Doctoral theses at NTNU, 2023:421

Håkon Hukkelås

Deep Generative Models for Realistic Image Anonymization

Doctoral thesis

NTNU
Norwegian University of Science and Technology
Thesis for the Degree of Philosophiae Doctor
Faculty of Information Technology and Electrical Engineering
Department of Computer Science

NTNU
Norwegian University of
Science and Technology

Håkon Hukkelås

Deep Generative Models for Realistic Image Anonymization

Thesis for the Degree of Philosophiae Doctor

Trondheim, December 2023

Norwegian University of Science and Technology
Faculty of Information Technology and Electrical Engineering
Department of Computer Science

NTNU
Norwegian University of
Science and Technology

NTNU
Norwegian University of Science and Technology

Thesis for the Degree of Philosophiae Doctor

Faculty of Information Technology and Electrical Engineering
Department of Computer Science

ISBN 978-82-326-7542-5 (printed ver.)
ISBN 978-82-326-7541-8 (electronic ver.)
ISSN 1503-8181 (printed ver.)
ISSN 2703-8084 (online ver.)

Doctoral theses at NTNU, 2023:421

Printed by NTNU Grafisk senter

Abstract

The following pages explore the use of generative models for realistic image anonymization. In summary, this thesis aims to address two primary objectives. First, develop generative models for synthesizing human figures for the purpose of anonymization. Secondly, evaluate the impact of anonymization on the development of computer vision algorithms.

This thesis culminates into four key contributions. First, it introduces DeepPrivacy, an open-source framework for realistic anonymization of human faces and bodies. DeepPrivacy is the first framework to effectively handle the challenges of in-the-wild image anonymization, such as handling overlapping objects, partial bodies, and extreme poses. Secondly, a variety of Generative Adversarial Networks (GANs) are proposed for synthesizing realistic human figures. To the best of our knowledge, the proposed GANs are the first to synthesize human figures in-the-wild effectively. The third contribution comprises of two open-source datasets, namely Flickr Diverse Faces (FDF) and Flickr Diverse Humans (FDH). Unlike previous datasets, FDF and FDH are large-scale and diverse datasets consisting of unfiltered images that capture the complexities of realistic image anonymization. Finally, the thesis presents an empirical evaluation of DeepPrivacy and compare it to traditional anonymization. Specifically, the impact of anonymization is evaluated for training computer vision models, with a focus on autonomous vehicle settings.

This thesis demonstrates that realistic anonymization is a superior alternative to traditional methods and a promising method to replace privacy-sensitive data with artificial data. We are confident that our open-source framework and datasets will be highly useful for practitioners and researchers seeking to anonymize their visual data.

Structure of Thesis

This thesis is a collection of papers organized into two parts.

Part I provides an overview of the research contributions, covers the relevant background, and discusses the contributions.

Part II includes the published papers. For improved readability, the format of the papers has been altered from their original published forms, but their content remains unchanged. Supplementary material for each paper is accessible online via links in Part II. The six papers included in this thesis are listed below.

Paper A. **DeepPrivacy: A Generative Adversarial Network for Face Anonymization**
Håkon Hukkelås, Rudolf Mester, Frank Lindseth
14th International Symposium on Visual Computing, 2019
Won best paper award

Paper B. **Image Inpainting with Learnable Feature Imputation**
Håkon Hukkelås, Frank Lindseth, Rudolf Mester
42nd DAGM German Conference on Pattern Recognition, 2020

Paper C. **Realistic Full-Body Anonymization with Surface-Guided GANs**
Håkon Hukkelås, Morten Smebye, Rudolf Mester, Frank Lindseth
IEEE/CVF Winter Conference on Applications of Computer Vision, 2023

Paper D. **DeepPrivacy2: Towards Realistic Full-Body Anonymization**
Håkon Hukkelås, Frank Lindseth
IEEE/CVF Winter Conference on Applications of Computer Vision, 2023

Paper E. **Synthesizing Anyone, Anywhere, in Any Pose**
Håkon Hukkelås, Frank Lindseth
IEEE/CVF Winter Conference on Applications of Computer Vision, 2024

Paper F. **Does Image Anonymization Impact Computer Vision Training?**
Håkon Hukkelås, Frank Lindseth
CVPR Workshop on Autonomous Driving, 2023

Contents

I Research Overview

II Publications

Acknowledgements

First, I would like to express my gratitude to my supervisors, Frank Lindseth and Rudolf Mester, for their guidance and encouragement. This thesis would not be possible without their support, especially without Frank's unexpected offer of a Ph.D. position early in my master's studies. Furthermore, thanks to all my great colleagues for making the fourth floor a great workplace over the last couple of years! A special thanks goes to Johannes and Bart for sharing their LaTeX files for writing this thesis. You saved me countless hours of frustration.

Furthermore, I am grateful to all my friends and family that has supported, distracted, and given me a life outside of research through my 9 years in Trondheim. Finally, my most heartfelt thanks goes to Ingvild for her support and patience during the last years of my Ph.D. You have made this period of my life much more enjoyable!

Part I

Research Overview

Chapter 1

Introduction

Collecting and storing images is ubiquitous in our modern society, everywhere from communication to the development of advanced autonomous agents acting freely in the world. However, such collection raises concerns regarding the individual's right to privacy, as visual data is rich in privacy-sensitive information (*e.g.* persons, documents, license plates). Recent privacy legislation (*e.g.* GDPR (Council of European Union, 2016)) restricts the collection of personal data, requiring entities to collect consent from recorded individuals or anonymize the data. For some domains, collecting consent from all individuals is infeasible (*e.g.* recording a crowded street), leaving anonymization as the only option for preserving privacy rights. The prominent anonymization technique, anonymization by obfuscation (*e.g.* blurring), distorts the data, possibly reducing its utility for its intended purpose. For example, learning a car to detect blurred people will not perform well in real-world driving scenarios and could pose a serious threat to human safety. This can be viewed as a barrier to research and development, especially for the data-dependent field of computer vision.

For the computer vision community, collecting task-specific datasets has become critical in tailoring models for their intended purpose. However, collecting such datasets can pose challenges when they involve sensitive data, as anonymization can lead to a degradation in data quality. Current computer vision algorithms are not designed to handle visual artifacts from anonymization, and they assume access to undistorted datasets [1]. Consequently, if anonymization by obfuscation is the way to comply with privacy laws, it is likely that the model's overall performance will suffer as a result of the reduced data quality and therefore data utility. This introduces the need for *realistic image anonymization*.

[1] The leading computer vision datasets employ no form of anonymization. Some datasets anonymize faces by obfuscation (listed in Section 2.5.1). However, the general use of anonymization for computer vision development is limited.

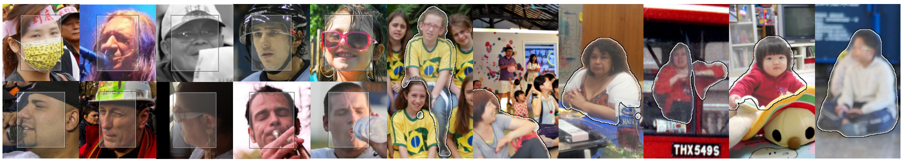


Figure 1.1: In-the-wild synthesis is difficult, where the generative model has to handle challenging cases such as partial bodies, occlusions, image distortions, and more. Images from the FDF and FDH datasets presented in Paper A and D.

1.1 What is Realistic Image Anonymization?

Before delving deeper into this thesis, the question arises; *what is realistic image anonymization?* Realistic image anonymization aims to replace privacy-sensitive information with semantically equivalent information suited for the application. For autonomous vehicles, the goal might be to replace people in the image with a synthesized realistic-looking identity. The term "utility-preserving anonymization" (Gross *et al.*, 2006a) is frequently used and similar in meaning, but it specifies the anonymization process dependent on the intended use of the data. For example, anonymization by obfuscation is utility-preserving for many tasks, as the realism of the data is irrelevant in many cases. A shortcoming of the term "realistic anonymization" with respect to this thesis is its focus on realism without considering distribution preservation. For example, anonymization can be done via content removal (Uittenbogaard *et al.*, 2019) (*e.g.* replacing persons with their background), which is equally realistic. However, this drastically alters the distribution of the data, which might reduce its utility for some computer vision tasks. Thus, for this thesis, "realistic anonymization" refers to techniques that aim to: 1) replace privacy-sensitive information with application-based semantically equivalent information, and 2) retain the data distribution to preserve data utility.

1.2 The Key Difficulties of In-the-Wild Synthesis

Generative models (*i.e.* models that can generate new instances from a data distribution) designed for anonymization are required to handle the complexities of in-the-wild synthesis. However, typical generative models do not address these challenges. The majority of previous research focuses on curated datasets [2], where "poor quality" images are filtered out to improve synthesis quality. For example, recent datasets for full-body synthesis filter out overlapping, occluded, or partial bodies (Fruhstuck *et al.*, 2022; Fu *et al.*, 2022). However, handling these challenging cases is a requirement for realistic anonymization. If an image is blurred or if a body is partially visible, the anonymization model should effectively handle this. Thus, we refer to these challenges as the *key difficulties of in-the-wild synthesis*, which include overlapping/partial bodies, occlusions, complex backgrounds, extreme poses, distorted images, and more.

1.3 Research Goals

This thesis explores the use of deep generative models for realistic anonymization of human figures, focusing on the following three goals.

> Research Goal 1
>
> Explore the use of generative models for realistic replacement of *faces* in images.

This research goal aims to develop a generative model for synthesizing realistic faces to replace the original identity. Generative models are known to generate close-to-photorealistic faces unconditionally (Karras *et al.*, 2018). However, their use for realistic image anonymization is under-explored. Previous research (Sun *et al.*, 2018a,b) has indicated that Generative Adversarial Networks (GANs) (Goodfellow *et al.*, 2014) can realistically anonymize faces. Nevertheless, no open-source tool exists for realistic image anonymization, and the aforementioned studies show few qualitative examples of complex scenarios, such as difficult poses or occluding objects. Thus, its practical use in real-world scenarios is restricted. Considering this, research goal 1 focuses on developing a generative model for realistic face anonymization that can handle the complexities of in-the-wild synthesis.

[2] For example, CelebA-HQ, FFHQ (Karras *et al.*, 2018, 2019), AFHQ (Choi *et al.*, 2020).

Research Goal 2

Explore the use of generative models for realistic replacement of *human figures* in images.

The human body is recognizable even though the face is anonymized through other identifying attributes (Wilber *et al.*, 2016; Lander *et al.*, 2001; McPherson *et al.*, 2016), such as gait (Jain *et al.*, 2008), clothing (Gallagher and Chen, 2008), and body appearance (Zhang *et al.*, 2015a; Oh *et al.*, 2016). Therefore, research goal 2 aims to develop a generative model for synthesizing realistic human figures to replace full-body humans in images.

Synthesizing full-body human figures is a much more challenging task than face synthesis, as the human body is a deformable surface that interacts with complex objects in the world. Previous work often focus on simpler tasks, such as transferring a known appearance into a given pose (Chan *et al.*, 2019; Balakrishnan *et al.*, 2018), transferring garments (Han *et al.*, 2018; Sarkar *et al.*, 2020), or full-body synthesis into a plain background (Fruhstuck *et al.*, 2022). These studies often disregard the key difficulties of in-the-wild synthesis. As far as we know, only a few methods are suitable for full-body anonymization (Ma *et al.*, 2018; Maximov *et al.*, 2020; Ma *et al.*, 2017). However, these methods focus primarily on full bodies (not partial bodies) in very low resolution.

Research Goal 3

Evaluate the impact of anonymization on the development of computer vision algorithms.

The impact of data anonymization on training computer vision models is under-explored. Previous work study the effect of face anonymization for classification (Yang *et al.*, 2022b), semantic segmentation (Geyer *et al.*, 2020; Zhou and Beyerer, 2022), object detection (Dvořáček and Hurtik, 2022), action recognition (Tomei *et al.*, 2021), and face detection (Klomp *et al.*, 2021). These studies find that realistic face anonymization is more effective in utility preservation for computer vision development (Zhou and Beyerer, 2022; Dvořáček and Hurtik, 2022). Still, the impact of anonymization is unclear for key tasks such as instance segmentation and human pose estimation. Furthermore, the impact of full-body anonymization is not explored in the current literature.

Chapter 2

Background

This chapter introduces the underlying background knowledge to understand the context of this thesis' contributions. A basic understanding of machine learning theory, neural networks, and their applications for computer vision is recommended to comprehend the following material.

2.1 Generative Adversarial Nets

Generative Adversarial Networks (GANs) (Goodfellow *et al.*, 2014) is a generative model that learns to sample from the data distribution by creating a competitive adversarial game between a generator and a discriminator. Commonly, these adversaries are modeled as deep neural nets. The task of the generator is to sample artificial data points (*e.g.* images resembling faces) from random noise while the discriminator tries to distinguish real examples from generated ones. In essence, the generator can be viewed as an "art forger" that tries to convince the "police" (discriminator) that the artificial images are real. In this way, by competing over several thousands of examples, the generator learns to generate more and more realistic examples (Figure 2.1). Formally, the adversarial objective is given by,

$$\min_G \max_D \mathbb{E}_{x \sim p_{data}}[\log D(x)] + \mathbb{E}_{z \sim p_z}[\log(1 - D(G(z)))], \tag{2.1}$$

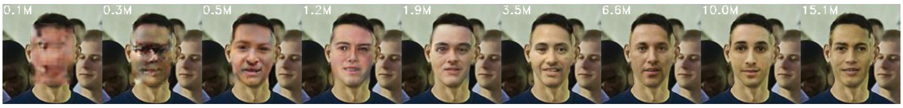


Figure 2.1: The figure shows generated images from the method in Paper A during training. The number in the top left corner is the number of images that the generator has trained on (in millions). Note the progressive improvement of image quality during training.

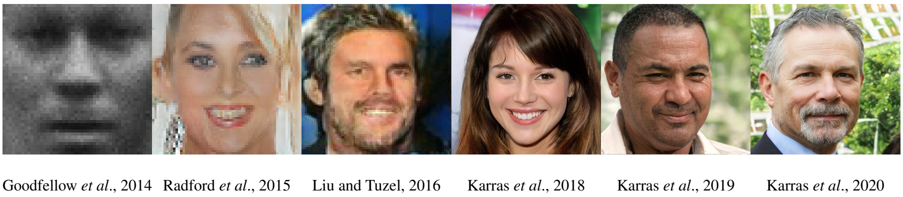


Figure 2.2: 6 years of GAN progress. Figure inspiration: `https://twitter.com/goodfellow_ian/status/1084973596236144640`.

where $\boldsymbol{z} \in \mathbb{R}^d$ is a latent vector with size d, drawn from a random noise distribution p_z (*e.g.* $p_z \in \mathcal{N}(0,1)$), and $\boldsymbol{x}$ is a sample drawn from the real data distribution p_{data}. Note that Equation (2.1) can be extended to a conditional GAN (Mirza and Osindero, 2014) by including conditional information to D and G.

GANs are notoriously difficult to train, and a notable research focus is placed on ensuring stable training of the generator. Since their conception, they have evolved from generating low-resolution grayscale images to becoming the leading image generation method (Figure 2.2). Arguably, the majority of these advances have focused on basic engineering efforts, where "hand-designed" network architectures (Karras *et al.*, 2019, 2020; Isola *et al.*, 2017) combined with a well-designed training strategy (Karras *et al.*, 2018; Sauer *et al.*, 2022; Karnewar and Wang, 2020) can improve results significantly. Furthermore, a range of "tricks" have emerged in recent years to improve training stability, such as minibatch discrimination (Salimans *et al.*, 2016), exponential moving averages (Chandrasekhar *et al.*, 2018), gradient/epsilon penalties for the discriminator (Mescheder *et al.*, 2018; Gulrajani *et al.*, 2017; Karras *et al.*, 2018), specialized Adam parameters (Radford *et al.*, 2015; Kingma and Ba, 2015), and different learning rates for the generator and discriminator (Heusel *et al.*, 2017). The major limitation of current GANs is their inability to handle complex multi-modal distributions (*e.g.* ImageNet (Deng *et al.*, 2009)), where scaling GANs up to handle these complex distributions results in unstable training (Sauer *et al.*, 2022). Recently, these issues can be diminished by employing pre-trained feature networks for the discriminator (Sauer *et al.*, 2022).

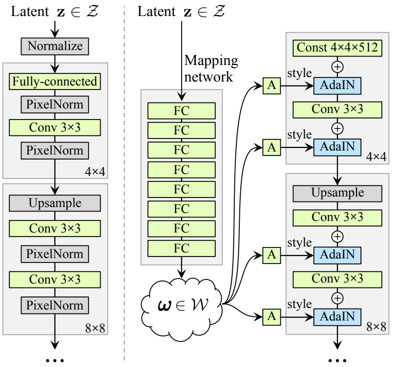


Figure 2.3: Comparison between the traditional GAN architecture (**left**) and the StyleGAN architecture (**right**) (Karras *et al.*, 2019). The traditional generator inputs the latent code exclusively to the input layer. StyleGAN maps the latent code (z) to an intermediate representation ω, which controls the generator through Adaptive Instance Normalization (Huang and Belongie, 2017) at every layer. Here "A" is a learned linear transformation. Original figure simplified from Figure 1 in (Karras *et al.*, 2019).

Style-based GANs The StyleGAN (Karras *et al.*, 2019, 2020) family of architectures has been the leading architecture for GANs, and its techniques are employed in Paper C-E. The significant advance of StyleGAN lies in how the latent code (z) is injected into the generator, illustrated in Figure 2.3. The traditional GAN inputs z *exclusively* to the input layer, whereas StyleGAN inputs z at multiple resolutions throughout the generator. StyleGAN maps the latent code z to an intermediate representation ω through a non-linear fully connected network. This allows the generator to learn a disentangled representation [1] in ω which is not required to follow any fixed distribution (*i.e.* as p_z). Furthermore, the multi-resolution input of ω throughout the generator allows StyleGAN to control different factors at different resolutions.

StyleGAN significantly improved image quality compared to previous work, but the major advance was (arguably) the editability of the generated images.

[1] Here "disentangled representation" refers to ω being divided into linear subspaces controlling different factors of variations. For example, when you write a text in a text editor, you have multiple factors of variations to edit the appearance of the text (*e.g.* font size or font color). If you change the font size, the color won't change as the factors of variations are disentangled. If font size and color were entangled, changing the color would also change the font size.

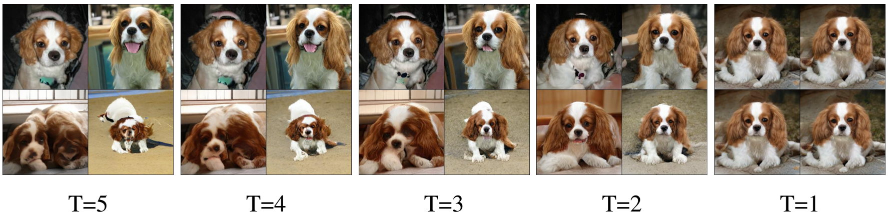


Figure 2.4: The Truncation Trick. Image diversity can be traded off for image quality by truncating the sampled latent towards the mode of the distribution. Here, truncation is done by resampling all values falling above the threshold T until they fall within the range [-T, T]. Original figure source: Figure 2 in Brock *et al.* (2019).

Due to the multi-resolution input of ω, StyleGAN allows for controlling specific attributes at different resolutions. For example, it can mix styles between generated images at different resolutions, where coarse styles (*i.e.* ω inputted at low-resolution layers) represent "coarse features," *e.g.* structural information such as head position/rotation. Finer styles (*i.e.* ω inputted at high resolution) represent "finer features," such as hair color. For a visualization, see Figure 3 in Karras *et al.* (2019). Finally, this style-based generator can edit attributes given text prompts (Kocasari *et al.*, 2022), which Paper D extends to conditional GANs.

The Truncation Trick Given a GAN trained with $p_z \sim \mathcal{N}(0,1)$, Brock *et al.* (2019) proposed the truncation trick. The truncation trick can trade off the generated diversity for final image quality (Figure 2.4) by truncating the sampled latent towards the mode of the distribution (0 for p_z). Originally, Brock *et al.* (2019) proposed to resample all values falling above a given threshold until they fall in the range. The truncation trick was later extended for StyleGAN (Karras *et al.*, 2019), where truncation can be done in ω, by linearly interpolating a sampled ω to the approximated mode of ω. Recently, the truncation trick was extended to multi-modal truncation (Mokady *et al.*, 2022), enabling sampling of high-quality images while minimizing the loss of diversity. Instead of approximating a single mode of the distribution ω, multi-modal truncation estimates a set of cluster centers by employing KMeans clustering. Here, sampled latent codes are truncated toward their closest cluster centers.

2.2 Evaluating Generative Models

This section thoroughly covers the evaluation metrics employed in Paper A-E. For generative models, there are three evaluation criteria of interest; image quality, diversity, and disentanglement, where the latter often correlates with the editability of the generated images. Note that image quality and diversity are subjective measurements, and no perfect metric exists to assess this accurately.

Distribution Similarity Distribution similarity evaluates image quality and diversity by estimating the similarity of the generated data distribution to the original data distribution. The prominent metrics are Fréchet Inception Distance (FID) (Heusel *et al.*, 2017) and Inception Score (IS) (Salimans *et al.*, 2016). All papers in this thesis focus on FID, as the use of IS for purposes other than ImageNet generation is questionable (Barratt and Sharma, 2018).

Fréchet Inception Distance (FID) (Heusel *et al.*, 2017) estimates the distribution similarity by comparing the features from generated/real images embedded with an Inception Network (Szegedy *et al.*, 2016). The similarity is estimated with the Frèchet distance (Dowson and Landau, 1982), which compares the mean and covariance of the generated/real images. The mean and covariance are estimated from the features of a number of generated/real images (generally, 50K images each).

FID is far from a perfect metric, as it assumes that the features from an Inception Network align with human judgment of image quality. However, FID is known to often disagree with human judgment, and it is shown that FID is sensitive to the presence of ImageNet objects in the image (Kynkäänniemi *et al.*, 2022). For example, for face synthesis on FFHQ (Karras *et al.*, 2019), FID focuses primarily outside the face region and score faces containing ImageNet classes better (*e.g.* bow tie, sunglasses). Kynkäänniemi *et al.* (2022) recommends that FID improvements should be verified using non-ImageNet features (*e.g.* CLIP (Radford *et al.*, 2021) or uninitialized networks (Naeem *et al.*, 2020)). Furthermore, they recommend that extra care should be exercised when using pre-trained networks for training GANs (Sauer *et al.*, 2022). Note that the method in Paper E employs pre-trained networks for the discriminator.

Image Similarity & Diversity Learned Perceptual Image Patch Similarity (LPIPS) (Zhang *et al.*, 2018) estimates the similarity between two images using features from convolutional networks. Specifically, LPIPS computes a weighted sum of l_2 distances from multi-resolution features extracted from a VGG16 network (Simonyan and Zisserman, 2014). Here, the weights are fitted such that LPIPS aligns with perceptual human judgment. Zhang *et al.* (2018) shows that LPIPS align better with human judgment of perceptual similarity than traditional metrics (*e.g.* PSNR). Similarly, LPIPS can measure the diversity of generated images for conditional GANs (Zhu *et al.*, 2017). Conditional GANs can suffer from mode collapse given a condition where the GAN generates the same image for different latents. Zhu *et al.* (2017) propose to estimate the sample diversity for a given condition by measuring the LPIPS distance for different points in the latent space p_z.

Perceptual Path Length The ability of the generator to disentangle the latent space correlates with the stability and consistency of generated shapes (Karras *et al.*, 2020). Intuitively, linear interpolation in the latent space should produce linear interpolations in the image space. For example, the linear interpolation between a woman with "black hair" and "blond hair" should not produce "red hair." If so, it indicates that the latent space is entangled, and factors of variations are not separated into linear subspaces.

Perceptual Path Length (PPL) (Karras *et al.*, 2019) approximates the disentanglement of the latent space z by computing the LPIPS distance between two close latent points. A close latent point to $z_1 \sim p_z$ is found by randomly sampling $z_2 \sim p_z$, then linearly interpolating $z_{\text{close}} = z_1 + \varepsilon \cdot z_2$ (generally $\varepsilon = 10^{-4}$). If the latent space is disentangled, the perceptual difference (LPIPS) of the generated images should be minimal.

2.3 Full-Body Synthesis

The research field of full-body synthesis has a range of applications with a large variation of high-level goals. This section categorizes human synthesis into the following two categories: transfer-based and synthesis-based models. *Transfer-based methods* transfers a source appearance (or garment (Han *et al.*, 2018; Sarkar *et al.*, 2020)) into a **new pose** (Balakrishnan *et al.*, 2018; Li *et al.*, 2019b; Sarkar *et al.*, 2020; Pumarola *et al.*, 2018; Ma *et al.*, 2017; Si *et al.*, 2018), **motion** (Chan *et al.*, 2019), or **scene** (Siarohin *et al.*, 2018). While some of these methods are applicable for in-the-wild human figure synthesis (Yang *et al.*, 2022a; Siarohin *et al.*, 2018), they require a source appearance that limits the synthesized identities to a texture bank or image dataset of appearances. For the latter goal, *synthesis-based methods* can synthesize the appearance either conditioned on a **pose** (Song *et al.*, 2021; Ma *et al.*, 2018; Yang *et al.*), a **scene** (Esser and Sutter, 2018), or **unconditionally** (Fruhstuck *et al.*, 2022; Chaudhuri *et al.*, 2021; Fu *et al.*, 2022). Furthermore, some methods focus on the reverse task of synthesis, reconstruction of the 3D surface, and texture (Natsume *et al.*, 2019; Saito *et al.*, 2019; Weng *et al.*, 2020). These reconstructions can later be rendered to the scene given a camera view (Weng *et al.*, 2020).

Independent of the goal, most methods use a form of pose information to improve synthesis quality through **dense pose annotations** (Sarkar *et al.*, 2020; Neverova *et al.*, 2018; Yang *et al.*, 2022a), **semantic segmentations** (Song *et al.*, 2021; Chaudhuri *et al.*, 2021; Yang *et al.*, 2022a), **sparse keypoints** (Han *et al.*, 2018; Balakrishnan *et al.*, 2018; Li *et al.*, 2019b; Pumarola *et al.*, 2018; Ma *et al.*, 2017; Si *et al.*, 2018; Chan *et al.*, 2019; Siarohin *et al.*, 2018; Ma *et al.*, 2018; Esser and Sutter, 2018), or a **3D pose of the body** (Lassner *et al.*, 2017; Yang *et al.*).

The primary limitation of the aforementioned studies w.r.t. full-body anonymization is the lack of handling in-the-wild synthesis. Most of these studies disregard the key difficulties of in-the-wild-synthesis, such as overlapping objects, partial bodies, complex backgrounds, and extreme poses. Recent studies filter out these difficult cases from their dataset to improve synthesis quality (Fruhstuck *et al.*, 2022; Fu *et al.*, 2022). Note that several studies (Song *et al.*, 2021; Ma *et al.*, 2018; Maximov *et al.*, 2020; Ma *et al.*, 2017; Esser and Sutter, 2018) perform experiments on the Market1501 dataset (Zheng *et al.*, 2015), which includes bodies in a large variety of poses and different backgrounds. However, the Market1501 dataset consists primarily of full bodies (not partial) with few occluding objects.

2.4 Image Anonymization

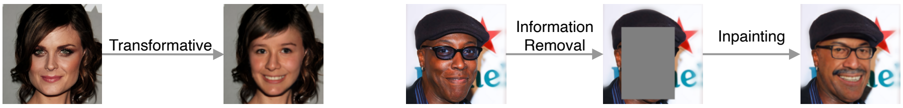
The goal of image anonymization is to remove any privacy-sensitive information contained in the image. Image anonymization can be categorized into *traditional anonymization* and *realistic anonymization*. Traditional anonymization is widely adopted in practice, where methods anonymize the image via obfuscation (*e.g*. blurring or masking), encryption (He *et al.*, 2016a), or k-means (Gross *et al.*, 2006b; Jourabloo *et al.*, 2015; Newton *et al.*, 2005). Often, these methods are sufficient to protect privacy; however, they degrade the quality of the data reducing its utility for downstream tasks.

This section focuses on methods targeting realistic anonymization of human figures in images. Note that several papers focus on anonymization of other objects, such as license plates (Kacmaz *et al.*, 2021), documents (Orekondy *et al.*, 2018), or medical images (Kim *et al.*, 2021). Furthermore, some methods focus on predicting the privacy-sensitive parts of the image (Gupta *et al.*, 2021). Finally, it is worth mentioning that some papers investigate the use of adversarial attacks to insert noise invisible to the human eye that can impede face recognition models from accurately identifying individuals (Oh *et al.*, 2017). While these methods do preserve the realism of the data, they do not anonymize the data, as the biometric information remains present in the image.

2.4.1 Realistic Image Anonymization

The goal of realistic image anonymization is to remove any privacy-sensitive information from the original image while generating realistic images that retain the utility of the data. Preserving utility depends on the task that the data is collected for. For example, collecting data for classroom studies can require the retention of specific attributes (*e.g*. facial expressions). In comparison, collecting data for autonomous vehicles has softer requirements for utility preservation, where the main requirement is the realism of the generated data.

Current methods in the literature provide different guarantees with respect to privacy and utility preservation. In the following section, the literature is cate-



(a) (b)

Figure 2.5: **(a)** Tranformative anonymization observes the original image and anonymizes the identity by altering privacy-sensitive attributes. **(b)** Inpainting-based anonymization separates anonymization into information removal and inpainting of missing regions.

gorized into two different methodologies: *anonymization by transformation* and *anonymization by inpainting*, illustrated in Figure 2.5. This section focuses on prominent anonymization techniques that provide different trade-offs between privacy guarantee and utility preservation. Note that all models in Paper A-E anonymizes by inpainting.

Anonymization by Transformation Anonymization by transformation refers to methods that observe the original image and transform it to remove privacy-sensitive information (Figure 2.5a). Transformative anonymization provides no formal guarantee of privacy as a "black box" model is responsible for removing privacy-sensitive information. Therefore, transformative methods require quantitative experiments for validating that the anonymization model can confuse both human and machine evaluators (Ren *et al.*, 2018; Gafni *et al.*, 2019).

Transformative anonymization yields high utility preservation, where current models can preserve non-identifying attributes (*e.g.* facial hair). For example, Ren *et al.* (2018) proposes a model to anonymize the identity while preserving the performed action in a video. Similarly, Gafni *et al.* (2019) proposes a model that removes privacy-sensitive attributes while preserving all other attributes. These models (Wu *et al.*, 2019; Ren *et al.*, 2018; Gafni *et al.*, 2019) learns privacy-sensitive attributes empirically based on what attributes a face recognition system uses for identification. Recent methods explore face swapping for anonymization, where the original face is swapped with a new face dissimilar to the original (Ciftci *et al.*, 2023). The aforementioned methods empirically show that they confuse humans (Gafni *et al.*, 2019; Ren *et al.*, 2018) and machine evaluators (Gafni *et al.*, 2019; Ren *et al.*, 2018; Ciftci *et al.*, 2023). However, empirical validation provides no formal guarantee of anonymization.

Anonymization by Inpainting Anonymization by inpainting masks out the original identity before generating a new identity. Therefore, inpainting-based methods never observes the original identity unless the identity is recognizable outside the masked-out region. As a result, inpainting-based methods provide stronger privacy guarantees than transformative methods, as the identity is only recognizable when the detection system fails. However, current inpainting-based techniques often yield poorer utility preservation than transformative-based methods.

Sun *et al.* (2018a) propose an inpainting-based model for head obfuscation (not only the face region), where their model is guided on 68 facial landmarks. They later extended this with a parametric model for face anonymization (Sun *et al.*, 2018b). Both of these models can retain the pose of the face (given the 68 facial landmarks), where Sun *et al.* (2018b) can retain other non-identifying attributes. Similarly, CIAGAN (Maximov *et al.*, 2020) can retain the pose given facial landmarks, while also specifying which identity to synthesize [2]. The major limitation of the aforementioned methods is the dataset used to train the models. Sun *et al.* (2018a,b) use a filtered version of the PIPA dataset (Zhang *et al.*, 2015b), where extreme poses are removed. Similarly, CIAGAN (Maximov *et al.*, 2020) uses the CelebA dataset (Liu *et al.*, 2015), which has a limited diversity in ethnicity, extreme poses, ages, *etc*. Therefore, these models struggle with in-the-wild anonymization.

2.5 Anonymized Data in Computer Vision

A key motivation of this this thesis is to use anonymized data for computer vision development. Thus, the following section summarizes the current use of anonymized data for computer vision development. In addition, this section covers previous work that explores the impact of anonymization on developing computer vision models.

[2] Identity selection is based on a pre-defined set of different identities, where the authors use a set of 10K unique identities.

2.5.1 Public Anonymized Datasets

Most computer vision datasets employ no form of anonymization with only a few exceptions. The literature survey in Paper F found five prominent datasets that employed anonymization for computer vision development. NuScenes (Caesar *et al.*, 2020) contain images from vehicles driving in Singapore and Boston, where faces and license plates are anonymized via blurring. A2D2 (Geyer *et al.*, 2020) includes data from southern Germany, where license plates and heads are blurred to comply with German privacy regulations. AViD (Piergiovanni and Ryoo, 2020) is a video dataset for action recognition, where heads are blurred. P3M (Li *et al.*, 2021) is a portrait matting dataset, where every face is blurred. Uittenbogaard *et al.* (2019) propose a dataset containing street view scenes, where cars and pedestrians are removed via content removal using image inpainting.

2.5.2 Anonymization and Its Impact on Computer Vision

There exists a limited set of studies exploring the effect that anonymization has on training computer vision models. For ImageNet training (Deng *et al.*, 2009), face obfuscation (blurring) has little effect on top-5 accuracy, and no impact on feature transferability to scene recognition, object localization, or face attribute classification (Yang *et al.*, 2022b). Nevertheless, anonymization slightly degrades accuracy in classes appearing together with faces (*e.g.* facial masks). For autonomous vehicle datasets, some studies find that face obfuscation degrades instance segmentation on Cityscapes (Cordts *et al.*, 2016; Zhou and Beyerer, 2022). In contrast, Dvořáček and Hurtik (2022) finds little impact of face anonymization on object detection on the same dataset. Geyer *et al.* (2020) finds that face obfuscation does not affect instance segmentation on A2D2. For action recognition, face obfuscation significantly degrades performance (Tomei *et al.*, 2021), where the authors propose a teacher-student self-distillation framework to mitigate the degradation. Klomp *et al.* (2021) finds that realistic anonymization performs substantially better than traditional methods for training face detectors.

Other studies focus on the effect that anonymization has on evaluation. For example, Wilber *et al.* (2016) presents a black-box study of Facebook's face detection model, and finds that the model is robust to severe face obfuscation.

Finally, some studies focus on the human perspective. For example, Hasan *et al.* (2018) systematically studies how anonymization affects the user's perceived utility of the anonymized image. Similarly, Li *et al.* (2017) evaluates the human perception of different anonymization techniques w.r.t. image satisfaction, information sufficiency, enjoyment, and social presence.

Chapter 3

Research Contributions

This thesis presents the iterative development of DeepPrivacy. This chapter first present and contextualizes each paper to previous work, followed by a summarization of the main contributions. Finally, the DeepPrivacy framework is described in detail. The following papers are included in this thesis:

Paper A. **DeepPrivacy: A Generative Adversarial Network for Face Anonymization**
Håkon Hukkelås, Rudolf Mester, Frank Lindseth
14th International Symposium on Visual Computing, 2019
Won best paper award

Paper B. **Image Inpainting with Learnable Feature Imputation**
Håkon Hukkelås, Frank Lindseth, Rudolf Mester
42nd DAGM German Conference on Pattern Recognition, 2020

Paper C. **Realistic Full-Body Anonymization with Surface-Guided GANs**
Håkon Hukkelås, Morten Smebye, Rudolf Mester, Frank Lindseth
IEEE/CVF Winter Conference on Applications of Computer Vision, 2023

Paper D. **DeepPrivacy2: Towards Realistic Full-Body Anonymization**
Håkon Hukkelås, Frank Lindseth
IEEE/CVF Winter Conference on Applications of Computer Vision, 2023

Paper E. **Synthesizing Anyone, Anywhere, in Any Pose**
Håkon Hukkelås, Frank Lindseth
IEEE/CVF Winter Conference on Applications of Computer Vision, 2024

Paper F. **Does Image Anonymization Impact Computer Vision Training?**
Håkon Hukkelås, Frank Lindseth
CVPR Workshop on Autonomous Driving, 2023

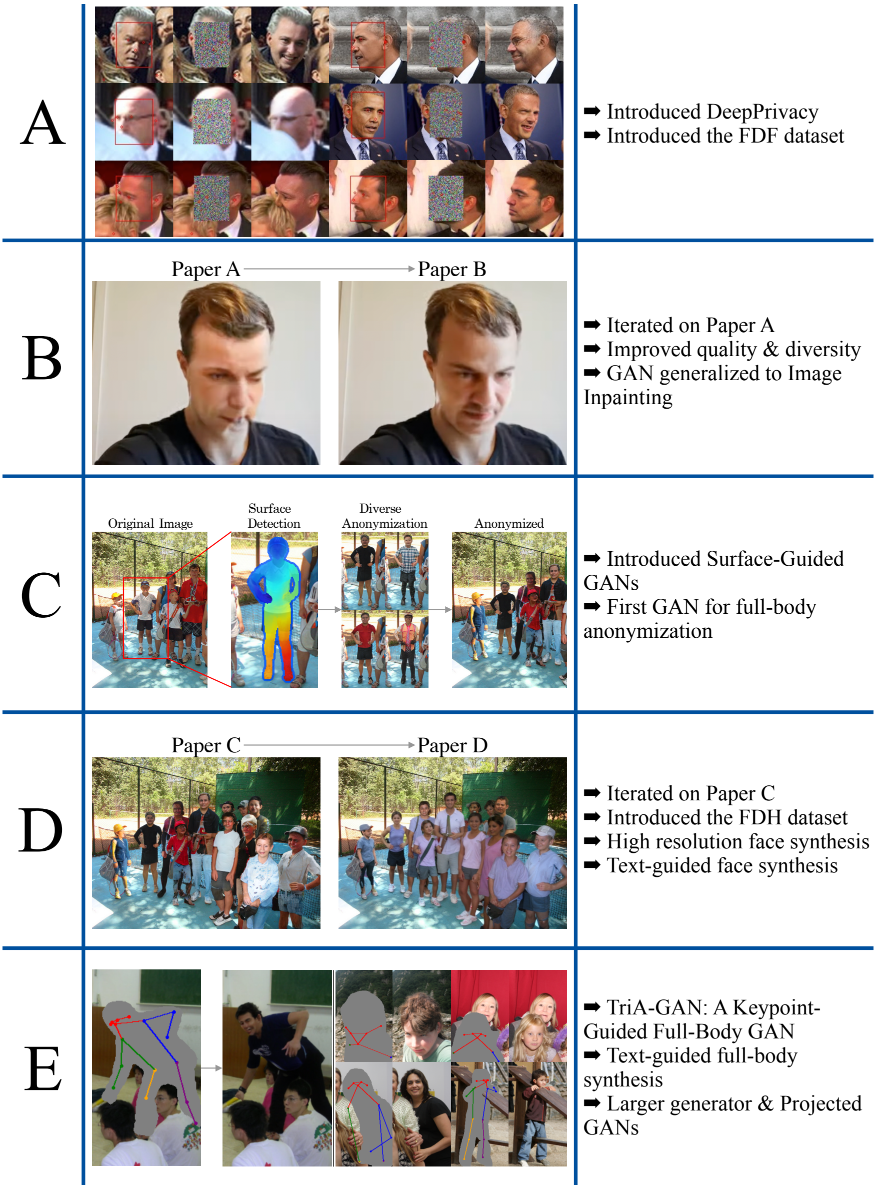

A
➡ Introduced DeepPrivacy
➡ Introduced the FDF dataset
B
Paper A
Paper B
➡ Iterated on Paper A
➡ Improved quality & diversity
➡ GAN generalized to Image Inpainting
C
Original Image
Surface Detection
Diverse Anonymization
Anonymized
➡ Introduced Surface-Guided GANs
➡ First GAN for full-body anonymization
D
Paper C
Paper D
➡ Iterated on Paper C
➡ Introduced the FDH dataset
➡ High resolution face synthesis
➡ Text-guided face synthesis
E
➡ TriA-GAN: A Keypoint-Guided Full-Body GAN
➡ Text-guided full-body synthesis
➡ Larger generator & Projected GANs

Figure 3.1: A brief overview of papers A-E, showcasing the improvements of each paper for anonymizing faces and full bodies.

3.1 Summary of Thesis Papers

Paper A-E present the iterative development of DeepPrivacy (Figure 3.1) and Paper F evaluates the impact of anonymization for training computer vision models. A brief summary of each paper is given here.

Paper A - DeepPrivacy: A Generative Adversarial Network for Face Anonymization

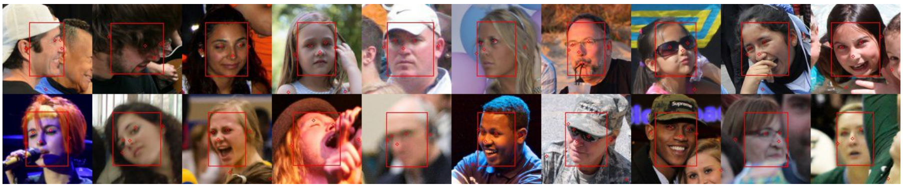


Figure 3.2: The FDF dataset. Each image is annotated with 7 keypoints and a tight bounding box indicating the face region.

This paper introduced the DeepPrivacy framework for face anonymization. It formulated anonymization as an inpainting task, where a Conditional Generative Adversarial Network (C-GAN) (Mirza and Osindero, 2014) fills in a square missing region in the face, illustrated in Figure 3.1. To stabilize training of the C-GAN, DeepPrivacy adapted techniques from Karras *et al.* (2018), such as progressive growing and discriminator regularization. Furthermore, the paper introduced the Flickr Diverse Faces (FDF) dataset (Figure 3.2), which is a large and diverse dataset of human faces, including unconventional poses, occluded faces, and a vast variability in backgrounds. In comparison to previous realistic anonymization techniques (Sun *et al.*, 2018a,b), the main improvement of DeepPrivacy was the FDF dataset. Previous papers had focused on smaller and filtered datasets with a small diversity of extreme poses (covered in detail in Section 2.4). As an illustration, training DeepPrivacy on the Celeb-A dataset (used in Maximov *et al.* (2020)) causes severe artifacts for simple head rotations [1].

[1] Video comparing FDF training vs Celeb-A: https://youtu.be/k-SpRVc6nOc.

Paper B - Image Inpainting with Learnable Feature Imputation

This paper iterated on the method from Paper A. The key improvements were a revised gradient penalty specialized for the inpainting task and the replacement of progressive growing with Multi-Scale Gradient GANs (Karnewar and Wang, 2020). These contributions significantly improved synthesis quality. Notably, the new model achieved a better FID score than the model in Paper A while only using 10% of the parameters. The paper evaluated the method on general image inpainting on the Places2 dataset (Zhou *et al.*, 2017) and face inpainting. In addition, the method was competitive with state-of-the-art for general image inpainting at the time.

This paper improved many of the failure cases from Paper A (discussed in Section A-6), demonstrated in the following videos (Figure 3.3).

youtu.be/nOJVqgvGwkU

youtu.be/K8n-Ck0YHxc

Figure 3.3: Videos comparing the method in Paper B to Paper A. Note the significant improvement in image quality, stability of generated identities, and synthesis quality for extreme poses. Note that the label "DeepPrivacyV2" in the videos refers to the method in Paper B, not the one in Paper D.

Paper C - Realistic Full-Body Anonymization with Surface-Guided GANs

Paper A-B demonstrate that GANs can generate close-to-photorealistic faces to anonymize individuals. However, the human body is often recognizable from many other cues in the image other than the face. Therefore, this paper addresses the full-body anonymization task, which was an under-explored task at the time.

This paper introduces Surface-Guided GANs, which condition the generator on dense pixel-to-surface correspondences between the image and a canonical 3D surface (T-shaped 3D body). Key to the method is Variational Surface-Adaptive Modulation (V-SAM) which embeds surface information throughout the generator. Combining this with the proposed discriminator surface supervision loss, the generator can synthesize high-quality humans with diverse appearances in complex and varying scenes. This method showed promising results for full-body anonymization, but it often generated human figures containing visually annoying artifacts. The key limitation of this method was the small dataset (40K images from COCO (Lin *et al.*, 2014)), where the discriminator overfitted early in training.

Paper D - DeepPrivacy2: Towards Realistic Full-Body Anonymization

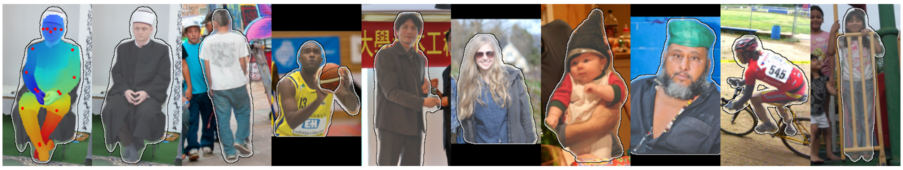


Figure 3.4: Examples from the FDH dataset. Each image is annotated with keypoints, pixel-to-vertex correspondences (from CSE (Neverova *et al.*, 2020)), and a segmentation mask. The leftmost image shows annotations for the first image.

This paper iterated on the method from Paper C. In summary, the key improvements can be summarized into the following four points. First, the paper introduces the Flickr Diverse Humans (FDH) dataset (Figure 3.4). The FDH dataset consists of 1.87M images, where each image includes a single human figure as the subject. Note that the same image can contain several individuals. This removed the issue of overfitting from Paper C, substantially improving generated image quality. Secondly, it introduced an updated version of the FDF dataset (FDF256), consisting of higher resolution images (256×256 *vs.* 128×128). Thirdly, the paper adopted a StyleGAN architecture

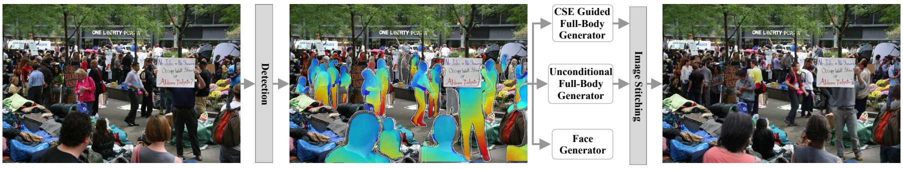

Detection
CSE Guided Full-Body Generator
Unconditional Full-Body Generator
Face Generator
Image Stitching

Figure 3.5: DeepPrivacy2 supports multi-modal anonymization, where three detection and synthesis networks are employed: (1) a CSE-guided generator for individuals detected with dense pose (by CSE (Neverova *et al.*, 2020)), (2) an unconditional full-body generator for cases where CSE fails to detect (note the segmented persons without color-coded CSE detections), and (3) a face generator for the remaining individuals (marked in red). The original image is from Wider-Face (Yang *et al.*, 2016).

enabling attribute-guided anonymization via text prompts. Finally, it improved the anonymization pipeline, including stitching of generated images into the original image and support for multi-modal anonymization (Figure 3.5) [2].

Paper E - Synthesizing Anyone, Anywhere, in Any Pose

This paper explored full-body synthesis conditioned on sparse 2D-keypoints, eliminating the need for expensive dense pose annotations. The primary limitation of the method introduced in Paper D is its reliance on dense pose estimation. Detecting dense pose correspondences can be challenging, particularly for long-range detection, which is common in autonomous vehicles. In addition, the available datasets with such annotations are either limited in size (Guler *et al.*, 2018) or automatically annotated (*e.g.* the FDH dataset). Replacing dense pose correspondences with keypoints increases the modeling complexity considerably, as the generative model must now infer both the body's texture *and* its structure. This paper addresses the challenge of scaling up GANs to handle in-the-wild full-body synthesis without dense pose correspondences.

This paper introduces TriA-GAN, a keypoint-guided GAN that can synthesize Anyone, Anywhere, in Any given pose. The advances of TriA-GAN can be

[2] Multi-modal anonymization refers to using different generators for different detection types. For example, some people are detected by full-body segmentation models, whereas others are only detected by a face detector.

summarized into the three following points. First, TriA-GAN replaces the conventional GAN discriminator with Projected GANs (Sauer *et al.*, 2021) which employ pre-trained feature networks to discriminate images. The paper thoroughly evaluates different pre-trained networks and finds that the previously used classification networks (Sauer *et al.*, 2021, 2022) are poorly suited for discriminating human figures. Instead, TriA-GAN uses a combination of self-supervised feature networks for the discriminator, which significantly improves sample quality. Furthermore, the paper introduces a progressive training scheme for U-nets (Ronneberger *et al.*, 2015), enabling TriA-GAN to easily scale up to higher resolutions and large model sizes. Finally, the paper demonstrates that TriA-GAN can be used with unconditional editing methods for GANs, enabling text-guided synthesis for human figures.

Paper F - Does Image Anonymization Impact Computer Vision Training?

Paper A-E focus on the development of generative models for anonymization. In contrast, Paper F studies the impact of image anonymization on the training of computer vision models. Note that Paper A, C, D presented smaller experiments exploring the impact of anonymization. However, they rely on automatic detection of regions to anonymize, which raises questions about whether the performance degradation is due to detection errors or the anonymization model.

This paper explores the impact of image anonymization on the Cityscapes (Cordts *et al.*, 2016), BDD100k (Yu *et al.*, 2020), and COCO (Lin *et al.*, 2014) datasets. Specifically, the paper benchmarks traditional and realistic anonymization techniques for faces and bodies that are implemented in DeepPrivacy. The findings in the paper can be summarized into the following. First, traditional image anonymization substantially impacts final model performance, particularly when anonymizing the full body. Secondly, realistic anonymization can mitigate this decrease in performance, where the presented experiments reflect a minimal performance drop for face anonymization. The paper concludes that realistic anonymization can enable privacy-preserving computer vision development with minimal performance degradation in some settings. However, the experiments reflect that realistic image anonymization still is far from being a perfect substitute to the original data, and it highlights several limitations of current methods. Chapter 4 further discuss these limitations.

3.2 Primary Contributions

Building upon the iterative development presented in the previous section, the question arises: *what are the main contributions of this thesis?* The presented papers culminate into four primary contributions.

Primary Contribution 1

The DeepPrivacy Anonymization Framework

To the best of our knowledge, Paper A presented the first open-source framework for realistic anonymization. Note that previous studies had introduced closed-source anonymization frameworks previously (Sun *et al.*, 2018a; Gafni *et al.*, 2019). We consider the framework presented in Paper A (later improved in Paper B-E) to be a significant contribution for practitioners who need to anonymize images while preserving their realism. Today, the framework is continuously used and has garnered over 1300 stars on GitHub, with 50-100 downloads per month (as of March 2023).

Primary Contribution 2

Generative Models for Face and Full-Body Synthesis In-the-Wild

Paper A-E all introduced novel methods for handling in-the-wild anonymization, summarized into the following. Paper A introduced the first generative model for face anonymization that could handle the difficulties of in-the-wild synthesis. Furthermore, to the best of our knowledge, Paper C was the first method to address in-the-wild full-body anonymization, and Paper D introduced the first model to generate nearly photorealistic human figures. The final model, TriA-GAN (Paper E), is the current state-of-the-art for synthesizing human figures in-the-wild.

Primary Contribution 3

Large-Scale Anonymization Datasets

Paper A introduced the FDF dataset for face synthesis and Paper D introduced the FDH dataset for full-body synthesis. In contrast to previously used datasets for faces (*e.g.* CelebA, FFHQ) and bodies (*e.g.* DeepFashion, Market1501), the FDF/FDH datasets represent the difficulties of in-the-wild synthesis. Previous datasets often filtered out these cases, such as removing partial subjects, blurred images, extreme poses, and occluded subjects. Paper A, D showed that the large-scale datasets substantially improved performance, reflecting that such a diverse and large dataset was necessary to tackle in-the-wild synthesis.

Primary Contribution 4

Quantitative Analysis of the Impact of Image Anonymization

The impact of image anonymization for computer vision development is under-explored, which Paper F address. The literature review in Paper F found two unanswered questions w.r.t. the use of anonymized data for training computer vision models. First, is realistic anonymization more effective in retaining the utility of images compared to traditional methods? Secondly, to what extent does full-body anonymization impact the training of computer vision models? The former question was previously addressed for specific tasks and datasets (discussed in Section 2.5.2). The latter was unanswered, where Paper F was the first to address it.

3.3 The DeepPrivacy Framework

The contributions of Paper A-E culminate into the open-source framework DeepPrivacy [3]. The framework includes a range of generative models for realistic anonymization of human faces and bodies (listed in Table 3.1). Additionally, it supports traditional obfuscation techniques (blurring, masking, pixelation).

[3] First open-sourced at `https://github.com/hukkelas/DeepPrivacy` and later improved in `https://github.com/hukkelas/deep_privacy2`.

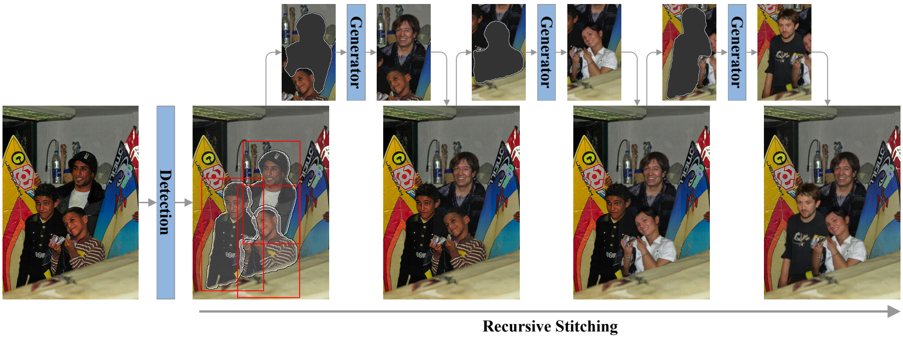


Figure 3.6: DeepPrivacy anonymizes recursively, where one instance is synthesized at a time and pasted into the original image. Note that the generator may rely on additional information, such as surface information or keypoints, which are not depicted here.

This section gives a brief overview of the three stages of DeepPrivacy shown in Figure 3.6; detection, synthesis, and image stitching.

DeepPrivacy employs instance-wise generative models that synthesize one individual (face or body) at a time. The motivation for instance-wise synthesis is threefold; first, it is simpler than synthesizing multiple individuals at the same time. Secondly, it allows for explicit instance-wise editability. Thirdly, it is easy to process high-resolution images (*e.g.* 2048 × 1024 for Cityscapes anonymization (Cordts *et al.*, 2016)). However, instance-wise synthesis requires stitching the generated individuals into the original image, which can introduce visual artifacts, particularly when individuals overlap. This issue is further discussed in Section 3.3.3.

Recommended demo We recommend the reader to try out the framework with our website demos on Hugginface. Note that the demos are open-source if you want to test them on your local machine.

Face Anonymization:
huggingface.co/spaces/haakohu/deep_privacy2_face.

Full-Body Anonymization:
huggingface.co/spaces/haakohu/deep_privacy2.

Modality	Resolution	Detection Type	Synthesis Method
Face	128×128	Face bounding box + 7 Keypoints	Paper B*
Face	128×128	Face bounding box	Paper D
Face	256×256	Face bounding box	Paper D
Full-body	288×160	Segmentation mask	Paper D
Full-body	288×160	DensePose + Segmentation mask	Paper D
Full-body	288×160	Segmentation mask + 17 Keypoints	Paper E

Table 3.1: An overview of the different generative models provided in DeepPrivacy. "DensePose" detection refers to surface maps from Continuous Surface Embeddings (Neverova *et al.*, 2020). The 7 facial keypoints follow the COCO format and are the shoulder, ears, and head keypoints. The 17 keypoints for full-body are all keypoints following the COCO format. * Models from Paper B are not possible to train in the current framework, but the weights are ported from the source code of Paper B.

3.3.1 Detecting Human Figures

The DeepPrivacy framework separates detection into face detection via bounding boxes, and full-body detection via segmentation masks. Some synthesis models require auxiliary information, such as facial landmarks, DensePose estimations, or full-body keypoints. Paper D describes in detail the different detection networks used for each modality. Note that the framework supports processing multiple detection modalities at the same time, allowing anonymization of faces in cases where the full body is not detected. Kalman filtering is used to track all detections for video processing (using motpy (Muron, 2022)), allowing for a single latent variable to be assigned to each identity. This improves temporal consistency between frames [4]. Note that the generative model processes frame by frame and does not include any modeling choices to ensure temporal consistency.

3.3.2 TriA-GAN - Synthesizing Human Figures

Paper A-E all introduce novel methods for synthesizing faces or full bodies. This section does not aim to provide an extensive description of each network, as they are already described in detail in their respective papers. Rather, it

[4] Video demo of tracking: `https://youtu.be/Kt3au719hhk`.

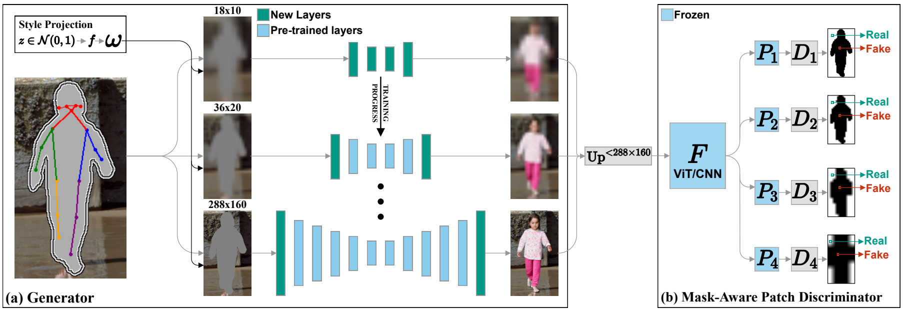


Figure 3.7: (a) TriA-GAN fills in the missing region given a masked-out image conditioned on 17 keypoints. The generator layers employ adaptive instance normalization (Huang and Belongie, 2017) to condition the generator on ω, where ω is the output of the style mapping network. TriA-GAN is trained progressively starting at 18×10 resolution, then increased by adding layers to the start/end of the encoder/decoder. (b) Each feature network F use four shallow patch discriminators operating on its features (with different spatial resolutions), where each feature is projected through random differentiable operations (P_1-P_4). Given the projected features, each discriminator predicts if a given patch corresponds to a real or fake image region.

offers a brief summary of TriA-GAN (Figure 3.7) introduced in Paper E, with an aim to provide the reader with a general understanding of its capabilities and limitations. Note that TriA-GAN is a general method for synthesizing faces or bodies, and conditioning the model on other factors than keypoints (used in Paper E), such as surface maps, is straightforward.

The TriA-GAN Generator The generator is a StyleGAN-based (Karras *et al.*, 2019) U-Net architecture (Ronneberger *et al.*, 2015), first introduced in Paper D and slightly revised in Paper E. Each layer in the encoder/decoder consists of a set of residual blocks, where layers with matching resolutions are connected between the encoder and decoder. Note that the decoder employs output skip-connections at every resolution to improve training stability, following Karnewar and Wang (2020). The latent code (z) is injected via adaptive instance normalization (Huang and Belongie, 2017), by first mapping it to ω with a style mapping network (described in Section 2.1). The generator is progressively trained, as described in Paper E.

Injecting the latent code with adaptive instance normalization results in a

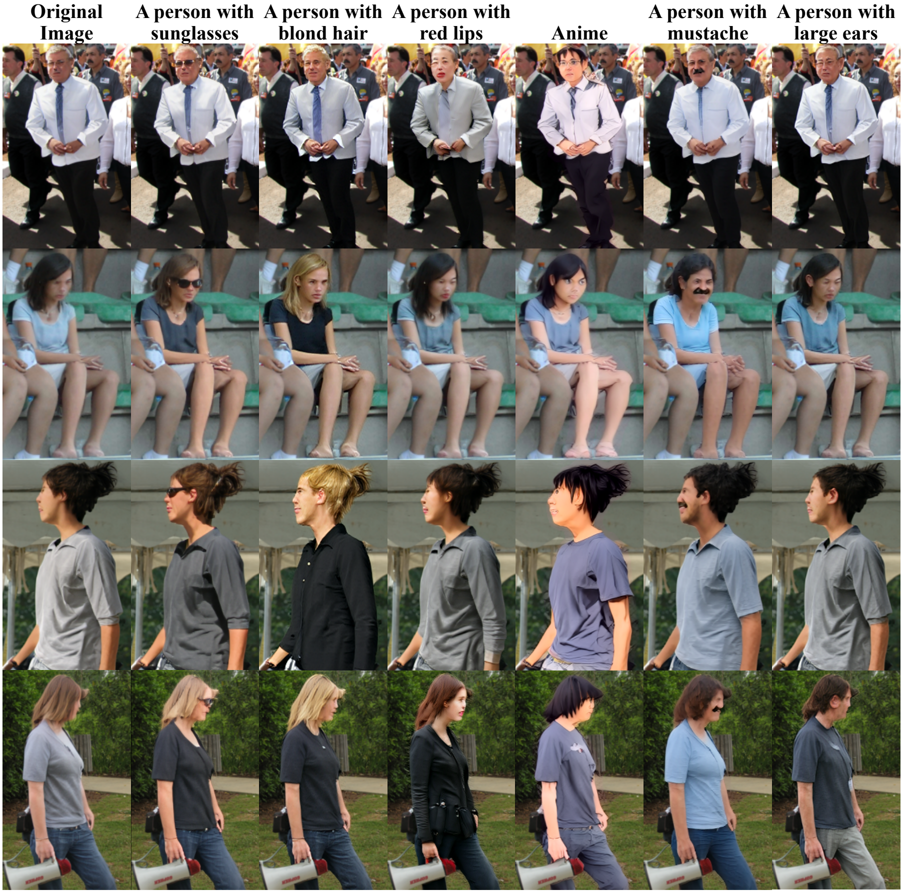


Figure 3.8: StyleMC (Kocasari *et al.*, 2022) edits with TriA-GAN, where a global direction (from the text prompt above each column) is added to the style code of the original (leftmost) image.

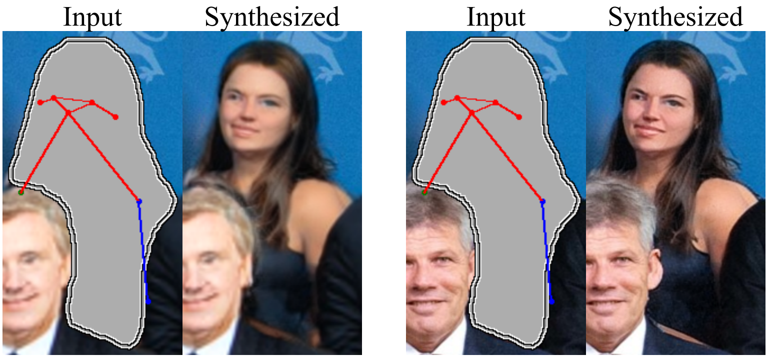
disentangled latent space which is easy to edit. Figure 3.8 demonstrate that StyleMC (Kocasari *et al.*, 2022) is effective with TriA-GAN to find global semantic directions in the GAN latent space. StyleMC finds global directions by manipulating random images towards a text prompt using a CLIP encoder (Radford *et al.*, 2021).

The TriA-GAN Discriminator The major advance of TriA-GAN was the adoption of Projected GANs (Sauer *et al.*, 2021) for the discriminator. TriA-GAN uses two feature networks to discriminate human figures, specifically ResNet-50 (He *et al.*, 2016b) with CLIP pre-trained weights (Radford *et al.*, 2021) and ViT-L/16 (Dosovitskiy *et al.*, 2021) pre-trained with masked autoencoding (He *et al.*, 2022). Paper E demonstrate that the combination of these networks is well-suited to discriminate human figures and significantly improves over previously used classification networks (Sauer *et al.*, 2022, 2021). Furthermore, the use of Projected GANs significantly simplified the adversarial objective. Paper A-D combined the GAN objective with regularizing objectives to stabilize training, such as gradient penalties (Gulrajani *et al.*, 2017) and epsilon penalties (Karras *et al.*, 2018). In contrast, TriA-GAN exclusively optimizes the adversarial objective.

3.3.3 Stitching it All Up

The final stage is stitching the synthesized identities into the final image. If not handled correctly, the stitching process can generate visually annoying artifacts, especially in regions with overlapping instances. Paper A-C adopts a naive approach by synthesizing all identities first and then stitching them in. However, in cases of overlapping detections, the resulting synthesized identities have sharp boundaries at the points of overlap. This occurs because the boundaries transition smoothly to the original boundary rather than to the other overlapping synthesized person. To compensate for this, Paper D introduces a recursive stitching process.

Recursive Stitching Recursive stitching involves synthesizing individual instances one at a time and then incorporating them into the image before



(a) Descending Order **(b)** Ascending Order

Figure 3.9: The order of synthesis matters. DeepPrivacy synthesizes in ascending order w.r.t. the number of pixels a person covers. Note that the reverse order (**a**) can lead to identities that appear "blurred". This occurs when the generator observes upsampling artifacts in the input, as higher resolution images are synthesized first (note that the person in the foreground appears blurred in (a)).

moving on to the next instance. In this way, the generative model handles overlapping artifacts when generating each individual. For recursive stitching, it is crucial to consider the order in which synthesis is conducted to achieve optimal image quality. Paper D proposes to synthesize individuals depending on the number of pixels a person covers in ascending order. The motivation for this ordering is twofold. First, the ordering assumes that objects in the foreground cover a larger area, where foreground objects are stitched in last. The reverse order (foreground objects first) results in background objects ”overwriting” foreground objects, as the detections can overlap. This naive assumption that foreground objects cover a larger area is not always true, however, it is a straightforward estimation. Secondly, by stitching in higher-resolution identities last minimizes the possibility of introducing upsampling artifacts when the original resolution is larger than the resolution of the generative model. These artifacts are commonly not visually annoying, but the generative model can react to such artifacts, as demonstrated in Figure 3.9.

3.4 Further Contributions

Technical Contributions In addition to the presented contributions, a range of libraries was made open-source for the community.

- *A High-Performance Pytorch Implementation of face detection models, including RetinaFace and DSFD*, `https://github.com/hukkelas/DSFD-Pytorch-Inference`.

  A library containing efficient, lightweight, and state-of-the-art face detection models in Pytorch and experimental ports to TensorRT.

- *High-performance Keypoint-Mask RCNN Models*, `https://github.com/hukkelas/keypoint_mask_rcnn`.

  A library containing pre-trained, efficient, and high-performing Mask R-CNN models for keypoint and instance segmentation of human figures.

Other Contributions Concurrently with this work, the Ph.D. work has contributed to other scientific contributions.

- *Realistic Image Anonymisation*.
  Håkon Hukkelås and Frank Lindseth.
  In van der Sloot, B. and van Schendel, S. *The boundaries of data: Technical, practical and regulatory perspectives*.
  Accepted in Amsterdam University Press.

- *DeepPrivacy: A Framework for Realistic Image Anonymization*.
  Håkon Hukkelås.
  Presented demo at NorwAI Innovate 2022.
  Won best demo award.

- *Autonomous Vehicle Control: End-to-end Learning in Simulated Environments*.
  Hege Haavaldsen, Max Aasboe, Håkon Hukkelås, Frank Lindseth.
  Norsk IKT-konferanse for forskning og utdanning 2019.

- *Deep Active Learning for Autonomous Perception*.
  Navjot Singh, Håkon Hukkelås, Frank Lindseth.
  Norsk IKT-konferanse for forskning og utdanning 2020.

Chapter 4

Discussion and Conclusion

This chapter evaluates how the papers address the research goals and examine any present limitations that may provide opportunities for future research.

4.1 Synthesis Limitations

The section commences with a discussion of the capabilities and limitations of current synthesis methods for realistic image anonymization.

4.1.1 Face Synthesis

Research Goal 1

Explore the use of generative models for realistic replacement of *faces* in images.

Paper A,B,D address research goal 1. Paper A proposed an open-source generative model for realistic face anonymization, which Paper B,D further improved. Furthermore, Paper A introduced the FDF dataset, which Paper D extended for higher-resolution face anonymization. The face generative models presented in Paper A,B,D reflect a remarkable improvement in synthesis quality and the ability to handle challenging settings. The final model synthesizes nearly photorealistic human faces and allows manipulation of specific attributes through user-given text prompts (*e.g.* eye color, see Paper D). However, there is a range of limitations to the current model. The model in Paper A had significant issues for all the limitations listed below, whereas subsequent iterations (Paper B,D) demonstrated progressive improvement. Considering this trend, we expect that more advanced generative models will further mitigate these issues.

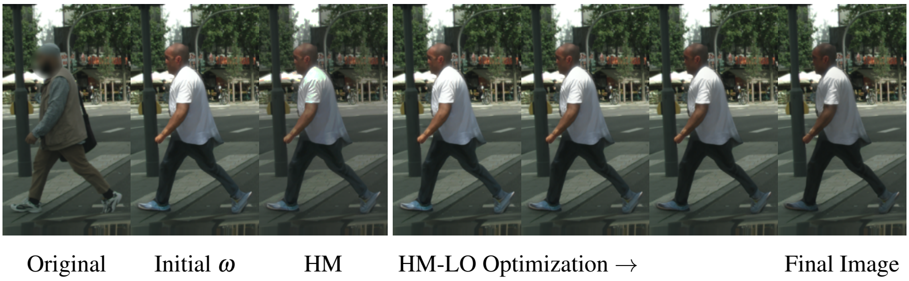


Figure 4.1: The global context mismatch problem refers to the issue where the synthesized identity may not align with the global context of the image, making the synthesized identity "stick out". Note that the illumination of the synthesized identity ("initial ω") is different from the original identity. Paper F explore two options to address this issue: naive histogram matching (**HM**), and histogram matching via latent optimization (**HM-LO**), which iteratively adjusts the initial ω to better fit the histogram of the original image (in HSV). See Paper F for further details.

Limitation 1: Temporal Consistency

This thesis does not address the modeling of temporal consistency. However, the proposed model produces identities that are somewhat consistent over time. The sole design choice to enhance temporal consistency involves tracking and sampling the identical latent variable for each individual (described in Section 3.3.1). Although this enhances consistency, the current approach does not provide temporal smoothness, and the identity can vary across different contexts or head rotations. To address these challenges, it may be useful to draw insights from analogous tasks in video generation. For example, by introducing temporal blocks into the generative model (Skorokhodov *et al.*, 2022) or eliminating specific points in the pipeline that contribute to temporal inconsistencies (Tzaban *et al.*, 2022).

Limitation 2: Occluding Objects

Occluding objects can be extremely challenging to handle, especially complex objects covering the face (*e.g.* hands). A potential solution to this issue is more fine-grained masking techniques (Kirillov *et al.*, 2020) or more robust generative models that are better at synthesizing the borders of the occluding object.

Limitation 3: Extreme Poses

Extreme poses are difficult to handle. Similar to Limitation 2, better generative models have demonstrated progressive improvement in handling extreme poses. Furthermore, having more detailed pose information (*e.g.* 67 landmarks used in Maximov *et al.* (2020)) could further improve the model. However, obtaining a dataset containing such annotations is time-consuming, as current detection models often struggle with extreme/unusual poses.

4.1.2 Full-body Synthesis

Research Goal 2

Explore the use of generative models for realistic replacement of *human figures* in images.

Paper C,D,E all address research goal 2. Paper C introduced the first generative model for in-the-wild full-body anonymization, and Paper D,E improved upon this method. Furthermore, Paper D introduced the FDH dataset, which significantly improved synthesis quality compared to the dataset used in Paper C. Synthesizing full bodies in the wild is an exceptionally demanding task, yet Paper C,D,E indicate a noteworthy advancement in terms of image quality. Naturally, the full-body generative models exhibit all the limitations previously mentioned for face synthesis. In addition, full-body synthesis highlights further limitations that become more apparent due to the complexity of the task.

Limitation 4: Global Context Mismatch

A pressing limitation for full-body synthesis is the *global context mismatch problem*. This issue originates from instance-wise cropping, where a synthesized identity matches the local context given to the generative model, but not to the global context. This is due to the instance-wise cropping removing critical scene factors from the perspective of the generative model. For example, the lighting of a scene might only be observable in the area outside the crop. Paper F proposed two naive approaches to address this issue (Figure 4.1), which involve matching the histogram of the synthesized image to the original. Furthermore, note that this issue persists for handling factors that are

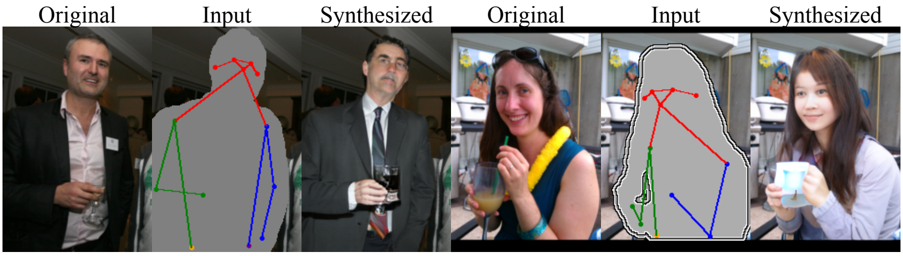


Figure 4.2: The generator struggles to synthesize realistic objects when the person appears to be holding something.

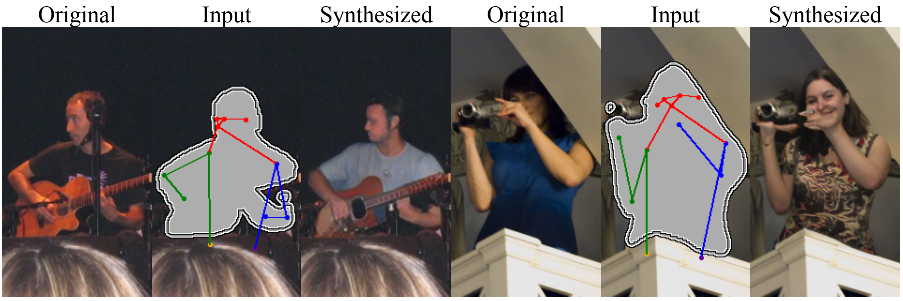


Figure 4.3: The generator struggles to synthesize realistic interactions with objects on the boundary of the missing region.

partially visible in the instance crop. For example, handling lens distortions or transparency (*e.g.* standing in a window).

Limitation 5: Interacting Objects

Humans interact with objects constantly, and generating realistic interactions where both the human and the object look natural can be extremely challenging (Figure 4.2 and 4.3). The models in Paper D,E are often able to handle interactions with objects on the boundary (*e.g.* bicycles or chairs). However, it struggles with complex objects (*e.g.* guitars), or in cases where the pose indicates that the human is carrying something (*e.g.* a mug). Similar to Limitation 2, this can be alleviated with more fine-grained masking or a stronger generative model.

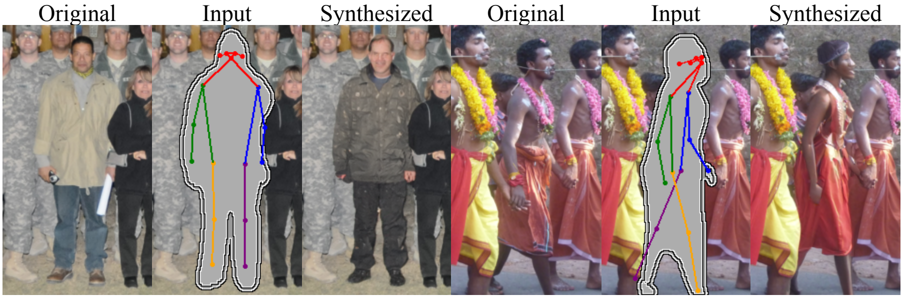


Figure 4.4: The generator samples from a small variation of possible identities/appearances for certain contexts.

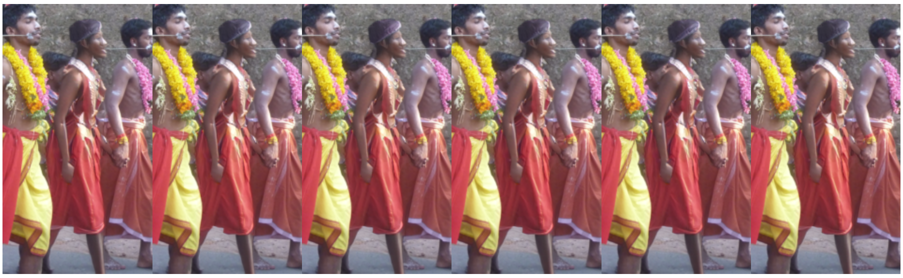
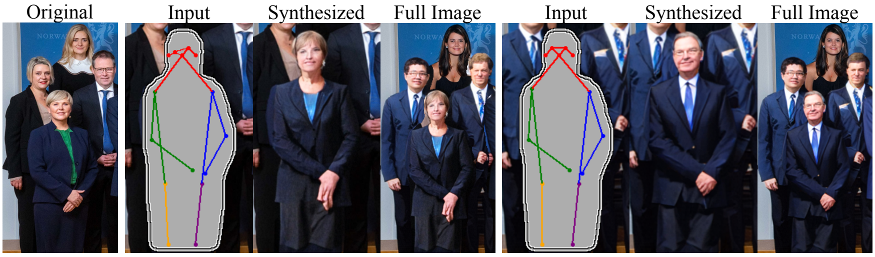


Figure 4.5: Synthesized identities for 8 random latent variables. Note that the generator samples near-identical appearances for certain contexts. Figure 4.4 illustrates the original image.

Limitation 6: Condition Dependency

The generated identity and appearance are highly dependent on the context, which narrows the sampling space of the generator (Figure 4.4 and 4.5). For example, on football fields, the generator samples primarily players wearing sports uniforms. Similarly, if a person in the background wears a suit, the sampled person will likely wear a similar outfit. This feature often results in more realistic images, but it can pose certain problems. It severely limits temporal consistency, where changes in the background or the pose can drastically alter the appearance. Furthermore, this condition dependency narrows the sampling space, often resulting in a very limited diversity of appearances for some contexts.

A potential solution to this condition dependency is to disentangle the pose,



(a) Not Recursive Synthesis **(b)** Recursive Synthesis

Figure 4.6: The recursive synthesis impacts the synthesized identity as the input of the generator is drastically changed. Note that severe artifacts are inserted at the border of overlapping bodies if synthesis is not done recursively (a).

background, and appearance into separate factors (Ma *et al.*, 2018). However, current methods require paired datasets, which are limited in dataset size and variation of background/identities.

Limitation 7: Overlapping Bodies

Overlapping bodies are particularly difficult to handle, due to the recursive synthesis used by DeepPrivacy. The recursive synthesis inserts sampling artifacts when the cropped image does not match the resolution of the generator, as discussed in Section 3.3.3. The generator subsequently reacts to these artifacts, potentially leading to significant distortions, particularly in the case of high-resolution images. Furthermore, the recursive synthesis alters the input of the generator for adjacent identities, causing the generated identity to vary depending on the order of synthesis. As demonstrated in Figure 4.6, when directly pasting identities into the image rather than recursively synthesizing them, the identity changes drastically.

4.1.3 General Limitations

Limitation 8: Utility-Privacy Trade-Off

The utility-privacy trade-off problem refers to the issue where improved image utility compromises the guarantee of privacy. In many cases, improving utility requires the retention of more attributes from the original person. For example, in a classroom study, retention of gestures or motion of an individual is essential to maintain image utility. However, this approach creates the potential for individuals to be recognized from motion. DeepPrivacy does not prioritize retaining specific attributes since it is not crucial for image recognition in autonomous vehicles. Nevertheless, note that previous work focus on the retention of detailed facial pose (Gafni *et al.*, 2019) or other attributes such as gender (Jourabloo *et al.*, 2015).

Limitation 9: Controlled Sampling

Paper D,E demonstrate that the generative model is controllable through user-given text prompts. This serves as a proof-of-concept for controlling synthesized identities, but it is not used in the current anonymization pipeline. Nonetheless, it presents exciting avenues for further research on realistic anonymization in autonomous vehicles. For example, ensuring that the synthesized demography matches that of the original data. As far as we know, this is an unexplored area of research.

4.2 Using Anonymized Data for Computer Vision

Research Goal 3

Evaluate the impact of anonymization on the development of computer vision algorithms.

Paper F evaluated the impact of training typical computer vision methods on anonymized data with a focus on autonomous vehicle datasets and tasks. Specifically, it benchmarked instance segmentation methods on Cityscapes (Cordts *et al.*, 2016) and BDD100K (Yu *et al.*, 2020), and pose estimation on COCO (Lin *et al.*, 2014). This section discusses limitations to the analysis in Paper F. See Section 3.1 for a summary of the paper's findings.

There are four primary limitations to the evaluation protocol in Paper F, which are summarized here. See Section 5.1 in Paper F for a more thorough discussion. First, the experiments relies on automatic annotations, which introduce ambiguity in the results. This raises the question of whether the current performance degradation is due to annotation errors or synthesis limitations.

Secondly, due to the filtering criteria presented in the paper, the anonymization model is not able to anonymize all individuals in the images. Thirdly, the analysis is restricted to models using ResNet (He *et al.*, 2016b) and R-CNN (Ren *et al.*, 2015). Other architectures are not explored (*e.g.* YOLO (Bochkovskiy *et al.*, 2020)), which might respond differently to anonymization artifacts. Finally, Paper F limits the analysis to image-based detection methods. For example, datasets requiring temporal consistency (*e.g.* tracking) or multi-view consistency are not studied, as such capabilities are not present in the current anonymization framework.

4.3 Privacy Considerations

Detection Limitations DeepPrivacy relies on a two-stage system; detection of privacy-sensitive regions and anonymization of the respective regions. Methods following this regime cannot guarantee the privacy of individuals without human validation, as current detection networks are far from perfect. Nevertheless, current state-of-the-art can detect most individuals, where up to 90% of all persons that take up a "large" portion of the image are recognized [1]. In terms of faces, state-of-the-art methods detect well above 90% of all faces in an image (Li *et al.*, 2019a). Furthermore, detection networks are vulnerable to adversarial attacks, where malicious actors can insert objects into the physical world that can prevent the detection model from recognizing individuals (Kurakin *et al.*, 2018b). However, there is currently a large focus in the community on developing defenses against such attacks (Kurakin *et al.*, 2018a).

Identity Leakage The identity of individuals can leak through other means of recognition. State-of-the-art methods focus primarily on face anonymization. However, as discussed previously, the human body is identifiable through other attributes than the face. The most pressing limitation of current anonymization techniques is gait recognition. The gait of a person is a behavioral biometric (Jain *et al.*, 2008), where the pattern of shape and motion of a walking person

[1] Following the top-ranked COCO (Lin *et al.*, 2014) object detection submission as of March 2022, where a "large" portion is regions larger than 96x96 pixels in the image.

is a discriminative feature for long-range video recognition. Furthermore, the GANs presented in Paper C-D employ dense surface information to generate the individual. This information is likely reproduced in the generated identity, which could make the individual recognizable through its surface map.

GDPR and the Limitations of DeepPrivacy The GDPR (Council of European Union, 2016) affects the ability of entities to collect and store data containing identifiable information ('biometric data') and personal data. Thus, entities are required to collect consent from recorded individuals or anonymize the data containing such information. This raises the question; what are anonymized images, and does the DeepPrivacy framework provide this? GDPR article 4.14 defines "biometric data" as:

> 'biometric data' means personal data resulting from specific technical processing relating to the physical, physiological or behavioural characteristics of a natural person, which allow or confirm the unique identification of that natural person, such as facial images or dactyloscopic data. (Council of European Union (2016), Article 4.14).

Article 4.14 specifies facial images as biometric data. However, it does not provide a clear answer regarding the full-body. In December 2021, the Belgian Data Protection Authority specified that 'biometric data' does cover behavioral characteristics, such as gait patterns (Brodahl *et al.*, 2022). Furthermore, GDPR regulates the processing of "personal data", defined as:

> 'personal data' means any information relating to an identified or identifiable natural person ('data subject'); an identifiable natural person is one who can be identified, directly or indirectly, in particular by reference to an identifier such as a name, an identification number, location data, an online identifier or to one or more factors specific to the physical, physiological, genetic, mental, economic, cultural or social identity of that natural person; (Council of European Union (2016), Article 4.1).

This article does not clearly define what is and is not personal data. By combining information (*e.g*. location, clothes, body shape), you might be able to identify who the information belongs to, which qualifies it as personal data.

With this in mind, does DeepPrivacy comply with GDPR? Before addressing this, note that the response to this question was provided by computer scientists who do not possess a law degree. As a result, it is recommended to approach the following answer with a degree of skepticism. As discussed previously, DeepPrivacy provides no formal guarantee of anonymization without a human in the loop to verify that all individuals are detected. On the assumption that all humans are detected, does DeepPrivacy comply with GDPR? In most cases, the face anonymizer of DeepPrivacy does not, as the human body is still identifiable from other cues than the face. For full-body anonymization, it depends on the context. Considering video data, DeepPrivacy does not adjust the gait in any way, thus gait patterns are likely to be similar to the original identity. Considering image data, the generator guided on keypoints provides no additional information compared to masking the identity out, except the location of the 17 keypoints of the human body. Thus, this provides similar privacy guarantees as masking the area out, and it is unlikely to reliably reproduce information to identify the original individual.

4.4 Ethical Considerations

Realistic anonymization focuses on synthesizing realistic humans, creating a potential for misuse. A typical example is the misuse of DeepFakes, where generative models can be used to create manipulated content to misinform. In contrast to realistic anonymization, typical DeepFake methods observe the original identity (Zakharov *et al.*, 2019) or perform computationally expensive finetuning on a specific individual (Thies *et al.*, 2016). Furthermore, there exist several solutions to mitigate the potential for misuse. The DeepFake Detection Challenge (Dolhansky *et al.*, 2020) has increased the ability of automatic models to detect manipulated content. In addition, pre-emptive solutions, such as model watermarking (Yu *et al.*, 2021) can embed a synthetic "fingerprint" on the image data to identify it as fake.

Similar to all learning-based generative models, the synthesized human figures adhere to the sampling probability of the dataset. For all generative models proposed in this thesis, the dataset originates from Flickr. Thus, these generators follows the biases from Flickr and is less likely to synthesize people from underrepresented groups on this website.

Finally, overfitting of the generative model can imply privacy risks to individuals in the training dataset. This privacy risk arises from the fact that the GAN has a higher probability of generating images of people from the training dataset compared to those who are not part of it. Nevertheless, recent studies have found that overfitting of GANs is minimal for face synthesis (Marriott *et al.*, 2020) and less prone to overfitting than diffusion models (Carlini *et al.*, 2023).

4.5 Conclusion

This thesis presents DeepPrivacy, the first open-source framework for realistic image anonymization of human figures and faces. The primary contributions include a variety of generative models and datasets for face and full-body synthesis. The proposed generative models are capable of addressing the challenges of in-the-wild synthesis that were previously unexplored. Furthermore, the presented findings indicate that realistic image anonymization is a superior alternative to traditional methods in cases where the realism of the data is essential. The experiments reflect that training computer vision models on traditionally anonymized data severely impact model performance, whereas realistic anonymization can mitigate this decrease. Nonetheless, realistic anonymization is not a complete substitute for real data, especially for full-body anonymization, as current generative models still struggle with complex scenarios. Moreover, the presented analysis of generative models for realistic anonymization and its impact on computer vision development has identified several exciting and challenging areas for future research. For example, handling multi-view and temporal consistency or ensuring that the synthesized demography matches that of the original data. Finally, it is worth noting that this research coincides with the ongoing generative model revolution, and the quality of synthesized human figures has significantly improved in recent years. Given the present trend, it is not far-fetched to presume that synthesized individuals will soon become a near-perfect alternative to the original data.

Bibliography

Guha Balakrishnan, Amy Zhao, Adrian V Dalca, Fredo Durand, and John Guttag. Synthesizing Images of Humans in Unseen Poses. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8340–8348. IEEE, jun 2018. ISBN 978-1-5386-6420-9. doi: 10.1109/CVPR.2018.00870.

Shane Barratt and Rishi Sharma. A Note on the Inception Score. *arXiv preprint arXiv:1801.01973*, 2018.

Alexey Bochkovskiy, Chien-Yao Wang, and Hong-Yuan Mark Liao. YOLOv4: Optimal Speed and Accuracy of Object Detection. *arXiv preprint arXiv:2004.10934*, 2020.

Andrew Brock, Jeff Donahue, and Karen Simonyan. Large Scale GAN Training for High Fidelity Natural Image Synthesis. In *International Conference on Learning Representations*, 2019.

Brodahl, Laura De Boel, and Joanna Juzak. Belgian data protection authority clarifies key rules on biometric data processing, jan 2022. URL `https://www.wsgrdataadvisor.com/2022/01/belgian-data-protection-authority-clarifies-key-rules-on-biometric-data-processing/#_ftnref1`.

Holger Caesar, Varun Bankiti, Alex H. Lang, Sourabh Vora, Venice Erin Liong, Qiang Xu, Anush Krishnan, Yu Pan, Giancarlo Baldan, and Oscar Beijbom. nuScenes: A Multimodal Dataset for Autonomous Driving. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 11618–11628. IEEE, jun 2020. ISBN 978-1-7281-7168-5. doi: 10.1109/CVPR42600.2020.01164.

Nicholas Carlini, Jamie Hayes, Milad Nasr, Matthew Jagielski, Vikash Sehwag, Florian Tramèr, Borja Balle, Daphne Ippolito, and Eric Wallace. Extracting Training Data from Diffusion Models. *arXiv preprint arXiv:2301.13188*, 2023. URL `http://arxiv.org/abs/2301.13188`.

Caroline Chan, Shiry Ginosar, Tinghui Zhou, and Alexei Efros. Everybody Dance Now. In *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, volume 49, pages 5932–5941. IEEE, oct 2019. ISBN 978-1-7281-4803-8. doi: 10.1109/ICCV.2019.00603.

Yasin Yazıcı Chandrasekhar, Chuan-Sheng Foo, Stefan Winkler, Kim-Hui Yap, Georgios Piliouras, and Vijay. The Unusual Effectiveness of Averaging in GAN Training. In *International Conference on Learning Representations*, 2018.

Bindita Chaudhuri, Nikolaos Sarafianos, Linda Shapiro, and Tony Tung. Semi-supervised Synthesis of High-Resolution Editable Textures for 3D Humans. In *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pages 7987–7996, 2021. ISBN 9781665445092. doi: 10.1109/CVPR46437.2021.00790.

Yunjey Choi, Youngjung Uh, Jaejun Yoo, and Jung-Woo Ha. StarGAN v2: Diverse Image Synthesis for Multiple Domains. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 8185–8194. IEEE, jun 2020. ISBN 978-1-7281-7168-5. doi: 10.1109/CVPR42600.2020.00821.

Umur A. Ciftci, Gokturk Yuksek, and Ilke Demir. My Face My Choice: Privacy Enhancing Deepfakes for Social Media Anonymization. In *2023 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, 2023.

Marius Cordts, Mohamed Omran, Sebastian Ramos, Timo Rehfeld, Markus Enzweiler, Rodrigo Benenson, Uwe Franke, Stefan Roth, and Bernt Schiele. The Cityscapes Dataset for Semantic Urban Scene Understanding. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3213–3223. IEEE, jun 2016. ISBN 978-1-4673-8851-1. doi: 10.1109/CVPR.2016.350.

Council of European Union. Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (General Data Protection Regulation), 2016. URL `http://eur-lex.europa.eu/legal-content/EN/TXT/?uri=OJ:L:2016:119:TOC`.

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. ImageNet: A large-scale hierarchical image database. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pages 248–255. IEEE, jun 2009. ISBN 978-1-4244-3992-8. doi: 10.1109/CVPR.2009.5206848.

Brian Dolhansky, Joanna Bitton, Ben Pflaum, Jikuo Lu, Russ Howes, Menglin Wang, and Cristian Canton Ferrer. The DeepFake Detection Challenge (DFDC) Dataset. *arXiv preprint arXiv:2006.07397*, jun 2020.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. In *International Conference on Learning Representations*, 2021.

DC Dowson and BV Landau. The Fréchet distance between multivariate normal distributions. *Journal of multivariate analysis*, 12(3):450–455, 1982.

Petr Dvořáček and Petr Hurtik. What Is the Cost of Privacy? In *Communications in Computer and Information Science*, volume 1602 CCIS, pages 696–706. Springer International Publishing, 2022. ISBN 9783031089732. doi: 10.1007/978-3-031-08974-9_55.

Patrick Esser and Ekaterina Sutter. A Variational U-Net for Conditional Appearance and Shape Generation. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8857–8866. IEEE, jun 2018. ISBN 978-1-5386-6420-9. doi: 10.1109/CVPR.2018.00923.

Anna Fruhstuck, Krishna Kumar Singh, Eli Shechtman, Niloy J Mitra, Peter Wonka, and Jingwan Lu. InsetGAN for Full-Body Image Generation. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 7713–7722. IEEE, jun 2022. ISBN 978-1-6654-6946-3. doi: 10.1109/CVPR52688.2022.00757.

Jianglin Fu, Shikai Li, Yuming Jiang, Kwan-Yee Lin, Chen Qian, Chen Change Loy, Wayne Wu, and Ziwei Liu. StyleGAN-Human: A Data-Centric Odyssey of Human Generation. In *Computer Vision - ECCV 2022*, volume 13676, pages 1–19. Springer, Cham, 2022. doi: 10.1007/978-3-031-19787-1_1.

Oran Gafni, Lior Wolf, and Yaniv Taigman. Live Face De-Identification in Video. In *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 9377–9386. IEEE, oct 2019. ISBN 978-1-7281-4803-8. doi: 10.1109/ICCV.2019.00947.

Andrew C Gallagher and Tsuhan Chen. Clothing cosegmentation for recognizing people. In *2008 IEEE Conference on Computer Vision and Pattern Recognition*, pages 1–8. IEEE, jun 2008. ISBN 978-1-4244-2242-5. doi: 10.1109/CVPR.2008.4587481.

Jakob Geyer, Yohannes Kassahun, Mentar Mahmudi, Xavier Ricou, Rupesh Durgesh, Andrew S. Chung, Lorenz Hauswald, Viet Hoang Pham, Maximilian Mühlegg, Sebastian Dorn, Tiffany Fernandez, Martin Jänicke, Sudesh Mirashi, Chiragkumar Savani, Martin Sturm, Oleksandr Vorobiov, Martin Oelker, Sebastian Garreis, and Peter Schuberth. A2D2: Audi Autonomous Driving Dataset. *arXiv preprint arXiv:2004.06320*, 2020.

Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In *Advances in neural information processing systems*, pages 2672–2680, 2014.

Ralph Gross, Edoardo Airoldi, Bradley Malin, and Latanya Sweeney. Integrating Utility into Face De-identification. In *Proceedings of the 5th international conference on Privacy Enhancing Technologies*, pages 227–242. 2006a. ISBN 3540347453. doi: 10.1007/11767831_15.

Ralph Gross, Latanya Sweeney, F. de la Torre, and Simon Baker. Model-Based Face De-Identification. In *2006 Conference on Computer Vision and Pattern Recognition Workshop (CVPRW'06)*, pages 161–161. IEEE, 2006b. ISBN 0-7695-2646-2. doi: 10.1109/CVPRW.2006.125.

Riza Alp Guler, Natalia Neverova, and Iasonas Kokkinos. DensePose: Dense Human Pose Estimation in the Wild. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7297–7306. IEEE, jun 2018. ISBN 978-1-5386-6420-9. doi: 10.1109/CVPR.2018.00762.

Ishaan Gulrajani, Faruk Ahmed, Martin Arjovsky, Vincent Dumoulin, and Aaron C Courville. Improved training of wasserstein gans. In *Advances in Neural Information Processing Systems*, pages 5767–5777, 2017.

Aayush Gupta, Ayush Jaiswal, Yue Wu, Vivek Yadav, and Pradeep Natarajan. Adversarial Mask Generation for Preserving Visual Privacy. In *2021 16th IEEE International Conference on Automatic Face and Gesture Recognition (FG 2021)*, pages 1–5. IEEE, dec 2021. ISBN 978-1-6654-3176-7. doi: 10.1109/FG52635.2021.9666933.

Xintong Han, Zuxuan Wu, Zhe Wu, Ruichi Yu, and Larry S Davis. VITON: An Image-Based Virtual Try-on Network. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7543–7552. IEEE, jun 2018. ISBN 978-1-5386-6420-9. doi: 10.1109/CVPR.2018.00787.

Rakibul Hasan, Eman Hassan, Yifang Li, Kelly Caine, David J. Crandall, Roberto Hoyle, and Apu Kapadia. Viewer Experience of Obscuring Scene Elements in Photos to Enhance Privacy. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*, volume 2018-April, pages 1–13, New York, NY, USA, apr 2018. ACM. ISBN 9781450356206. doi: 10.1145/3173574.3173621.

Jianping He, Bin Liu, Deguang Kong, Xuan Bao, Na Wang, Hongxia Jin, and George Kesidis. PUPPIES: Transformation-Supported Personalized Privacy Preserving Partial Image Sharing. In *2016 46th Annual IEEE/IFIP International Conference on Dependable Systems and Networks (DSN)*, pages 359–370. IEEE, jun 2016a. ISBN 978-1-4673-8891-7. doi: 10.1109/DSN.2016.40.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep Residual Learning for Image Recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 770–778. IEEE, jun 2016b. ISBN 978-1-4673-8851-1. doi: 10.1109/CVPR.2016.90.

Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollar, and Ross Girshick. Masked Autoencoders Are Scalable Vision Learners. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, volume 2022-June, pages 15979–15988. IEEE, jun 2022. ISBN 978-1-6654-6946-3. doi: 10.1109/CVPR52688.2022.01553.

Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. In *Advances in Neural Information Processing Systems*, pages 6626–6637, 2017.

Xun Huang and Serge Belongie. Arbitrary Style Transfer in Real-Time with Adaptive Instance Normalization. In *2017 IEEE International Conference on Computer Vision (ICCV)*, pages 1510–1519. IEEE, oct 2017. ISBN 978-1-5386-1032-9. doi: 10.1109/ICCV.2017.167.

Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A Efros. Image-to-Image Translation with Conditional Adversarial Networks. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5967–5976. IEEE, jul 2017. ISBN 978-1-5386-0457-1. doi: 10.1109/CVPR.2017.632.

Anil K Jain, Patrick Flynn, and Arun A Ross. *Handbook of Biometrics*. Springer US, Boston, MA, 2008. ISBN 978-0-387-71040-2. doi: 10.1007/978-0-387-71041-9.

Amin Jourabloo, Xi Yin, and Xiaoming Liu. Attribute preserved face de-identification. In *2015 International Conference on Biometrics (ICB)*, pages 278–285. IEEE, may 2015. ISBN 978-1-4799-7824-3. doi: 10.1109/ICB.2015.7139096.

Ufuk Kacmaz, Jan Melchior, Daniela Horn, Andreas Witte, Sebastian Schoenen, and Sebastian Houben. Fully Automated, Realistic License Plate Substitution in Real-Life Images. In *IEEE Conference on Intelligent Transportation Systems, Proceedings, ITSC*, volume 2021-Septe, pages 2972–2978. Institute of Electrical and Electronics Engineers Inc., sep 2021. ISBN 9781728191423. doi: 10.1109/ITSC48978.2021.9564769.

Animesh Karnewar and Oliver Wang. MSG-GAN: Multi-Scale Gradients for Generative Adversarial Networks. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 7796–7805. IEEE, jun 2020. ISBN 978-1-7281-7168-5. doi: 10.1109/CVPR42600.2020.00782.

Tero Karras, Timo Aila, Samuli Laine, and Jaakko Lehtinen. Progressive growing of gans for improved quality, stability, and variation. In *International Conference on Learning Representations*, 2018.

Tero Karras, Samuli Laine, and Timo Aila. A Style-Based Generator Architecture for Generative Adversarial Networks. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4396–4405. IEEE, jun 2019. ISBN 978-1-7281-3293-8. doi: 10.1109/CVPR.2019.00453.

Tero Karras, Samuli Laine, Miika Aittala, Janne Hellsten, Jaakko Lehtinen, and Timo Aila. Analyzing and Improving the Image Quality of StyleGAN. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 8107–8116. IEEE, jun 2020. ISBN 978-1-7281-7168-5. doi: 10.1109/CVPR42600.2020.00813.

Bach Ngoc Kim, Jose Dolz, Pierre-Marc Jodoin, and Christian Desrosiers. Privacy-Net: An Adversarial Approach for Identity-Obfuscated Segmentation of Medical Images. *IEEE Transactions on Medical Imaging*, 40(7): 1737–1749, jul 2021. ISSN 0278-0062. doi: 10.1109/TMI.2021.3065727.

Diederik P Kingma and Jimmy Ba. Adam: A Method for Stochastic Optimization. In *International Conference on Learning Representations*, 2015.

Alexander Kirillov, Yuxin Wu, Kaiming He, and Ross Girshick. Pointrend: Image segmentation as rendering. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 9796–9805, 2020. doi: 10.1109/CVPR42600.2020.00982.

Sander R. Klomp, Matthew Van Rijn, Rob G.J. Wijnhoven, Cees G.M. Snoek, and Peter H.N. De With. Safe Fakes: Evaluating Face Anonymizers for Face Detectors. In *2021 16th IEEE International Conference on Automatic Face and Gesture Recognition (FG 2021)*, pages 1–8. IEEE, dec 2021. ISBN 978-1-6654-3176-7. doi: 10.1109/FG52635.2021.9666936.

Umut Kocasari, Alara Dirik, Mert Tiftikci, and Pinar Yanardag. StyleMC: Multi-Channel Based Fast Text-Guided Image Generation and Manipulation. In *2022 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pages 3441–3450. IEEE, jan 2022. ISBN 978-1-6654-0915-5. doi: 10.1109/WACV51458.2022.00350.

Alexey Kurakin, Ian Goodfellow, Samy Bengio, Yinpeng Dong, Fangzhou Liao, Ming Liang, Tianyu Pang, Jun Zhu, Xiaolin Hu, Cihang Xie, Jianyu Wang, Zhishuai Zhang, Zhou Ren, Alan Yuille, Sangxia Huang, Yao Zhao, Yuzhe Zhao, Zhonglin Han, Junjiajia Long, Yerkebulan Berdibekov, Takuya Akiba, Seiya Tokui, and Motoki Abe. Adversarial Attacks and Defences Competition, 2018a.

Alexey Kurakin, Ian J. Goodfellow, and Samy Bengio. Adversarial Examples in the Physical World. In *Artificial Intelligence Safety and Security*, pages 99–112. Chapman and Hall/CRC, jul 2018b. doi: 10.1201/9781351251389-8.

Tuomas Kynkäänniemi, Tero Karras, Miika Aittala, Timo Aila, and Jaakko Lehtinen. The Role of ImageNet Classes in Frechet Inception Distance. *arXiv preprint arXiv:2203.06026*, 2022.

Karen Lander, Vicki Bruce, and Harry Hill. Evaluating the effectiveness of pixelation and blurring on masking the identity of familiar faces. *Applied Cognitive Psychology*, 15(1):101–116, jan 2001. ISSN 0888-4080. doi: 10.1002/1099-0720(200101/02)15:1<101::AID-ACP697>3.0.CO;2-7.

Christoph Lassner, Gerard Pons-Moll, and Peter V. Gehler. A Generative Model of People in Clothing. In *2017 IEEE International Conference on Computer Vision (ICCV)*, volume 2017-Octob, pages 853–862. IEEE, oct 2017. ISBN 978-1-5386-1032-9. doi: 10.1109/ICCV.2017.98.

Jian Li, Yabiao Wang, Changan Wang, Ying Tai, Jianjun Qian, Jian Yang, Chengjie Wang, Jilin Li, and Feiyue Huang. DSFD: Dual Shot Face Detector. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5055–5064. IEEE, jun 2019a. ISBN 978-1-7281-3293-8. doi: 10.1109/CVPR.2019.00520.

Jizhizi Li, Sihan Ma, Jing Zhang, and Dacheng Tao. Privacy-Preserving Portrait Matting. In *Proceedings of the 29th ACM International Conference on Multimedia*, pages 3501–3509, New York, NY, USA, oct 2021. ACM. ISBN 9781450386517. doi: 10.1145/3474085.3475512.

Yifang Li, Nishant Vishwamitra, Bart P. Knijnenburg, Hongxin Hu, and Kelly Caine. Blur vs. Block: Investigating the Effectiveness of Privacy-Enhancing Obfuscation for Images. In *2017 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, volume 2017-July, pages 1343–1351. IEEE, jul 2017. ISBN 978-1-5386-0733-6. doi: 10.1109/CVPRW.2017.176.

Yining Li, Chen Huang, and Chen Change Loy. Dense Intrinsic Appearance Flow for Human Pose Transfer. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3688–3697. IEEE, jun 2019b. ISBN 978-1-7281-3293-8. doi: 10.1109/CVPR.2019.00381.

Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. Microsoft COCO: Common Objects in Context. In *European conference on computer vision*, volume 8693 LNCS, pages 740–755. Springer, Cham, 2014. doi: 10.1007/978-3-319-10602-1_48.

Ziwei Liu, Ping Luo, Xiaogang Wang, and Xiaoou Tang. Deep Learning Face Attributes in the Wild. In *2015 IEEE International Conference on Computer Vision (ICCV)*, volume 15, pages 3730–3738. IEEE, dec 2015. ISBN 978-1-4673-8391-2. doi: 10.1109/ICCV.2015.425.

Liqian Ma, Xu Jia, Qianru Sun, Bernt Schiele, Tinne Tuytelaars, and Luc Van Gool. Pose guided person image generation. In *Advances in Neural Information Processing Systems*, volume 2017-Decem, pages 406–416, 2017.

Liqian Ma, Qianru Sun, Stamatios Georgoulis, Luc Van Gool, Bernt Schiele, and Mario Fritz. Disentangled Person Image Generation. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 99–108. IEEE, jun 2018. ISBN 978-1-5386-6420-9. doi: 10.1109/CVPR.2018.0001 8.

Richard T. Marriott, Safa Madiouni, Sami Romdhani, Stephane Gentric, and Liming Chen. An Assessment of GANs for Identity-related Applications. In *2020 IEEE International Joint Conference on Biometrics (IJCB)*, pages 1–10. IEEE, sep 2020. ISBN 978-1-7281-9186-7. doi: 10.1109/IJCB48548.2020.9304879.

Maxim Maximov, Ismail Elezi, and Laura Leal-Taixe. CIAGAN: Conditional Identity Anonymization Generative Adversarial Networks. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5446–5455. IEEE, jun 2020. ISBN 978-1-7281-7168-5. doi: 10.1109/CVPR42600.2020.00549.

Richard McPherson, Reza Shokri, and Vitaly Shmatikov. Defeating Image Obfuscation with Deep Learning. *arXiv preprint arXiv:1609.00408*, sep 2016.

Lars Mescheder, Andreas Geiger, and Sebastian Nowozin. Which training methods for GANs do actually Converge? In *International Conference on Machine Learning*, pages 3478–3487, 2018.

Mehdi Mirza and Simon Osindero. Conditional generative adversarial nets. *arXiv preprint arXiv:1411.1784*, 2014.

Ron Mokady, Omer Tov, Michal Yarom, Oran Lang, Inbar Mosseri, Tali Dekel, Daniel Cohen-Or, and Michal Irani. Self-Distilled StyleGAN: Towards Generation from Internet Photos. In *ACM SIGGRAPH 2022 Conference Proceedings*, pages 1–9. ACM, aug 2022. ISBN 9781450393379. doi: 10.1145/3528233.3530708.

Wiktor Muron. motpy - simple multi object tracking library, 2022. URL `https://github.com/wmuron/motpy`.

Muhammad Ferjad Naeem, Seong Joon Oh, Youngjung Uh, Yunjey Choi, and Jaejun Yoo. Reliable Fidelity and Diversity Metrics for Generative Models. In Hal Daumé III and Aarti Singh, editors, *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pages 7176–7185. PMLR, 2020.

Ryota Natsume, Shunsuke Saito, Zeng Huang, Weikai Chen, Chongyang Ma, Hao Li, and Shigeo Morishima. SiCloPe: Silhouette-Based Clothed People. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, volume 2019-June, pages 4475–4485. IEEE, jun 2019. ISBN 978-1-7281-3293-8. doi: 10.1109/CVPR.2019.00461.

Natalia Neverova, Riza Alp Güler, and Iasonas Kokkinos. Dense Pose Transfer. In *European conference on computer vision*, volume 11207 LNCS, pages 128–143, 2018. ISBN 9783030012182. doi: 10.1007/978-3-030-01219-9_8. URL `http://link.springer.com/10.1007/978-3-030-01219-9_8`.

Natalia Neverova, David Novotny, Vasil Khalidov, Marc Szafraniec, Patrick Labatut, and Andrea Vedaldi. Continuous Surface Embeddings. In *Advances in Neural Information Processing Systems*, volume 33, pages 17258–17270. Curran Associates, Inc., nov 2020.

E.M. Newton, Latanya Sweeney, and Bradley Malin. Preserving privacy by de-identifying face images. *IEEE Transactions on Knowledge and Data Engineering*, 17(2):232–243, feb 2005. ISSN 1041-4347. doi: 10.1109/TK DE.2005.32.

Seong Joon Oh, Rodrigo Benenson, Mario Fritz, and Bernt Schiele. Faceless Person Recognition: Privacy Implications in Social Media. In *Computer Vision - ECCV 2016*, pages 19–35. Springer Verlag, 2016. ISBN 9783319464862. doi: 10.1007/978-3-319-46487-9_2.

Seong Joon Oh, Mario Fritz, and Bernt Schiele. Adversarial Image Perturbation for Privacy Protection A Game Theory Perspective. In *2017 IEEE International Conference on Computer Vision (ICCV)*, volume 2017-Octob, pages 1491–1500. IEEE, oct 2017. ISBN 978-1-5386-1032-9. doi: 10.1109/ICCV.2017.165.

Tribhuvanesh Orekondy, Mario Fritz, and Bernt Schiele. Connecting Pixels to Privacy and Utility: Automatic Redaction of Private Information in Images. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8466–8475. IEEE, jun 2018. ISBN 978-1-5386-6420-9. doi: 10.110 9/CVPR.2018.00883.

A. J. Piergiovanni and Michael S. Ryoo. AViD dataset: Anonymized videos from diverse countries. In *Advances in Neural Information Processing Systems*, volume 2020-Decem, 2020.

Albert Pumarola, Antonio Agudo, Alberto Sanfeliu, and Francesc Moreno-Noguer. Unsupervised Person Image Synthesis in Arbitrary Poses. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8620–8628. IEEE, jun 2018. ISBN 978-1-5386-6420-9. doi: 10.1109/CV PR.2018.00899.

Alec Radford, Luke Metz, and Soumith Chintala. Unsupervised representation learning with deep convolutional generative adversarial networks. *arXiv preprint arXiv:1511.06434*, 2015.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning Transferable Visual

Models From Natural Language Supervision. In *International Conference on Machine Learning*, pages 8748–8763, 2021.

Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. In *Advances in neural information processing systems*, pages 91–99, 2015.

Zhongzheng Ren, Yong Jae Lee, and Michael S Ryoo. Learning to Anonymize Faces for Privacy Preserving Action Detection. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 639–655. Springer International Publishing, 2018. ISBN 978-3-030-01246-5. doi: 10.1007/978-3-030-01246-5_38.

Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-Net: Convolutional Networks for Biomedical Image Segmentation. In *International Conference on Medical image computing and computer-assisted intervention*, pages 234–241. Springer, 2015. doi: 10.1007/978-3-319-24574-4_28.

Shunsuke Saito, Zeng Huang, Ryota Natsume, Shigeo Morishima, Hao Li, and Angjoo Kanazawa. PIFu: Pixel-Aligned Implicit Function for High-Resolution Clothed Human Digitization. In *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, volume 2019-Octob, pages 2304–2314. IEEE, oct 2019. ISBN 978-1-7281-4803-8. doi: 10.1109/ICCV.2019.00239.

Tim Salimans, Ian Goodfellow, Wojciech Zaremba, Vicki Cheung, Alec Radford, and Xi Chen. Improved techniques for training gans. In *Advances in Neural Information Processing Systems*, pages 2234–2242, 2016.

Kripasindhu Sarkar, Dushyant Mehta, Weipeng Xu, Vladislav Golyanik, and Christian Theobalt. Neural Re-rendering of Humans from a Single Image. In *European conference on computer vision*, volume 12356 LNCS, pages 596–613. Springer Science and Business Media Deutschland GmbH, 2020. ISBN 9783030586201. doi: 10.1007/978-3-030-58621-8_35.

Axel Sauer, Kashyap Chitta, Jens Müller, and Andreas Geiger. Projected GANs Converge Faster. In *Advances in Neural Information Processing Systems*, pages 17480–17492, 2021.

Axel Sauer, Katja Schwarz, and Andreas Geiger. StyleGAN-XL: Scaling StyleGAN to Large Diverse Datasets. In *ACM SIGGRAPH 2022 Conference Proceedings*, pages 1–10, Vancouver, BC, Canada, aug 2022. Association for Computing Machinery. ISBN 9781450393379. doi: 10.1145/3528233.3530738.

Chenyang Si, Wei Wang, Liang Wang, and Tieniu Tan. Multistage Adversarial Losses for Pose-Based Human Image Synthesis. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 118–126. IEEE, jun 2018. ISBN 978-1-5386-6420-9. doi: 10.1109/CVPR.2018.00020.

Aliaksandr Siarohin, Enver Sangineto, Stephane Lathuiliere, and Nicu Sebe. Deformable GANs for Pose-Based Human Image Generation. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3408–3416. IEEE, jun 2018. ISBN 978-1-5386-6420-9. doi: 10.1109/CVPR.2018.00359.

Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.

Ivan Skorokhodov, Sergey Tulyakov, and Mohamed Elhoseiny. StyleGAN-V: A Continuous Video Generator with the Price, Image Quality and Perks of StyleGAN2. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, volume 2022-June, pages 3616–3626. IEEE, jun 2022. ISBN 978-1-6654-6946-3. doi: 10.1109/CVPR52688.2022.00361.

Sijie Song, Wei Zhang, Jiaying Liu, Zongming Guo, and Tao Mei. Unpaired Person Image Generation With Semantic Parsing Transformation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 43(11):4161–4176, nov 2021. ISSN 0162-8828. doi: 10.1109/TPAMI.2020.2992105.

Qianru Sun, Liqian Ma, Seong Joon Oh, Luc Van Gool, Bernt Schiele, and Mario Fritz. Natural and Effective Obfuscation by Head Inpainting. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5050–5059. IEEE, jun 2018a. ISBN 978-1-5386-6420-9. doi: 10.1109/CVPR.2018.00530.

Qianru Sun, Ayush Tewari, Weipeng Xu, Mario Fritz, Christian Theobalt, and Bernt Schiele. A Hybrid Model for Identity Obfuscation by Face Replacement. In *Proceedings of the European Conference on Computer*

Vision (ECCV), pages 570–586. Springer International Publishing, 2018b. doi: 10.1007/978-3-030-01246-5_34.

Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jon Shlens, and Zbigniew Wojna. Rethinking the inception architecture for computer vision. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2818–2826, 2016. doi: 10.1109/CVPR.2016.308.

Justus Thies, Michael Zollhofer, Marc Stamminger, Christian Theobalt, and Matthias Niessner. Face2Face: Real-Time Face Capture and Reenactment of RGB Videos. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2387–2395. IEEE, jun 2016. ISBN 978-1-4673-8851-1. doi: 10.1109/CVPR.2016.262.

Matteo Tomei, Lorenzo Baraldi, Simone Bronzin, and Rita Cucchiara. Estimating (and fixing) the Effect of Face Obfuscation in Video Recognition. In *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 3257–3263. IEEE, jun 2021. ISBN 978-1-6654-4899-4. doi: 10.1109/CVPRW53098.2021.00364.

Rotem Tzaban, Ron Mokady, Rinon Gal, Amit Bermano, and Daniel Cohen-Or. Stitch it in Time: GAN-Based Facial Editing of Real Videos. In *SIGGRAPH Asia 2022 Conference Papers*, pages 1–9, New York, NY, USA, nov 2022. ACM. ISBN 9781450394703. doi: 10.1145/3550469.3555382.

Ries Uittenbogaard, Clint Sebastian, Julien Vijverberg, Bas Boom, Dariu M. Gavrila, and Peter H.N. de With. Privacy Protection in Street-View Panoramas Using Depth and Multi-View Imagery. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, volume 2019-June, pages 10573–10582. IEEE, jun 2019. ISBN 978-1-7281-3293-8. doi: 10.1109/CVPR.2019.01083.

Chung-Yi Weng, Brian Curless, and Ira Kemelmacher-Shlizerman. Vid2Actor: Free-viewpoint Animatable Person Synthesis from Video in the Wild. *arXiv preprint arXiv:2012.12884*, 2020.

Michael J. Wilber, Vitaly Shmatikov, and Serge Belongie. Can we still avoid automatic face detection? In *2016 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 1–9. IEEE, mar 2016. ISBN 978-1-5090-0641-0. doi: 10.1109/WACV.2016.7477452.

Yifan Wu, Fan Yang, Yong Xu, and Haibin Ling. Privacy-Protective-GAN for Privacy Preserving Face De-Identification. *Journal of Computer Science and Technology*, 34(1):47–60, jan 2019. ISSN 1000-9000. doi: 10.1007/s11390-019-1898-8.

Chaojie Yang, Hanhui Li, Shengjie Wu, Shengkai Zhang, Haonan Yan, Nianhong Jiao, Jie Tang, Runnan Zhou, Xiaodan Liang, and Tianxiang Zheng. BodyGAN: General-purpose Controllable Neural Human Body Generation. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 7723–7732. IEEE, jun 2022a. ISBN 978-1-6654-6946-3. doi: 10.1109/CVPR52688.2022.00758.

Kaiyu Yang, Jacqueline Yau, Li Fei-Fei, Jia Deng, and Olga Russakovsky. A Study of Face Obfuscation in ImageNet. In *International Conference on Machine Learning*, pages 25313–25330, mar 2022b.

Shuo Yang, Ping Luo, Chen Change Loy, and Xiaoou Tang. WIDER FACE: A Face Detection Benchmark. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, volume 2016-Decem, pages 5525–5533. IEEE, jun 2016. ISBN 978-1-4673-8851-1. doi: 10.1109/CVPR.2016.596.

Zhuoqian Yang, Shikai Li, Wayne Wu, and Bo Dai. 3DHumanGAN: Towards Photo-Realistic 3D-Aware Human Image Generation. *arXiv preprint arXiv:2212.07378v1*.

Fisher Yu, Haofeng Chen, Xin Wang, Wenqi Xian, Yingying Chen, Fangchen Liu, Vashisht Madhavan, and Trevor Darrell. BDD100K: A Diverse Driving Dataset for Heterogeneous Multitask Learning. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2633–2642. IEEE, jun 2020. ISBN 978-1-7281-7168-5. doi: 10.1109/CVPR42600.2020.00271.

Ning Yu, Vladislav Skripniuk, Sahar Abdelnabi, and Mario Fritz. Artificial Fingerprinting for Generative Models: Rooting Deepfake Attribution in Training Data. In *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 14428–14437. IEEE, oct 2021. ISBN 978-1-6654-2812-5. doi: 10.1109/ICCV48922.2021.01418.

Egor Zakharov, Aliaksandra Shysheya, Egor Burkov, and Victor Lempitsky. Few-shot adversarial learning of realistic neural talking head models. In

Proceedings of the IEEE International Conference on Computer Vision, volume 2019-Octob, pages 9458–9467, 2019. ISBN 9781728148038. doi: 10.1109/ICCV.2019.00955.

Ning Zhang, Manohar Paluri, Yaniv Taigman, Rob Fergus, and Lubomir Bourdev. Beyond frontal faces: Improving Person Recognition using multiple cues. In *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, volume 07-12-June, pages 4804–4813. IEEE, jun 2015a. ISBN 978-1-4673-6964-0. doi: 10.1109/CVPR.2015.7299113. URL `http://ieeexplore.ieee.org/document/7299113/`.

Ning Zhang, Manohar Paluri, Yaniv Taigman, Rob Fergus, and Lubomir Bourdev. Beyond frontal faces: Improving Person Recognition using multiple cues. In *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, volume 07-12-June, pages 4804–4813. IEEE, jun 2015b. ISBN 978-1-4673-6964-0. doi: 10.1109/CVPR.2015.7299113.

Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The Unreasonable Effectiveness of Deep Features as a Perceptual Metric. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 586–595. IEEE, jun 2018. ISBN 978-1-5386-6420-9. doi: 10.1109/CVPR.2018.00068.

Liang Zheng, Liyue Shen, Lu Tian, Shengjin Wang, Jingdong Wang, and Qi Tian. Scalable Person Re-identification: A Benchmark. In *2015 IEEE International Conference on Computer Vision (ICCV)*, pages 1116–1124. IEEE, dec 2015. ISBN 978-1-4673-8391-2. doi: 10.1109/ICCV.2015.133.

Bolei Zhou, Agata Lapedriza, Antonio Torralba, and Aude Oliva. Places: An Image Database for Deep Scene Understanding. *Journal of Vision*, 17(10): 296, aug 2017. ISSN 1534-7362. doi: 10.1167/17.10.296.

Jingxing Zhou and Jurgen Beyerer. Impacts of Data Anonymization on Semantic Segmentation. In *2022 IEEE Intelligent Vehicles Symposium (IV)*, volume 2022-June, pages 997–1004. IEEE, jun 2022. ISBN 978-1-6654-8821-1. doi: 10.1109/IV51971.2022.9827262.

Jun-Yan Zhu, Richard Zhang, Deepak Pathak, Trevor Darrell, Alexei A Efros, Oliver Wang, and Eli Shechtman. Toward multimodal image-to-image translation. In *Advances in neural information processing systems*, pages 465–476, 2017.

Part II

Publications

Paper A

DeepPrivacy: A Generative Adversarial Network for Face Anonymization

Authors:

Håkon Hukkelås, Rudolf Mester, Frank Lindseth

Published at conference:

14th International Symposium on Visual Computing

Copyright:

Source Code:

`https://github.com/hukkelas/DeepPrivacy`

DeepPrivacy: A Generative Adversarial Network for Face Anonymization

Håkon Hukkelås, Rudolf Mester, Frank Lindseth

Department of Computer Science
Norwegian University of Science and Technology
Trondheim, Norway
{hakon.hukkelas, rudolf.mester, frankl}@ntnu.no

Abstract

We propose a novel architecture which is able to automatically anonymize faces in images while retaining the original data distribution. We ensure total anonymization of all faces in an image by generating images exclusively on privacy-safe information. Our model is based on a conditional generative adversarial network, generating images considering the original pose and image background. The conditional information enables us to generate highly realistic faces with a seamless transition between the generated face and the existing background. Furthermore, we introduce a diverse dataset of human faces, including unconventional poses, occluded faces, and a vast variability in backgrounds. Finally, we present experimental results reflecting the capability of our model to anonymize images while preserving the data distribution, making the data suitable for further training of deep learning models. As far as we know, no other solution has been proposed that guarantees the anonymization of faces while generating realistic images.

1 Introduction

Privacy-preserving data-processing is becoming more critical every year; however, no suitable solution has been found to anonymize images without degrading the image quality. The General Data Protection Regulation (GDPR) came to effect as of 25th of May, 2018, affecting all processing of personal data across Europe. GDPR requires regular consent from the individual for any use of their personal data. However, if the data does not allow to identify an

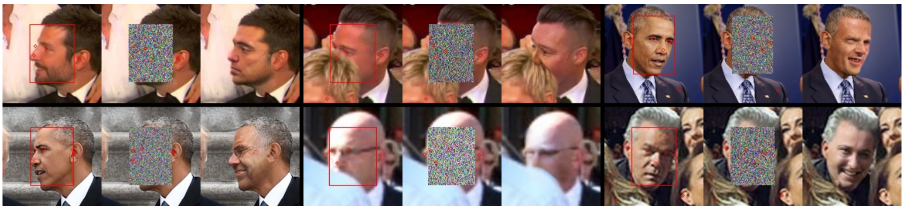


Figure 1: DeepPrivacy Results on a diverse set of images. The left image is the original image annotated with bounding box and keypoints, the middle image is the input image to our GAN, and the right image is the generated image. Note that our generator never sees any privacy-sensitive information.

individual, companies are free to use the data without consent. To effectively anonymize images, we require a robust model to replace the original face, without destroying the existing data distribution; that is: the output should be a realistic face fitting the given situation.

Anonymizing images, while retaining the original distribution, is a challenging task. The model is required to remove all privacy-sensitive information, generate a highly realistic face, and the transition between original and anonymized parts has to be seamless. This requires a model that can perform complex semantic reasoning to generate a new anonymized face. For practical use, we desire the model to be able to manage a broad diversity of images, poses, backgrounds, and different persons. Our proposed solution can successfully anonymize images in a large variety of cases, and create realistic faces to the given conditional information.

Our proposed model, called *DeepPrivacy*, is a conditional generative adversarial network [3, 18]. Our generator considers the existing background and a sparse pose annotation to generate realistic anonymized faces. The generator has a U-net architecture [23] that generates images with a resolution of 128×128. The model is trained with a progressive growing training technique [12] from a starting resolution of 8×8 to 128×128, which substantially improves the final image quality and overall training time. By design, our generator never observes the original face, ensuring removal of any privacy-sensitive information.

For practical use, we assume no demanding requirements for the object and keypoint detection methods. Our model requires two simple annotations of the face: (1) a bounding box annotation to identify the privacy-sensitive area, and (2) a sparse pose estimation of the face, containing keypoints for the ears, eyes, nose, and shoulders; in total seven keypoints. This keypoint annotation is identical to what Mask R-CNN [6] provides.

We provide a new dataset of human faces, *Flickr Diverse Faces* (FDF), which consists of 1.47M faces with a bounding box and keypoint annotation for each face. This dataset covers a considerably large diversity of facial poses, partial occlusions, complex backgrounds, and different persons. We will make this dataset publicly available along with our source code and pre-trained networks[1][2].

We evaluate our model by performing an extensive qualitative and quantitative study of the model's ability to retain the original data distribution. We anonymize the validation set of the WIDER-Face dataset [27], then run face detection on the anonymized images to measure the impact of anonymization on Average Precision (AP). DSFD [14] achieves 99.3% (95.9% out of 96.6% AP), 99.3% (95.0%/95.7%), and 99.3% (89.8%/90.4%) of the original AP on the easy, medium, and hard difficulty, respectively. On average, it achieves 99.3% of the original AP. In contrast, traditional anonymization techniques, such as 8*x*8 pixelation achieves 96.7%, heavy blur 90.5%, and black-out 41.4% of the original performance. Additionally, we present several ablation experiments that reflect the importance of a large model size and conditional pose information to generate high-quality faces.

In summary, we make the following contributions:

- We propose a novel generator architecture to anonymize faces, which ensures 100% removal of privacy-sensitive information in the original face. The generator can generate realistic looking faces that have a seamless transition to the existing background for various sets of poses and contexts.

- We provide the FDF dataset, including 1.47M faces with a tight bounding box and keypoint annotation for each face. The dataset covers a considerably larger diversity of faces compared to previous datasets.

[1] Code: `www.github.com/hukkelas/DeepPrivacy`

[2] FDF Dataset: `www.github.com/hukkelas/FDF`

2 Related Work

De-Identifying Faces: Currently, there exists a limited number of research studies on the task of removing privacy-sensitive information from an image including a face. Typically, the approach chosen is to alter the original image such that we remove all the privacy-sensitive information. These methods can be applied to all images; however, there is no assurance that these methods remove all privacy-sensitive information. Naive methods that apply simple image distortion have been discussed numerous times in literature [1, 19, 5, 20, 4], such as pixelation and blurring; but, they are inadequate for removing the privacy-sensitive information [4, 19, 20], and they alter the data distribution substantially.

K-same family of algorithms [4, 11, 20] implements the k-anonymity algorithm [25] for face images. Newton *et al.* prove that the k-same algorithm can remove all privacy-sensitive information; but, the resulting images often contain "ghosting" artifacts due to small alignment errors [4].

Jourabloo *et al.* [11] look at the task of de-identification grayscale images while preserving a large set of facial attributes. This is different from our work, as we do not directly train our generative model to generate faces with similar attributes to the original image. In contrast, our model is able to perform complex semantic reasoning to generate a face that is coherent with the overall context information given to the network, yielding a highly realistic face.

Generative Adversarial Networks (GANs) [3] is a highly successful training architecture to model a natural image distribution. GANs enables us to generate new images, often indistinguishable from the real data distribution. It has a broad diversity of application areas, from general image generation [2, 12, 13, 30], text-to-photo generation [31], style transfer [8, 24] and much more. With the numerous contributions since its conception, it has gone from a beautiful theoretical idea to a tool we can apply for practical use cases. In our work, we show that GANs are an efficient tool to remove privacy-sensitive information without destroying the original image quality.

Ren *et al.* [22] look at the task of anonymizing video data by using GANs. They perform anonymization by altering each pixel in the original image to hide the identity of the individuals. In contrast to their method, we can ensure the removal of all privacy-sensitive information, as our generative model never observes the original face.

Progressive Growing of GANs [12] propose a novel training technique to generate faces progressively, starting from a resolution of 4*x*4 and step-wise increasing it to 1024*x*1024. This training technique improves the final image quality and overall training time. Our proposed model uses the same training technique; however, we perform several alterations to their original model to convert it to a conditional GAN. With these alterations, we can include conditional information about the context and pose of the face. Our final generator architecture is similar to the one proposed by Isola *et al.* [9], but we introduce conditional information in several stages.

Image Inpainting is a closely related task to what we are trying to solve, and it is a widely researched area for generative models [10, 15, 17, 29]. Several research studies have looked at the task of face completion with a generative adversarial network [15, 29]. They mask a specific part of the face and try to complete this part with the conditional information given. From our knowledge, and the qualitative experiments they present in their papers, they are not able to mask a large enough section to remove all privacy-sensitive information. As the masked region grows, it requires a more advanced generative model that understands complex semantic reasoning, making the task considerably harder. Also, their experiments are based on the Celeb-A dataset [17], primarily consisting of celebrities with low diversity in facial pose, making models trained on this dataset unsuitable for real-world applications.

3 The Flickr Diverse Faces Dataset

FDF (Flickr Diverse Faces) is a new dataset of human faces, crawled from the YFCC-100M dataset [26]. It consists of 1.47M human faces with a minimum resolution of 128×128, containing facial keypoints and a bounding box annotation for each face. The dataset has a vast diversity in terms of age, ethnicity, facial pose, image background, and face occlusion. Randomly picked examples from the dataset can be seen in Figure 2. The dataset is extracted from scenes related to traffic, sports events, and outside activities. In comparison to the FFHQ [13] and Celeb-A [17] datasets, our dataset is more diverse in facial

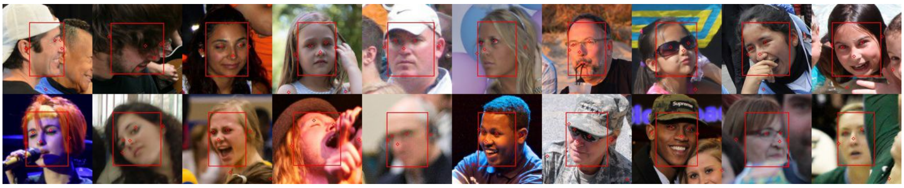


Figure 2: The FDF dataset. Each image has a sparse keypoint annotation (7 keypoints) of the face and a tight bounding box annotation. We recommend the reader to zoom in.

poses and it contains significantly more faces; however, the FFHQ dataset has a higher resolution.

The FDF dataset is a high-quality dataset with few annotation errors. The faces are automatically labeled with state-of-the-art keypoint and bounding box models, and we use a high confidence threshold for both the keypoint and bounding box predictions. The faces are extracted from 1.08M images in the YFCC100-M dataset. For keypoint estimation, we use Mask R-CNN [6], with a ResNet-50 FPN backbone [16]. For bounding box annotation, we use the Single Shot Scale-invariant Face Detector [32]. To combine the predictions, we match a keypoint with a face bounding box if the eye and nose annotation are within the bounding box. Each bounding box and keypoint has a single match, and we match them with a greedy approach based on descending prediction confidence.

4 Model

Our proposed model is a conditional GAN, generating images based on the surrounding of the face and sparse pose information. Figure 1 shows the conditional information given to our network, and Appendix A has a detailed description of the pre-processing steps. We base our model on the one proposed by Karras *et al*. [12]. Their model is a non-conditional GAN, and we perform several alterations to include conditional information.

We use seven keypoints to describe the pose of the face: left/right eye, left/right ear, left/right shoulder, and nose. To reduce the number of parameters in the

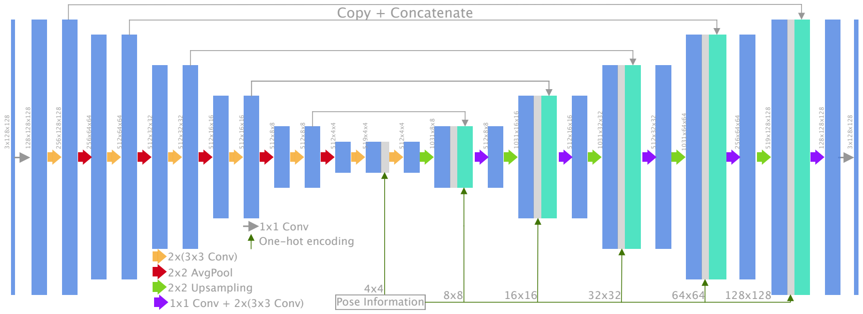


Figure 3: Generator Architecture for 128×128 resolution. Each convolutional layer is followed by pixel normalization [12] and LeakyReLU($\alpha = 0.2$). After each upsampling layer, we concatenate the upsampled output with pose information and the corresponding skip connection.

network, we pre-process the pose information into a one-hot encoded image of size $K \times M \times M$, where K is the number of keypoints and M is the target resolution.

Progressive growing training technique is crucial for our model's success. We apply progressive growing to both the generator and discriminator to grow the networks from a starting resolution of 8. We double the resolution each time we expand our network until we reach the final resolution of 128×128. The pose information is included for each resolution in the generator and discriminator, making the pose information finer for each increase in resolution.

4.1 Generator Architecture

Figure 3 shows our proposed generator architecture for 128×128 resolution. Our generator has a U-net [23] architecture to include background information. The encoder and decoder have the same number of filters in each convolution, but the decoder has an additional 1×1 bottleneck convolution after each skip connection. This bottleneck design reduces the number of parameters in the decoder significantly. To include the pose information for each resolution, we concatenate the output after each upsampling layer with pose information and the corresponding skip connection. The general layer structure is identical to Karras *et al.* [12], where we use pixel replication for upsampling, pixel normalization and LeakyReLU after each convolution, and equalized learning rate instead of careful weight initialization.

Progressive Growing: Each time we increase the resolution of the generator, we add two 3×3 convolutions to the start of the encoder and the end of the decoder. We use a transition phase identical to Karras *et al.* [12] for both of these new blocks, making the network stable throughout training. We note that the network is still unstable during the transition phase, but it is significantly better compared to training without progressive growing.

4.2 Discriminator Architecture

Our proposed discriminator architecture is identical to the one proposed by Karras *et al.* [12], with a few exceptions. First, we include the background information as conditional input to the start of the discriminator, making the input image have six channels instead of three. Secondly, we include pose information at each resolution of the discriminator. The pose information is concatenated with the output of each downsampling layer, similar to the decoder in the generator. Finally, we remove the mini-batch standard deviation layer presented by Karras *et al.* [12], as we find the diversity of our generated faces satisfactory.

The adjustments made to the generator doubles the number of total parameters in the network. To follow the design lines of Karras *et al.* [12], we desire that the complexity in terms of the number of parameters to be similar for the discriminator and generator. We evaluate two different discriminator models, which we will name the *deep discriminator* and the *wide discriminator*. The deep discriminator doubles the number of convolutional layers for each resolution. To mimic the skip-connections in the generator, we wrap the convolutions for each resolution in residual blocks. The wider discriminator keeps the same architecture; however, we increase the number of filters in each convolutional layer by a factor of $\sqrt{2}$.

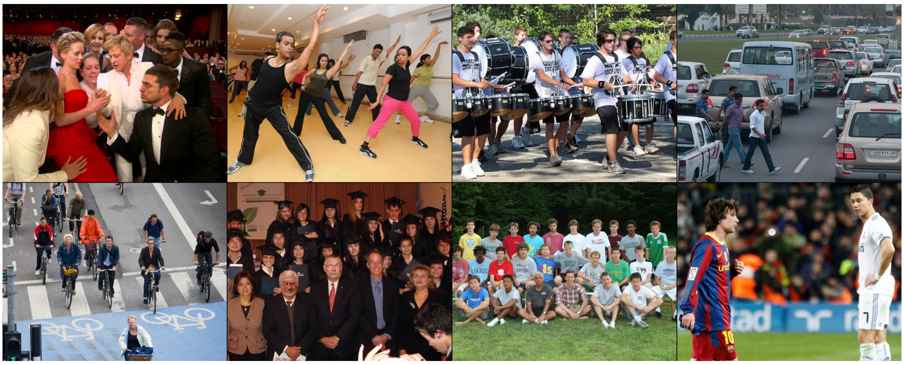


Figure 4: Anonymized Images from DeepPrivacy. Every single face in the images has been generated. We recommend the reader to zoom in.

5 Experiments

DeepPrivacy can robustly generate anonymized faces for a vast diversity of poses, backgrounds, and different persons. From qualitative evaluations of our generated results on the WIDER-Face dataset [27], we find our proposed solution to be robust to a broad diversity of images. Figure 4 shows several results of our proposed solution on the WIDER-Face dataset. Note that the network is trained on the FDF dataset; we do not train on any images in the WIDER-Face dataset.

We evaluate the impact of anonymization on the WIDER-Face [27] dataset. We measure the AP of a face detection model on the anonymized dataset and compare this to the original dataset. We report the standard metrics for the different difficulties for WIDER-Face. Additionally, we perform several ablation experiments on our proposed FDF dataset.

Our final model is trained for 17 days, 40M images, until we observe no qualitative differences between consecutive training iterations. It converges to a Frèchect Inception Distance (FID) [7] of 1.53. Specific training details and input pre-processing are given in Appendix A.

Table 1: Face Detection AP on the WIDER Face [27] validation dataset. The face detection method used is DSFD [14], the current state-of-the-art on WIDER-Face.

Anonymization method	Easy	Medium	Hard
No Anonymization [14]	96.6%	95.7%	90.4%
Blacked out	24.9%	36.3%	54.8%
Pixelation (16x16)	95.3%	94.9%	**90.2%**
Pixelation (8x8)	91.4%	92.3%	88.9%
9x9 Gaussian Blur ($\sigma = 3$)	95.3%	92.8%	84.7%
Heavy Blur (filter size = 30% face width)	83.4%	86.3%	86.1%
DeepPrivacy (Ours)	**95.9%**	**95.0%**	89.8%

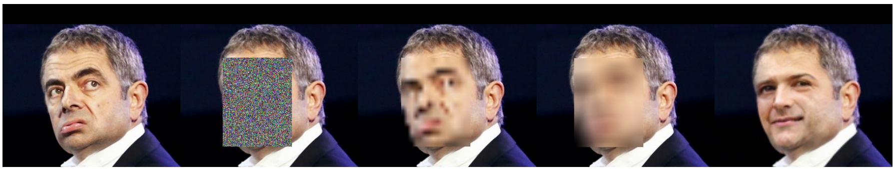


Figure 5: Different Anonymization Methods on a face in the WIDER Face validation set.

5.1 Effect of Anonymization for Face Detection

Table 1 shows the AP of different anonymization techniques on the WIDER-Face validation set. In comparison to the original dataset, DeepPrivacy only degrades the AP by 0.7%, 0.7%, and 0.6% on the easy, medium, and hard difficulties, respectively.

We compare DeepPrivacy anonymization to simpler anonymization methods; black-out, pixelation, and blurring. Figure 5 illustrates the different anonymization methods. DeepPrivacy generally achieves a higher AP compared to all other methods, with the exception of 16×16 pixelation.

Note that 16×16 pixelation does not affect a majority of the faces in the dataset. For the "hard" challenge, 0% of the faces has a resolution larger than 16×16. For the easy and medium challenge, 43% and 29.9% has a resolution larger than 16×16. The observant reader might notice that for the "hard" challenge, 16×16 pixelation should have no effect; however, the AP is degraded in comparison to the original dataset (see Table 1). We believe that the AP on the "hard" challenge is degraded due to anonymizing faces in easy/medium challenge can affect the model in cases where faces from "hard" and easy/medium are present in the same image.

Experiment Details: For the face detector we use the current state-of-the-art, Dual Shot Face Detector (DSFD) [14]. The WIDER-Face dataset has no facial keypoint annotations; therefore, we automatically detect keypoints for each face with the same method as used for the FDF dataset. To match keypoints with a bounding box, we use the same greedy approach as earlier. Mask R-CNN [6] is not able to detect keypoints for all faces, especially in cases with high occlusion, low resolution, or faces turned away from the camera. Thus, we are only able to anonymize 43% of the faces in the validation set. Of the faces that are not anonymized, 22% are partially occluded, and 30% are heavily occluded. For the remaining non-anonymized faces, 70% has a resolution smaller than $14x14$. Note that for each experiment in Table 1, we anonymize the same bounding boxes.

5.2 Ablation Experiments

We perform several ablation experiments to evaluate the model architecture choices. We report the Frèchet Inception Distance [7] between the original images and the anonymized images for each experiment. We calculate FID from a validation set of $50,000$ faces from the FDF dataset. The results are shown in Table 2 and discussed in detail next.

Effect of Pose Information: Pose of the face provided as conditional information improves our model significantly, as seen in Table 2a. The FDF dataset has a large variance of faces in different poses, and we find it necessary to include sparse pose information to generate realistic faces. In contrast, when trained on the Celeb-A dataset, our model completely ignores the given pose information.

Discriminator Architecture: Table 2b compares the quality of images for a deep and wide discriminator. With a deeper network, the discriminator struggles to converge, leading to poor results. We use no normalization layers in the discriminator, causing deeper networks to suffer from exploding forward

Table 2: Ablation Experiments with our model. We report the Frèchet Inception Distance (FID) on the FDF validation dataset, after showing the discriminator 30.0*M* images (lower is better). For results in Table 2a and Table 2b, we use a model size of 12*M* parameters for both the generator and discriminator. *Reported after 20.0*M* images, as the deep discriminator diverged after this.

(a) Result of using conditional pose.

Model	FID
With Pose	**2.71**
Without Pose	3.36

(b) Result of the deep and wide discriminator.

Discriminator	FID
Deep Discriminator*	9.327
Wide Discriminator*	**3.86**

(c) Result of different model sizes.

#parameters	FID
12M	2.71
46M	**1.84**

passes and vanishing gradients. Even though, Brock *et al.* [2] also observe similar results; a deeper network architecture degrades the overall image quality. Note that we also experimented with a discriminator with no modifications to number of parameters, but this was not able to generate realistic faces.

Model Size: We empirically observe that increasing the number of filters in each convolution improves image quality drastically. As seen in Table 2c, we train two models with 12*M* and 46*M* parameters. Unquestionably, increasing the number of parameters generally improves the image quality. For both experiments, we use the same hyperparameters; the only thing changed is the number of filters in each convolution.

6 Limitations

Our method proves its ability to generate objectively good images for a diversity of backgrounds and poses. However, it still struggles in several challenging scenarios. Figure 6 illustrates some of these. These issues can impact the generated image quality, but, by design, our model ensures the removal of all privacy-sensitive information from the face.

Faces occluded with high fidelity objects are extremely challenging when generating a realistic face. For example, in Figure 6, several images have persons covering their faces with hands. To generate a face in this scenario

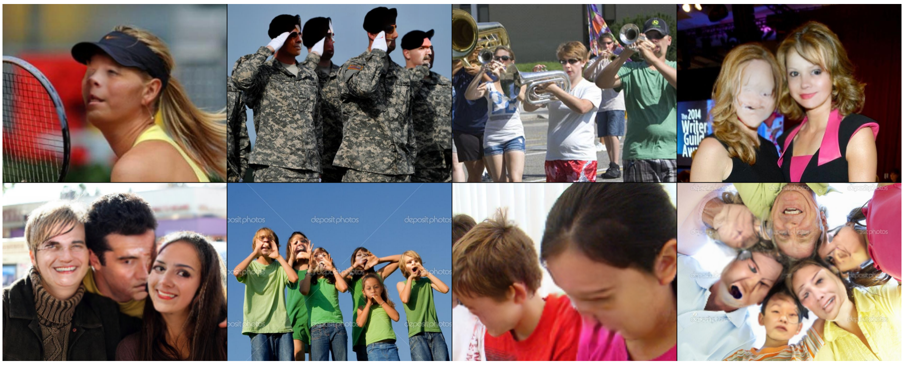


Figure 6: Failure Cases of DeepPrivacy Our proposed solution can generate unrealistic images in cases of high occlusion, difficult background information, and irregular poses.

requires complex semantic reasoning, which is still a difficult challenge for GANs.

Handling non-traditional poses can cause our model to generate corrupted faces. We use a sparse pose estimation to describe the face pose, but there is no limitation in our architecture to include a dense pose estimation. A denser pose estimation would, most likely, improve the performance of our model in cases of irregular poses. However, this would set restrictions on the pose estimator and restrict the practical use case of our method.

7 Conclusion

We propose a conditional generative adversarial network, *DeepPrivacy*, to anonymize faces in images without destroying the original data distribution. The presented results on the WIDER-Face dataset reflects our model's capability to generate high-quality images. Also, the diversity of images in the WIDER-Face dataset shows the practical applicability of our model. The current state-of-the-art face detection method can achieve 99.3% of the original average precision on the anonymized WIDER-Face validation set. In comparison to previous solutions, this is a significant improvement to both the generated image quality and the certainty of anonymization. Furthermore, the presented ablation experiments on the FDF dataset suggests that a larger model size and inclusion of sparse pose information is necessary to generate high-quality images.

DeepPrivacy is a conceptually simple generative adversarial network, easily extendable for further improvements. Handling irregular poses, difficult occlusions, complex backgrounds, and temporal consistency in videos is still a subject for further work. We believe our contribution will be an inspiration for further work into ensuring privacy in visual data.

Appendix A - Training Details

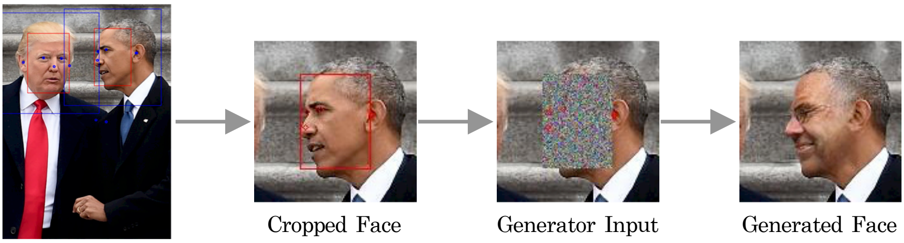
We use the same hyperparameters as Karras *et al.* [12], except the following: We use a batch size of 256, 256, 128, 72 and 48 for resolution 8, 16, 32, 64, and 128. We use a learning rate of 0.00175 with the Adam optimizer. For each expansion of the network, we have a transition and stabilization phase of 1.2M images each. We use an exponential running average for the weights of the generator as this improves overall image quality [28]. For the running average, we use a decay β given by:

$$\beta = 0.5^{\frac{B}{10^4}}, \tag{1}$$

where B is the batch size. Our final model was trained for 17 days on two NVIDIA V100-32GB GPUs.

Image Pre-Processing

Figure 7 shows the input pre-processing pipeline. For each detected face with a bounding box and keypoint detection, we find the smallest possible square bounding box which surrounds the face bounding box. Then, we resize the expanded bounding box to the target size (128×128). We replace the pixels within the face bounding box with a constant pixel value of 128. Finally, we shift the pixel values to the range $[-1,1]$.



Cropped Face

Generator Input

Generated Face

Figure 7: Input Pipeline: Each detected face is cropped to a quadratic image, then we replace the privacy-sensitive information with a constant value, and feed it to the generator. The keypoints are represented as a one-hot encoded image.

Tensor Core Modifications

To utilize tensor cores in NVIDIA's new Volta architecture, we do several modifications to our network, following the requirements of tensor cores. First, we ensure that each convolutional block use number of filters that are divisible by 8. Secondly, we make certain that the batch size for each GPU is divisible by 8. Further, we use automatic mixed precision for pytorch [21] to significantly improve our training time. We see an improvement of 220% in terms of training speed with mixed precision training.

References

[1] Boyle, M., Edwards, C., Greenberg, S.: The effects of filtered video on awareness and privacy. In: Proceedings of the 2000 ACM conference on Computer supported cooperative work. pp. 1–10. ACM (2000). https://doi.org/10.1145/358916.358935

[2] Brock, A., Donahue, J., Simonyan, K.: Large scale GAN training for high fidelity natural image synthesis. In: International Conference on Learning Representations (2019), `https://openreview.net/forum?id=B1xsqj09Fm`

[3] Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., Bengio, Y.: Generative adversarial nets. In: Ghahramani, Z., Welling, M., Cortes, C., Lawrence, N.D., Weinberger,

K.Q. (eds.) Advances in Neural Information Processing Systems 27. pp. 2672–2680. Curran Associates, Inc. (2014), http://papers.nips.cc/paper/5423-generative-adversarial-nets.pdf

[4] Gross, R., Sweeney, L., de la Torre, F., Baker, S.: Model-based face de-identification. In: 2006 Conference on Computer Vision and Pattern Recognition Workshop (CVPRW '06). IEEE (2006). https://doi.org/10.1109/cvprw.2006.125

[5] Gross, R., Sweeney, L., Cohn, J., de la Torre, F., Baker, S.: Face de-identification. In: Protecting Privacy in Video Surveillance, pp. 129–146. Springer London (2009). https://doi.org/10.1007/978-1-84882-301-3_8

[6] He, K., Gkioxari, G., Dollar, P., Girshick, R.: Mask r-CNN. In: 2017 IEEE International Conference on Computer Vision (ICCV). IEEE (oct 2017). https://doi.org/10.1109/iccv.2017.322

[7] Heusel, M., Ramsauer, H., Unterthiner, T., Nessler, B., Hochreiter, S.: Gans trained by a two time-scale update rule converge to a local nash equilibrium. In: Guyon, I., Luxburg, U.V., Bengio, S., Wallach, H., Fergus, R., Vishwanathan, S., Garnett, R. (eds.) Advances in Neural Information Processing Systems 30. pp. 6626–6637. Curran Associates, Inc. (2017)

[8] Huang, X., Belongie, S.: Arbitrary style transfer in real-time with adaptive instance normalization. In: 2017 IEEE International Conference on Computer Vision (ICCV). pp. 1501–1510. IEEE (oct 2017). https://doi.org/10.1109/iccv.2017.167

[9] Isola, P., Zhu, J.Y., Zhou, T., Efros, A.A.: Image-to-image translation with conditional adversarial networks. In: 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). IEEE (jul 2017). https://doi.org/10.1109/cvpr.2017.632

[10] Jo, Y., Park, J.: SC-FEGAN: Face Editing Generative Adversarial Network with User's Sketch and Color. arXiv preprint arXiv:1902.06838 (2019)

[11] Jourabloo, A., Yin, X., Liu, X.: Attribute preserved face de-identification. Proceedings of 2015 International Conference on Biometrics, ICB 2015 pp. 278–285 (2015). https://doi.org/10.1109/ICB.2015.7139096

[12] Karras, T., Aila, T., Laine, S., Lehtinen, J.: Progressive growing of gans for improved quality, stability, and variation. In: International Conference on Learning Representations (2018), `https://openreview.net/forum?id=Hk99zCeAb`

[13] Karras, T., Laine, S., Aila, T.: A style-based generator architecture for generative adversarial networks. In: 2019 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 4401–4410 (2019)

[14] Li, J., Wang, Y., Wang, C., Tai, Y., Qian, J., Yang, J., Wang, C., Li, J., Huang, F.: DSFD: Dual shot face detector. In: 2019 IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2019)

[15] Li, Y., Liu, S., Yang, J., Yang, M.H.: Generative face completion. In: 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 5892 – 5900. IEEE (jul 2017). https://doi.org/10.1109/cvpr.2017.624

[16] Lin, T.Y., Dollár, P., Girshick, R., He, K., Hariharan, B., Belongie, S.: Feature pyramid networks for object detection. In: 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 2117–2125. IEEE (jul 2017). https://doi.org/10.1109/cvpr.2017.106

[17] Liu, G., Reda, F.A., Shih, K.J., Wang, T.C., Tao, A., Catanzaro, B.: Image inpainting for irregular holes using partial convolutions. In: Proceedings of the European Conference on Computer Vision (ECCV). pp. 89–105. Springer International Publishing (2018). https://doi.org/10.1007/978-3-030-01252-6_6

[18] Mirza, M., Osindero, S.: Conditional generative adversarial nets. arXiv preprint arXiv:1411.1784 (2014)

[19] Neustaedter, C., Greenberg, S., Boyle, M.: Blur filtration fails to preserve privacy for home-based video conferencing. ACM Transactions on Computer-Human Interaction **13**(1), 1–36 (mar 2006). https://doi.org/10.1145/1143518.1143519

[20] Newton, E.M., Sweeney, L., Malin, B.: Preserving privacy by de-identifying face images. IEEE transactions on Knowledge and Data Engineering **17**(2), 232–243 (feb 2005). https://doi.org/10.1109/tkde.2005.32

[21] NVIDIA: A pytorch extension: Tools for easy mixed precision and distributed training in pytorch (2019), https://github.com/NVIDIA/apex

[22] Ren, Z., Lee, Y.J., Ryoo, M.S.: Learning to anonymize faces for privacy preserving action detection. In: Proceedings of the European Conference on Computer Vision (ECCV). pp. 639–655. Springer International Publishing (2018). https://doi.org/10.1007/978-3-030-01246-5_38

[23] Ronneberger, O., Fischer, P., Brox, T.: U-net: Convolutional networks for biomedical image segmentation. In: Lecture Notes in Computer Science. pp. 234–241. Springer International Publishing (2015). https://doi.org/10.1007/978-3-319-24574-4_28

[24] Ruder, M., Dosovitskiy, A., Brox, T.: Artistic style transfer for videos. In: German Conference on Pattern Recognition. pp. 26–36. Springer (2016). https://doi.org/10.1007/978-3-319-45886-1_3

[25] Sweeney, L.: k-anonymity: A model for protecting privacy. International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems **10**(05), 557–570 (2002)

[26] Thomee, B., Shamma, D.A., Friedland, G., Elizalde, B., Ni, K., Poland, D., Borth, D., Li, L.J.: YFCC100M: The new data in multimedia research. arXiv preprint arXiv:1503.01817 (2015), http://arxiv.org/abs/1503.01817

[27] Yang, S., Luo, P., Loy, C.C., Tang, X.: WIDER FACE: A face detection benchmark. In: 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). IEEE (jun 2016). https://doi.org/10.1109/cvpr.2016.596

[28] Yazıcı, Y., Foo, C.S., Winkler, S., Yap, K.H., Piliouras, G., Chandrasekhar, V.: The unusual effectiveness of averaging in GAN training. In: International Conference on Learning Representations (2019), https://openreview.net/forum?id=SJgw_sRqFQ

[29] Yeh, R.A., Chen, C., Lim, T.Y., Schwing, A.G., Hasegawa-Johnson, M., Do, M.N.: Semantic image inpainting with deep generative models. In: 2017 IEEE Conference on Computer Vision and

Pattern Recognition (CVPR). pp. 6882 – 6890. IEEE (jul 2017). https://doi.org/10.1109/cvpr.2017.728

[30] Zhang, H., Goodfellow, I., Metaxas, D., Odena, A.: Self-attention generative adversarial networks. In: Proceedings of the 36th International Conference on Machine Learning. vol. 97, pp. 7354–7563. PMLR (2019), `http://proceedings.mlr.press/v97/zhang19d.html`

[31] Zhang, H., Xu, T., Li, H., Zhang, S., Wang, X., Huang, X., Metaxas, D.: StackGAN: Text to photo-realistic image synthesis with stacked generative adversarial networks. Proceedings of the IEEE International Conference on Computer Vision pp. 5908–5916 (2017). https://doi.org/10.1109/ICCV.2017.629

[32] Zhang, S., Zhu, X., Lei, Z., Shi, H., Wang, X., Li, S.Z.: S^3FD: Single shot scale-invariant face detector. In: 2017 IEEE International Conference on Computer Vision (ICCV). pp. 192–201. IEEE (oct 2017). https://doi.org/10.1109/iccv.2017.30

Paper B

Image Inpainting with Learnable Feature Imputation

Authors:

Håkon Hukkelås, Frank Lindseth, Rudolf Mester

Published at conference:

42nd DAGM German Conference on Pattern Recognition 2020

Copyright:

Source Code:

`https://github.com/hukkelas/DeepPrivacy/blob/master/GCPR.md`

Appendix:

`https://bird.unit.no/resources/670be073-5b75-4ed2-ab11-089287d98040/content`

Paper B

Image Inpainting with Learnable Feature Imputation

Håkon Hukkelås, Frank Lindseth, Rudolf Mester

Department of Computer Science
Norwegian University of Science and Technology
Trondheim, Norway
{hakon.hukkelas, rudolf.mester, frankl}@ntnu.no

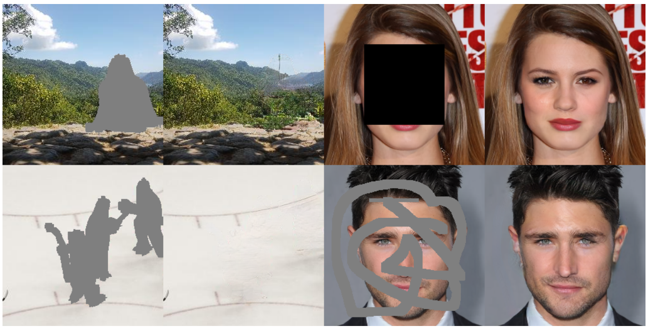


Figure 1: Masked images and corresponding generated images from our proposed single-stage generator.

Abstract

A regular convolution layer applying a filter in the same way over known and unknown areas causes visual artifacts in the inpainted image. Several studies address this issue with feature re-normalization on the output of the convolution. However, these models use a significant amount of learnable parameters for feature re-normalization [36, 42], or assume a binary representation of the certainty of an output [11, 25].

We propose (layer-wise) feature imputation of the missing input values to a convolution. In contrast to learned feature re-normalization [36, 42], our method is efficient and introduces a minimal number of parameters. Furthermore, we propose a revised gradient penalty for image inpainting, and a novel GAN architecture trained exclusively on adversarial loss. Our quantitative evaluation on the FDF dataset reflects that our revised gradient penalty and alternative convolution improves generated image quality significantly. We present comparisons on CelebA-HQ and Places2 to current state-of-the-art to validate our model.

1 Introduction

Image inpainting is the task of filling in missing areas of an image. Use cases for image inpainting are diverse, such as restoring damaged images, removing unwanted objects, or replacing information to preserve the privacy of individuals. Prior to deep learning, image inpainting techniques were generally examplar-based. For example, pattern matching, by searching and replacing with similar patches [4, 8, 22, 26, 33, 38], or diffusion-based, by smoothly propagating information from the boundary of the missing area [3, 5, 6].

Convolutional Neural Networks (CNNs) for image inpainting have led to significant progress in the last couple of years [1, 23, 37]. In spite of this, a standard convolution does not consider if an input pixel is missing or not, making it ill-fitted for the task of image inpainting. Partial Convolution (PConv) [25] propose a modified convolution, where they zero-out invalid (missing) input pixels and re-normalizes the output feature map depending on the number of valid pixels in the receptive field. This is followed by a hand-crafted certainty propagation step, where they assume an output is valid if one or more features in the receptive field are valid. Several proposed improvements replace the hand-crafted components in PConv with fully-learned components [36, 42]. However, these solutions use $\sim 50\%$ of the network parameters to propagate the certainties through the network.

We propose *Imputed Convolution (IConv)*; instead of re-normalizing the output feature map of a convolution, we replace uncertain input values with an estimate from spatially close features (see Figure 2). IConv assumes that a single spatial location (with multiple features) is associated with a single certainty. In contrast, previous solutions [36, 42] requires a certainty *for each feature* in a spatial location, which allocates half of the network parameters for certainty representation and propagation. Our simple assumption enables certainty representation and propagation to be minimal. In total, replacing all convolution layers with IConv increases the number of parameters by only $1-2\%$.

We use the DeepPrivacy [15] face inpainter as our baseline and suggest several improvements to stabilize the adversarial training: (1) We propose an improved version of gradient penalties to optimize Wasserstein GANs [2], based on the simple observation that standard gradient penalties causes training instability for image inpainting. (2) We combine the U-Net [30] generator with Multi-Scale-Gradient GAN (MSG-GAN) [19] to enable the discriminator to attend to multiple resolutions simultaneously, ensuring global and local consistency. (3) Finally, we replace the inefficient representation of the pose-information for the FDF dataset [15]. In contrast to the current state-of-the-art, our model requires no post-processing of generated images [16, 24], no refinement network [41, 42], or any additional loss term to stabilize the adversarial training [36, 42]. From our knowledge, our model is the first to be trained exclusively on adversarial loss for image-inpainting.

Our main contributions are the following:

1. We propose IConv which utilize a learnable feature estimator to impute uncertain input values to a convolution. This enables our model to generate visually pleasing images for free-form image inpainting.

2. We revisit the standard gradient penalty used to constrain Wasserstein GANs for image inpainting. Our simple modification significantly improves training stability and generated image quality at no additional computational cost.

3. We propose an improved U-Net architecture, enabling the adversarial training to attend to local and global consistency simultaneously.

2 Related Work

In this section, we discuss related work for generative adversarial networks (GANs), GAN-based image-inpainting, and the recent progress in free-form image-inpainting.

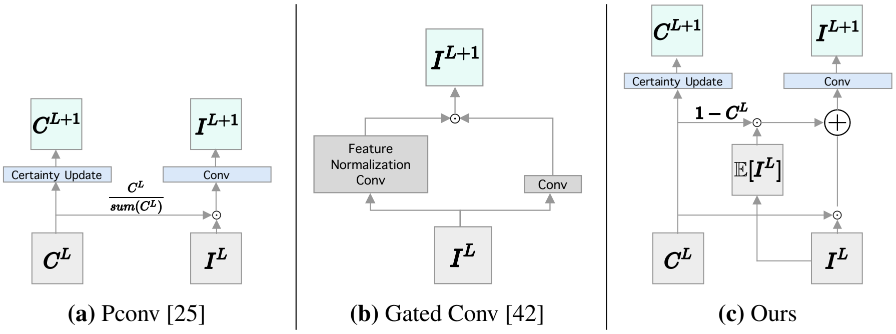
2.0.1 Generative Adversarial Networks

Generative Adversarial Networks [9] is a successful unsupervised training technique for image-based generative models. Since its conception, a range of techniques has improved convergence of GANs. Karras *et al.* [21] propose a *progressive growing* training technique to iteratively increase the network complexity to stabilize training. Karnewar *et al.* [19] replace progressive growing with Multi-Scale Gradient GAN (MSG-GAN), where they use skip connections between the matching resolutions of the generator and discriminator. Furthermore, Karras *et al.* [20] propose a modification of MSG-GAN in combination with residual connections [12]. Similar to [20], we replace progressive growing in the baseline model [15] with a modification of MSG-GAN for image-inpainting.

2.0.2 GAN-based Image Inpainting

GANs have seen wide adaptation for the image inpainting task, due to its astonishing ability to generate semantically coherent results for missing regions. There exist several studies proposing methods to ensure global and local consistency; using several discriminators to focus on different scales [16, 24], specific modules to connect spatially distant features [34, 39, 40, 41], patch-based discriminators [42, 43], multi-column generators [35], or progressively inpainting the missing area [11, 44]. In contrast to these methods, we ensure consistency over multiple resolutions by connecting different resolutions of the generator with the discriminator. Zheng *et al.* [46] proposes a probabilistic framework to address the issue of mode collapse for image inpainting, and they generate several plausible results for a missing area. Several methods propose combining the input image with auxiliary information, such as user sketches [17], edges [27], or examplar-based inpainting [7]. Hukkelås *et al.* [15] propose a U-Net based generator conditioned on the pose of the face.

GANs are notoriously difficult to optimize reliably [31]. For image inpainting, the adversarial loss is often combined with other objectives to improve training stability, such as pixel-wise reconstruction [7, 16, 24, 28], perceptual loss [34, 45], semantic loss [24], or style loss [36]. In contrast to these methods, we optimize exclusively on the adversarial loss. Furthermore, several studies [17, 35, 36, 41] propose to use Wasserstein GAN [2] with gradient penalties [10]; however, the standard gradient penalty causes training instability for image-inpainting models, as we discuss in Section 3.2.



(a) Pconv [25] **(b)** Gated Conv [42] **(c)** Ours

Figure 2: Illustration of partial convolution, gated convolution and our proposed solution. $\odot$ is element-wise product and $\oplus$ is addition. Note that C^L is binary for partial convolution.

2.0.3 Free-Form Image-Inpainting

Image Inpainting with irregular masks (often referred to as free-form masks) has recently caught more attention. Liu *et al.* [25] propose Partial Convolutions (PConv) to handle irregular masks, where they zero-out input values to a convolution and then perform feature re-normalization based on the number of valid pixels in the receptive field. Gated Convolution [42] modifies PConv by removing the binary-representation constraint, and they combine the mask and feature representation within a single feature map. Xie *et al.* [36] propose a simple modification to PConv, where they reformulate it as "attention" propagation instead of certainty propagation. Both of these PConv adaptations [36, 42] doubles the number of parameters in the network when replacing regular convolutions.

3 Method

In this section, we describe a) our modifications to a regular convolution layer, b) our revised gradient penalty suited for image inpainting, and c) our improved U-Net architecture.

3.1 Imputed Convolution (IConv)

Consider the case of a regular convolution applied to a given feature map $I \in \mathbb{R}^N$:

$$f(I) = W_F * I, \tag{1}$$

where $*$ is the convolution and $W_F \in \mathbb{R}^D$ is the filter. To simplify notation, we consider a single filter applied to a single one-dimensional feature map. The generalization to a regular multidimensional convolution layer is straightforward. A convolution applies this filter to all spatial locations of our feature map, which works well for general image recognition tasks. For image inpainting, there exists a set of known and unknown pixels; therefore, a regular convolution applied to all spatial locations is primarily undefined (“unknown" is not the same as 0 or any other fixed value), and naive approaches cause annoying visual artifacts [25].

We propose to replace the missing input values to a convolution with an estimate from spatially close values. To represent known and unknown values, we introduce a certainty C_x for each spatial location x, where $C \in \mathbb{R}^N$, and $0 \leq C_x \leq 1$. Note that this representation enables a single certainty to represent several values in the case of having multiple channels in the input. Furthermore, we define $\tilde{I}_x$ as a random variable with discrete outcomes $\{I_x, h_x\}$, where I_x is the feature at spatial location x, and h_x is an estimate from spatially close features. In this way, we want the output of our convolution to be given by,

$$O = \phi(f(\mathbb{E}[\tilde{I}_x])), \tag{2}$$

where ϕ is the activation function, and O the output feature map. We approximate the probabilities of each outcome using the certainty C_x; that is, $P(\tilde{I}_x = I_x) \approx C_x$ and $P(\tilde{I}_x = h_x) \approx 1 - C_x$, yielding the expected value of $\tilde{I}_x$,

$$\mathbb{E}[\tilde{I}_x] \ = C_x \cdot I_x + (1 - C_x) \cdot h_x. \tag{3}$$

We assume that a missing value can be approximated from spatially close values. Therefore, we define h_x as a learned certainty-weighted average of the surrounding features:

$$h_x = \frac{\sum_{i=1}^{K} I_{x+i} \cdot C_{x+i} \cdot \omega_i}{\sum_{i=1}^{K} C_{x+i}}, \tag{4}$$

where $\omega \in R^K$ is a learnable parameter. In a sense, our convolutional layer will try to learn the outcome space of $\tilde{I}_x$. Furthermore, h_x is efficient to implement in standard deep learning frameworks, as it can be implemented as a depth-wise separable convolution [32] with a re-normalization factor determined by C.

Propagating Certainties Each convolutional layer expects a certainty for each spatial location. We handle propagation of certainties as a learned operation,

$$C^{L+1} = \sigma(W_C * C^L), \tag{5}$$

where $*$ is a convolution, $W_C \in \mathbb{R}^D$ is the filter, and σ is the sigmoid function. We constraint W_C to have the same receptive field as f with no bias, and initialize C^0 to 0 for all unknown pixels and 1 else.

The proposed solution is minimal, efficient, and other components of the network remain close to untouched. We use LeakyReLU as the activation function ϕ, and average pooling and pixel normalization [21] after each convolution f. Replacing all convolutional layers with O_x in our baseline network increases the number of parameters by $\sim 1\%$. This is in contrast to methods based on learned feature re-normalization [36, 42], where replacing a convolution with their proposed solution doubles the number of parameters. Similar to partial convolution [25], we use a single scalar to represent the certainty for each spatial location; however, we do not constrain the certainty representation to be binary, and our certainty propagation is fully learned.

U-Net Skip Connection U-Net [30] skip connection is a method to combine shallow and deep features in encoder-decoder architectures. Generally, the skip connection consists of concatenating shallow and deep features, then followed by a convolution. However, for image inpainting, we only want to propagate certain features.

To find the combined feature map for an input in layer L and $L+l$, we find a weighted average. Assuming features from two layers in the network, (I^L, C^L), (I^{L+l}, C^{L+l}), we define the combined feature map as;

$$I^{L+l+1} = \gamma \cdot I^L + (1-\gamma) \cdot I^{L+l}, \tag{6}$$

and likewise for C^{L+l+1}. γ is determined by

$$\gamma = \frac{C^L \cdot \beta_1}{C^L \cdot \beta_1 + C^{L+l} \cdot \beta_2}, \tag{7}$$

where $\beta_1, \beta_2 \in \mathbb{R}^+$ are learnable parameters initialized to 1. Our U-Net skip connection is unique compared to previous work and designed for image inpainting. Equation 6 enables the network to only propagate features with a high certainty from shallow layers. Furthermore, we include β_1 and β_2 to give the model the flexibility to learn if it should attend to shallow or deep features.

3.2 Revisiting Gradient Penalties for Image Inpainting

Improved Wasserstein GAN [2, 10] is widely used in image inpainting [17, 35, 36, 41]. Given a discriminator D, the objective function for optimizing a Wasserstein GAN with gradient penalties is given by,

$$\mathscr{L}_{total} = \mathscr{L}_{adv} + \lambda \cdot (||\nabla D(\hat{x})||_p - 1)^2, \tag{8}$$

where $\mathscr{L}_{adv}$ is the adversarial loss, p is commonly set to 2 (L^2 norm), λ is the gradient penalty weight, and $\hat{x}$ is a randomly sampled point between the real image, x, and a generated image, $\tilde{x}$. Specifically, $\hat{x} = t \cdot x + (1-t) \cdot \tilde{x}$, where t is sampled from a uniform distribution [10].

Previous methods enforce the gradient penalty only for missing areas [17, 35, 41]. Given a mask M to indicate areas to be inpainted in the image x, where M is 0 for missing pixels and 1 otherwise (note that $M = C^0$), Yu *et al.* [41] propose the gradient penalty:

$$\bar{g}(\hat{x}) = (||\nabla D(\hat{x}) \odot (1-M)||_p - 1)^2, \tag{9}$$

where $\odot$ is element-wise multiplication. This gradient penalty cause significant training instability, as the gradient sign of $\bar{g}$ shifts depending on the cardinality of M. Furthermore, Equation 9 impose $||\nabla D(\hat{x})|| \approx 1$, which leads to a lower bound on the Wasserstein distance [18].

Imposing $||\nabla D(\hat{x})|| \leq 1$ will remove the issue of shifting gradients in Equation 9. Furthermore, imposing the constrain $||\nabla D(\hat{x})|| \leq 1$ is shown to properly

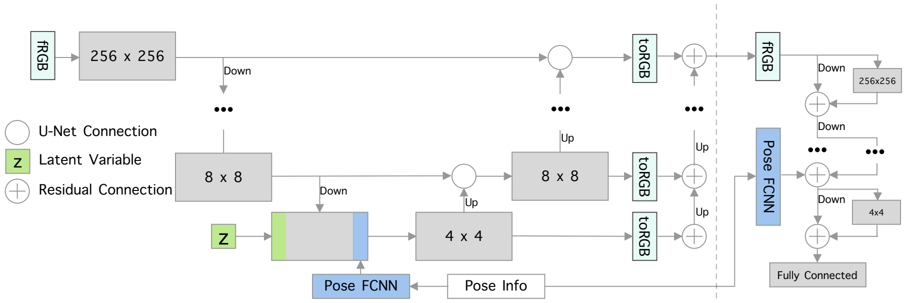


Figure 3: Illustration of the generator (left of the dashed line) and discriminator architecture. Up and down denotes nearest neighbor upsampling and average pool. The pose information in the discriminator is concatenated to the input of the first convolution layer with 32×32 resolution. Note that pose information is only used for the FDF dataset [15].

estimate the Wasserstein distance [18]. Therefore, we propose the following gradient penalty:

$$g(\hat{x}) = \max(0, ||\nabla D(\hat{x}) \odot (1-M))||_p - 1) \tag{10}$$

Previous methods enforce the L^2 norm [17, 35, 41]. Jolicoeur-Martineau *et al.* [18] suggest that replacing the L^2 gradient norm with L^∞ can improve robustness. From empirical experiments (see Appendix 1), we find L^∞ more unstable and sensitive to choice of hyperparameters; therefore, we enforce the L^2 norm (p=2).

In total, we optimize the following objective function:

$$\mathscr{L}_{total} = \mathscr{L}_{adv} + \lambda \cdot \max(0, ||\nabla D(\hat{x}) \odot (1-M))||_p - 1) \tag{11}$$

3.3 Model Architecture

We propose several improvements to the baseline U-Net architecture [15]. See Figure 3 for our final architecture. We replace all convolutions with Equation 2, average pool layer with a certainty-weighted average and U-Net skip connections with our revised skip connection (see Equation 6). Furthermore, we replace progressive growing training [21] with Multi-Scale Gradient GAN

(MSG-GAN) [19]. For the MSG-GAN, instead of matching different resolutions from the generator with the discriminator, we upsample each resolution and sum up the contribution of the RGB outputs [20]. In the discriminator we use residual connections, similar to [20]. Finally, we improve the representation of pose information in the baseline model (pose information is only used on the FDF dataset [15]).

Representation of Pose Information The baseline model [15] represents pose information as one-hot encoded images for each resolution in the network, which is extremely memory inefficient and a fragile representation. The pose information, $P \in \mathbb{R}^{K \cdot 2}$, represents K facial keypoints and is used as conditional information for the generator and discriminator. We propose to replace the one-hot encoded representation, and instead pre-process P into a $4 \times 4 \times 32$ feature bank using two fully-connected layers. This feature bank is concatenated with the features from the encoder. Furthermore, after replacing progressive growing with MSG-GAN, we include the same pose pre-processing architecture in the discriminator, and input the pose information as a $32 \times 32 \times 1$ feature map to the discriminator.

4 Experiments

We evaluate our proposed improvements on the Flickr Diverse Faces (FDF) dataset [15], a lower resolution (128×128) face dataset. We present experiments on the CelebA-HQ [21] and Places2 [47] datasets, which reflects that our suggestions generalizes to standard image inpainting. We compare against current state-of-the art [36, 42, 46, 29]. Finally, we present a set of ablation studies to analyze the generator architecture. [1]

[1] To prevent ourselves from cherry-picking qualitative examples, we present several images (with corresponding masks) chosen by previous state-of-the-art papers [11, 36, 42, 46], thus copying their selection. Appendix 5 describes how we selected these samples. The only hand-picked examples in this paper are Figure 1, Figure 5, Figure 6, and Figure 7. No examples in the Supplementary Material are cherry-picked.

Table 1: Quantitative results on the FDF dataset [15]. We report standard metrics after showing the discriminator 20 million images on the FDF and Places2 validation sets. We report L1, L2, and SSIM in Appendix 3. Note that Config E is trained with MSG-GAN, therefore, we separate it from Config A-D which are trained with progressive growing [21]. * Did not converge. † Same as Config B

Configuration		FDF			Places2		
		LPIPS ↓	PSNR ↑	FID ↓	LPIPS ↓	PSNR ↑	FID ↓
A	Baseline [15]	0.1036	22.52	6.15	–*	–*	–*
B	+ Improved Gradient penalty	0.0757	23.92	1.83	0.1619	20.99	7.96
C	+ Scalar Pose Information	0.0733	24.01	1.76	– †	–†	– †
D	+ Imputed Convolution	0.0739	23.95	1.66	0.1563	21.21	6.81
E	+ No Growing, MSG	**0.0728**	**24.01**	**1.49**	**0.1491**	**21.42**	**5.24**

Quantitative Metrics For quantitative evaluations, we report commonly used image inpainting metrics; pixel-wise distance (L1 and L2), peak signal-to-noise ratio (PSNR), and structural similarity (SSIM). Neither of these reconstruction metrics are any good indicators of generated image quality, as there often exist several possible solutions to a missing region, and they do not reflect human nuances [45]. Recently proposed deep feature metrics correlate better with human perception [45]; therefore, we report the Frèchet Inception Distance (FID) [13] (lower is better) and Learned Perceptual Image Patch Similarity (LPIPS) [45] (lower is better). We use LPIPS as the main quantitative evaluation.

4.1 Improving the Baseline

We iteratively add our suggestions to the baseline [15] (Config A-E), and report quantitative results in Table 1. First, we replace the gradient penalty term with Equation 10, where we use the L^2 norm ($p = 2$), and impose the following constraint (Config B):

$$G_{out} = G(I, C^0) \cdot (1 - C^0) + I \cdot C^0, \tag{12}$$

where C^0 is the binary input certainty and G is the generator. Note that we are not able to converge Config A while imposing G_{out}. We replace the one-hot encoded representation of the pose information with two fully connected layers in the generator (Config C). Furthermore, we replace the input to all convolutional layers with Equation 3 (Config D). We set the receptive field of h_x to 5×5 ($K = 5$ in Equation 4). We replace the progressive-growing training technique with MSG-GAN [19], and replace the one-hot encoded pose-information in the discriminator (Config E). These modifications combined improve the LPIPS score by *30.0%*. The authors of [15] report a FID of 1.84 on the FDF dataset with a model consisting of 46M learnable parameters. In comparison, we achieve a FID of 1.49 with 2.94M parameters (config E). For experimental details, see Appendix 2.

4.2 Generalization to Free-Form Image Inpainting

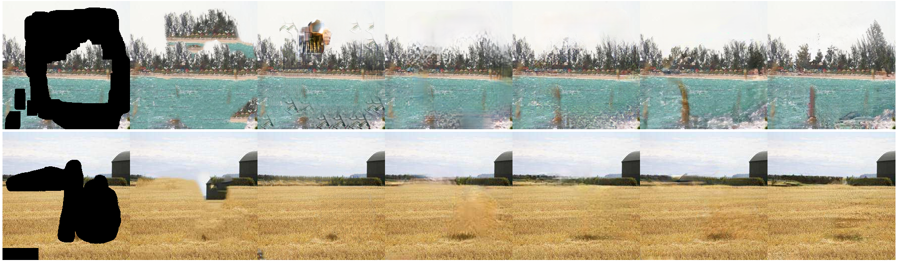
We extend Config E to general image inpainting datasets; CelebA-HQ [21] and Places2 [47]. We increase the number of filters in each convolution by a factor of 2, such that the generator has 11.5M parameters. In comparison, Gated Convolution [42] use 4.1M, LBAM [36] 68.3M, StructureFlow [29] 159M, and PIC [46] use 3.6M parameters. Compared to [42, 46], our increase in parameters improves semantic reasoning for larger missing regions. Also, compared to previous solutions, we achieve similar inference time since the majority of the parameters are located at low-resolution layers (8×8 and 16×16). In contrast, [42] has no parameters at a resolution smaller than 64×64. For single-image inference time, our model matches (or outperforms) previous models; on a single NVIDIA 1080 GPU, our network runs at ~ 89 ms per image on 256×256 resolution, 2× faster than LBAM [36], and PIC [46]. GatedConvolution [42] achieves ~ 62 ms per image. [2] See Appendix 2.1 for experimental details.

Quantitative Results Table 2 shows quantitative results for the CelebA-HQ and Places2 datasets. For CelebA-HQ, we improve LPIPS and FID significantly compared to previous models. For Places2, we achieve comparable results to [42] for free-form and center-crop masks. Furthermore, we compare our model with and without IConv and notice a significant improvement in generated image quality (see Figure 1 in Appendix 3). See Appendix 5.1 for examples of the center-crop and free-form images.

[2] We measure runtime for [42, 46] with their open-source code, as they do not report inference time for 256×256 resolution in their paper.

Table 2: Quantitative results on the CelebA-HQ and Places2 datasets. We use the official frameworks to reproduce results from [42, 46]. For the (Center) dataset we use a 128 × 128 center mask, and for (Free-Form) we generate free-form masks for each image following the approach in [42]. We report L1, L2, and SSIM in Appendix 3.

Method	Places2 (Center)			Places2 (Free Form)			CelebA-HQ (Center)			CelebA-HQ (Free Form)		
	PSNR	LPIPS	FID	PSNR	LPIPS	FID	PSNR	LPIPS	FID	PSNR	LPIPS	FID
Gated Convolutions [42]	21.56	**0.1407**	4.14	**27.59**	**0.0579**	0.90	**25.55**	0.0587	6.05	30.26	0.0366	2.98
Plurastic Image Inpainting [46]	21.04	0.1584	7.23	26.66	0.0804	2.76	24.59	0.0644	7.50	29.30	0.0394	3.30
Ours	**21.70**	0.1412	**3.99**	27.33	0.0597	0.94	25.29	**0.0522**	**4.43**	**30.32**	**0.0300**	**2.38**



(a) Input **(b)** PM [4] **(c)** PIC [46] **(d)** PC [25] **(e)** BA [36] **(f)** GC [42] **(g)** Ours

Figure 4: Places2 comparison to PatchMatch (PM) [4], Pluralistic Image Completion (PIC) [46], Partial Convolution (PC) [25], Bidirectional Attention (BA) [36], and Gated Convolution (GC) [42]. Examples selected by authors of [36] (images extracted from their supplementary material). Results of [42, 46] generated by using their open-source code and models. We recommend the reader to zoom-in on missing regions.

Qualitative Results Figure 5 shows a set of hand-picked examples, Figure 4 shows examples selected by [36], and Appendix 5 includes a large set of examples selected by the authors of [11, 36, 42, 46]. We notice less visual artifacts than models using vanilla convolutions [46, 29], and we achieve comparable results to Gated Convolution [42] for free-form image inpainting. For larger missing areas, our model generates more semantically coherent results compared to previous solutions [11, 36, 42, 46].

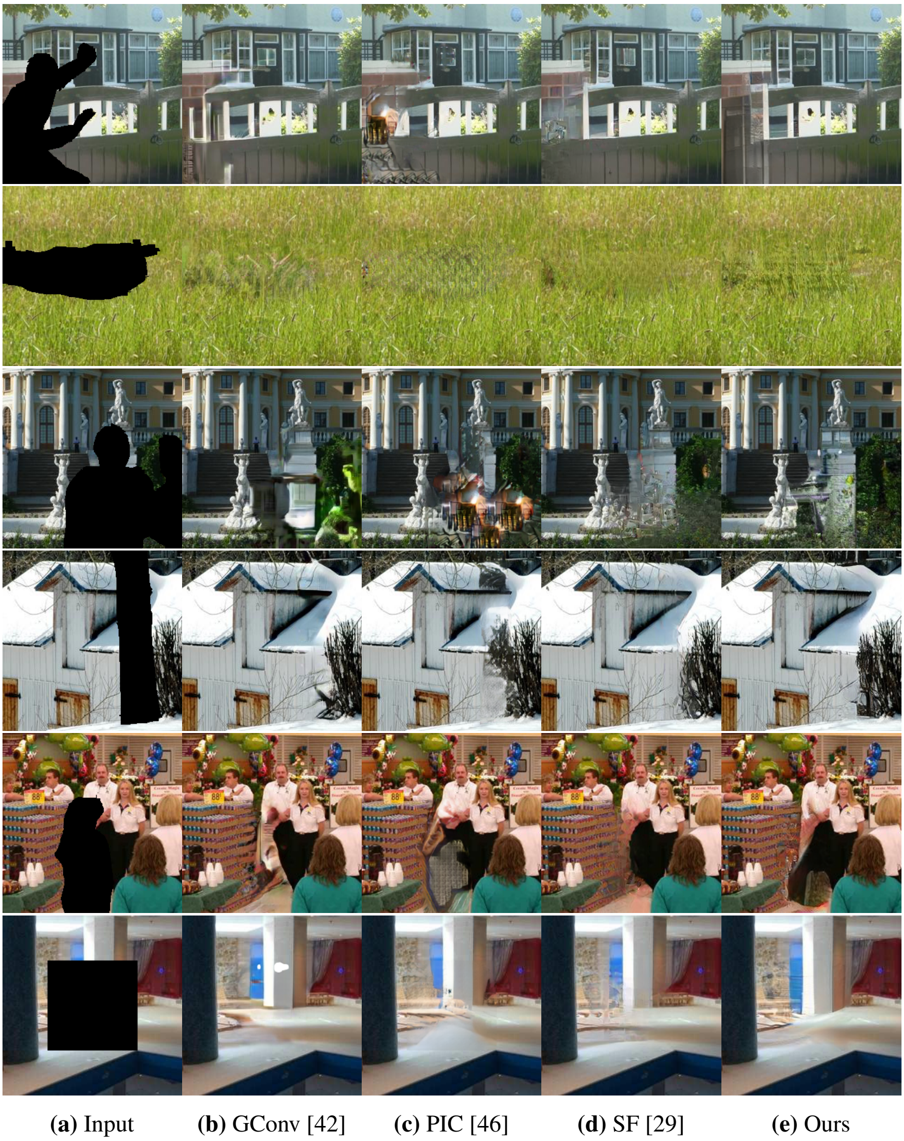


Figure 5: Qualitative examples on the Places2 validation set with comparisons to Gated Convolution (GConv) [42], StructureFlow (SF) [29], and Pluralistic Image Completion (PIC) [46]. We recommend the reader to zoom-in on missing regions. For non hand-picked qualitative examples, see Appendix 5.

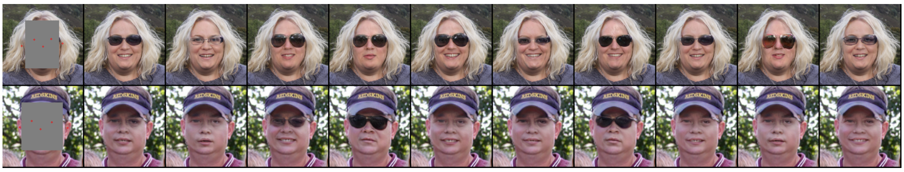


Figure 6: Diverse Plausible Results: Images from the FDF validation set [15]. Left column is the input image with the pose information marked in red. Second column and onwards are different plausible generated results. Each image is generated by randomly sampling a latent variable for the generator (except for the second column where the latent variable is set to all 0's). For more results, see Appendix 6.

4.3 Ablation Studies

Pluralistic Image Inpainting Generating different possible results for the same conditional image (pluralistic inpainting) [46] has remained a problem for conditional GANs [14, 48]. Figure 6 illustrates that our proposed model (Config E) generates multiple and diverse results. Even though, for Places2, we observe that our generator suffers from mode collapse early on in training. Therefore, we ask the question; *does a deterministic generator impact the generated image quality for image-inpainting?* To briefly evaluate the impact of this, we train Config D without a latent variable, and observe a 7% degradation in LPIPS score on the FDF dataset. We leave further analysis of this for further work.

Propagation of Certainties Figure 7 visualizes if the generator attends to shallow or deep features in our encoder-decoder architecture. Our proposed U-Net skip connection enables the network to select features between the encoder and decoder depending on the certainty. Notice that our network attends to deeper features in cases of uncertain features, and shallower feature otherwise.

5 Conclusion

We propose a simple single-stage generator architecture for free-form image inpainting. Our proposed improvements to GAN-based image inpainting

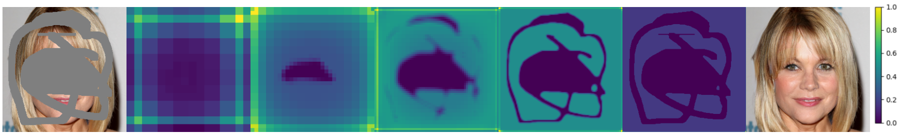


Figure 7: U-Net Skip Connections. Visualization of γ from Equation 6. The left image is the input image, second column and onwards are the values of γ for resolution 8 to 256. Rightmost image is the generated image. Smaller values of γ indicates that the network selects deep features (from the decoder branch).

significantly stabilizes adversarial training, and from our knowledge, we are the first to produce state-of-the-art results by exclusively optimizing an adversarial objective. Our main contributions are; a revised convolution to properly handle missing values in convolutional neural networks, an improved gradient penalty for image inpainting which substantially improves training stability, and a novel U-Net based GAN architecture to ensure global and local consistency. Our model achieves state-of-the-art results on the CelebA-HQ and Places2 datasets, and our single-stage generator is much more efficient compared to previous solutions.

Acknowledgements.

The computations were performed on resources provided by the Tensor-GPU project led by Prof. Anne C. Elster through support from The Department of Computer Science and The Faculty of Information Technology and Electrical Engineering, NTNU. Furthermore, Rudolf Mester acknowledges the support obtained from DNV GL.

References

[1] Aghdam, H.H., Heravi, E.J.: Convolutional neural networks. In: Guide to Convolutional Neural Networks, pp. 85–130. Springer International Publishing (2017). https://doi.org/10.1007/978-3-319-57550-6_3

[2] Arjovsky, M., Chintala, S., Bottou, L.: Wasserstein gan. arXiv preprint arXiv:1701.07875 (2017)

[3] Ballester, C., Bertalmio, M., Caselles, V., Sapiro, G., Verdera, J.: Filling-in by joint interpolation of vector fields and gray levels. IEEE Transactions on Image Processing **10**(8), 1200–1211 (2001). https://doi.org/10.1109/83.935036

[4] Barnes, C., Shechtman, E., Finkelstein, A., Goldman, D.B.: PatchMatch. In: ACM SIGGRAPH 2009 papers on - SIGGRAPH 09. ACM Press (2009). https://doi.org/10.1145/1576246.1531330

[5] Bertalmio, M., Sapiro, G., Caselles, V., Ballester, C.: Image inpainting. In: Proceedings of the 27th annual conference on Computer graphics and interactive techniques. pp. 417–424 (2000)

[6] Criminisi, A., Perez, P., Toyama, K.: Region filling and object removal by exemplar-based image inpainting. IEEE Transactions on Image Processing **13**(9), 1200–1212 (sep 2004). https://doi.org/10.1109/tip.2004.833105

[7] Dolhansky, B., Ferrer, C.C.: Eye in-painting with exemplar generative adversarial networks. In: 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. IEEE (jun 2018). https://doi.org/10.1109/cvpr.2018.00824

[8] Efros, A.A., Freeman, W.T.: Image quilting for texture synthesis and transfer. In: Proceedings of the 28th annual conference on Computer graphics and interactive techniques - SIGGRAPH 01. ACM Press (2001). https://doi.org/10.1145/383259.383296

[9] Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., Bengio, Y.: Generative adversarial nets.

In: Advances in neural information processing systems. pp. 2672–2680 (2014)

[10] Gulrajani, I., Ahmed, F., Arjovsky, M., Dumoulin, V., Courville, A.C.: Improved training of wasserstein gans. In: Advances in neural information processing systems. pp. 5767–5777 (2017)

[11] Guo, Z., Chen, Z., Yu, T., Chen, J., Liu, S.: Progressive image inpainting with full-resolution residual network. In: Proceedings of the 27th ACM International Conference on Multimedia. ACM (oct 2019). https://doi.org/10.1145/3343031.3351022

[12] He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). IEEE (jun 2016). https://doi.org/10.1109/cvpr.2016.90

[13] Heusel, M., Ramsauer, H., Unterthiner, T., Nessler, B., Hochreiter, S.: Gans trained by a two time-scale update rule converge to a local nash equilibrium. In: Advances in Neural Information Processing Systems. pp. 6626–6637 (2017)

[14] Huang, X., Liu, M.Y., Belongie, S., Kautz, J.: Multimodal unsupervised image-to-image translation. In: Proceedings of the European Conference on Computer Vision (ECCV). pp. 172–189 (2018)

[15] Hukkelås, H., Mester, R., Lindseth, F.: Deepprivacy: A generative adversarial network for face anonymization. In: Advances in Visual Computing. pp. 565–578. Springer International Publishing (2019). https://doi.org/10.1007/978-3-030-33720-9_44

[16] Iizuka, S., Simo-Serra, E., Ishikawa, H.: Globally and locally consistent image completion. ACM Transactions on Graphics **36**(4), 1–14 (jul 2017). https://doi.org/10.1145/3072959.3073659

[17] Jo, Y., Park, J.: Sc-fegan: Face editing generative adversarial network with user's sketch and color. In: The IEEE International Conference on Computer Vision (ICCV) (October 2019). https://doi.org/10.1109/ICCV.2019.00183

[18] Jolicoeur-Martineau, A., Mitliagkas, I.: Connections between support vector machines, wasserstein distance and gradient-penalty gans. arXiv preprint arXiv:1910.06922 (2019)

[19] Karnewar, A., Wang, O., Iyengar, R.S.: Msg-gan: multi-scale gradient gan for stable image synthesis. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. vol. 6 (2019). https://doi.org/10.1109/CVPR42600.2020.00782

[20] Karras, T., Laine, S., Aittala, M., Hellsten, J., Lehtinen, J., Aila, T.: Analyzing and improving the image quality of stylegan. In: 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 8107–8116 (2020). https://doi.org/10.1109/CVPR42600.2020.00813

[21] Karras, T., Aila, T., Laine, S., Lehtinen, J.: Progressive growing of GANs for improved quality, stability, and variation. In: International Conference on Learning Representations (2018)

[22] Kwatra, V., Essa, I., Bobick, A., Kwatra, N.: Texture optimization for example-based synthesis. In: ACM SIGGRAPH 2005 Papers on - SIGGRAPH 05. ACM Press (2005). https://doi.org/10.1145/1186822.1073263, `https://doi.org/10.1145%2F1186822.1073263`

[23] Köhler, R., Schuler, C., Schölkopf, B., Harmeling, S.: Mask-specific inpainting with deep neural networks. In: Lecture Notes in Computer Science, pp. 523–534. Springer International Publishing (2014). https://doi.org/10.1007/978-3-319-11752-2_43

[24] Li, Y., Liu, S., Yang, J., Yang, M.H.: Generative face completion. In: 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 5892 – 5900. IEEE (jul 2017). https://doi.org/10.1109/cvpr.2017.624

[25] Liu, G., Reda, F.A., Shih, K.J., Wang, T.C., Tao, A., Catanzaro, B.: Image inpainting for irregular holes using partial convolutions. In: Computer Vision – ECCV 2018, pp. 89–105. Springer International Publishing (2018). https://doi.org/10.1007/978-3-030-01252-6_6

[26] Meur, O.L., Gautier, J., Guillemot, C.: Examplar-based inpainting based on local geometry. In: 2011 18th IEEE International Conference on Image Processing. IEEE (sep 2011). https://doi.org/10.1109/icip.2011.6116441

[27] Nazeri, K., Ng, E., Joseph, T., Qureshi, F.Z., Ebrahimi, M.: Edgeconnect: Generative image inpainting with adversarial edge learning. arXiv preprint arXiv:1901.00212 (2019)

[28] Pathak, D., Krahenbuhl, P., Donahue, J., Darrell, T., Efros, A.A.: Context encoders: Feature learning by inpainting. In: 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). IEEE (jun 2016). https://doi.org/10.1109/cvpr.2016.278

[29] Ren, Y., Yu, X., Zhang, R., Li, T.H., Liu, S., Li, G.: StructureFlow: Image inpainting via structure-aware appearance flow. In: 2019 IEEE/CVF International Conference on Computer Vision (ICCV). IEEE (oct 2019). https://doi.org/10.1109/iccv.2019.00027

[30] Ronneberger, O., Fischer, P., Brox, T.: U-net: Convolutional networks for biomedical image segmentation. In: International Conference on Medical image computing and computer-assisted intervention. pp. 234–241. Springer (2015). https://doi.org/10.1007/978-3-319-24574-4_28

[31] Salimans, T., Goodfellow, I., Zaremba, W., Cheung, V., Radford, A., Chen, X.: Improved techniques for training gans. In: Advances in neural information processing systems. pp. 2234–2242 (2016)

[32] Sifre, L., Mallat, S.: Rigid-motion scattering for image classification. Ph. D. thesis (2014)

[33] Simakov, D., Caspi, Y., Shechtman, E., Irani, M.: Summarizing visual data using bidirectional similarity. In: 2008 IEEE Conference on Computer Vision and Pattern Recognition. IEEE (jun 2008). https://doi.org/10.1109/cvpr.2008.4587842

[34] Song, Y., Yang, C., Lin, Z., Liu, X., Huang, Q., Li, H., Kuo, C.C.J.: Contextual-based image inpainting: Infer, match, and translate. In: Computer Vision – ECCV 2018, pp. 3–18. Springer International Publishing (2018). https://doi.org/10.1007/978-3-030-01216-8_1

[35] Wang, Y., Tao, X., Qi, X., Shen, X., Jia, J.: Image inpainting via generative multi-column convolutional neural networks. In: Advances in neural information processing systems. pp. 331–340 (2018)

[36] Xie, C., Liu, S., Li, C., Cheng, M.M., Zuo, W., Liu, X., Wen, S., Ding, E.: Image inpainting with learnable bidirectional attention maps. In: Proceedings of the IEEE International Conference on Computer Vision. pp. 8858–8867 (2019). https://doi.org/10.1109/ICCV.2019.00895

[37] Xie, J., Xu, L., Chen, E.: Image denoising and inpainting with deep neural networks. In: Advances in neural information processing systems. pp. 341–349 (2012)

[38] Xu, Z., Sun, J.: Image inpainting by patch propagation using patch sparsity. IEEE Transactions on Image Processing **19**(5), 1153–1165 (may 2010). https://doi.org/10.1109/tip.2010.2042098

[39] Yan, Z., Li, X., Li, M., Zuo, W., Shan, S.: Shift-net: Image inpainting via deep feature rearrangement. In: Computer Vision – ECCV 2018, pp. 3–19. Springer International Publishing (2018). https://doi.org/10.1007/978-3-030-01264-9_1

[40] Yang, C., Lu, X., Lin, Z., Shechtman, E., Wang, O., Li, H.: High-resolution image inpainting using multi-scale neural patch synthesis. In: 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). IEEE (jul 2017). https://doi.org/10.1109/cvpr.2017.434

[41] Yu, J., Lin, Z., Yang, J., Shen, X., Lu, X., Huang, T.S.: Generative image inpainting with contextual attention. In: 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. IEEE (jun 2018). https://doi.org/10.1109/cvpr.2018.00577

[42] Yu, J., Lin, Z., Yang, J., Shen, X., Lu, X., Huang, T.S.: Free-form image inpainting with gated convolution. In: Proceedings of the IEEE International Conference on Computer Vision. pp. 4471–4480 (2019). https://doi.org/10.1109/ICCV.2019.00457

[43] Zeng, Y., Fu, J., Chao, H., Guo, B.: Learning pyramid-context encoder network for high-quality image inpainting. In: 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). IEEE (jun 2019). https://doi.org/10.1109/cvpr.2019.00158

[44] Zhang, H., Hu, Z., Luo, C., Zuo, W., Wang, M.: Semantic image inpainting with progressive generative networks. In: 2018 ACM Multimedia Conference on Multimedia Conference. ACM Press (2018). https://doi.org/10.1145/3240508.3240625

[45] Zhang, R., Isola, P., Efros, A.A., Shechtman, E., Wang, O.: The unreasonable effectiveness of deep features as a perceptual metric. In: Proceedings of the IEEE conference on computer vision and pattern recognition (2018). https://doi.org/10.1109/CVPR.2018.00068

[46] Zheng, C., Cham, T.J., Cai, J.: Pluralistic image completion. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 1438–1447 (2019). https://doi.org/10.1109/CVPR.2019.00153

[47] Zhou, B., Lapedriza, A., Khosla, A., Oliva, A., Torralba, A.: Places: A 10 million image database for scene recognition. IEEE Transactions on Pattern Analysis and Machine Intelligence (2017). https://doi.org/10.1109/TPAMI.2017.2723009

[48] Zhu, J.Y., Zhang, R., Pathak, D., Darrell, T., Efros, A.A., Wang, O., Shechtman, E.: Toward multimodal image-to-image translation. In: Advances in neural information processing systems. pp. 465–476 (2017)

Paper C

Realistic Full-Body Anonymization with Surface-Guided GANs

Authors:

Håkon Hukkelås, Morten Smebye, Rudolf Mester, Frank Lindseth

Published at conference:

2023 IEEE/CVF Winter Conference on Applications of Computer Vision

Copyright:

Source Code:

`https://github.com/hukkelas/full_body_anonymization`

Appendix:

`https://openaccess.thecvf.com/content/WACV2023/supplemental/Hukkelas_Realistic_Full-Body_Anonymization_WACV_2023_supplemental.pdf`

Paper C

Realistic Full-Body Anonymization with Surface-Guided GANs

Håkon Hukkelås Morten Smebye Rudolf Mester Frank Lindseth

Department of Computer Science
Norwegian University of Science and Technology
Trondheim, Norway
hakon.hukkelas@ntnu.no

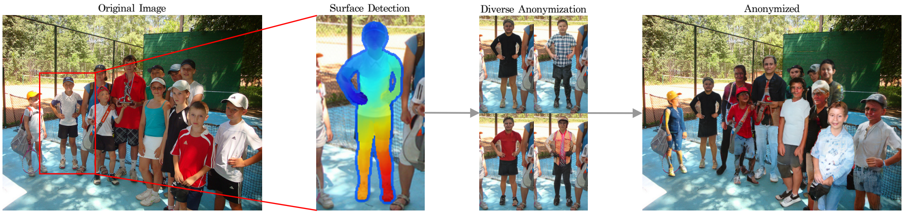


Figure 1: Our model performs in-the-wild anonymization by first detecting pixel-to-surface correspondences with CSE [41], then Surface-Guided GAN individually anonymizes each person. Original image from COCO [32].

Abstract

Recent work on image anonymization has shown that generative adversarial networks (GANs) can generate near-photorealistic faces to anonymize individuals. However, scaling up these networks to the entire human body has remained a challenging and yet unsolved task. We propose a new anonymization method that generates realistic humans for in-the-wild images. A key part of our design is to guide adversarial nets by dense pixel-to-surface correspondences between an image and a canonical 3D surface. We introduce Variational Surface-Adaptive Modulation (V-SAM) that embeds surface information throughout the generator. Combining this with our novel discriminator surface supervision loss, the generator can synthesize high quality humans with diverse appearances in complex and varying scenes. We demonstrate that surface guidance significantly improves image quality and diversity of samples, yielding a highly practical generator. Finally, we show that our method preserves data usability without infringing privacy when collecting image datasets for training computer vision models. Source code and appendix is available at: github.com/hukkelas/full_body_anonymization

1 Introduction

Privacy regulations constitute a significant obstacle against using image data taken in public for training computer vision algorithms. Recent work reflects that generative adversarial networks (GANs) [21, 35, 50] can realistically anonymize faces, where the anonymized datasets perform similarly to the original for future computer vision development. However, these methods [21, 35, 50] focus solely on face anonymization, leaving several primary identifiers(*e.g.* ears [23]) and soft identifiers (*e.g.* gender) on the human body untouched.

Generative adversarial networks are great at synthesizing high-resolution images in many domains, including humans [27]. Despite this success, previous work on full-body generative modeling focuses on simplified tasks, such as motion transfer [6], pose transfer [3, 31], garment swapping [16], or rendering a body with known 3D structure of the scene [55]. These methods do not directly apply to in-the-wild anonymization, as they do not handle variations in the background. As far as we know, our work is the first to address the task of synthesizing humans into in-the-wild images without simplifying the task (*e.g.* having a source texture to transfer, known 3D structure of the scene, or assuming a static background) [1].

Our contributions address the unexplored and challenging task of full-body anonymization for in-the-wild images. Our goal is to ensure the privacy of the anonymized individual; thus, we pose the anonymization task as an image inpainting problem. Modeling anonymization as image inpainting has stronger privacy guarantees than previous human synthesis methods, which rely on a source body texture or the original identity.

In this work, we propose *Surface-guided GANs* that utilize Continuous Surface Embeddings (CSE) [41] to guide the generator with pixel-to-surface correspondences. The compact, high-fidelity, and continuous representation of CSE excels for synthesizing human figures, as it allows for simple modeling choices without compromising fine-grained details. We show that surface guidance

[1] Although, we note that CIAGAN [35] ablates their method for low-resolution human synthesis.

significantly improves image quality, whereas current state-of-the-art GANs struggle with generating human figures without it.

We summarize our contributions into three points.

First, to efficiently utilize the powerful CSE representation, we propose *Variational Surface Adaptive Modulation (V-SAM)*. V-SAM projects the input latent space of the generator to an intermediate surface-adaptive latent space. This allows the generator to directly map the latent factors of variations to relevant surface locations (*e.g.* relate "red shirt" to the upper body independent of its spatial position), resulting in a latent space disentangled from the spatial image. The explicit disentangled representation is unique to V-SAM, which significantly improves latent disentanglement and image fidelity compared to previous spatially-invariant [27, 62] and spatial-adaptive modulation [43].

Secondly, we propose *Discriminator Surface Supervision* that incentivizes the discriminator to learn pixel-to-surface correspondences. The surface awareness of the discriminator provides higher-fidelity feedback to the generator, which significantly improves image quality. In fact, we find that the surface-aware feedback from the discriminator is a key factor to the powerful representation learned by V-SAM, where similar semantic-based supervision [48] yields suboptimal results.

Thirdly, we present a novel full-body anonymization framework that produces close-to-photorealistic images. We demonstrate that surface-guided anonymization significantly improves upon traditional methods (*e.g.* pixelation) in terms of data usability and privacy. For example, pixelation degrades the person average precision by 14.4 for Mask R-CNN [17] instance segmentation. In contrast, surface-guided anonymization yields only a 2.8 degradation.

2 Related Work

Anonymization of Images Naive anonymization methods that apply simple image distortions (*e.g.* blurring) are known to be inadequate for removing privacy-sensitive information [14, 39], and severely distorts the data. Recent work reflects that deep generative models can realistically anonymize faces by inpainting [2, 21, 35, 50, 51] or transforming the original image [10]. These

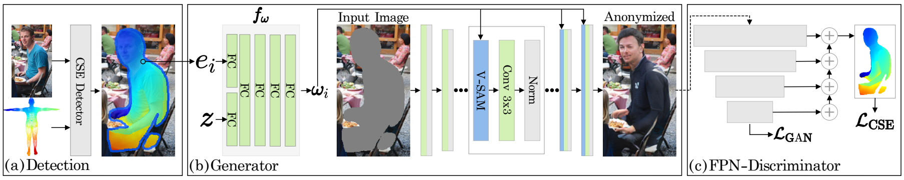


Figure 2: (a) A CSE-detector [41] predicts pixel-to-surface correspondences represented as a continuous positional embedding e_i. For simplicity, we show the pipeline with a single person, but multi-person detection is done by cropping each person (see Figure 1). **(b)** The mapping network (f_ω) transform surface locations and the latent variable ($z \sim \mathcal{N}(0,1)$) into an intermediate surface-adaptive latent space (ω_i) (section 3.1, section 3.2). Then, w_i controls the generator with pixel-wise modulation and normalization after each convolution. **(c)** Our FPN-discriminator predicts the surface embedding and optimizes a surface-regression loss ($\mathscr{L}_{\text{CSE}}$, section 3.3) along with the adversarial loss ($\mathscr{L}_{\text{GAN}}$).

methods demonstrate that retaining the original data distribution is important for future computer vision development (*e.g*. evaluation of face detection [21]). However, prior work focuses on face anonymization, leaving several primary and secondary identifiers untouched. Some methods anonymize the entire body [4, 35], but these methods are limited to low-resolution images [35] or generate images with visual artifacts [4].

Conditional Image Synthesis Current state-of-the-art for conditional image synthesis generates highly realistic images in many domains, such as image-to-image translation [22, 48]. An emerging approach is to introduce conditional information to the generator via *adaptive modulation* (also known as adaptive normalization [19]). This is known to be effective for unconditional synthesis [27], semantic synthesis [43], and style transfer [19]. Adaptive modulation conditions the generator by layer-wise shifting and scaling feature maps of the generator, where the shifting and scaling parameters are adaptive with respect to the condition. In contrast to prior semantic-modulation methods [43, 52, 53], V-SAM conditions the modulation parameters on dense surface information and generates global modulation parameters instead of independent layer-wise parameters. Conditional modulation is adapted for human synthesis, where prior methods adapt spatially-invariant [36, 46], or spatially-variant modulation [1, 59]. However, these methods are conditioned on a source appearance, yielding softer privacy guarantees compared to V-SAM.

Human Synthesis Prior work for person image generation often focus on resynthesizing humans with user-guided input, such as rendering persons in novel poses [3, 31], with different garments [16], or with a new motion [6]. Recent work [7, 13, 30, 40, 47] employ dense pixel-to-surface correspondences in the form of DensePose UV-maps [15]. These methods "fill in" UV texture maps, then render the person in new camera views [7] or poses [13, 30, 40, 47]. In contrast, CSE is a much more compact representation, and the continuous representation eases modeling complexity (*e.g.* downsampling of DensePose is not straightforward) and removes the need to handle borders. In other cases, the aim is to reconstruct the 3D surface and texture [38, 45, 55], which can be rendered to the scene given a camera view [55]. A limited amount of work focuses on human synthesis without a source image, where Ma *et al.* [34] maps background, pose, and person style into Gaussian variables, enabling synthesis of novel persons. None of these methods are directly applicable for human anonymization, as they require information about a source identity or the camera position to render the person. Additionally, none of them account for modeling background variations in the scene, which is the challenge of in-the-wild anonymization.

3 Method

We describe the anonymization task as an inpainting task. The objective of the generator is to inpaint the missing regions in the image $I \odot M$, where $M_i = 0$ for missing pixels and 1 otherwise. For each missing pixel, the surface embedding $e_i \in \mathbb{R}^{16}$ (the output of a CSE-detector [41]) represents the position of pixel i on a canonical 3D surface S (*i.e.* the position on a "T-shaped" human body). The surface S is discretized with $27K$ vertices, where each vertex has a positional embedding e_k obtained from the CSE-detector [41]. From this, pixel-to-vertex correspondences are found from euclidean nearest neighbor search between e_i and e_k [2]. Figure 2 shows the overall architecture.

[2] Finding pixel-to-vertex correspondence is not strictly necessary. However, replacing the regressed embedding e_i with the nearest e_k prohibits the generator from directly observing

3.1 Surface Adaptive Modulation

Inspired by the effectiveness of semantic-adaptive modulation [43], we introduce *Surface Adaptive Modulation* (SAM). SAM normalizes and modulates convolutional feature maps with respect to dense pixel-to-surface correspondences between the image and a fixed 3D surface. Given the continuous positional embedding e_i, a non-linear mapping f_ω transforms e_i to an intermediate surface-adaptive representation ω_i;

$$\omega_i = \begin{cases} f_\omega(e_i) & \text{if } M_i = 0 \\ \omega_M & \text{otherwise} \end{cases}, \tag{1}$$

where $\omega_i \in \mathbb{R}^D$, and $\omega_M \in \mathbb{R}^D$ is a pixel-independent learned parameter for all pixels that do not correspond to the surface ($D = 512$ for all experiments). Given ω_i, a learned affine operation transforms ω_i to layer-wise "styles" γ_i^ℓ (we use the word "style" following prior work [19, 27]) to scale the feature map x^ℓ;

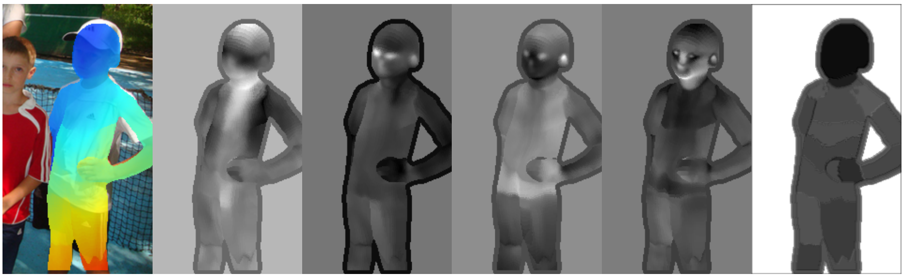
$$\text{SAM}(x_i^\ell, \gamma_i^\ell) = \gamma_i^\ell \cdot x_i^\ell, \tag{2}$$

where each pixel i is modulated by γ_i^ℓ independently. Note that we follow StyleGAN2 design [28], with modulation before convolution and normalization after.

The global mapping network (f_ω) adapts the smooth surface embedding into semantically meaningful surface-adaptive styles, which are not necessarily smooth. For instance, this enables the generator to learn part-wise continuous styles with clearly defined semantic borders (*e.g.* between two pieces of clothing). We observe that a deeper mapping network learns higher-fidelity styles (Figure 3), which improves image quality (shown in Section 4.1).

Unlike prior semantic-based modulation [43, 52, 53], SAM uses a denser and more informative representation that excels at human synthesis. Semantic-based modulation learns spatially-invariant (but semantic-variant) styles [53], which is reflected in Figure 3. These spatially-invariant parameters are efficient for natural image synthesis but translate poorly to the highly fine-grained task of human figure synthesis. In contrast, SAM learns semantically detailed styles independent of pre-defined semantic regions.

embeddings regressed from the original image. This can mitigate identity leaking through CSE-embeddings.



(a) $n = 0$ **(b)** $n = 2$ **(c)** $n = 4$ **(d)** $n = 6$ **(e)** SPADE

Figure 3: Visualization of the norm of γ for SAM where f_ω has n layers (a-d), and (e) show SPADE [43] with 26 semantic regions. Note that SAM learns much more fine-grained details (*e.g.* zoom in on head or fingers) than its semantic counterpart [43].

3.2 Variational Surface Adaptive Modulation

A key limitation to SAM is that the appearance of the synthesized body depends on its spatial position. Typically, an image-to-image generator inputs a latent code (z) directly to a 2D feature map through concatenation or additive noise. However, this entangles the latent code with the spatial feature map, making the appearance of the generated person dependent on the position in the image.

Instead of inputting z to a 2D feature map, we extend SAM to condition the mapping network on z; $\omega_i = f_\omega(e_i, z)$. Now, f_ω transforms the latent variable z to a surface-adaptive intermediate latent space (ω), which is modulated onto the spatial feature map. This naive extension of SAM allows the generator to directly relate latent factors of variations (*e.g.* color of the shirt) to specific positions on the body. Note that the variational modulation of V-SAM is independent of the spatial position of the body in the image, as γ_i^ℓ is determined solely from (z, e_i). This enables V-SAM to modulate the style of the body invariant to image rotation and translation, improving the ability of the generator to synthesize the same person independent of its spatial position [3].

[3] Note that rotational invariance is not retained in the generator, as the generator is not rotationally invariant itself. However, adapting V-SAM with StyleGAN3-R [26] produces a surface-guided rotationally invariant generator.

Adaptive modulation is an established technique in the literature for unconditional [19, 27] and conditional modulation [43, 62]. However, the design of V-SAM is more expressive than current methods, and the explicit adaption of latent variables to surface locations independent of spatial position is unique to V-SAM. The naive design of V-SAM originates from the plain representation of CSE, where equally expressive modulation techniques based on other representations (*e.g*. DensePose or semantic maps) require much more engineering effort. For example, current variational semantic-based modulation [52, 66] does not directly translate to human synthesis [4] and the styles generated by V-SAM are of higher fidelity. Furthermore, the expressiveness of V-SAM significantly improves quality and disentanglement compared to previous methods, which we experimentally validate in section 4.2.

3.3 Discriminator Surface Supervision

Supervising the discriminator by teaching it to predict conditional information (instead of inputting it), is known to improve image quality and training stability [42, 48]. We propose a similar objective for surface embeddings.

We formulate the surface embedding prediction as a regression task. We extend the discriminator with an FPN-head that outputs a continuous embedding for each pixel; $\hat{e}_i$. Along with the adversarial objective, the discriminator optimizes a masked version of the smooth L_1 loss [11];

$$\mathscr{L}_{CSE}(\hat{e}, e) = \sum_{i \in h,w} (1 - M_i) \odot \text{smooth}_{L_1}(\hat{e}_i, e_i). \tag{3}$$

Similarly, the generator objective is extended with the regression loss with respect to the generated image. Unlike the original CSE loss [41], our objective is simpler as we assume a fixed embedding e which is learned in advance.

Discriminator surface supervision explicitly encourages the discriminator to learn pixel-to-surface correspondences. This yields a discriminator that provides highly detailed gradient signals to the generator, which considerably improves image quality. In comparison to semantic-based supervision [48],

[4] *E.g*. adapting [52, 66] for body parts requires class-specific latent variables that have to semantically match between related regions

surface-supervision provides higher-fidelity feedback without relying on predefined semantic regions. Finally, we found that additionally predicting "real" and "fake" areas (as in OASIS [48]) negatively affects training stability and that a FPN-head is more stable to train compared to a U-Net [44] architecture (as used in [48]).

3.4 The Anonymization Pipeline

Our proposed anonymization framework consists of two stages. Initially, a CSE-based [41] detector computes the location of humans, including a dense 2D-3D correspondence between the 2D image and a fixed 3D human surface. Given the detected human, we zero-out pixels covering the human body and complete the partial image with a generative model. Note that the masks generated from CSE [41] do not cover areas that are "outside" of the human body, thus we dilate the mask to ensure that it covers clothing and hair. We extend eq. (1) with an additional pixel-independent learned parameter for the dilated regions (similar to ω_M), to ensure a smooth transition between known areas and unknown dilated areas (without a surface embedding).

4 Experiments

We validate our design choices in Section 4.1 and compare V-SAM to alternative methods in Section 4.2. Section 4.3 ablate on the DeepFashion [33] dataset for scene-independent human synthesis. Finally Section 4.4 evaluates the impact of anonymization for future computer vision development. Appendix C and D include further evaluation.

Architecture Details We follow the implementation of StyleGAN2 [28] for our training setup. The generator is a U-Net [44], previously adapted for image-to-image translation [22], and the discriminator is similar to the one of StyleGAN2. The generator uses instance normalization for each convolution, operating only on standard deviation (*i.e.* the mean is not used for normalization). The latent variable (z) is linearly projected and concatenated to the input of the decoder of the generator, unless it is inputted through modulation. The

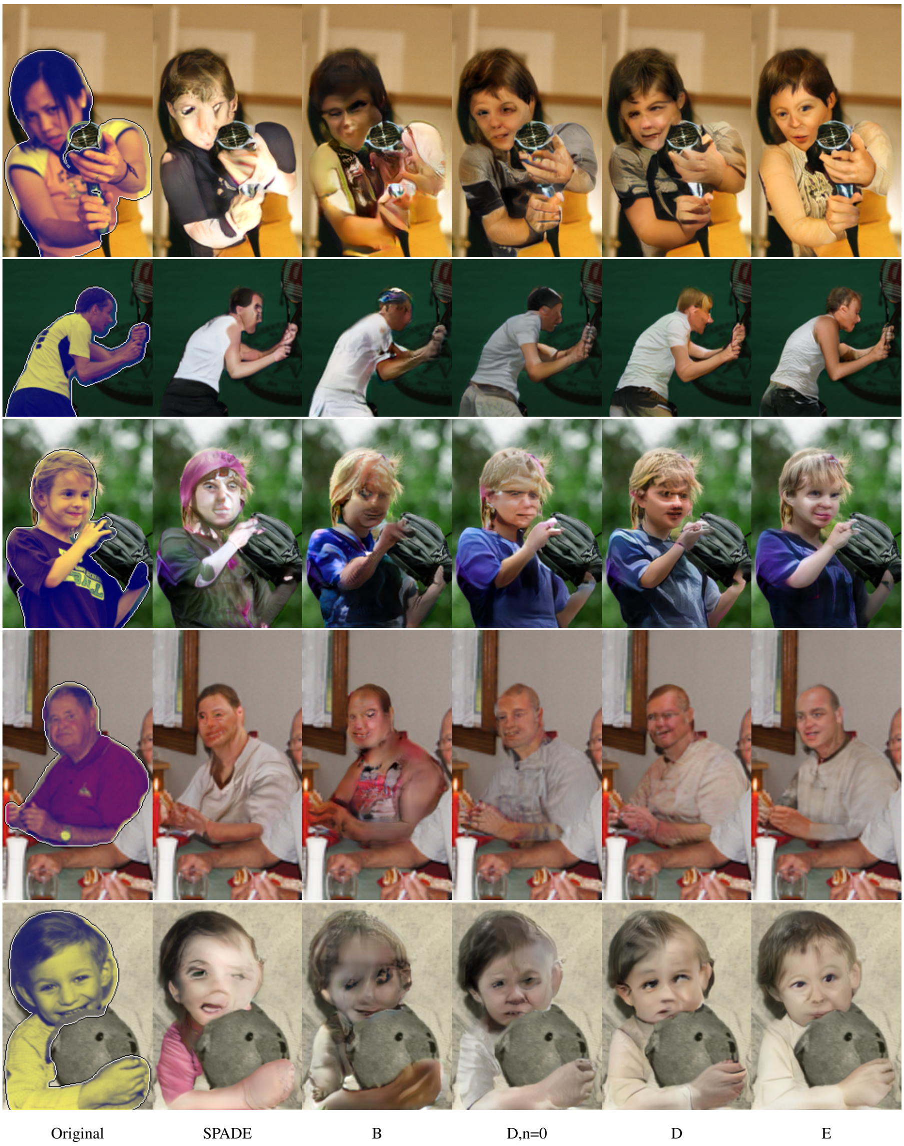


Figure 4: Synthesized images for the different model iterations in Table 1. Appendix D includes random examples.

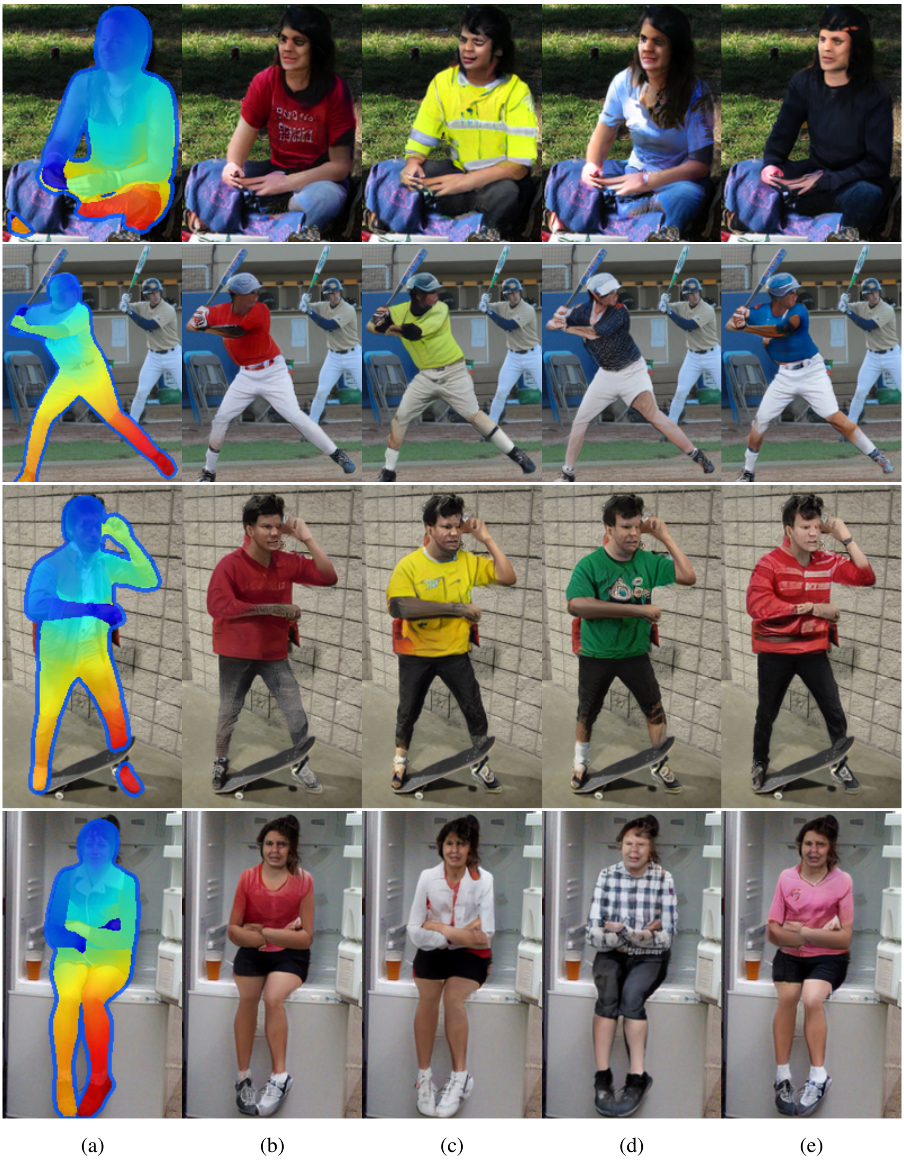


Figure 5: Config E diverse synthesis. (a) is the input, (b) is the generated image with truncation (t=0), and (c-e) are without truncation. Appendix D includes random examples.

baseline discriminator and generator has 8.5M and 7.4M parameters, respectively. We use the non-saturating adversarial loss [12] with epsilon penalty [24] and r1-regularization [37]. We mask the r1-regularization by M, similar to [57, 20]. Data augmentation is used for COCO-Body, including geometrical transforms, and color transforms. Otherwise, we keep the training setup simple, with no feature matching loss [54], or path length regularization [28]. We set the dimensionality of ω_i and the fully-connected layers in f_ω to 512, and use 6 layers in f_ω unless stated otherwise. Appendix A includes further details.

Dataset Details We validate our method on two datasets; a derived version of the COCO-dataset [32] (named *COCO-Body*) for full-body anonymization and DeepFashion [33] for static scene synthesis. We will open-source the CSE-annotations for both datasets.

- *COCO-Body* contains cropped images from COCO [32], where a single human is in the center of each image. Each image has automatically annotated CSE embeddings and a boolean mask indicating the area to be replaced. Note that each mask is dilated from the original CSE-embedding such that the mask covers all parts of the body. The dataset contains 43,053 training images and 10,777 validation images, with a resolution of 288 × 160. See Appendix B for more details.
- *DeepFashion-CSE* includes images from the In-shop Clothes Retrieval Benchmark of DeepFashion [33], where we have annotated each image with a CSE embedding. It has 40,625 training images and 10,275 validation images, where each image is downsampled to 384 × 256. The dataset includes some errors in annotations, as no annotation validation is done.

Evaluation Details We follow typical evaluation practices for generative modeling. We report Fréchet Inception Distance (FID) [18], Learned Perceptual Image Patch Similarity (LPIPS) [61], LPIPS Diversity [65], and Perceptual Path Length (PPL) [27]. FID, LPIPS, and LPIPS diversity is found by generating 6 images per validation sample, where the reported LPIPS is the average. In addition, we report the face quality by evaluating FID for the face region (see Appendix A). Appendix C includes all metrics for each model.

Table 1: Iterative addition of surface guidance to the baseline. * $\mathscr{L}_{CSE}$ is applied to G and D, where G receives CSE-information by concatenation with the image

	Method	LPIPS ↓	FID ↓	PPL ↓	Diversity ↑
A:	Baseline	0.237	7.4	26.7	0.162
B:	A + $\mathscr{L}_{\text{CSE}}$*	0.220	5.8	19.0	0.140
C:	B + SAM	0.219	5.6	19.2	0.143
D:	B + V-SAM	0.220	5.2	**13.7**	**0.166**
E:	D + Larger D/G	**0.211**	**4.8**	15.1	0.161

Table 2: Config D with different number of layers (n) in the mapping network (f_ω). All other experiments use 6 layers.

f_ω depth (n)	Face FID ↓	FID ↓	PPL ↓
0	7.7	5.4	24.9
2	8.0	5.4	19.7
4	7.9	5.5	19.8
6	**7.4**	**5.2**	**13.7**

4.1 Attributes of Surface-Guided GANs

We iteratively develop the baseline architecture to introduce surface guidance. Table 1 (and Figure 4) reflects that the addition of discriminator surface-supervision (config B) and surface modulation (config C/D) drastically improves image quality. Note that adaptive modulation is only applied for the convolutional layers in the decoder. Config E increases the model size of the generator and discriminator to 33M and 34M parameters, respectively. The final generator produces high-quality and diverse results (Figure 5). In addition, the conditional intermediate latent space ω is amenable to similar techniques as the latent space of StyleGAN [27], *e.g.* the truncation trick [5] and latent interpolation (ablated in Appendix C). Figure 5 includes generated images with latent truncation.

Mapping Network Depth A deeper mapping network allows the generator to learn finer-grained modulation parameters, which we find to significantly improve image quality and latent disentanglement (Table 2). Qualitatively, we observe that this significantly improves the quality of fine-grained regions (*e.g.* the face and fingers, see Figure 4). We quantitatively validate this improvement through the FID of the upsampled face region (Face FID).

Furthermore, a deeper mapping network allows the generator to better disentangle the latent space [5], which is reflected by PPL. The improved disentanglement is rooted in two design choices; first, SAM explicitly disentangles the variations of pose into surface-adaptive modulation. Secondly, V-SAM allows the generator to easier control specific areas of the human body disentangled of the spatial image, by "unwarping" the fixed distribution z to the surface-conditioned distribution ω.

Affine Invariance Studies V-SAM is invariant to affine image-plane transformations, and thus, improves the ability of the generator to disentangle the latent representation from such transforms. We quantitatively evaluate this with Peak Signal-to-Noise Ratio (PSNR), following [60],

$$\mathbb{E}_{I,M,E,t\sim T}\text{PSNR}[t(G(\bar{I},E)),G(t(\bar{I}),t(E))], \tag{4}$$

where $\bar{I} = I \odot M$, E is the CSE embedding, G is the generator, and T is the distribution of vertical/horizontal image shifts. T is limited to translate by a maximum $\frac{1}{8}$ of the image width/height. Similarly, we evaluate rotational invariance (limited to $\pm 90^\circ$) and horizontal flip.

V-SAM significantly improves the baseline w.r.t. invariance to affine transformations (table 3), as V-SAM is invariant to such transformations. In comparison, SAM achieves similar scores as the baseline. The aspect of affine-invariance is important for realistic anonymization, as the detection can induce slight shifts across frames.

Computational Complexity V-SAM consists of two stages, the mapping network and layer-wise linear transformations. Each layer-wise transformation is efficiently implemented as 1×1 convolution. The mapping network is a sequence of fully-connected layers, which can be implemented as 1×1

[5] Following Karras *et al.* [27], "disentangled latent space" refers to that the latent factors of variations are separated into linear subspaces.

Table 3: Comparison of V-SAM to alternative adaptive normalization methods. All methods are applied on top of config B.

	Affine Transformation			Quality			
Method	Translation ↑	Rotation ↑	Hflip ↑	Diversity ↑	FID ↓	PPL ↓	Face FID ↓
Config B	23.1	20.3	21.7	0.140	5.8	19.0	9.1
B + SPADE [43]	22.5	19.8	20.7	0.150	5.9	20.6	9.7
B + INADE [52]	24.1	20.2	20.9	0.140	5.8	19.5	9.4
B + CLADE [53]	22.9	20.1	21.3	0.138	5.7	16.9	8.9
B + StyleGAN [28]	25.5	20.9	21.6	0.155	5.7	48.2	9.4
B + CoMod [62]	24.5	20.6	21.6	0.154	5.5	17.5	8.0
B + SAM	23.8	20.7	21.4	0.143	5.6	19.2	**7.4**
B + V-SAM	**26.1**	**21.4**	**22.5**	**0.166**	**5.2**	**13.7**	**7.4**

convolution by using the spatial embedding map e_i for each pixel i. However, in practice, we find the nearest vertex embedding e_k for each embedding e_i, and transform the 27K vertex-embeddings to w_k. This results in a mapping network independent of image resolution.

4.2 The Expressiveness of V-SAM

We now analyze the expressiveness of V-SAM compared to well-established modulation techniques. Specifically, we compare against adaptive instance normalization from StyleGAN2 [28], co-modulation (CoMod) [62], and variational [52]/non-variational [43, 53] semantic-based methods. All methods are applied on top of Config B.

table 3 shows that V-SAM significantly improves upon previous modulation methods. V-SAM generates higher-fidelity styles than both semantic-based modulation [28, 62] and spatially invariant modulation [43, 52, 53], yielding a substantial improvement in image quality (FID). This is especially prominent in semantically complex areas of the body (Face FID). Note that the improvement of V-SAM over co-modulation [62] is significant, as it is approximately the same as increasing the number of parameters by 20M (config E vs D, table 1). Furthermore, V-SAM improves latent disentanglement (PPL), originating from the expressive and explicit design of V-SAM. Finally, V-SAM is more invariant to affine image-plane transformations.

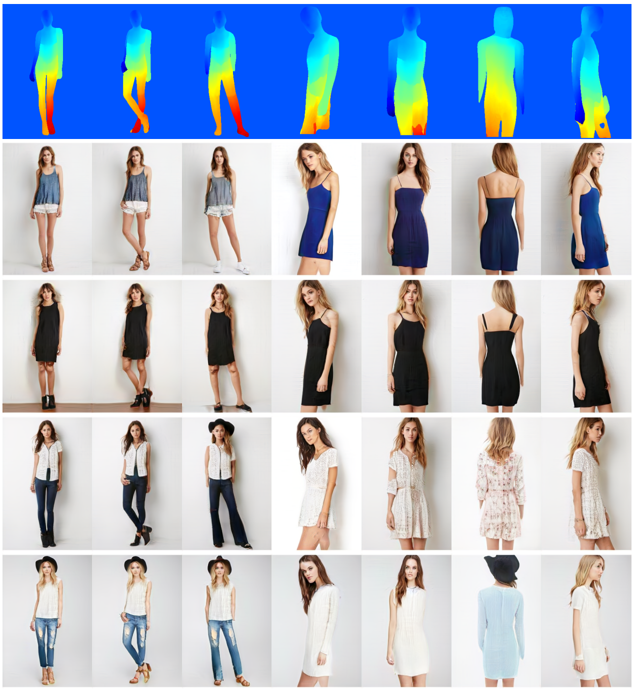
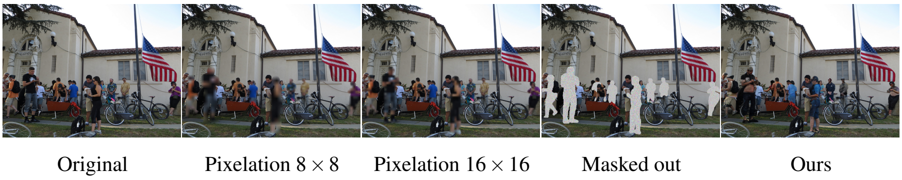


Figure 6: V-SAM can transfer attributes between poses by simply sampling the same latent variable z. Each row shows synthesized images with the same latent variable, but different input pose.

4.3 Synthesis of Humans in Static Scenes

We demonstrate that V-SAM excels at human synthesis for the DeepFashion [33] dataset. Following the design of SPADE [43], we design a decoder-only generator that synthesizes humans independent of any background image.

The disentangled and spatially-invariant latent space of V-SAM allows the generator to transfer attributes between poses. By sampling the same latent variable z for different poses, V-SAM is able to perform pose/motion transfer of synthesized humans (Figure 6) without any task-specific modeling choices (*e.g.* including a texture encoder [59]). However, V-SAM is variant to 3D affine transformations that are not parallel to the imaging plane (*e.g.* changing the depth of the scene). This is reflected in Figure 6, where changing the depth



Original Pixelation 8 × 8 Pixelation 16 × 16 Masked out Ours

Figure 7: Different anonymization methods for an image from COCO [32] val2017. Appendix D includes random examples.

Table 4: Instance segmentation mask AP on the COCO validation set [32]. The results are from a pre-trained Mask R-CNN [17] R50-FPN-3x from detectron2 [56] evaluated on different anonymized datasets.

Validation Dataset	$AP_{50:95} \uparrow$	$AP_{50} \uparrow$	$AP_{75} \uparrow$	$AP_{s} \uparrow$	$AP_{m} \uparrow$	$AP_{l} \uparrow$	$AP_{Person} \uparrow$
Original	37.2	58.6	39.9	18.6	39.5	53.3	47.7
Mask Out	32.8	52.0	35.1	16.3	34.6	47.3	27.5
8 × 8 Pixelation	32.8	51.8	35.2	16.4	34.6	47.2	33.3
16 × 16 Pixelation	33.4	53.0	35.7	16.7	35.0	48.1	38.4
Ours	**34.6**	**55.0**	**37.0**	**17.1**	**36.8**	**50.0**	**44.9**

of the scene significantly changes the synthesized person. We believe that combining V-SAM with task-specific modeling choices from the pose/motion transfer literature [36, 59] could resolve these issues.

4.4 Effect of Anonymization for Computer Vision

Data Usability We analyze the effect of anonymization for future computer vision development by evaluating a pre-trained Mask R-CNN [17] on the COCO dataset (results on PASCAL VOC [9] are included in Appendix B). We anonymize all individuals that are detected by a pre-trained CSE-detector [41], where we use all detections with a confidence score higher than 0.1. We compare our framework to traditional anonymization methods (Figure 7).

Our method significantly improves AP_{person} compared to traditional anonymization (Table 4), even pixelation, which is known to be questionable for anonymization [14, 39]. However, we observe a notable drop in average precision for other object classes, which originates from two sources of error. First, full-body anonymization removes objects that often appear with human figures.

Table 5: Re-identification mAP and rank-1 accuracy on Market1501 [63] using the official code of OSNet [64].

Anonymization	R1 ↓	mAP ↓
Original	94.4	82.5
Pixelation 8 × 8	67.8	39.6
Pixelation 16 × 16	86.6	66.4
Mask-out	28.2	10.4
Face Anonymization [21]	82.1	50.7
Ours	**31.1**	**14.4**

For example, the "tie" class drops from 31% AP to 1% and "toothbrush" drops from 14.6% to 6.2%. Secondly, the detections include false positives, yielding highly corrupted images when anonymizing these. For example, the "zebra" class drops from 56.2% to 48.0%. We observe insignificant degradation for objects that are rarely detected as person (*e.g.* car, train, elephant). Finally, surface-guided anonymization improves over traditional techniques for training purposes, which we validated on the anonymized COCO dataset (Appendix B).

Anonymization Quality Table 5 evaluates the effect of anonymization for person re-identificaiton on the Market1501 [63] dataset. Surface-guided GANs provide similar anonymization guarantees as masking out the region. Meanwhile, face anonymization and pixelation yields a much higher re-identification rate, reflecting its worse anonymization guarantee.

5 Conclusion

We present a novel full-body anonymization framework that generates close-to-photorealistic and diverse humans in varying and complex scenes. Our experiments show that guiding adversarial nets with dense pixel-to-surface correspondences strongly improves synthesis of high-fidelity textures for varying poses and scenes. Finally, we demonstrate that our anonymization framework better retains data usability for future computer vision development compared to traditional anonymization.

Limitations Our contributions significantly improve the usability of anonymized data and generate new identities independent of the original. However, our method has limitations that can compromise the privacy of individuals. As with any anonymization method, our method relies on detection that is far from perfect [6] and vulnerable to adversarial attacks. Detection is improving every year and defense against adversarial attacks is currently a large focus in the community [29]. We believe that potential errors in detection can be circumvented with face detection as a fallback.

With the assumption of perfect detections, identification is still possible through gait recognition (when anonymizing videos), or through identity leaks in the CSE-embeddings. We speculate that gait recognition can be mitigated by slightly randomizing the original pose between frames. Furthermore, identity leaking through surface embeddings is possible, as they are regressed from the original image and could include identifying information. We reduce this possibility by discretizing the regressed embedding into one of the 27K vertex-specific embeddings (Section 3).

Surface-guided GANs significantly improve human figure synthesis for in-the-wild image anonymization. Nevertheless, human synthesis is a complicated task, and many of the images generated by our method are recognizable as artificial by a human evaluator. One of the limiting factors of our model is the dataset, where COCO-Body contains 40K images with a large variety. This is relatively small compared to the 70K images in FFHQ [27], which is a considerably simpler task. Our method applies data augmentation to mitigate this. However, further extension with adaptive augmentation [25] or transfer learning could be fruitful.

Societal Impact We live in the age of Big Data, where personal information is the business model for many companies. Recently introduced legislation has complicated data collection, requiring consent to store any data that contains personal information. This can be a barrier to research and development, especially for the data-dependent field of computer vision. We present a method that can better preserve the privacy of individuals, while retaining the

[6] Current CSE-based detectors (R-101-FPN-DL-s1x [56]) has an average recall rate of 96.65% (AR50) for human segmentation on COCO-DensePose [15]. Note that COCO-DensePose contains primarily high-resolution human figures.

usability of the data. Nevertheless, our work focus on the synthesis of realistic humans, which has a potential for misuse. The typical example is misuse of DeepFakes, where generative models can be used to create manipulated content with an intention to misinform. Several solutions have been proposed, where the DeepFake Detection Challenge [8] has increased the ability of models to detect manipulated content, and pre-emptive solutions such as model watermarking [58] can mitigate the potential for misuse.

Acknowledgement The computations were performed on resources provided by the NTNU IDUN/EPIC computing cluster [49] and the Tensor-GPU project led by Prof. Anne C. Elster through support from The Department of Computer Science and The Faculty of Information Technology and Electrical Engineering, NTNU. Furthermore, Rudolf Mester acknowledges the support obtained from DNV GL.

References

[1] Badour AlBahar, Jingwan Lu, Jimei Yang, Zhixin Shu, Eli Shechtman, and Jia-Bin Huang. Pose with style: Detail-preserving pose-guided image synthesis with conditional stylegan. *arXiv preprint arXiv:2109.06166*, 2021.

[2] Thangapavithraa Balaji, Patrick Blies, Georg Göri, Raphael Mitsch, Marcel Wasserer, and Torsten Schön. Temporally coherent video anonymization through gan inpainting. *arXiv preprint arXiv:2106.02328*, 2021.

[3] Guha Balakrishnan, Amy Zhao, Adrian V Dalca, Fredo Durand, and John Guttag. Synthesizing images of humans in unseen poses. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 8340–8348, 2018.

[4] Karla Brkic, Ivan Sikiric, Tomislav Hrkac, and Zoran Kalafatic. I know that person: Generative full body and face de-identification of people in images. In *2017 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 1319–1328. IEEE, 2017.

[5] Andrew Brock, Jeff Donahue, and Karen Simonyan. Large Scale GAN Training for High Fidelity Natural Image Synthesis. In *International Conference on Learning Representations*, 2019.

[6] Caroline Chan, Shiry Ginosar, Tinghui Zhou, and Alexei A Efros. Everybody dance now. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 5933–5942, 2019.

[7] Bindita Chaudhuri, Nikolaos Sarafianos, Linda Shapiro, and Tony Tung. Semi-supervised synthesis of high-resolution editable textures for 3d humans. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7991–8000, 2021.

[8] Brian Dolhansky, Joanna Bitton, Ben Pflaum, Jikuo Lu, Russ Howes, Menglin Wang, and Cristian Canton Ferrer. The deepfake detection challenge dataset. *arXiv e-prints*, pages arXiv–2006, 2020.

[9] M. Everingham, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman. The pascal visual object classes (voc) challenge. *International Journal of Computer Vision*, 88(2):303–338, June 2010.

[10] Oran Gafni, Lior Wolf, and Yaniv Taigman. Live face de-identification in video. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9378–9387, 2019.

[11] Ross Girshick. Fast r-CNN. In *2015 IEEE International Conference on Computer Vision (ICCV)*. IEEE, dec 2015.

[12] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In *Advances in neural information processing systems*, pages 2672–2680, 2014.

[13] Artur Grigorev, Artem Sevastopolsky, Alexander Vakhitov, and Victor Lempitsky. Coordinate-based texture inpainting for pose-guided human image generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12135–12144, 2019.

[14] R. Gross, L. Sweeney, F. de la Torre, and S. Baker. Model-based face de-identification. In *Conference on Computer Vision and Pattern Recognition Workshop*. IEEE, 2006.

[15] Riza Alp Guler, Natalia Neverova, and Iasonas Kokkinos. DensePose: Dense human pose estimation in the wild. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*. IEEE, jun 2018.

[16] Xintong Han, Zuxuan Wu, Zhe Wu, Ruichi Yu, and Larry S Davis. Viton: An image-based virtual try-on network. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7543–7552, 2018.

[17] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask r-cnn. In *Proceedings of the IEEE international conference on computer vision*, pages 2961–2969, 2017.

[18] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. In *Advances in Neural Information Processing Systems*, pages 6626–6637, 2017.

[19] Xun Huang and Serge Belongie. Arbitrary style transfer in real-time with adaptive instance normalization. In *2017 IEEE International Conference on Computer Vision (ICCV)*. IEEE, oct 2017.

[20] Håkon Hukkelås, Frank Lindseth, and Rudolf Mester. Image inpainting with learnable feature imputation. *arXiv preprint arXiv:2011.01077*, 2020.

[21] Håkon Hukkelås, Rudolf Mester, and Frank Lindseth. Deepprivacy: A generative adversarial network for face anonymization. In *Advances in Visual Computing*, pages 565–578. Springer International Publishing, 2019.

[22] Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A. Efros. Image-to-image translation with conditional adversarial networks. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, jul 2017.

[23] Anil Jain, Patrick Flynn, and Arun Ross. *Handbook of Biometrics*. Springer New York, NY, 01 2008.

[24] Tero Karras, Timo Aila, Samuli Laine, and Jaakko Lehtinen. Progressive growing of GANs for improved quality, stability, and variation. In *International Conference on Learning Representations*, 2018.

[25] Tero Karras, Miika Aittala, Janne Hellsten, Samuli Laine, Jaakko Lehtinen, and Timo Aila. Training generative adversarial networks with limited data. In *Proc. NeurIPS*, 2020.

[26] Tero Karras, Miika Aittala, Samuli Laine, Erik Härkönen, Janne Hellsten, Jaakko Lehtinen, and Timo Aila. Alias-free generative adversarial networks. *arXiv preprint arXiv:2106.12423*, 2021.

[27] Tero Karras, Samuli Laine, and Timo Aila. A style-based generator architecture for generative adversarial networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4401–4410, 2019.

[28] Tero Karras, Samuli Laine, Miika Aittala, Janne Hellsten, Jaakko Lehtinen, and Timo Aila. Analyzing and improving the image quality of stylegan. *arXiv preprint arXiv:1912.04958*, 2019.

[29] Alexey Kurakin, Ian Goodfellow, Samy Bengio, Yinpeng Dong, Fangzhou Liao, Ming Liang, Tianyu Pang, Jun Zhu, Xiaolin Hu, Cihang Xie, et al. Adversarial attacks and defences competition. In *The NIPS'17 Competition: Building Intelligent Systems*, pages 195–231. Springer, 2018.

[30] Verica Lazova, Eldar Insafutdinov, and Gerard Pons-Moll. 360-degree textures of people in clothing from a single image. In *2019 International Conference on 3D Vision (3DV)*, pages 643–653. IEEE, 2019.

[31] Yining Li, Chen Huang, and Chen Change Loy. Dense intrinsic appearance flow for human pose transfer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3693–3702, 2019.

[32] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *European conference on computer vision*, pages 740–755. Springer, 2014.

[33] Ziwei Liu, Ping Luo, Shi Qiu, Xiaogang Wang, and Xiaoou Tang. Deepfashion: Powering robust clothes recognition and retrieval with rich annotations. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2016.

[34] Liqian Ma, Qianru Sun, Stamatios Georgoulis, Luc Van Gool, Bernt Schiele, and Mario Fritz. Disentangled person image generation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 99–108, 2018.

[35] Maxim Maximov, Ismail Elezi, and Laura Leal-Taixé. Ciagan: Conditional identity anonymization generative adversarial networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5447–5456, 2020.

[36] Yifang Men, Yiming Mao, Yuning Jiang, Wei-Ying Ma, and Zhouhui Lian. Controllable person image synthesis with attribute-decomposed gan. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5084–5093, 2020.

[37] Lars Mescheder, Sebastian Nowozin, and Andreas Geiger. Which training methods for gans do actually converge? In *International Conference on Machine Learning (ICML)*, 2018.

[38] Ryota Natsume, Shunsuke Saito, Zeng Huang, Weikai Chen, Chongyang Ma, Hao Li, and Shigeo Morishima. Siclope: Silhouette-based clothed people. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4480–4490, 2019.

[39] Carman Neustaedter, Saul Greenberg, and Michael Boyle. Blur filtration fails to preserve privacy for home-based video conferencing. *ACM Transactions on Computer-Human Interaction*, 13(1):1–36, mar 2006.

[40] Natalia Neverova, Riza Alp Guler, and Iasonas Kokkinos. Dense pose transfer. In *Proceedings of the European conference on computer vision (ECCV)*, pages 123–138, 2018.

[41] Natalia Neverova, David Novotny, Marc Szafraniec, Vasil Khalidov, Patrick Labatut, and Andrea Vedaldi. Continuous surface embeddings. *Advances in Neural Information Processing Systems*, 33, 2020.

[42] Augustus Odena, Christopher Olah, and Jonathon Shlens. Conditional image synthesis with auxiliary classifier gans. In *Proceedings of the 34th International Conference on Machine Learning-Volume 70*, pages 2642–2651. JMLR. org, 2017.

[43] Taesung Park, Ming-Yu Liu, Ting-Chun Wang, and Jun-Yan Zhu. Semantic image synthesis with spatially-adaptive normalization. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, jun 2019.

[44] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical image computing and computer-assisted intervention*, pages 234–241. Springer, 2015.

[45] Shunsuke Saito, Zeng Huang, Ryota Natsume, Shigeo Morishima, Angjoo Kanazawa, and Hao Li. Pifu: Pixel-aligned implicit function for high-resolution clothed human digitization. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2304–2314, 2019.

[46] Kripasindhu Sarkar, Vladislav Golyanik, Lingjie Liu, and Christian Theobalt. Style and pose control for image synthesis of humans from a single monocular view. *arXiv preprint arXiv:2102.11263*, 2021.

[47] Kripasindhu Sarkar, Dushyant Mehta, Weipeng Xu, Vladislav Golyanik, and Christian Theobalt. Neural re-rendering of humans from a single image. In *European Conference on Computer Vision*, pages 596–613. Springer, 2020.

[48] Edgar Schönfeld, Vadim Sushko, Dan Zhang, Juergen Gall, Bernt Schiele, and Anna Khoreva. You only need adversarial supervision for semantic image synthesis. In *International Conference on Learning Representations*, 2020.

[49] Magnus Själander, Magnus Jahre, Gunnar Tufte, and Nico Reissmann. EPIC: An energy-efficient, high-performance GPGPU computing research infrastructure, 2019.

[50] Qianru Sun, Liqian Ma, Seong Joon Oh, Luc Van Gool, Bernt Schiele, and Mario Fritz. Natural and effective obfuscation by head inpainting. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5050–5059, 2018.

[51] Qianru Sun, Ayush Tewari, Weipeng Xu, Mario Fritz, Christian Theobalt, and Bernt Schiele. A hybrid model for identity obfuscation by face replacement. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 553–569, 2018.

[52] Zhentao Tan, Menglei Chai, Dongdong Chen, Jing Liao, Qi Chu, Bin Liu, Gang Hua, and Nenghai Yu. Diverse semantic image synthesis via probability distribution modeling. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7962–7971, 2021.

[53] Zhentao Tan, Dongdong Chen, Qi Chu, Menglei Chai, Jing Liao, Mingming He, Lu Yuan, Gang Hua, and Nenghai Yu. Efficient semantic image synthesis via class-adaptive normalization. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pages 1–1, 2021.

[54] Ting-Chun Wang, Ming-Yu Liu, Jun-Yan Zhu, Andrew Tao, Jan Kautz, and Bryan Catanzaro. High-resolution image synthesis and semantic manipulation with conditional GANs. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*. IEEE, jun 2018.

[55] Chung-Yi Weng, Brian Curless, and Ira Kemelmacher-Shlizerman. Vid2actor: Free-viewpoint animatable person synthesis from video in the wild. *arXiv preprint arXiv:2012.12884*, 2020.

[56] Yuxin Wu, Alexander Kirillov, Francisco Massa, Wan-Yen Lo, and Ross Girshick. Detectron2. `https://github.com/facebookresearch/detectron2`, 2019.

[57] Jiahui Yu, Zhe Lin, Jimei Yang, Xiaohui Shen, Xin Lu, and Thomas S. Huang. Generative image inpainting with contextual attention. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*. IEEE, jun 2018.

[58] Ning Yu, Vladislav Skripniuk, Sahar Abdelnabi, and Mario Fritz. Artificial fingerprinting for generative models: Rooting deepfake attribution in training data. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 14448–14457, 2021.

[59] Jinsong Zhang, Kun Li, Yu-Kun Lai, and Jingyu Yang. Pise: Person image synthesis and editing with decoupled gan. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7982–7990, 2021.

[60] Richard Zhang. Making convolutional networks shift-invariant again. In *International conference on machine learning*, pages 7324–7334. PMLR, 2019.

[61] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *CVPR*, 2018.

[62] Shengyu Zhao, Jonathan Cui, Yilun Sheng, Yue Dong, Xiao Liang, Eric I-Chao Chang, and Yan Xu. Large scale image completion via co-modulated generative adversarial networks. In *International Conference on Learning Representations*, 2021.

[63] Liang Zheng, Liyue Shen, Lu Tian, Shengjin Wang, Jingdong Wang, and Qi Tian. Scalable Person Re-identification: A Benchmark. In *2015 IEEE International Conference on Computer Vision (ICCV)*, pages 1116–1124. IEEE, dec 2015.

[64] Kaiyang Zhou, Yongxin Yang, Andrea Cavallaro, and Tao Xiang. Omni-Scale Feature Learning for Person Re-Identification. In *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 3701–3711. IEEE, oct 2019.

[65] Jun-Yan Zhu, Richard Zhang, Deepak Pathak, Trevor Darrell, Alexei A Efros, Oliver Wang, and Eli Shechtman. Toward multimodal image-to-image translation. In *Advances in neural information processing systems*, pages 465–476, 2017.

[66] Peihao Zhu, Rameen Abdal, Yipeng Qin, and Peter Wonka. Sean: Image synthesis with semantic region-adaptive normalization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5104–5113, 2020.

Paper D

DeepPrivacy2: Towards Realistic Full-Body Anonymization

Authors:

Håkon Hukkelås, Frank Lindseth

Published at conference:

2023 IEEE/CVF Winter Conference on Applications of Computer Vision

Copyright:

Source Code:

`https://github.com/hukkelas/deep_privacy2`

Appendix:

`https://openaccess.thecvf.com/content/WACV2023/supplemental/Hukkelas_DeepPrivacy2_Towards_Realistic_WACV_2023_supplemental.pdf`

Paper D

DeepPrivacy2: Towards Realistic Full-Body Anonymization

Håkon Hukkelås Frank Lindseth

Department of Computer Science
Norwegian University of Science and Technology
Trondheim, Norway
hakon.hukkelas@ntnu.no

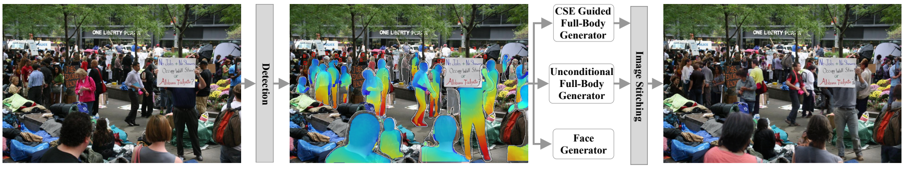

Detection
CSE Guided Full-Body Generator
Unconditional Full-Body Generator
Face Generator
Image Stitching

Figure 1: DeepPrivacy2 detects and anonymizes individuals via three detection and synthesis networks; (1) a CSE-guided generator for individuals detected with dense pose (by CSE [30]), (2) an unconditional full-body generator for cases where CSE fails to detect (note the segmented persons without color-coded CSE detections), and (3) a face generator for the remaining individuals (marked in red). The original image is from Wider-Face [45].

Abstract

Generative Adversarial Networks (GANs) are widely adopted for anonymization of human figures. However, current state-of-the-art limits its anonymization to the task of face anonymization. In this paper, we propose a novel anonymization framework (DeepPrivacy2) for realistic anonymization of human figures and faces. We introduce a new large and diverse dataset for full-body synthesis, which significantly improves image quality and diversity of generated images. Furthermore, we propose a style-based GAN that produces high-quality, diverse, and editable anonymizations. We demonstrate that our full-body anonymization framework provides stronger privacy guarantees than previously proposed methods. Source code and appendix is available at: github.com/hukkelas/deep_privacy2.

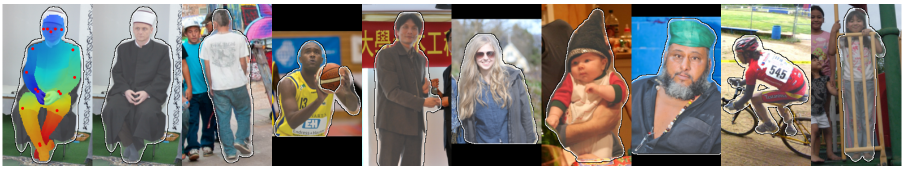


Figure 2: Examples from the FDH dataset. Each image is annotated with keypoints, pixel-to-vertex correspondences (from CSE [30]) and a segmentation mask. The leftmost image shows annotations for the first image.

1 Introduction

Collecting and storing images is ubiquitous in our modern society, where a range of applications requires collecting privacy-sensitive data. However, collecting such data without anonymization or consent from the individual is troublesome due to recently introduced legislation in many areas (*e.g*. GDPR in EU). Traditional image anonymization (*e.g*. blurring) is widely adopted in practice; however, it severely distorts the data, making it unusable for future applications. Recently, *realistic anonymization* has been introduced as an alternative to traditional methods, where generative models can generate realistic faces fitting into a given context [7, 13, 26, 40]. However, current methods focus on face anonymization, which does not prevent recognition through identifiers outside the face, including both primary (*e.g*. ears, gait [15]) and secondary (*e.g*. gender) identifiers.

Surface Guided GANs (SG-GAN) [14] propose a full-body anonymization GAN guided on dense pixel-to-surface correspondences from Continuous Surface Embeddings (CSE) [30]. SG-GAN shows promising results for full-body anonymization, but their method often includes visual artifacts, degrading the image quality. The authors attribute the limited visual quality to the dataset, where they use a derivate of COCO [23] containing 40K human figures. Furthermore, the CSE segmentation used for anonymization does not include accessories/hair on the human body; thus, the anonymized individual often "wears" these unsegmented areas (see fig. 3). Additionally, SG-GAN fails to anonymize many individuals, as the CSE detector often fails to detect persons that are further away from the camera.

In this work, we extend Surface Guided GANs to address the limited visual quality and the insufficient anonymization due to poor segmentation. Furthermore, we address cases where the CSE detector fails to detect individuals. We summarize our contributions in the following.

First, we introduce the Flickr Diverse Humans (FDH) dataset. The FDH dataset consists of 1.5M images of human figures in diverse contexts extracted from the YFCC100M [42] dataset. We show that the larger dataset greatly benefits the visual quality of generated human figures.

Secondly, we propose a novel anonymization framework that combines detections from multiple modalities to improve the segmentation and detection of human figures. Our anonymization framework divides image anonymization into three individual anonymizers; (1) for human figures that are detected with a dense pose estimation (CSE), (2) for human figures that CSE does not detect, and (3) for the remaining faces (see fig. 1). For each category, our framework employs a simple inpainting GAN that follows established GAN training techniques for unconditional image generation [17, 18]. We show that our GAN generates high-quality and diverse identities with few task-specific modeling choices.

Finally, we extend our GAN for face anonymization on an updated version of the Flickr Diverse Faces (FDF) dataset [13]. In contrast to previous face anonymization techniques [7, 13, 26, 40], our GAN uses no pose guidance, enabling it to anonymize individuals where pose information is challenging to detect. Furthermore, we show that our style-based generator can adapt methods from unconditional GANs to find global semantically meaningful directions in the GAN latent space. This enables text-guided attribute editing for our anonymization pipeline.

DeepPrivacy2 surpasses all previous state-of-the-art realistic anonymization methods in terms of image quality and anonymization guarantees. We validate the effectiveness of DeepPrivacy2 with extensive qualitative and quantitative evaluation. Our code, pre-trained models, and the FDH dataset is available at github.com/hukkelas/deep_privacy2.

2 Related Work

Image Anonymization Naive image anonymization (*e.g*. masking, blurring, pixelation) is widely adopted in practice; however these methods severely degrade the quality of the anonymized image, making the data unusable for many applications. Early work focused on the K-same family of algorithms [8, 16, 31], which provides better privacy guarantees and data usability than naive methods , but generate highly corrupted images. Recent work on deep generative models reflects that learning-based anonymization can realistically anonymize data while retaining its usability for downstream applications. These methods anonymize face regions by either inpainting missing regions [13, 26, 40, 41] or transforming [7, 35] the original face. Our method anonymizes by inpainting, as inpainting-based methods provide stronger privacy guarantees than transformative methods, as they never observe the original privacy-sensitive information. The majority of prior work focuses on face anonymization, which compromises privacy for many use cases, as they leave several primary (*e.g*. ears, gait) and secondary (*e.g*. gender) identifiers on the human body untouched. There is a limited amount of work focusing on full-body anonymization [2, 14, 26], where prior methods are limited to low-resolution images [26] or generate images with visual artifacts [2, 14].

Full-body Synthesis Recent work on full-body synthesis focus on limited tasks, such as transferring source appearances into new poses [1, 3, 22, 29, 38], with different garments [9, 37, 38], or with new motion [3]. These methods are often guided on dense pixel-to-surface correspondences or sparse keypoints annotations. In contrast to these methods, our anonymization approach does not rely on a source appearance to transfer, and the majority of the aforementioned methods do not handle large variations in background contexts. Furthermore, a number of these methods focus on low-variance datasets (*e.g*. DeepFashion [24]), which consists of a limited number of identities in similar poses and a close-to static context (white background). There is a limited amount of work focusing on full-body synthesis without a source appearance, where Ma *et al.* [25] proposes a pose-guided GAN for novel full-body synthesis.

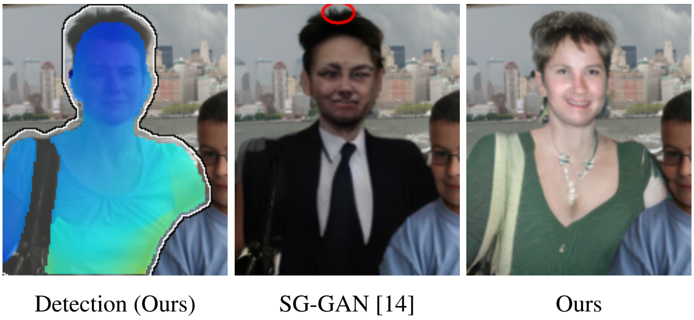


Figure 3: SG-GAN [14] anonymizes only the area from a CSE segmentation (marked in blue tint), which does not include accessories/hair. This results in SG-GAN [14] anonymization often wearing the original hair/accessories of the original identity (marked in red). In contrast, DeepPrivacy2 anonymizes the segmentation from Mask R-CNN (outlined), which includes hair and clothing.

3 The Flickr Diverse Humans Dataset

The Flickr Diverse Humans (FDH) dataset consists of 1.53M images of human figures from the YFCC100M [42] dataset. Figure 2 shows examples from the dataset. Each image contains a single human figure in the center, with a pixel-wise dense pose estimation from CSE [30], 17 keypoint annotations from a keypoint R-CNN model [10], and a segmentation mask. The segmentation mask is the union of the mask from a CSE detector and Mask R-CNN [10] trained on COCO. The dataset is automatically filtered through confidence thresholding, automatic image quality assessment, the number of body parts visible in the image, and overlap between keypoint and CSE predictions (see Appendix A for more details). Otherwise, we perform no further filtering such that the dataset includes individuals in all various contexts. The resolution of each image is 288 × 160, and the dataset is split into 1,524,845 images for training and 30K images for validation. Compared to previously adopted datasets for full-body synthesis [14, 24, 47], FDH is much larger and contains a diverse set of individuals from in-the-wild images. Additionally, FDH is less curated than typical datasets for generative modeling, where it includes human figures with unusual poses, perspectives, lighting conditions, and contexts. This is to ensure that our anonymization method can handle such conditions.

4 The DeepPrivacy2 Anonymization Pipeline

This section outlines the core technologies used for our anonymization pipeline. First, we present our ensemble detection pipeline, then our GAN-based synthesis method.

4.1 Detection

The main objective of the detection module is to ensure that all individuals in the image are detected. DeepPrivacy2 uses an ensemble of three detection networks from different modalities; DSFD [21] for face detection, CSE [30] for dense pose estimation, and Mask R-CNN [10] for instance segmentation. The pipeline categorizes the detections into three categories; individuals with dense pose estimation (detection w/ CSE), individuals not detected by CSE (detection w/o CSE), and faces that are not included in the former categories. For each category, we propose individual anonymization methods, introduced in section 4.2. For human figures, we anonymize the union of Mask R-CNN and CSE segmentations, such that accessories/hair detected by Mask R-CNN (but not CSE) are anonymized (see fig. 3). Note that dense pose estimation is not essential for privacy, but it substantially improves synthesized image quality. Furthermore, the detections are tracked with a bounding box tracker, such that the anonymization can retain the same identity over a sequence of frames. Compared to SG-GAN [14], the ensemble of detectors significantly improves detection recall, as DeepPrivacy2 uses Mask R-CNN and DSFD for fallback detection when the CSE detector fails.

Implementation Details. Instance segmentation and CSE segmentation are combined via simple Intersection over Union (IoU) thresholding, where we assume all detections with an IoU higher than 0.4 are the same individual. All instance segmentations from Mask R-CNN not combined with a CSE detection are categorized as a detection without CSE. All face detections within the CSE or Mask R-CNN segmentation are discarded. All detections are tracked with simple Kalman filtering on bounding boxes, following the implementation of motpy [28]. We use Mask R-CNN and the CSE implementations from detectron2 [43], specifically, the ResNeXt-101 FPN [44] Mask R-CNN, and

ResNet-101 [11] CSE. We adapt DSFD [21] from the official implementation of the authors.

4.2 Synthesis Method

DeepPrivacy2 uses three independently trained generators for the three different detection categories introduced in section 4.1 (detection w/ cse, detection w/o cse, faces). While the tasks significantly differ in complexity, they share training setup and architecture to a high degree. Here, we first present our style-based generator, then present task-specific modeling choices for full-body and face synthesis. Each generator frame the anonymization task as an image inpainting task, where we remove areas to be anonymized and let a generator complete the missing region. Specifically, the input and output of each generator is given by [1],

$$\tilde{I} = G(I \odot M, M, z) \odot M + (1 - M) \odot I, \tag{1}$$

$\odot$ is element-wise multiplication, I the original image, and M indicates missing regions (M is 1 for known pixels and 0 for pixels to be anonymized).

4.2.1 A Style-Based U-Net Generator

Our synthesis method follows the implementation of Surface-guided GANs [14]. The generator is a U-Net [36] with limited task-specific modeling choices, consisting of a context encoder and a style-based decoder. The context encoder uses a sequence of convolutions and downsampling layers, with residual connections at every feature map resolution. We use no normalization layers in the encoder, as it performs similarly without it. However, we find it essential to apply instance normalization for the features in the U-net skip connections, where we combine features from the encoder and decoder as additive residuals. The decoder follows the design of Stylegan2 [18], with the operation order *instance normalization* → *convolution* → *style modulation*. Note that

[1] For the CSE-guided generator, the CSE-embedding is concatenated with $I \odot M$ to the input of the generator. See section 4.2.1.

we replace the baked-in weight demodulation in Stylegan2 with instance normalization, as we find that normalization on expected statistics works poorly when large areas of the input are missing.

Furthermore, we increase the depth of the U-net to 5 downsampling layers, such that the minimum feature resolution is 9×5. In contrast, SG-GAN has three downsampling blocks, where most parameters are at the 36×20 resolution. We observe no performance degradation by increasing the depth while improving inference speed, as more parameters are located at lower resolution layers. Finally, we remove V-SAM and discriminator surface supervision used in SG-GAN. Appendix C includes further details.

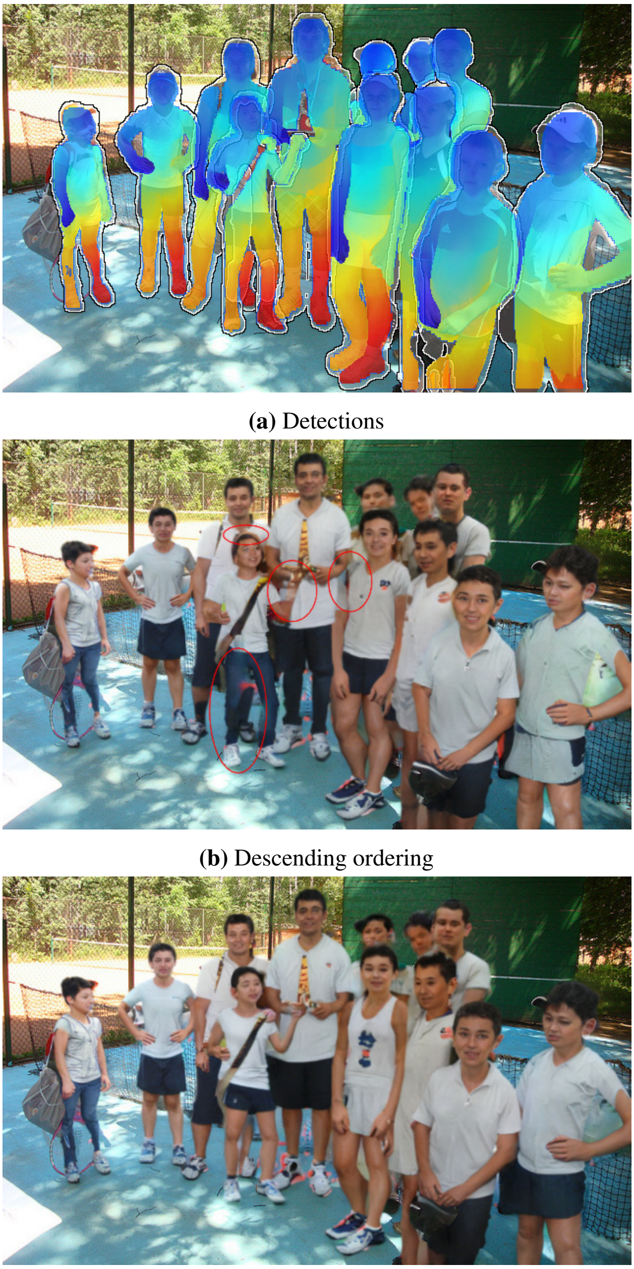
Full-Body Synthesis The full-body generator is trained on the FDH dataset at a resolution of 288×160. We train two independent generators for full-body synthesis; one that concatenates the CSE embedding to the input image and one that does not. The CSE embedding has a resolution of $16 \times 288 \times 160$, where we use the pixel-to-vertex embedding map released in the official implementation of CSE [30, 43].

Face Synthesis In contrast to previous face anonymization methods [13, 26, 40], we propose a generator that does not use keypoints for synthesis. Removing the keypoint detector improves detection recall in cases where keypoints are difficult to detect. We train the face generator on an updated version of the FDF dataset [13], which increases the image resolution to 256×256 from the original 128×128.

4.3 Recursive Stitching

The final stage of our pipeline is pasting the anonymized identities into the original image. Unlike face anonymization, full-body anonymization has many detection overlaps. If not handled correctly, these overlaps generates visually annoying artifacts at the border between individuals.

Our stitching approach recursively stitches each individual in ascending order depending on the number of pixels the person covers. The recursive stitching



(a) Detections

(b) Descending ordering

(c) Ascending ordering

Figure 4: Anonymization results comparing our method with descending and ascending image stitching order. The ascending ordering stitches foreground objects last, which improves image quality at detection borders (*e.g*. marked in red).

assumes that the synthesis method handles overlapping artifacts when generating each individual. Additionally, our ordering assumes that objects in the foreground cover a larger area, where foreground objects are stitched in last. The reverse order (foreground objects first) results in background objects "overwriting" foreground objects, as the detections can overlap (see fig. 4). This naive ordering significantly reduces visual artifacts at borders between individuals.

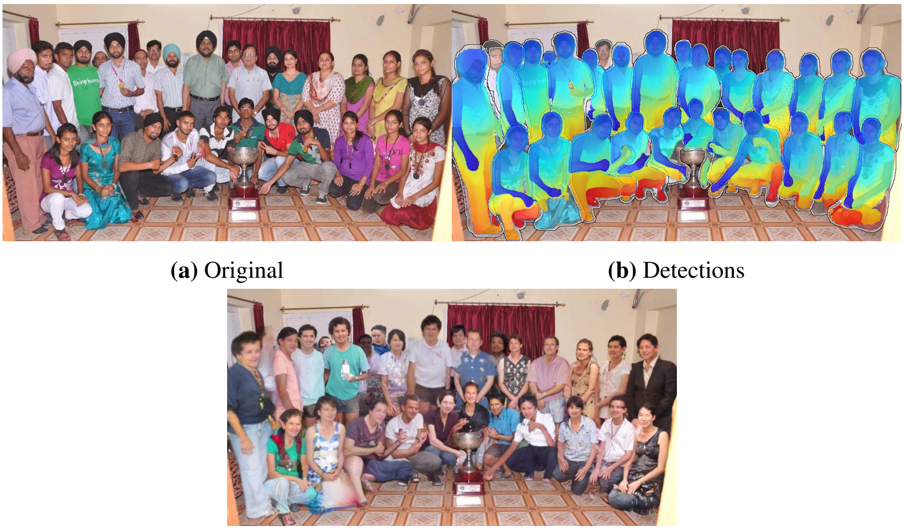
5 Experimental Evaluation

We validate our proposed anonymization pipeline in terms of synthesis quality, using anonymized data for future development, and anonymization guarantees. There are no standard baselines to compare against for realistic anonymization of data. Thus, we compare against traditional anonymization techniques, and DeepPrivacy [13], a widely adopted realistic face anonymizer. Additionally, we compare our full-body generator to Surface Guided GANs [14]. Appendix D includes random anonymized images on Cityscapes [4], COCO [23], FDH, and FDF256.

Experimental Details. All models are trained with Pytorch 1.10 [33] on 4 NVIDIA V100-32GB. For qualitative examples, we use multi-modal truncation to improve image quality while retaining diversity [27]. We report Fréchet Inception Distance (FID) [12] and FID_{CLIP} [2] [20] to evaluate image quality using Torch Fidelity [32]. The three generators (for CSE-guided, unconditional and face) has 43M parameters each.

Datasets. For training, we use the FDH dataset (see section 3) for full-body synthesis, and FDF256 for face synthesis. The FDF256 dataset is an updated version of FDF [13], where the image resolution is increased from 128 × 128 to 256 × 256 (see Appendix B). For evaluation, we use Market1501 [47], Cityscapes [4], and COCO [23]. We follow the standard train/validation split for all datasets.

[2] FID_{CLIP} is less sensitive to ImageNet classes. ImageNet-FID is insensitive to faces and scores images containing ImageNet objects (*e.g.* tie) higher [20].



(a) Original

(b) Detections

(c) Anonymized

Figure 5: DeepPrivacy2 anonymization result on an image from WIDER-Face. Appendix D includes random examples.

Runtime Analysis. The DeepPrivacy2 architecture is computationally efficient, where the CSE-guided generator processes ~ 11.6 frames per second (FPS), and the face generator at ~ 7.9 FPS on an NVIDIA 1080 8GB GPU. In contrast, the SG-GAN [14] generator processes ~ 7.3 FPS, where our improved runtime originate from the removal of V-SAM and moving the majority of parameters to lower resolution layers. The entire pipeline (detection, synthesis and stitching) require ~ 2.8 seconds to process an image with 12 persons on an NVIDIA 1080 8GB GPU.

5.1 Synthesis Quality

Full-Body Synthesis. Figure 6 shows diverse synthesized examples on the FDH dataset. Our model generates high-quality human figures that seamlessly transition into the original image. Furthermore, the model can handle a large variety of background contexts, poses, and overlapping objects. We find CSE guidance necessary for high-quality anonymization, where the un-

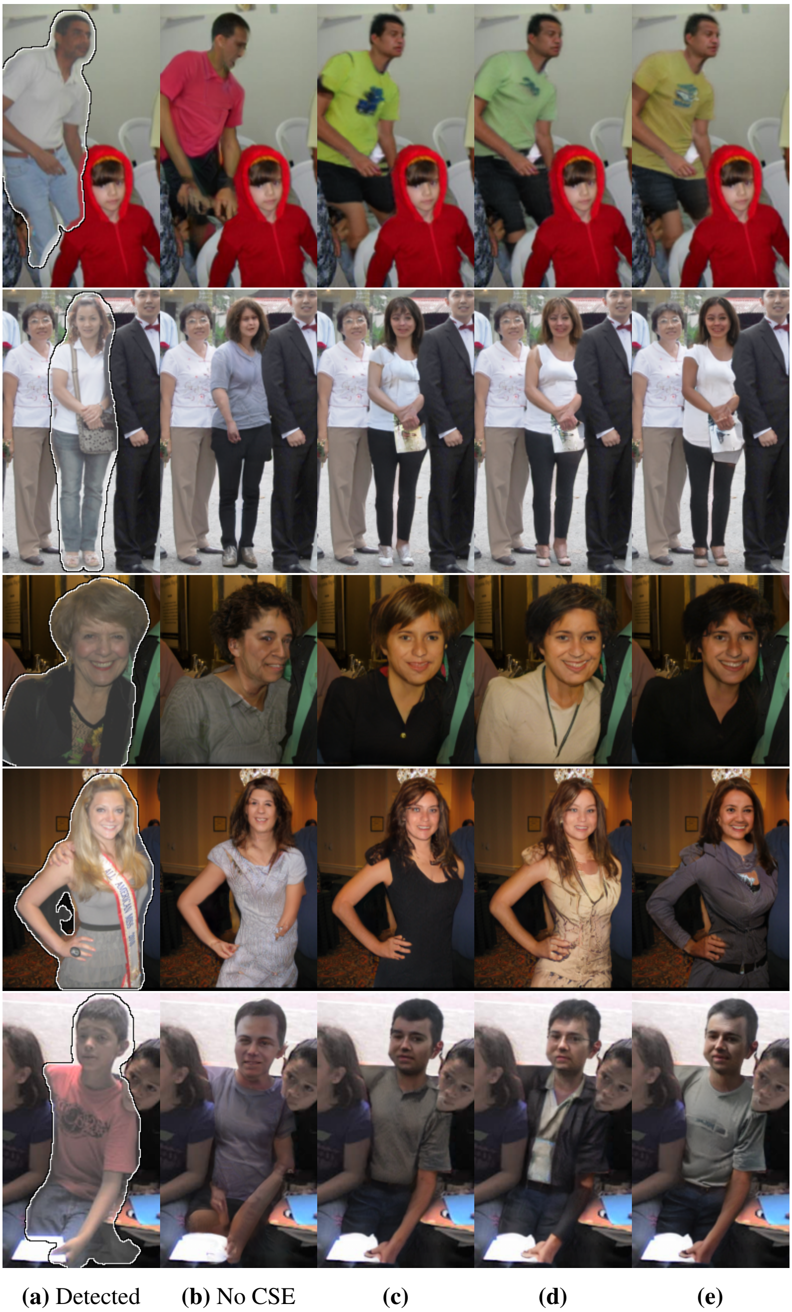


Figure 6: Synthesis results on FDH. (a) is the original identity and the anonymization mask, (b) is the unconditional generator, and (c-e) is the CSE-guided generator

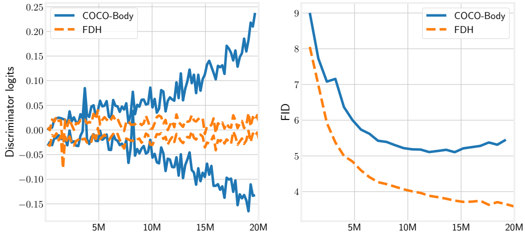


Figure 7: Our generator overfits early when trained on the small COCO-Body dataset (blue) [14], which consists of $\sim$ 40K images of human figures. Note that this occurs even with the strong data augmentation used by [14]. No data augmentation, other than horizontal flip, is used for FDH.

conditional generator often generates human figures with unnatural poses (see fig. 6). This is reflected in quantitative metrics, where the CSE-guided generator (FID=5.6, FID_{CLIP}=1.7) substantially improves over the unconditional generator (FID=6.1, FID_{CLIP}=2.30). Furthermore, the primary improvement of our model compared to Surface-guided GANs [14] is the larger and more diverse FDH dataset. The same model trained on the COCO-Body dataset [14] starts to overfit early in training, reflected by the diverging discriminator logits and increasing FID (fig. 7).

Face Synthesis. Figure 8 shows generated results on the FDF256 dataset. Directly comparing our face anonymizer to DeepPrivacy [13] is not straightforward, as we synthesize higher resolution images (256 × 256, not 128 × 128). Additionally, the FDF256 dataset does not include the same images as the original dataset, as FDF256 filter out lower resolution images. Nevertheless, to validate our modeling choices, we retrain our GAN for 128 × 128 synthesis on FDF [13]. Our GAN achieves a FID of 0.56, a significant improvement compared to DeepPrivacy (FID=0.68) [13]. Note that this is without using face keypoints, which the original DeepPrivacy uses to improve quality.

Figure 9 compares the open-source DeepPrivacy [13] to our method. Our method generates higher quality faces and handles overlaps between detections better. Also, note that DeepPrivacy does not anonymize all faces in the image, as it is unable to detect keypoints for all individuals [3].

[3] Even with confidence threshold of 0.05, DeepPrivacy is unable to detect keypoints for all

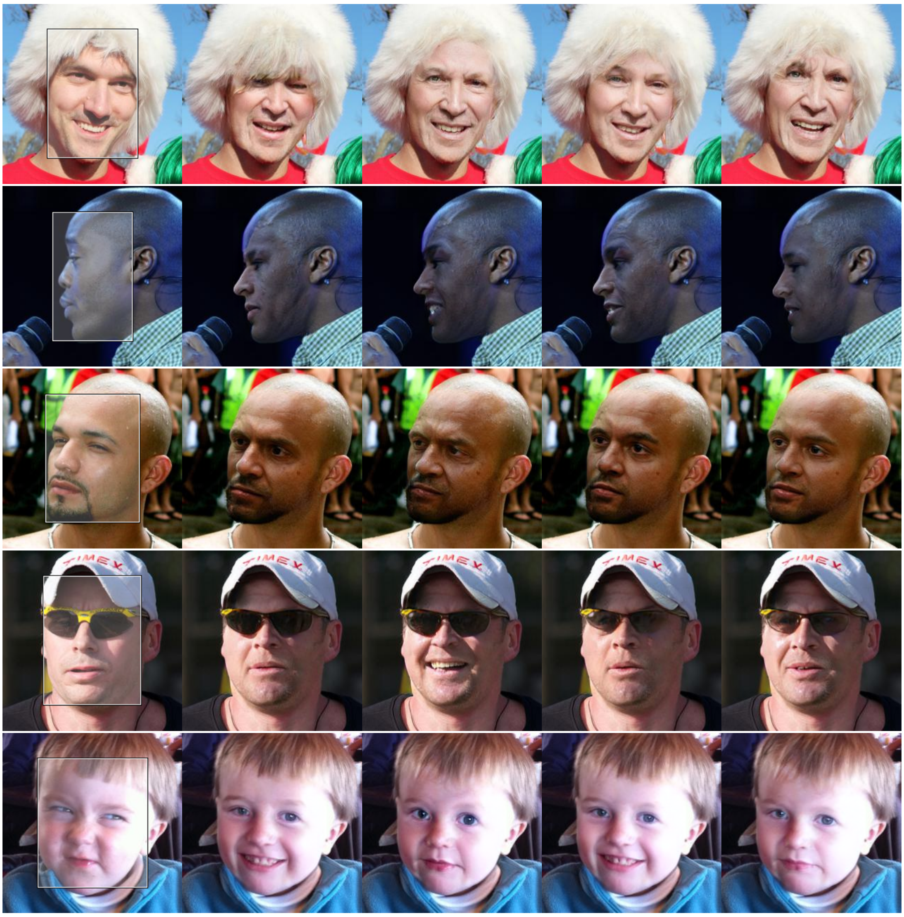
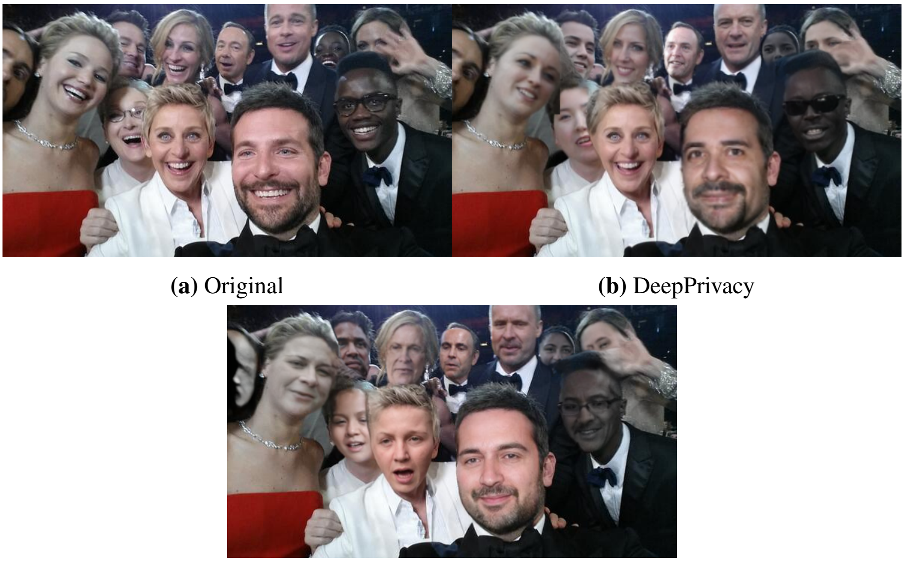


Figure 8: Synthesis results on FDF256. First column shows the original identity and the anonymization mask. Columns 2-5 shows generated identities from DeepPrivacy2.



(a) Original

(b) DeepPrivacy

(c) Ours

Figure 9: Face anonymization comparison between our method and DeepPrivacy [13].

Attribute-Guided Anonymization DeepPrivacy2 allows for controllable anonymization through text prompts by adapting StyleMC [19]. StyleMC finds global semantically meaningful directions in the GAN latent space by manipulating images towards a given text prompt with a CLIP-based [34] loss. Figure 10 shows attribute-guided anonymization, where the global directions are found over 256 images. As far as we know, DeepPrivacy2 is the first to enable controllable anonymization through text prompts, whereas previous methods are limited to no control or attribute preservation from the original identity [7] .

5.2 Anonymization Evaluation

Anonymization Guarantee To evaluate the anonymization guarantee of DeepPrivacy2, we evaluate how well automatic re-identification tools can

individuals.

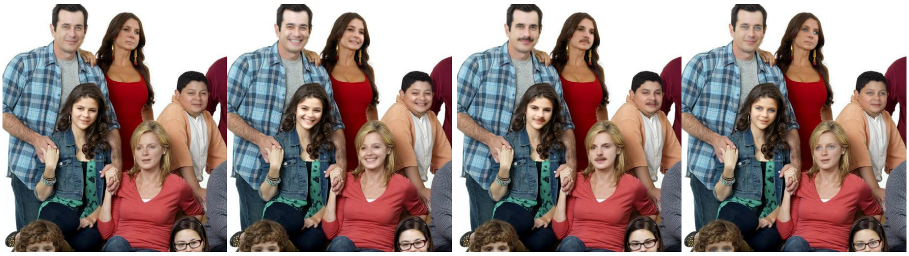


Figure 10: Latent manipulations with StyleMC [19]. The text prompt used for each edit is shown below each image. See our video for an interactive demo: https://youtu.be/faoNyaaORts.

Anonymization	R1 ↓	mAP ↓
Original	94.4	82.5
Pixelation 8 × 8	54.6	16.1
Pixelation 16 × 16	70.3	36.6
Mask-out	45.5	**8.0**
SG-GAN [14]	74.4	30.2
Full-body anonymization (Ours)	**44.7**	8.5

Table 1: Re-identification mAP and rank-1 accuracy on Market1501 [47] using OSNet [48].

identify anonymized individuals. Specifically, we evaluate the re-identification rate of OSNet [48] by anonymizing Market1501 [47]. In this case, a lower re-identification mAP and rank-1 accuracy (R1) reflects worse re-identification, indicating improved anonymization. Appendix C details the experiment further.

Table 1 reflects that pixelation enables re-identification of several of the anonymized individuals. Our full-body anonymizer yields similar anonymization guarantees as masking out the area and significantly improve compared to pixelation. Furthermore, the full-body anonymization of SG-GAN [14] provides poorer anonymization results than DeepPrivacy2. This is caused by the CSE detector failing in several cases and the poor segmentation of accessories/hair in SG-GAN, where the anonymized identity often "wears" parts of

	Train w/ Anon. Data		Validate w/ Anon. Data	
Dataset	Box AP ↑	Kp. AP ↑	Box AP ↑	Kp. AP↑
Original	53.6	64.0	53.6	64.0
Masked Out	10.1	0.5	17.0	1.8
Pixelation 8 × 8	10.4	1.0	29.1	2.2
Pixelation 16 × 16	10.1	1.5	36.5	12.0
DeepPrivacy2 (w/o CSE)	21.4	10.2	**49.9**	11.5
DeepPrivacy2	**26.0**	**31.9**	49.4	**48.4**

Table 2: Keypoint (Kp.) AP on the COCO [23] validation set with a Keypoint R-50 FPN R-CNN [10].

the original identity (see section 4.1).

5.2.1 Training and Evaluating on Anonymized Data

A typical use case for anonymization is collecting and anonymizing data for the development of computer vision systems. We evaluate DeepPrivacy2 on two established computer vision benchmarks: COCO [23] person keypoint estimation and Cityscapes [4] instance segmentation. We evaluate two use cases; (1) using anonymized data for training, and (2) using anonymized data for validation with a pre-trained model. For the former, we report evaluation metrics on the original validation dataset.

COCO Person Keypoint Estimation. Table 2 analyzes the effect of anonymization on the COCO dataset for person keypoint estimation. Pixelation greatly affects model training for the fine-grained task of keypoint estimation, whereas DeepPrivacy2 significantly improves over traditional methods. Note that CSE fails to detect many individuals in the COCO dataset, which yields poor pose preservation for individuals anonymized by the unconditional generator.

Cityscapes Instance Segmentation. Table 3 analyzes the effect of anonymization on the Cityscapes dataset. DeepPrivacy2 improves over pixelation and

	Train w/ Anon. Data		Validate w/ Anon. Data	
Dataset	mAP ↑	$\text{AP}_{\text{person}}$ ↑	mAP ↑	$\text{AP}_{\text{person}}$ ↑
Original	36.5	35.0	36.5	35.0
Masked Out	34.0	26.4	27.7	4.7
Pixelation 8×8	34.7	27.1	29.4	10.2
Pixelation 16×16	34.7	29.6	32.0	21.8
DeepPrivacy2 (w/o CSE)	33.4	27.5	33.1	**27.8**
DeepPrivacy2	**35.2**	**30.3**	**33.2**	27.3

Table 3: Instance segmentation AP on the Cityscapes [4] validation set with a Mask R-CNN [10] R-50 FPN.

mask-out, but the gap is less prevalent than for keypoint estimation. We believe this originates from model weight initialization [4].

Is surface guidance necessary? Section 5.1 established that the CSE-guided generator improves image quality compared to the unconditional generator. We now ask the question; does the improved image quality translate to improvements when using the anonymized data? In table 3 and table 2, we replace the CSE-guided generator with the unconditional generator, such that all persons are anonymized without CSE-guidance (denoted *DeepPrivacy2 w/o CSE*). Removing CSE-guidance severly hurts performance, especially when using the anonymized data for training.

6 Conclusion

DeepPrivacy2 is an automatic realistic anonymization framework for human figures and faces, and is a practical tool for anonymization without degrading the image quality. Compared to previously proposed anonymization frameworks, we show that DeepPrivacy2 substantially improves image quality and privacy guarantees. Furthermore, we introduce the FDH dataset, a large-scale full-body synthesis dataset that includes a wide variety of identities in different

[4] The Cityscapes model is initialized from a COCO pre-trained Mask R-CNN, while the keypoint R-CNN from an ImageNet [5] backbone.

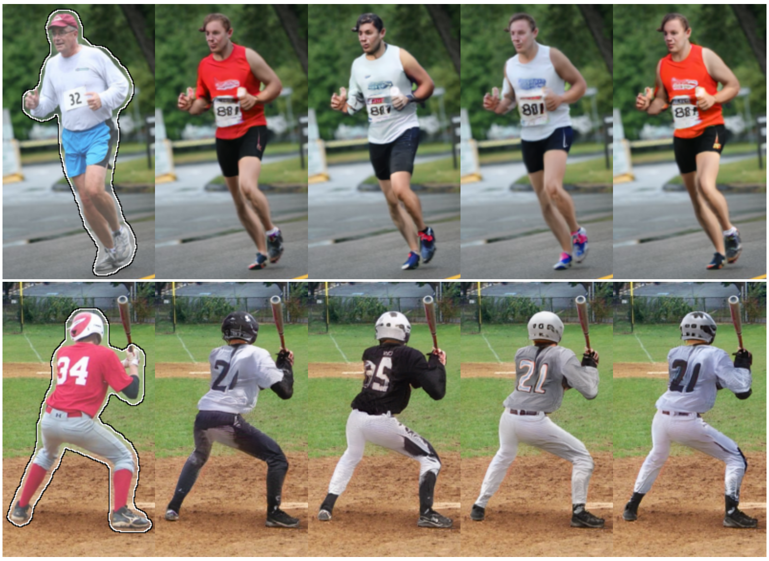


Figure 11: The generator samples from a small subset of different identities given the condition.

poses and contexts. Our new FDH dataset, combined with our simple style-based GAN, improves image quality and diversity of human figure synthesis for in-the-wild images. Furthermore, we show that our simple style-based GAN generates high-quality human faces that are controllable through user-guided anonymization via text prompts. We believe that our open-source framework will be a useful tool for computer vision researchers and other entities requiring anonymization while retaining image quality.

Societal Impact Recently introduced legislation in many regions has complicated collecting privacy-sensitive data, where consent from individuals is required for storing the data. This can act as a barrier for developing applications relying on high-quality images, such as computer vision models. This paper proposes an automatic realistic image anonymization framework that simplifies the collection of privacy-sensitive data while retaining the original image quality. We believe this will be a highly useful tool for the computer vision field. Nevertheless, our work focuses on synthesizing realistic humans, which has a potential for misuse (*e.g.* DeepFakes). There is a large focus in the community to mitigate this, for example, the DeepFake Detection Challenge [6] and model watermarking [46].

6.1 Limitations

DeepPrivacy2 generates a limited set of identities given a particular input condition. The input condition is highly descriptive of the shape of the original identity and the context that the identity should fit into. Thus, the generator learns a sampling probability of identities given the condition. For example, if the generator observes a baseball field, the synthesized identity is likely to be a baseball player (fig. 11).

As with any anonymization framework, DeepPrivacy2 cannot guarantee anonymization without human supervision, as the detector can fail. However, DeepPrivacy2 uses a set of detectors from different modalities to improve detection in cases where one or more of the detectors fail. Also, DeepPrivacy2 uses dense pose description for anonymization, which allows identity recognition through gait [15].

Synthesis Quality DeepPrivacy2 significantly improves full-body synthesis for in-the-wild images; however, it struggles in several scenarios. First, DeepPrivacy2 relies on dense pose estimation to synthesize high-quality human figures, where the image quality is severely degraded in cases where the pose description is incorrect. Furthermore, we find our full-body GAN harder to edit (*e.g.* attribute edit via text prompts [19]), and we observe that common directions in the latent space do not translate to semantically equivalent transformations for different poses/background contexts.

Acknowledgement The computations were performed on resources provided by the NTNU IDUN/EPIC computing cluster [39] and the Tensor-GPU project led by Prof. Anne C. Elster through support from The Department of Computer Science and The Faculty of Information Technology and Electrical Engineering, NTNU.

References

[1] Guha Balakrishnan, Amy Zhao, Adrian V Dalca, Fredo Durand, and John Guttag. Synthesizing Images of Humans in Unseen Poses. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 8340–8348, 2018.

[2] Karla Brkic, Ivan Sikiric, Tomislav Hrkac, and Zoran Kalafatic. I Know That Person: Generative Full Body and Face De-identification of People in Images. In *2017 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 1319–1328, 2017.

[3] Caroline Chan, Shiry Ginosar, Tinghui Zhou, and Alexei Efros. Everybody Dance Now. In *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 5932–5941, 2019.

[4] Marius Cordts, Mohamed Omran, Sebastian Ramos, Timo Rehfeld, Markus Enzweiler, Rodrigo Benenson, Uwe Franke, Stefan Roth, and Bernt Schiele. The Cityscapes Dataset for Semantic Urban Scene Understanding. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3213–3223, 2016.

[5] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. ImageNet: A large-scale hierarchical image database. In *2009 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 248–255, 2009.

[6] Brian Dolhansky, Joanna Bitton, Ben Pflaum, Jikuo Lu, Russ Howes, Menglin Wang, and Cristian Canton Ferrer. The DeepFake Detection Challenge (DFDC) Dataset. 2020.

[7] Oran Gafni, Lior Wolf, and Yaniv Taigman. Live Face De-Identification in Video. In *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 9377–9386, 2019.

[8] Ralph Gross, Latanya Sweeney, F. de la Torre, and Simon Baker. Model-Based Face De-Identification. In *2006 Conference on Computer Vision and Pattern Recognition Workshop (CVPRW)*, pages 161–161, 2006.

[9] Xintong Han, Zuxuan Wu, Zhe Wu, Ruichi Yu, and Larry S Davis. VITON: An Image-Based Virtual Try-on Network. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 7543–7552, 2018.

[10] Kaiming He, Georgia Gkioxari, Piotr Dollar, and Ross Girshick. Mask R-CNN. In *2017 IEEE International Conference on Computer Vision (ICCV)*, pages 2980–2988. IEEE, oct 2017.

[11] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep Residual Learning for Image Recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 770–778, 2016.

[12] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. In *Advances in Neural Information Processing Systems*, pages 6626–6637, 2017.

[13] Håkon Hukkelås, Rudolf Mester, and Frank Lindseth. DeepPrivacy: A Generative Adversarial Network for Face Anonymization. In George Bebis, Richard Boyle, Bahram Parvin, Darko Koracin, Daniela Ushizima, Sek Chai, Shinjiro Sueda, Xin Lin, Aidong Lu, Daniel Thalmann, Chaoli Wang, and Panpan Xu, editors, *Advances in Visual Computing*, pages 565–578. Springer International Publishing, 2019.

[14] Håkon Hukkelås, Morten Smebye, Rudolf Mester, and Frank Lindseth. Realistic full-body anonymization with surface-guided gans. *arXiv preprint arXiv:2201.02193*, 2022.

[15] Arun Jain, Anil and Flynn, Patrick and Ross. *Handbook of Biometrics*. Springer US, Boston, MA, 2008.

[16] Amin Jourabloo, Xi Yin, and Xiaoming Liu. Attribute preserved face de-identification. *2015 International Conference on Biometrics (ICB)*, pages 278–285, 2015.

[17] Tero Karras, Samuli Laine, and Timo Aila. A Style-Based Generator Architecture for Generative Adversarial Networks. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4396–4405, 2019.

[18] Tero Karras, Samuli Laine, Miika Aittala, Janne Hellsten, Jaakko Lehtinen, and Timo Aila. Analyzing and Improving the Image Quality of StyleGAN. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 8107–8116, 2020.

[19] Umut Kocasari, Alara Dirik, Mert Tiftikci, and Pinar Yanardag. StyleMC: Multi-Channel Based Fast Text-Guided Image Generation and Manipulation. In *2022 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pages 3441–3450, 2022.

[20] Tuomas Kynkäänniemi, Tero Karras, Miika Aittala, Timo Aila, and Jaakko Lehtinen. The Role of ImageNet Classes in Frechet Inception Distance. *arXiv preprint arXiv:2203.06026*, 2022.

[21] Jian Li, Yabiao Wang, Changan Wang, Ying Tai, Jianjun Qian, Jian Yang, Chengjie Wang, Jilin Li, and Feiyue Huang. DSFD: Dual Shot Face Detector. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5055–5064, 2019.

[22] Yining Li, Chen Huang, and Chen Change Loy. Dense Intrinsic Appearance Flow for Human Pose Transfer. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3688–3697, 2019.

[23] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. Microsoft COCO: Common Objects in Context. In *Computer Vision – ECCV 2014*, pages 740–755. Springer International Publishing, 2014.

[24] Ziwei Liu, Ping Luo, Shi Qiu, Xiaogang Wang, and Xiaoou Tang. DeepFashion: Powering Robust Clothes Recognition and Retrieval with Rich Annotations. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1096–1104, 2016.

[25] Liqian Ma, Qianru Sun, Stamatios Georgoulis, Luc Van Gool, Bernt Schiele, and Mario Fritz. Disentangled Person Image Generation. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 99–108, 2018.

[26] Maxim Maximov, Ismail Elezi, and Laura Leal-Taixe. CIAGAN: Conditional Identity Anonymization Generative Adversarial Networks. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5446–5455, 2020.

[27] Ron Mokady, Michal Yarom, Omer Tov, Oran Lang, Daniel Cohen-Or, Tali Dekel, Michal Irani, and Inbar Mosseri. Self-Distilled StyleGAN: Towards Generation from Internet Photos. 2022.

[28] Wiktor Muron. motpy - simple multi object tracking library. `https://github.com/wmuron/motpy`, 2022.

[29] Natalia Neverova, Rıza Alp Güler, and Iasonas Kokkinos. Dense Pose Transfer. In *Computer Vision – ECCV 2018*, pages 128–143. Springer International Publishing, 2018.

[30] Natalia Neverova, David Novotny, Vasil Khalidov, Marc Szafraniec, Patrick Labatut, and Andrea Vedaldi. Continuous Surface Embeddings. In *Advances in Neural Information Processing Systems*, pages 17258—-17270. Curran Associates, Inc., 2020.

[31] E.M. Newton, L. Sweeney, and B. Malin. Preserving privacy by de-identifying face images. *IEEE Transactions on Knowledge and Data Engineering*, 17(2):232–243, 2005.

[32] Anton Obukhov, Maximilian Seitzer, Po-Wei Wu, Semen Zhydenko, Jonathan Kyl, and Elvis Yu-Jing Lin. High-fidelity performance metrics for generative models in PyTorch. `https://github.com/toshas/torch-fidelity`, 2020.

[33] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, and Others. Pytorch: An imperative style, high-performance deep learning library. *Advances in neural information processing systems*, 32, 2019.

[34] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning Transferable Visual Models From Natural Language Supervision. In *International Conference on Machine Learning*, pages 8748—-8763, 2021.

[35] Zhongzheng Ren, Yong Jae Lee, and Michael S Ryoo. Learning to Anonymize Faces for Privacy Preserving Action Detection. In *Computer Vision – ECCV 2018*, pages 639–655. Springer International Publishing, Cham, 2018.

[36] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-Net: Convolutional Networks for Biomedical Image Segmentation. In *International Conference on Medical image computing and computer-assisted intervention*, pages 234–241. Springer, 2015.

[37] Kripasindhu Sarkar, Vladislav Golyanik, Lingjie Liu, and Christian Theobalt. Style and Pose Control for Image Synthesis of Humans from a Single Monocular View. *arXiv preprint arXiv:2102.11263*, feb 2021.

[38] Kripasindhu Sarkar, Dushyant Mehta, Weipeng Xu, Vladislav Golyanik, and Christian Theobalt. Neural Re-rendering of Humans from a Single Image. In *Computer Vision – ECCV 2020*, pages 596–613. Springer International Publishing, Cham, 2020.

[39] Magnus Själander, Magnus Jahre, Gunnar Tufte, and Nico Reissmann. EPIC: An energy-efficient, high-performance GPGPU computing research infrastructure, 2019.

[40] Qianru Sun, Liqian Ma, Seong Joon Oh, Luc Van Gool, Bernt Schiele, and Mario Fritz. Natural and Effective Obfuscation by Head Inpainting. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5050–5059, 2018.

[41] Qianru Sun, Ayush Tewari, Weipeng Xu, Mario Fritz, Christian Theobalt, and Bernt Schiele. A Hybrid Model for Identity Obfuscation by Face Replacement. In *Computer Vision – ECCV 2018*, pages 570–586. Springer International Publishing, Cham, 2018.

[42] Bart Thomee, David A. Shamma, Gerald Friedland, Benjamin Elizalde, Karl Ni, Douglas Poland, Damian Borth, and Li-Jia Li. YFCC100M: The new data in multimedia research. *Commun. ACM*, 59(2):64–73, jan 2016.

[43] Yuxin Wu, Alexander Kirillov, Francisco Massa, Wan-Yen Lo, and Ross Girshick. Detectron2. `https://github.com/facebookresearch/detectron2`, 2019.

[44] Saining Xie, Ross Girshick, Piotr Dollar, Zhuowen Tu, and Kaiming He. Aggregated Residual Transformations for Deep Neural Networks. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5987–5995, 2017.

[45] Shuo Yang, Ping Luo, Chen Change Loy, and Xiaoou Tang. WIDER FACE: A Face Detection Benchmark. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5525–5533, 2016.

[46] Ning Yu, Vladislav Skripniuk, Sahar Abdelnabi, and Mario Fritz. Artificial Fingerprinting for Generative Models: Rooting Deepfake Attribution in Training Data. In *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 14428–14437, 2021.

[47] Liang Zheng, Liyue Shen, Lu Tian, Shengjin Wang, Jingdong Wang, and Qi Tian. Scalable Person Re-identification: A Benchmark. In *2015 IEEE International Conference on Computer Vision (ICCV)*, pages 1116–1124, 2015.

[48] Kaiyang Zhou, Yongxin Yang, Andrea Cavallaro, and Tao Xiang. Omni-Scale Feature Learning for Person Re-Identification. In *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 3701–3711, 2019.

Paper E

Synthesizing Anyone, Anywhere, in Any Pose

Authors:

Håkon Hukkelås, Frank Lindseth

Published at conference:

2024 IEEE/CVF Winter Conference on Applications of Computer Vision

Copyright:

© 2024 IEEE

Source Code:

`https://github.com/hukkelas/deep_privacy2`

Appendix:

`https://bird.unit.no/resources/0e91e613-e863-4fa4-83af-22c59ee1b595/content`

Paper E

Synthesizing Anyone, Anywhere, in Any Pose

Håkon Hukkelås Frank Lindseth

Norwegian University of Science and Technology
Trondheim, Norway
hakon.hukkelas@ntnu.no

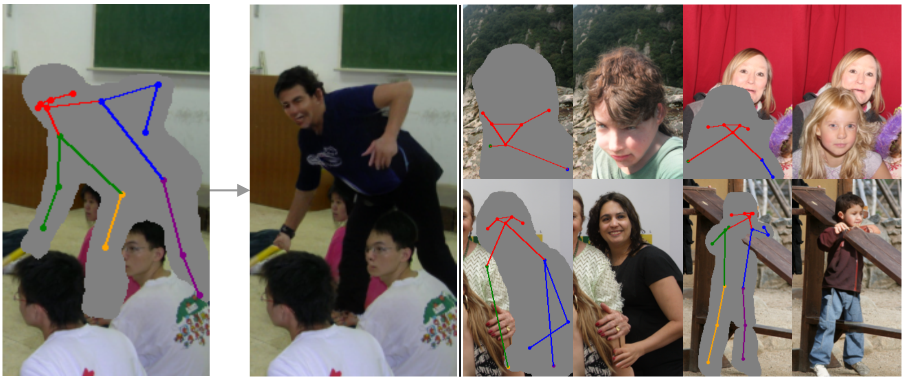


Figure 1: TriA-GAN can synthesize realistic human figures given a masked image and a sparse set of keypoints.

Abstract

We address the task of in-the-wild human figure synthesis, where the primary goal is to synthesize a full body given any region in any image. In-the-wild human figure synthesis has long been a challenging and under-explored task, where current methods struggle to handle extreme poses, occluding objects, and complex backgrounds.

Our main contribution is TriA-GAN, a keypoint-guided GAN that can synthesize Anyone, Anywhere, in Any given pose. Key to our method is projected GANs combined with a well-crafted training strategy, where our simple generator architecture can successfully handle the challenges of in-the-wild full-body synthesis. We show that TriA-GAN significantly improves over previous in-the-wild full-body synthesis methods, all while requiring less conditional information for synthesis (keypoints *vs*. DensePose). Finally, we show that the latent space of TriA-GAN is compatible with standard unconditional editing techniques, enabling text-guided editing of generated human figures.

1 Introduction

Given any image with a missing region, can you imagine a human appearance fitting into it? If there is a football next to the missing region, does your imaginary person change? This is a fascinating and difficult problem because countless possible solutions could fit the context. We refer to this task as in-the-wild human figure synthesis. Addressing this problem requires a complex understanding of human appearances and how they vary based on different environmental conditions, viewpoints, poses, and sizes of the missing region. Such a system would have widespread applications in content creation, fashion [37], or even for anonymization purposes [18].

Human figure synthesis is a well-established research field with many high-level goals. However, *in-the-wild* human figure synthesis is a difficult and under-explored task. Previous methods focus on simpler tasks, such as transferring a known appearance into a given pose [2, 4], transferring garments [14, 52], or full-body synthesis into a plain background [9]. Often they disregard the key difficulties of in-the-wild-synthesis, such as overlapping objects, partial bodies, complex backgrounds, and extreme poses. In fact, recent studies filter out these difficult cases from their dataset to improve synthesis quality [9, 10]. To the best of our knowledge, only a handful of research studies have tackled these challenges, with a focus on full-body synthesis for anonymization [18, 20] [1]. While previous methods [18] generate visually pleasing results, they heavily rely on DensePose estimation and struggle in complex scenarios. In addition, the generated images are hard to edit [18].

A key issue of current methods for in-the-wild human figure synthesis is their reliance on DensePose annotations [18, 20]. The available datasets with such annotations are either limited in size [12, 20] or automatically annotated [18]. We argue that this reliance constrain these methods, either by overfitting on small datasets [20] or by the numerous annotation errors arising from DensePose [18].

This paper explores full-body synthesis conditioned on sparse 2d-keypoints, eliminating the need for expensive DensePose annotations. However, this

[1] Note that other studies address similar tasks [40, 60], but they focus on simpler datasets (*i.e.* Market1501 [75], DeepFasion [37]) with few overlapping/occluding objects.

increases the modeling complexity considerably, as the generative model must now infer both the body's texture *and* its structure. We find that current GANs [18] struggle to synthesize realistic human figures without DensePose correspondences.

Our contributions address the challenge of scaling up GANs to handle in-the-wild full-body synthesis without DensePose correspondences. Key to our method is replacing the conventional GAN discriminator with Projected GANs [53]. By combining Projected GANs with a thoughtfully designed training strategy, our method can generate coherent bodies with visually pleasing textures.

Our contributions can be summarized as follows. First, we adapt Projected GANs [53] for image inpainting (section 3.1), and propose a novel mask-aware patch discriminator (section 3.2). Secondly, we investigate the representational power of pre-trained feature networks used by the discriminator (section 3.3). Our experiments reflect that the previously used classification networks [53, 54] are poorly suited for discriminating human figures. Instead, we use a combination of self-supervised feature networks for the discriminator, which significantly improves sample quality. Finally, we propose a progressive training technique for U-Net [50] architectures (section 3.4), enabling us to easily scale up to high resolutions and larger model sizes.

Our contributions culminate into a new state-of-the-art for in-the-wild human figure synthesis. As far as we know, our approach is the first to generate nearly photorealistic humans without DensePose annotations while effectively dealing with extreme poses, complex backgrounds, partial bodies, and occlusions. Source code: `http://github.com/hukkelas/deep_privacy2/`.

2 Related Work

2.1 Full-body Human Synthesis

Synthesizing human bodies has a range of applications, and previous studies have a large variety of high-level goals. We categorize human synthesis into *transfer-based* and *synthesis-based* models. *Transfer-based* methods transfers a source appearance (or garment [14, 52]) into a new pose [2, 33, 39, 47, 52, 56], motion [4] or scene [57]. While some of these methods are applicable for in-the-wild human figure synthesis [57, 67], they require a source appearance that limits the synthesized identities to a texture bank or an image dataset of appearances. In contrast, our method can directly synthesize novel identities. For the latter goal, *synthesis-based* methods can synthesize the appearance either conditioned on a pose [40, 60, 68], scene [8, 18, 20], or unconditionally [9, 5, 10]. Several of these methods are applicable for in-the-wild human synthesis [18, 40, 60], but they are limited to low-resolution [8, 40], struggle to handle complex backgrounds [9, 60], and only a few handles overlapping objects [18, 20].

Independent of the goal, most methods use a form of pose information to enhance synthesis quality through DensePose annotations [18, 20, 43, 52], semantic segmentations [5, 60, 67], sparse keypoints [2, 4, 8, 14, 33, 39, 40, 47, 56, 57, 67], or a 3d pose of the body [32, 68].

Previous studies primarily focus on GAN-based methods, but recent studies have employed diffusion models [59] for human figure synthesis [22]. Our work focuses on GANs as they offer fast sampling of high-quality images.

2.2 Generative Adversarial Networks

Generative Adversarial Networks [11] (GANs) have long been a leading generative model for a range of full-body synthesis tasks. GANs are notoriously difficult to train, and a notable research focus has been on achieving stable training of the generator, where different techniques such as novel objectives [1], architectures [24, 26, 27, 28], training strategies [25], and regularization [13, 41] has been proposed to improve stability and synthesis quality. Recently introduced Projected GANs [53] use pre-trained feature networks for the discriminator to reduce training time and improve image quality, which was later extended for high-resolution image synthesis on the ImageNet [6] dataset [54]. We continue this line of research, where we adapt projected GANs for conditional synthesis.

2.3 Image Inpainting

Image inpainting [3] aims to complete missing regions in natural images. Unlike general image inpainting, we complete missing regions that contain human figures appearing at random regions in natural images. GANs have long been the leading methodology for free-form image inpainting [46, 70], where most prior work focuses on architectural changes to the generator. For example, to handle missing values [19, 35, 70], generate higher resolution [69], utilize auxiliary information [23, 31, 42], or improve the receptive field via attention mechanisms [71] or fourier convolutions [62]. Previous methods adapt a traditional GAN discriminator, often patch discriminators [21, 36, 49, 65, 70], combined with perceptual image similarity losses [36, 49] and pixel-wise l_1 loss [49, 65]. As far as we know, we are the first to adapt Projected GANs [53] for image inpainting, where we exclusively train on the adversarial objective.

3 TriA-GAN - A Keypoint-Guided GAN

In this section, we gradually introduce changes to improve synthesis quality (table 1). **Config A** (section 3.1) starts with a StyleGAN-based [27] U-Net [50] architecture, similar to the architecture used in [18], trained with Projected GANs [54] using EfficientNet-Lite0 [63]. **Config B** introduces our Mask-Aware Discriminator objective (section 3.2), and **Config C** replaces EfficientNet-lite0 with ViT-L16$_{\text{MAE}}$ and RN50$_{\text{CLIP}}$ (section 3.3). **Config D** introduces our progressive training technique (section 3.4) and finally, **Config E** increases the generator model size. To reduce training time, we ablate our method on low-resolution images (72×40). Finally, section 3.5 increases the resolution to 288×160. Appendix A includes experimental and architecture details.

Problem Formulation We formulate in-the-wild full-body synthesis as an image inpainting task. Our goal is to complete the missing regions of a corrupted image $\bar{I} = I \odot M$, where I is the ground truth image, M is the mask indicating missing regions ($M_i = 1$ for known pixels and 0 for missing), and $\odot$

Table 1: Iterative development of our method. Each addition is added on top of the previous. Config A-C are trained until the discriminator has observed 50M images.

Configuration	FID ↓	FID_{CLIP} ↓	PPL ↓	OKS ↑
A: Baseline	1.73	1.74	55.8	0.916
B: + Mask-Aware Discriminator	1.65	1.63	52.8	0.912
C: + Improved Feature Nets	1.79	0.47	49.2	0.951
D: + Progressive Growing	1.66	0.40	52.0	**0.954**
E: + Larger G (62M → 110M)	**1.62**	**0.30**	**52.0**	0.948

is element-wise multiplication. To improve synthesis quality, we condition the generator on 17 keypoints following the COCO [34] keypoint format

Dataset We conduct our experiments on the FDH dataset [18]. The FDH dataset is a large unfiltered dataset, where models trained on FDH adapt well to in-the-wild settings [18]. The dataset consists of 1.87M training images and 30K validation images. Each image includes a single human figure as the subject, but the same image can include several individuals. Each image is annotated with a 2d keypoint annotation, a segmentation mask indicating the human to be inpainted, and pixel-to-surface correspondences (*i.e.* surface of a T-shaped 3D body). Note that TriA-GAN does not use pixel-to-surface correspondences.

We find that a large amount of the keypoint annotations in the FDH dataset are incorrect. Thus, we automatically re-annotate all images with ViTPose [66] (see Appendix B).

Pose Representation We represent keypoints as a one-hot encoded spatial map, specifically $P \in \{0,1\}^{K\times H\times W}$ where $K = 17$ and $P_{k,y,x} = 1$ for keypoint k with location (x,y) and P is 0 otherwise. In addition, we include a spatial map (S) drawing the human skeleton. Specifically, the spatial map $S \in \{0,1\}^{6\times H\times W}$ is one-hot encoded into 6 categories, where lines connect closeby joints in the body, separated into 6 classes (left/right arm/leg, torso, head). The one-hot encoded pose and the skeleton map are concatenated with the input image of the generator.

Evaluating Sample Quality We evaluate sample quality with Fréchet Inception Distance (FID) [16] and FID_{CLIP}[2]. Additionally, we report latent disentanglement via Perceptual Path Length (PPL) [27], which correlates with consistency and stability of shapes [28].

Furthermore, we introduce a new metric for assessing the sample quality of generated human figures, namely Object Keypoint similarity (*OKS*), that compares the generated pose to the ground truth keypoints. The motivation behind this metric is to obtain a metric that is not influenced by the feature network used by the discriminator. Projected GANs [53] are known to achieve artificially good scores on feature-based metrics [30], which makes it challenging to make quantitative comparisons across different types of feature networks. This is evident from our experiments, where Config B (which uses ImageNet features for the discriminator) generates severely more corrupted images than Config E but still achieves a similar ImageNet FID.

Object Keypoint Similarity (*OKS*) is calculated by predicting keypoints with ViTPose [66], then computing the OKS to the ground truth keypoints following COCO [34]. Compared to direct Euclidean distance, OKS considers that "correct" keypoints can deviate slightly from the ground truth keypoints, where the acceptable deviation varies for different keypoints (*e.g.* the shoulder keypoint can deviate more than the eye keypoint).

3.1 Projected GANs for Image Inpainting

Projected GANs [53] employ pre-trained feature networks to discriminate between real and fake images. Given an image I, the adversarial objective is formulated as

$$\begin{aligned} \min_{G} \max_{D_\ell} \sum_{\ell \in \mathscr{L}} & \mathbb{E}_{I \sim p_{data}} [\log (D_\ell (P_\ell (I)))] + \\ & \mathbb{E}_{z \sim p_z} [\log (1 - D_\ell (P_\ell (G(z, \bar{I}))))] , \end{aligned} \tag{1}$$

[2] ImageNet-FID scores images containing ImageNet objects higher and is insensive to faces [30]. These issues are diminished with FID_{CLIP}, where we use features from a CLIP [48] pre-trained ViT-B/32.

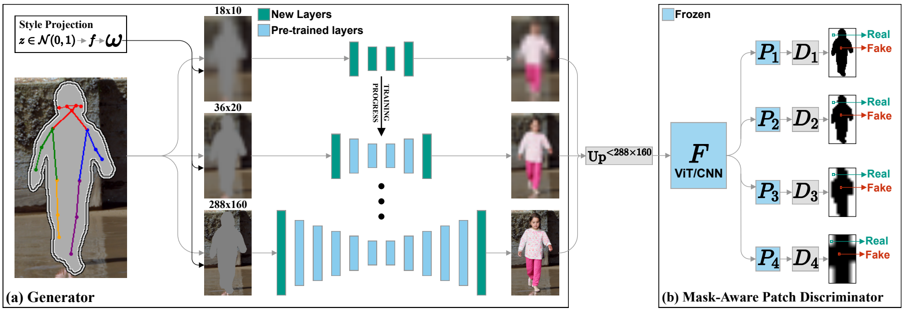


Figure 2: (a) Our generator fills in the missing region given 17 keypoints. The generator layers employ adaptive instance normalization to condition the generator on ω, where ω is the output of the style mapping network. Config D&E is trained progressively starting at 18×10 resolution, then increased by adding layers to the start/end of the encoder/decoder. Note that all layers remain trainable throughout training. (b) For each feature network F, we use four shallow patch discriminators operating its features (with different spatial resolutions), where each feature is projected through random differentiable operations (P_1-P_4). Given the projected features, each discriminator predicts if a given patch corresponds to a real or fake image region.

where $\{D_\ell\}$ is a set of independent discriminators operating on its feature projector P_ℓ. Each projector is frozen during training and consists of a pretrained feature network F, where features from F are randomly projected with differentiable operations. For the baseline (**Config A**), we use EfficientNet-Lite0 [63] as F following [53], which we later revisit in Section 3.3. For each discriminator D_ℓ, we adopt a patch discriminator architecture, described in Section 3.2.

Equation (1) does not enforce consistency between the condition ($\bar{I}$) and the generated image, yielding a generator that learns to completely ignore $\bar{I}$ in practice. Thus, we enforce condition consistency by masking the output of the generator. Specifically, we set $G(z,\bar{I}) = \tilde{I} \odot (1-M) + \bar{I} \odot M$, where $\tilde{I}$ is the output of the last layer in G.

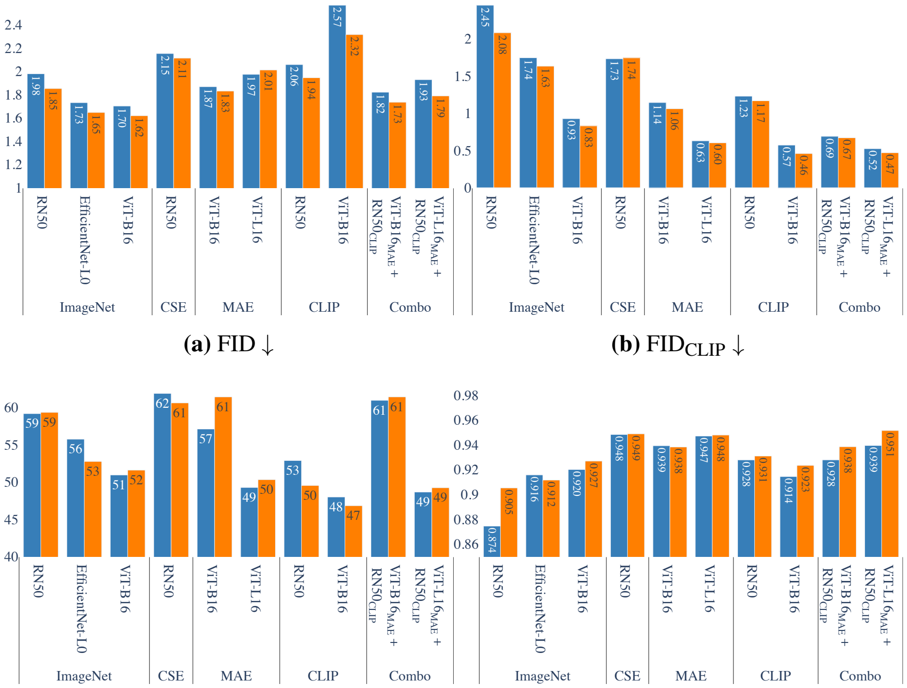
3.1.1 Stabilizing the Generator

Naively adopting projected GANs for image inpainting is unstable to train and prone to mode collapse early in training. This originates from the generator struggling to keep up with the pre-trained discriminator, where the discriminator overpowers the generator early in training. To improve stability, we introduce several modifications to the adversarial setup. First, we blur images inputted to the discriminator at the start of training, where the blur is linearly faded over 4M images. The long blur prevents the discriminator from focusing on the high-frequency edges caused by the masking of the generator output. Previous methods apply discriminator blurring over the first 200k images [26, 54], whereas we find it beneficial to significantly increase this period. Furthermore, the U-net architecture injects the latent code (z) via a mapping network and style modulation following StyleGAN2 [28]. We set the mapping network to 2 layers and reduce the dimensionality of z to 64, following [54]. Furthermore, we scale residual skip connections by $^1/\sqrt{2}$ (similar to [28]), and $^1/\sqrt{3}$ for skip connections where residual U-net connections are present. Finally, we use instance normalization instead of weight demodulation [28], as we find it more stable to train.

3.2 Mask-Aware Patch Discriminator

Projected GANs [53, 54] adapt four shallow discriminators operating on different feature projections (P_ℓ) with different spatial resolutions. Each discriminator output logits at the same resolution (4×4). In contrast, we find patch discriminators to work better for the image inpainting task, where each discriminator tries to classify local patches instead of the global image. Specifically, each D_ℓ (inputting features from the projection P_ℓ) consists of three convolutions, where the output of D_ℓ is half the spatial resolution of P_ℓ. We find that replacing the discriminator from [53] with a patch discriminator substantially improves performance.

Patch discriminators are widely adapted for image inpainting [62, 70, 73, 74]. Typically, each patch is classified as belonging to the class of the original image, such that all patches corresponding to a real image are classified as real. However, this introduces ambiguity for the image inpainting task, as certain features (*e.g.* shallow features from CNNs) might exclusively depend on real pixels even though the image is fake due to a limited receptive field. Thus, we propose a mask-aware discriminator objective, where the discriminator's patches are categorized as belonging to the real or fake class based on whether



(a) FID ↓

(b) FID$_{\text{CLIP}}$ ↓

(c) PPL ↓

(d) OKS ↑

Figure 3: Comparison of different feature networks with the standard projected GAN objective (eq. (1)) and mask-aware discriminator objective (eq. (2)). All models are trained until the discriminator has observed 50M images.

they correspond to a real or fake region in the image. The new objective is given by

$$\begin{aligned}\min_G \max_{D_\ell} \sum_{\ell \in \mathscr{L}} \mathbb{E}_{I \sim p_{data}} \left[\log\left(D_\ell\left(P_\ell\left(I\right)\right)\right)\right] + \\ \mathbb{E}_{z \sim p_z}\left[\sum_y^{H_\ell}\sum_x^{W_\ell} M_\ell^{y,x} \cdot \log\left(D_\ell^{y,x}\left(P_\ell\left(G\left(z,\bar{I}\right)\right)\right)\right) + \right. \\ \left. \left(1 - M_\ell^{y,x}\right) \cdot \log(1 - D_\ell^{y,x}(P_\ell(G(z,\bar{I})))) \right],\end{aligned} \tag{2}$$

where $D_\ell \in \mathbf{R}^{H_\ell \times W_\ell}$, and M_ℓ is downsampled from M to $H_\ell \times W_\ell$ via min-pooling.

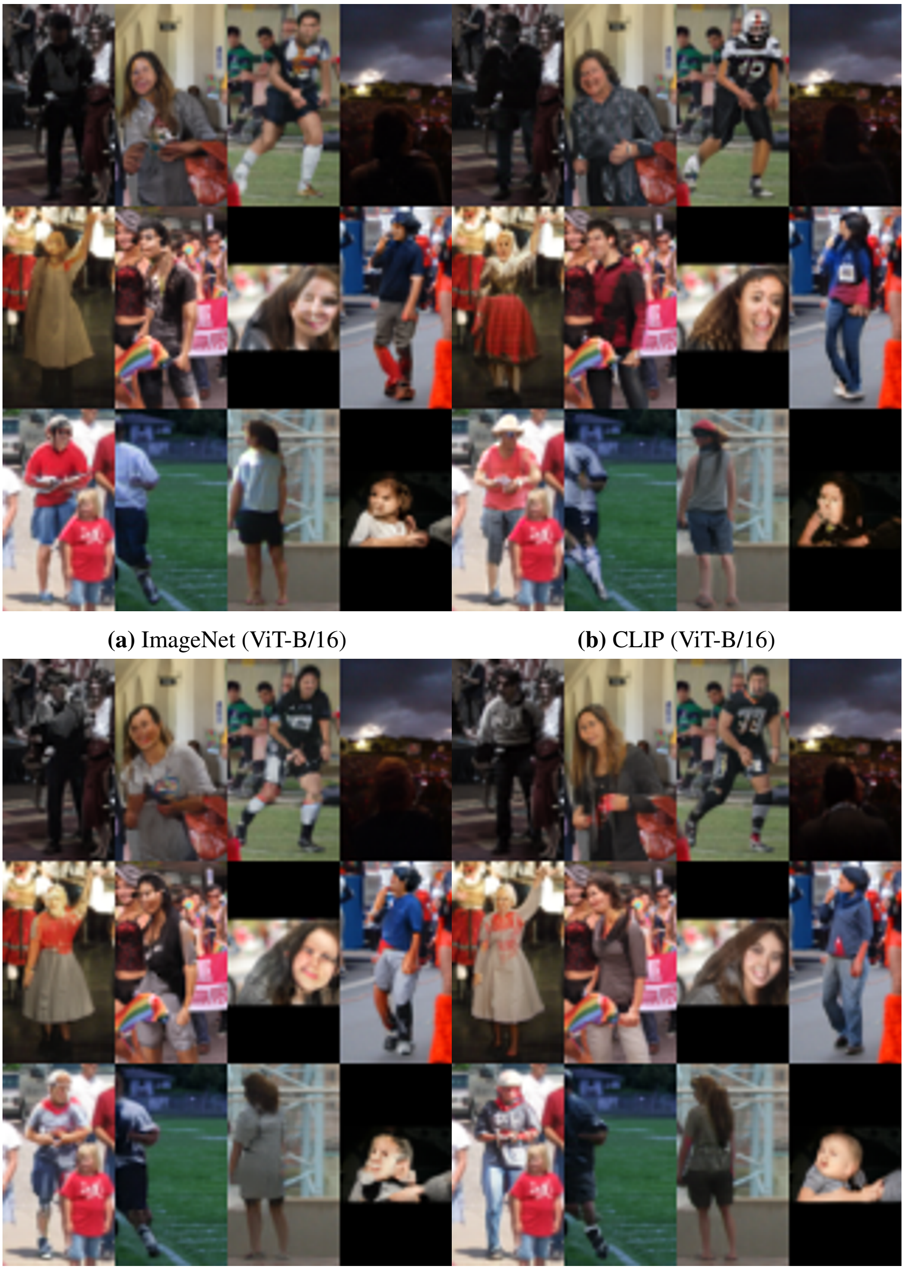
Equation (2) removes the ambiguous classification of patches due to global class allocation, which provides more detailed and spatial coherent responses to the generator. Furthermore, it introduces an auxiliary task to the discriminator, which is known to improve synthesis quality [45]. In our case, the auxiliary task is to spatially segment the region that corresponds to the generated area.

Figure 3 confirms that Equation (2) improves image quality (FID/FID$_{\text{CLIP}}$) and OKS across a range of feature networks. This includes feature networks with different pre-training tasks and architectures (CNNs and ViTs). Similar segmentation discriminators have been explored before for other tasks [55, 61, 68]. Our work further validate that this concept generalizes to extremely shallow discriminator architectures leveraging pre-trained feature networks, independent on the feature network used as F.

3.3 Discriminative Feature Networks for Human Synthesis

GANs have historically generated impressive results for aligned human synthesis, especially on the FFHQ [27] and CelebA-HQ [25, 38] datasets. However, projected GANs are known to generate artifacts for face synthesis on FFHQ [53] and struggle to generate realistic images of unaligned humans [54] [3]. We find that the poor human synthesis quality originates from an invariance in the pre-trained feature space used by the discriminator. Earlier work [53, 54]

[3] See the appendix in [54].



(a) ImageNet (ViT-B/16)

(b) CLIP (ViT-B/16)

(c) MAE (ViT-B/16)

(d) ViT-L/16$_{\text{MAE}}$ + RN50$_{\text{CLIP}}$

Figure 4: Qualitative comparison of various feature networks used for the discriminator. It is worth noting that these examples are not curated but selected from the first 12 images from the validation set.

has utilized pre-trained ImageNet [6] classification networks. These feature networks learn feature representations for the sole goal of classification; mapping an image to the top-1 class. Hence, they learn to ignore features that are irrelevant to the goal of classification. While this invariance benefits image classification, we find it to hurt discriminative representation for human synthesis.

We explore different feature networks (including variants of CNNS/ViTs) with widely different pre-training tasks for the discriminator. Specifically, Figure 3 ablate the following feature nets with the following pre-training tasks:

- **IN**: ImageNet Classification: ResNet50 (**RN50**), **ViT-B16** (DeIT variant), EfficientNet-Lite0 (**EN-L0**).
- **CLIP**: Contrastive Language Image Pre-training [48]: **RN50**, **ViT-B16**.
- **MAE**: Masked Autoencoders [15]: **ViT-B16**, **ViT-L16**.
- **CSE**: DensePose estimation [44]: ResNet50 (**RN50**).

We refer to each model as $architecture_{task}$, *e.g.* $\text{RN50}_{\text{CLIP}}$ refers to ResNet-50 with CLIP pre-trained weights. Directly selecting the best feature network from standard generative metrics ($\text{FID}/\text{FID}_{\text{CLIP}}$) is ambiguous, as projected GANs are known to achieve unnatural high scores on feature-based metrics [30]. We find that ImageNet models achieve unnatural high FID due to matching pre-training tasks, and ViT scores better on FID_{CLIP} due to matching architecture [4].

Independent of the architecture, we observe that all ImageNet [6] models generate highly corrupted faces, illustrated in Figure 4. This is most likely due to the invariance of facial descriptors in these feature networks, a phenomenon that has also been observed in [30]. Note that Appendix C includes comparison for all networks in Figure 3.

From the results in Figure 3, **Config C** replaces EfficientNet-Lite0 with $\text{ViT-L16}_{\text{MAE}}$ and $\text{RN50}_{\text{CLIP}}$. The motivation for pairing these networks is to exploit features with completely different architectures and pre-training tasks. In addition, these networks scores among the best w.r.t. OKS, FID_{CLIP}, and PPL.

[4] FID_{CLIP} is calculated from features of ViT-B/32 following [30].

Finally, $\text{RN50}_{\text{CLIP}}$ supplements ViT well, as RN50 operates on the original aspect ratio (288 × 160), whereas ViT is fixed to 224 × 224 [5].

3.4 Progressive Growing

Progressive training [25] is known to improve training stability of GANs and was recently re-introduced for unconditional synthesis with projected GANs [54]. StyleGAN-XL [54] first trains at 16 × 16 resolution, then increases the resolution by adding new layers to the end of the decoder. Note that StyleGAN-XL freezes already trained layers and the style network when training the next stage.

We adopt a straightforward extension to the image-to-image translation case, where we progressively train the U-net architecture by adding layers to the start/end of the encoder and decoder, respectively (see fig. 2). We observe that adding new blocks to the start of the encoder leads to training instability as it results in significant changes to the input of already-trained layers. To mitigate this, we introduce LayerScale [64] for each residual block with an initial value of 10^{-5} to lessen the contribution of new blocks. Furthermore, we include output skip connections following [27]. Unlike StyleGAN-XL, we avoid freezing any blocks during training as the computational benefit is minimal, given that we need to calculate gradients for layers at the beginning of the encoder. Introducing these changes substantially improves the final image quality (**Config D**)

We note that we experimented with more advanced techniques for progressive training, such as cascaded U-nets [17], or assymetric training of the encoder/decoder (*i.e.* start with a full-resolution encoder and a low-resolution decoder). However, we found that the straightforward progressive training technique was superior in terms of training time and final image quality.

[5] ViT input resolution is set to 224 × 224 for all models, as ViT features are less robust to changes in resolution from the training resolution.

Table 2: Quantitative comparison of SG-GAN [18] *vs*. ours.

Method	FID ↓	FID_{CLIP} ↓	PPL ↓	OKS ↑
SG-GAN [18]	1.97	1.25	70.2	0.950
TriA-GAN (ours)	**1.68**	**0.43**	**47.8**	**0.972**

3.5 Scaling Up the Generator

Config E double the number of residual blocks for each resolution in the encoder/decoder, resulting in 110.4M parameters in the generator compared to the previous 62.2M. This model trains stable up to 288 × 160 resolution, which is the maximum resolution of the FDH dataset.

4 Comparison to Surface-Guided GANs

Table 2 compares TriA-GAN to Surface Guided GANs (SG-GAN) [20] trained following DeepPrivacy2 [18], the current state-of-the-art for in-the-wild full-body synthesis. Figure 1 shows synthesis results with TriA-GAN, and Figure 5, Figure 6, compares TriA-GAN to SG-GAN. Appendix D include randomly selected samples.

The main difference between TriA-GAN and SG-GAN [18] is the improved training strategy of TriA-GAN, and the sparser conditional information (keypoints *vs*. dense surface correspondences). TriA-GAN improves at handling overlapping objects, partial bodies (*e.g*. intersection with image edges), and synthesis of texture (*e.g*. hair, clothing). Furthermore, TriA-GAN improves at context handling, *e.g*. inferring that an elderly lady is likely to sit at the table (top row, fig. 5), or that there is a motorcyclist on the bike (3rd row, fig. 5).

Finally, TriA-GAN is easier to use for downstream tasks, as our method does not rely on DensePose detections. For example, keypoints are easier to edit for interactive editing applications. Furthermore, detecting DensePose is challenging and unreliable for long-range detection, restricting its use in many scenarios (*e.g*. anonymizing pedestrians on the street). See Appendix D for examples of failure cases.

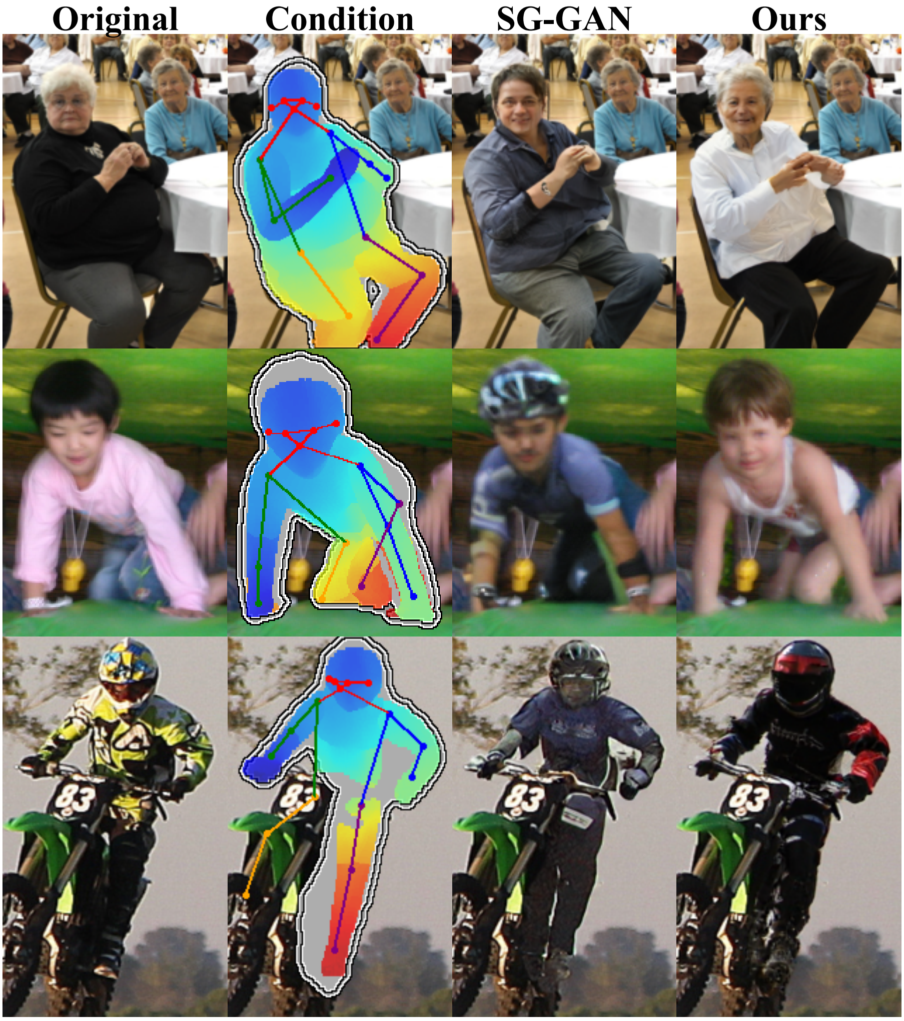


Figure 5: Curated examples comparing Surface Guided GAN [18] to TriA-GAN. Note that surface information is not used for TriA-GAN (shown in blue-yellow tint).

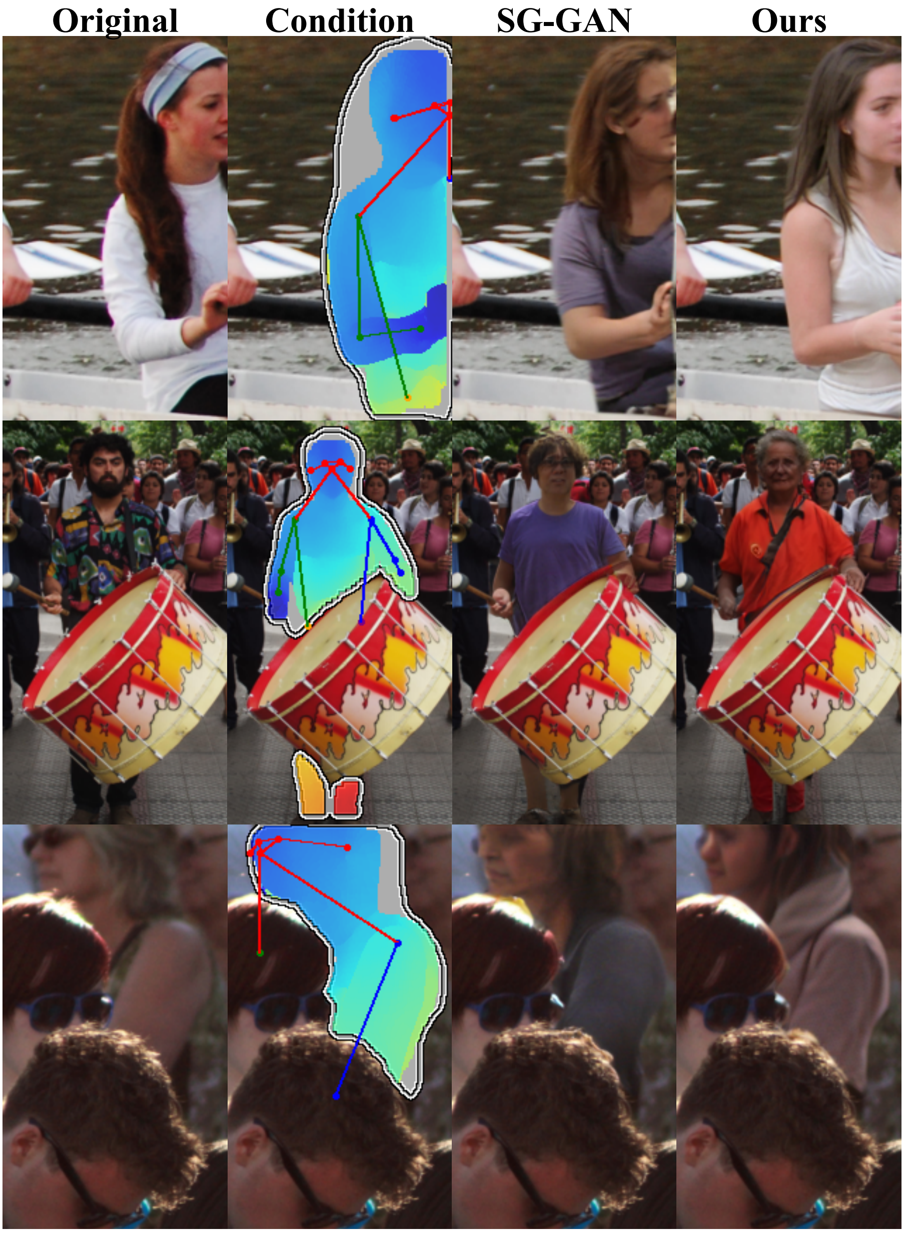


Figure 6: Curated examples comparing Surface Guided GAN [18] to TriA-GAN. Note that surface information is not used for TriA-GAN (shown in blue-yellow tint).

5 Editability of TriA-GAN

StyleGAN [27] is known for its disentangled latent space, and it is widely used for user-guided image editing, such as modifying images through text prompts [29]. However, most methods for editing images focus on unconditional GANs (or class-conditional GANs), and their application to image inpainting is less explored. StyleMC [29] is effective for editing faces with inpainting methods [18], but the same study finds editing human figures in-the-wild much harder [18]. We believe this limitation originates from the DensePose condition, where descriptive conditions can be correlated with specific attributes. This narrows the sampling probability, which makes it harder to find meaningful directions for randomly sampled images.

Figure 7 demonstrate that StyleMC [29] is effective with TriA-GAN to find semantically meaningful directions in the GAN latent space. StyleMC finds global directions by manipulating random images towards a text prompt using a CLIP encoder [48], where the directions are found over 1280 images. We find that StyleMC combined with TriA-GAN can edit a wide range of attributes, even quite specific attributes such as the size of the ears. However, we do note that editing some attributes results in changes to other correlated attributes. For example, the edit "blond hair" induces slight changes to the skin color. Furthermore, some attributes are more challenging to edit. For instance, introducing "red lips" to a body inferred as a male can result in significant semantic changes (top row, fig. 7). It is unclear whether this limitation is a result from the editing technique or TriA-GAN itself. We believe these correlations are inherent in the training datasets of CLIP or TriA-GAN.

6 Conclusion

TriA-GAN has enabled the generation of human figures in any desirable pose and location given a sparse set of keypoints, resulting in a new state-of-the-art for person synthesis on the FDH dataset. Key to our method is leveraging pretrained feature networks for the discriminator. We demonstrate that a carefully designed training strategy combined with feature networks suited to discriminate human figures substantially improves synthesis quality. TriA-GAN is the

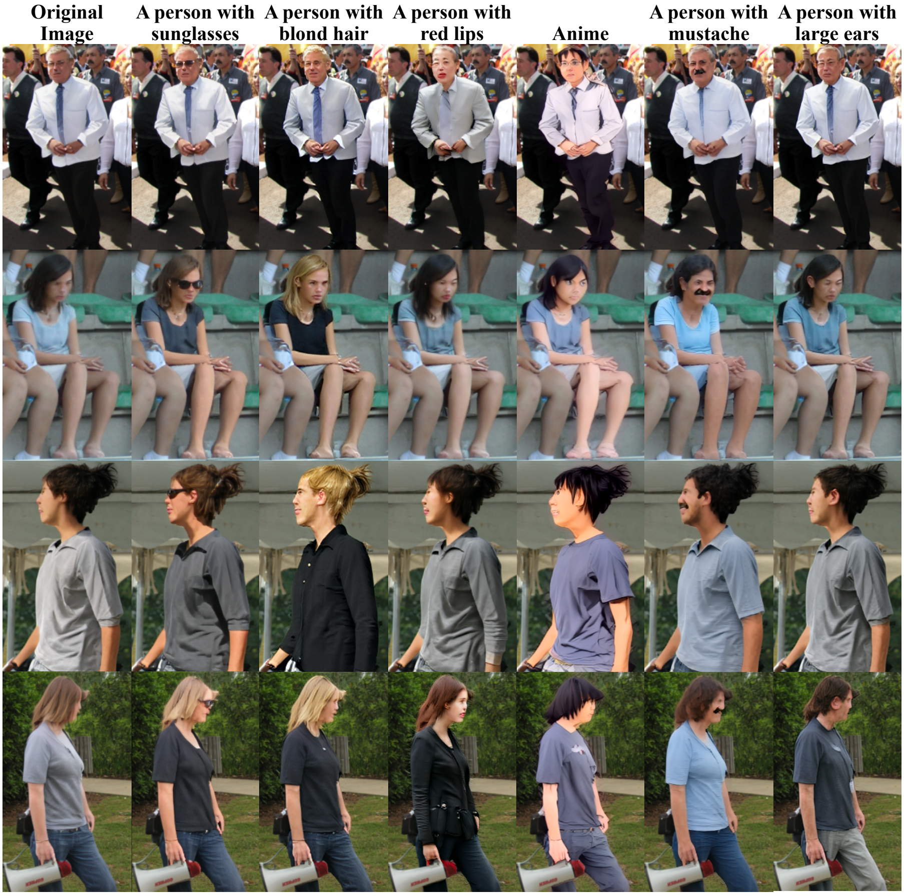


Figure 7: StyleMC [29] edits with TriA-GAN, where a global direction (from text prompt above each column) is added to the style code of the original (leftmost) image.

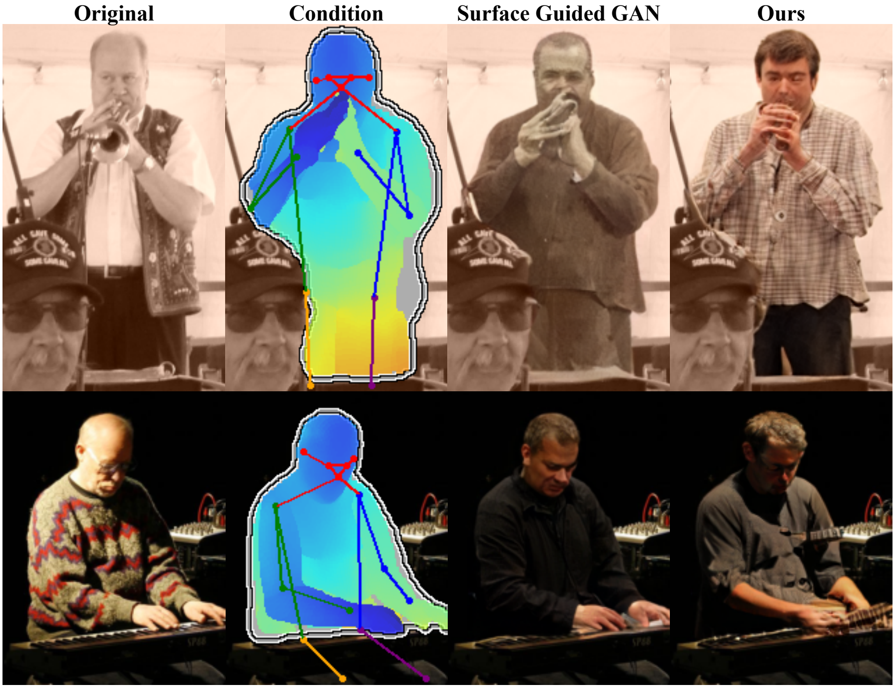


Figure 8: Failure cases of TriA-GAN.

first to demonstrate reliable attribute editing of human figures via text prompts, which we believe will be highly practical for many applications.

Societal Impact Synthesizing human figures has a range of useful applications everywhere, from content creation to anonymization purposes. However, similar to all learning-based generative models, the synthesized human figures adhere to the sampling probability of the dataset. In our case, the dataset originates from Flickr, which means that our generator follows its biases and is less likely to synthesize people from underrepresented groups on the website. Furthermore, our work focuses on generating lifelike humans, which carries the potential for abuse (*e.g*. DeepFakes). We note that the community has made a concerted effort to address this issue, through initiatives like the DeepFake Detection Challenge [7], or embedding watermarks into images from generative models [72].

6.1 Limitations

TriA-GAN sets a new state-of-the-art for human figure synthesis in-the-wild. Exploring methods for disentangling the latent space from the pose, body shape, and environment are exciting future avenues. Currently, the sampling space of TriA-GAN is highly dependent on the conditional information, where it can collapse into a single synthesized identity given certain conditions. Disentangled person image generation can mitigate this, by disentangle pose, appearance, and context. However, current methods require datasets with paired images [40, 51], which are less diverse and small.

The key limitation of TriA-GAN is handling more complex interactions with objects (fig. 8). This is particularly true for generating realistic hands/fingers, *e.g.* when playing the piano. SG-GAN [18] often improve on TriA-GAN in such scenarios if the DensePose information explicitly describes the interaction. But, it still struggles in cases where it is not clear (*e.g.* playing the masked-out trumpet).

TriA-GAN is hard to edit for attributes that are less frequent in the FDH dataset. For example, many images do not contain the lower body and attempting to find editing directions for "a person wearing red pants" results in editing other attributes as well. Whether this is a limitation to the editing method, or TriA-GAN is an open question.

Acknowledgement The computations were performed on the NTNU IDUN/EPIC computing cluster [58] and the Tensor-GPU project led by Prof. Anne C. Elster. Furthermore, we thank Rudolf Mester for his general support and thoughtful discussion.

References

[1] Martin Arjovsky, Soumith Chintala, and Léon Bottou. Wasserstein generative adversarial networks. In *International Conference on Machine Learning*, pages 214–223, 2017.

[2] Guha Balakrishnan, Amy Zhao, Adrian V Dalca, Fredo Durand, and John Guttag. Synthesizing Images of Humans in Unseen Poses. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8340–8348. IEEE, jun 2018.

[3] M. Bertalmio, G. Sapiro, V. Caselles, and C. Ballester. Image inpainting. In *Proceedings of the ACM SIGGRAPH Conference on Computer Graphics*, pages 417–424, 2000.

[4] Caroline Chan, Shiry Ginosar, Tinghui Zhou, and Alexei Efros. Everybody Dance Now. In *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, volume 49, pages 5932–5941. IEEE, oct 2019.

[5] Bindita Chaudhuri, Nikolaos Sarafianos, Linda Shapiro, and Tony Tung. Semi-supervised Synthesis of High-Resolution Editable Textures for 3D Humans. In *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pages 7987–7996, 2021.

[6] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. ImageNet: A large-scale hierarchical image database. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pages 248–255. IEEE, jun 2009.

[7] Brian Dolhansky, Joanna Bitton, Ben Pflaum, Jikuo Lu, Russ Howes, Menglin Wang, and Cristian Canton Ferrer. The DeepFake Detection Challenge (DFDC) Dataset. *arXiv preprint arXiv:2006.07397*, 2020.

[8] Patrick Esser and Ekaterina Sutter. A Variational U-Net for Conditional Appearance and Shape Generation. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8857–8866. IEEE, jun 2018.

[9] Anna Fruhstuck, Krishna Kumar Singh, Eli Shechtman, Niloy J Mitra, Peter Wonka, and Jingwan Lu. InsetGAN for Full-Body Image Generation. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 7713–7722. IEEE, jun 2022.

[10] Jianglin Fu, Shikai Li, Yuming Jiang, Kwan-Yee Lin, Chen Qian, Chen Change Loy, Wayne Wu, and Ziwei Liu. StyleGAN-Human: A Data-Centric Odyssey of Human Generation. In *Computer Vision - ECCV 2022*, volume 13676, pages 1–19. Springer, Cham, 2022.

[11] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In *Advances in neural information processing systems*, pages 2672–2680, 2014.

[12] Riza Alp Guler, Natalia Neverova, and Iasonas Kokkinos. DensePose: Dense Human Pose Estimation in the Wild. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7297–7306. IEEE, jun 2018.

[13] Ishaan Gulrajani, Faruk Ahmed, Martin Arjovsky, Vincent Dumoulin, and Aaron C Courville. Improved training of wasserstein gans. In *Advances in Neural Information Processing Systems*, pages 5767–5777, 2017.

[14] Xintong Han, Zuxuan Wu, Zhe Wu, Ruichi Yu, and Larry S Davis. VITON: An Image-Based Virtual Try-on Network. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7543–7552. IEEE, jun 2018.

[15] Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollar, and Ross Girshick. Masked Autoencoders Are Scalable Vision Learners. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, volume 2022-June, pages 15979–15988. IEEE, jun 2022.

[16] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. In *Advances in Neural Information Processing Systems*, pages 6626–6637, 2017.

[17] Jonathan Ho, Chitwan Saharia, William Chan, David J. Fleet, Mohammad Norouzi, and Tim Salimans. Cascaded Diffusion Models for High Fidelity Image Generation. *Journal of Machine Learning Research*, 23:1–33, 2022.

[18] Håkon Hukkelås and Frank Lindseth. DeepPrivacy2: Towards Realistic Full-Body Anonymization. In *2023 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pages 1329–1338. IEEE, jan 2023.

[19] Håkon Hukkelås, Frank Lindseth, and Rudolf Mester. Image Inpainting with Learnable Feature Imputation. In *DAGM German Conference on Pattern Recognition*, pages 388–403. Springer-Verlag, 2021.

[20] Håkon Hukkelås, Morten Smebye, Rudolf Mester, and Frank Lindseth. Realistic Full-Body Anonymization with Surface-Guided GANs. In *2023 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pages 1430–1440. IEEE, jan 2023.

[21] Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A Efros. Image-to-Image Translation with Conditional Adversarial Networks. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5967–5976. IEEE, jul 2017.

[22] Yuming Jiang, Shuai Yang, Haonan Qiu, Wayne Wu, Chen Change Loy, and Ziwei Liu. Text2human: Text-driven controllable human image generation. *ACM Transactions on Graphics*, 41(4):1–11, jul 2022.

[23] Youngjoo Jo and Jongyoul Park. SC-FEGAN: Face Editing Generative Adversarial Network With User's Sketch and Color. In *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 1745–1753. IEEE, oct 2019.

[24] Animesh Karnewar and Oliver Wang. MSG-GAN: Multi-Scale Gradients for Generative Adversarial Networks. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 7796–7805. IEEE, jun 2020.

[25] Tero Karras, Timo Aila, Samuli Laine, and Jaakko Lehtinen. Progressive growing of gans for improved quality, stability, and variation. In *International Conference on Learning Representations*, 2018.

[26] Tero Karras, Miika Aittala, Samuli Laine, Erik Härkönen, Janne Hellsten, Jaakko Lehtinen, and Timo Aila. Alias-Free Generative Adversarial Networks. In *Advances in Neural Information Processing Systems*, volume 34, pages 852–863, 2021.

[27] Tero Karras, Samuli Laine, and Timo Aila. A Style-Based Generator Architecture for Generative Adversarial Networks. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4396–4405. IEEE, jun 2019.

[28] Tero Karras, Samuli Laine, Miika Aittala, Janne Hellsten, Jaakko Lehtinen, and Timo Aila. Analyzing and Improving the Image Quality of StyleGAN. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 8107–8116. IEEE, jun 2020.

[29] Umut Kocasari, Alara Dirik, Mert Tiftikci, and Pinar Yanardag. StyleMC: Multi-Channel Based Fast Text-Guided Image Generation and Manipulation. In *2022 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pages 3441–3450. IEEE, jan 2022.

[30] Tuomas Kynkäänniemi, Tero Karras, Miika Aittala, Timo Aila, and Jaakko Lehtinen. The Role of ImageNet Classes in Frechet Inception Distance. *arXiv preprint arXiv:2203.06026*, 2022.

[31] Avisek Lahiri, Arnav Kumar Jain, Sanskar Agrawal, Pabitra Mitra, and Prabir Kumar Biswas. Prior Guided GAN Based Semantic Inpainting. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, jun 2020.

[32] Christoph Lassner, Gerard Pons-Moll, and Peter V. Gehler. A Generative Model of People in Clothing. In *2017 IEEE International Conference on Computer Vision (ICCV)*, volume 2017-Octob, pages 853–862. IEEE, oct 2017.

[33] Yining Li, Chen Huang, and Chen Change Loy. Dense Intrinsic Appearance Flow for Human Pose Transfer. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3688–3697. IEEE, jun 2019.

[34] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. Microsoft COCO: Common Objects in Context. In *European conference on computer vision*, volume 8693 LNCS, pages 740–755. Springer, Cham, 2014.

[35] Guilin Liu, Fitsum A Reda, Kevin J Shih, Ting-Chun Wang, Andrew Tao, and Bryan Catanzaro. Image inpainting for irregular holes using partial convolutions. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 85–100, 2018.

[36] Hongyu Liu, Bin Jiang, Yi Xiao, and Chao Yang. Coherent Semantic Attention for Image Inpainting. In *2019 IEEE/CVF International*

Conference on Computer Vision (ICCV), pages 4169–4178. IEEE, oct 2019.

[37] Ziwei Liu, Ping Luo, Shi Qiu, Xiaogang Wang, and Xiaoou Tang. DeepFashion: Powering Robust Clothes Recognition and Retrieval with Rich Annotations. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1096–1104. IEEE, jun 2016.

[38] Ziwei Liu, Ping Luo, Xiaogang Wang, and Xiaoou Tang. Deep Learning Face Attributes in the Wild. In *2015 IEEE International Conference on Computer Vision (ICCV)*, pages 3730–3738. IEEE, dec 2015.

[39] Liqian Ma, Xu Jia, Qianru Sun, Bernt Schiele, Tinne Tuytelaars, and Luc Van Gool. Pose guided person image generation. In *Advances in Neural Information Processing Systems*, volume 2017-Decem, pages 406–416, 2017.

[40] Liqian Ma, Qianru Sun, Stamatios Georgoulis, Luc Van Gool, Bernt Schiele, and Mario Fritz. Disentangled Person Image Generation. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 99–108. IEEE, jun 2018.

[41] Takeru Miyato, Toshiki Kataoka, Masanori Koyama, and Yuichi Yoshida. Spectral normalization for generative adversarial networks. *International Conference on Learning Representations*, 2018.

[42] Kamyar Nazeri, Eric Ng, Tony Joseph, Faisal Qureshi, and Mehran Ebrahimi. EdgeConnect: Structure Guided Image Inpainting using Edge Prediction. In *2019 IEEE/CVF International Conference on Computer Vision Workshop (ICCVW)*, pages 3265–3274. IEEE, oct 2019.

[43] Natalia Neverova, Riza Alp Güler, and Iasonas Kokkinos. Dense Pose Transfer. In *European conference on computer vision*, volume 11207 LNCS, pages 128–143, 2018.

[44] Natalia Neverova, David Novotny, Vasil Khalidov, Marc Szafraniec, Patrick Labatut, and Andrea Vedaldi. Continuous Surface Embeddings. In *Advances in Neural Information Processing Systems*, volume 33, pages 17258–17270. Curran Associates, Inc., nov 2020.

[45] Augustus Odena, Christopher Olah, and Jonathon Shlens. Conditional image synthesis with auxiliary classifier gans. In *34th International Conference on Machine Learning (ICML)*, volume 6, pages 4043–4055, 2017.

[46] Deepak Pathak, Philipp Krahenbuhl, Jeff Donahue, Trevor Darrell, and Alexei A Efros. Context Encoders: Feature Learning by Inpainting. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2536–2544. IEEE, jun 2016.

[47] Albert Pumarola, Antonio Agudo, Alberto Sanfeliu, and Francesc Moreno-Noguer. Unsupervised Person Image Synthesis in Arbitrary Poses. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8620–8628. IEEE, jun 2018.

[48] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning Transferable Visual Models From Natural Language Supervision. In *International Conference on Machine Learning*, pages 8748–8763, 2021.

[49] Yurui Ren, Xiaoming Yu, Ruonan Zhang, Thomas H Li, Shan Liu, and Ge Li. StructureFlow: Image Inpainting via Structure-Aware Appearance Flow. In *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 181–190. IEEE, oct 2019.

[50] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-Net: Convolutional Networks for Biomedical Image Segmentation. In *International Conference on Medical image computing and computer-assisted intervention*, pages 234–241. Springer, 2015.

[51] Kripasindhu Sarkar, Vladislav Golyanik, Lingjie Liu, and Christian Theobalt. Style and Pose Control for Image Synthesis of Humans from a Single Monocular View. *arXiv preprint arXiv:2102.11263*, 2021.

[52] Kripasindhu Sarkar, Dushyant Mehta, Weipeng Xu, Vladislav Golyanik, and Christian Theobalt. Neural Re-rendering of Humans from a Single Image. In *European conference on computer vision*, volume 12356 LNCS, pages 596–613. Springer Science and Business Media Deutschland GmbH, 2020.

[53] Axel Sauer, Kashyap Chitta, Jens Müller, and Andreas Geiger. Projected GANs Converge Faster. In *Advances in Neural Information Processing Systems*, pages 17480–17492, 2021.

[54] Axel Sauer, Katja Schwarz, and Andreas Geiger. StyleGAN-XL: Scaling StyleGAN to Large Diverse Datasets. In *ACM SIGGRAPH 2022 Conference Proceedings*, pages 1–10, Vancouver, BC, Canada, aug 2022. Association for Computing Machinery.

[55] Edgar Schonfeld, Bernt Schiele, and Anna Khoreva. A U-Net Based Discriminator for Generative Adversarial Networks. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 8204–8213. IEEE, jun 2020.

[56] Chenyang Si, Wei Wang, Liang Wang, and Tieniu Tan. Multistage Adversarial Losses for Pose-Based Human Image Synthesis. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 118–126. IEEE, jun 2018.

[57] Aliaksandr Siarohin, Enver Sangineto, Stephane Lathuiliere, and Nicu Sebe. Deformable GANs for Pose-Based Human Image Generation. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3408–3416. IEEE, jun 2018.

[58] Magnus Själander, Magnus Jahre, Gunnar Tufte, and Nico Reissmann. EPIC: An energy-efficient, high-performance GPGPU computing research infrastructure, 2019.

[59] Jascha Sohl-Dickstein, Eric Weiss, Niru Maheswaranathan, and Surya Ganguli. Deep unsupervised learning using nonequilibrium thermodynamics. In *International Conference on Machine Learning*, pages 2256–2265. PMLR, 2015.

[60] Sijie Song, Wei Zhang, Jiaying Liu, Zongming Guo, and Tao Mei. Unpaired Person Image Generation With Semantic Parsing Transformation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 43(11):4161–4176, nov 2021.

[61] Vadim Sushko, Edgar Schönfeld, Dan Zhang, Juergen Gall, Bernt Schiele, and Anna Khoreva. OASIS: Only Adversarial Supervision for Semantic Image Synthesis. *International Journal of Computer Vision*, 130(12):2903–2923, dec 2022.

[62] Roman Suvorov, Elizaveta Logacheva, Anton Mashikhin, Anastasia Remizova, Arsenii Ashukha, Aleksei Silvestrov, Naejin Kong, Harshith Goka, Kiwoong Park, and Victor Lempitsky. Resolution-robust Large Mask Inpainting with Fourier Convolutions. In *2022 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pages 3172–3182. IEEE, jan 2022.

[63] Mingxing Tan and Quoc V Le. EfficientNet: Rethinking Model Scaling for Convolutional Neural Networks. In *International conference on machine learning*, pages 6105–6114, 2019.

[64] Hugo Touvron, Matthieu Cord, Alexandre Sablayrolles, Gabriel Synnaeve, and Herve Jegou. Going deeper with Image Transformers. In *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 32–42. IEEE, oct 2021.

[65] Wei Xiong, Jiahui Yu, Zhe Lin, Jimei Yang, Xin Lu, Connelly Barnes, and Jiebo Luo. Foreground-Aware Image Inpainting. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5833–5841. IEEE, jun 2019.

[66] Yufei Xu, Jing Zhang, Qiming Zhang, and Dacheng Tao. ViTPose: Simple Vision Transformer Baselines for Human Pose Estimation. *arXiv preprint arXiv:2204.12484*, 2022.

[67] Chaojie Yang, Hanhui Li, Shengjie Wu, Shengkai Zhang, Haonan Yan, Nianhong Jiao, Jie Tang, Runnan Zhou, Xiaodan Liang, and Tianxiang Zheng. BodyGAN: General-purpose Controllable Neural Human Body Generation. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 7723–7732. IEEE, jun 2022.

[68] Zhuoqian Yang, Shikai Li, Wayne Wu, and Bo Dai. 3DHumanGAN: Towards Photo-Realistic 3D-Aware Human Image Generation. *arXiv preprint arXiv:2212.07378*, 2022.

[69] Zili Yi, Qiang Tang, Shekoofeh Azizi, Daesik Jang, and Zhan Xu. Contextual Residual Aggregation for Ultra High-Resolution Image Inpainting. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 7505–7514. IEEE, jun 2020.

[70] Jiahui Yu, Zhe Lin, Jimei Yang, Xiaohui Shen, Xin Lu, and Thomas Huang. Free-Form Image Inpainting With Gated Convolution. In *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 4470–4479. IEEE, oct 2019.

[71] Jiahui Yu, Zhe Lin, Jimei Yang, Xiaohui Shen, Xin Lu, and Thomas S. Huang. Generative Image Inpainting with Contextual Attention. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5505–5514. IEEE, jun 2018.

[72] Ning Yu, Vladislav Skripniuk, Sahar Abdelnabi, and Mario Fritz. Artificial Fingerprinting for Generative Models: Rooting Deepfake Attribution in Training Data. In *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 14428–14437. IEEE, oct 2021.

[73] Tao Yu, Zongyu Guo, Xin Jin, Shilin Wu, Zhibo Chen, Weiping Li, Zhizheng Zhang, and Sen Liu. Region Normalization for Image Inpainting. *Proceedings of the AAAI Conference on Artificial Intelligence*, 34(07):12733–12740, apr 2020.

[74] Yu Zeng, Zhe Lin, Jimei Yang, Jianming Zhang, Eli Shechtman, and Huchuan Lu. High-Resolution Image Inpainting with Iterative Confidence Feedback and Guided Upsampling. In *European conference on computer vision*, pages 1–17. Springer-Verlag, 2020.

[75] Liang Zheng, Liyue Shen, Lu Tian, Shengjin Wang, Jingdong Wang, and Qi Tian. Scalable Person Re-identification: A Benchmark. In *2015 IEEE International Conference on Computer Vision (ICCV)*, pages 1116–1124. IEEE, dec 2015.

Paper F

Does Image Anonymization Impact Computer Vision Training?

Authors:

Håkon Hukkelås, Frank Lindseth

Published at conference:

2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Workshop on Autonomous Driving

Source Code:

https://github.com/hukkelas/deep_privacy2/blob/master/docs/anonymizing_datasets.md

Appendix:

https://openaccess.thecvf.com/content/CVPR2023W/WAD/supplemental/Hukkelas_Does_Image_Anonymization_CVPRW_2023_supplemental.pdf

Paper F

Does Image Anonymization Impact Computer Vision Training?

Håkon Hukkelås Frank Lindseth

Norwegian University of Science and Technology
Trondheim, Norway
hakon.hukkelas@ntnu.no

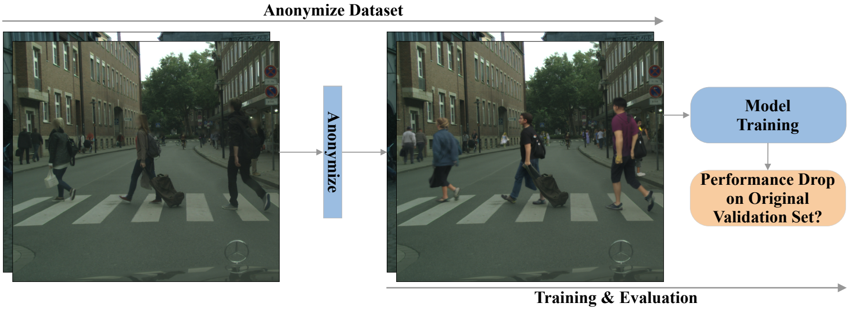


Figure 1: To assess the impact of anonymization, we first anonymize common computer vision datasets, then train various models using the anonymized data, and finally evaluate the models on the original validation datasets. The figure depicts our Cityscapes [8] full-body anonymization experiment. Note that the leftmost image is anonymized with face blurring, following Cityscapes [8] terms of use.

Abstract

Image anonymization is widely adapted in practice to comply with privacy regulations in many regions. However, anonymization often degrades the quality of the data, reducing its utility for computer vision development. In this paper, we investigate the impact of image anonymization for training computer vision models on key computer vision tasks (detection, instance segmentation, and pose estimation). Specifically, we benchmark the recognition drop on common detection datasets, where we evaluate both traditional and realistic anonymization for faces and full bodies. Our comprehensive experiments reflect that traditional image anonymization substantially impacts final model performance, particularly when anonymizing the full body. Furthermore, we find that realistic anonymization can mitigate this decrease in performance, where our experiments reflect a minimal performance drop for face anonymization. Our study demonstrates that realistic anonymization can enable privacy-preserving computer vision development with minimal performance degradation across a range of important computer vision benchmarks.

1 Introduction

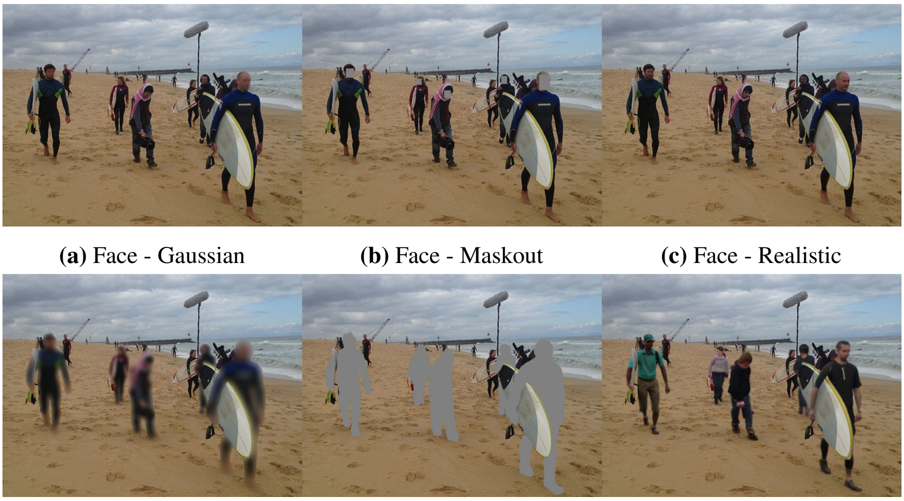
Collecting and storing large amounts of visual data is a fundamental task in developing robust and efficient computer vision algorithms. However, this raises concerns regarding the individual's right to privacy, as visual data is rich in privacy-sensitive information, *e.g.* persons, license plates, and street signs. Recent privacy legislation (*e.g.* GDPR [9] in the European Union) requires anonymization when collecting visual data or consent from individuals, which is often infeasible. This can be viewed as a barrier to research and development, particularly for the data-dependent field of Autonomous Vehicle (AV) research. To compensate for these restrictions, practitioners have adopted traditional image anonymization (*e.g.* blurring) for collecting AV datasets [15, 6] and street view images [12].

Traditional image anonymization can protect privacy, but it severely distorts the visual data, potentially reducing its utility for computer vision development. Despite this, face obfuscation (*e.g.* blurring) is the standard method employed to anonymize public autonomous vehicle datasets [15, 6], and its impact on final model performance is currently unclear. Previous work analyzed the impact of face anonymization for classification [59], semantic segmentation [15, 63], object detection [11], action recognition [54], and face detection [30]. In summary, their findings reveal that face anonymization can impact visual recognition related to the human class, and it can severely hurt tasks where the human is in focus [30, 54].

Our literature review, detailed in section 2, resulted in two unanswered questions, which we address in this study.

First, *is realistic anonymization more effective to preserve image utility compared to traditional methods?* Realistic anonymization replaces privacy-sensitive information with synthesized content from generative models, which are found to better preserve utility compared to traditional methods [52, 25]. Previous work has found realistic anonymization to improve utility preservation for semantic segmentation [30, 63]. Our work builds upon this by investigating different objectives and datasets.

Secondly, *to what extent does full-body anonymization impact the training of computer vision models?* The human body is recognizable from many cues



(a) Face - Gaussian **(b)** Face - Maskout **(c)** Face - Realistic

(d) Body - Gaussian **(e)** Body - Mask out **(f)** Body - Realistic

Figure 2: The different anonymization methods evaluated in this paper. Image from COCO train2017 [37], image id=000000097507.

outside the face (*e.g.* gait, clothes, ear, body shape), often requiring full-body anonymization to protect privacy. A few studies explore the impact of full-body anonymization [26, 23], where they find it to improve over traditional methods. However, they rely on automatic detection methods, which opens the question if the performance degradation is due to detection errors or the anonymization model. Furthermore, their model requires dense pose estimation [18, 43], which limits anonymization to individuals close to the camera due to limited long-range detection recall of dense pose models.

In this paper, we focus on key computer vision tasks related to autonomous vehicles, namely instance segmentation and human pose estimation. We evaluate the full-body and face anonymization models built in DeepPrivacy2 [23] and compare realistic anonymization to traditional methods. See `https://github.com/hukkelas/deep_privacy2/blob/master/docs/anonymizing_datasets.md` to reproduce our experiments.

2 Related Work

Image Anonymization The goal of image anonymization is to remove any privacy-sensitive information contained in the image. Traditional anonymization is widely adopted in practice, where methods anonymize the image via obfuscation (*e.g.* blurring, masking), encryption [20], or k-means [17, 28, 44]. Often, these methods are sufficient to protect privacy; however, they degrade the quality of the data reducing its utility for downstream tasks.

Recent work has introduced *realistic image anonymization*, where anonymization is done by replacing persons with synthesized identities from a generative model. The majority of previous work focuses on face anonymization, where current methods anonymize by *inpainting* a masked out region [25, 38, 52, 53], or *transforming* [13, 50, 7] the original identity to remove privacy-sensitive information. Transformative models often maintain higher utility (*e.g.* preserving facial expression) but offer no formal guarantee of removing the original identity from the image, making them vulnerable to adversarial attacks. A few methods explore anonymizing the full-body [23, 26, 4, 38], where the current state-of-the-art [23, 24] can generate convincing full-bodies given sparse keypoints [24] or dense pose annotations [23]. Finally, some methods insert adversarial perturbation in the image, which is invisible to the human eye but able to fool face recognition systems [46].

Privacy Guarantees of Anonymization Most current anonymization systems offer no formal guarantee of anonymization, and the identity can often be recognized from other cues in the image. Image blurring is discussed numerous times in the literature [35, 36, 3, 42, 16, 44], where the identity is often recognizable due to limited blurring. Furthermore, the identity is recognizable even though the face is anonymized through other identifying attributes of the human body [56, 32, 39], such as gait [27], clothing [14], and body appearance [62, 45]. This makes full-body anonymization more effective than face anonymization in terms of privacy. Finally, most anonymization systems rely on automatic detection, which is far from perfect and vulnerable to adversarial attacks [31].

Public Anonymized Datasets The prominent computer vision datasets employ no form of anonymization, where only a few datasets are anonymized. NuScenes [6] contains images from vehicles driving in Singapore and Boston, where faces and license plates are anonymized via blurring. A2D2 [15] includes data from southern Germany, where license plates and heads are blurred to comply with German privacy regulations. AViD [48] is a video dataset for action recognition with blurred heads. P3M [33] is a portrait matting dataset where every face is blurred. [55] propose a dataset containing street view scenes where cars and pedestrians are removed via image inpainting.

Visual Recognition on Anonymized Data There exists a limited set of studies exploring the effect that anonymization has on training computer vision models. For ImageNet [10] training, face obfuscation (blurring) has little effect on top-5 accuracy and no impact on feature transferability to scene recognition, object localization, and face attribute classification. Nevertheless, anonymization slightly degrades accuracy in classes appearing together with faces (*e.g.* facial masks). For autonomous vehicle datasets, traditional face anonymization can degrade instance segmentation on Cityscapes [8, 63], whereas realistic face anonymization has no noticeable negative impact. Furthermore, they find that larger backbones and multi-scale features are more robust to image anonymization [63]. Dvoracek *et al.* [11] finds little impact of face anonymization on object detection on the same dataset. Geyer [15] finds that face anonymization has little effect on semantic segmentation on the A2D2 dataset. For face detection, realistic anonymization performs substantially better than traditional methods for training face detectors [30]. For action recognition, face obfuscation significantly degrades performance [54], where the authors propose a teacher-student self-distillation framework to mitigate the degradation.

Finally, we note that some studies focus on the human perspective and investigate the effect of different anonymization techniques on the users' perceived experience [19, 35].

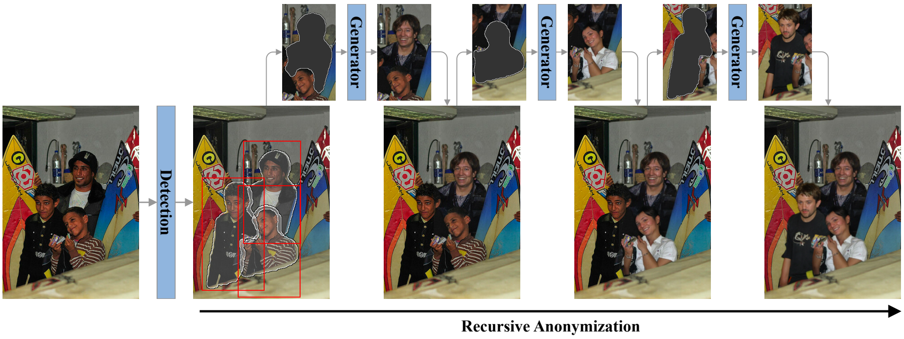


Figure 3: DeepPrivacy2 [23] anonymizes one instance at a time, then paste each synthesized individual into the original image. For our experiments, detection is not performed, as segmentation masks are defined from pre-defined annotations (see section 3.1). Note that the generator relies on keypoint annotations, which are not depicted here.

3 Anonymization Method

In this paper, we explore three different anonymization techniques for full-body and face anonymization; blurring, mask-out, and realistic anonymization (see fig. 2). Given the image I and a mask M indicating the region to be anonymized, the goal of each method is to remove any privacy-sensitive information within M. In this section, we first define M for face and full-body anonymization (Section 3.1), then introduce the anonymization methods in Section 3.2 and Section 3.3.

3.1 Anonymization Region

To define the anonymization region, we employ the pre-defined instance segmentation annotations for the person/pedestrian class, as every dataset in this paper includes such annotations. Note that we do not anonymize annotations marked as "crowd" or "ignored" in the datasets, nor classes that often contain a person (*e.g.* bicycle, motorcycle), as the realistic anonymization techniques require distinct instance-wise annotations. Given the two aforementioned filtering criteria, it is important to note that we are not able to anonymize all individuals in the dataset. An alternative option is to obtain instance-wise annotations by manual annotation or automatic detection. However, we decided against this approach, as the former is too time-consuming, and the latter may introduce detection errors, making it unclear if performance degradation is due to detection errors or poor anonymization.

Face Region As none of the benchmark datasets include annotated faces, we define the face anonymization region following a standard face detection dataset, WIDER-Face [60]. Specifically, the region is the minimal bounding box containing the forehead, chin, and cheek. We annotate each dataset with a pre-trained face detector (DSFD [34]), where we filter the detections by matching them with annotated instance segmentations. We match boxes to segmentations via Intersection over Union (IoU), where we select the match with the highest IoU and bounding box score. Any matches with an IoU $< 1\%$ are removed.

Full-Body Anonymization Since all benchmark datasets include annotated instance segmentations, we use these to define the full-body anonymization region. To compensate for annotations where the segmentations don't fully encompass the body (often segmentation does not include bordering pixels), we slightly dilate the segmentation following [23].

3.2 Traditional Anonymization

We evaluate two commonly used obfuscation techniques for traditional anonymization, namely blurring and masking out. Note that we employ the same method for both face and full-body anonymization.

Mask-Out Mask-out defines the anonymized image as $I_{new} = I \odot (1 - M) + M \odot 127$, where $\odot$ is element-wise multiplication.

Gaussian Blur Gaussian blur defines the anonymized image as $I_{new} = I \odot (1 - M) + M \odot I_{blur}$. Here, I_{blur} is the blurred image with a Gaussian filter ($\sigma = 7$, k-size= $3 \cdot \sigma$).

3.3 Realistic Anonymization

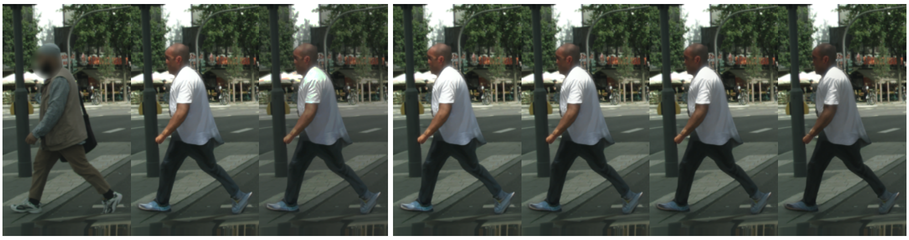
For realistic anonymization, we employ pre-trained models from DeepPrivacy2 [23]. Note that DeepPrivacy2 anonymizes by inpainting (illustrated in fig. 3), such that it never observes the masked region in I. Thus, it provides similar privacy protection as mask-out anonymization.

Face Anonymization For face anonymization, we employ the face anonymization model in DeepPrivacy2 [23], which is a U-Net GAN trained on FDF [25] that synthesizes faces at 128 × 128 resolution. This model does not rely on keypoint annotations, which enables it to anonymize all faces detected.

Full-Body Anonymization For full-body anonymization, we employ a U-Net GAN [24] relying on keypoint annotations following the COCO format [37]. This model is trained on the FDH dataset [23], and the model is integrated into the DeepPrivacy2 framework [23]. For datasets without keypoint annotations, we use a top-down pose estimation network (ViTPose [58]) which estimates the pose given the image and the minimal bounding box encompassing the instance segmentation. All keypoints with a confidence $\geq 30\%$ are assumed to be visible.

3.4 Global Context for Full-Body Synthesis

In our preliminary experiments, we observed that the full-body generative model often generated human bodies that fit the local context of the generative model but did not align with the global context. We believe this is not a limitation of the generative model itself but a limitation to the crop-based anonymization method used by DeepPrivacy2 (see fig. 3). In this paper, we explore two solutions to this issue; ad-hoc histogram equalization and histogram matching via latent optimization illustrated in fig. 4



Original Initial ω HM HM-LO Optimization → Final Image

Figure 4: The initial synthesized identity ("initial ω") may not align with the global context of the image, making the synthesized identity "stick out" compared to the original identity. We explore two options to address this issue: naive histogram matching (**HM**), and Histogram matching via latent optimization (**HM-LO**), which iteratively adjusts the initial ω to better fit the histogram of the original image (in HSV)

Histogram Matching (HM) A naive approach for matching the generated body to the global context is naive histogram equalization. Specifically, we match the synthesized (cropped) image to the original (cropped) image by using skimage match_histogram. This adjusts the synthesized image such that each color channel (RGB) matches the cumulative histogram of the original image. To reduce bordering effects when pasting the equalized image into the original image, we smoothly transition the border by slightly blurring the mask with a gaussian filter. That is, given the cropped image x, the corresponding mask M_c, and the synthesized image y, the new image is given by; $y_{new} = x \odot (1 - M_c^{blurred}) + y \odot M_c^{blurred}$, where $M_c^{blurred}$ is M_c blurred with a gaussian filter with size=$[19, 19]$ and $\sigma = 9$. We note that this is far from an optimal solution, where naive histogram matching can introduce severe visual artifacts fig. 5.

Histogram Matching via Latent Optimization (HM-LO) An alternative approach to post-processing the output is a search in the latent space of the generator. Conceptually, if the exact environmental context (*e.g.* scene lightning) is not given by the cropped image, it should be possible to adjust such factors through the latent space of the generator. Therefore, we suggest utilizing gradient descent to modify the latent vector of the generator, aligning the histogram of the generated image with that of the original image

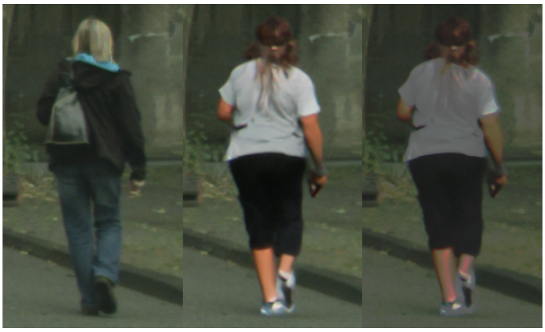

Original Anonymized Final after HM

Figure 5: Naive histogram matching can introduce visual artifacts.

Given the cropped image x and the mask M_c, the generated image is $y = G(x \odot M_C, \omega)$, where ω is the latent space of the generator, following StyleGAN [29]. Given x, we adjust a sampled ω via gradient descent such that y matches the histogram of x in the S and V channel of the HSV transform of x and y. Specifically, we optimize;

$$\begin{aligned}\mathscr{L}(x_{hsv}, y_{hsv}) = \mathbb{W}(P_S(x_{hsv}), P_S(y_{hsv})) + \\ \mathbb{W}(P_V(x_{hsv}), P_V(y_{hsv})),\end{aligned} \tag{1}$$

where $\mathbb{W}$ is the Wasserstein-1 distance, and P_V, P_S is the histogram of the S and V color channel in the HSV transformed image of x and y. Then, we perform gradient descent on ω for 100 steps or until $\mathscr{L}(x_{hsv}, y_{hsv}) < 0.02$.

Often, HM-LO induces slight adjustments to the generated image such that it better matches the context of the image (fig. 4). However, we note that HM-LO can induce significant semantic changes if the original sampled colors deviate from the original identity (fig. 6).

4 Experiments

In this section, we report results for training on anonymized data. We train each model on the anonymized dataset and report standard evaluation metrics on the original validation set. To reduce randomness, we report the average and

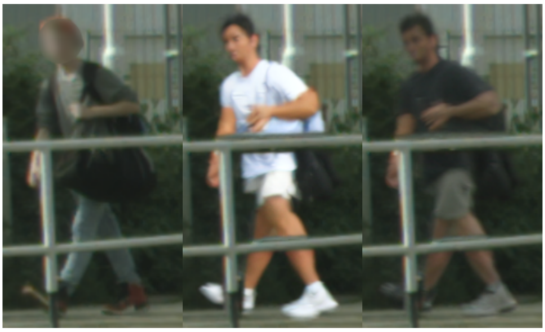


Figure 6: Histogram Matching via Latent Optimization can induce significant semantic changes to the synthesized identity, due to directly optimizing ω to match the HSV histogram (S/V channels).

standard error over three independent training runs using seeds 0, 1, and 2. All experiments are done with Pytorch 1.12 [47] on a single NVIDIA A100-40GB. Random qualitative examples from our experiments are given in Appendix B.

4.1 Experimental Details

COCO Pose Estimation We train a Keypoint R-50 FPN R-CNN using detectron2 [57] on the COCO2017 dataset [37]. The training dataset contains 118,287 images with 149,813 person instances (after filtration following section 3), and we evaluate on the original validation dataset (5K images). Out of 149,813 instances, 95,295 are detected by the face detector. Detectron2 is run with commit: 58e472e076

Cityscapes Instance Segmentation We train Mask R-CNN [21] R-50 FPN using detectron2 [57] on the Cityscapes dataset [8]. The training dataset contains 2,975 images with 17,919 person instances (after filtration following section 3), and we evaluate on the original validation dataset (500 images). Out of 17,919 instances, 4,456 were detected by the face detector. Interestingly, this is a noticeably smaller percentage than for the COCO dataset, which we

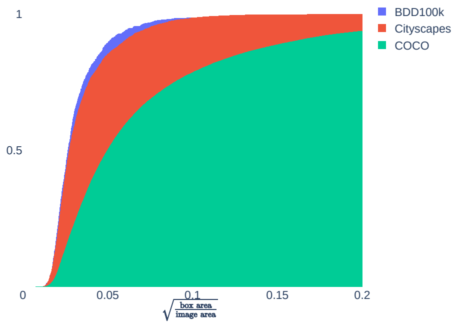


Figure 7: Cumulative histogram of average bounding box length normalized to image size.

speculate is due to the dataset distribution (persons in COCO often face the camera, while they often do not in Cityscapes).

BDD100K Instance Segmentation We train Mask R-CNN [21] R-50 FPN using MMDetection [40] on the BDD100K dataset [61]. The training dataset contains 7K images with 9,954 person instances (after filtration following section 3), and we evaluate on the original validation dataset (1K images). Out of 9,954 instances, 687 were detected by the face detector. MMdetection is run with commit: b95583270c.

4.2 Effect of Face Anonymization

We start our analysis by focusing on face anonymization. On Cityscapes and BDD100k (table 1, 2), we observe no significant performance difference from any type of face anonymization. We note that realistic anonymization slightly outperforms mask-out anonymization for both datasets. In Figure 7, we find that the majority of boxes in BDD100K/Cityscapes cover less than 1% of the image area. Thus, it is not surprising that face anonymization has little impact on these datasets.

For COCO pose estimation (table 3), face anonymization severely impacts performance, where both mask-out and blurring degrade keypoint AP by $> 10\%$. This performance drop is significant for bounding box AP as well, reflecting

Table 1: Instance segmentation AP on the Cityscapes [8] validation set with a Mask R-CNN [21] R-50 FPN. **HM**=Histogram matching (section 3.4). **HM-LO**=Histogram matching via Latent Optimization (section 3.4).

	Anonymization Method	**AP** ↑	**AP50** ↑	**AP$_{\text{person}}$**
	Original	36.7 ± 0.1 (Δ)	62.8 ± 0.2	35.0 ± 0.2 (Δ)
Face	Blur	36.4 ± 0.2 (-0.3)	62.5 ± 0.2 (-0.3)	34.9 ± 0.1 (-0.1)
	Mask-out	**36.7** ± 0.2 (0.0)	**63.1** ± 0.2 (0.3)	34.9 ± 0.1 (-0.1)
	Realistic	36.6 ± 0.1 (-0.1)	62.8 ± 0.3 (0.0)	**35.0** ± 0.1 (0.0)
Body	Blur	31.4 ± 0.2 (-5.3)	54.5 ± 0.4 (-8.3)	2.1 ± 0.1 (-32.9)
	Mask-out	31.2 ± 0.1 (-5.5)	53.2 ± 0.1 (-9.6)	0.7 ± 0.1 (-34.3)
	Realistic	34.6 ± 0.1 (-2.1)	59.0 ± 0.3 (-3.8)	20.3 ± 0.2 (-14.7)
	Realistic + HM	34.3 ± 0.2 (-2.4)	58.9 ± 0.2 (-3.9)	21.3 ± 0.3 (-13.7)
	Realistic + HM-LO	**34.8** ± 0.2 (-1.9)	**60.0** ± 0.3 (-2.8)	**21.5** ± 0.1 (-13.5)

Table 2: Instance segmentation AP on the BDD100K [61] validation set with a Mask R-CNN [21] R-50 FPN.

	Anonymization Method	**AP** ↑	**AP50** ↑	**AP$_{\text{person}}$**
	Original	20.2 ± 0.2 (Δ)	34.9 ± 0.4 (Δ)	32.0 ± 0.0 (Δ)
Face	Blur	20.5 ± 0.1 (0.3)	35.9 ± 0.1 (1.0)	31.7 ± 0.1 (-0.3)
	Mask-out	20.3 ± 0.1 (0.1)	35.3 ± 0.3 (0.4)	31.4 ± 0.1 (-0.6)
	Realistic	**20.6** ± 0.1 (0.4)	**35.8** ± 0.3 (0.9)	**31.6** ± 0.2 (-0.4)
Body	Blur	15.4 ± 0.1 (-4.8)	26.3 ± 0.2 (-8.6)	0.5 ± 0.0 (-31.5)
	Mask-out	15.3 ± 0.0 (-4.9)	25.5 ± 0.1 (-9.4)	0.0 ± 0.0 (-32.0)
	Realistic	**17.0** ± 0.1 (-3.2)	**28.9** ± 0.4 (-6.0)	**12.8** ± 0.1 (-19.2)

Table 3: Keypoint (Kp.) AP on the COCO [37] validation set with a Keypoint R-50 FPN R-CNN [21].

	Anonymization Method	**Box AP ↑**	**Kp. AP ↑**
	Original	55.7 ± 0.0 (Δ)	65.2 ± 0.0 (Δ)
Face	Blur	50.3 ± 0.2 (-5.4)	53.5 ± 0.2 (-11.7)
	Mask-out	49.9 ± 0.2 (-5.8)	52.0 ± 0.3 (-13.2)
	Realistic	54.3 ± 0.1 (-1.4)	60.6 ± 0.1 (-4.6)
	Realistic + HR Faces	**54.4** ± 0.0 (-1.3)	**60.8** ± 0.2 (-4.4)
Body	Blur	17.8 ± 0.0 (-37.9)	4.4 ± 0.1 (-60.8)
	Mask-out	17.4 ± 0.1 (-38.3)	2.0 ± 0.1 (-63.2)
	Realistic	**24.0** ± 0.1 (-31.7)	**15.6** ± 0.1 (-49.6)

that the performance difference is not due to the inability to predict keypoints in the facial region. Likely, this is due to learning that blurring/masking artifacts correlate to the human body. Furthermore, we hypothesize that the major performance drop compared to Cityscapes and BDD100k is due to dataset distribution and not the task at hand. To validate this, we train an instance segmentation model on the anonymized COCO datasets and observe a similar performance drop [1].

Refining COCO Faces Although realistic anonymization significantly improves over traditional methods, there remains a considerable degradation between it and the original COCO dataset. We hypothesize that this degradation results from the following factors; limited synthesis quality, facial keypoint mismatch, and low-resolution synthesis. As the generative model is not conditioned on facial keypoints, the synthesized identity will likely not match the annotated keypoints. There exists keypoint guided anonymization models [38, 25, 52], which we leave for further work to investigate. Furthermore, the generative model synthesizes faces at 128 × 128 resolution, introducing

[1] For mask-out, we observe a 6.7% performance drop for Box AP for COCO instance segmentation, compared to a 10.4% drop for Box AP for Keypoint R-CNN in table 3. See Appendix A.2 for more details.

upsampling artifacts for any face above. In total, we found 14,688 faces with an area larger than 128^2. To remove these upsampling artifacts, we employ a higher resolution (256×256) face synthesis model from DeepPrivacy2 [23] to anonymize any face larger than 128×128. This slightly improved downstream use (marked *Realistic + HR Faces* in table 3), supporting our hypothesis that upsampling artifacts can degrade image utility for COCO keypoint detection training.

4.3 Effect of Full-Body Anonymization

For full-body anonymization, we observe a substantial decline in performance for both traditional and realistic anonymization methods (table 1, 2, 3). Traditional anonymization leads to a complete degradation in performance, whereas realistic anonymization improves this significantly. Interestingly, the performance of realistic full-body anonymization on BDD100K [61] is noticeably worse than for Cityscapes [8], which we discuss further below.

Clearly, realistic full-body anonymization significantly degrades the performance compared to the original dataset, which we attribute to the following three issues: keypoint detection errors, synthesis limitations, and global context mismatch. Synthesizing realistic human bodies is difficult, and current models may introduce severe visual artifacts for many contexts. Furthermore, current methods rely on a crop-based anonymization method (discussed in Section 3.4), which can result in synthesized identities that do not fit the global context of the image. Section 3.4 introduced naive histogram matching and HM-LO to mitigate this issue, which we find to significantly improve results on the Cityscapes dataset (Table 1).

BDD100k *vs.* Cityscapes The decline in performance is significantly more prominent for BDD100k than Cityscapes, despite both datasets being collected for the same purpose. We suspect this discrepancy stems from two sources; keypoint annotations and dataset resolution. First, ViTPose [58] detects keypoints for 95.8% of the instances in the Cityscapes dataset, whereas it only detects for 85.5% in the BDD100k dataset. Secondly, the BDD100k images are of lower resolution (720p) than Cityscapes (2048×1024). This results in 36% of the instance crops having an area $< 32^2$, compared to 24% for

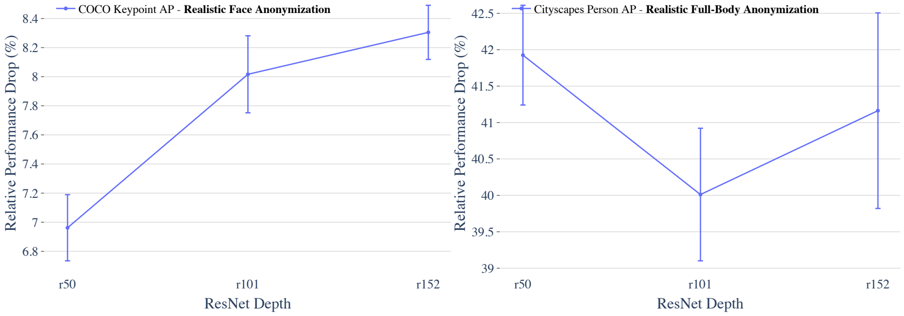


Figure 8: The relative performance drop of realistic anonymization (face or body) for different ResNet depths.

Cityscapes. While lower-resolution bodies are easier to synthesize in theory, the employed generative model operates at the resolution 288 × 160, and major deviations from this resolution can induce visual artifacts. For example, if we do not anonymize any detections $< 32^2$, BDD100k $\text{AP}_{\text{person}}$ is increased from 12.8% to 19.9%. In contrast, this increases $\text{AP}_{\text{person}}$ from 20.3% to 23.4% for Cityscapes.

4.4 Ablations

Do Larger Models Generalize Better? Zhou *et al.* [63] observes that deeper models are less impacted by realistic image anonymization. In our experiments, we observed the reverse to be true. We train a ResNet-50, 101, and 152 and compare the relative performance drop of realistic anonymization compared to the original dataset. We investigate this for realistic face anonymization on COCO and full-body anonymization for Cityscapes. Figure 8 reflects that larger models perform worse for both the COCO, whereas it is not clear for the Cityscapes dataset.

Diversity vs. Quality Trade-off GANs can trade off the diversity of samples with quality through the truncation trick [5]. Specifically, by interpolating the input latent variable $z \sim \mathcal{N}(0,1)$ towards the mode of $\mathcal{N}(0,1)$, generated diversity is traded off for improved quality. This leaves the question, what is

Table 4: Instance segmentation AP on the Cityscapes [8] validation set with full-body anonymization using different latent sampling strategies. Results from Mask R-CNN [21] R-50 FPN.

Anonymization Method Original	**AP ↑** 36.7 ± 0.1 (Δ)	**AP50 ↑** 62.8 ± 0.2 (Δ)	**AP$_{\text{person}}$** 35.0 ± 0.2 (Δ)
No Truncation	34.0 ± 0.2 (-2.7)	57.7 ± 0.5 (-5.1)	18.6 ± 0.2 (-16.4)
Unimodal Truncation	33.9 ± 0.2 (-2.8)	58.1 ± 0.3 (-4.7)	19.7 ± 0.5 (-15.3)
Multi-modal Truncation (**Default**)	**34.6** ± 0.1 (-2.1)	**59.0** ± 0.3 (-3.8)	**20.3** ± 0.2 (-14.7)

best for anonymization purposes? Limited diversity might result in a detector primarily being able to detect a small diversity of the population, whereas limited quality might reduce transferability to real-world data.

We explore the use of the truncation trick for anonymization purposes, where we investigate the use of no truncation, multi-modal truncation [41] [2], and standard truncation [5]. Note that in all other experiments, multi-modal truncation is used for full-body anonymization, while we use no truncation for face anonymization.

Table 4 reflects that both standard and multi-modal truncation performs substantially better than no truncation for AP_{person}. Furthermore, we observe that multi-modal truncation further improves over standard truncation.

Does Anonymization Impact Other Classes? For many tasks, person detection is not the intended task of the anonymized data (*e.g.* road damage detection [1]). Thus, we investigate the impact of anonymization where person detection is not part of the task. To answer this, we re-train the instance segmentation for the Cityscapes dataset and exclude the "person" class from the segmentation task.

Our experiment (see Appendix A.3) reflects that full-body anonymization does not impact the detection of the following classes: bus, car, motorcycle, train,

[2] Multi-modal truncation [41] approximates multiple modes of the latent distribution, enabling sampling high-quality images while minimizing the loss of diversity. We estimate 512 cluster centers following [23].

or truck. However, we do notice a performance drop for detecting "rider" and "bicycle". We believe this is due to detection overlaps.

5 Conclusion

In this work, we investigated the impact of anonymization for training computer vision models, with a focus on autonomous vehicle datasets. Our experiments reflect that face anonymization (obfuscation and realistic) has little to no impact for instance segmentation on the BDD100K [61] and Cityscapes [8] datasets. In contrast, face obfuscation severely degrades the performance of keypoint detection models on the COCO [37] dataset, as faces are more prevalent in comparison to the BDD100k and Cityscapes datasets. We find that realistic face anonymization can significantly reduce this performance drop. Furthermore, we find that full-body obfuscation severely impairs performance on all datasets, where realistic full-body anonymization can notably alleviate this issue. In summary, our findings reflect that realistic anonymization is a superior option compared to traditional methods. However, they are not a complete substitute for real data, especially for full-body anonymization, as current generative models can often produce unnatural humans that do not fit the given context.

Societal Impact Computer vision models are becoming increasingly adopted for solving challenging tasks everywhere in our society, from manufacturing to driving our cars. These models require task-specific training data to specialize for the task at hand. Collecting such data is troublesome due to privacy legislation, especially for autonomous vehicles which operate in environments where individuals appear everywhere. Our findings indicate that realistic anonymization can effectively substitute the original data, encouraging companies to protect individuals' privacy without compromising model performance. Our main societal concern is that we do not advocate that the anonymization methods studied in this paper give any sort of privacy guarantee. The detailed discussion in Section 2 clarifies that face anonymization and image blurring are questionable with respect to privacy. Furthermore, anonymized bodies could still be identified, *e.g.* from gait recognition [27].

5.1 Limitations and Further Work

Limitations The primary limitation of our study is the reliance on automatic annotations, where we use DSFD [34] for face detections, and ViTPose [58] for keypoint annotations. While the performance of these methods is impressive, they introduce ambiguity in our results, questioning if the current performance degradation is due to annotation errors or synthesis limitations. Furthermore, due to the filtering criteria for full-body anonymization and automatic annotation of faces, we are not able to anonymize all individuals in the images. Finally, it is also worth mentioning that our analysis is restricted to ResNet [22] and R-CNN [49] based models and that other architectures (*e.g.* YOLO [2]) may respond differently to anonymization artifacts.

Further Work Our explorative analysis of current realistic anonymization techniques highlights several areas of improvement and limitations. To the best of our knowledge, all current anonymization techniques rely on a crop-based anonymization method to improve synthesis quality. However, this can result in a mismatch between the synthesized identity and the global image. For example, the synthesized identity may not align with the global context of the image despite fitting the local crop given to the generative model. To mitigate this, we show that histogram equalization can reduce the impact of this, but we note that histogram equalization is far from the optimal solution. Furthermore, our experiments reflect that there are major practical difficulties remaining in effectively employing generative models for anonymization. For example, current anonymization techniques operate at a fixed synthesis resolution, where large deviations from the operating resolution (*e.g.* bodies smaller than 32^2) result in unnatural images, which impacts performance. Finally, we note that there are several intriguing and unexplored challenges to handle for synthesizing human figures for anonymization in autonomous vehicles. *E.g.* handling multi-view consistency, temporal consistency, or ensuring that the synthesized demography matches the demography of the original data.

Acknowledgement The computations were performed on resources provided by the NTNU IDUN/EPIC computing cluster [51]. Furthermore, we thank Rudolf Mester for his general support and thoughtful discussion.

References

[1] Deeksha M Arya, Hiroya Maeda, S Ghosh, Durga Toshniwal, Yoshihide Sekimoto Indian Institute of Technology Roorkee, India, T U O Tokyo, Japan., UrbanX Technologies, Inc., and Tokyo. RDD2022: A multinational image dataset for automatic Road Damage Detection. *arXiv preprint arXiv:2209.08538*, 2022.

[2] Alexey Bochkovskiy, Chien-Yao Wang, and Hong-Yuan Mark Liao. YOLOv4: Optimal Speed and Accuracy of Object Detection. *arXiv preprint arXiv:2004.10934*, 2020.

[3] Michael Boyle, Christopher Edwards, and Saul Greenberg. The effects of filtered video on awareness and privacy. In *Proceedings of the 2000 ACM conference on Computer supported cooperative work*, pages 1–10. ACM, 2000.

[4] Karla Brkic, Ivan Sikiric, Tomislav Hrkac, and Zoran Kalafatic. I Know That Person: Generative Full Body and Face De-identification of People in Images. In *2017 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, volume 2017-July, pages 1319–1328. IEEE, jul 2017.

[5] Andrew Brock, Jeff Donahue, and Karen Simonyan. Large Scale GAN Training for High Fidelity Natural Image Synthesis. In *International Conference on Learning Representations*, 2019.

[6] Holger Caesar, Varun Bankiti, Alex H. Lang, Sourabh Vora, Venice Erin Liong, Qiang Xu, Anush Krishnan, Yu Pan, Giancarlo Baldan, and Oscar Beijbom. nuScenes: A Multimodal Dataset for Autonomous Driving. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 11618–11628. IEEE, jun 2020.

[7] Umur A. Ciftci, Gokturk Yuksek, and Ilke Demir. My Face My Choice: Privacy Enhancing Deepfakes for Social Media Anonymization. In *2023 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, 2023.

[8] Marius Cordts, Mohamed Omran, Sebastian Ramos, Timo Rehfeld, Markus Enzweiler, Rodrigo Benenson, Uwe Franke, Stefan Roth, and

Bernt Schiele. The Cityscapes Dataset for Semantic Urban Scene Understanding. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3213–3223. IEEE, jun 2016.

[9] Council of European Union. Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (General Data Protection Regulation), 2016.

[10] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. ImageNet: A large-scale hierarchical image database. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pages 248–255. IEEE, jun 2009.

[11] Petr Dvořáček and Petr Hurtik. What Is the Cost of Privacy? In *Communications in Computer and Information Science*, volume 1602 CCIS, pages 696–706. Springer International Publishing, 2022.

[12] Andrea Frome, German Cheung, Ahmad Abdulkader, Marco Zennaro, Bo Wu, Alessandro Bissacco, Hartwig Adam, Hartmut Neven, and Luc Vincent. Large-scale privacy protection in Google Street View. In *2009 IEEE 12th International Conference on Computer Vision*, pages 2373–2380. IEEE, sep 2009.

[13] Oran Gafni, Lior Wolf, and Yaniv Taigman. Live Face De-Identification in Video. In *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 9377–9386. IEEE, oct 2019.

[14] Andrew C. Gallagher and Tsuhan Chen. Clothing cosegmentation for recognizing people. In *2008 IEEE Conference on Computer Vision and Pattern Recognition*, pages 1–8. IEEE, jun 2008.

[15] Jakob Geyer, Yohannes Kassahun, Mentar Mahmudi, Xavier Ricou, Rupesh Durgesh, Andrew S. Chung, Lorenz Hauswald, Viet Hoang Pham, Maximilian Mühlegg, Sebastian Dorn, Tiffany Fernandez, Martin Jänicke, Sudesh Mirashi, Chiragkumar Savani, Martin Sturm, Oleksandr Vorobiov, Martin Oelker, Sebastian Garreis, and Peter Schuberth. A2D2: Audi Autonomous Driving Dataset. *arXiv preprint arXiv:2004.06320*, 2020.

[16] Ralph Gross, Latanya Sweeney, Jeffrey Cohn, Fernando de la Torre, and Simon Baker. Face De-identification. In *Protecting Privacy in Video Surveillance*, pages 129–146. Springer London, London, 2009.

[17] Ralph Gross, Latanya Sweeney, F. de la Torre, and Simon Baker. Model-Based Face De-Identification. In *2006 Conference on Computer Vision and Pattern Recognition Workshop (CVPRW'06)*, pages 161–161. IEEE, 2006.

[18] Riza Alp Guler, Natalia Neverova, and Iasonas Kokkinos. DensePose: Dense Human Pose Estimation in the Wild. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7297–7306. IEEE, jun 2018.

[19] Rakibul Hasan, Eman Hassan, Yifang Li, Kelly Caine, David J. Crandall, Roberto Hoyle, and Apu Kapadia. Viewer Experience of Obscuring Scene Elements in Photos to Enhance Privacy. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*, volume 2018-April, pages 1–13, New York, NY, USA, apr 2018. ACM.

[20] Jianping He, Bin Liu, Deguang Kong, Xuan Bao, Na Wang, Hongxia Jin, and George Kesidis. PUPPIES: Transformation-Supported Personalized Privacy Preserving Partial Image Sharing. In *2016 46th Annual IEEE/IFIP International Conference on Dependable Systems and Networks (DSN)*, pages 359–370. IEEE, jun 2016.

[21] Kaiming He, Georgia Gkioxari, Piotr Dollar, and Ross Girshick. Mask R-CNN. In *2017 IEEE International Conference on Computer Vision (ICCV)*, pages 2980–2988. IEEE, oct 2017.

[22] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep Residual Learning for Image Recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 770–778. IEEE, jun 2016.

[23] Håkon Hukkelås and Frank Lindseth. DeepPrivacy2: Towards Realistic Full-Body Anonymization. In *2023 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pages 1329–1338. IEEE, jan 2023.

[24] Håkon Hukkelås and Frank Lindseth. Synthesizing anyone, anywhere, in any pose. *arXiv preprint arXiv:2304.03164*, 2023.

[25] Håkon Hukkelås, Rudolf Mester, and Frank Lindseth. DeepPrivacy: A Generative Adversarial Network for Face Anonymization. In George Bebis, Richard Boyle, Bahram Parvin, Darko Koracin, Daniela Ushizima, Sek Chai, Shinjiro Sueda, Xin Lin, Aidong Lu, Daniel Thalmann, Chaoli Wang, and Panpan Xu, editors, *Advances in Visual Computing*, pages 565–578. Springer International Publishing, 2019.

[26] Håkon Hukkelås, Morten Smebye, Rudolf Mester, and Frank Lindseth. Realistic Full-Body Anonymization with Surface-Guided GANs. In *2023 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pages 1430–1440. IEEE, jan 2023.

[27] Anil K Jain, Patrick Flynn, and Arun A Ross. *Handbook of Biometrics*. Springer US, Boston, MA, 2008.

[28] Amin Jourabloo, Xi Yin, and Xiaoming Liu. Attribute preserved face de-identification. In *2015 International Conference on Biometrics (ICB)*, pages 278–285. IEEE, may 2015.

[29] Tero Karras, Samuli Laine, and Timo Aila. A Style-Based Generator Architecture for Generative Adversarial Networks. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4396–4405. IEEE, jun 2019.

[30] Sander R. Klomp, Matthew Van Rijn, Rob G.J. Wijnhoven, Cees G.M. Snoek, and Peter H.N. De With. Safe Fakes: Evaluating Face Anonymizers for Face Detectors. In *2021 16th IEEE International Conference on Automatic Face and Gesture Recognition (FG 2021)*, pages 1–8. IEEE, dec 2021.

[31] Alexey Kurakin, Ian J. Goodfellow, and Samy Bengio. Adversarial Examples in the Physical World. In *Artificial Intelligence Safety and Security*, pages 99–112. Chapman and Hall/CRC, jul 2018.

[32] Karen Lander, Vicki Bruce, and Harry Hill. Evaluating the effectiveness of pixelation and blurring on masking the identity of familiar faces. *Applied Cognitive Psychology*, 15(1):101–116, jan 2001.

[33] Jizhizi Li, Sihan Ma, Jing Zhang, and Dacheng Tao. Privacy-Preserving Portrait Matting. In *Proceedings of the 29th ACM International Conference on Multimedia*, pages 3501–3509, New York, NY, USA, oct 2021. ACM.

[34] Jian Li, Yabiao Wang, Changan Wang, Ying Tai, Jianjun Qian, Jian Yang, Chengjie Wang, Jilin Li, and Feiyue Huang. DSFD: Dual Shot Face Detector. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5055–5064. IEEE, jun 2019.

[35] Yifang Li, Nishant Vishwamitra, Bart P. Knijnenburg, Hongxin Hu, and Kelly Caine. Blur vs. Block: Investigating the Effectiveness of Privacy-Enhancing Obfuscation for Images. In *2017 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, volume 2017-July, pages 1343–1351. IEEE, jul 2017.

[36] Yifang Li, Nishant Vishwamitra, Bart P. Knijnenburg, Hongxin Hu, and Kelly Caine. Effectiveness and Users' Experience of Obfuscation as a Privacy-Enhancing Technology for Sharing Photos. *Proceedings of the ACM on Human-Computer Interaction*, 1(CSCW):1–24, dec 2017.

[37] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. Microsoft COCO: Common Objects in Context. In *European conference on computer vision*, volume 8693 LNCS, pages 740–755. Springer, Cham, 2014.

[38] Maxim Maximov, Ismail Elezi, and Laura Leal-Taixe. CIAGAN: Conditional Identity Anonymization Generative Adversarial Networks. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5446–5455. IEEE, jun 2020.

[39] Richard McPherson, Reza Shokri, and Vitaly Shmatikov. Defeating Image Obfuscation with Deep Learning. *arXiv preprint arXiv:1609.00408*, sep 2016.

[40] MMDetection Contributors. OpenMMLab Detection Toolbox and Benchmark, 2018.

[41] Ron Mokady, Omer Tov, Michal Yarom, Oran Lang, Inbar Mosseri, Tali Dekel, Daniel Cohen-Or, and Michal Irani. Self-Distilled StyleGAN:

Towards Generation from Internet Photos. In *ACM SIGGRAPH 2022 Conference Proceedings*, pages 1–9. ACM, aug 2022.

[42] Carman Neustaedter, Saul Greenberg, and Michael Boyle. Blur filtration fails to preserve privacy for home-based video conferencing. *ACM Transactions on Computer-Human Interaction*, 13(1):1–36, mar 2006.

[43] Natalia Neverova, David Novotny, Vasil Khalidov, Marc Szafraniec, Patrick Labatut, and Andrea Vedaldi. Continuous Surface Embeddings. In *Advances in Neural Information Processing Systems*, volume 33, pages 17258–17270. Curran Associates, Inc., nov 2020.

[44] E.M. Newton, Latanya Sweeney, and Bradley Malin. Preserving privacy by de-identifying face images. *IEEE Transactions on Knowledge and Data Engineering*, 17(2):232–243, feb 2005.

[45] Seong Joon Oh, Rodrigo Benenson, Mario Fritz, and Bernt Schiele. Faceless Person Recognition: Privacy Implications in Social Media. In *Computer Vision - ECCV 2016*, pages 19–35. Springer Verlag, 2016.

[46] Seong Joon Oh, Mario Fritz, and Bernt Schiele. Adversarial Image Perturbation for Privacy Protection A Game Theory Perspective. In *2017 IEEE International Conference on Computer Vision (ICCV)*, volume 2017-Octob, pages 1491–1500. IEEE, oct 2017.

[47] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, and Others. Pytorch: An imperative style, high-performance deep learning library. *Advances in neural information processing systems*, 32, 2019.

[48] A. J. Piergiovanni and Michael S. Ryoo. AViD dataset: Anonymized videos from diverse countries. In *Advances in Neural Information Processing Systems*, volume 2020-Decem, 2020.

[49] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. In *Advances in neural information processing systems*, pages 91–99, 2015.

[50] Zhongzheng Ren, Yong Jae Lee, and Michael S Ryoo. Learning to Anonymize Faces for Privacy Preserving Action Detection. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 639–655. Springer International Publishing, 2018.

[51] Magnus Själander, Magnus Jahre, Gunnar Tufte, and Nico Reissmann. EPIC: An energy-efficient, high-performance GPGPU computing research infrastructure, 2019.

[52] Qianru Sun, Liqian Ma, Seong Joon Oh, Luc Van Gool, Bernt Schiele, and Mario Fritz. Natural and Effective Obfuscation by Head Inpainting. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5050–5059. IEEE, jun 2018.

[53] Qianru Sun, Ayush Tewari, Weipeng Xu, Mario Fritz, Christian Theobalt, and Bernt Schiele. A Hybrid Model for Identity Obfuscation by Face Replacement. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 570–586. Springer International Publishing, 2018.

[54] Matteo Tomei, Lorenzo Baraldi, Simone Bronzin, and Rita Cucchiara. Estimating (and fixing) the Effect of Face Obfuscation in Video Recognition. In *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 3257–3263. IEEE, jun 2021.

[55] Ries Uittenbogaard, Clint Sebastian, Julien Vijverberg, Bas Boom, Dariu M. Gavrila, and Peter H.N. de With. Privacy Protection in Street-View Panoramas Using Depth and Multi-View Imagery. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, volume 2019-June, pages 10573–10582. IEEE, jun 2019.

[56] Michael J. Wilber, Vitaly Shmatikov, and Serge Belongie. Can we still avoid automatic face detection? In *2016 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 1–9. IEEE, mar 2016.

[57] Yuxin Wu, Alexander Kirillov, Francisco Massa, Wan-Yen Lo, and Ross Girshick. Detectron2, 2019.

[58] Yufei Xu, Jing Zhang, Qiming Zhang, and Dacheng Tao. ViTPose: Simple Vision Transformer Baselines for Human Pose Estimation. *arXiv preprint arXiv:2204.12484v3*, 2022.

[59] Kaiyu Yang, Jacqueline Yau, Li Fei-Fei, Jia Deng, and Olga Russakovsky. A Study of Face Obfuscation in ImageNet. In *International Conference on Machine Learning*, pages 25313–25330, mar 2022.

[60] Shuo Yang, Ping Luo, Chen Change Loy, and Xiaoou Tang. WIDER FACE: A Face Detection Benchmark. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5525–5533. IEEE, 2016.

[61] Fisher Yu, Haofeng Chen, Xin Wang, Wenqi Xian, Yingying Chen, Fangchen Liu, Vashisht Madhavan, and Trevor Darrell. BDD100K: A Diverse Driving Dataset for Heterogeneous Multitask Learning. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2633–2642. IEEE, jun 2020.

[62] Ning Zhang, Manohar Paluri, Yaniv Taigman, Rob Fergus, and Lubomir Bourdev. Beyond frontal faces: Improving Person Recognition using multiple cues. In *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, volume 07-12-June, pages 4804–4813. IEEE, jun 2015.

[63] Jingxing Zhou and Jurgen Beyerer. Impacts of Data Anonymization on Semantic Segmentation. In *2022 IEEE Intelligent Vehicles Symposium (IV)*, volume 2022-June, pages 997–1004. IEEE, jun 2022.

ISBN 978-82-326-7542-5 (printed ver.)
ISBN 978-82-326-7541-8 (electronic ver.)
ISSN 1503-8181 (printed ver.)
ISSN 2703-8084 (online ver.)

NTNU
Norwegian University of
Science and Technology