Master's thesis

NTNU
Norwegian University of Science and Technology
Faculty of Information Technology and Electrical Engineering
Department of ICT and Natural Sciences

Mahdis Saharkhizan

Automatic Face Anonymization in Videos Data Based on YOLOv7

Master's thesis in Simulation and Visualization
Supervisor: Prof. Ibrahim A. Hameed
Co-supervisor: Muhammad Umair Hassan
June 2023

NTNU
Norwegian University of
Science and Technology

Mahdis Saharkhizan

Automatic Face Anonymization in Videos Data Based on YOLOv7

Master's thesis in Simulation and Visualization
Supervisor: Prof. Ibrahim A. Hameed
Co-supervisor: Muhammad Umair Hassan
June 2023

Norwegian University of Science and Technology
Faculty of Information Technology and Electrical Engineering
Department of ICT and Natural Sciences

NTNU
Norwegian University of
Science and Technology

Mahdis Saharkhizan

Automatic Face Anonymization in Videos Data Based on YOLOv7

Master's thesis in Simulation and Visualization
Supervisor: Prof. Ibrahim A. Hameed
Co-supervisor: Muhammad Umair Hassan
June 2023

Norwegian University of Science and Technology
Faculty of Information Technology and Electrical Engineering
Department of ICT and Natural Sciences

ABSTRACT

With the widespread availability of digital visual data across various industries, such as video surveillance, social networks, and media, concerns regarding privacy have grown significantly. Face-anonymization techniques have emerged as a means to safeguard individuals' privacy while preserving the quality and intelligibility of the primary visual content. Common methods for obscuring human identities include black masking, pixelization, and blurring. Prior to applying these techniques, it is crucial to accurately identify faces within each frame or image.

In this study, we present the effectiveness of YOLOv7, a rapid and precise face detection algorithm, in identifying faces in images and videos across diverse real-world scenarios. We evaluate the performance of our face-anonymization methodology on the identified faces using qualitative measures, such as FID and reverse image search of obscured celebrity faces. Our experimental results on the WIDER Face dataset demonstrate a final mAP of 0.746 for face detection and an FID of 17.88 for face anonymization. These results showcase the capability of our method in effectively anonymizing faces in real-time surveillance videos.

The practical applications of our work extend to fields such as ecology, public and private spaces, and dataset preparation and distribution. Notably, our method stands out among existing studies in this domain by enabling face masking in real-time videos under diverse weather conditions, highlighting its novelty and exceptional performance.

Keywords: face detection, face anonymization, YOLOv7, privacy preservation, real-time surveillance, visual data.

ACKNOWLEDGMENTS

I would like to express my deepest gratitude to my supervisor, Prof. Ibrahim A. Hameed, and my co-supervisor Muhammad Umair Hassan, for their unwavering support, guidance, and invaluable expertise throughout the entire duration of this research project. Their encouragement, constructive feedback, and dedication have been instrumental in shaping the direction and quality of this thesis.

I extend my heartfelt appreciation to the faculty members and staff at NTNU-Department of ICT and Engineering, who have contributed to my academic journey. Their commitment to excellence in education, their willingness to share knowledge, and their passion for research have profoundly influenced my growth as a scholar.

In particular, I would like to acknowledge the support of my beautiful lovely, and kind family, my dearest friends, and most specifically, my beloved husband. Words cannot express the depth of my gratitude for his unwavering encouragement, understanding, and constant presence throughout this process. His love, belief in my abilities, unconditional support, patience, and willingness to lend an ear during moments of doubt have provided me with the strength and resilience needed to complete this thesis. His unwavering faith in my potential has been a constant source of inspiration and motivation.

In conclusion, I am profoundly grateful to everyone who has contributed to this master thesis. Your support, encouragement, and collaboration have played an integral role in its successful completion. Thank you for being a part of this journey and for helping me grow both personally and professionally.

CONTENTS

LIST OF FIGURES

LIST OF TABLES

ABBREVIATIONS

List of all abbreviations in alphabetic order:

- **NTNU** Norwegian University of Science and Technology
- **AI** Artificial Intelligence
- **ML** Machine Learning
- **YOLO** You only look once
- **YOLOv5** You Only Look Once version 5
- **YOLOv7** You Only Look Once version 7
- **E-ELAN** Extended Efficient Layer Aggregation Network
- **MAP** Mean Average Percentage
- **GDPR** General Data Protection Regulation
- **GAN** Generative Adversarial Network
- **ITS** Intelligent Transportation Systems
- **R-CNN** Region-based Convolutional Neural Networks
- **FPN** Feature Pyramid Network
- **lr** learning rate
- **UFDD** Unconstrained Face Detection Dataset
- **IoU** Intersection over Union
- **FID** Fréchet Inception Distance
- **AV** Autonomous vehicle

CHAPTER

ONE

INTRODUCTION

1.1 Motivation

The growing application of face images and modern Artificial Intelligent (AI) technology has raised an important concern in privacy protection. In numerous real-world contexts, including scientific research, social sharing, and commercial applications, many images and videos are made available without identity protection, posing security and privacy risks to individuals. Once someone's visual data, particularly their face images, is publicly disclosed, they may run into trouble if this information is collected and utilized illegally by other parties using AI techniques like DeepFake [1, 2]. Therefore, it is essential to create efficient algorithms to preserve people's privacy when sharing their visual data.

The General Data Protection Regulation (GDPR) was established by the European Union in 2016 [3] to secure individuals' privacy rights [4]. As a consequence of this legislation, it is necessary to eliminate, blur, or pixelate human faces from images of people taken in public areas in order to make them anonymous [5, 6]. Indeed, anonymization is ensured by using these approaches without destroying the video quality, i.e., only the human faces will be obscured, and the whole concept of the video or images will be clear and recognizable (see Fig. 4.3.2).

Figure 1.1.1: Face anonymization at large-scale.

Face anonymization entails the process of face detection, which is crucial. First,

we must find faces in videos or images and then apply anonymization techniques, e.g., blurring, pixelation, and blackened, to conceal identifiable faces. Although developing a model that works for detecting faces in real-time surveillance videos is substantially difficult, considering that our main objective is not face detection. Therefore, having a model that performs well enough and fast is sufficient for the main target of anonymization in real-time videos. So, finding faces in various real-world scenarios is a significant challenge. We require fast model processing, especially when employing real-time cameras, to obscure everyone's personas and make them anonymous.

1.2 Problem Statement

To begin with, it is important to acknowledge that our primary objective revolves around safeguarding individuals' privacy and preventing the disclosure of information captured by surveillance cameras. Specifically, our focus is on anonymizing faces in images or real-time videos. Hence, what we require is a fast and precise approach that can effectively identify faces and promptly apply anonymization techniques.

To expand, although deep learning has developed efficient techniques that accurately detect faces in this regard [7]. However, most developed techniques only operate effectively when people's faces are occluded, too near or far from the camera lens, or when the weather could be better. In light of the numerous real-world scenarios that may arise while recording or archiving videos or images that contain individuals, it is crucial to hide their faces and preserve their privacy. In the computer vision and facial detection world, YOLOv7 [8] (You Only Look Once version 7) is one of our best options for overcoming these challenges and keeping faces obscured in the conditions as mentioned earlier.

In fact, among all existing real-time object detectors, YOLOv7 is the fastest and most precise real-time object detection model for computer vision applications. YOLOv7 surpasses previously known object detectors in both speed and accuracy in the range from 5 FPS to 160 FPS. It has the highest accuracy, 56.8% AP, among all known real-time object detectors with 30 FPS or higher on GPU V100. YOLOv7-E6 object detector (56 FPS V100, 55.9% AP) outperforms both transformer-based detector SWINL Cascade-Mask R-CNN (9.2 FPS A100, 53.9% AP) by 509% in speed and 2% in accuracy, and convolutional-based detector ConvNeXt-XL Cascade-Mask R-CNN (8.6 FPS A100, 55.2% AP) by 551% in speed and 0.7% AP in accuracy. Moreover, in this research, we train and evaluate this model on three separate datasets (WIDER Face [9], CelebA [10], and UFDD [11]) in order to fulfill our objective, face-anonymization in real-time.

1.3 Research questions

In summary, the following main research questions are addressed in this thesis:

1. How to anonymize the faces in video and images of surveillance cameras to protect individuals' privacy?

2. How to measure the effectiveness of proposed anonymization models?
3. How to reach the goal of making anonymized available for real-time videos?

1.4 Contributions

Our principal contributions to this work include the following:

1. Train deep learning models to achieve better accuracy in detection and blurring.
2. Blur/pixelate/blackened the faces in the surveillance videos:
    - Identifying out-of-focus/view, people far away from the closer view, and anonymizing them all.
    - Identifying people under different weather conditions (sun, rain, fog, snow, night, etc.) and anonymizing them all.
    - Identifying people even in low-resolution videos like real-time surveillance cameras and anonymizing them all.
    - Anonymizing all people in a crowded area, even occluded faces.
    - Anonymizing all people in the swift movement.
3. Testing our model with diverse images and real-time surveillance videos.
4. Validate the results through quantitative and qualitative evaluation.

1.5 Thesis outline

The overview of each chapter in the thesis is as per following:

- **Chapter 1** introduces the thesis, it includes motivation, problem statement, research question, and contributions.
- **Chapter 2** presents relevant background.
- **Chapter 3** provides the necessary theoretical background for AI, ethics in AI, Deep learning, and computer vision.
- **Chapter 4** discusses an overview of our proposed methodology in detail including the suggested neural network architecture, and the complete pipeline for 'face anonymizer'.
- **Chapter 5** is the implementation details of our proposed method and evaluation methods.
- **Chapter 6** summarizes the experiments and analyzes our results performed during this thesis.
- **Chapter 7** is about the final conclusion of the thesis.

- **Chapter 8** The potential use cases and challenges to face anonymization are discussed in this section and possible future works are available there as well.

CHAPTER

TWO

LITERATURE REVIEW

2.1 Face Detection

The queries about face detection arise after standard object detection. Indeed, the WIDER Face dataset [9] brought a new wave in the face detection field thanks to its challenging and accomplished data. Using this dataset, face detection creates quickly centers on the extraordinary and genuine variety of issues, counting scale, pose, occlusion, expression, makeup, light, obscure, etc.

These problems sparked researchers' interest in face detection, which led to numerous technique proposals in this area, especially issues related to the scale, context, and anchor to detect small faces. Among the relevant attempts, we can name MTCNN [12], FaceBox [13], S3FD [14], DSFD [15], RetinaFace [16], RefineFace [17], and the most recent ASFD [18], MaskFace [19], TinaFace [20], MogFace [21], and SCRFD [22]. Also, many well-done works can be found on the WIDER Face webpage [23]. In [24], the authors proposed a face detector based on YOLOv3 to improve real-time facial detection performance. They addressed the problem of varied detection accuracy for different face scales. Their proposed algorithm includes the anchor boxes and a regression loss function for appropriate face detection. They tested their model on WIDER Face and FDDB datasets.

Whereas many challenges in face detection, such as those mentioned above, are effectively tended to, there still exist a few issues not particularly captured by existing strategies or datasets; issues like weather-based degradations, motion blur, focus blur and focus blur, and several others. Therefore, except for some of the face detectors that explore unique characteristics of the human face, others are just general object detectors adopted and modified for face detection.

2.2 Real-time Face Detectors

Today, real-time object detectors are considered crucial topics, and the state-of-the-art ones are mainly based on YOLO [25]. The main goals of these works are to improve the following aspects: faster and stronger network architecture, more precise feature integration method [26, 27], more accurate object detection

method [28, 29], more robust and general loss function [30, 31], more efficient label assignment method [32, 33], and also more effective training method. YOLOv7 is the most recent version of YOLO, which overcomes all the issues in object detection so far and achieves state-of-the-art improvements in accuracy and speed. It performs well over varied datasets and a wide range of real-world situations to detect objects from the background. Accordingly, we chose this model for our face anonymization process.

2.3 Face Anonymization

Face de-identification [34] is a technique that has developed through time that focuses on maintaining facial characteristics like gender, age, and race while de-identifying face images. Today, only a few research works address the problem of eliminating personally identifiable information from images that include faces. In order to preserve users' privacy, some AI techniques have been unveiled that purport to anonymize faces in images "while the original data distribution stays unbroken" [34]. The most recent works are deep learning-based methods like Li et al. did in [34]. In this work, Li et al., based on an identity-aware class activation heatmap of the original faces, tried to generate the privacy-preserving face.

The introduction of the Generative Adversarial Network (GAN) and its variations [35] represents the largest change. These neural networks can create new faces out of context or noise, allaying severe privacy issues. Modern face-generating GANs, like [36], create highly detailed, high-resolution faces. These faces are not reasonable to be utilized to anonymize faces inside a more prominent image since they do not mix the faces with the background within the original image. In [37], Sander et al. investigate the impact of both the conventional methods and GAN-based face anonymizers.

[38] presents a novel deep learning-based pipeline for face anonymization in the context of intelligent transportation systems (ITS). By leveraging diffusion models instead of GANs [37], the proposed pipeline offers an approach to realistic face in-painting. The two-stage method, comprising face detection and latent diffusion, provides an effective solution for anonymizing data for segmentation tasks and achieves comparable performance to recent GAN-based methods. Furthermore, the anonymized faces generated by the pipeline enhance the performance of face detection models. While this study adds to the continuous endeavors in safeguarding privacy while preserving the practicality of facial data in intelligent transportation systems, it focuses solely on anonymized faces found in high-resolution images. However, it does not encompass small faces and is restricted to favorable weather and lighting conditions.

Recent research has concentrated primarily on employing a DL-based method to create a new face to anonymize them. Most techniques only operate with images, even though they disregarded several practical concerns for de-identifying individuals, such as anonymizing people on real-time surveillance cameras that may capture people's movements in various climates and with various people in various

poses. Therefore, in our study, we trusted the standard anonymization techniques such as blurring, pixelization, and blacking to safeguard people's privacy in real life. With the substantial facial covered, these techniques are practical and reliable, so we may use them in any circumstance. These approaches are also easy to use and may be used widely.

CHAPTER

THREE

THEORY

3.1 AI

The goal of AI, a multidisciplinary science, is to develop intelligent machines that can carry out tasks that traditionally call for human intelligence. AI methods are used in the field of computer vision to create models and algorithms that can evaluate and comprehend visual data, such as pictures and movies. These methods allow computers to comprehend visual inputs and derive meaningful information from them. AI algorithms can be used, for instance, to automatically discover and anonymize faces in movies by recognizing facial traits and performing privacy-preserving changes like blurring or pixelation.

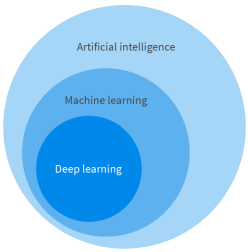


Figure 3.1.1: World of artificial intelligence, machine learning (ML), and deep learning.[39]

3.2 Ethics in AI

The field of ethics in artificial intelligence covers the ethical considerations and implications that are associated with the creation and use of AI systems. In the context of face anonymization, ethical considerations arise due to the sensitive nature of personal data and the potential privacy risks that are involved in the process. It is of the utmost importance to guarantee that methods of face anonymization respect the individual privacy rights of people, uphold fairness, openness, and accountability, and prevent the perpetuation of biases. For instance, while designing

a face anonymization system, ethical principles, and regulations should be followed. These guidelines and laws include things like gaining informed consent for the usage of data and adopting steps to prevent the re-identification of anonymized faces.

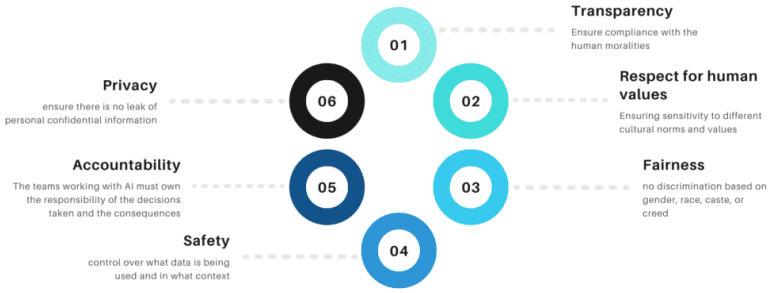


Figure 3.2.1: The interrelation between data and AI ethics.[40]

3.3 Deep Learning Fundamentals

Deep learning is a subfield of artificial intelligence that focuses on training deep neural networks to learn hierarchical representations from the input. This is accomplished through a process known as "deep training". Layers of a neural network are interconnected, and each layer processes incoming data in order to extract progressively more sophisticated characteristics. Because these networks are trained with a significant amount of data that has been tagged, it is possible for them to recognize trends and generate accurate forecasts. Deep learning is a technology that may be leveraged in the context of face anonymization to construct models that can automatically detect and anonymize faces in movies. These models can then be used. Deep neural networks that have been trained on datasets that have been annotated, for instance, are able to learn to identify facial features and perform anonymization strategies, such as substituting facial areas with features that are either synthetic or changed.

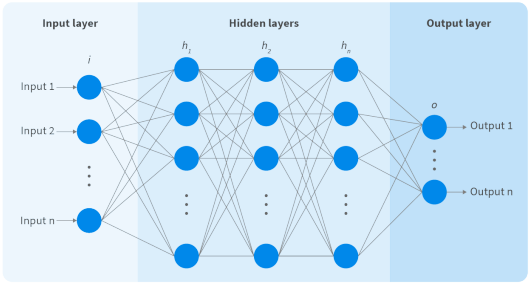


Figure 3.3.1: Deep Learning Network.[39]

3.4 Computer Vision

Computer vision is a subfield of computer science that focuses on making computers capable of comprehending and analyzing visual data. The process involves the development of algorithms and procedures that can extract useful information from visual media such as photographs and movies. In the context of face anonymization, computer vision techniques are essential for effectively detecting and localizing faces within movies. This can only be accomplished with their assistance. The identification of facial areas can be accomplished by face recognition algorithms using a variety of approaches, such as template matching, methods based on features, or methods based on machine learning. In order to differentiate faces from the background or other objects in the image, these algorithms examine visual clues such as color, texture, and shape.

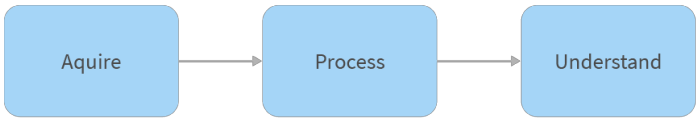


Figure 3.4.1: Three basic steps in the computer vision process: 1- Acquire the image data (Through video, pictures, or 3D technologies, images, even massive collections, can be collected in real-time for analysis.) 2- Vectorize the image (Although deep learning models automate a large portion of this procedure, the models are frequently trained by being first fed a large number of labeled or previously classified photos.) 3- Understand the image (The interpretive step, which identifies or categorizes the object, is the last stage.)[39]

3.4.1 Object Detection

Object detection is a fundamental problem in computer vision that involves recognizing and localizing things inside images or videos. This can be accomplished by analyzing the content of an image or video. Object detection is utilized in the context of face anonymization in order to discover faces that are candidates for anonymization. Object detection can be accomplished in a variety of ways, such as with the use of single-stage and two-stage detectors.

- **Single-stage detectors:** Object detection can be completed in a single pass when using single-stage detectors like YOLO (which stands for "You Only Look Once"). They do this by first dividing the image into a grid, after which they immediately anticipate the bounding box coordinates and class labels. Real-time detection is enabled by YOLO at the expense of some degree of precision in location determination.

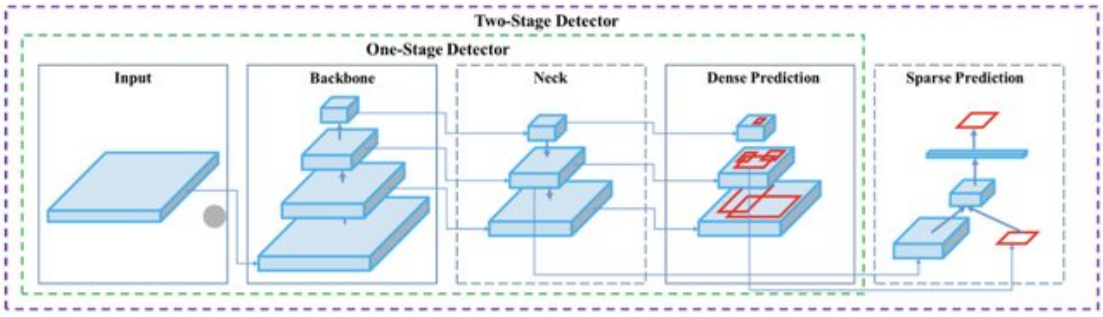


Figure 3.4.2: The difference between the two-stage and one-stage detectors according to [41]

In the context of face anonymization, selecting an object detection method that is both accurate and efficient is essential to the achievement of the desired outcomes. The utilization of single-stage detectors, such as the YOLO system, is one method that sees widespread application. It is able to provide detection capabilities in real-time, making it suited for applications that require fast and immediate processing. However, it is essential to keep in mind that the YOLO method may result in a reduction in localization accuracy compared to other approaches.

For our face anonymization work, leveraging YOLO-based models can yield significant advantages. The efficiency of YOLO enables quick identification of faces in videos, facilitating the seamless and real-time application of anonymization techniques. By rapidly detecting faces, YOLO streamlines the subsequent anonymization process, contributing to improved efficiency and responsiveness. This is particularly beneficial when working with large video datasets or in scenarios that demand timely face anonymization, such as real-time video surveillance.

- **Two-stage detectors:** For the purpose of object detection, two-stage detectors, such as Faster R-CNN (Region-based Convolutional Neural Networks), utilize a procedure that is split into two stages. The first thing that is done is the generation of possible object regions, often known as proposals. In the second step, these hypotheses are examined more closely and categorized in order to acquire precise item detections. In general, two-stage detectors offer improved precision; however, this benefit comes at the expense of increased computing complexity.

3.5 Choosing network - A Review

Although faster R-CNN-based models can produce more accurate face detection results for face anonymization, their real-time processing capabilities may be constrained by their computational complexity. YOLO-based models, on the other hand, are excellent at accurately detecting faces, which makes them particularly

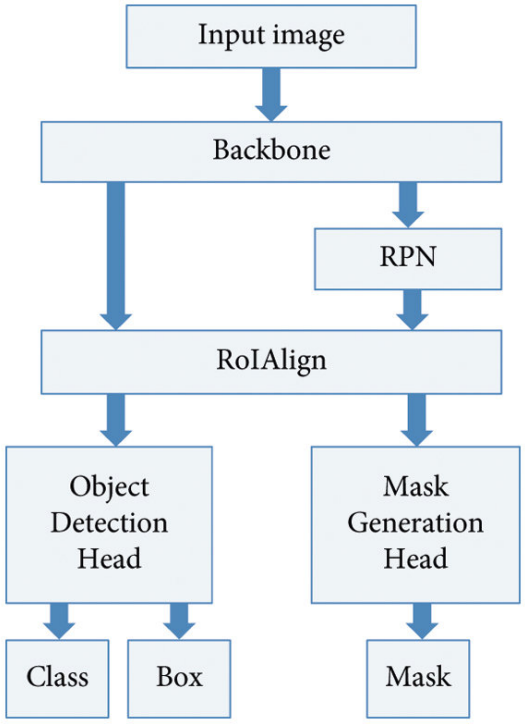


Figure 3.4.3: Mask R-CNN structure.[42]

ideal for real-time face anonymization tasks. With YOLO, faces can be quickly and precisely detected, supporting the smooth integration of anonymization methods and guaranteeing accurate and dependable anonymization of identified faces in movies.

Our face anonymization work can achieve quick and accurate face detection for efficient application of anonymization techniques in videos by carefully weighing the advantages and disadvantages of various object detection methods, with a focus on the advantages of YOLO's efficiency and real-time capabilities.

CHAPTER

FOUR

RESEARCH METHODOLOGY

Here you will find my thesis methodology and proposed pipeline that will be followed by implementation details of my work.

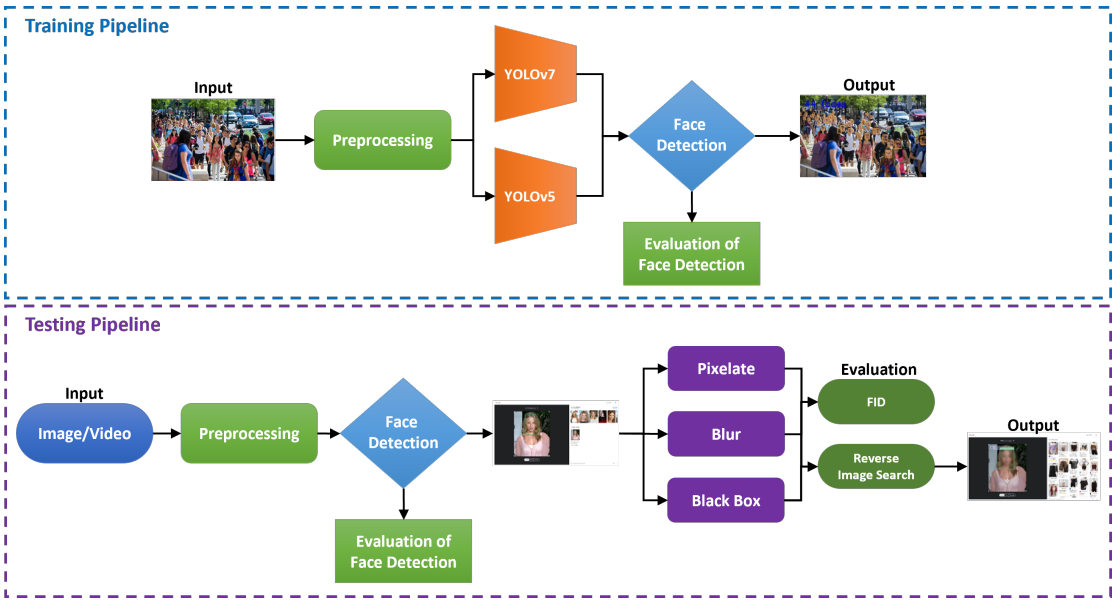


Figure 4.0.1: Training and testing pipelines of our proposed face anonymization work.

4.1 The Anonymization Pipeline

The proposed anonymization framework pipeline that we have developed is intended to protect the personal information of users of real-time videos. The process has two stages: the first is face detection, and the second is de-identification.

When it comes to face detection, we use two detectors based on the YOLOv7 and YOLOv5 models. These detectors give us the ability to accurately recognize human faces. After the face has been identified, we select one of the de-identification methods and use it to either completely cover the face or turn all of the pixels into zeros. Note that our masks only cover the human face and not any other parts

of the head or neck. This is a crucial point to keep in mind. The mask is also stretched out in order to provide complete coverage.

We train our algorithm to recognize and anonymize faces in order to accomplish our goal of concealing people's identities in films captured in real-time. After that, we put the model through its paces by putting it through a series of tests, including a variety of photos and video frames. Our model is trained on photos, and we test it with movies and images that it has not seen before.

In addition, we have a different setup for the prediction portion of our process that allows us to use movies as real-time input. We are able to preserve the privacy of individuals in a variety of contexts thanks to this setup, which guarantees that the model will perform successfully and efficiently in real-time scenarios.

In general, the anonymization framework pipeline that we have developed is intended to offer robust and efficient protection of individuals' privacy when it comes to real-time videos. We are able to accomplish this goal with a high degree of precision and productivity because of the combination of algorithms for face detection and de-identification.

4.2 Face Detection

4.2.1 YOLO in Face Detection

Progressive Advancements of YOLO in Face Detection
Face detection is a challenging task in computer vision due to the small size and complex structure of the face region. However, YOLO [43] has shown significant progress in this field over the years.

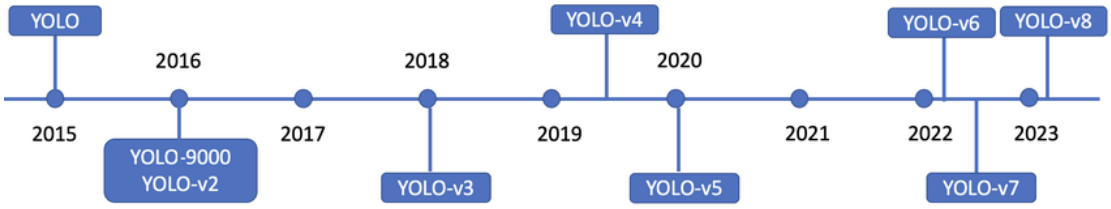


Figure 4.2.1: Timeline of YOLO variant [44].

YOLO is a very efficient method for identifying objects. It is a single-stage detection model that can recognize several items inside an image and categorize them with a high level of precision while doing so in real-time. When in detection mode, YOLO will only make one forward pass through the network, which is the basis of its neural network architecture known as Darknet.

There are five different iterations of the YOLO acronym: YOLOv1 [43], YOLOv2 [45], YOLOv3 [25], and YOLOv4 [46]. The very first version of YOLO was called YOLOv1, and it represented a major step forward in the field of object identification. On the other hand, it had several deficiencies in terms of accuracy as well as

faults in localization and that was the reason that it struggled with face detection. In YOLOv2, several of the problems that had been present in YOLOv1 were fixed with improved performance, higher input image size, and batch normalization, and in YOLOv3, a considerable improvement was achieved over YOLOv2 with the addition of various new features that made it more accurate and resilient. It introduced multi-scale features (FPN) [47] prediction, a superior backbone network (Darknet53), the substitution of the binary cross-entropy loss for the softmax classification loss, feature extraction, and advanced data augmentation techniques. These features made it more accurate and robust in face detection tasks. YOLOv4, the newest and most advanced version of YOLO, features a variety of enhancements such as CSPDarknet-53 as the backbone network, Spatial Pyramid Pooling, and Mish activation function that, when compared to earlier versions, resulting in a considerable improvement in the program's overall performance.

To be more specific, YOLO has made considerable strides in terms of face detection during the course of its development. YOLO's earlier iterations had difficulty recognizing faces because of the constrained space and intricate nature of the facial region in the image. However, YOLOv3 and YOLOv4 have shown impressive accuracy in face detection, and they have been successfully implemented in a variety of applications, including security surveillance systems and technology for facial recognition.

The most recent group of researchers, who came up with the YOLOv4 algorithm and released it in 2020, examined a vast number of potential outcomes for every facet of the YOLOv3 algorithm, including the backbone and what they referred to as "bags of freebies" and "bags of specials." Using a Tesla V100, YOLOv4 was able to achieve 43.5% AP (65.7% AP50) for the MS-COCO dataset while maintaining a real-time frame rate of 65 FPS.

To summarize, Yolo is now one of the most advanced object identification models that can be used, and with each new edition, it becomes better in terms of both its functionality and its appearance. Its speed and accuracy make it a great instrument in the area of computer vision, and its improvement in face identification has made it a vital tool in a variety of applications. Its speed and precision make it a valuable tool in the field of computer vision.

4.2.2 YOLOv5

YOLOv5: Architecture, Advancements, and Applications in Privacy Enhancement for Image and Video Analysis
YOLOv5 was released a month after YOLOv4. Both tackled the problem of grid sensitivity and can now detect easily bounding boxes having center points in the edges. Compared to the previous versions of YOLO, in YOLOv5, the model size was significantly reduced. Although it functions similarly to YOLOv4, it is often faster and lighter than versions released before.

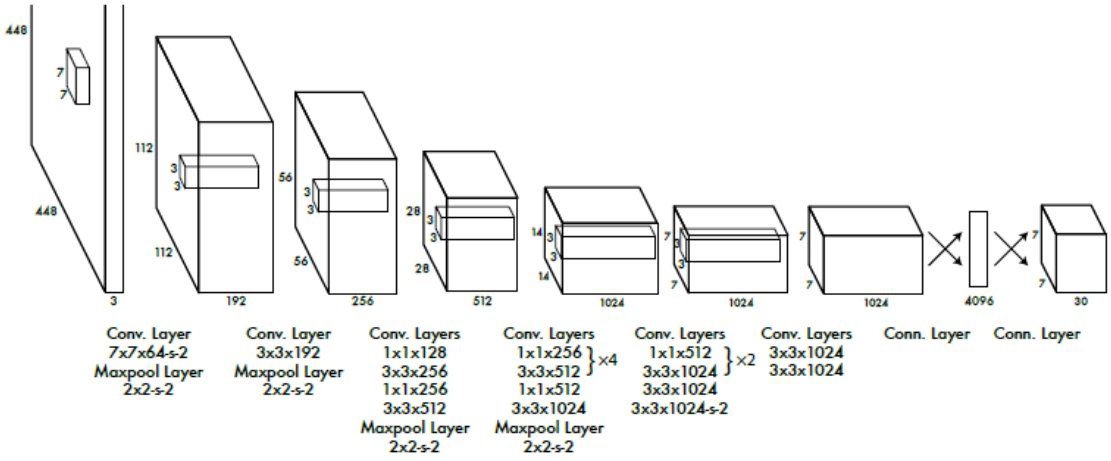


Figure 4.2.2: YOLO Architecture from the original paper [48].The architecture works as follows:
Resizes the input image into 448x448 before going through the convolutional network. A 1x1 convolution is first applied to reduce the number of channels, which is then followed by a 3x3 convolution to generate a cuboidal output. The activation function under the hood is ReLU, except for the final layer, which uses a linear activation function. Some additional techniques, such as batch normalization and dropout, respectively regularize the model and prevent it from overfitting.

Figure 4.2.3: Sample of detected faces using YOLOv5.

4.2.3 YOLOv5-Excel over Previous Versions of YOLO

YOLOv5 features a number of enhancements over earlier iterations of the software, namely YOLOv4, and YOLOv3, making it a superior option for face identification in security cameras. Here are some of the main benefits of YOLOv5 over earlier YOLO versions:

- Improved accuracy: On a number of object identification benchmarks, including COCO and PASCAL VOC, YOLOv5 achieves state-of-the-art precision. This is partly because YOLOv5 is able to extract more precise characteristics from pictures thanks to the implementation of a more effective backbone network and neck network.

- Faster inference: YOLOv5 is significantly quicker than the earlier versions of YOLO, making it an excellent choice for real-time applications like security cameras. This speed is made possible through YOLOv5's utilization of a

backbone network and head network that are both more effective, in addition to a number of other enhancements.

- Smaller model size: When compared to earlier versions of YOLO, the model size of YOLOv5 is more compact, which makes it simpler to implement on devices with limited resources, such as surveillance cameras.

- Improved handling of small objects: When compared to earlier iterations of YOLO, the most recent version, YOLOv5, is more adept than its predecessors at locating small objects. This is because YOLOv5 makes use of a redesigned anchor box method, which makes it more capable of dealing with objects of a smaller size.

- Improved generalization: In comparison to earlier iterations of YOLO, the most recent version, YOLOv5, is superior when it comes to generalizing to new object classes and environments. This is in part because of the utilization of a more effective backbone network and neck network, both of which enable YOLOv5 to extract more generic information from pictures.

Due to these benefits, YOLOv5 is a good option to consider when it comes to face detection in security cameras. Because of its increased accuracy and its ability to handle small objects, it is well-suited for recognizing faces in congested surroundings. Additionally, because of its quicker inference and lower model size, it is easier to deploy on devices with limited resources. In addition, as a result of the increased generalization, it is now capable of detecting faces in a wider range of settings, including those with varying degrees of illumination. But consider that it is all about in comparison with the previous versions of YOLO! In continuing we will see how YOLOv7 outperforms YOLOv5 in our work.

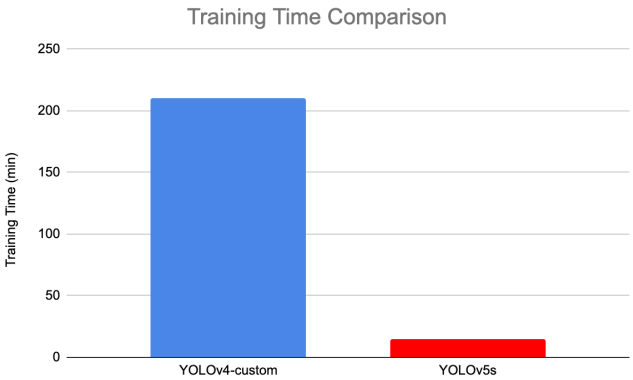


Figure 4.2.4: Training time comparison between YOLOv5 and YOLOv4 [49].

In addition to the above-mentioned benefits of using YOLOv5 in comparison with the previous versions of YOLO in our work is the **running environment** of YOLOv5 which **PyTorch**. One of the reasons that I started to work with YOLOv5 for my project is thanks to running YOLOv5 in PyTorch, it provides several benefits (more efficient to develop and deploy) for face detection specifically in surveillance cameras:

- Customization for face detection: PyTorch lets users adapt YOLOv5 for security camera face identification. Users can vary loss functions or anchor box sizes to better recognize faces of varied sizes.
- Improved accuracy: PyTorch lets users experiment with model architectures and hyperparameters to enhance face identification. Faster iteration and experimentation can improve performance.
- Real-time performance: PyTorch's efficient YOLOv5 implementation enables real-time security camera face identification. The efficient use of memory and compute resources allows real-time face identification in video streams with low latency.
- Easy deployment: PyTorch simplifies surveillance camera face identification using YOLOv5 models. TorchScript and ONNX let users transform PyTorch models to formats that work on mobile and embedded devices.
- Transfer learning: PyTorch simplifies pre-trained model transfer learning. This lets users utilize pre-trained YOLOv5 models learned on big datasets like COCO and fine-tune them for security camera face identification using smaller datasets.

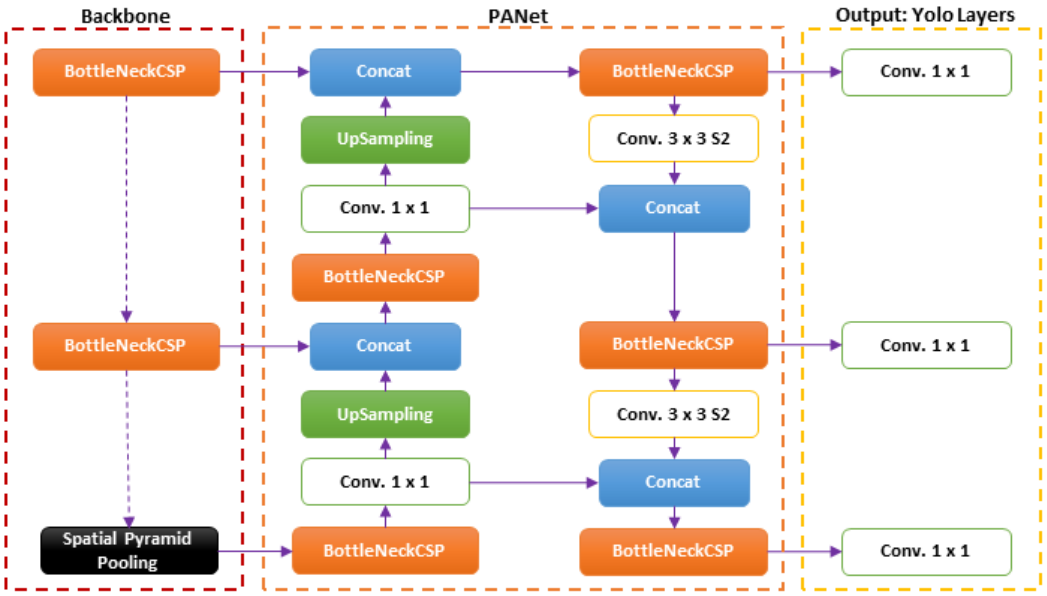


Figure 4.2.5: The network architecture of Yolov5. It consists of three parts: (1) Backbone: CSPDarknet, (2) Neck: PANet, and (3) Head: Yolo Layer. The data are first input to CSPDarknet for feature extraction and then fed to PANet for feature fusion. Finally, Yolo Layer outputs detection results (class, score, location, size) [50].

4.2.4 YOLOv5-Architecture

The architecture of YOLOv5 is based on a deep convolutional neural network (CNN) that has a backbone of CSPDarknet53 layers followed by a YOLOv5 head.

The CSPDarknet53 architecture is a modification of the Darknet architecture that uses a cross-stage partial connection to reduce the number of parameters and improve the accuracy of the model. The YOLOv5 head is a lightweight network that consists of convolutional layers, upsampling layers, and detection layers. The architecture of YOLOv5 is depicted in 4.2.5.
As you see in 4.2.5, the architecture is divided into three main parts: a backbone network, a neck network, and a head network. Here are the details of each part:

- **Backbone network:** The task of separating out features from the input image falls under the purview of the backbone network. The CSPDarknet backbone is used in its modified form in YOLOv5, which consists of a succession of convolutional layers followed by residual connections. Because it has been programmed to be both deep and efficient, the backbone network is able to extract meaningful features from the input image while simultaneously keeping the model size to a minimum.

- **Neck network:** The neck network is in charge of fusing data from many scales into a single representation in order to increase the accuracy of object recognition. The neck network in YOLOv5 is a feature pyramid network (FPN), which comprises lateral connections that incorporate characteristics from various levels of the backbone network. This type of network was chosen because of its flexibility. YOLOv5 is able to identify objects at a variety of sizes and resolutions as a result of this, which makes it more resilient to differences in object size.

- **Head network:** Based on the characteristics that the backbone network and the neck network have retrieved, it is the responsibility of the head network to forecast item bounding boxes and class probabilities. The YOLOv5 head network is based on a modified version of the YOLOv3 head network. This network is comprised of numerous convolutional layers that predict item bounding boxes and class probabilities. In addition, anchor boxes are utilized by the head network in order to enhance the precision of object detection.

In addition to these three primary components, YOLOv5 makes use of a variety of methods to enhance the precision and effectiveness of its object identification capabilities. For instance, YOLOv5 improves the effectiveness of the backbone network by utilizing cross-stage partial connections (CSP), and it improves the accuracy of object recognition at varying sizes by utilizing multi-scale training and testing.

In general, the architecture of YOLOv5 is meant to be efficient and precise, which enables it to recognize objects in photos and videos with a high level of precision while maintaining a speed that is close to real-time. It is one of the most successful object identification algorithms that are currently accessible because of the collaborative efforts of its backbone, neck, and head networks, which work together to extract features, fuse information, and create predictions.

Going through more details, we know that in any deep learning model choosing an **activation function** is crucial. For YOLOv5 the authors went with SiLU and Sigmoid activation function. In 4.2.6 you will see the performance of SiLU vs ReLU.

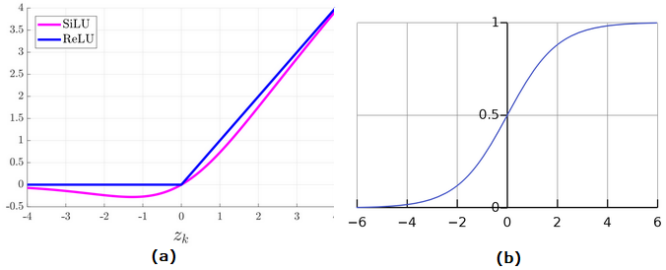


Figure 4.2.6: Graph depicting the activation functions that are utilized by YOLOv5. (a) The graph of the SiLU function. (b) A graph of the sigmoid function. The Sigmoid Linear Unit, commonly known as the Swish Activation Function, is an abbreviation for the Sigmoid Linear Unit. Convolutional procedures, which are employed in the hidden layers, have been performed using it. Although the Sigmoid activation function was applied to the convolutional operations carried out in the output layer, the results were not satisfactory [49].

In terms of **loss function**, the classes of the detected objects, the bounding boxes of those items, and their objectness scores are the three outputs that YOLOv5 provides. Therefore, in order to compute the class loss and the objectness loss, it uses BCE, which stands for binary cross entropy. While CIoU loss, which stands for complete intersection over union loss, is used to compute location loss. The following equation provides a formula for determining the total amount of loss:

$$Loss = \lambda_1 L_{cls} + \lambda_2 L_{obj} + \lambda_3 L_{loc}. \tag{4.1}$$

4.2.5 YOLOv5-Implementing in Anonymization Processes: Enhancing Privacy in Image and Video Analysis

We use the base YOLOv5 object detector [8] as one of the methods for the face detection phase by dividing images into a grid system. Each cell in the grid is responsible for detecting objects within itself. The network architecture of the YOLOv5 face detector includes the model Backbone, model Neck, and model Head as YOLO layers. A problem in this process is altering the hyperparameters to improve the accuracy and efficiency of our model. In the end, we test our model only over the varied test images in different conditions, as we mentioned earlier. In the most recent attempt, a batch size of 32 over 60 epochs was run in 3 hours, and the .yaml file was used for configuration. To optimize the model, we used an SGD optimizer with a learning rate (lr) of 0.1. We train this model with the WIDER Face dataset. In the experiment results, we will see that a weak point of YOLOv5 is its ability to detect small objects that are closed together, far from the camera, or faces in occlusion. The performance of face anonymization in real-time applications, such as those we will deploy in edge devices in the future, will be impacted by this flaw.

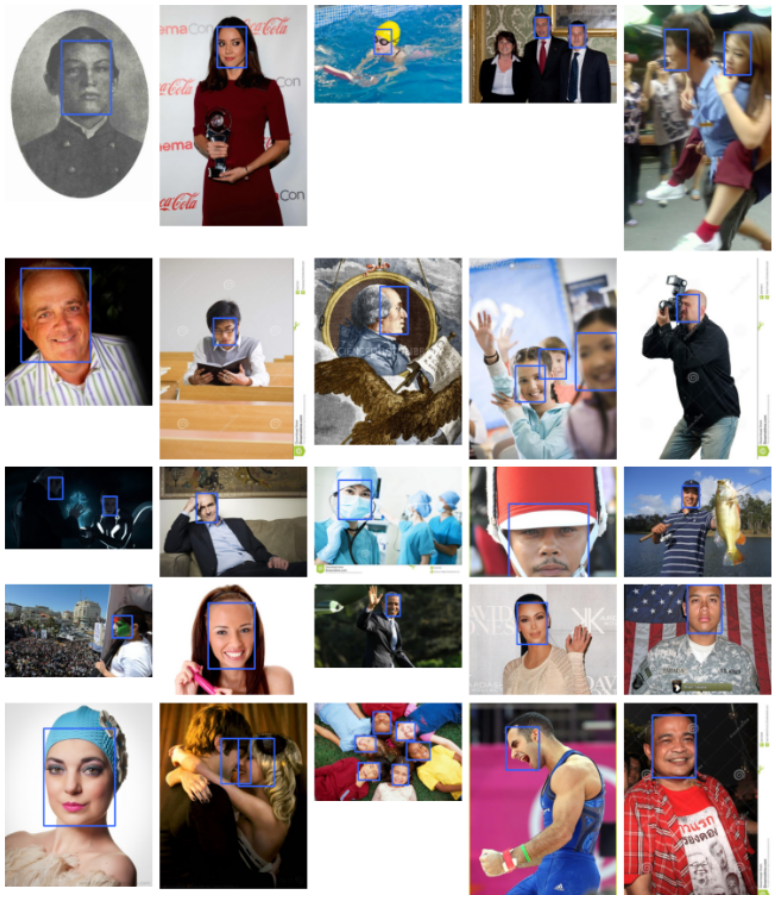


Figure 4.2.7: Face detection by YOLOv5. It is not working well in detecting very small faces in crowded areas. However, its performance is not terrible in a variety of lighting conditions.

In summary, we became acquainted with YOLOv5 and its advantages in face detection compared to previous versions of YOLO. We learned about the architecture of YOLOv5 and how to use PyTorch to run it on our own computers. By adjusting the model's hyperparameters, we were able to improve its performance on the WiderFace dataset. Through this process, we gained insight into the strengths and limitations of YOLOv5 for our specific goal of face anonymization which will be described more in the experimental results section.

4.2.6 YOLOv7

What is the most powerful object detection algorithm?

The YOLOv7 algorithm is making big waves in the computer vision, and machine learning communities and was released in July 2022 [7].

When it comes to real-time object detection for computer vision applications, YOLOv7 is the model that is both the **quickest** and **most accurate**. Chien-Yao Wang, Alexey Bochkovskiy, and Hong-Yuan Mark Liao were the authors of the official YOLOv7 study, which was titled "YOLOv7: Trainable Bag-of-Freebies Sets New State-of-the-Art for Real-Time Object Detectors." This paper was published

in July 2022. Within a few short days, the YOLOv7 research paper has amassed an incredible amount of interest among readers. The source code was made available to the public as open source and distributed under the GPL-3.0 license, a free copyleft license. It may be located on the official YOLOv7 GitHub repository, which received over 4.3 thousand stars in the first month following its release. In addition, the YOLOv7 document has an appendix that is fully comprehensive.

With this brief introduction to YOLOv7, let's go a bit deeper into its architecture and see its new benefits, which suit us in this project.

Figure 4.2.8: Real-Time Object Detection Inference in Python with YOLOv7 [51]

4.2.7 YOLOv7-What makes it different?

The primary objective of the authors of YOLOv7 was to advance the state of the art in object detection by developing a network architecture that could properly predict bounding boxes at inference speeds comparable to those of its competitors while maintaining the same level of precision.

4.2.7.1 E-ELAN (Extended Efficient Layer Aggregation Network)

The performance of the convolutional layers that make up the backbone of YOLO networks plays a critical part in determining how quickly and effectively the network can make inferences. Through the creation of cross-stage partial networks, WongKinYiu was the first person to begin investigating ways to significantly improve the effectiveness of these layers.

Optimizing many aspects, such as the number of parameters, computational effort, and computational density of the model, is the main goal of building an efficient architecture. By investigating the effects of the input/output channel ratio, the number of branches in the architecture, and the element-wise operations on the speed of network inference, the VovNet model goes one step further.

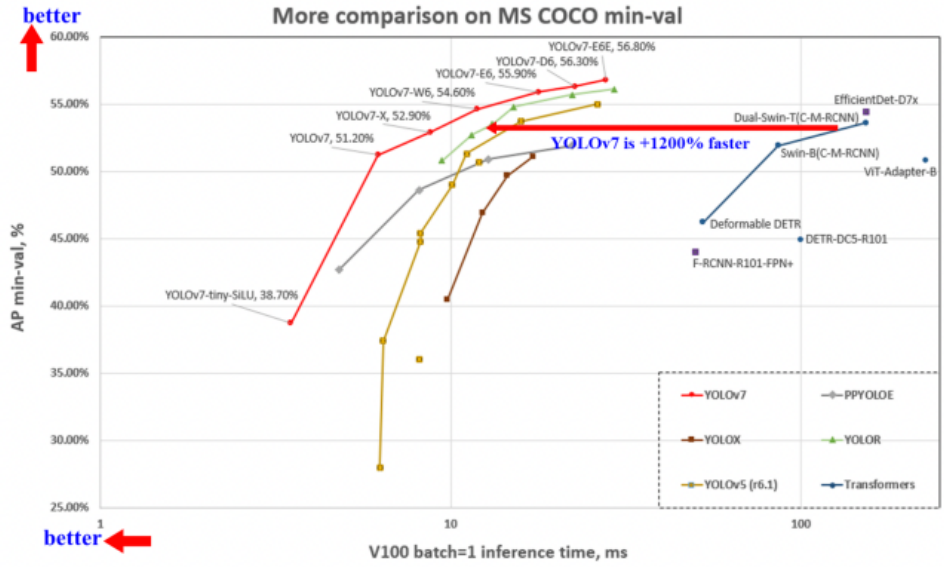


Figure 4.2.9: YOLOv7 evaluates to the upper left - it is quicker and more accurate than its peer networks. [52]

ELAN is the next major development in architecture search, and YOLOv7 builds on it by referring to it as E-ELAN. According to the ELAN paper's findings, a deeper network can successfully train and converge by managing the shortest and longest gradient paths.

The large-scale ELAN has stabilized despite the length of the gradient path and the stacking of computing blocks. This stable state, however, may be disturbed if an excessive number of computational blocks are stacked without restriction, leading to a decline in the rate of parameter use. Expand, shuffle, and merge cardinality algorithms are included in E-ELAN to overcome this issue and enable continual improvement of the network's learning capabilities without compromising the initial gradient path.

Only the architecture of the computational blocks is changed in E-ELAN; the architecture of the transition layer is left alone. E-ELAN employs group convolution as a strategy to increase the cardinality and channel of processing blocks. It applies the same channel multiplier and group parameter to every computational block inside a computational layer. The feature maps computed by each computational block are then concatenated after being divided into "g" groups based on the set group parameter "g". Currently, each group of feature maps has the same number of channels as there were in the original architecture. The feature maps from the "g" groups are then combined using merge cardinality.

The network can direct various groups of computing blocks to learn more diverse features by using the E-ELAN approach while keeping the original ELAN design architecture, improving its overall performance.

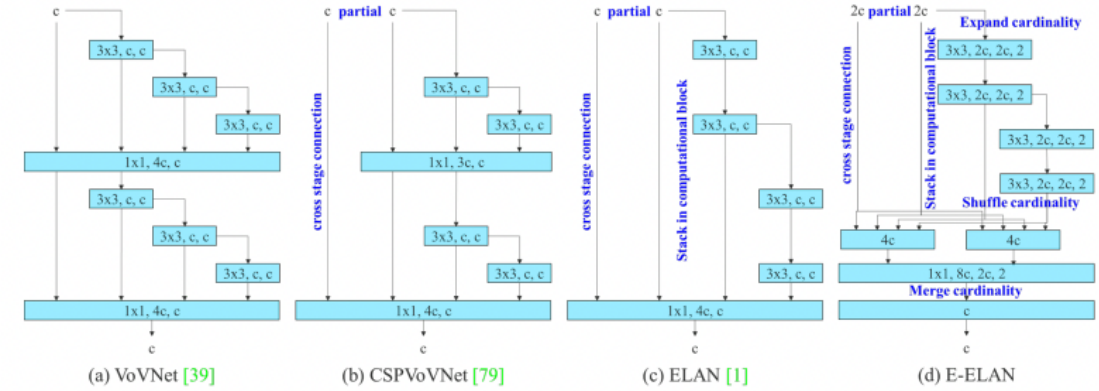


Figure 4.2.10: Extended networks with efficient layer aggregation. The proposed extended ELAN (E-ELAN) does not alter the gradient transmission path of the original architecture but uses group convolution to increase the cardinality of the added features and shuffle and merge cardinality to combine the features of distinct groups. This mode of operation can increase the features learned by various feature maps and the parameter and calculation usage. [7]

4.2.7.2 Model Scaling Techniques

Model scaling's primary goal is to modify certain model properties, resulting in models of various scales to meet various inference speed requirements. Scaling is accomplished using a significant Google architecture called EfficientNet by altering the model's breadth, depth, and resolution. However, additional studies looked into how group convolution and vanilla convolution affected parameter count and computational load during the scaling process.

Concatenation-based architectures might not be a good fit for EfficientNet's scaling method. The intensity of a transition layer immediately following a concatenation-based computational block will either rise or decrease when the depth of the model is scaled up or down. For instance, changing the depth factor changes the ratio between a transition layer's input and output channels, which may result in less hardware being used. To solve this, it is important to determine the corresponding change in the output channel of a computational block while increasing the depth factor of that block. The model's ideal structure is then preserved by applying this adjustment to the transition layers using width factor scaling.

While maintaining an ideal structure, the YOLOv7 compound scaling method successfully maintains the model's original design attributes. The results of using the width factor scaling methodology with the proper modifications in transition layers are shown in the figure below, which is consistent with the underlying principles of the suggested method.

YOLOv7 makes efficient scale adjustments while maintaining the model's original design characteristics, thus enabling optimal performance.

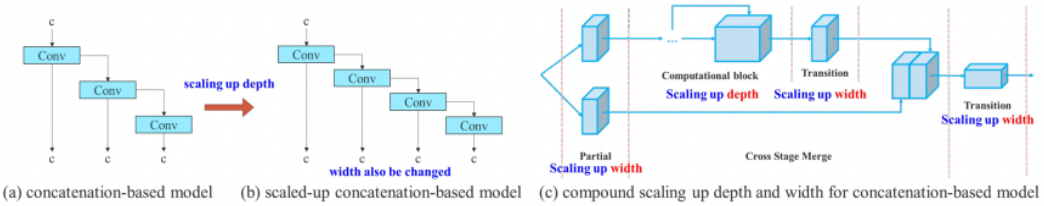


Figure 4.2.11: Model scalability for concatenated models. We observe, from (a) to (b), that when depth scaling is conducted on concatenation-based models, the output width of a computational block increases as well. This phenomenon will result in an increase in the input breadth of the subsequent transmission layer. Therefore, we propose (c), which states that, when conducting model scaling on concatenation-based models, only the depth in a computational block needs to be scaled, and the remainder of the transmission layer is scaled with corresponding width scaling. [7]

4.2.7.3 Trainable bag-of-freebies

Planned re-parameterized convolution
Re-parameterization approaches are crucial for improving a model's robustness by averaging a group of model weights. By using this strategy, the model is better equipped to identify and generalize patterns. Module-level re-parameterization, in which various components or modules of the network apply their own re-parameterization procedures customized to their particular characteristics and requirements, has received increasing attention in recent studies.

The authors of YOLOv7 carefully identify the network modules that would benefit from re-parameterization techniques by using gradient flow propagation channels. They gather knowledge about which modules would benefit from such strategies and which modules can be left undisturbed by looking at the channels via which gradients travel during the training process.

With the aid of this technique, the YOLOv7 writers are able to decide which modules in the network should receive a selective application of re-parameterization strategies. They maximize the effectiveness and overall performance of the YOLOv7 model by identifying the modules that need more robustness and changing the re-parameterization approaches accordingly.

The YOLOv7 authors make sure that the re-parameterization procedures are targeted and used where they will be most helpful by utilizing gradient flow propagation pathways. By using this strategy, the network is able to balance flexibility with computational effectiveness, creating an optimal and reliable model that is excellent at identifying and comprehending the patterns it is intended to handle.

Auxiliary Head Coarse-to-Fine
Deep supervision is a method that is frequently applied when deep networks are being trained. The shallow network weights using assistant loss as the guidance, and the main idea is to increase the number of auxiliary heads in the middle levels of the network. Deep supervision can nevertheless considerably enhance the

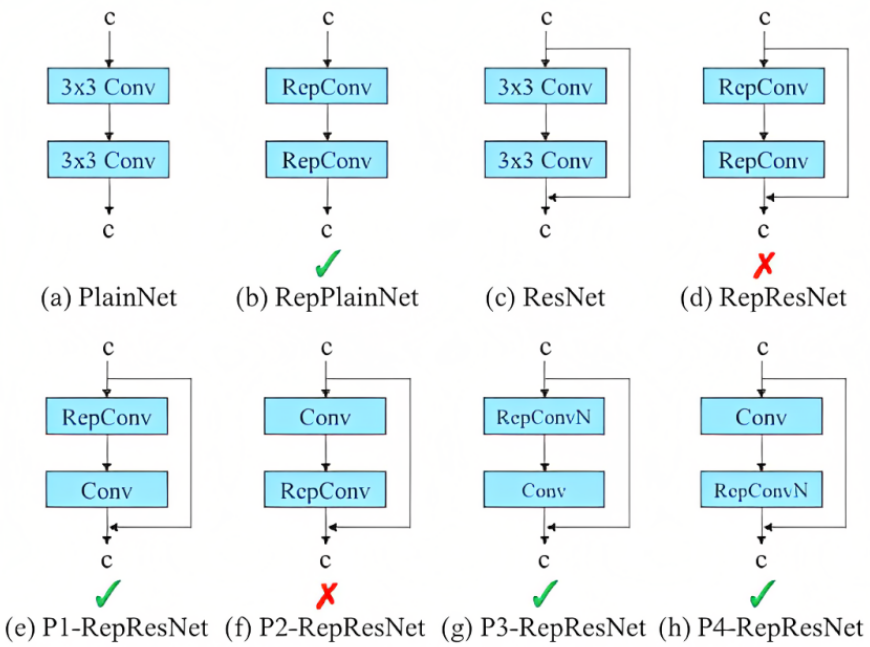


Figure 4.2.12: Planned re-parameterized model. In the planned re-parameterized model, we discovered that a layer with individual or concatenation connections should not have an identity connection for its RepConv. RepConvN, which contains no identity connections, can be substituted in these circumstances. [7]

model's performance on many tasks, even for designs like ResNet and DenseNet that often converge well. The object detector architecture is depicted below in both its "without" and "with" deep supervision states. In the YOLOv7 architecture, the lead head is in charge of producing the output, and the auxiliary head is in charge of assisting in training.

In the past, label assignment during deep network training typically made direct reference to the ground truth and produced hard labels in accordance with the prescribed rules. However, if we use object identification as an example, researchers now frequently include the quality and distribution of the network's prediction output in addition to the ground truth when using some calculation and optimization approaches to produce a trustworthy soft label.

However, here is a question: "How can a soft label be assigned to the auxiliary head and the lead head?" YOLOv7 makes use of lead head prediction as a kind of guidance in order to generate coarse-to-fine hierarchical labels. These labels are then used for auxiliary head learning and lead head learning, respectively. The graphic that may be found below depicts the two different recommended ways for assigning the deep supervision label.

Lead head guided label assigned: By allowing the shallower auxiliary head to directly learn the information that the lead head has learned, the lead head will be better able to concentrate on acquiring residual information that has not yet been taught.

Coarse-to-fine lead head guided label assigned: The relevance of the fine label and the coarse label may be constantly modified throughout the learning process thanks to this method, which also ensures that the optimizable upper bound of the fine label is always greater than that of the coarse label.

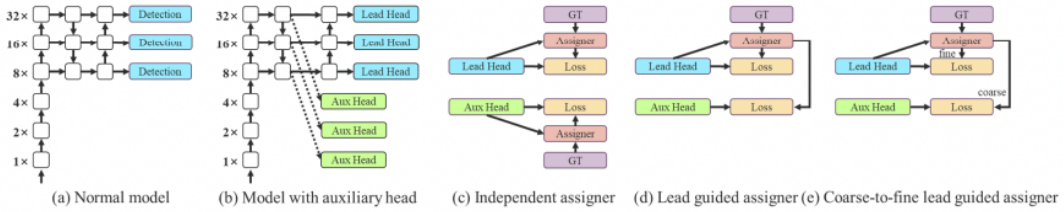


Figure 4.2.13: Coarse for auxiliary and fine for the assigner of lead head labels. Unlike the normal model (a), (b) contains an auxiliary cranium. We propose (d) lead head guided label assigner and (e) coarse-to-fine lead head guided label assigner as alternatives to the standard independent label assigner (c). The proposed label assigner is optimized by lead head prediction and the ground truth to simultaneously assign labels to the training lead head and auxiliary head. [7]

To conclude, in the range of 5 FPS to 160 FPS, YOLOv7 outperforms all other known object detectors in terms of speed and accuracy, and on GPU V100, it has the greatest accuracy of 56.8% AP of all real-time object detectors with 30 FPS or more. Both the transformer-based SWINL Cascade-Mask R-CNN (9.2 FPS A100, 53.9% AP) and the convolutional-based ConvNeXt-XL Cascade-Mask R-CNN (8.6 FPS A100, 55.2% AP) are outperformed by the YOLOv7-E6 object detector (56 FPS V100, 55.9% AP), which performs 509% faster and

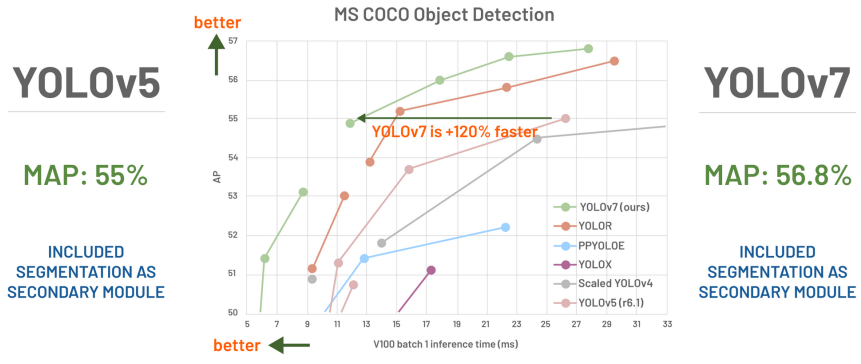


Figure 4.2.14: YOLOv5 and YOLOv7 Accuracy Comparison. In terms of accuracy, YOLOv7 is superior to YOLOv5. On the COCO dataset for YOLOv5, the MAP (mean average precision) is 55.0%, and for YOLov7, it is 56.8%. According to research, the shift was brought about by cutting the parameters by 35–40% and the computations for each (normal and embedded systems) by half. [53]

YOLOv5	YOLOv7
Fast processing in training on custom dataset	Slow processing in training on custom dataset
Fast inference on CPU systems	Slow inference on CPU systems
Stable memory utilization during training on custom data	Unstable memory utilization during training on custom data
Fast Inference speed on normal GPU's (i.e., Quadro P2200, Nvidia RTX 1650)	Fast Inference speed on latest GPU's (i.e., Nvidia RTX 3090, Tesla A100)
Good Accuracy (less than YOLOv7)	Good Accuracy (better than YOLOv5)
Used darknet backbone with cross-stage partial network	Used E-Elan (Extended Efficient layer aggregation networks) backbone
Developed in Python with PyTorch framework	Developed in Python with PyTorch framework
Provided support, for instance, segmentation as a secondary module	Provided support, for instance, segmentation as a secondary module
Provided supported for image classification as a secondary module	Provided support for image classification as a secondary module
Single stage Detector	Single Stage Detector
Used less floating-point operations (less computational)	used more floating-point operations (More computational)
Provided support for multi-GPU's training	Provided support for Multi GPU's training
Automatic hyperparameters optimization using genetic algorithm	Automatic hyperparameters optimization using genetic algorithm

Table 4.2.1: General comparison. YOLOv5 vs YOLOv7.

4.2.8 YOLOv7-An Efficient Architecture for Face Detection and Anonymization in Security Cameras

The YOLOv7 architecture requires 45% fewer parameters compared to YOLOv5 and 63% less computation while achieving a 47% faster inference speed [7]. Due to the points mentioned above, we invested more time in YOLOv7. The experimental results will provide more practical differences between these two models. You can find the general comparison between YOLOv5 and YOLOv7 in table 4.2.1.

Another critical factor supporting our decision to use YOLOv7 in this study is its effective use of security cameras to recognize and track objects in a specific area. Therefore, we may instantaneously anonymize faces after detection using anonymization techniques.

4.2.9 YOLOv7-Implementing in Anonymization Processes: The Most Powerful Face Detection Algorithm

Figure 4.2.15: Sample screen-shot of face detection using YOLOv7 on video.

In this study, we used the pre-trained YOLOv7, which is trained on thousands of image data, and then we re-trained it on our three independent datasets.
In detail, for the re-training phase, we first gather data and store it in the specie structure based on YOLOv7 on a drive or local disc. Next, labels and bounding boxes (bbox) are converted to the YOLOv7 format. Indeed, in contrast to our

Figure 4.2.16: Sample of face detection using YOLOv7 on image.

Parameters\Dataset	WIDER Face	CelebA	UFDD
batch_size	32	24	22
lr0	0.01	0.025	0.01
lrf	0.15	0.25	0.2
img_size	640	640	640

Table 4.2.2: Hyperparameters to train YOLOv7.

data bbox format in xyxy and xywh, YOLOv7 only supports a few special bboxes. So, to comply with YOLOv7, we had to modify all labels to the appropriate format. We iteratively trained our model by changing hyper-parameters like batch sizes. We looked at the performance data for resistance training to adjust the parameters again and get the best model results. Our best model for the WIDER Face data set is compatible with batch sizes of 32, image sizes of 640 x 640, lr of 0.15, and over 60 epochs. The runtime is 4.769 hours, which is not bad. The hyperparameter details are available in Table 4.2.2. In inference, it can quickly get a pre-trained model to make predictions on given images or videos, storing or displaying the results, even for a live video using the webcam. We can also control the anonymization processes by selecting the methods and levels of anonymization that suit us.

Thus, our face detection model can detect faces in images and videos and deliver bounding boxes with pixel information for each face. This pixel information can then modify the image by, for example, blurring out the respective areas or masking them. This is what we call face-anonymization and will describe more in the following Section 4.3.

4.3 Face Anonymization

This section provides in-depth details of face anonymization. We often utilize face detection in several research applications, typically as the first step in a face recognition pipeline, but what if we wanted to do the "opposite" of face recognition? What if we instead wanted to anonymize the face by, for example, blurring it, thereby making it impossible to identify the face? An example comparison is illustrated in Fig. 4.3.1 for face anonymization.

(a) (b)

Figure 4.3.1: Visual performance comparison of face anonymization based (a) YOLOv7, and (b) YOLOv5.

As we indicated above, anonymization may be used to increase user privacy or filter out images containing people for regulated processes. It can also be used for downstream processes that do not require personally identifiable information.

Knowing the primary objective of face anonymization, we tended to anonymize faces in a variety of situations, including various weather conditions like rain, snow, haze, day, and night, as well as various positions of individuals in each image or frame, such as in occlusion or motion stance. The second factor we consider in this research is the head size of persons, which is associated with their proximity to the camera. In this work, all of these actual situations can be identified and anonymized. As a result, our primary goal and effort were to conceal identities in any situation. Our workflow makes it straightforward to detect faces, after which we can quickly apply one of the following techniques to the bounding box of recognized faces. In other words, we can use these anonymization methods as an extension module for all faces that have been discovered. By the way, Our real-time video anonymization outcomes in this procedure heavily rely on our dataset variety and the face identification algorithm. Figure 4.0.1 demonstrates our workflow to do anonymization.

We utilized traditional anonymization methods such as image blurring, pixelation, and black masking that can remove valuable information and are fast enough to fulfill our goal.

4.3.1 Blur

Face blurring is a computer vision method used to anonymize faces in images and video. Indeed, we used facial blurring to help protect a person's identity in an image or video. Figure 4.3.2 (a) illustrates face blurring anonymization after face detection by YOLOv7; note how the face is covered and the person's identity is unrecognizable.
Indeed, in order to automatically detect and blur faces, the first step is to determine whether or not a face is present in each image or video; for this study, we are using YOLOv5 or YOLOv7 to make this determination. In a generic sense,

(a) (b)

Figure 4.3.2: Face anonymization (a) School (b) Snow.

the region of interest refers to the section of an image that is selected for a particular reason. Our face detection method generates the face bounding box, which is then used in the face blurring process. These coordinates (bounding box) have to include the starting and ending points of the face both in the x and y directions.

After identifying the region of interest in an image, here a person's face, the next step in the process of face anonymization is to blur it. There are various methods that can be employed to blur the detected face, with one popular approach being the Gaussian blurring method.

The Gaussian blur technique involves reducing the level of detail in an image to create a blurred effect. This is achieved by combining a Gaussian FIR kernel with the original image. A Gaussian kernel is essentially a matrix of numerical values that defines the shape and size of the blur. The kernel is centered on each pixel in the image, and the values in the kernel are multiplied by the corresponding pixel values in the image. The resulting values are then added together to create the new pixel value for that location in the blurred image.
To apply the Gaussian blur, a common method is to use a two-dimensional kernel that is applied in both the horizontal and vertical directions of the image. However, this approach can be computationally expensive. An alternative method is to use two one-dimensional kernels, one for the horizontal direction and one for the vertical direction. Applying two one-dimensional kernels has the same effect as a two-dimensional kernel but with less computational cost.

The next step is to overlay the blurred face back onto the original image after identifying and blurring the region of interest in the image, which is typically a person's face. To do this, the blurred face must be carefully positioned back in the image's original spot so that it blurs in and looks consistent with the background.

To do this, the original coordinates of the region of interest before it is blurred are carefully recorded. The blurred area is then returned to its original location in the picture using these coordinates. To guarantee that the blurred face is positioned appropriately and preserves its blurry look, the technique involves exact alignment and scaling.

It's critical to make sure that the overall image quality is maintained once the

(a) (b)

Figure 4.3.3: (a) Original image, (b) Face anonymization (blurring method) using YOLOv5.

blurred face has been reinserted into the original picture. To achieve a smooth transition between the blurred area and the remainder of the image, it may be necessary to modify the image's contrast, brightness, or color balance. The objective is to maintain the privacy and anonymity of the subject in the blurred area while producing a final photograph that seems natural and untouched.

A crucial phase in the face anonymization process is adjusting the blurred area into the original image. To make sure that the final result fulfills the appropriate criteria for privacy and image quality, it takes close attention to detail and a high degree of technical skill.

(a) (b)

Figure 4.3.4: (a) Face detection using YOLOv7, (b) Face anonymization in meeting occasion.

Therefore, referring to what we understood from the steps of face anonymization, to do this part, we used Gaussian blur. To be more specific, a Gaussian filter is a low-pass filter used to blur specific areas of an image and reduce noise (high-frequency components). To get the desired result, the filter is constructed as an odd-sized symmetric kernel (DIP version of a matrix) and passed through each pixel in the region of interest. The filter is convolved over the image, and the value of the center pixel is replaced by the average of all the pixels under the

kernel. A 3×3 Gaussian Kernel Approximation(two-dimensional) with Standard Deviation = 1 appears in the matrix below (see matrix 4.2). In this example, all filter elements have the same weight, giving the average of all 9 pixels. Size and weight have an impact on the filter's effectiveness.

$$\frac{1}{16}\begin{bmatrix}1 & 1 & 1\\ 1 & 1 & 1\\ 1 & 1 & 1\end{bmatrix} \tag{4.2}$$

The values inside the kernel are computed by the Gaussian function, which is as follows:

$$G(x,y) = \frac{1}{2\pi\sigma^2}e^{-\frac{x^2+y^2}{2\sigma^2}}. \tag{4.3}$$

Where, x → X coordinate value, y → Y coordinate value, and σ → Standard Deviation.
Here, we applied Gaussian blurring kernel size in the form of (height, width) with defaults value to (11, 11). The degree of blurring varies depending on the user or the issue.

To conclude, the Gaussian blurring method is widely used in face anonymization techniques because it effectively obscures the facial features of an individual while retaining the overall shape of the face. By reducing the level of detail in the image, it makes it more difficult to identify the person's face and maintain their privacy and anonymity.

4.3.2 Pixelation

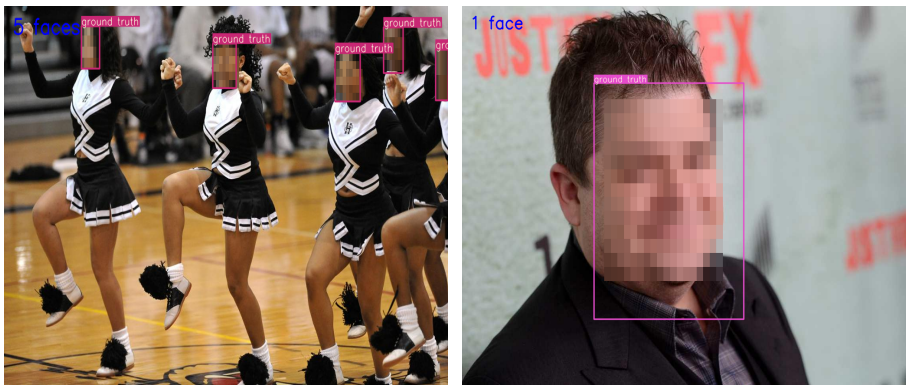


Figure 4.3.5: Face anonymization using pixelation method after face detection phase by YOLOv7.

Pixelation is a visual effect that can occur when raster or non-vector images are resized to a smaller size, such that individual pixels become too small to be perceived. This results in a loss of detail, causing the image to appear blurry or fuzzy. Essentially, pixelation is a technique used to reduce the amount of information in an image by decreasing the number of pixels used.
To achieve pixelation, an image is divided into a grid of square-shaped blocks,

with each block consisting of a set number of pixels. The average color of each block is then calculated by adding together the colors of all the pixels within that block. This results in an image that appears to be composed of blocks of uniform color rather than detailed shapes and lines.

There are two main methods for resizing images to achieve pixelation: static and dynamic. With the static method, images are resized to a fixed size that is determined by the "pixelate size" parameter. On the other hand, with the dynamic method, images are resized by a specified ratio that is determined by the "pixelate-ratio" parameter. In our work, we default to the static method using a scale of (8,8). However, if dynamic pixelation is desired, we use a scaling ratio of 10. In Fig. 4.3.6, we can see the sample of anonymization by the pixelating filter.

$$I_P(x,y) = \frac{1}{b^2}\sum_{i=0}^{b-1}\sum_{j=0}^{b-1} I(\left\lfloor\frac{x}{b}\right\rfloor + \left\lfloor\frac{y}{b}\right\rfloor + j). \tag{4.4}$$

where x and y are the pixel coordinates and b is the block size.
To apply pixelation for anonymization purposes, images are typically divided into non-overlapping squares of size NxN, with N being a parameter that is manually adjusted based on the level of anonymity desired. All the pixels within each square are then replaced with the average color of that square. This technique is commonly used in situations where privacy concerns are paramount, such as in news broadcasts or medical research.

While pixelation can be an effective way to obscure details in an image, it is not foolproof. In some cases, contextual clues such as clothing or facial features can still be used to identify individuals. Additionally, advanced image processing techniques can potentially be used to reconstruct the original image. Despite these limitations, pixelation remains a popular and accessible technique for anonymizing images and videos.

4.3.3 Blackened

Figure 4.3.6: Face anonymization using blackened method after face detection phase by YOLOv7. Here also it is obvious that our detection is how accurate referring to the ground truth bounding box.

Another traditional face anonymization method is the Blackening of the ROI. This method is easy and very useful in this process. By replacing the ROI patches with black pixels, we can obscure faces and protect individuals' identities. The face segment will be masked with an intensity value (between 0 and 255). In our research, the default intensity value is 0 (black). However, depending on the background or the illumination of the image or video, you can vary this intensity value. The level of privacy protection is controlled by varying parameters in each method above.

In other words, this filter is likely created by applying the following formula to the image data, resulting in the desired blackening effect:

$$ImgBlackened = originalImg * (1 - \alpha). \tag{4.5}$$

with α representing the opacity, the bigger is α the stronger is the impact of the filter.

CHAPTER

FIVE

IMPLEMENTATION DETAILS

5.1 Dataset

Based on our main objective, each of the following datasets was chosen for our work. Our models were trained using each dataset on its own, and the experimental results will show you how obviously the intrinsics of each dataset can have an effect on the overall result.

In our research, we conducted a thorough examination of various datasets and carefully selected them based on their suitability for our primary goal, which is face anonymization under diverse conditions. Each dataset presents unique challenges and characteristics that allow us to evaluate the performance of our model in different scenarios, enabling us to compare and analyze the results.

It is worth noting that combining multiple datasets can sometimes yield the best outcomes as it provides a broader range of diversity and variability in the data. By incorporating different datasets into our analysis, we aim to enhance the robustness and generalization capabilities of our face anonymization model.

In our work, it was crucial for us to have a dataset that encompassed a wide range of conditions, including variations in lighting, poses, backgrounds, occlusions, and other factors that can affect face detection and anonymization. We also took into account the importance of resolution and scale, as these aspects influence the accuracy and effectiveness of our model in handling different image sizes and levels of detail.

Let's delve further into the specifics of the three datasets we selected and the reasons behind our choices. Each dataset offers unique advantages and aligns with our research objectives, allowing us to comprehensively evaluate the performance and efficacy of our face anonymization algorithm.

5.1.1 WIDER Face dataset

The WIDER Face dataset [9] is a benchmark dataset for face identification tasks and is generally acknowledged for its quality. It includes a substantial amount of

Figure 5.1.1: WIDER FACE: A Face Detection Benchmark. [9]

photos, particularly 32,203 images, with a total of 393,703 faces that have been tagged. The dataset contains a wide range of variations in facial features, including size, position, occlusion, expression, makeup, and illumination. As a result, it accurately reflects a variety of situations that may occur in the real world.

The dataset is broken up into 61 different event categories so that it can be utilized to its full potential. For each event class, a random selection method is used to determine how much of the data will be used for training, how much will be used for validation, and how much will be used for testing. This division makes it possible to conduct an all-encompassing analysis and comparison of face identification algorithms across a wide variety of event categories.

In this specific piece of research, the YOLOv7 model is trained using a portion of the dataset that consists of 12,880 pictures with a total of 158,230 faces contained inside them. This carefully chosen collection of images covers a wide range of scenarios, ensuring that the YOLOv7 model is presented with a variety of image types that require the use of anonymization methods. The scenarios are covered by this collection of photographs.

It is easier to build and evaluate effective methods for face detection and anonymization when one combines the wide and rich WIDER Face dataset with the training of the YOLOv7 model using a varied collection of photos. This combination is what makes the dataset so useful.

5.1.2 CelebA dataset

The second dataset that we exploited in this procedure was the large-scale CelebFaces Attributes (CelebA) dataset [10]. A massive face attributes collection, it includes over 200,000 photos of celebrities, each of which is manually annotated with 40 attribute labels, that describe various facial characteristics. These attributes include gender, age, presence of glasses, facial expression, and more. The attribute annotations provide valuable information for tasks such as face attribute analysis and facial recognition.

This photo album features a diverse selection of stances, many of which are juxtaposed against busy backdrops. CelebA has a vast diversity, a massive number, and

Figure 5.1.2: Large-scale CelebFaces Attributes (CelebA) Dataset. [10]

Datasets	WIDER Face		CelebA		UFDD	
	#Images	#Faces	#Images	#Faces	#Images	#Faces
Train	12880	158230	25229	25228	4111	6940
Validation	3224	39560	6249	6249	1028	1854
Test	-	-	10000	10000	1285	2100

Table 5.1.1: Data portion used in this work.

rich annotations thanks to its 10,177 identities, 202,599 face photos, 5 landmark locations, and 40 binary attribute annotations per image.

The portion of data we use in training, testing, and validation is in Table 5.1.1.

5.1.3 UFDD dataset

Figure 5.1.3: UFDD-Illumination situation.[11]

The Unconstrained Face Detection Dataset [11], also known as the UFDD, was selected to be the third dataset utilized in our study. This dataset is comprised of a collection of facial photos that contain circumstances that are not normally addressed in traditional facial image datasets. These problems include motion blur, focus blur, and a variety of others besides. Some of them are caused by the weather.

The selection of this difficult dataset was motivated by the significance of the aforementioned conditions in a variety of applications, including biometric surveillance, maritime surveillance, and long-range surveillance, all of which place a premium on accurate detection.

However, despite the solid justification we had for using the UFDD dataset, we were unable to achieve real-time anonymization due to the low number of photos that were accessible and the very few faces that were visible in each image. This was a hurdle for us. Despite this, the UFDD dataset continues to be useful for developing research in the field of unconstrained face identification. It provides a comprehensive testbed for testing and developing face detection algorithms, particularly in tough and realistic settings, which will ultimately enhance their accuracy and robustness in crucial surveillance application contexts.

Figure 5.1.4: UFDD-Rainy weather.[11]

5.2 Preprocessing

In the data preparation phase, our first step is to extract the dataset and then divide it into three subsets: training, testing, and validation. Following the YOLOv7 format, as outlined in [8], we need to convert the labels, which represent the bounding box positions, for each subset of the data. This conversion is crucial for ensuring compatibility with the YOLOv7 model.

After converting the bounding box positions to the YOLOv7 format, we save them as individual *.txt* files, following the specifications required by YOLOv7. Each *.txt* file is given the same filename as the corresponding image file and contains the converted bounding box positions along with the respective class label for that image. This file format facilitates the integration of the labeled bounding box information into the training pipeline of the YOLOv7 model.

For uniformity and consistency, the input size for all images is set to 640×640. This standardization ensures that the images are of a consistent size and aligns with the requirements of the YOLOv7 model.

Additionally, in the case of video test data, a crucial preprocessing step is to convert the video into individual frames. This is necessary because the subsequent

face detection and anonymization steps operate on individual images. Although the final output will be a processed video generated from these images, initially, we need to extract the frames from the input video.

To simplify the process of converting videos to frames, we employ a Python library package that automates this task. This package handles the extraction of frames from the video, allowing us to efficiently process each frame for subsequent face detection and anonymization steps. It is worth noting that while our model is capable of handling videos in various formats, it generally performs better with MP4 files that are no longer than 10 seconds.

By incorporating these data preprocessing steps, we ensure that the dataset is appropriately formatted and prepared for training and evaluating our face detection and anonymization model.

5.3 Evaluation Metrics

In computer vision, several evaluation metrics are commonly used to assess the performance of algorithms and models. These metrics provide quantitative measures to evaluate the accuracy, robustness, and generalization capabilities of computer vision systems. The following criteria are used to evaluate the utility of deep learning-based techniques in face detection and face anonymization that we used:

5.3.1 Evaluation on Face Detection

Precision

Precision is a fundamental evaluation metric used in computer vision to assess the accuracy and reliability of predictions. It quantifies the probability of the predicted bounding boxes accurately matching the actual ground truth boxes, thus measuring the positive predictive value. A high precision score indicates that a significant proportion of the detected faces align with the ground truth.

The precision score is calculated as the ratio of true positive predictions (correctly detected faces) to the sum of true positive and false positive predictions (incorrectly detected faces). This ratio represents the precision of the model's predictions and ranges between 0 and 1. A precision score of 1 implies that all detected faces perfectly match the ground truth, while a score of 0 signifies that none of the predicted bounding boxes align with the actual objects of interest.

$$Precision = \frac{TP}{TP + FP}. \tag{5.1}$$

TP stands for "true positives" in this equation, which means the prediction was as positive as it was correct. False positives (FP) are predictions that were incorrectly classified as positive.

A high precision score is desirable in many computer vision applications, especially those that require accurate object localization and identification. For instance, in face detection tasks, a high precision score indicates that the model successfully identifies and localizes faces without too many false positive detections. This is particularly crucial in scenarios where precision is paramount, such as surveillance systems or biometric applications where accurate face recognition is essential.

Achieving a high precision score requires a model to minimize false positive detections, avoiding the inclusion of irrelevant or non-existent objects in the predictions. It signifies that the model is providing accurate and reliable results, increasing confidence in the detected faces. Evaluating precision alongside other metrics such as recall (which measures the ability to detect all relevant objects) provides a comprehensive understanding of the model's performance, enabling researchers and practitioners to fine-tune and optimize their computer vision algorithms accordingly.

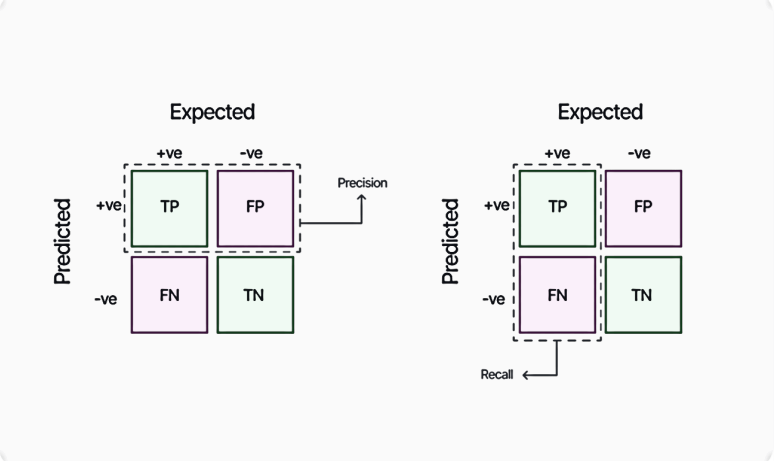


Figure 5.3.1: Precision-Recall.

Recall

Also known as sensitivity, is a crucial evaluation metric in face detection that assesses the effectiveness of a model in correctly identifying all the faces present in an image or a dataset. It measures the proportion of ground truth faces that are successfully detected by the model.

The recall score is calculated by dividing the number of true positive detections (correctly identified faces) by the sum of true positive detections and false negative detections (missed faces). This ratio represents the recall of the model's predictions and varies between 0 and 1. A high recall score implies that a significant proportion of the ground truth faces were successfully detected by the model, indicating a strong ability to capture most of the faces in the image.

$$Recall = \frac{TP}{TP + FN}. \tag{5.2}$$

FN stands for False Negative, which shows Failed to predict an object that was there.

In face detection tasks, achieving a high recall score is crucial, especially in applications where it is essential to capture all instances of faces accurately. For example, in surveillance systems or security applications, it is vital to identify as many faces as possible to ensure comprehensive monitoring and identification. A high recall score indicates that the model successfully captures a large portion of the faces, minimizing the number of missed or undetected faces.

However, it is important to note that optimizing recall often comes at the cost of increased false positive detections. In an attempt to capture as many faces as possible, the model may also detect non-face objects or artifacts incorrectly. Balancing recall with other metrics such as precision is necessary to ensure a well-performing face detection system. Precision measures the accuracy of the detected faces, focusing on minimizing false positive detections.

By considering both recall and precision, researchers and practitioners can evaluate the trade-off between detecting all relevant faces (high recall) and minimizing false positive detections (high precision). This comprehensive analysis allows for the development of face detection models that achieve a balance between sensitivity and accuracy, ensuring reliable and efficient face detection in various applications.

F1

To determine true positive and false positive face detection, the F1 score is the harmonic mean of the Precision and recall measures when compared to ground truth regions.

$$F1 = 2 \times \frac{Precision \times Recall}{Precision + Recall}. \tag{5.3}$$

mAP

The mean Average Precision (mAP) is a widely adopted metric for evaluating the accuracy of object detectors, including face detection systems. It provides a comprehensive assessment by calculating the area under the precision-recall curve. The precision-recall curve is generated by varying the decision threshold of the detector and plotting the corresponding precision and recall values.

To compute mAP, the Average Precision (AP) is first calculated for each individual class, representing the precision-recall performance for that class. These AP values are then averaged across all the classes, yielding the mean Average Precision (see Eq. 5.4). This approach allows for a holistic evaluation of the model's performance across different object classes.

$$mAP = \frac{1}{N} \sum_{i=1}^{N} AP_i. \tag{5.4}$$

In this study, we examined the performance of our face detection model using different ranges of decision thresholds. Specifically, we reported the mAP scores at a single threshold of 0.5 (mAP@.5) and across a range of thresholds from 0.5 to 0.95, with an increment of 0.05 (mAP@[.5:.95]) [54]. The latter provides an average mAP value over various IOU (Intersection over Union) thresholds, which measure the overlap between the predicted bounding box and the ground truth. By considering multiple IOU thresholds, we obtain a more comprehensive evaluation of the model's accuracy at different levels of bounding box overlap.

Figure 5.3.2: Detected faces vs Ground-truth. The effectiveness of YOLOv7 in accurately detecting faces within the specified ground truth bounding box is evident.

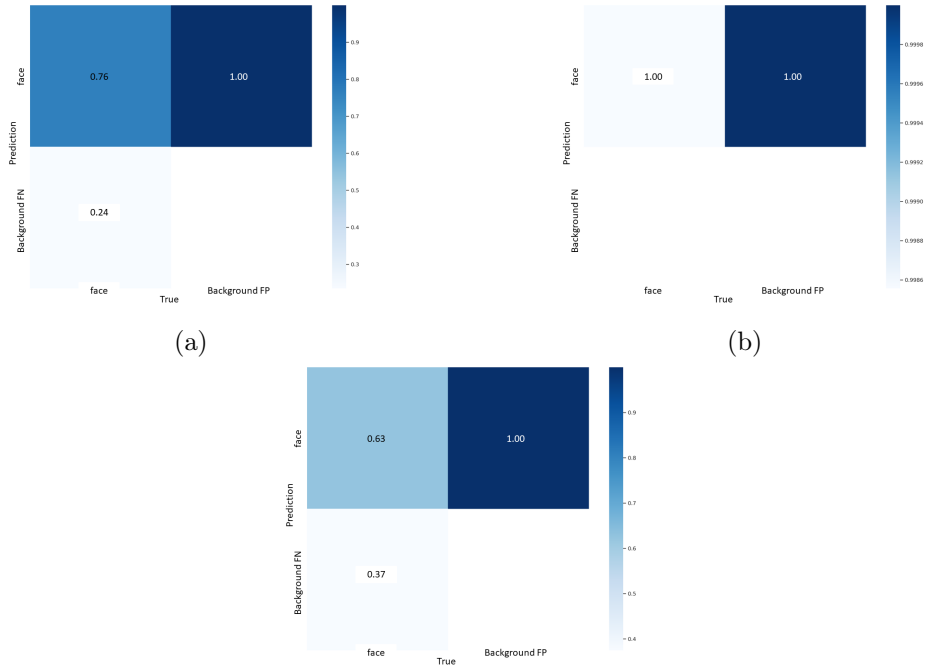
Based on our experiments, we set the IOU threshold at 0.45, indicating that a predicted bounding box must have a significant overlap with the ground truth to be considered a true positive detection. Additionally, we set the confidence threshold at 0.2, which serves as a measure of the model's confidence in its predictions. Bounding boxes with a confidence score below this threshold are considered false positives or low-confidence detections.

These threshold values are critical in face detection tasks as they directly impact the trade-off between detection sensitivity and precision. The IOU threshold ensures that detected faces have sufficient overlap with the ground truth, ensuring accurate localization. The confidence threshold helps filter out low-confidence or ambiguous detections, enhancing the precision of the system.

By analyzing the performance of our face detection model with respect to these thresholds, we can gain insights into its capability to accurately detect and localize faces while controlling the false positive rate. This assessment provides valuable information for optimizing the performance of the model in real-world face detection scenarios.

Confusion Matrix

We only use the Item Confusion Matrix to show how many faces are correctly detected in accordance with the ground truth and how many faces are incorrectly detected in the background (see the Confusion Matrix Fig. 5.3.3).



(a) (b)

(c)

Figure 5.3.3: CM of (a) WIDER Face, (b) CelebA, and (c) UFDD. The proportion of faces that were accurately identified as faces and those that were mistakenly identified as backgrounds.

5.3.2 Evaluation on Face Anonymization

FID

The Frechet Inception Distance score, or FID for short, is a statistic that determines how far feature vectors produced for real images are from feature vectors calculated for images with anonymous faces. FID score may still be used as a measure to assess our anonymization model even though it is often used to assess the quality of pictures produced by generative adversarial networks, and lower scores have been demonstrated to correspond strongly with higher-quality images. Higher FID ratings are associated with faces that have additional hiding techniques applied to them, such as blurring or pixelation. This statistic is used to assess our anonymized faces across test images generally.

Reverse Image Search

It is challenging to quantify anonymization objectively. In this study, we evaluate anonymized faces by utilizing the Google search engine to do reverse image searches. To do this, we randomly choose a few well-known celebrities from the CelebA collection whose images are readily available on search engines. Then, if Google could identify them this way, our anonymization method would be ineffective, and vice versa. The anonymizing model may now be checked and evaluated using this way because it is quick. Fortunately, we show that all anonymized celebrities remain unknown using reverse search.

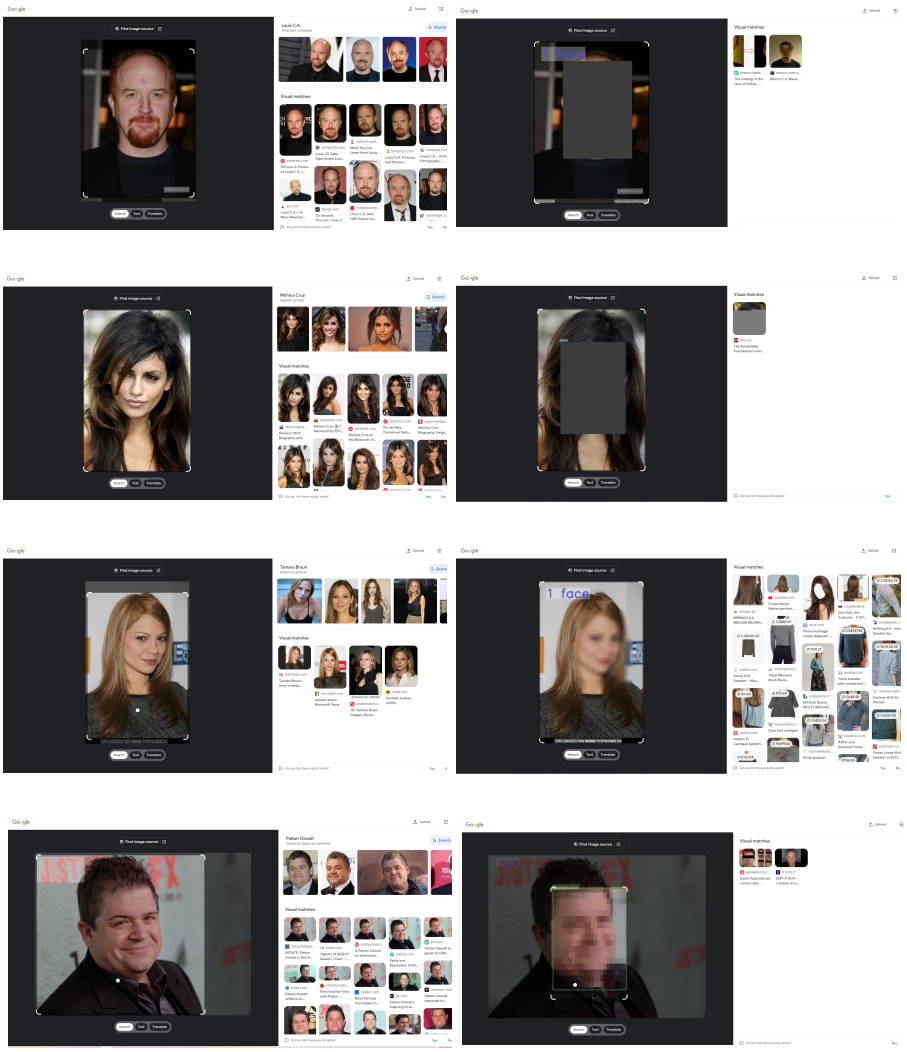


Figure 5.3.4: Reverse Image Search using the Google search engine. From left to right, each column shows the (a) Original celebrities image search and (b) Reverse search after anonymization using different methods. The fact that the Google search engine is unable to recognize the identities of celebrities whose faces have been anonymized demonstrates that our techniques for protecting privacy are effective.

CHAPTER

SIX

EXPERIMENTAL RESULTS

As is evident, this research aims to de-identify each person in every frame of a real-time video by detecting and anonymizing their faces. This is true even if our best model can identify and anonymize faces in every case encountered in the real world.

During the first stage of the assessment process, we meticulously trained two highly efficient and accurate detection models, YOLOv7 and YOLOv5, utilizing three diverse datasets: WIDER Face, CelebA, and UFDD. This comprehensive approach allowed us to capture a wide range of facial variations and appearances, ensuring robust performance across different scenarios. To further enhance privacy protection, we ventured into exploring three distinct obscuring techniques: Blur, Pixelation, and Blackened. These techniques were carefully applied during the anonymization phase, resulting in the complete fulfillment of our objectives.

The assessment process comprises two distinct steps, each serving a crucial purpose in evaluating the performance of our models and the effectiveness of our anonymization techniques. In the first stage, we focused on assessing the models' proficiency in face detection. To obtain a comprehensive understanding of their capabilities, we conducted rigorous quantitative analyses, and the findings are presented in Table 6.1.3. This table provides valuable insights into the models' precision, recall, accuracy, and other relevant metrics, shedding light on their overall performance.

Moving to the second phase, we sought to validate the outcomes of the anonymization process through qualitative analysis. Recognizing the significance of visual perception and the human ability to discern faces, we employed two specific evaluation methods in this phase: Fréchet Inception Distance (FID) and reverse-search images. These methods allowed us to gauge the effectiveness of our anonymization techniques from different perspectives.

By employing the FID metric, we were able to quantitatively measure the dissimilarity between the anonymized images and the original face images. Lower FID scores indicate a higher level of resemblance between the two sets, indicating successful anonymization. Additionally, we conducted reverse searches using the

Dataset	WIDER Face				
Split Set\Metric	FID	Precision	Recall	mAP@.5	mAP@.5:.95
Training	20.11	0.887	0.742	0.765	0.445
Validation	17.88	0.886	0.719	0.746	0.416
Test	-	-	-	-	-

Table 6.1.1: Face-detection evaluation metric results. WIDER-Face.

Dataset	CelebA				
Split Set\Metric	FID	Precision	Recall	mAP@.5	mAP@.5:.95
Training	16.67	0.991	0.996	0.998	0.874
Validation	18.03	0.992	0.996	0.997	0.873
Test	20.48	0.992	0.993	0.997	0.872

Table 6.1.2: Face-detection evaluation metric results. CelebA.

anonymized images to assess their potential for identifying the individuals behind them. This process involved utilizing facial recognition algorithms on publicly available image databases and comparing the anonymized images to identify any potential matches.

Through this meticulous assessment process, we have gained valuable insights into the performance of our detection models and the effectiveness of our anonymization techniques. These findings serve as a foundation for ensuring the privacy and confidentiality of individuals in various applications, such as data sharing, research, and public image databases.

6.1 Experiment on Face detection models

We conducted a detailed comparison between our two detection models, YOLOv7 and YOLOv5, to assess their respective performance. The results of this comparison revealed notable distinctions, showcasing the superior capabilities of YOLOv7 over YOLOv5 in face detection tasks.

When examining the performance of the two models, YOLOv7 consistently exhibited less fluctuation and higher confidence in detecting individuals. It displayed a greater level of stability and reliability, delivering more accurate and reliable results. Conversely, YOLOv5 demonstrated a higher degree of variability in its detections, leading to less consistent outcomes.

One significant observation was that YOLOv5 exhibited limitations in capturing

Dataset	UFDD				
Split Set\Metric	FID	Precision	Recall	mAP@.5	mAP@.5:.95
Training	8.49	0.807	0.585	0.605	0.331
Validation	8.92	0.851	0.542	0.577	0.303
Test	7.91	0.803	0.6	0.605	0.323

Table 6.1.3: Face-detection evaluation metric results. UFDD.

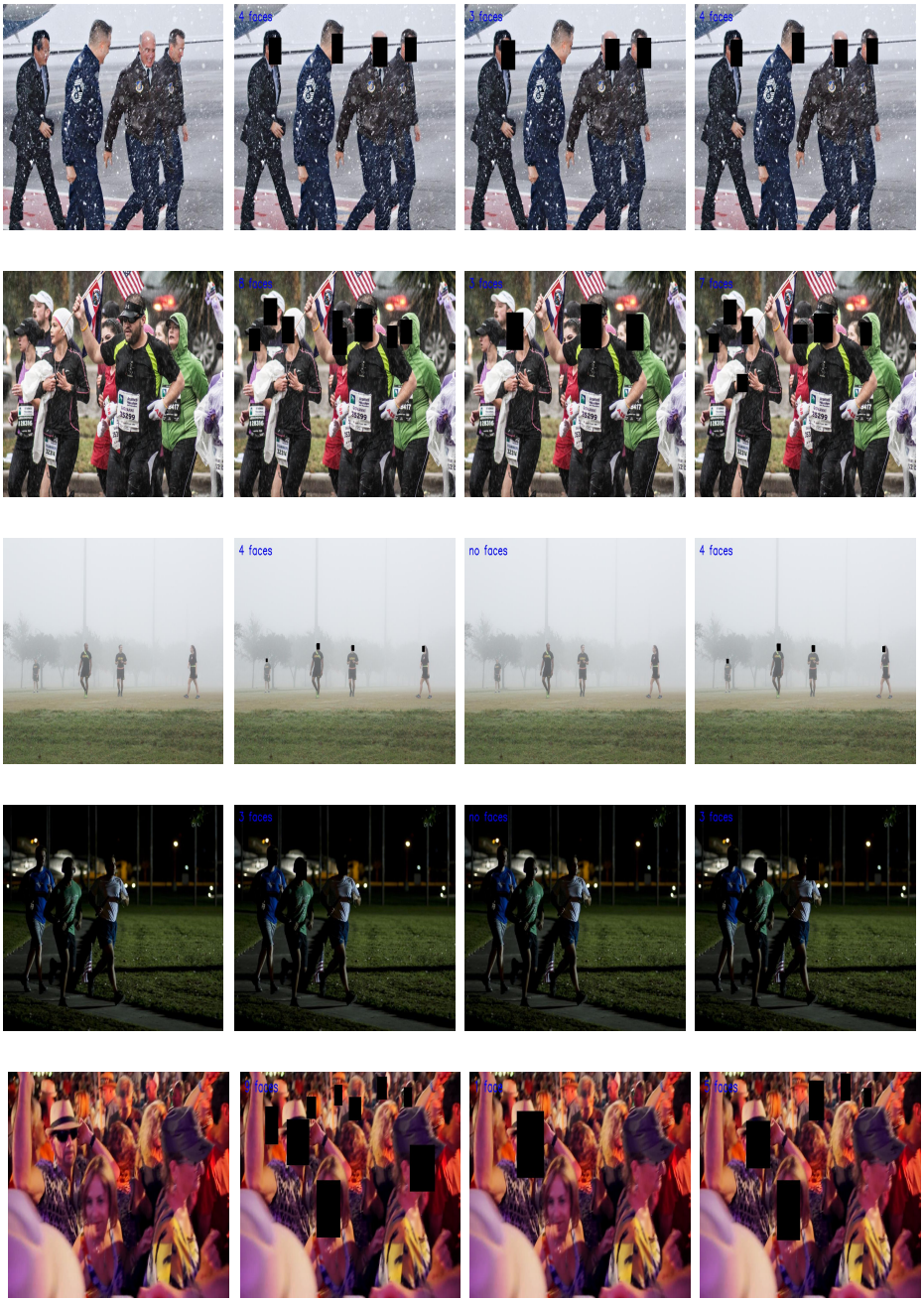


Figure 6.1.1: Qualitative comparison of anonymization on image test samples in different situations and weather conditions. From left to right, each column shows the (a) Original and anonymized faces for each dataset, (b) WIDER Face, (c) CelebA, and (d) UFDD. Each row depicts a unique sample on snow, under rain, haze, at night, and in motion situations, respectively. The results for the WIDER Face dataset are the best, then the UFDD, and the last CelebA.

certain elements present in the background as well as individuals positioned at a distance. This deficiency resulted in missed detections and a reduced ability to accurately identify people who were farther away or situated in complex visual contexts. In contrast, YOLOv7 demonstrated superior performance by effectively capturing individuals across various distances and successfully accounting for background elements. These differences became evident when comparing visual results from the same sample images, as depicted in Fig.6.1.1 and Fig.6.2.3.

The visual results displayed in these figures serve as tangible evidence of YOLOv7's superior performance in the face detection process. The images demonstrate its ability to detect individuals with greater precision, accuracy, and comprehensiveness compared to YOLOv5. These results reinforce our conclusion that YOLOv7 outperforms YOLOv5 in terms of face detection, providing enhanced capabilities for various applications where accurate identification and tracking of individuals are essential.

6.2 Experiment on Dataset

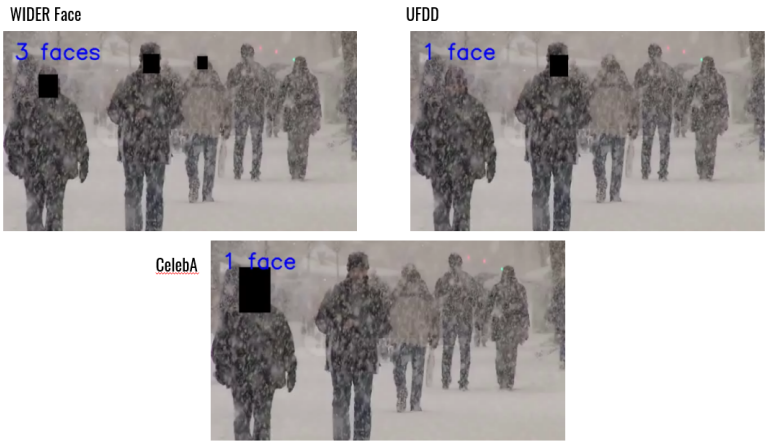


Figure 6.2.1: First frame of video anonymization under heavy snowy weather. Model performance on our three different datasets.

The performance of detection and anonymization processes can be significantly influenced by the choice of dataset and model utilized. In our study, we took this into careful consideration and employed three distinct datasets for training our model. In addition to the test data provided by these datasets, we also included additional photos and videos to comprehensively evaluate the performance of our model.

When it comes to real-time videos, the WIDER Face dataset showcased the most promising overall performance among all the datasets utilized. It effectively fulfilled the requirements for real-time anonymization as outlined in the introduction. However, the UFDD dataset also exhibited favorable outcomes, albeit with certain imperfections. Our model trained on UFDD demonstrated better capabilities in

Figure 6.2.2: First frame of video anonymization in crowded stadium with motions. Model performance on our three different datasets.

identifying and anonymizing faces of smaller sizes. This suggests that it excels in recognizing individuals who are facing the camera or positioned near the camera lens.

Nevertheless, the UFDD-trained model displayed varying levels of success in identifying and anonymizing all faces within real-time videos. While it occasionally misidentified objects in the background as faces, it proved to be highly proficient in de-identifying individuals in fast-moving scenarios. The limitations in the results produced by UFDD can be attributed, in part, to the relatively small number of training images available in this dataset. This scarcity of diverse training samples impacts the performance of deep-based detection models. Despite this drawback, UFDD exhibited commendable performance in various challenging and constrained situations. Conversely, YOLOv7 demonstrated poor performance on this dataset, while results obtained from CelebA were the least satisfactory. YOLOv7 struggled to learn features associated with occluded faces, multiple faces within a single frame and faces in low-resolution images or videos. These limitations arise from the fact that UFDD predominantly consists of non-diverse samples, each containing a large single face.

As a consequence, the choice of dataset and model plays a crucial role in the efficacy of face detection and anonymization processes. Our findings indicate that the WIDER Face dataset yields highly favorable results, particularly for real-time videos, meeting the demands set forth in our study. Despite certain imperfections, the UFDD dataset demonstrates considerable potential, excelling in specific scenarios and capturing smaller-sized faces effectively. However, the limited diversity within the UFDD dataset poses challenges for deep-based detection models, impacting their performance in certain aspects. Recognizing these factors enables us to make informed decisions regarding dataset selection and model training to achieve optimal results in face detection and anonymization applications.

When it comes to anonymization techniques, we have the flexibility to choose a suitable method and adjust the extent of concealment based on various factors

Figure 6.2.3: Anonymization methods. From left to right, each column shows the (a) Blur, (b) Blackened (black box), and Pixelation for each dataset. Rows are samples of WIDER Face (first row), CelebA (second row), and UFDD (third row) datasets, respectively. These are unique samples of whether people are close to the camera or far from the camera with small head sizes and in crowded areas.

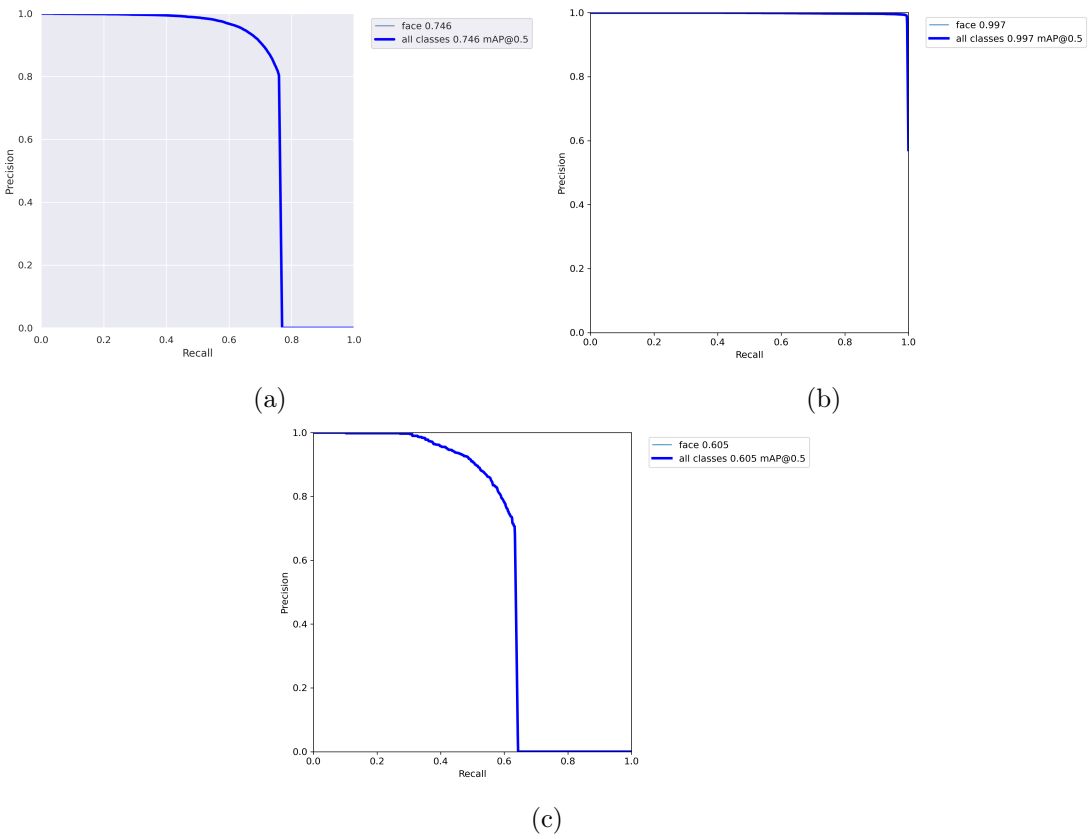


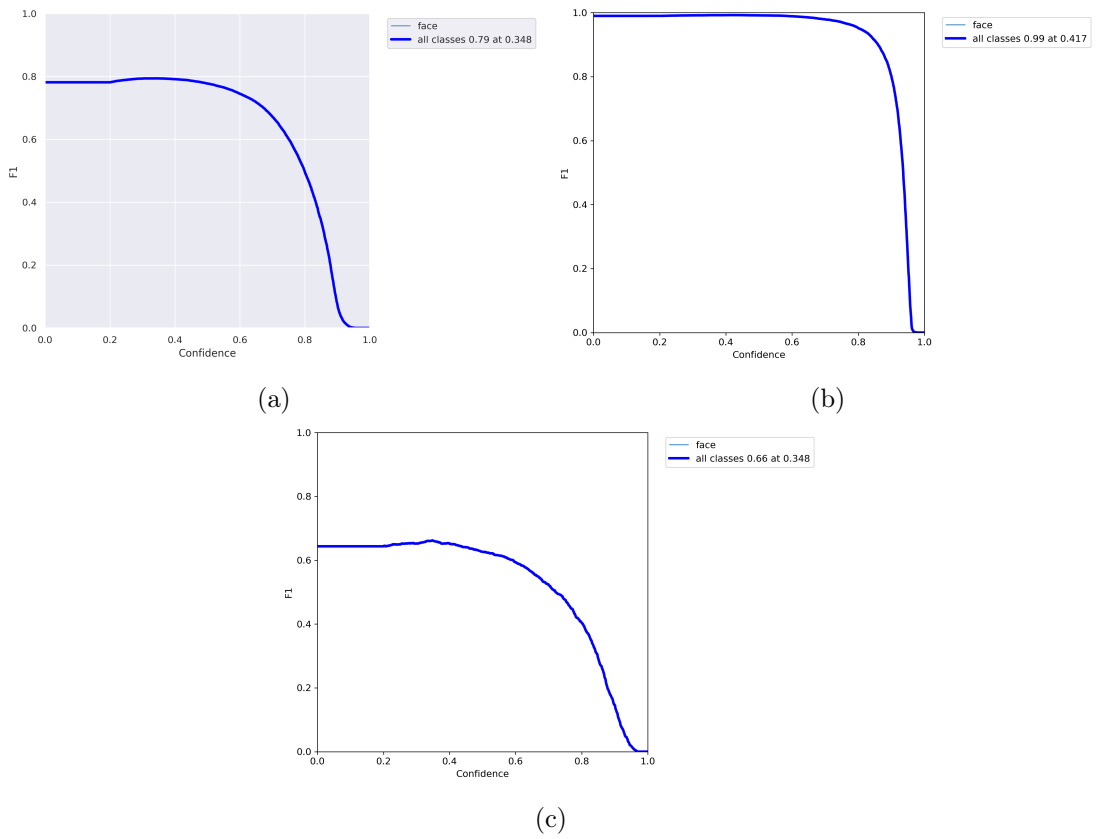
Figure 6.2.4: PR curve for (a) WIDER Face, (b) CelebA, and (c) UFDD.

such as background theme color, lighting conditions, presence of other individuals, and other relevant aspects. This adaptability allows us to tailor the anonymization process to each specific scenario, ensuring effective privacy protection while maintaining the necessary level of visibility and context.

6.3 Experiment on anonymization methods

Objective quantification of anonymization effectiveness presents a significant challenge. In our research, we have adopted a visual demonstration to showcase the efficacy of our anonymization techniques. By applying these techniques to the faces of celebrities, we conducted reverse image searches using well-known search engines like Google, effectively demonstrating that the anonymized faces cannot be identified. This reverse image search approach serves as a qualitative assessment, providing tangible evidence of the anonymization's success in preventing face recognition and preserving anonymity.

To further evaluate the performance of our proposed methodology, we utilize quantitative measures such as precision-recall curves and F1-scores. These metrics allow us to assess the precision, recall, and overall performance of the anonymization process across all three datasets. The precision-recall curve, as depicted in Fig. 6.2.4, provides insights into the trade-off between precision and recall, illustrating the effectiveness of our methodology at different thresholds. Additionally,



(a) (b)

(c)

Figure 6.3.1: F1-score for (a) WIDER Face (b) CelebA, and (c) UFDD.

the F1-score, shown in Fig. 6.3.1, offers a comprehensive evaluation by considering both precision and recall, providing a single metric to gauge the overall performance.

To sum up, this research successfully achieves face detection in images and real-time videos while ensuring the anonymity of individuals by concealing facial features such as the eyes, mouth, and nose. The level of anonymization can be customized based on specific requirements, allowing for the adjustment of parameters like the degree of blurring. Visual demonstrations, including reverse image searches, provide compelling evidence of the effectiveness of our anonymization techniques in preventing face recognition. Furthermore, quantitative evaluations through precision-recall curves and F1-scores offer objective assessments, further validating the robustness of our proposed methodology across different datasets.

CHAPTER

SEVEN

CONCLUSION

Facial anonymization has emerged as a prominent research area within computer vision, driven by the growing need for privacy protection and the increasing stringency of regulations governing data privacy. The significance of this topic has resulted in a surge of interest and research efforts, with a promising future in terms of advancements and practical applications.

In our study, we focused on adapting the Yolov7 detection model for face anonymization specifically tailored for video data. By leveraging the capabilities of Yolov7, we trained our model to detect and anonymize faces effectively. To demonstrate the effectiveness of our proposed approach, we conducted experiments using three diverse datasets: WIDER Face, CelebA, and UFDD.

Throughout our experiments, we evaluated the performance of our model under various challenging conditions, including different weather conditions and occlusion scenarios. This allowed us to assess the robustness and adaptability of our approach in real-world scenarios where faces may be partially obscured or subject to adverse environmental factors. Our model demonstrated commendable detection and anonymization capabilities across these challenging conditions.

To further validate the effectiveness of our proposed work, we conducted comprehensive evaluations using both quantitative and qualitative assessment measures. We employed standard evaluation metrics to quantitatively measure the performance of our model, including precision, recall, and F1-score. Additionally, we employed visual analysis techniques to qualitatively assess the anonymization results on diverse images and real-time surveillance videos.

Our experimental results showcased the superiority of our proposed approach in achieving reliable face detection and effective anonymization. The quantitative evaluation results demonstrated high precision, recall, and F1 scores, affirming the robustness and accuracy of our model. Furthermore, the qualitative analysis revealed visually compelling anonymization results, substantiating the efficacy of our approach in preserving privacy while maintaining the utility of the data.

In conclusion, our project study presents a novel adaptation of the Yolov7 model

for face anonymization in video data. Through comprehensive experiments on diverse datasets, we have demonstrated the effectiveness and versatility of our proposed approach under challenging conditions. The quantitative and qualitative evaluations further validate the superior performance of our model. The promising outcomes of our research contribute to the advancement of facial anonymization techniques and hold significant implications for privacy protection in various domains.

CHAPTER

EIGHT

DISCUSSION AND FUTURE WORK

In the following section, several important use cases of face anonymization and its applicability in real-world scenarios will be discussed. In addition, the various difficulties that may arise in the process of face de-identification are explored in this section as well. Afterward, we will discuss the possible future works that we had studied over them.

8.1 Face Anonymization Use Cases

8.1.1 Ecological Data Collection

The field of animal ecology aims to study and gather data on wild species in their natural habitats. In ecological research, the use of internet-of-things (IoT) sensors is prevalent for collecting various types of ecological data [55]. Additionally, cameras are often deployed in forest environments to observe wildlife behavior and gather data for scientific investigation [56, 57]. However, it is common for humans to engage in activities within these natural environments, such as hunting, camping, hiking, and more. Often, these individuals are unaware of their presence being recorded by the data collection devices.

To protect the privacy of individuals and comply with ethical considerations, it becomes necessary to remove or de-identify human appearances from the collected ecological data. Currently, researchers typically perform this task manually, which can be time-consuming and labor-intensive. To address this challenge, our proposed method offers a solution by automating the detection and anonymization of human subjects in the visual data collected by IoT devices.

By implementing our proposed methodology, it becomes feasible to automatically identify and de-identify human subjects captured by IoT devices in real-time. This can be accomplished by deploying lightweight deep learning models in deep learning-based cameras or edge devices such as the *Jetson Nano*. These models are designed to detect and blur or obscure the faces of humans present in the visual data, ensuring their anonymity without the need for manual intervention.

With the integration of low-weight models into the data collection devices, the process of de-identifying humans in ecological data can be streamlined and efficient. Researchers would no longer need to manually delete human appearances from the collected data, saving time and resources. Moreover, the implementation of such automated de-identification methods enhances the privacy protection of individuals while allowing researchers to focus on analyzing and interpreting ecological data.

It is worth noting that while face anonymization in ecological data is essential for privacy preservation, careful consideration should be given to balancing the need for anonymization with the preservation of valuable data. Researchers must strike a balance between protecting the identities of individuals and maintaining the scientific integrity of the ecological data collected. Proper anonymization techniques, such as the use of blurring or obscuring methods, should be applied to ensure privacy while preserving the usefulness of the data for ecological studies.

Therefore, the application of face anonymization in animal ecology presents an opportunity to automate the detection and anonymization of human subjects in ecological data collected through IoT devices. By implementing lightweight deep learning models in cameras or edge devices, such as the *Jetson Nano*, researchers can streamline the process of de-identifying humans in ecological data, reducing the need for manual labor and ensuring privacy protection. However, it is crucial to strike a balance between privacy preservation and data usability to uphold the scientific value of the collected ecological data.

8.1.2 Data Monetization

Marketing firms and governmental organizations now routinely gather enormous volumes of data about people and societies in order to monetize that data. The information is subsequently sold to outside parties, who can use it to gain knowledge, establish wise judgments, and generate value. However, the ethical considerations surrounding data privacy necessitate the anonymization of personal identities before selling the data.

Traditionally, the anonymization process has been time-consuming and resource-intensive, involving careful removal or masking of personally identifiable information. With our proposed methodology, a faster and more efficient anonymization of data becomes possible, enabling the data to be fully anonymized before being sold to third parties.

By leveraging our proposed methodology, the process of data anonymization can be automated, ensuring that all sensitive personal information is appropriately concealed. This automation significantly reduces the time and effort required to anonymize large datasets, enabling a more streamlined and scalable approach to data monetization.

The value of collected data is significantly enhanced by anonymizing it before its

monetization. By removing personal identities, individuals' privacy is protected, mitigating potential risks associated with the unauthorized use or disclosure of sensitive information. This increased privacy assurance instills greater confidence among data providers, contributing to a higher willingness to share data for monetization purposes.

Furthermore, anonymizing data promotes compliance with regulations and privacy standards, such as the General Data Protection Regulation (GDPR) and other data protection laws. Adhering to these regulations is crucial for maintaining trust with data subjects and avoiding legal implications.

From the perspective of data buyers, receiving fully anonymized data allows them to leverage the information without the risk of re-identification or compromising the privacy of individuals. They can confidently explore and analyze the data, uncovering valuable insights and patterns that can drive business strategies, innovation, and societal improvements.

In this study, our proposed methodology facilitates faster and more efficient anonymization of data, allowing for full anonymization before its monetization. By ensuring the protection of personal identities, data providers can create greater value from their collected data while adhering to ethical and legal considerations. The availability of fully anonymized data enhances trust, compliance, and the usability of data for data buyers, promoting responsible data monetization practices.

8.1.3 License Plate Anonymization

Figure 8.1.1: License Plate Anonymization. [58]

Autonomous vehicles (AVs) are equipped with an array of sensors, including rear and front cameras, which capture a substantial volume of data during operation. This data is typically stored on cloud servers and within the vehicles themselves. One crucial aspect of data management in AVs is ensuring the protection of sensitive information, such as license plates, to prevent accidental exposure to unauthorized individuals.

Real-time anonymization of license plates from AVs data is aligned with the principles of the General Data Protection Regulation (GDPR) regarding recording archiving. The GDPR emphasizes the importance of safeguarding personal data and preventing its unauthorized disclosure. By applying our proposed methodology, the real-time anonymization of license plates and other sensitive data becomes achievable without the need for manual labor-intensive data anonymization processes within autonomous and other vehicles.

Our methodology enables the automatic detection and anonymization of license plates in real-time, directly within the AVs' data processing pipeline. Through the use of computer vision algorithms and machine learning techniques, the license plates can be quickly identified and obfuscated, preserving the privacy of vehicle owners and occupants.

By implementing real-time anonymization, sensitive information, such as license plates, can be promptly concealed before the data is stored or transmitted. This proactive approach minimizes the risk of accidental visibility and unauthorized access to sensitive data. It ensures compliance with data protection regulations, mitigates potential privacy breaches, and reinforces public trust in the responsible use of autonomous vehicle technology.

Moreover, the advantage of our proposed methodology is that it eliminates the need for manual labor-intensive data anonymization procedures within AVs. Automating the anonymization process saves time and resources, allowing AVs to efficiently handle and protect sensitive data without relying on human intervention.

Hence, real-time anonymization of license plates and sensitive data in AVs aligns with GDPR guidelines for recording archiving, and protecting personal information from accidental exposure to unauthorized parties. Our proposed methodology offers a seamless and automated approach to anonymizing license plates, ensuring compliance with data protection regulations, and enhancing privacy in autonomous and other vehicles.

8.2 Face Anonymization Potential Challenge

Despite the significant advantages of facial anonymization in real-world scenarios, several notable challenges need to be addressed. Currently, the models used for facial anonymization are not integrated into surveillance cameras, necessitating the installation of low-weight models on edge devices to process data. One possible solution to overcome this challenge is to leverage cloud servers for data processing, enabling efficient anonymization.

Furthermore, the increasing prevalence of large-scale visual data presents difficulties in managing anonymization for such vast quantities of information. As the volume of data grows, the risk of inadvertently identifying individuals within the data also increases. To tackle this challenge, it is crucial to explore potential so-

lutions such as individualization and correlation techniques.

In terms of individualization, developing techniques that make it practically impossible to identify specific individuals in video streaming data can enhance the effectiveness of facial anonymization. By implementing measures that ensure the anonymity of individuals in real-time video streams, privacy can be safeguarded.

Correlation is another critical aspect to consider in facial anonymization. Developing criteria that prevent cross-checking of information, making it challenging to link anonymized data with external sources, contributes to maintaining the privacy of individuals. By ensuring that the anonymized data is distributed in a manner that meets these criteria, the risk of re-identification is significantly reduced.

Another significant challenge in facial anonymization lies in the potential reversibility of blurring techniques. In certain cases, blurring can be reversed, compromising the effectiveness of anonymization. To address this issue, one possible solution is to propose the use of GAN-based fake faces for real-time videos. GAN-based approaches can generate realistic synthetic faces that serve as substitutes for the original faces, effectively protecting the anonymity of individuals. By incorporating GAN-based techniques, the de-anonymization of faces in images and video data can be mitigated.

In general, while facial anonymization offers numerous benefits, several challenges need to be tackled for its successful implementation. These challenges include the integration of models into surveillance cameras, managing the anonymization of large-scale visual data, addressing the potential reversibility of blurring techniques, and exploring techniques such as individualization and correlation to ensure robust anonymity. By addressing these challenges, facial anonymization can be enhanced, contributing to improved privacy protection in various real-world scenarios.

8.3 Future Work

In our future endeavors, we are dedicated to advancing the field of data anonymization and face detection by focusing on enhancing model processing power and speed. Our goal is to make these techniques applicable to edge devices, such as the Raspberry Pi, enabling efficient and real-time processing of data right at the source.
To achieve this, we plan to explore and refine the YOLOv7 model, which is renowned for its accuracy and speed in object detection. By leveraging diverse datasets like WIDER Face, UFDD, and CelebA, we can train and adapt the YOLOv7 model to perform optimized face detection and anonymization. This combination of datasets allows us to cover a wide range of scenarios and improve the model's ability to detect and anonymize faces in real-time videos.

However, we recognize that in video processing, face detection and anonymization can vary over time due to factors like changes in lighting conditions, camera an-

gles, and facial expressions. To address this, we intend to explore techniques such as face tracking, which can track and follow faces across frames, ensuring consistent and accurate detection and anonymization throughout the duration of a video.

Furthermore, according to the latest published work in this field which is mentioned earlier in Literature Review, there is a potential future work for face anonymization that could be extending the current two-stage anonymization pipeline to address full-body anonymization in order to further enhance privacy protection. This research would involve exploring various methods and optimizations specific to full-body anonymization, considering factors such as body pose, clothing, and potential challenges in preserving segmentation details. By successfully extending the method to full-body anonymization and evaluating its impact on segmentation tasks, this future work can contribute to the development of more comprehensive privacy protection techniques on images. It would address the need for anonymizing not only faces but also other identifiable body features, making the proposed approach more versatile and applicable in a wider range of scenarios where individual recognition reduction is essential.
Indeed, this approach presents an efficient solution for anonymizing image data while preserving its utility. By ensuring that the anonymized data retains the appearance of real data, we can maintain the integrity and functionality of the information. As a result, the synthetic images generated through this method can serve as a generated valuable dataset in various machine learning applications, prioritizing human privacy considerations.
To conclude, by conducting cutting-edge research in this fascinating area, we aim to push the boundaries of data privacy and security, providing more robust and efficient solutions for real-time video anonymization. Our focus on optimizing processing power, speed, and the integration of edge devices will enable practical and scalable applications of these techniques in various domains, safeguarding individual privacy while allowing the monetization of valuable data.

REFERENCES

[1] Shruti Agarwal et al. "Protecting World Leaders Against Deep Fakes." In: *CVPR workshops*. Vol. 1. 2019, p. 38.

[2] Zhenzhong Kuang et al. "Integrating multi-level deep learning and concept ontology for large-scale visual recognition". In: *Pattern Recognition* 78 (2018), pp. 198–214.

[3] *General Data Protection Regulation EUR-Lex - 32016R0679 - EN*. `https://eur-lex.europa.eu/eli/reg/2016/679/oj`. [Accessed 20-Dec-2022].

[4] Aikaterini Soumelidou and Aggeliki Tsohou. "Towards the creation of a profile of the information privacy aware user through a systematic literature review of information privacy awareness". In: *Telematics and Informatics* 61 (2021), p. 101592.

[5] Michael Boyle, Carman Neustaedter, and Saul Greenberg. "Privacy factors in video-based media spaces". In: *Media Space 20+ Years of Mediated Life*. Springer, 2009, pp. 97–122.

[6] James L Crowley, Joëlle Coutaz, and François Bérard. "Perceptual user interfaces: things that see". In: *Communications of the ACM* 43.3 (2000), 54–ff.

[7] Chien-Yao Wang, Alexey Bochkovskiy, and Hong-Yuan Mark Liao. "YOLOv7: Trainable bag-of-freebies sets new state-of-the-art for real-time object detectors". In: *arXiv preprint arXiv:2207.02696* (2022).

[8] Glenn Jocher. *YOLOv5 by Ultralytics*. Version 7.0. May 2020. DOI: `10.5281/zenodo.3908559`. URL: `https://github.com/ultralytics/yolov5`.

[9] Shuo Yang et al. "WIDER FACE: A Face Detection Benchmark". In: *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 2016.

[10] Ziwei Liu et al. "Deep Learning Face Attributes in the Wild". In: *Proceedings of International Conference on Computer Vision (ICCV)*. Dec. 2015.

[11] Hajime Nada et al. "Pushing the Limits of Unconstrained Face Detection: a Challenge Dataset and Baseline Results". In: *arXiv preprint arXiv:1804.10275* (2018).

[12] Kaipeng Zhang et al. "Joint face detection and alignment using multitask cascaded convolutional networks". In: *IEEE signal processing letters* 23.10 (2016), pp. 1499–1503.

[13] Shifeng Zhang et al. “Faceboxes: A CPU real-time face detector with high accuracy”. In: *2017 IEEE International Joint Conference on Biometrics (IJCB)*. IEEE. 2017, pp. 1–9.

[14] Siqi Yang et al. “Using lip to gloss over faces in single-stage face detection networks”. In: *Proceedings of the European Conference on Computer Vision (ECCV)*. 2018, pp. 640–656.

[15] Jian Li et al. “DSFD: dual shot face detector”. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2019, pp. 5060–5069.

[16] Jiankang Deng et al. “RetinaFace: Single-stage Dense Face Localisation in the Wild”. In: *CoRR* abs/1905.00641 (2019). arXiv: 1905.00641. URL: http://arxiv.org/abs/1905.00641.

[17] Shifeng Zhang et al. “Refineface: Refinement neural network for high performance face detection”. In: *IEEE transactions on pattern analysis and machine intelligence* 43.11 (2020), pp. 4008–4020.

[18] Bin Zhang et al. “Asfd: Automatic and scalable face detector”. In: *arXiv preprint arXiv:2003.11228* (2020).

[19] Dmitry Yashunin, Tamir Baydasov, and Roman Vlasov. “MaskFace: multi-task face and landmark detector”. In: *arXiv preprint arXiv:2005.09412* (2020).

[20] Yanjia Zhu et al. “Tinaface: Strong but simple baseline for face detection”. In: *arXiv preprint arXiv:2011.13183* (2020).

[21] Yang Liu et al. “MogFace: Rethinking scale augmentation on the face detector”. In: *arXiv preprint arXiv:2103.11139* (2021).

[22] Jia Guo et al. “Sample and computation redistribution for efficient face detection”. In: *arXiv preprint arXiv:2105.04714* (2021).

[23] *WIDER FACE: A Face Detection Benchmark*. http://shuoyang1213.me/WIDERFACE/WiderFace_Results.html. [Accessed 20-Dec-2022].

[24] Weijun Chen et al. “YOLO-face: a real-time face detector”. In: *The Visual Computer* 37.4 (2021), pp. 805–813.

[25] Joseph Redmon and Ali Farhadi. “Yolov3: An incremental improvement”. In: *arXiv preprint arXiv:1804.02767* (2018).

[26] Golnaz Ghiasi, Tsung-Yi Lin, and Quoc V Le. “Nas-fpn: Learning scalable feature pyramid architecture for object detection”. In: *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 2019, pp. 7036–7045.

[27] Alexander Kirillov et al. “Panoptic feature pyramid networks”. In: *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 2019, pp. 6399–6408.

[28] Zhi Tian et al. “Fcos: Fully convolutional one-stage object detection”. In: *Proceedings of the IEEE/CVF international conference on computer vision*. 2019, pp. 9627–9636.

[29] Zhi Tian et al. “Fcos: A simple and strong anchor-free object detector”. In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* (2020).

[30] Kemal Oksuz et al. "A ranking-based, balanced loss function unifying classification and localisation in object detection". In: *Advances in Neural Information Processing Systems* 33 (2020), pp. 15534–15545.

[31] Hamid Rezatofighi et al. "Generalized intersection over union: A metric and a loss for bounding box regression". In: *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 2019, pp. 658–666.

[32] Zheng Ge et al. "Ota: Optimal transport assignment for object detection". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2021, pp. 303–312.

[33] Chengjian Feng et al. "Tood: Task-aligned one-stage object detection". In: *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*. IEEE Computer Society. 2021, pp. 3490–3499.

[34] Håkon Hukkelås, Rudolf Mester, and Frank Lindseth. "Deepprivacy: A generative adversarial network for face anonymization". In: *International symposium on visual computing*. Springer. 2019, pp. 565–578.

[35] Ian Goodfellow et al. "Generative adversarial networks". In: *Communications of the ACM* 63.11 (2020), pp. 139–144.

[36] Tero Karras, Samuli Laine, and Timo Aila. "A style-based generator architecture for generative adversarial networks". In: *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 2019, pp. 4401–4410.

[37] Sander R Klomp et al. "Safe Fakes: Evaluating Face Anonymizers for Face Detectors". In: *2021 16th IEEE International Conference on Automatic Face and Gesture Recognition (FG 2021)*. IEEE. 2021, pp. 1–8.

[38] Marvin Klemp et al. *LDFA: Latent Diffusion Face Anonymization for Self-driving Applications*. 2023. arXiv: 2302.08931 [cs.CV].

[39] In: (). URL: https://blog.griddynamics.com/what-is-computer-vision-and-what-can-it-do/.

[40] Emre Kazim and Adriano Koshiyama. "The interrelation between data and AI ethics in the context of impact assessments". In: *AI and Ethics* 1 (Aug. 2021), pp. 1–7. DOI: 10.1007/s43681-020-00029-w.

[41] Alexey Bochkovskiy, Chien-Yao Wang, and Hong-Yuan Mark Liao. "YOLOv4: Optimal Speed and Accuracy of Object Detection". In: *CoRR* abs/2004.10934 (2020). arXiv: 2004.10934. URL: https://arxiv.org/abs/2004.10934.

[42] Alwin Poulose, Jung Hwan Kim, and Dong Han. "HIT HAR: Human Image Threshing Machine for Human Activity Recognition Using Deep Learning Models". In: *Computational Intelligence and Neuroscience* 2022 (Oct. 2022), pp. 1–21. DOI: 10.1155/2022/1808990.

[43] Joseph Redmon et al. "You only look once: Unified, real-time object detection". In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2016, pp. 779–788.

[44] Youshan Zhang. *Stall Number Detection of Cow Teats Key Frames*. 2023. arXiv: 2303.10444 [cs.CV].

[45] Joseph Redmon and Ali Farhadi. "YOLO9000: better, faster, stronger". In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2017, pp. 7263–7271.

[46] Wael Omar et al. "Aerial Dataset Integration For Vehicle Detection Based on YOLOv4". In: *Korean Journal of Remote Sensing* 37.4 (2021), pp. 747–761.

[47] Tsung-Yi Lin et al. "Feature pyramid networks for object detection". In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2017, pp. 2117–2125.

[48] Joseph Redmon et al. "You Only Look Once: Unified, Real-Time Object Detection". In: *CoRR* abs/1506.02640 (2015). arXiv: 1506.02640. URL: http://arxiv.org/abs/1506.02640.

[49] In: (). URL: https://blog.roboflow.com/yolov4-versus-yolov5/.

[50] Renjie Xu et al. "A forest fire detection system based on ensemble learning". In: *Forests* 12.2 (2021), p. 217.

[51] In: (). URL: https://stackabuse.com/real-time-object-detection-inference-in-python-with-yolov7/.

[52] In: (). URL: https://blog.roboflow.com/yolov7-breakdown/.

[53] In: (). URL: https://www.cameralyze.co/blog/yolov7-and-yolov5-comparison-on-embedded-devices-and-computer-systems.

[54] Rafael Padilla, Sergio L Netto, and Eduardo AB Da Silva. "A survey on performance metrics for object-detection algorithms". In: *2020 international conference on systems, signals and image processing (IWSSIP)*. IEEE. 2020, pp. 237–242.

[55] Hjalte MR Mann et al. "Automatic flower detection and phenology monitoring using time-lapse cameras and deep learning". In: *Remote Sensing in Ecology and Conservation* (2022).

[56] Marc Besson et al. "Towards the fully automated monitoring of ecological communities". In: *Ecology Letters* (2022).

[57] Juan Li et al. "Deep learning for visual recognition and detection of aquatic animals: A review". In: *Reviews in Aquaculture* (2022).

[58] In: (). URL: https://platerecognizer.com/blur/.

APPENDICES

.1 Project repository

The project video results and images, the LaTeXcode, and the related **paper** can be found in the following shared links:

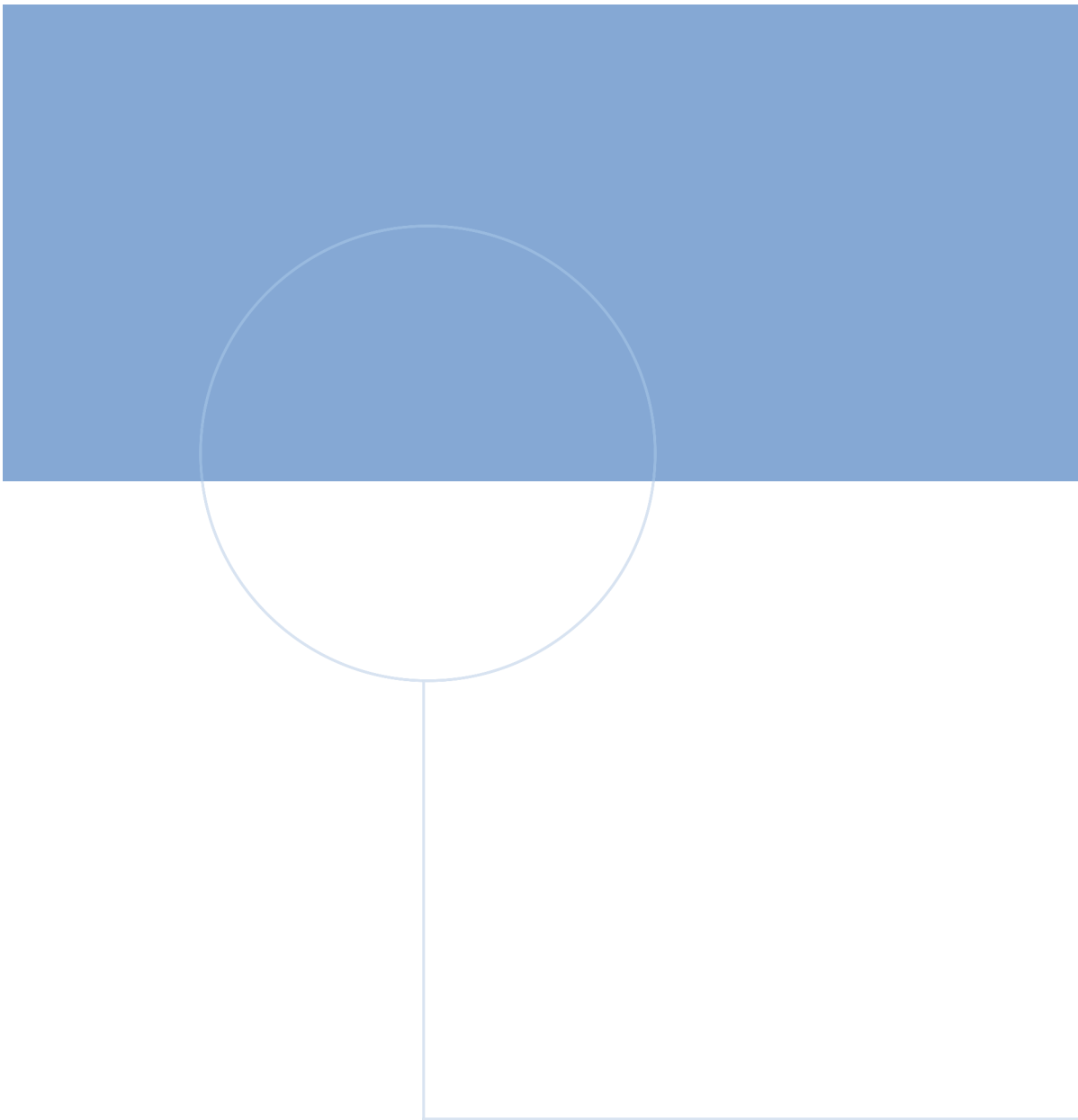
- **Google Drive link**

- **GitHub Repository**

.2 Side Research

During this research, we focused on the face-anonymization aspect of the study and employed transfer learning through fine-tuning. The approach involved utilizing a pre-trained Dual Condition Diffusion Model to generate synthetic faces and anonymize individuals' faces in images or real-time videos.
In an effort to improve the performance of the model for our specific task, we attempted further training using the WIDER-Face dataset. However, we encountered challenges and limitations in achieving the desired results. Our findings indicate that the fine-tuning process did not yield satisfactory outcomes, as the generated synthetic faces closely resembled the original detected faces, making it discernible to humans who the original person was. This qualitative analysis revealed that face anonymization was not effectively accomplished.

Moreover, it became apparent that the model relied on high-resolution images to function optimally and struggled to perform adequately with low-resolution, unseen data. Due to the lack of time for further investigation and the need to explore more effective transfer learning methods, we decided to conclude this study at this point. Additionally, the inherent limitations of the model itself, which hindered the fulfillment of our main objectives and alignment with our research questions, influenced this decision.
It is worth noting that the model and the related paper, which provide further details on the employed technique, can be found in the following link. Its paper is a recent publication and offers insights into the approach used in this research.
link to the source code



NTNU

Norwegian University of
Science and Technology