Image Inpainting with Learnable Feature Imputation

Håkon Hukkelås[0000−0001−9830−4931], Frank Lindseth[0000−0002−4979−9218], and Rudolf Mester[0000−0002−6932−0606]

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

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

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

We propose (layer-wise) feature imputation of the missing input values to a convolution. In contrast to learned feature re-normalization [36,42], our method is efficient and introduces a minimal number of parameters. Furthermore, we propose a revised gradient penalty for image inpainting, and a novel GAN architecture trained exclusively on adversarial loss. Our quantitative evaluation on the FDF dataset reflects that our revised gradient penalty and alternative convolution improves generated image quality significantly. We present comparisons on CelebA-HQ and Places2 to current state-of-the-art to validate our model. [1]

[1] Code is available at: github.com/hukkelas/DeepPrivacy. Supplementary material can be downloaded from: folk.ntnu.no/haakohu/GCPR_supplementary.pdf

1 Introduction

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

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

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

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

Our main contributions are the following:

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

2. We revisit the standard gradient penalty used to constrain Wasserstein GANs for image inpainting. Our simple modification significantly improves training stability and generated image quality at no additional computational cost.
3. We propose an improved U-Net architecture, enabling the adversarial training to attend to local and global consistency simultaneously.

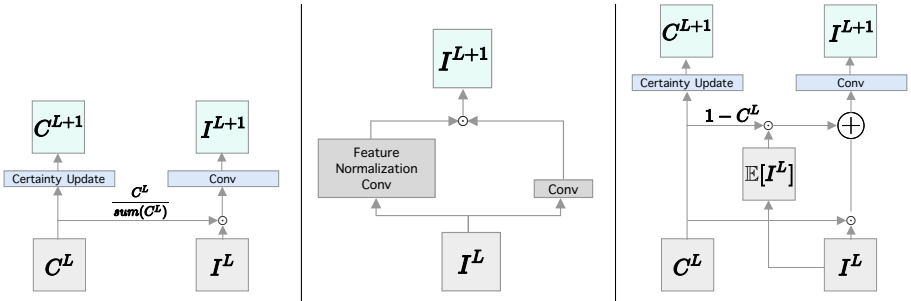
2 Related Work

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

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

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

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



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

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

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

3 Method

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

3.1 Imputed Convolution (IConv)

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

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

where $*$ is the convolution and $W_F \in \mathbb{R}^D$ is the filter. To simplify notation, we consider a single filter applied to a single one-dimensional feature map. The generalization to a regular multidimensional convolution layer is straightforward. A convolution applies this filter to all spatial locations of our feature map, which works well for general image recognition tasks. For image inpainting, there exists a set of known and unknown pixels; therefore, a regular convolution applied to all spatial locations is primarily undefined

("unknown" is not the same as 0 or any other fixed value), and naive approaches cause annoying visual artifacts [25].

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

3.2 Revisiting Gradient Penalties for Image Inpainting

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

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

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

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

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

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

Imposing $||\nabla D(\hat{x})|| \leq 1$ will remove the issue of shifting gradients in Equation 9. Furthermore, imposing the constrain $||\nabla D(\hat{x})|| \leq 1$ is shown to properly estimate the Wasserstein distance [18]. Therefore, we propose the following gradient penalty:

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

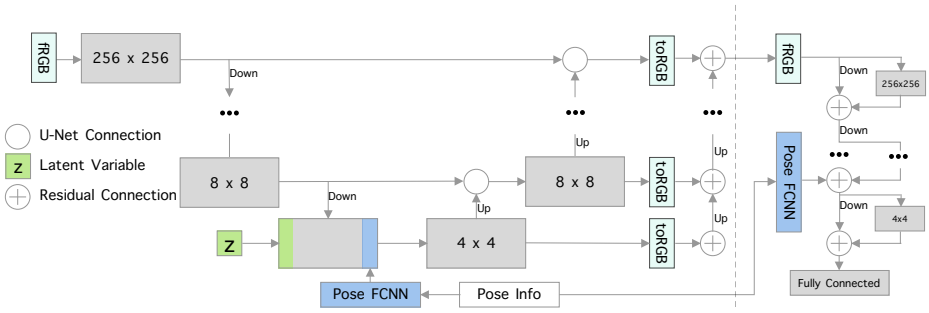


Fig. 3: Illustration of the generator (left of the dashed line) and discriminator architecture. Up and down denotes nearest neighbor upsampling and average pool. The pose information in the discriminator is concatenated to the input of the first convolution layer with 32×32 resolution. Note that pose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

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

In total, we optimize the following objective function:

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

3.3 Model Architecture

We propose several improvements to the baseline U-Net architecture [15]. See Figure 3 for our final architecture. We replace all convolutions with Equation 2, average pool layer with a certainty-weighted average and U-Net skip connections with our revised skip connection (see Equation 6). Furthermore, we replace progressive growing training [21] with Multi-Scale Gradient GAN (MSG-GAN) [19]. For the MSG-GAN, instead of matching different resolutions from the generator with the discriminator, we upsample each resolution and sum up the contribution of the RGB outputs [20]. In the discriminator we use residual connections, similar to [20]. Finally, we improve the representation of pose information in the baseline model (pose information is only used on the FDF dataset [15]).

Representation of Pose Information The baseline model [15] represents pose information as one-hot encoded images for each resolution in the network, which is extremely memory inefficient and a fragile representation. The pose information, $P \in \mathbb{R}^{K \cdot 2}$, represents K facial keypoints and is used as conditional information for the generator and discriminator. We propose to replace the one-hot encoded representation, and instead pre-process P into a $4 \times 4 \times 32$ feature bank using two fully-connected

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

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

layers. This feature bank is concatenated with the features from the encoder. Furthermore, after replacing progressive growing with MSG-GAN, we include the same pose pre-processing architecture in the discriminator, and input the pose information as a $32 \times 32 \times 1$ feature map to the discriminator.

4 Experiments

We evaluate our proposed improvements on the Flickr Diverse Faces (FDF) dataset [15], a lower resolution (128×128) face dataset. We present experiments on the CelebA-HQ [21] and Places2 [47] datasets, which reflects that our suggestions generalizes to standard image inpainting. We compare against current state-of-the art [36,42,46,29]. Finally, we present a set of ablation studies to analyze the generator architecture. [2]

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

[2] To prevent ourselves from cherry-picking qualitative examples, we present several images (with corresponding masks) chosen by previous state-of-the-art papers [11,36,42,46], thus copying their selection. Appendix 5 describes how we selected these samples. The only hand-picked examples in this paper are Figure 1, Figure 4, Figure 6, and Figure 7. No examples in the Supplementary Material are cherry-picked.

4.1 Improving the Baseline

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

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

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

4.2 Generalization to Free-Form Image Inpainting

We extend Config E to general image inpainting datasets; CelebA-HQ [21] and Places2 [47]. We increase the number of filters in each convolution by a factor of 2, such that the generator has 11.5M parameters. In comparison, Gated Convolution [42] use 4.1M, LBAM [36] 68.3M, StructureFlow [29] 159M, and PIC [46] use 3.6M parameters. Compared to [42,46], our increase in parameters improves semantic reasoning for larger missing regions. Also, compared to previous solutions, we achieve similar inference time since the majority of the parameters are located at low-resolution layers (8×8 and 16×16). In contrast, [42] has no parameters at a resolution smaller than 64×64. For single-image inference time, our model matches (or outperforms) previous models; on a single NVIDIA 1080 GPU, our network runs at ~ 89 ms per image on 256×256 resolution, $2\times$ faster than LBAM [36], and PIC [46]. GatedConvolution [42] achieves ~ 62 ms per image. [3] See Appendix 2.1 for experimental details.

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

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

(a) Input (b) GConv [42] (c) PIC [46] (d) SF [29] (e) Ours

Fig. 4: Qualitative examples on the Places2 validation set with comparisons to Gated Convolution (GConv) [42], StructureFlow (SF) [29], and Pluralistic Image Completion (PIC) [46]. We recommend the reader to zoom-in on missing regions. For non hand-picked qualitative examples, see Appendix 5.

Qualitative Results Figure 4 shows a set of hand-picked examples, Figure 5 shows examples selected by [36], and Appendix 5 includes a large set of examples selected by

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

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

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

Fig. 5: Places2 comparison to PatchMatch (PM) [4], Pluralistic Image Completion (PIC) [46], Partial Convolution (PC) [25], Bidirectional Attention (BA) [36], and Gated Convolution (GC) [42]. Examples selected by authors of [36] (images extracted from their supplementary material). Results of [42,46] generated by using their open-source code and models. We recommend the reader to zoom-in on missing regions.

the authors of [11,36,42,46]. We notice less visual artifacts than models using vanilla convolutions [46,29], and we achieve comparable results to Gated Convolution [42] for free-form image inpainting. For larger missing areas, our model generates more semantically coherent results compared to previous solutions [11,36,42,46].

4.3 Ablation Studies

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

Propagation of Certainties Figure 7 visualizes if the generator attends to shallow or deep features in our encoder-decoder architecture. Our proposed U-Net skip connection

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

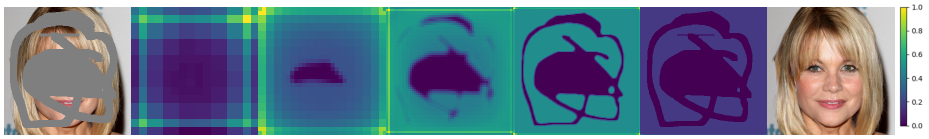


Fig. 7: **U-Net Skip Connections.** Visualization of γ from Equation 6. The left image is the input image, second column and onwards are the values of γ for resolution 8 to 256. Rightmost image is the generated image. Smaller values of γ indicates that the network selects deep features (from the decoder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

enables the network to select features between the encoder and decoder depending on the certainty. Notice that our network attends to deeper features in cases of uncertain features, and shallower feature otherwise.

5 Conclusion

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

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

References

1. Aghdam, H.H., Heravi, E.J.: Convolutional neural networks. In: Guide to Convolutional Neural Networks, pp. 85–130. Springer International Publishing (2017). https://doi.org/10.1007/978-3-319-57550-6_3
2. Arjovsky, M., Chintala, S., Bottou, L.: Wasserstein gan. arXiv preprint arXiv:1701.07875 (2017)
3. Ballester, C., Bertalmio, M., Caselles, V., Sapiro, G., Verdera, J.: Filling-in by joint interpolation of vector fields and gray levels. IEEE Transactions on Image Processing **10**(8), 1200–1211 (2001). https://doi.org/10.1109/83.935036
4. Barnes, C., Shechtman, E., Finkelstein, A., Goldman, D.B.: PatchMatch. In: ACM SIGGRAPH 2009 papers on - SIGGRAPH 09. ACM Press (2009). https://doi.org/10.1145/1576246.1531330
5. Bertalmio, M., Sapiro, G., Caselles, V., Ballester, C.: Image inpainting. In: Proceedings of the 27th annual conference on Computer graphics and interactive techniques. pp. 417–424 (2000)
6. Criminisi, A., Perez, P., Toyama, K.: Region filling and object removal by exemplar-based image inpainting. IEEE Transactions on Image Processing **13**(9), 1200–1212 (sep 2004). https://doi.org/10.1109/tip.2004.833105
7. Dolhansky, B., Ferrer, C.C.: Eye in-painting with exemplar generative adversarial networks. In: 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. IEEE (jun 2018). https://doi.org/10.1109/cvpr.2018.00824
8. Efros, A.A., Freeman, W.T.: Image quilting for texture synthesis and transfer. In: Proceedings of the 28th annual conference on Computer graphics and interactive techniques - SIGGRAPH 01. ACM Press (2001). https://doi.org/10.1145/383259.383296
9. Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., Bengio, Y.: Generative adversarial nets. In: Advances in neural information processing systems. pp. 2672–2680 (2014)
10. Gulrajani, I., Ahmed, F., Arjovsky, M., Dumoulin, V., Courville, A.C.: Improved training of wasserstein gans. In: Advances in neural information processing systems. pp. 5767–5777 (2017)
11. Guo, Z., Chen, Z., Yu, T., Chen, J., Liu, S.: Progressive image inpainting with full-resolution residual network. In: Proceedings of the 27th ACM International Conference on Multimedia. ACM (oct 2019). https://doi.org/10.1145/3343031.3351022
12. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). IEEE (jun 2016). https://doi.org/10.1109/cvpr.2016.90
13. Heusel, M., Ramsauer, H., Unterthiner, T., Nessler, B., Hochreiter, S.: Gans trained by a two time-scale update rule converge to a local nash equilibrium. In: Advances in Neural Information Processing Systems. pp. 6626–6637 (2017)
14. Huang, X., Liu, M.Y., Belongie, S., Kautz, J.: Multimodal unsupervised image-to-image translation. In: Proceedings of the European Conference on Computer Vision (ECCV). pp. 172–189 (2018)
15. Hukkelås, H., Mester, R., Lindseth, F.: Deepprivacy: A generative adversarial network for face anonymization. In: Advances in Visual Computing. pp. 565–578. Springer International Publishing (2019). https://doi.org/10.1007/978-3-030-33720-9_44
16. Iizuka, S., Simo-Serra, E., Ishikawa, H.: Globally and locally consistent image completion. ACM Transactions on Graphics **36**(4), 1–14 (jul 2017). https://doi.org/10.1145/3072959.3073659

17. Jo, Y., Park, J.: Sc-fegan: Face editing generative adversarial network with user's sketch and color. In: The IEEE International Conference on Computer Vision (ICCV) (October 2019). https://doi.org/10.1109/ICCV.2019.00183
18. Jolicoeur-Martineau, A., Mitliagkas, I.: Connections between support vector machines, wasserstein distance and gradient-penalty gans. arXiv preprint arXiv:1910.06922 (2019)
19. Karnewar, A., Wang, O., Iyengar, R.S.: Msg-gan: multi-scale gradient gan for stable image synthesis. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. vol. 6 (2019). https://doi.org/10.1109/CVPR42600.2020.00782
20. Karras, T., Laine, S., Aittala, M., Hellsten, J., Lehtinen, J., Aila, T.: Analyzing and improving the image quality of stylegan. In: 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 8107–8116 (2020). https://doi.org/10.1109/CVPR42600.2020.00813
21. Karras, T., Aila, T., Laine, S., Lehtinen, J.: Progressive growing of GANs for improved quality, stability, and variation. In: International Conference on Learning Representations (2018)
22. Kwatra, V., Essa, I., Bobick, A., Kwatra, N.: Texture optimization for example-based synthesis. In: ACM SIGGRAPH 2005 Papers on - SIGGRAPH 05. ACM Press (2005). https://doi.org/10.1145/1186822.1073263, https://doi.org/10.1145%2F1186822.1073263
23. Köhler, R., Schuler, C., Schölkopf, B., Harmeling, S.: Mask-specific inpainting with deep neural networks. In: Lecture Notes in Computer Science, pp. 523–534. Springer International Publishing (2014). https://doi.org/10.1007/978-3-319-11752-2_43
24. Li, Y., Liu, S., Yang, J., Yang, M.H.: Generative face completion. In: 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 5892 – 5900. IEEE (jul 2017). https://doi.org/10.1109/cvpr.2017.624
25. Liu, G., Reda, F.A., Shih, K.J., Wang, T.C., Tao, A., Catanzaro, B.: Image inpainting for irregular holes using partial convolutions. In: Computer Vision – ECCV 2018, pp. 89–105. Springer International Publishing (2018). https://doi.org/10.1007/978-3-030-01252-6_6
26. Meur, O.L., Gautier, J., Guillemot, C.: Examplar-based inpainting based on local geometry. In: 2011 18th IEEE International Conference on Image Processing. IEEE (sep 2011). https://doi.org/10.1109/icip.2011.6116441
27. Nazeri, K., Ng, E., Joseph, T., Qureshi, F.Z., Ebrahimi, M.: Edgeconnect: Generative image inpainting with adversarial edge learning. arXiv preprint arXiv:1901.00212 (2019)
28. Pathak, D., Krahenbuhl, P., Donahue, J., Darrell, T., Efros, A.A.: Context encoders: Feature learning by inpainting. In: 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). IEEE (jun 2016). https://doi.org/10.1109/cvpr.2016.278
29. Ren, Y., Yu, X., Zhang, R., Li, T.H., Liu, S., Li, G.: StructureFlow: Image inpainting via structure-aware appearance flow. In: 2019 IEEE/CVF International Conference on Computer Vision (ICCV). IEEE (oct 2019). https://doi.org/10.1109/iccv.2019.00027
30. Ronneberger, O., Fischer, P., Brox, T.: U-net: Convolutional networks for biomedical image segmentation. In: International Conference on Medical image computing and computer-assisted intervention. pp. 234–241. Springer (2015). https://doi.org/10.1007/978-3-319-24574-4_28
31. Salimans, T., Goodfellow, I., Zaremba, W., Cheung, V., Radford, A., Chen, X.: Improved techniques for training gans. In: Advances in neural information processing systems. pp. 2234–2242 (2016)
32. Sifre, L., Mallat, S.: Rigid-motion scattering for image classification. Ph. D. thesis (2014)
33. Simakov, D., Caspi, Y., Shechtman, E., Irani, M.: Summarizing visual data using bidirectional similarity. In: 2008 IEEE Conference on Computer Vision and Pattern Recognition. IEEE (jun 2008). https://doi.org/10.1109/cvpr.2008.4587842

34. Song, Y., Yang, C., Lin, Z., Liu, X., Huang, Q., Li, H., Kuo, C.C.J.: Contextual-based image inpainting: Infer, match, and translate. In: Computer Vision – ECCV 2018, pp. 3–18. Springer International Publishing (2018). https://doi.org/10.1007/978-3-030-01216-8_1
35. Wang, Y., Tao, X., Qi, X., Shen, X., Jia, J.: Image inpainting via generative multi-column convolutional neural networks. In: Advances in neural information processing systems. pp. 331–340 (2018)
36. Xie, C., Liu, S., Li, C., Cheng, M.M., Zuo, W., Liu, X., Wen, S., Ding, E.: Image inpainting with learnable bidirectional attention maps. In: Proceedings of the IEEE International Conference on Computer Vision. pp. 8858–8867 (2019). https://doi.org/10.1109/ICCV.2019.00895
37. Xie, J., Xu, L., Chen, E.: Image denoising and inpainting with deep neural networks. In: Advances in neural information processing systems. pp. 341–349 (2012)
38. Xu, Z., Sun, J.: Image inpainting by patch propagation using patch sparsity. IEEE Transactions on Image Processing **19**(5), 1153–1165 (may 2010). https://doi.org/10.1109/tip.2010.2042098
39. Yan, Z., Li, X., Li, M., Zuo, W., Shan, S.: Shift-net: Image inpainting via deep feature rearrangement. In: Computer Vision – ECCV 2018, pp. 3–19. Springer International Publishing (2018). https://doi.org/10.1007/978-3-030-01264-9_1
40. Yang, C., Lu, X., Lin, Z., Shechtman, E., Wang, O., Li, H.: High-resolution image inpainting using multi-scale neural patch synthesis. In: 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). IEEE (jul 2017). https://doi.org/10.1109/cvpr.2017.434
41. Yu, J., Lin, Z., Yang, J., Shen, X., Lu, X., Huang, T.S.: Generative image inpainting with contextual attention. In: 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. IEEE (jun 2018). https://doi.org/10.1109/cvpr.2018.00577
42. Yu, J., Lin, Z., Yang, J., Shen, X., Lu, X., Huang, T.S.: Free-form image inpainting with gated convolution. In: Proceedings of the IEEE International Conference on Computer Vision. pp. 4471–4480 (2019). https://doi.org/10.1109/ICCV.2019.00457
43. Zeng, Y., Fu, J., Chao, H., Guo, B.: Learning pyramid-context encoder network for high-quality image inpainting. In: 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). IEEE (jun 2019). https://doi.org/10.1109/cvpr.2019.00158
44. Zhang, H., Hu, Z., Luo, C., Zuo, W., Wang, M.: Semantic image inpainting with progressive generative networks. In: 2018 ACM Multimedia Conference on Multimedia Conference. ACM Press (2018). https://doi.org/10.1145/3240508.3240625
45. Zhang, R., Isola, P., Efros, A.A., Shechtman, E., Wang, O.: The unreasonable effectiveness of deep features as a perceptual metric. In: Proceedings of the IEEE conference on computer vision and pattern recognition (2018). https://doi.org/10.1109/CVPR.2018.00068
46. Zheng, C., Cham, T.J., Cai, J.: Pluralistic image completion. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 1438–1447 (2019). https://doi.org/10.1109/CVPR.2019.00153
47. Zhou, B., Lapedriza, A., Khosla, A., Oliva, A., Torralba, A.: Places: A 10 million image database for scene recognition. IEEE Transactions on Pattern Analysis and Machine Intelligence (2017). https://doi.org/10.1109/TPAMI.2017.2723009
48. Zhu, J.Y., Zhang, R., Pathak, D., Darrell, T., Efros, A.A., Wang, O., Shechtman, E.: Toward multimodal image-to-image translation. In: Advances in neural information processing systems. pp. 465–476 (2017)