Probabilistic Diverse GAN Based Human Face Inpainting

SHI, Haochen WANG, Yucheng ZHANG, Juntao

Hong Kong University of Science and Techonology
{hshiah,ywangls,jzhangfq}@connect.ust.hk

Abstract

Image inpainting is the task of reconstructing missing regions corrupted by irregular holes or Mosaics which has been an active topic in computer vision research. Although various models that have emerged in recent years can effectively restore corrupted images, most of them focus on restoring landscape photos. For human faces with a large amount of high-frequency information, it usually takes a long time to train but difficult to achieve good results. Our work is aimed at restoring human face images covered with irregular holes with better inpainting effects. Based on the PD-GAN model, we modified the network structure, network layer implementation, and added specific loss functions for face restoration. To validate our network performance, we completed ablation experiments for different variables, and performed qualitative and quantitative analysis concerning face restoration effects.

1. Introduction

Face inpainting is an important research direction in the field of computer vision, which has many real-world application scenarios. How to repair the face covered by masks or mosaics, that is, to repair the irregular holes becomes particularly important. Many image inpainting model helps deal with image restoration and context removal tasks, which utilizes corrupted images and tries to restore the whole facial features. Most of these models are based on the principle of partial convolution, and these methods can effectively repair landscape photos. However, for high-frequency information such as faces, problems such as distortion, blurring, and irreparable facial features often occur.

In recent year, Generative Adversarial Network (GAN) model [5] has better performance in face restoration than traditional image restoration methods. However, the problem that the facial features cannot be repaired still exists, and we found in the experiment that the GAN model will produce artifacts in the repaired part. In order to solve these problems in face restoration, we propose a new model based on PD-GAN model [8], which is dedicated to more perfect restoration of faces.

Our model is based on the PD-GAN model [8], but modified the network structure, network layer implementation, and added specific loss functions, which enables our model to perform better in face restoration. Given masked facial images, our model would generate the whole image with similar pixel values and semantic meanings to the original one. Then, we validated our network performance with ablation experiments for different variables and evaluated the model by comparing the original and generated images. Qualitatively, inpainted images would be displayed to demonstrate our inpainting performance. At the same time, quantitatively, PSNR, SSIM [13] would be utilized to evaluate how the generative model is performed pixel-wisely.

In summary, the major works we have done are as follows:

- We modified the network structure and network layer implementation of PD-GAN to make the probability change between pixels more continuous and reduce the generation of artifacts.
- We proposed a MASK loss function and utilized the Heatmap loss function to focus the inpainting target more on human face and achieve better facial restoration results.
- Ablation experiments are conducted to validate the proposed network. The experiment result demonstrates the satisfactory level of our model in comparison to several other existing models.

2. Related Work

2.1. Image inpainting

Image inpainting is a reconstruction technology aiming to restore the missing regions corrupted by irregular holes or Mosaics. The traditional approach is to utilize Partial Convolution models [7], proposed by Nvidia, which has a long-term impact on subsequent.

In recent research, many of the frameworks proposed for the image inpainting task is based on the Generative Adversarial Network [3]. There are serval approaches to the image inpainting problem. PD-GAN [8] trains a GAN using the proposed SPDNorm to generate masked regions, and UctGAN [14] utilizes an encoder-decoder network similar to conditional VAE [12]. Our model is based on the principle of PD-GAN with improvements on the SPDNorm layer.

2.2. PD-GAN

The PD-GAN network has two stages. The first stage pre-trains an inpainting network using Partial Convolutions [7], which can generate a coarse prediction as the prior information. In the second stage, The model samples the latent vector z from a Random Noise and modulates z with the prior information based on the SPDNorm Residual block.

The SPDNorm layer is based on the idea of Batch Norm. In the former model inspried PD-GAN [8] - SPADE [11], the masks are first projected onto an embedding space and then convolved to produce the modulation parameters γ and β. The produced γ and β are multiplied and added to the normalized activation element-wise.

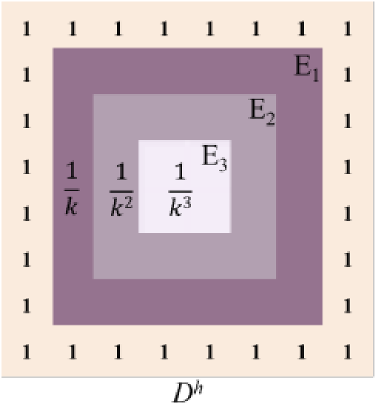


Figure 1. This figure is from PD-GAN: Probabilistic Diverse GAN for Image Inpainting and illustrates how the Hard SPDNorm decides probability which decreases from boundary to center. The probability decreases at a constant valve k, but it ignores the number of surrounding ground truth pixels will affect the probability of hole pixels.

In the PD-GAN [8] case, the SPDNorm Residual block consists of Hard SPDNorm and Soft SPDNorm. Figure 1 shows an example of Hard SPDNorm, which is a direct probability map. For the pixels masked by irregular holes, find the nearest ground truth distance i , And the probability that this pixel can obtain information from the surrounding ground truth is set to $\frac{1}{k^i}$.

However, this arbitrary method of setting the probability ignores the number of surrounding ground truth pixels will affect the probability of hole pixels. The probability changes between pixels are discrete and the interval is large. As the main probability assignment method, Hard SPDNorm may be one of the main causes of artifacts. In our model, a relatively continuous value is taken for the quantitative relationship between the pixel and the surrounding ground truth.

2.3. Super-FAN

Super-FAN is a face super-resolution model proposed in [2] aiming to super-resolve very low-resolution faces. Super-FAN proposed a Face Alignment Network (FAN) with 2 Hourglass models [9], a network for facial landmark localization through heatmap regression to enforce facial structural consistency. FAN utilizes the concept of heatmap regression [1] which represents each landmark by an output channel containing a 2D Gaussians centered at the landmark's location [2].

3. Data

3.1. Datasets

We trained our model to perform diverse inpainting tasks on facial images on CelebAMask-HQ [6] dataset, and use irregular masks provided by [7]. The choice of Ground truth and masks has a great impact on the accuracy of training. We carefully selected and designed 100 masks instead of random generation, these masks can effectively cover facial features, which allows us to be more specific to facial features. At the same time, these masks contain large block occlusions and interval small occlusions, which avoids overfitting of our model and makes it perfect for various tasks.

Name	Number of images
CelebAMask-HQ [6]	7,000
CelebAMask-HQ(Skin) [6]	7,000
Masks [7]	100

Table 1. Our datasets

Shown in Table 1, For the ground truth, we selected 7000 images from the CelebAMask-HQ [6] dataset.Before we fed them to our model, there were inspection and pre-

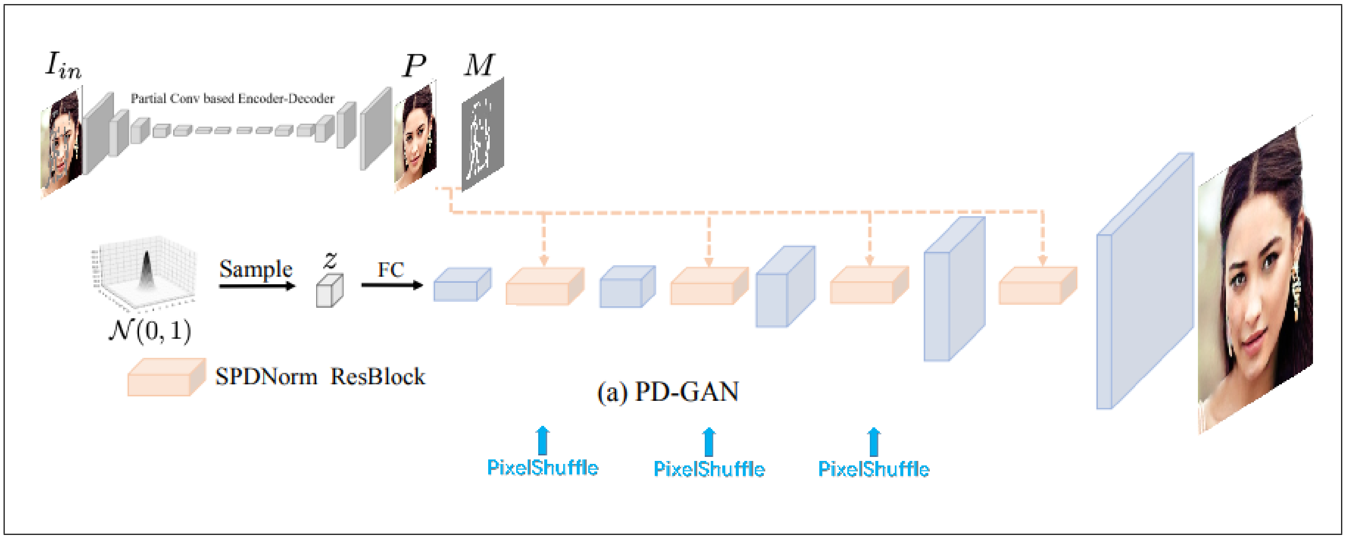


Figure 2. Our model pipeline

processing steps. Skin masks (the value of the face is 1) are also fed to help us quickly obtain the facial pixel area, which is easy for the calculation of the subsequent loss function.

3.2. Preprocessing

The preprocessing process includes multiple stages, the first is the processing of ground truth and mask images to make these images the same size. In each epoch, we select the ground truth image and masks in a certain order,which helps to compare the effects of different methods on the same image and mask in our subsequent ablation experiments. After selecting a batch of ground truth images, we stack them together with pixel-wise multiplication.

4. Method

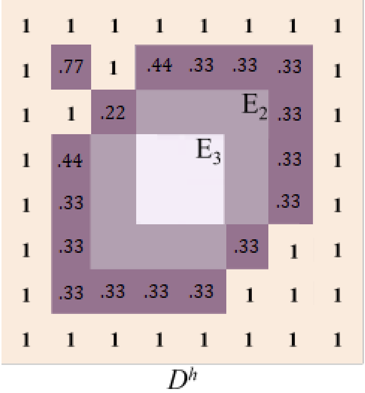
4.1. Network Architectures

In order to ensure the diversity of generated images, our model has a similar structure to PD-GAN [8], which does not use an Encoder-Decoder structure to inpaint the corrupted parts, but generates diverse image content according to different Random Noise inputs. In this section, we will introduce the pipeline of our model, as well as the improvements in response to the potential problems.

Specific improvement measures include: In the upsampling process, we use PixelShuffle instead of Transpose convolution to avoid the ”checkerboard pattern”. As shown in [10], transpose convolution may result in ”checkerboard patterns” when dealing with bright-colored images. This pattern is caused by the overlapping of transpose convolution operations. To tackle this, we use two more convolution layers and a pixel-shuffle layer to replace the original transpose convolution layers.

4.2. Hard SPDNorm Inprovement

Intuitively, the masked pixel is to the hole boundary, their should be stronger constraints of the pixel value. [8], which is because more prior information is needed to keep the pixel value consistent with its neighbouring ground truth as the Figure 3 shows.



1	1	1	1	1	1	1	1
1	.77	1	.44	.33	.33	.33	1
1	1	.22				.33	1
1	.44					.33	1
1	.33					.33	1
1	.33				.33	1	1
1	.33	.33	.33	.33	1	1	1
1	1	1	1	1	1	1	1

Figure 3. Improved Hard SPDNorm example, the ivoroy color pixels stand for irregular boundary of ground truth. For the masked pixels (purple), its pixel value depends on the number of neighboring ground truth. We acquire pixel value by convolution then divided by kernel size.

As an improvement to Hard SPDNorm, For the masked M region, we apply n iterative operations to it. The mask after the i-th dilation operation is M_i. During the update process, M_i is decided by the kernal convolution values of M_{i-1}. $\mathcal{N}(x,y)$ denotes the neighbouring region of centered at (x,y), the size of neighbouring region would increase as the upsampling continues. Mathematically, our mask update process can be expressed as:

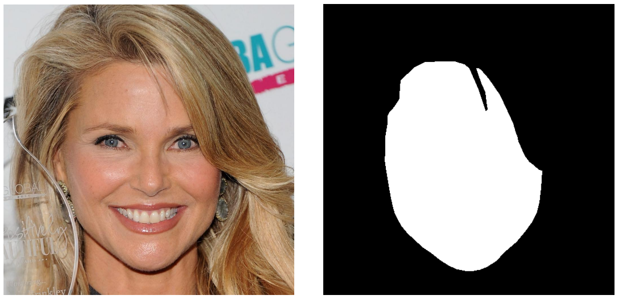
$$M_i(x,y) = \begin{cases} \frac{M_{i-1}(a,b)}{\#\ of\ \mathcal{N}(x,y)} & if \sum_{(a,b)\in\mathcal{N}(x,y)} M_{i-1}(a,b) > 0 \\ 0 & otherwise \end{cases}$$

4.3. Loss Function

The loss functions in the PD-GAN model [8] consist of perceptual diversity loss, reconstruction loss, feature matching loss [7] and hinge adversarial loss. These loss functions optimize the network concerning diversity of generated images based on the prior information. However, these general loss functions do not have a significant effect on the human face. On these basis, we propose and use Mask loss function and FAN loss function.

4.3.1 Skin-Mask Loss Function

As referred in table 1, we also loaded 7000 corresponding skin images from the CelebAMask-HQ [6] dataset. These masks help us quickly obtain the facial area with pixel-wise multiplication. We can add a specific loss function to the pixels in this area, which enables our model have a better repairing effect on facial details.



(a) Image

(b) Skin mask

Figure 4. Figure 4b displays the facial mask of the person on the left, which is called the skin image in the dataset.

For better facial features inpainting, it is appropriate to add an L1 loss:

$$\mathcal{L}_{L1}(G) = \sum_{All\ G} \|G - Gt\|_1 * Mask$$

4.3.2 Heatmap Loss Function

From preliminary experiments, we found that some inpainted images have significant misalignments in the facial structures compared to the ground truth. And some inpainted images have weak boundaries on facial features. Since the primary loss function overlooks facial structures, the model is unaware that facial feature misalignment causes significant semantic differences. Therefore, we utilize the Heatmap Loss proposed in [2]. And the heatmap loss is defined as:

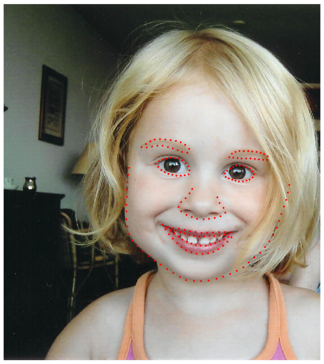


Figure 5. from [4] Landmarks (red dots) of a human face

$$\mathcal{L}_{heatmap} = \frac{1}{N}\sum_{i=1}^{N}\sum_{ij}\left(\widehat{M_{i,j}^n} - M_{i,j}^n\right)^2$$

We obtain $M_{i,j}^n$ and $\widehat{M_{i,j}^n}$ by running FAN on the ground truth images and our inpainted images. And the heatmap corresponding to the n-th landmark at pixel (i,j) is $M_{i,j}^n$.

4.4. Overall Training Loss

The overall training loss of our model:

$$\mathcal{L}_{total} = \alpha * \mathcal{L}_{Mask} + \beta * \mathcal{L}_{heatmap} + \mathcal{L}_{PDGAN}$$

where α and β are the corresponding weights

5. Experiment

5.1. Experiment Setup

We evaluated our proposed model on first 7000 images in CelebAMAask-HQ [6] and 100 irregular masks. We set the inpainting result of Partial Convolution (PC) [7] as prior information, using pretrained model provided in PDGAN. The mask and image are resized to 256*256 as network input. Our model is optimized using Adam optimizer. The initial learning rate is 0.001 and we use linear learning rate decay schedule to train the model. The latent vector dimension is 128. We train the network for 120 epochs with batch size of 8 and it completes one epoch in 9 minutes. For the loss function, we maintain the original setting in PDGAN and set α as 0.5, β as 10.

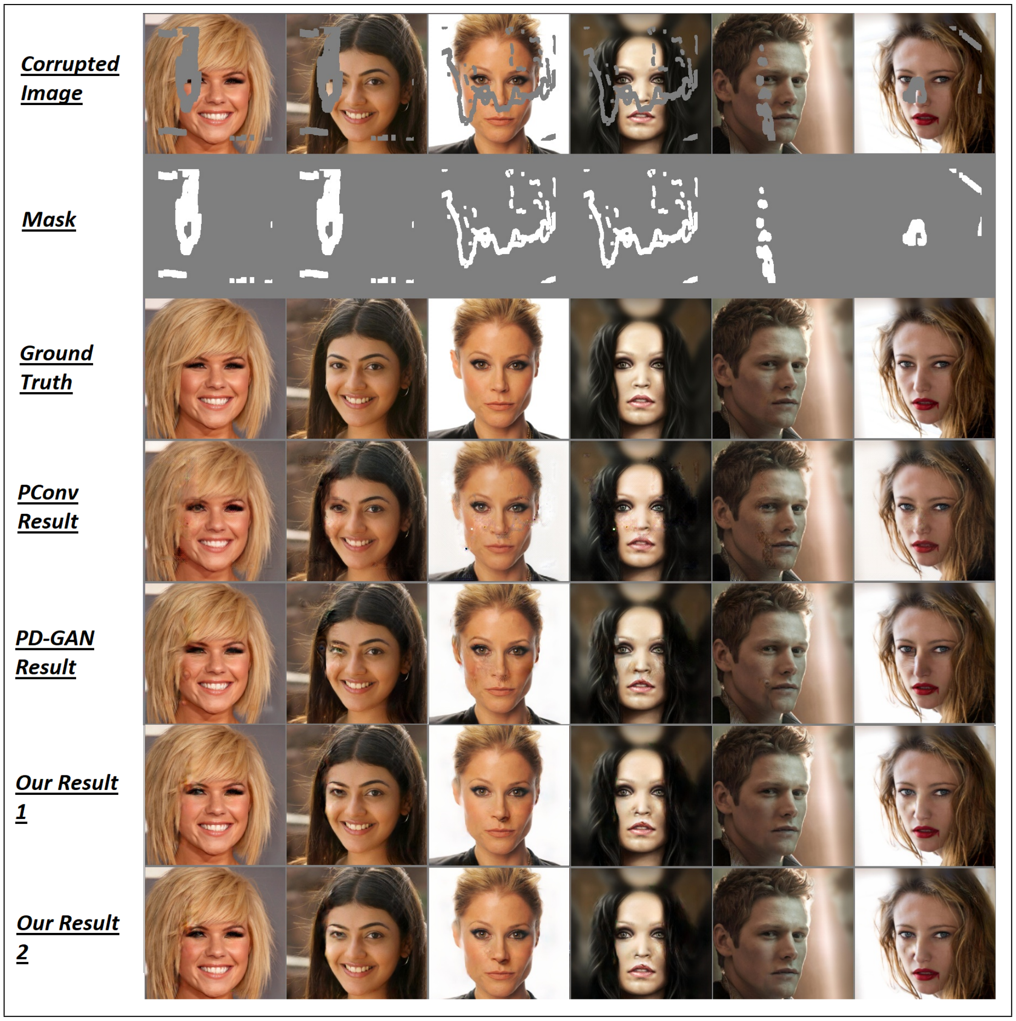


Figure 6. Comparison of inpainting image.

5.2. Comparing with Existing Work

We compare with the following inpainting models: PConv [7], PD-GAN [8] and our model and they are trained using official config. The result of inpainting irregular holes on CelebAMask-HQ [6] is shown in Figure 5.

The first three rows are masked image, mask and ground truth. The next two rows are the generated image of PConv and PD-GAN, while the last part is result of our model. PConv and PD-GAN can generate the roughly correct structure of originally image, but the result contains blurry content and sprial stain especially when the masked regions are dispersive and eyes or lip are masked. The result generated by our model is more natural. The intensity on the edge of mask is much smoother and the blurry problem is alleviated due to restrictions on intensity values in skin mask region.

5.3. Ablation Study

We perform ablation study on FAN loss and Mask loss to evaluate their effects on experiment results.

Model Name	PSNR	SSIM
PConv [7]	21.3485	0.9307
PD-GAN [8]	22.4371	0.9448
FAN1 MASK0	22.7748	0.9390
FAN0 MASK1	23.4279	0.9562
Our Model	24.0219	0.9593

Table 2. Quantitative ablation study on FAN loss and Mask loss

To test the performance of image inpainting, we use Peak Signal-to-Noise Ratio (PSNR) and Structural Similarity (SSIM) [13] as evaluation criteria for image quality and

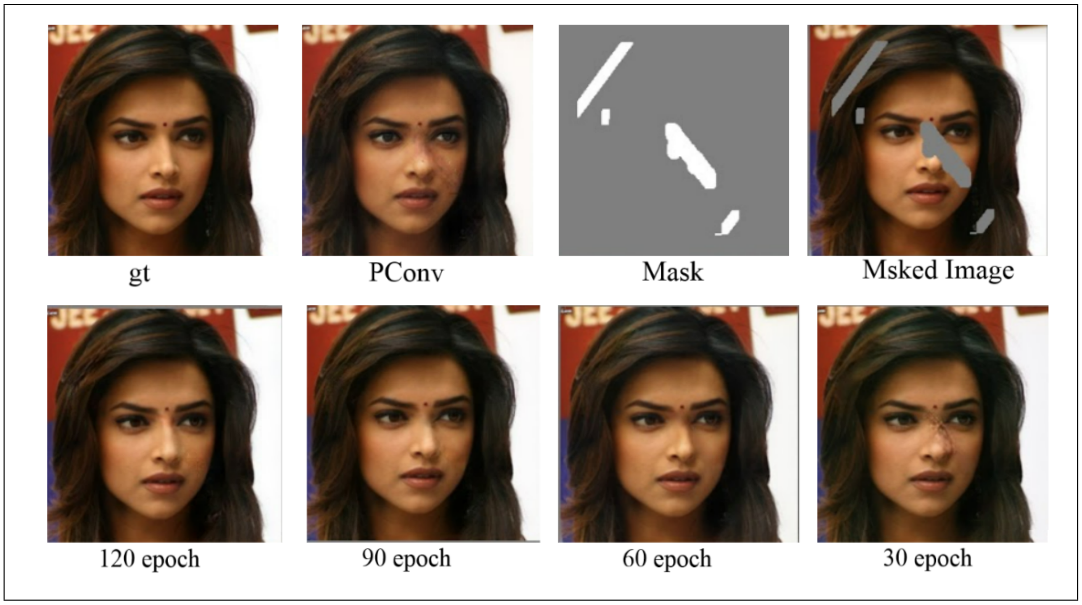


Figure 7. Results in different epoches

similarity among two images respectively. FAN1 MASK0 only involves Fan loss and PDGAN loss and FAN0 MASK1 includes Mask loss and PD-GAN loss. To avoid effects of other module, remaining parameters are same. The evaluation result is shown in Table 2. Our model outperforms PDGAN and PConv and Mask Loss contributes significantly towards improvement on similarity, while Fan Loss promotes image quality.

5.4. Learning Process

We found some interesting facts during the training process by comparing the outputs for a single gt-mask pair at different training stages. As shown in Figure 7, our model first learns to use priors given by the Partial Conv based network. However, priors often gives blurred and unobvious Facial structures.

Around the 60th epoch, our model gradually learns to sharpen those areas. Instead of giving blurring patch-like results, our model focuses on facial structures. Then in about the 90th epoch, our model learned to enhance those missing facial features. However, at this stage, the recovered features are not entirely reasonable (e.g. the non-continuous nose bridge).

Toward the end of the training, our model starts to output clear and reasonable results. The reconstructed facial structure is more reasonable, and the transition between masked-unmasked areas becomes smooth.

5.5. Fail cases

Two typical fail cases are shown in Figure 8. The leading cause of such failures is that the model cannot distinguish objects (e.g., lollipops) and poster words from human faces. The model regards those parts as components of human faces and tries to extend features to the generated parts—this results in mixing up facial features with non-face features and is unreasonable to humans.

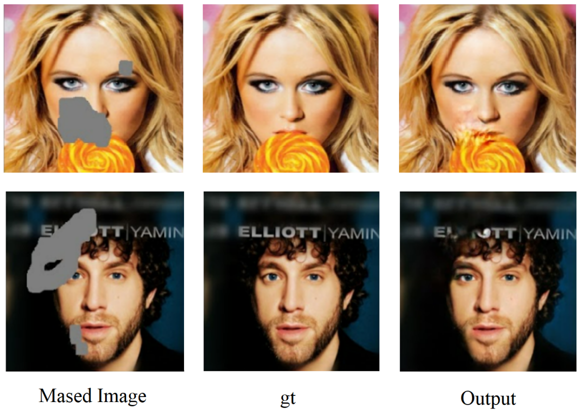


Figure 8. Fail cases of our model

The solution to such fails may need further enhancement of the ability to understand more real-world objects, words, and phrases.

6. Conclusion

We modified the PD-GAN network structure and improved the network layer implementation of diversity map. By aggregating Skin-Mask loss and Fan loss, our model has better capacity to inpaint irregular holes on facial image. In the process of testing, our model could restore the image by generating semantically meaningful results on the masked areas given a single masked facial image. Qualitative and quantitative analysis are conducted and demonstrate that our model generates high-quality reconstruction image with less blurry content. In future work, we will train our model on more cases to enhance the ability to understand real-world objects and further balance the diversity and image quality.

References

[1] Adrian Bulat and Georgios Tzimiropoulos. Human pose estimation via convolutional part heatmap regression. *CoRR*, abs/1609.01743, 2016. 2

[2] Adrian Bulat and Georgios Tzimiropoulos. Super-fan: Integrated facial landmark localization and super-resolution of real-world low resolution faces in arbitrary poses with gans. 12 2017. 2, 4

[3] Ugur Demir and Gozde Unal. Patch-based image inpainting with generative adversarial networks. *arXiv preprint arXiv:1803.07422*, 2018. 2

[4] J. Brandt F. Zhou and Z. Lin. Exemplar-based graph matching for robust facial landmark localization. In *IEEE International Conference on Computer Vision (ICCV)*, 2013. 4

[5] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In Z. Ghahramani, M. Welling, C. Cortes, N. Lawrence, and K.Q. Weinberger, editors, *Advances in Neural Information Processing Systems*, volume 27. Curran Associates, Inc., 2014. 1

[6] Cheng-Han Lee, Ziwei Liu, Lingyun Wu, and Ping Luo. Maskgan: Towards diverse and interactive facial image manipulation. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020. 2, 4, 5

[7] Guilin Liu, Fitsum A. Reda, Kevin J. Shih, Ting-Chun Wang, Andrew Tao, and Bryan Catanzaro. Image inpainting for irregular holes using partial convolutions, 2018. 2, 4, 5

[8] Hongyu Liu, Ziyu Wan, Wei Huang, Yibing Song, Xintong Han, and Jing Liao. Pd-gan: Probabilistic diverse gan for image inpainting, 2021. 1, 2, 3, 4, 5

[9] Alejandro Newell, Kaiyu Yang, and Jia Deng. Stacked hourglass networks for human pose estimation. In Bastian Leibe, Jiri Matas, Nicu Sebe, and Max Welling, editors, *Computer Vision – ECCV 2016*, pages 483–499, Cham, 2016. Springer International Publishing. 2

[10] Augustus Odena, Vincent Dumoulin, and Chris Olah. Deconvolution and checkerboard artifacts. *Distill*, 2016. 3

[11] Taesung Park, Ming-Yu Liu, Ting-Chun Wang, and Jun-Yan Zhu. Semantic image synthesis with spatially-adaptive normalization. 2019. 2

[12] Jacob Walker, Carl Doersch, Abhinav Gupta, and Martial Hebert. An uncertain future: Forecasting from static images using variational autoencoders. *CoRR*, abs/1606.07873, 2016. 2

[13] Zhou Wang and A.C. Bovik. A universal image quality index. *IEEE Signal Processing Letters*, 9(3):81–84, 2002. 1, 5

[14] Lei Zhao, Qihang Mo, Sihuan Lin, Zhizhong Wang, Zhiwen Zuo, Haibo Chen, Wei Xing, and Dongming Lu. Uctgan: Diverse image inpainting based on unsupervised cross-space translation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020. 2