ISFRNet: A Deep Three-stage Identity and Structure Feature Refinement Network for Facial Image Inpainting

Yan Wang, and Jitae Shin*
Department of Electrical and Computer Engineering, Sungkyunkwan University
Suwon, Gyeonggi-do 16419, Korea
[e-mail: {wy137568, jtshin}@skku.edu]
*Corresponding author: Jitae Shin

Received November 18, 2022; revised February 9, 2023; accepted March 12, 2023; published March 31, 2023

Abstract

Modern image inpainting techniques based on deep learning have achieved remarkable performance, and more and more people are working on repairing more complex and larger missing areas, although this is still challenging, especially for facial image inpainting. For a face image with a huge missing area, there are very few valid pixels available; however, people have an ability to imagine the complete picture in their mind according to their subjective will. It is important to simulate this capability while maintaining the identity features of the face as much as possible. To achieve this goal, we propose a three-stage network model, which we refer to as the identity and structure feature refinement network (ISFRNet). ISFRNet is based on 1) a pre-trained pSp-styleGAN model that generates an extremely realistic face image with rich structural features; 2) a shallow structured network with a small receptive field; and 3) a modified U-net with two encoders and a decoder, which has a large receptive field. We choose structural similarity index (SSIM), peak signal-to-noise ratio (PSNR), L1 Loss and learned perceptual image patch similarity (LPIPS) to evaluate our model. When the missing region is 20%-40%, the above four metric scores of our model are 28.12, 0.942, 0.015 and 0.090, respectively. When the lost area is between 40% and 60%, the metric scores are 23.31, 0.840, 0.053 and 0.177, respectively. Our inpainting network not only guarantees excellent face identity feature recovery but also exhibits state-of-the-art performance compared to other multi-stage refinement models.

Keywords: Deep Learning, Image Inpainting, GAN.

1. Introduction

Image inpainting is a process used to fill in missing areas. The goal is to make modifications in an image that are semantically reasonable and more detailed. This process is similar to when humans visually observe a damaged image, and automatically imagine what the complete image looks like based on valid information. However, when the missing region is too large, the complete image looks like a blur or rough outline in our mind, or only the category of the image is known. Although the convolutional neural network model mimics the operation of the human brain, it also suffers from the problem of not being able to complement the image very well. The location, size and shape of the damaged area are arbitrary, and as the damaged area increases, the effective pixel information that can be used becomes insufficient. Therefore, when the missing region is large, it is still a great challenge to generate a more complete structure and more detailed and realistic content.

In recent years, especially after the emergence of generative adversarial network (GAN) [1] models, deep learning has achieved significant breakthroughs in the area of image inpainting. Image inpainting methods based on adversarial networks can learn effective high- and low-frequency feature information at a low level and learn the consistency of the image structure and texture at a high semantic level. These approaches can be broadly divided into single stage inpainting and progressive image inpainting. Moreover, several scholars even apply attention modules or a priori knowledge to refine the final image generation. However, a common limitation of these methods is that they are not as good as human beings who imagine the complete image based on their own subjective ideas or preferences. For example, when all facial identifiers of a face image are missing (no valid information is available), people who like double eyelids may complete a portrait with big eyes, and people who like single eyelids may fix it with the shape of a single eye. This capability is necessary to avoid blurred or artificial images when the damaged area is very large. As we know, StyleGAN [2] has the ability to do that. It can generate images with different styles by injecting different style vectors, and generates less blurred or artificial images even when applied to the field of image inpainting.

Therefore, we try to apply the pre-trained pSp-SyleGAN [3] as the first step of our framework (ISFRNet). However, although pSp-StyleGAN is able to generate very realistic images regardless of the size of the corrupted regions in the input image, it is difficult to maintain the consistency of the identity features. Thus, we add a second step (IFR) to recover identity features and add the third step (SFR) to balance structural features in a global image. As demonstrated by [4], networks with small receptive fields are more effective in repairing local structures and textures, while networks with large receptive fields are more effective in repairing details and structures over long distances. Most of the identity features belong to the local structure and texture, so the identity feature recovery network (IFR) is designed as a shallow network with a small receptive field, and structure feature repair (SFR) is designed as a large receptive field based on U-Net [5]. Furthermore, we also introduce a weighting mechanism to balance the input structure feature in the third step.

In summary, the contributions of our paper are summarized in the following three points:

A. We apply two refinement networks with different receptive fields for face identity recovery and global structure repair, respectively.
B. A U-net model with large receptive fields is adapted into a network with two encoders that extract the structural features of different images and inject them into a decoder for image generation.

C. We propose a weighting mechanism and apply it in the third step. The aim is to balance the input structural features while attenuating the negative impact of missing regions on global structural repair.

2. Related Work

2.1 Single-stage Inpainting

Most previous proposed single-stage inpainting models are based on an encoder-decoder or GAN structure, such as [6], [7]. The encoder extracts the features of the input image and maps them to the latent space, while the decoder expands the compressed feature map step by step to recover the size of the original image. Then, the L2 reconstruction loss function is used to reconstruct the structural features, and the adversarial loss function is used to make the generated images look more realistic. However, in this type of approach, the image features are mapped to a higher-level latent space, and while the overall structural features of the image are better extracted, much of the detailed feature information is lost.

Therefore, some scholars have tried to change the standard structure of the encoder-decoder to a pyramid structure [8], mutual encoder-decoder [9]. The Pyramid-Context Encoder Network (PEN-Net） was proposed by Zeng et a. [8]. The multi-scale encoder extracts features while using the extracted high-level features to guide the low-level feature generation. By skipping connections, similar features are learned by the attention transfer network and decoded together with latent features to obtain the restored image. This design not only improves training speed but also produces more realistic test images. Mutual Encoder-Decoder was introduced by Hongyu Liu et al. [9]. This model performs multi-scale hole filling in the feature space while equalizing the output features in the channel and spatial domains. The equalized features contain consistent structural and textural features at different feature levels. It is good to consider the consistency of the structure and texture during the image inpainting process to produce a more logical and detailed structure and texture. There are also some modules that can be used to enhance the similarity between pixels or to strengthen the constraint on the estimated deep pixels by using the correlation between adjacent pixels [10].

2.2 Progressive Inpainting

Given that convolutional neural networks are not good at modeling the correlation between long-term distant contextual information and damaged holes. Therefore, some researchers have proposed coarse-to-fine multi-stage network architectures for progressive image inpainting. The methods related to progressive inpainting can be used for both traditional low-resolution image inpainting and high-resolution image inpainting [10]. For progressive inpainting, some attention modules such as contextual attention [11] and coherent semantic attention [12] are often used. Unlike single-stage painting, these methods tend to generate a coarse or low-resolution image, which is then further refined or generated in high resolution. However, both single-stage painting and coarse-to-fine progressive painting methods often do not take full advantage of a priori knowledge for accurate texture inference.

Therefore, many improved methods have been proposed to guide GAN network models for refined image generation by considering image contours and adding structural priors. This results in reconstructed images with a more reasonable texture structure and accurate semantic information. There are two main categories of image inpainting methods based on prior knowledge: contour edge guided image inpainting [13], [14] and generative prior guided image inpainting [15], [16]. So, it is not rare that some additional generation tools or pre-

trained models are used to assist in the refinement of images. For example, DeepCut is used to predict a salient object mask [13], and the Canny edge detector is used to generate the edge map of the input image [14]. In general, the existing inpainting techniques based on deep learning are divided into single-stage inpainting and progressive inpainting. With the development of deep learning in the field of image inpainting, multi-stage network architectures have replaced single-stage models as the mainstream. Although the multi-stage model improves the problem of long-term distant information correlation that cannot be overcome by the single-stage, there is still room for further improvement.

3. Proposed Method

Our proposed method is a three-stage network consisting of a pre-trained pSp-styleGAN [3], an identity feature recovery network (IFR) and a structural feature refinement network (SFR). As shown in **Fig. 1**, we use the pre-trained pSp-styleGAN as the first step of the model. IFR and SFR are used as the second and third steps of the model, respectively. As the input images are injected, the pSp-styleGAN will synthesize a coarse result image with a rich structure. Then, IFR further recovers the face identity feature from the rough results. Finally, SFR extracts and repairs the overall structural features of the image.

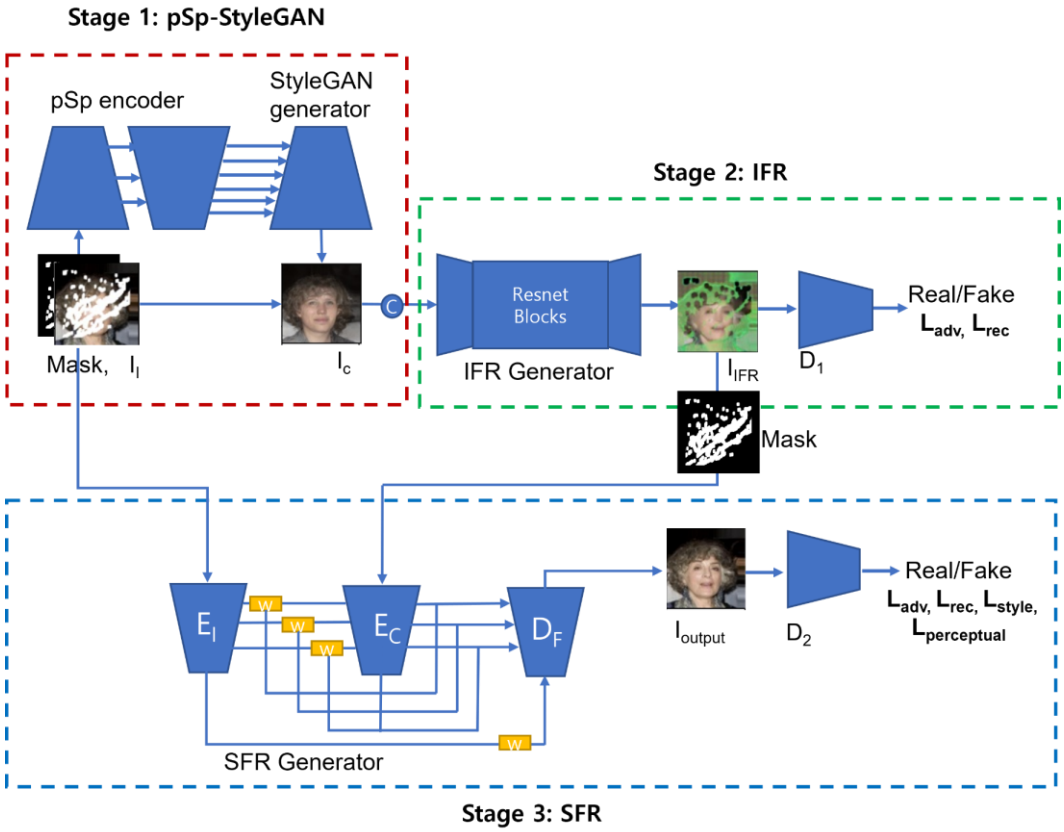

Fig. 1. Overview of ISFRNet

3.1 Stage One: Pretrained pSp-StyleGAN

The original pSp-StyleGAN is based on a standard pyramidal encoder (pSp) that generates a series of style vectors directly and then injects them into a pre-trained StyleGAN generator. pSp is used to extract a total of 18 target styles with ResNet as the backbone. 0-2 styles are extracted from the small feature map, 3-6 styles are from the medium feature map, and then 7-8 styles are extracted from the large feature map. Then, each style generates 512 vectors through an intermediate mapping network, which is injected into StyleGAN through affine transformations. Finally, StyleGAN generates the corresponding image based on the styles extracted from the input image. This model can solve a wide range of image-to-image translation tasks such as multi-modal conditional image synthesis, facial frontalization, inpainting, and super-resolution. However, when we apply this model to image inpainting, we find that although pSp-StyleGAN can generate complete and realistic images, it is difficult to maintain the same identity as the input image, as shown in the images of **Fig. 2** (c). Thus, we add two refinement networks based on the pSp-StyleGAN model to refine the identity feature and structural feature, respectively.

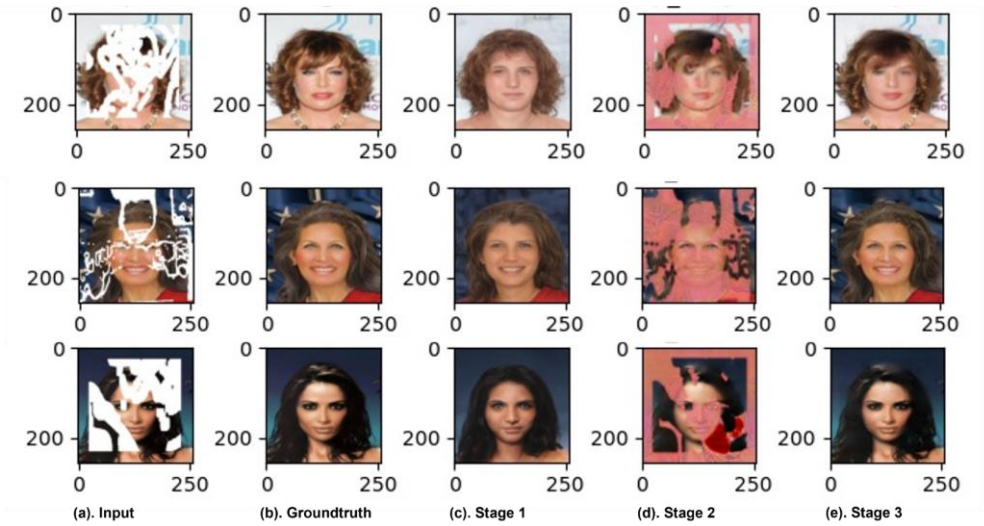

Fig. 2. Examples of the output image for each stage

3.2 Stage Two: IFR

Since the identity feature consists of the local structure and textures, we apply a shallow network (IFR) with a small receptive field in the second step (see **Fig. 3**). The purpose is to solve the problem where pSp-StyleGAN cannot maintain the identity feature in the first stage. The IFR generator is composed of an input layer, two downsampling layers, four residual blocks, two upsampling layers, and an output layer (Tanh as activation). D_1 and D_2 are depicted in **Fig. 4** as discriminators for IFR and SFR, respectively. They are composed of 5 convolutional blocks, where the first four blocks consisting of a convolutional layer, spectral normalization, the LeakyReLU. For the last block, we use a sigmoid for activation. We focus on repairing the information in the missing regions while restoring the identity features of the input image. The output images of IFR are shown in **Fig. 2** (d). As we can see, compared to the results of pSp-StyleGAN, the identity features and information within the mask of the image are restored by IFR.

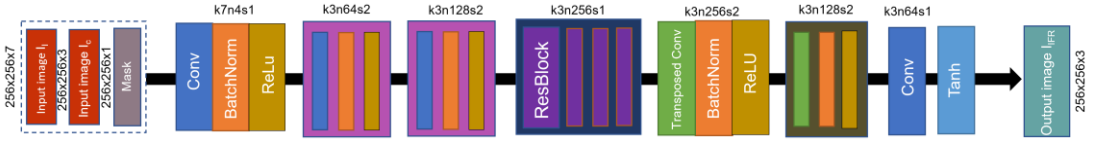

Fig. 3. The network architecture of the IFR generator, where k is the kernel size, n is the number of channels and s is the stride for each convolutional layer.

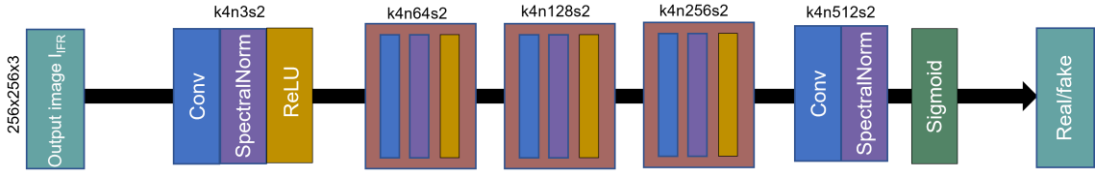

Fig. 4. The architecture of the D_1 and D_2 discriminator.

3.3 Stage Three: SFR

In the second stage, we recover the identity features of the image and some detailed features within the mask. To further refine the structural features of the output image from the second stage, we propose SFR, the architecture of which is depicted in **Fig. 5**. SFR is based on the original U-Net model by splitting the original encoder into two encoders and then connecting a decoder. The two encoders extract the structural features of the I_I image and I_{IFR} image respectively, and then generate the final output image with decoder D_F. A weighting mechanism is proposed in a part of the skip connection between the two encoders and the decoder. The available structural features in the input image decrease as the missing region increases. So, this mechanism is designed to be inversely proportional to the area of the missing region and then multiplied by matrix multiplication with the feature map of the input image to reduce the weight. This mechanism is demonstrated in (1), which is designed to be simple but effective:

$$\mathcal{F}_i^E = Concat(\frac{1}{N_M} \odot \mathcal{F}_i^{E_I}, \mathcal{F}_i^{E_C}) \tag{1}$$

Here, $\odot$ is the element-wise product operation, $\mathcal{F}_i^{E_I}$ and $\mathcal{F}_i^{E_C}$ are the feature maps extracted from the two encoders respectively, and N_M is the number of elements in the missing region.

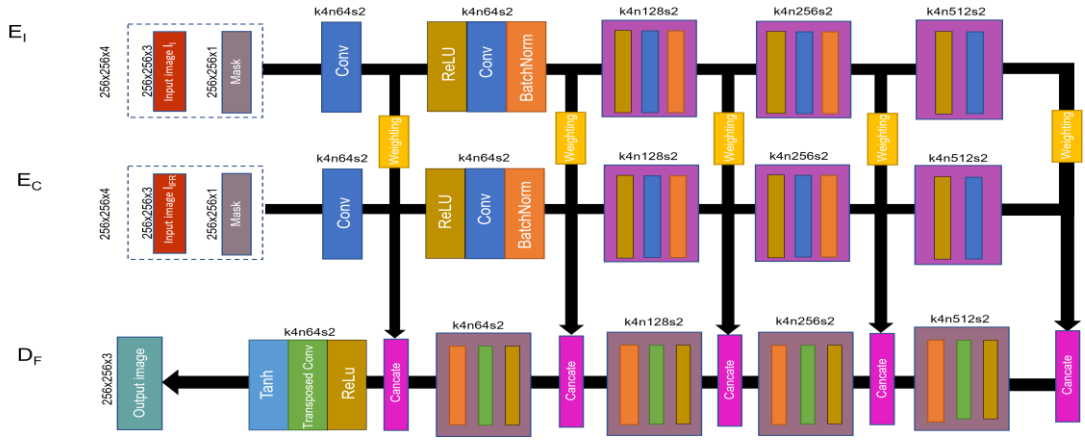

Fig. 5. The architecture of the SFR generator.

3.4 Loss Function

The second stage of the framework (IFR) is trained to recover the identity features of the face while generating better details of the missing parts. We use the reconstruction and adversarial losses. For the reconstruction loss (2), we use the L1 loss:

$$L_{rec}^{IFR} = \left\|\left(I_o - I_{gt}\right)\odot M\right\|_1 \tag{2}$$

Here, I_o is the output image from the IFR, I_{gt} is the ground truth image, $\odot$ is the element-wise product operation, and the M is a binary mask with an internal pixel value of 1 and an external pixel value of 0. The adversarial loss is defined in (3) and (4):

$$L_{D_1} = E_{I_{gt}}\left[\log D_1\left(I_{gt}\right)\right] + E_{I_c}\log[1 - D_1(I_c)] \tag{3}$$

$$L_{adv}^{D_1} = E_{I_c}\log[1 - D_1(I_c)] \tag{4}$$

Here, I_c is used as an input image of IFR. The total loss function for IFR training is defined as (5):

$$L_{total}^{IFR} = \lambda_1 L_{rec}^{IFR} + \lambda_2 L_{adv}^{D_1} \tag{5}$$

For our experiments, we use $\lambda_1 = 1$ and $\lambda_2 = 0.1$.

The third stage of the framework (SFR) is trained to further reconstruct the structural features of the full image. We need to ensure both structural integrity and clarity. Therefore, in addition to using the same adversarial loss function as the IRF, we add the reconstruction without the mask constraint, perceptual, style, and total variation (TV) loss functions. As with most image inpainting models, the perceptual and style loss functions we can apply are defined in VGG-16 and pre-trained on the ImageNet dataset. The reconstruction, perceptual and style loss can be written as (6), (7) and (8) respectively:

$$L_{rec}^{SFR} = \left\|I_{output} - I_{gt}\right\|_1 \tag{6}$$

$$L_{\text{perceptual}} = \sum_{p=0}^{P-1} \frac{\left\|\Psi_p^{I_{out}} - \Psi_p^{Igt}\right\|_1}{N_{\Psi_p^{Igt}}} \tag{7}$$

$$L_{\text{style}} = \sum_{p=0}^{P-1} \frac{\left\|\mathrm{K}_p\left(\left(\Psi_p^{I_{out}}\right)\left(\Psi_p^{I_{out}}\right) - \left(\Psi_p^{Igt}\right)\left(\Psi_p^{Igt}\right)\right)\right\|_1}{C_p C_p} \tag{8}$$

Here, I_{output} is the final output image from the third step of the framework, $\Psi_p^{I_*}$ denotes the activation of the p-th selected pre-trained VGG-16 layer. Here we use the 5th, 10th, and 17th layers of VGG-16 for calculations. In the style loss function $\left(\Psi_p^{I_{out}}\right)\left(\Psi_p^{I_{out}}\right)$ represents an auto-correlation (Gram matrix) is applied to each selected VGG feature map, $\left(C_p, H_p, W_p\right)$ is the shape of $\Psi_p^{I_*}$, and K_p equals $1/C_p\, H_p W_p$ is used for normalization. We also use the total variance (TV) loss as a smoothing penalty. It is expressed as (9):

$$L_{tv} = \sum_{(i,j)\in R,(i,j+1)\in R} \frac{\left\| I_{out}^{i,j+1} - I_{out}^{i,j} \right\|_1}{N_{I_{out}}} + \sum_{(i,j)\in R,(i+1,j)\in R} \frac{\left\| I_{out}^{i+1,j} - I_{out}^{i,j} \right\|_1}{N_{I_{out}}} \quad (9)$$

Here, R denotes a 1-pixel dilated hole region and $N_{I_{out}}$ is the number of all elements in final output image I_{out}. The full loss function for training of the SFR can be expressed by the formula (10):

$$L_{total}^{SFR} = \lambda_3 L_{rec}^{SFR} + \lambda_4 L_{per} + \lambda_5 L_{sty} + \lambda_6 L_{tv} + \lambda_7 L_{adv}^{D_1} \quad (10)$$

Here, λ_3, λ_4, λ_5, λ_6, and λ_7 are equal to 50, 120, 0.05, 0.1, and 1, respectively.

4. Experimental Results and Analysis

4.1 Experimental Data and Platform

During training, we scale the pixel values of the mask between [0, 1]. We train our model on an NVIDIA TITAN Xp GPU with an image resolution of 256 × 256 and a batch size of 1. We use Adams optimization with default momentum parameters and an initial learning rate of 1×10^{-4}.

We evaluate our network on a publicly available face dataset CelebA [17]. The CelebA is a dataset that includes 202,599 face images. We randomly select 182,339 images for training and 20,260 images for testing. For the mask dataset, we use the test set of NVIDIA Irregular Mask Dataset [18] to randomly combine with our CelebA training and test sets. This mask dataset contains 6000 masks with border constraints and 6000 irregular masks without border constraints, and it is evenly divided into six categories (0%-10%, 10%-20%, 20%- 30%, 30%-40%, 40%-50%, and 50%-60%) based on the size of the random missing region. Additionally, all images are resized to 256×256. We use the following metrics to measure the quality of our results: 1) structural similarity index (SSIM), 2) peak signal-to-noise ratio (PSNR), 3) L1 Loss and 4) learned perceptual image patch similarity (LPIPS) [19]. The higher the PSNR and SSIM values the better, however L1 Loss and LPIPS are the smaller the value the better.

4.2. Ablation Experiment

To clearly illustrate and analyze the impact of each stage of our three-stage deep network on the generation of the final image, we perform three ablation experiments.

4.2.1. Experiment 1: pSp + IFR

Experiment 1 was performed without using the SFR refinement network. The architecture of the compared method used pSp-StyleGAN as the first stage and only IFR as the second stage for identity feature recovery. The dataset used for training and processing remain the same as our ISFRNet. However, for the loss function, unlike the reconstructed loss function used in ISFRNet, we remove the mask constraint. This causes the obtained images to be more complete for a more objective and intuitive comparison. The purpose of Experiment 1 is to demonstrate the refinement effect of SFR on the structural features of the images.

Fig. 6 shows a visualization comparison of the results in Experiment 1. In the first row of **Fig. 6**, our model performs well on the contours of the teeth and the left eye, as compared to the pSp + IFR results in **Fig. 6** (b). The second row also shows that our model achieves better performance in repairing the mouth structure of the portrait and the recessed areas of the face.

For convenience, we selected 20%-40% and 40%-60% masked images for PSNR and SSIM score evaluation. These results are shown in **Table 1**. The PSNR and SSIM scores of our model are significantly higher than the network with StyleGAN and IFR.

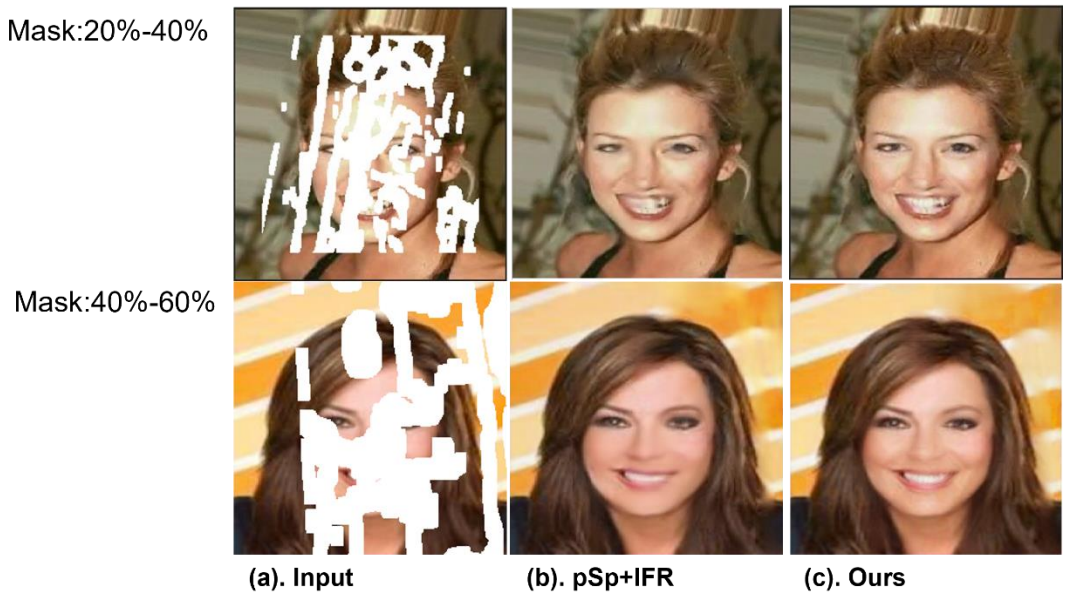

Fig. 6. Examples of the output image for pSp+IFR and ours

Table 1. Quantitative comparisons [PSNR(dB)/SSIM] between the compared methods and our approach

Mask	pSp [3]	IFR	SFR	Weighting	PSNR	SSIM
20-40%	√	√			27.81	0.902
40-60%					17.77	0.682
20-40%	√		√		17.76	0.683
40-60%					13.61	0.367
20-40%	√	√	√		26.73	0.915
40-60%					22.05	0.792
20-40%	√	√	√	√	**28.12**	**0.942**
40-60%					**13.31**	**0.840**

4.2.2 Experiment 2: pSp + SFR

Experiment 2 was performed without using IFR to refine the network. The overall framework used pSp-StyleGAN as the first stage and only SFR as the second stage. The processing of data, loss functions and training settings are the same as the processing for the ISFRNet. The input image I_I is injected into the pre-trained pSp-StyleGAN to generate the image I_c. Then, the structure features of I_I and I_c are extracted by E_I and E_C in SFR, respectively. Finally, a decoder D_F is connected to generate the result images, as **Fig. 7** (b). Since there is no SFR refinement, the result images are not detailed enough in terms of the texture and the identity feature of the image is not well recovered, especially in the first row. We apply PSNR and SSIM scores to evaluate the models. These results are shown in the fourth column of **Table 1**.

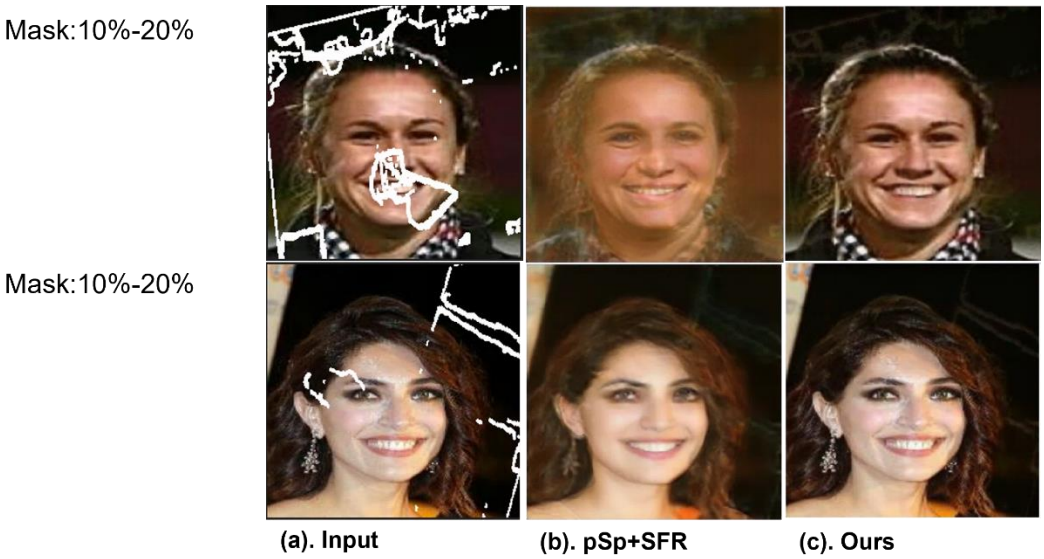

Fig. 7. Examples of output images for pSp + SFR and our approach

4.2.3 Experiment 3: pSp + IFR + SFR (without weighting mechanism)

Experiment 3 was performed by removing our weighting mechanism, but without changing any other conditions. The goal of this experiment is to verify whether our weighting mechanism is able to balance the structural features extracted between the two encoders, thus reducing in the negative impact of the missing regions. **Fig. 8** and **Table 1** show the visualization and quantitative comparison results, respectively. **Fig. 8** (b). remains partially masked with shadows and blurring due to the lack of weight balance of the feature map by the weighting mechanism. In **Table 1**, the PSNR and SSIM scores of our model are also much higher than those of the model without the weighting mechanism.

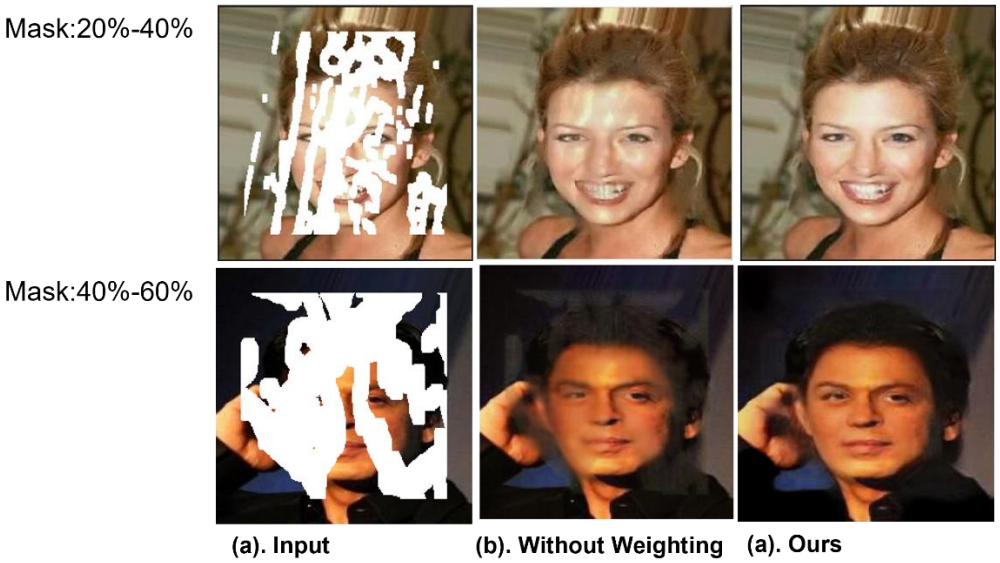

Fig. 8. Examples of output images without a weighting mechanism and our approach.

4.3. Comparison with the State-of-the-art Method:

In Experiment 4, we compared our method with three other coarse-to-fine painting methods. The visualization comparison is provided in **Fig. 9**. The first row of **Fig. 9** shows a scenario where the missing area of the input image is very large. We can observe that our method is able to maximize the complementary structural features while retaining good identity features. When the missing area of the input image is relatively small as in the second and third rows of **Fig. 9**, the result images of our method are much clearer and detailed with almost no artificial traces, as compared to LGNet and MADF model. The results of the quantitative comparison are shown in **Table 2**, We can see that our model is competitive in terms of PSNR, SSIM and L1 Loss. Although, when the missing region is relatively small, our model shows comparable performance compared to the Edgeconnect (EC) model. But, if we compare the results in the first row of **Fig. 9**, we can see that EC generates a very serious artificial structural feature, while our model generates a relatively complete structural feature. More comparisons with EC model are shown in **Fig. 10**. Our model focuses on balancing the refinement of structure and identity features to produce a more complete structure while maintaining identity consistency as much as possible. Meanwhile, it cannot be ignored that our model is not as good as EC and LGNet model in LPIPS measurement. Therefore, we measured and compared the LPIPS scores of the images separately. In Fig. 9, the LPIPS scores of the three output images from LGNet are 0.163, 0.035, and 0.065. The LPIPS scores of the three outputs from EC model are 0.135, 0.202, and 0.121. The three outputs from our model are 0.181, 0.081, and 0.093, respectively. We are able to observe that the results of our model are visually better than other models, but the LPIPS score is bad. In fact, whether the image is more realistic or not does not depend entirely on the LPIPS score. Since our model suffers from a certain degree of blurring, when the image is fed into the pre-trained Alex network to calculate the LPIPS distance, the blurred part will largely affect the calculation of LPIPS [19].

Fig. 9. Visualization comparison of our approach (ISFRNet) with LGNet [3], MADF [16], EC [14].

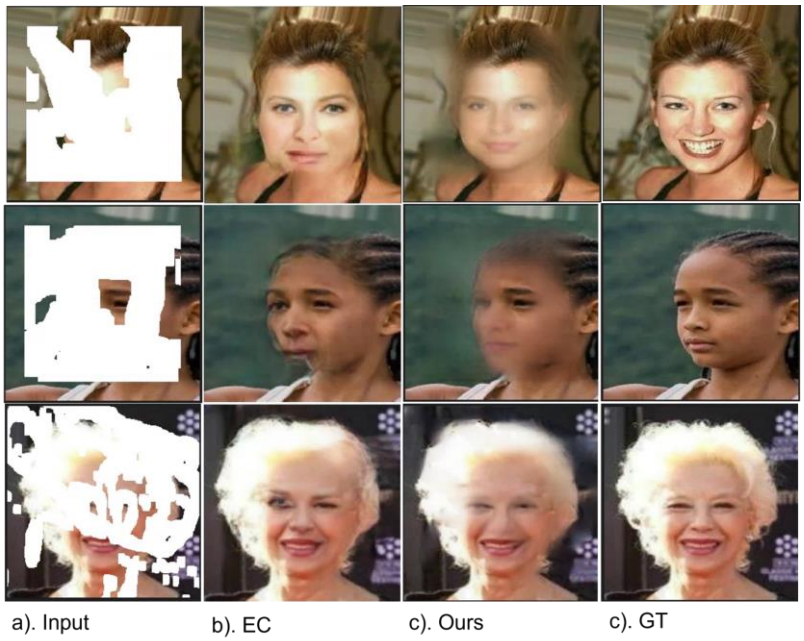

Fig. 10. Visualization comparison of our approach and EC [13] in the case of very large missing regions.

Table 2. Quantitative comparisons [PSNR(dB)/SSIM/ L_1 Loss/ LPIPS]

	Mask	pSp [3]	LGNet [4]	MADF [20]	EC [14]	Ours
PSNR	20%-40%	18.66	27.99	22.70	28.11	28.12
	40%-60%	17.53	22.93	19.72	22.83	23.31
SSIM	20%-40%	0.651	0.939	0.874	0.946	0.942
	40%-60%	0.581	0.832	0.764	0.836	0.840
L_1 Loss	20%-40%	0.113	0.020	0.043	0.029	0.015
	40%-60%	0.151	0.063	0.071	0.067	0.053
LPIPS	20%-40%	0.102	0.074	0.113	0.054	0.090
	40%-60%	0.205	0.144	0.183	0.126	0.177

5. Conclusion

In this paper, we present a three-stage adversarial network for face image inpainting. The first step is based on the pSp-StyleGAN model, which generates very detailed and realistic images regardless of the size of the corrupted area of the injected images. This ability simulates how people can imagine a complete image according to their own preferences when they look at an image with a large missing region. However, when the missing area is relatively small (i.e., valid information is sufficient), it is more important to keep the identity feature of the face. So, we apply the IFR and SFR as the refinement networks for the second and third stages. The initial input image is reused in SFR. However, as the area of the missing regions in the input image increases, the available effective structural information gradually decreases, so a weighting mechanism is applied to balance the weights of the features in the input image. The experimental results show that our network is effective and has outstanding and superior performance compared to other multi-refinement models. Although the problem of blurring still exists when the missing area is large. I believe that in the future, it is necessary to further improve the model's ability to use the valid pixel information and make the model generate new information autonomously.

Acknowledgement

This work was partly supported by the National Research Foundation of Korea (NRF) grant funded by the Korea government (MSIT) (NRF-2020R1F1A1065626) and was partly supported by the MSIT (Ministry of Science and ICT), Korea, under the Information Technology Research Center (ITRC) support program (IITP-2023-2018-0-01798) supervised by the Institute for Information & communications Technology Promotion(IITP).

References

[1] Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Bengio, Y., "Generative adversarial networks," *Communications of the ACM*, 63(11), 139-144, 2020. Article (CrossRef Link)

[2] Karras, T., Laine, S., & Aila, T., "A style-based generator architecture for generative adversarial networks," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 43, no. 12, pp. 4217-4228, 2021. Article (CrossRef Link)

[3] Richardson, E., Alaluf, Y., Patashnik, O., Nitzan, Y., Azar, Y., Shapiro, S., & Cohen-Or, D., "Encoding in style: a stylegan encoder for image-to-image translation," in *Proc. of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 2287-2296, 2021. Article (CrossRef Link)

[4] Quan, W., Zhang, R., Zhang, Y., Li, Z., Wang, J., & Yan, D. M., "Image inpainting with local and global refinement," *IEEE Transactions on Image Processing*, vol. 31, pp. 2405-2420, 2022. Article (CrossRef Link)

[5] Ronneberger, O., Fischer, P., Brox, T., "U-Net: Convolutional Networks for Biomedical Image Segmentation," in *Proc. of Medical Image Computing and Computer-Assisted Intervention—MICCAI 2015*, pp. 234–241, 2015. Article (CrossRef Link)

[6] Pathak, D., Krahenbuhl, P., Donahue, J., Darrell, T., & Efros, A. A., "Context encoders: Feature learning by inpainting," in *Proc. of the IEEE conference on computer vision and pattern recognition*, pp. 2536-2544, 2016. Article (CrossRef Link)

[7] Iizuka, S., Simo-Serra, E., & Ishikawa, H., "Globally and locally consistent image completion," *ACM Transactions on Graphics (ToG)*, 36(4), 1-14, 2017. Article (CrossRef Link)

[8] Zeng, Y., Fu, J., Chao, H., & Guo, B, "Learning pyramid-context encoder network for high-quality image inpainting," in *Proc. of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 1486-1494, 2019. Article (CrossRef Link)

[9] Liu, H., Jiang, B., Song, Y., Huang, W., & Yang, C., August, "Rethinking image inpainting via a mutual encoder-decoder with feature equalizations," in *Proc. of European Conference on Computer Vision*, pp. 725-741, 2020. Article (CrossRef Link)

[10] Qin, Z., Zeng, Q., Zong, Y., & Xu, F, "Image inpainting based on deep learning: A review," *Displays*, 69, 102028, 2021. Article (CrossRef Link)

[11] Yu, J., Lin, Z., Yang, J., Shen, X., Lu, X., & Huang, T. S., "Generative image inpainting with contextual attention," in *Proc. of the IEEE conference on computer vision and pattern recognition*, pp. 5505-5514, 2018. Article (CrossRef Link)

[12] Liu, H., Jiang, B., Xiao, Y., & Yang, C., "Coherent semantic attention for image inpainting," in *Proc. of the IEEE/CVF International Conference on Computer Vision*, pp. 4170-4179, 2019. Article (CrossRef Link)

[13] Xiong, W., Yu, J., Lin, Z., Yang, J., Lu, X., Barnes, C., & Luo, J., "Foreground-aware image inpainting," in *Proc. of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 5840-5848, 2019. Article (CrossRef Link)

[14] Nazeri, K., Ng, E., Joseph, T., Qureshi, F., & Ebrahimi, M., "Edgeconnect: Structure guided image inpainting using edge prediction," in *Proc. of the IEEE/CVF International Conference on Computer Vision Workshops*, pp. 3265-3274, 2019. Article (CrossRef Link)

[15] Lahiri, A., Jain, A. K., Agrawal, S., Mitra, P., & Biswas, P. K, "Prior guided gan based semantic inpainting," in *Proc. of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 13696-13705, 2020. Article (CrossRef Link)

[16] Pan, X., Zhan, X., Dai, B., Lin, D., Loy, C. C., & Luo, P., "Exploiting deep generative prior for versatile image restoration and manipulation," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 44, no. 11, pp. 7474-7489, 2022. Article (CrossRef Link)

[17] Liu, Z., Luo, P., Wang, X., & Tang, X., "Deep learning face attributes in the wild," in *Proc. of the IEEE international conference on computer vision*, pp. 3730-3738, 2015. Article (CrossRef Link)

[18] Liu, G., Reda, F. A., Shih, K. J., Wang, T. C., Tao, A., & Catanzaro, B, "Image inpainting for irregular holes using partial convolutions," in *Proc. of the European conference on computer vision (ECCV)*, pp. 85-105, 2018. Article (CrossRef Link)

[19] Zhang, R., Isola, P., Efros, A. A., Shechtman, E., & Wang, O., "The unreasonable effectiveness of deep features as a perceptual metric," in *Proc. of the IEEE conference on computer vision and pattern recognition*, pp. 586-595, 2018. Article (CrossRef Link)

[20] Zhu, M., He, D., Li, X., Li, C., Li, F., Liu, X., Zhang, Z., "Image inpainting by end-to-end cascaded refinement with mask awareness," *IEEE Transactions on Image Processing*, 30, 4855-4866, 2021. Article (CrossRef Link)

Yan Wang received the B.S. degree in electronic and electrical engineering from Sungkyunkwan University, Suwon, Korea, in 2021. She is a M.S. candidate in electronic and electrical computer engineering from Sungkyunkwan University, Suwon, Korea. Her research interests include computer vision, deep learning, image inpainting, and artificial intelligence.

Jitae Shin has been a Professor from 2002 with the School of Electronic and Electrical Engineering, Sungkyunkwan University, Suwon, South Korea. He received the B.S. degree from Seoul National University, in 1986, the M.S. degree from the Korea Advanced Institute of Science and Technology, in 1988, and the M.S. and Ph.D. degrees in electrical engineering from the University of Southern California, Los Angeles, USA, in 1998 and 2001, respectively. For former industrial experiences, he worked with Korea Electric Power Corporation and the Korea Atomic Energy Research Institute from 1988 to 1996. His current research interests include image/video signal processing using deep learning, medical image processing, and machine learning for wireless/mobile communication systems.