Generative Face Completion via edge learning and semantic attention

CAO SHILEI[1,a)] DANILO VASCONCELLOS VARGAS[2,b)] KOUICHI SAKURAI[2,c)]

Abstract: Face completion, which is to reproduce the missing region of an incomplete face image, have yield significant improvements through adaptation of neural network models. However, since the the discontinuity of the local pixels, the existing methods often fail to fill in semantically plausible and context aware details. To handle this problem, we propose a two-stage adversarial framework, for filling the missing regions with sharp edges and semantically plausible textures. Specifically, our model learns to predict the edges first, and then completion network fills the missing region using the predicted edges as a priori. To this end, we propose a novel U-net architecture, which including coherent semantic attention parts. Semantic attention parts not only preserve contextual structure but also serve as the estimation of the missing parts. We applied our model in generative face completion, and evaluated our method on CelebA datasets.

Keywords: Image Processing , Generative Advrsarial Network, Face Completion

1. Introduction

Face completion, which can be considered as a subset of image inpainting, is a research hotspot in computer vision and machine learning communities. It refers to reconstructing face images which have missing regions and can be utilized in many applications such as removing unwanted objects from face pictures or restoring damaged photographs. The core challenge of face completion is to maintain global semantic structure and generate realistic details.

In general, traditional image inpainting methods are based on the assumption that the missing area should contain similar patterns of the background region. Those methods typically fill missing pixels by matching and pasting patches based on low level features such as mean square difference of RGB, values or SIFT descriptors and so on. These methods recover well for highly repetitive texture images, but fail to produce images with complex structures. For example, as one of the once state-of-the-art methods, the PatchMatch[2] matches and copies the background patches into holes starting from low-resolution to high-resolution or propagating from hole boundaries. While this approach generally produces smooth results, especially in background inpainting tasks, it is limited by the available image statistics and not able to capture high-level semantics or global structure of the image. Furthermore, as the traditional diffusion-based and patch-based methods assume missing patches can be found somewhere in the background regions, they cannot produce novel image contents for complex inpainting regions where involve intricate structures like faces [3].

Over the last few years, the deep learning based methods have emerged as a promising alternative avenue by treating the problem as learning an end-to-end mapping from masked input to completed output. Especially, the adaption of deep convolutional neural networks (CNN)[5] and generative adversarial networks (GAN)[6] contributed to the vigorous development of inpainting. Context-encoder [4] is one of the first works that apply deep neural networks for image inpainting which cooperates autoencoders and GAN for image inpainting. this framework achieved amazing results, However, it only uses the local discriminator, although local discriminator facilitates exposing the local structure details, local discriminator only has main drawback that it pushes the generative network to produce independent textures that are incompatible with the whole image semantics Yang et al. [7] proposes to use style transfer for image inpainting. More specifically, it initializes the hole with the output of context-encoder, and then improves the texture by using style transfer techniques to propagate the high-frequency textures from the boundary to the hole. ACM2018 Semantic Inpainting[8] is an framework introduced LSTM to string all the subtaskes together. in this way, the essence learned from the previous subtasks are exploited to ease the learning of subsequent subtasks, which can also discribed as: this paper introduced LSTM architecture into PGN to string all subtasks to control the information flow in the PGN. It is able to avoid the information disturbed and improve the quality of the inpainting image

While these learning-based methods are significantly more effective in capturing high-level features than prior techniques, they still only useful for handling very low-solution inputs due to the

[1] Department of Informatics Graduate School of Information Science and Electrical Engineering, Kyushu University
[2] Department of Informatics Faculty of Information Science and Electrical Engineering, Kyushu University
[a)] 2IE19026W@s.kyushu-u.ac.jp
[b)] vargas@inf.kyushu-u.ac.jp
[c)] sakurai@inf.kyushu-u.ac.jp

memory limitations and difficulty in training. Even for slightly lager images, the inpainted regions would appear blurry and unpleasant boundries become visible. In order to achieve more fine-detail images, especiallly for face images, we proposed our framework. In our structure, we divided the inpainting process into two steps. The first step can be constructed by training network to detect the foreground contour of the corrupted image, and then completes the missing contours of the foreground objects with a contour completion module, rough out the missing contents. Then the refinement network leverages accurate contour prediction to guide image completion. In the second network, both predicted edges and original images with holes are fed to the U-net architecture. In order to get the filling holes semantic consistent, we propose a novel U-net architecture with pyramid semantic attention filling block. This semantic attention filling block can be considered as prediction and noises which can capture more effective features. With the purpose of maintaining more semantic information, we apply pyramid structure to semantic attention filling block. Our specific U-net connection utilizes high-level contextual information to fill the hole regions of low-level encoder feature maps progressively from bottle-neck layer to up. The fully filled encoder feature maps are concatenated with the corresponding decoder feature by skip connection. It can avoid the transmission of invalid information in hole regions of encoder feature maps when using skip connection, and make the model perceive high-level contextual information.

Based on our framework, we design our loss function specially, including consistence loss, reconstruction loss, edge loss, style loss and adversarial loss. To summarize, our contributions are as follows:

1.An end-to-end trainable network that combines edge generation and image completion to fill in missing regions exhibiting fine details.

2.A novel U-net architecture with pyramid semantic attention block fully filled in encoder feature maps.

3.A novel loss function for semantically plausible and context aware details

2. Related Work

2.1 GAN and improvement

Adversarial Neural Network (GAN) is a framework for estimating generative models via adversarial nets which first proposed by Ian Goodfellow in 2014.This framework corresponds to a minimax two-player game, looking for a point called Nash equilibrium that is simultaneously a minimum of the defending players cost and a maximum of the attackers attacking player cost. to Since the improvement of computing speed and the development of hardware, deep learning methods are widely used in image process. Deep learning and GAN-based approaches have emerged as a promising paradigm for image inpainting. However, since GAN was proposed in 2014, there have been some problems such as difficulty in training, difficulty in convergence, loss of generator and discriminator that cannot indicate the training process, and lack of diversity of generated samples. Since then, many researchers have proposed improvements that actually address some of the problems, such as DCGAN[9],Alec et al. introduced CNN into generator and discriminator

Wasserstein GAN (WGAN) [10],Martin Arjovsky et al. used Wasserstein distance (also known as Earth Mover distance) to replace jenson-shannon divergence. In this way, the gradient disappearance problem is solved theoretically. In addition, WGAN also theoretically gives the reason for mode collapse in naive GAN. WGAN-GP [11] optimized the structure of WGAN, using gradient penalty instead of weight clipping for satisfying the weight pruning constraint. BEGAN [12], introduce Proportional Control Theory to the GAN for more stable convergence. LSGAN [13] Mao et al. proposed least square GAN. The main idea is to provide a smooth and unsaturated gradient loss function for discriminator D for the sake of improving picture quality.

More relevant tasks to inpainting is conditional image generation. For example, Pix2Pix GAN[14], Phillip Isola proposed a networks structure not only learn the mapping from input image to output image, but also learn a loss function to train this mapping. This makes it possible to apply the same generic approach to problems that traditionally would require very different loss formulations. PatchGAN [14] rather the regular GAN maps from the image to a single scalar output, which signifies real or fake, the PatchGAN maps to an array of outputs each number in the array signifies whether the patch in the image is real or fake. It is widely used in discriminators. CycleGAN [15] consists of two pairs of generators and discriminators, and used cycle-consistent to map unpaired data.

2.2 Image inpainting with GAN

Using deep neural network for image inpainting has also been started by Pathak et al. [4], which architecture cooperate autoencoder and GAN for generating the predicted image. Iizuka et al. [16] propose local and global discriminators, assisted by dilated convolution [17] to improve the inpainting quality and to handle rectangular masks at any location. However, it requires the previous processing steps to enforce the color coherency near the hole boundaries. Guilin Liu et al. [18]proposed gated convolution, which can repair any non-central and irregular region. Hongyu Li et al. [19] put up with a fined deep generative model-based approach with a novel coherent semantic attention (CSA) layer, which can not only preserve contextual structure but also make more effective predictions of missing parts by modeling the semantic relevance between the holes features.

2.3 Attention Modeling

Attention model is first proposed by Bahdanau et.al.[20], Inspired by how human pay visual attention to different regions of an image. Human visual attention allows us to focus on a certain region, while ignoring some details of the surrounding regions, and then adjust the focal point or do the inference accordingly.

Due to this mechanism, when computing resources limited, attention modeling can be applied to solve the problem of information overload, thus it may allocate the computing resources to more important tasks.

In recent years, the attention model based on the relationship between the surrounding contextual and masked regions is widely used for inpainting tasks. Contextual Attention [21]proposes a

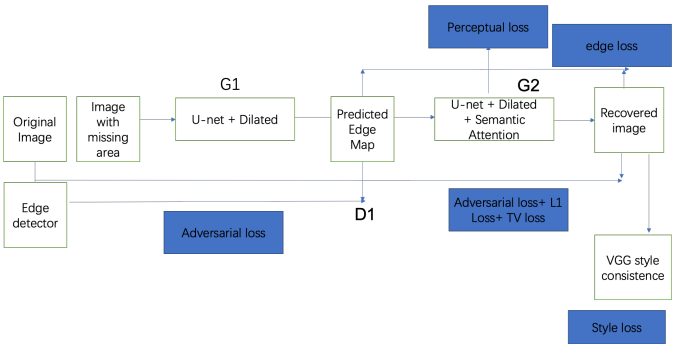

Fig. 1 The framework of the Network

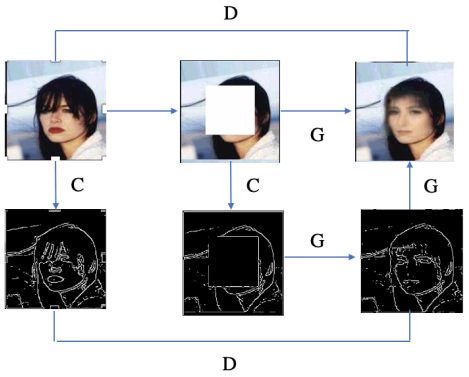

Fig. 2 Contour Detection and generator

contextual attention layer which searches for a collection of background patches with the highest cosine similarity to the coarse prediction. Yan et al. [22] introduce a shift-net powered by a shift operation and a guidance loss. The authors use shifted encoder features to estimate the prediction areas in the middle layer of decoder. Song et al. [23] introduce a patch-swap layer, which replaces each patch inside the missing regions of a feature map with the most similar patch on the contextual regions, and the feature map is extracted by VGG network. Ning Wang et al. [24] use a multi-scale image contextual attention learning for inpainting. Hongyu Li et al. [25]optimized the coherent semantic attention. generate the predicted patches with the consideration of surrounding predicted patches.

3. Approach

The purpose of our framework is to fill the incomplete image with a visually pleasing appearance. To this end, we adopt U-Net as the baseline network. It is a cascade of three modules: incomplete contour detection module, contour completion module and image completion module. In the following, we first introduce the guidance loss and Shift-Net, and then describe the model objective and learning algorithm.

3.1 Network structure

Our model consists of three steps: incomplete contour detect, contour complete and image complete. The overall framework of our inpainting system is shown in Fig. 1.

The contour completion model is composed of a generator and a discriminator. The generator is a coarse network aims to generate the complete edge image. This completed edges maps are not only the input of the refinement network but also the constraint of the recovered image.

There are many solutions for contour detection including traditional methods and deep learning methods including DeepCut, Holistically-Nested Edge Detection and so on. However, for the speed of image preprocessing, robustness, and ease of use, we adopt Canny edge maps.

Then we put the masked image and the edge image refinement network. This refinement network adopt our novel U-net structure which connection utilizes high-level contextual information to fill the hole regions of low-level encoder feature maps progressively from bottle-neck layer to up.

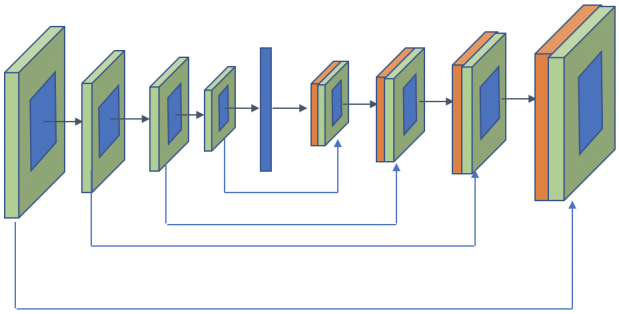

Fig. 3 Traditional U-net structure

3.2 U-net with pyramid semantic attention block

Recently, the spatial attention based on the relationship between contextual and hole regions is often used for image inpainting tasks. Contextual Attention [19] proposes a contextual attention layer which searches for a collection of background patches with the highest similarity to the coarse prediction. Yan et al. [22] introduce a shift-net powered by a shift operation and a guidance loss. The shift operation speculate the relationship between the contextual regions in the encoder layer and the associated hole region in the decoder layer. In this paper we propose a new novel semantic attention method.

3.2.1 U-Net

The U-net was developed by Olaf Ronneberger et al. for Bio Medical Image Segmentation. Now, this architecture is widely used for image process, which briefly showed in Fig.3. It contains two paths. First path is the contraction path (also called as the encoder) which is used to capture the context in the image. The encoder is just a traditional stack of convolutional and max pooling layers. The second path is the symmetric expanding path (also called as the decoder) which is used to enable precise localization using transposed convolutions. Thus it is an end-to-end fully convolutional network (FCN). Besides, in this symmetric architecture, the skip connection is introduced to concatenate the features from each layer of encoder and those of the corresponding layer of decoder. Such skip connection makes it convenient to utilize the information before and after bottleneck, which is valuable for image inpainting and other low level vision tasks in capturing localized visual details.

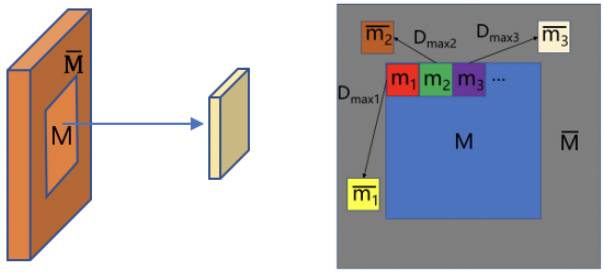

Fig. 4 semantic attention(SA)

3.2.2 Semantic Attention

In this paper, we applied semantic attention, also named as Shift-Net[22], to take into account the advantages of both exemplar-based and CNN-based methods for image inpainting.

In exemplar-based inpainting, the patch-based replication and filling process are iteratively performed to grow the texture and structure from the known region to the missing parts. And the patch processing order plays a key role in yielding plausible inpainting result. Guided by the salient structure produced by CNN, the filling process in the Shift-Net can be finished concurrently by introducing a shift-connection layer to connect the encoder feature of known region and the decoder feature of missing parts. Thus, our Shift-Net inherits the advantages of exemplar-based and CNN-based methods, and can produce inpainting result with both plausible semantics and fine detailed textures.

As shown in Figure 4, we first divide the M and $\overline{M}$ into patches. Then, we considers the similarity between features from the similarity of values. each neural patch in the hole M searches for the most similar neural patch on the boundary $\overline{M}$. In the last, we copy the information of the most similar patch in the $\overline{M}$ as the input of patch M. We measure with normalized inner product (cosine similarity)

$$D_{max_i} = \frac{\langle m_i, \overline{m_i} \rangle}{\|m_i\| * \|\overline{m_i}\|}$$

3.2.3 U-net with pyramid semantic attention block

However, for traditional U-net, it is obvious that the values of masked regions in the skip net is useless, which causes too much semantic information to be lost. Moreover, the feature information of each layer are from lower-level layer under the encoder-decoder architecture, which results in a lack of high-level semantics.

To solve this problem, we use semantic attention to borrow or copy feature information from known background patches to generate missing patches. It is differentiable, thus can be trained in deep models, and fully-convolutional, which allows testing on arbitrary resolutions.

For fully use of the semantic information, we propose a pyramid structure shows in Fig.5. In this figure, SA means the semantic attention method, UP means upsampling, since the information in the masked regions in the skip net is nearly 0, we use semantic attention to replaced these regions. Then we use a progressive strategy to fill remaining hole areas of other feature maps from deep to shallow. The original and filled feature maps are added by the short connection.

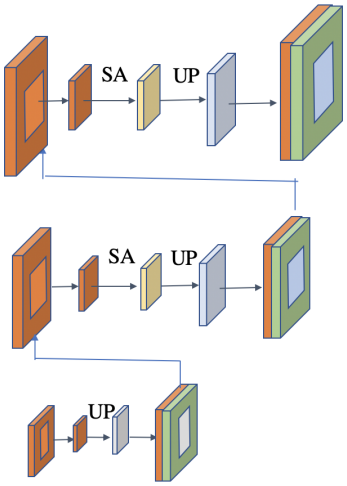

Fig. 5 pyramid semantic attention block

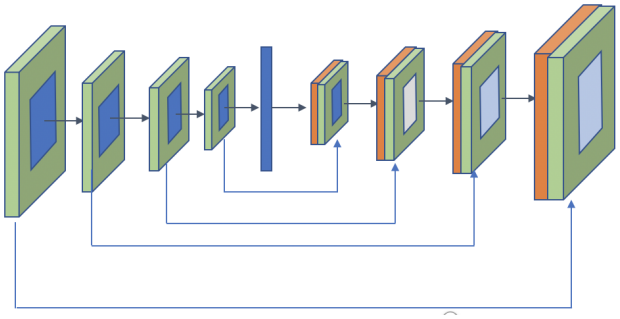

Fig. 6 U-net with pyramid semantic attention block

In the end, we get our novel U-net with pyramid semantic attention block. It is showed in Fig.6

3.3 Loss Function

Inspired by this process, we design our loss function specially, including consistence loss, reconstruction loss (TV+L_1), edge loss, style loss and adversarial loss. Here we assume I_o means the original image, I_r is the recovered image, I_m is the masked image, C_o means the original edge, C_m is the masked edge, C_r is the recovered edge.

3.3.1 TV loss

The Total Variation(TV) model can be regarded as the generalization of one-dimensional case, which was initially proposed by Rudin et al for image denoising, and then generalized to other image processing problems. TV(isotropic TV):

$$L_{TV} = \sum_{I_r} \sqrt{(s_{i+1,j} - s_{i,j})^2 + (s_{i,j+1} - s_{i,j})^2}$$

3.3.2 L_1 reconstruction loss

L_1 reconstruction loss is the distance between original image and recovered image with 1 norm.

$$L_{norm} = \|I_r - I_o\|_1$$

3.3.3 edge loss

The first U-Net is the adversarial network for generate the predict edges. Thus the edge loss is about adversarial loss.

$$L_{edge} = L_{C-GAN} + \lambda L_{C-L_1 loss}$$

3.3.4 style loss

When Convolutional Neural Networks are trained on object recognition, they develop a representation of the image that makes object information increasingly explicit along the processing hierarchy.Therefore, along the processing hierarchy of the network, the input image is transformed into representations that increasingly care about the actual content of the image compared to its detailed pixel values. Higher layers in the network capture the high-level content in terms of objects and their arrangement in the input image but do not constrain the exact pixel values of the reconstruction. In contrast, reconstructions from the lower layers simply reproduce the exact pixel values of the original image. We therefore refer to the feature responses in higher layers of the network as the content representation.

To obtain a representation of the style of an input image, we use a feature space originally designed to capture texture information. This feature space is built on top of the filter responses in each layer of the network. It consists of the correlations between the different filter responses over the spatial extent of the feature maps. By including the feature correlations of multiple layers, we obtain a stationary, multi-scale representation of the input image, which captures its texture information but not the global arrangement. Different from the direct operation of content representation, style representation uses the form of Gram matrix expanded into 1-dimensional vectors by Feature Map. The reason for using Gram matrix is that considering that the texture feature has nothing to do with the specific position of the image, this feature can be guaranteed by scrambling the position information of the texture. The definition of Gram matrix is as follows.

$$G^l_{i,j} = \sum_k F^l_{i,k} F^l_{j,k}$$

$$L_{style} = \frac{1}{r^2 WH} \Sigma_{x=1}^{rW} \Sigma_{y=1}^{rH} (I_{rx,y} - I_{ox,y})^2$$

3.3.5 adversarial loss

Wasserstein GAN (WGAN) makes progress toward stable training of GANs, but sometimes can still generate only low-quality samples or fail to converge due to the use of weight clipping in WGAN to enforce a Lipschitz constraint on the critic. We adopt WGAN-GP, which propose an alternative to clipping weights: penalize the norm of gradient of the critic with respect to its input. This method performs better than standard WGAN and enables stable training of a wide variety of GAN architectures with almost no hyperparameter tuning.

$$L = \underbrace{\mathop{\mathbb{E}}_{\tilde{\boldsymbol{x}} \sim \mathbb{P}_g} [D(\tilde{\boldsymbol{x}})] - \mathop{\mathbb{E}}_{\boldsymbol{x} \sim \mathbb{P}_r} [D(\boldsymbol{x})]}_{\text{Original critic loss}} + \underbrace{\lambda \mathop{\mathbb{E}}_{\hat{\boldsymbol{x}} \sim \mathbb{P}_{\hat{\boldsymbol{x}}}} \left[(\|\nabla_{\hat{\boldsymbol{x}}} D(\hat{\boldsymbol{x}})\|_2 - 1)^2 \right]}_{\text{Our gradient penalty}}.$$

3.3.6 loss function

$$L = \lambda_{style} * L_{style} + \lambda_{edge} * L_{edge} + \lambda_{norm} * L_{norm} + \lambda_{GAN} * L_{GAN} + \lambda_{TV} * L_{TV}$$

4. Experiment

4.1 Experiment environment

Models are implemented on Ubuntu : 18.04.2 LTS Python: 3.73, pytorch:1.3.1

Run in hardware CPU: Intel(R) Core(TM) i7-3960X CPU @ 3.30GHz

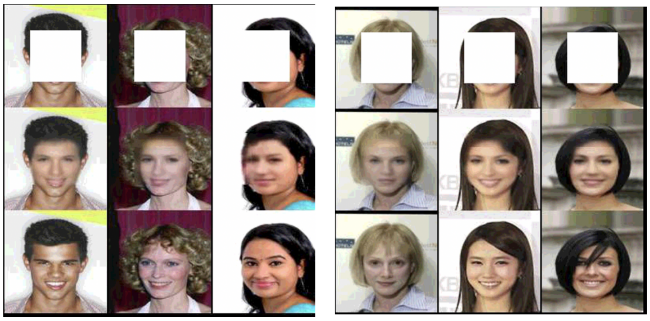

Fig. 7 The result of context encoder 1

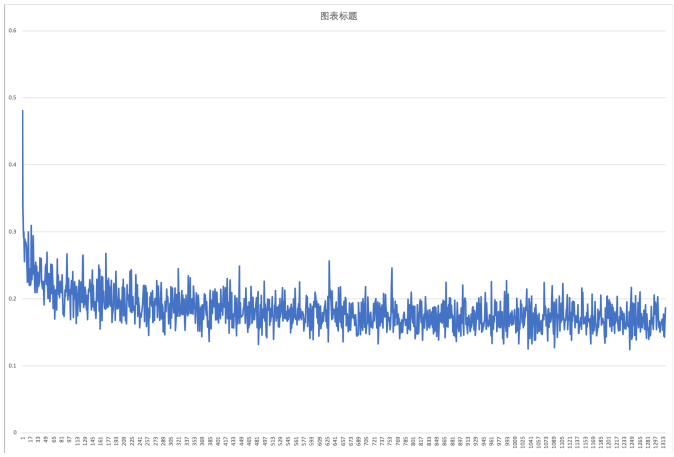

Fig. 8 The result of training

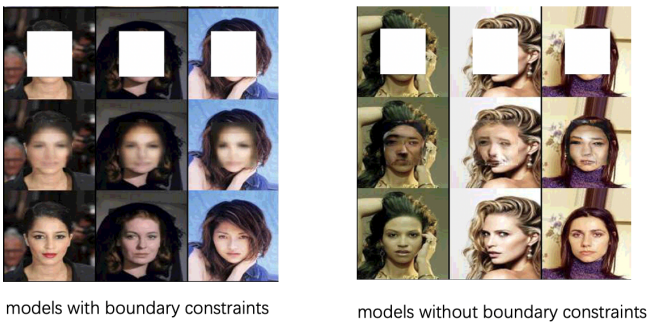

Fig. 9 A comparison of two models

4.2 Experiment result

dataset: CelebA

Fig.7 is the results after 750000 times training. The first line shows 128*128 pixel images with 64*64 pixels taken out as the unknown area in the middle. The second line in the image shows the recovered images and the third line shows the original images.

In Fig.8, the X-axis represents the number of iterations, and the Y-axis represents the MSE of the image.

Fig.9 shows the difference between models with boundary constraints and models without boundary constraints. Two models all trained for 1 epoch. It can be seen that models via edge learning is more stable.

5. Summary and future work

In this paper, we propose a two-stage adversarial framework with a special shift-connection and dilated convolutions, for filling the missing region with sharp edges and semantically plausible textures.

Due to the time limitation, we did not make more comparisons with the original model. The follow-up work is to process the data and compare more models. Besides, It is obviously that the human face image restoration tend to generate similar facial features, we can consider adding some noises to the bottleneck, in order for making the generated images more vivid.

Acknowledgment This work was supported by JST, ACT-I Grant Number JP-50166, Japan.

References

[1] Guillemot, C., Le Meur, O. (2013). Image inpainting: Overview and recent advances. IEEE signal processing magazine, 31(1), 127-144.

[2] Barnes, C., Shechtman, E., Finkelstein, A., Goldman, D. B. (2009, July). PatchMatch: A randomized correspondence algorithm for structural image editing. In ACM Transactions on Graphics (ToG) (Vol. 28, No. 3, p. 24). ACM.

[3] Wang, N., Li, J., Zhang, L., Du, B. (2019, August). MUSICAL: multi-scale image contextual attention learning for inpainting. In Proceedings of the 28th International Joint Conference on Artificial Intelligence (pp. 3748-3754). AAAI Press.

[4] Pathak, D., Krahenbuhl, P., Donahue, J., Darrell, T., Efros, A. A. (2016). Context encoders: Feature learning by inpainting. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 2536-2544).

[5] Shin, H. C., Roth, H. R., Gao, M., Lu, L., Xu, Z., Nogues, I., ... Summers, R. M. (2016). Deep convolutional neural networks for computer-aided detection: CNN architectures, dataset characteristics and transfer learning. IEEE transactions on medical imaging, 35(5), 1285-1298.

[6] Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., ... Bengio, Y. (2014). Generative adversarial nets. In Advances in neural information processing systems (pp. 2672-2680).

[7] Yang, C., Lu, X., Lin, Z., Shechtman, E., Wang, O., Li, H. (2017). High-resolution image inpainting using multi-scale neural patch synthesis. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (pp. 6721-6729).

[8] Zhang, H., Hu, Z., Luo, C., Zuo, W.,Wang, M. (2018, October). Semantic image inpainting with progressive generative networks. In Proceedings of the 26th ACM international conference on Multimedia (pp. 1939-1947).

[9] Radford, A., Metz, L., Chintala, S. (2015). Unsupervised representation learning with deep convolutional generative adversarial networks. arXiv preprint arXiv:1511.06434.

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

[11] Heusel, M., Ramsauer, H., Unterthiner, T., Nessler, B., Hochreiter, S. (2017). 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).

[12] Berthelot, D., Schumm, T., Metz, L. (2017). Began: Boundary equilibrium generative adversarial networks. arXiv preprint arXiv:1703.10717.

[13] Mao, X., Li, Q., Xie, H., Lau, R. Y., Wang, Z., Paul Smolley, S. (2017). Least squares generative adversarial networks. In Proceedings of the IEEE International Conference on Computer Vision (pp. 2794-2802).

[14] Isola, P., Zhu, J. Y., Zhou, T., Efros, A. A. (2017). Image-to-image translation with conditional adversarial networks. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 1125-1134).

[15] Zhu, J. Y., Park, T., Isola, P., Efros, A. A. (2017). Unpaired image-to-image translation using cycle-consistent adversarial networks. In Proceedings of the IEEE international conference on computer vision (pp. 2223-2232).

[16] Satoshi Iizuka, Edgar Simo-Serra, and Hiroshi Ishikawa. Globally and locally consistent image completion. ACM Transactions on Graphics, 36(4), 2017.

[17] Yu, F., Koltun, V. (2015). Multi-scale context aggregation by dilated convolutions. arXiv preprint arXiv:1511.07122.

[18] Liu, G., Reda, F. A., Shih, K. J., Wang, T. C., Tao, A., Catanzaro, B. (2018). Image inpainting for irregular holes using partial convolutions. In Proceedings of the European Conference on Computer Vision (ECCV) (pp. 85-100).

[19] Liu, H., Jiang, B., Xiao, Y., Yang, C. (2019). Coherent semantic attention for image inpainting. In Proceedings of the IEEE International Conference on Computer Vision (pp. 4170-4179).

[20] Bahdanau, D., Cho, K., Bengio, Y. (2014). Neural machine translation by jointly learning to align and translate. arXiv preprint arXiv:1409.0473.

[21] Yu, J., Lin, Z., Yang, J., Shen, X., Lu, X., Huang, T. S. (2018). Generative image inpainting with contextual attention. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 5505-5514).

[22] Yan, Z., Li, X., Li, M., Zuo, W., Shan, S. (2018). Shift-net: Image inpainting via deep feature rearrangement. In Proceedings of the European conference on computer vision (ECCV) (pp. 1-17).

[23] Song, Y., Yang, C., Lin, Z., Liu, X., Huang, Q., Li, H., Jay Kuo, C. C. (2018). Contextual-based image inpainting: Infer, match, and translate. In Proceedings of the European Conference on Computer Vision (ECCV) (pp. 3-19).

[24] Chen, L. C., Yang, Y., Wang, J., Xu, W., Yuille, A. L. (2016). Attention to scale: Scale-aware semantic image segmentation. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 3640-3649).

[25] Liu, H., Jiang, B., Huang, W., Yang, C. (2019). One-Stage Inpainting with Bilateral Attention and Pyramid Filling Block. arXiv preprint arXiv:1912.08642.