Image Extrapolation Based on Perceptual Loss and Style Loss

Yongpeng Ren[1], Xian Zhang[1], Hongping Ren[1], Lutao Wang[1], Guanrao Huang[2], Taisong Xiong[1], and Xiaojie Li[1(✉)]

[1] College of Computer Science,
Chengdu University of Information Technology, Chengdu 610103, China
lixj@cuit.edu.cn
[2] Chengdu Shengdaren Technology Co. Ltd., Chengdu 610000, China

Abstract. In recent years, deep learning-based image extrapolation has achieved remarkable improvements. Image extrapolation utilizes the structural and semantic information from the known area of an image to extrapolate the unknown area. In addition, these extrapolative parts not only maintain the consistency of spatial information and structural information with the known area, but also achieve a clear, beautiful, natural and harmonious visual effect. In view of the shortcomings of traditional image extrapolation methods, this paper proposes an image extrapolation method which is based on perceptual loss and style loss. In the paper, we use the perceptual loss and style loss to restrain the generation of the texture and style of images, which improves the distorted and fuzzy structure generated by traditional methods. The perceptual loss and style loss capture the semantic information and the overall style of the known area respectively, which is helpful for the network to grasp the texture and style of images. The experiments on the Places2 and Paris StreetView dataset show that our approach could produce better results.

Keywords: Image extrapolation · Perceptual loss · Style loss

1 Introduction

In computer vision tasks, deep learning-based image completion methods are widely used. Especially in recent years, image completion has made significant achieve-ments. Image completion is a special task, which actually falls between image editing and image generating. Traditional image completion methods can be divided into texture diffusion-based [3, 4, 16], distribution-based [13, 17] and generator-based models [8, 9, 14, 15]. Texture diffusion-based methods simply collect the similar pixels from the known area to fill the missing area. Due to directly searching similar pixels from a known area to complete an unknown area, the completed results usually have unnatural images and fuzzy boundaries. Based on the idea of data driving, distribution-based methods learn the relevant distribution information from large data to generate plausible structures.

X. Wu et al. (Eds.): QShine 2020, LNICST 381, pp. 179–191, 2021.
https://doi.org/10.1007/978-3-030-77569-8_13

However, these methods produce rough and fuzzy results. Generator-based methods generally employ neural network to extract high-level spatial features of the image and finally generate plausible structures for the missing regions. Since the semantic information of the image can be captured, these methods usually generate coherent, clear and authentic contents.

Image extrapolation, a specific application of image completion, utilizes the fragments of images to infer extensional parts, and finally generating the whole picture. It can be mainly used on texture synthesis, panorama synthesis and video expansion. However, it remains a challenging problem than image inpainting due to two main issues. First, fewer image proximity information can be used to infer the unknown regions. Second, the extrapolative results must have the realistic visual effect and natural structures. A mainstream deep learning-based technology for image extrapolation is the generative adversarial network (GAN) [2]. It as generative model performs remarkably on the unsupervised learning of complex distribution. The generator and discriminator networks are trained jointly with opposite goals: the former minimizes the objective function and the latter maximizes the objective function by adversarial training. This competitiveness helps them to mimic any distribution of data. In this way, generator can capture the data distribution once trained successfully.

How to make the network generate visually-realistic and semantically-plausible contents? In this paper, a new GAN-based method is proposed. In the proposed method, we use the perceptual loss [12] to extract features from both original images and generative images, as a result the network could obtain lower-level details and high-level abstract information. This finally helps network to produce clear and nature contents. Moreover, we utilize the style loss [11] to calculate Gram matrix of the extracted features for analysing the correlation of pairwise features. This eventually catches the overall style of images. Furthermore, the general reconstruction loss has some limitations, since it directly measures differences of the pixels from both original images and generative images. The general reconstruction results indeed have the higher signal-to-noise ratio, but it contains less high-frequency information. This would lead to blurry and distorted images. While perceptual loss acquires semantic information of images by extracting feature, which means that network is in the perception of the images. Owing to using low-level pixel information and high-level abstract features of images, which can restrain both texture and style of data, our proposed method could generate real and reasonable contents for the missing regions.

Our main contributions are as follows:

1. Perceptual loss is used to extract the details and abstract features of the images, ensuring the consistency of the semantic information between the known area and unknown area. Eventually, visually eliminating the blurry of boundary and generating natural and real results.
2. Style loss is used to obtain the overall style of the images through feature extractor and Gram matrix, ensuring the consistency of texture and style between the known area and unknown area. Finally, promoting the generation of real and natural style.

2 Related Work

Patch-based and diffusion-based methods are main non-deep-learning image completion approaches, which aim at attaining non-learning statistics information to complete images. They borrow similar pixels from undamaged images to fill missing parts. It usually generates implausible texture as it fails to understand high-level semantic information of images.

Context encoders (CE) [1] is an early deep learning-based method for image completion. It uses an encoder-decoder network architecture. Specially, the encoder maps the masked image to a low-dimensional feature space, then the decoder utilizes the features to reconstruct a complete image. Moreover, the encoder and decoder are connected by channel-wise fully-connected layer. Based on unsupervised learning method, CE utilizes the known feature information to complete the missing areas. However, it usually produces blurry textures and distorted structures, since the limitation of channel-wise fully-connected layer. Thus, CE needs to be improved in the network architecture.

Then, paper [5] could produce fine complete results by using multi-scale neural patch synthesis. Its network architecture consists of two branches. The first branch is responsible for generating the contents of missing regions, which utilizes the information from known areas. The second is used to produce textures for missing regions, which considers the differences of textures from both original images and generative images. Moreover, it employs the pre-trained VGG network [6] to extract texture features, then generates fine textures by restraining the gap of the feature maps. Due to fully considering the differences of textures, the method made a great improvement. However, owing to a large memory is required and only learning patch information from an image rather than a dataset, the method has some limitations.

Recently, Mark Sabini et al. [7] proposed the image outpainting with GANs (IOGnet). At present, the state-of-the-art approaches mainly apply GAN and CNN. Therefore, IOGnet employed CNN-based GAN to extrapolate the both sides of images, namely completing the regions beyond the border of images. In addition, it could produce a panorama through recursive extrapolation. In order to improve stability of the DCGAN, it changed the traditional training procedure. The training procedure is divided into three stages. The first stage is to train the generator by using mean square error. The second stage is to train the discriminator. The last stage is to train the generator and the discriminator at the same time. However, its extrapolative results are still much vague. Thus, the method also needs further improvements.

More recently, the pluralistic image completion (PICnet) proposed by Chuanxia Zheng et al. [10] achieved good complete results. Most image completion methods generate only one output result for a missing image. In this paper, they creatively proposed a pluralistic image completion method. Namely, they generated diverse and reasonable complete results for one input. However, the pluralistic image completion faced a great challenge. Sampling from CVAE would lead to the minimal diversity, due to a ground truth only provides one instance label in the training dataset. In order to solve this problem, PICnet proposed a probabilistic theory-based framework to maintain the diversity of sampling. Moreover, it used two parallel paths to train the model. One is the reconstructive path, which uses the instance label to obtain the prior distribution of missing regions and finally it reconstructs the image by the prior distribution. Another is the generative path, whose distribution is close to the prior distribution of the reconstructive path. The network achieves the trade-off between the reconstruction of original data and the variance of conditional distribution. Since utilizing the prior information of missing parts to guide the process of image completion, the approach could produce excellently clear and realistic results. In addition, the diversity of output results provides a sufficient condition to select high-quality complete images.

3 Image Extrapolation Based on Perceptual Loss and Style Loss

The encoder-decoder network architecture is used in our method (see Fig. 1). Specially, the encoder extracts high-level abstract features of the image, and the decoder utilizes the abstract features to up-sample a reconstructive image. Moreover, both encoder and decoder are connected by residual block to realize the inference network function. This could generate the mean and variance of VAE for the decoder's sampling in the latent space. To improve the stability of training procedure, we apply LSGAN that uses the least squares loss. In addition, the structure of discriminator is similar to that of the encoder, and we use the global discriminator to score the whole image. This could grasp the overall quality of the image.

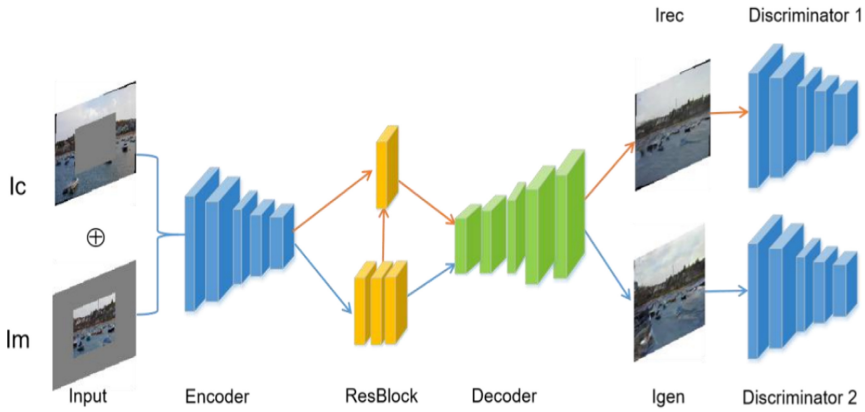


Fig. 1. Overview of our architecture.

In Fig. 1, Ic is the complement of the masked image, and Im is the masked image. ⊕ means to concatenate Ic and Im. The convolution kernels of Encoder, ResBlock, Decoder and Discriminator are all 3×3, and the stride is 1×1. During training, Ic and Im are concatenated and input into the Encoder. The mean and variance of VAE is generated by the Resblock, then utilizing the mean and variance to up-sample images. Thus, the prior information of reconstructive path is used to guide the process of image generation. Furthermore, the decoder generates the reconstructive image Irec and the generative image Igen respectively, finally sending the Irec and Igen to their own discriminators.

3.1 Perceptual Loss

The perceptual loss extracts the features from both generative images and original images through the pre-trained VGG network, and restrains the features of these images by the L1 norm. Since forcing the generative results to perceptually resemble these labels from pre-trained network, the perceptual loss could improve the quality of extrapolative areas. Formally,

$$L_p = E\left[\left\|\phi_j(I_{gt}) - \phi_j(I_{gen})\right\|_1\right] \tag{1}$$

where I_{gt} and I_{gen} are the original image and the generative image, respectively. $\phi_j(.)$ is the j-th layer feature map extracted by VGG network. Perceptual loss is to compare the features from the convolution of original images and generative images in the VGG network. This aims to make the extracted high-level feature information (such as the contents and structures of images) as close as possible. It also means that the network is perceiving the image. In the training of GAN network, perceptual loss can make the feature map of the generative image close to that of the original image, finally assists the image generation and improves the quality of generative images.

3.2 Style Loss

Style loss is similar to perceptual loss. We also use the pre-trained VGG network to extract features from both generative images and original images. However, the Gram matrix of extracted features is further calculated. As restraining the Gram matrix of features, the overall style of both generative images and original images could be as close as possible. Finally, the quality of the generative images also can be improved. Formally,

$$L_s = E\left[\left\| G_{\phi_i}(I_{gt}) - G_{\phi_i}(I_{gen}) \right\|_1\right] \tag{2}$$

where I_{gt} and I_{gen} are the original image and the generative image respectively, and $\phi_i(.)$ is the i-th layer feature map extracted by VGG network. $G(.)$ denotes the Gram matrix corresponding to the feature maps. Gram matrix calculates the inner product of any k vectors in the n-dimensional Euclidean space. It can be regarded as covariance matrix without subtracting mean between different feature maps. In the convolution network, the shallow layer network extracts low-level features of images, while the deep layer network extracts high-level abstract features. These low-level and high-level features are more like the overall style of an image, which determines the real attribute of an image. By calculating the Gram matrix of these feature maps, the correlation between pairwise eigenvectors can be estimated. In the training procedure, owing to the style of generative images can be gradually close to that of original images, the quality of generative images could be improved.

3.3 Other Loss

In addition, we use the loss of PICnet:

$$L = \alpha_{KL}(L^r_{KL} + L^g_{KL}) + \alpha_{app}(L^r_{app} + L^g_{app}) + \alpha_{ad}(L^r_{ad} + L^g_{ad}) \tag{3}$$

where the superscripts r and g denote the loss of reconstructive path and generative path respectively. L_{KL} is used to constrain the distribution of hidden layer and L_{app} is the reconstruction loss of images. L_{ad} is the adversarial loss.

Finally, we add the two losses to Eq. (3):

$$L_{total} = L + \lambda_1 L_p + \lambda_2 L_s \tag{4}$$

In the experiment, we set $\lambda_1 = 0.1$, $\lambda_2 = 250$.

3.4 Improvement of IOGnet

Furthermore, we also apply perceptual loss and style loss to the improvement of the IOGnet method. The loss of the original IOGnet is as follows:

$$L_{MSE} = ||M \odot (G(I_p) - I_n)||_2^2 \tag{5}$$

$$L_D = -[\log D(I_n) + \log(1 - D(G(I_p)))] \tag{6}$$

$$L_G = L_{MSE} - \gamma \log D(G(I_p)) \tag{7}$$

where M is the mask, I_n is ground truth, and I_p is concatenation of the masked I_n and the mask. D and G are the discriminator and generator, respectively. The training procedure is divided into three stages. The first stage is through Eq. (5) to train the generator. In the second stage, the discriminator is trained through Eq. (6). In the last stage, the generator and discriminator are trained at the same time through Eq. (7).

We modify the losses of Eq. (5) and Eq. (7) in IOGnet as follows:

$$L_r = L_{MSE} + \alpha L_p + \beta L_s \tag{8}$$

$$L_G = L_r - \gamma \log D(G(I_p)) \tag{9}$$

Namely, the perceptual loss and style loss are added in the first and third stages of training, and we set $\alpha = 10$, $\beta = 100$, and $\gamma = 0.0004$ in the experiment.

4 Experiment Results

The experimental metrics can be divided into the qualitative and quantitative comparison. The qualitative comparison is visually estimated for the quality of generative results. In addition, we measure the quantitative comparison by employing the following metrics: 1) Inception Score (IS) and Frechet Inception Distance (FID) are commonly used to evaluate the quality of the generative model, which can be used to measure the diversity and clarity of generative images; 2) peak signal-to-noise ratio (PSNR) is also a widely-used full-reference metric for objective estimation; 3) other metrics include ℓ_1 loss, root mean square error (RMSE) and structural similarity (SSIM). These metrics are based on pixel-wise independence.

$$IS = \exp(E_{x \sim pg} D_{KL}(p(y|x) \| p(y))) \tag{10}$$

where x is the generative image, g is the generator, and y is the label predicted by pre-trained Inception-V3 model.

$$FID = \left\| \mu_x - \mu_g \right\|_2^2 + Tr(\sum\nolimits_x + \sum\nolimits_g - 2(\sum\nolimits_x \sum\nolimits_g)^{1/2}) \tag{11}$$

where the superscripts x and g denote the ground truth and generative image respectively. μ is the mean of eigenvectors, and Σ is the covariance matrix of eigenvectors

$$PSNR = 10 \cdot \log_{10} \frac{MAX_I^2}{MSE} \tag{12}$$

where MAX_I^2 is the maximum pixel-value of the image, and MSE is mean square error.

The experiment is implemented in Python 3.6.9, PyTorch 1.2.0, and Ubuntu 16.04. The GPU is NVIDIA Geforce RTX 2080 Ti. In addition, the batch-size is 64, and the Adam optimizer is used to update the network parameters. The fixed learning rate is 10^{-4}. Furthermore, we train the network in the end-to-end style. The LSGAN is applied to make the training procedure more stable, and updating the discriminator once then updating the generator once. Our training procedure costs 100 epochs in total.

We evaluate the proposed model on the image dataset Paris StreetView [18] and Place2 [19]. All the images are resized to 128 $\times$ 128. The test input is masked images with the 64 $\times$ 64 center area. The experimental comparison method is the baseline of PICnet and IOGnet. Because of the multiple output results of PICnet, we choose the image with the highest score of discriminator for comparison.

4.1 Qualitative Comparison

Firstly, we evaluate our model on the Paris StreetView dataset. Figure 2 shows the extrapolative results generated by PIC and our method, which could be used to visually estimate these results. In Fig. 2(c), the PIC generates the results with distorted structures and even residual shadows in the extrapolative regions. Moreover, the results of PIC also exhibit slight blurriness and unnaturalness. In Fig. 2(d) and (e), we add style loss alone or add style loss and perceptual loss at the same time into PIC. As a result, the generative results have some improvements. Our model could basically eliminate the parts with residual shadows in PIC, and force the generative images to be closer to the style of ground truth.

Then, we also show the experimental results on the Places2 in Fig. 3. We can find the similar influence on generative images after adding the style loss and perceptual loss. By and large, the visual effects of Fig. 3(d) and (e) are better than Fig. 3(c). Compared to the original PIC, our model could produce more natural and more realistic images. As a result, the experimental results show that our method can improve the quality of image extrapolation.

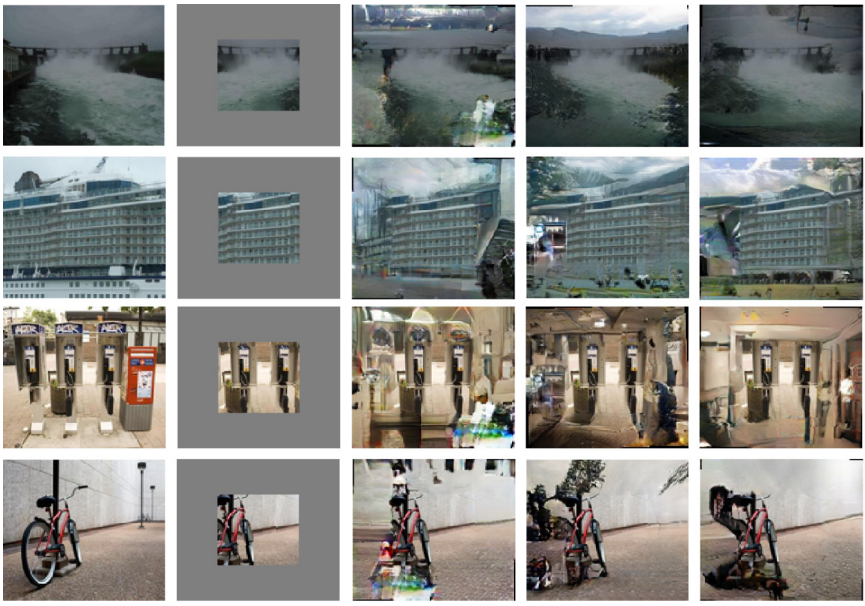
In order to further evaluate the effectiveness of our method, we also implement the experiments compared with the IOG method. Figure 4 and Fig. 5 show the visual effects of IOG method and our method on the Paris StreetView and Places2 dataset. Similarly, the generative results can be improved after adding the style loss and perceptual loss. Compared with the original IOG in Fig. 4(c) and Fig. 5(c), our model could smooth the coarseness of extrapolative parts, meanwhile enhance the clarity and aesthetics of generative images. On the whole, the images generated by our model are more visually-realistic and more plausible than the original IOG.

Fig. 2. Qualitative results of different methods on the Paris StreetView dataset. (a) ground truth, (b) input, (c) PIC, (d) PIC+style loss, and (e) PIC+style loss+perceptual loss.

In conclusion, we compared with the original PIC and IOG on the Paris StreetView and Place2 for qualitative comparison. The experiments prove that the style loss and perceptual loss can contribute to the improvement of image extrapolation, especially for these exhibit fuzzyness and unnaturalness.

4.2 Quantitative Comparison

Table 1 and 2 show the quantitative results of different methods on Places2 and Paris StreetView. These results of quantitative metrics show that after adding style loss and perceptual loss the generative images could be improved in some degrees. Table 1 is the quantitative metrics of 20000 test images on the Place2. The IS metric increases by 0.09, and FID decreases by 2.94. Thus, it shows that our methods could effectively improve the diversity and clarity of the generative images. Meanwhile, RMSE and ℓ_1 loss metrics decrease by 1.45 and 0.65 respectively, indicating that the overall pixel differences between the generative images and the original images are smaller. However, owing to the limitation of the 100 test images of Paris StreetView, we only evaluate the SSIM and RMSE. Furthermore, the SSIM and RMSE on the Paris StreetView are also better.



(a) (b) (c) (d) (e)

Fig. 3. Qualitative results of different methods on the Places2 dataset. (a) ground truth, (b) input, (c) PIC, (d) PIC+style loss, and (e) PIC+style loss+perceptual loss.

(a) (b) (c) (d) (e)

Fig. 4. Qualitative results of different methods on the Places2 dataset. (a) ground truth, (b) input, (c) IOG, (d) IOG+style loss, and (e) IOG+style loss+perceptual loss.

(a) (b) (c) (d) (e)

Fig. 5. Qualitative results of different methods on the Paris StreetView dataset. (a) ground truth, (b) input, (c) IOG, (d) IOG+style loss, and (e) IOG+style loss+perceptual loss.

Table 1. Quantitative results of different methods on the Places2 dataset.

Method	IS	FID	PSNR	ℓ_1 loss	SSIM	RMSE
PIC	5.60	34.81	12.95	38.29	0.4116	70.21
PIC+style loss	5.49	**31.87**	12.92	38.21	0.4109	70.54
PIC+style loss+perceptual loss	**5.69**	32.39	**13.11**	**37.64**	**0.4140**	**68.76**

Table 2. Quantitative results of different methods on Paris StreetView. Because the limitation of the 100 test images of Paris StreetView, we only evaluate the SSIM and RMSE.

Method	SSIM	RMSE
PIC	0.4248	55.63
PIC+style loss	**0.4441**	**55.58**
PIC+style loss+perceptual loss	0.4367	55.89

In addition, we have also implemented the comparative experiments on the IOG method. Table 3 and 4 show the quantitative results of different methods on Place2 and

Table 3. Quantitative results of different methods on Places2.

Method	IS	FID	PSNR	ℓ_1 loss	SSIM	RMSE
IOG	5.32	68.04	15.32	32.43	0.4455	52.56
IOG+style loss	5.72	64.71	15.43	31.82	0.4488	51.90
IOG+style loss+perceptual loss	**5.89**	**62.34**	**15.61**	**31.11**	**0.4660**	**51.14**

Table 4. Quantitative results of different methods on Paris StreetView. Because the limitation of the 100 test images of Paris StreetView, we only evaluate the SSIM and RMSE.

Method	SSIM	RMSE
IOG	0.4670	43.11
IOG+style loss	0.4730	43.37
IOG+style loss+perceptual loss	**0.4731**	**42.60**

Paris Streetview dataset. The experimental results on Place2 dataset show that our method increases by 0.57 and 0.29 in IS and PSNR respectively, and the FID and ℓ_1 loss decrease by 5.7 and 1.32. Thus, it proves that the style loss and perceptual loss can improve the quality of the generative images, and it can also improve the clarity and naturalness of the extrapolative results. Moreover, SSIM increases by 0.02 and RMSE decreases 1.42, indicating that our method could improve the quality of generative images. In addition, SSIM and RMSE metrics on the Paris StreetView dataset are also improved.

5 Conclusion

We proposed a new image extrapolation method. It combined the perceptual loss with style loss. After training and testing on the commonly-used image extrapolation dataset, experimental results show that our model can produce fine textures and natural contents for the missing images. In addition, no matter qualitative or quantitative comparison, our results could exhibit better than these methods. In the future, we will continue to explore the relevant aspects of image extrapolation, and further improve the quality of generative images in the field.

Acknowledgment. This work was supported by the National Key Research and Development Program of China (No. 2017YFC1502203), and the Sichuan Science and Technology program (2019JDJQ0002, 18MZGC0060, 2018RZ0072, and 2018GZ0184), the major Project of Education Department in Sichuan (17ZA0063).

References

1. Deepak, P., Philipp, K., Jeff, D., Trevor, D., Alexei, A.E.: Context encoders: feature learning by inpainting. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 2536–2544 (2016)

2. Goodfellow, I., Pouget-Abadie, J., Mirza, M.: Generative adversarial nets. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 654–656 (2014)
3. Coloma, B., Marcelo, B., Vient, C., Guillermo, S., Joan, V.: Filling-in by joint interpolation of vector fields and gray levels. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 1200–1211 (2001)
4. Sun, J., Yuan, L., Jia, J., Shum, H.-Y.: Image completion with structure propagation. ACM Trans. Graph. **24**, 861–868 (2005)
5. Chao, Y., Xin, L., Zhe, L.: High-Resolution image inpainting using multi-scale neural patch synthesis. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 1457–1460 (2017)
6. Simonyan, K., Zisserman, A.: Very deep convolutional networks for large-scale image recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 1409–1412 (2014)
7. Mark, S., Gili, R.: Painting outside the box: image outpainting with GANs. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 421–420 (2018)
8. Iizuka, S., Simo-Serra, E., Ishikawa, H.: Globally and locally consistent image completion. ACM Trans. Graph. **36**, 1–14 (2017)
9. Wang, C., Liu, M.Y., Zhu, J.Y., Tao, A., Kautz, J.: High-resolution image synthesis and semantic manipulation with conditional GANs. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 5–10 (2018)
10. Chuanxia, Z., Tat-Jen, C., Jianfei, C.: Pluralistic image completion. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 1–5 (2019)
11. Gatys, L.A., Ecker, A.S., Bethge, M.: Image style transfer using convolutional neural networks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 2414–2423 (2016)
12. Johnson, J., Alahi, A., Fei-Fei, L.: Perceptual losses for real-time style transfer and super-resolution. In: Leibe, B., Matas, J., Sebe, N., Welling, M. (eds.) Computer Vision – ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part II, pp. 694–711. Springer, Cham (2016). https://doi.org/10.1007/978-3-319-46475-6_43
13. 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)
14. Kamyar, N., Eric, B., Tony, J., Faisal, Z., Qureshi, M.: EdgeConnect: generative image inpainting with adversarial edge learning. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 5–7 (2019)
15. Dolhansky, B., Ferrer, C.C.: Eye in-painting with exemplar generative adversarial networks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 7902–7911 (2018)
16. Ballester, C., Bertalmio, M., Caselles, V., Sapiro, G., Verdera, J.: Filling-in by joint interpolation of vector fields and gray levels. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 1200–1211 (2001)
17. Hays, J., Efros, A.: Scene completion using millions of photographs. ACM Trans. Graph. **26**(3), 4 (2007)
18. Doersch, C., Singh, S., Gupta, A., Sivic, J., Efros, A.: What makes Paris look like Paris? ACM Trans. Graph. **31**(4), 1–9 (2012). https://doi.org/10.1145/2185520.2185597
19. Zhou, B., Lapedriza, A., Khosla, A., Oliva, A., Torralba, A.: Places: A 10 million image database for scene recognition. IEEE Trans. Pattern Anal. Mach. Intell. **40**(6), 1452–1464 (2018)