Computers, Materials & Continua
DOI:10.32604/cmc.2022.023071
Article

Tech Science Press

Perceptual Image Outpainting Assisted by Low-Level Feature Fusion and Multi-Patch Discriminator

Xiaojie Li[1], Yongpeng Ren[1], Hongping Ren[1], Canghong Shi[2], Xian Zhang[1], Lutao Wang[1], Imran Mumtaz[3] and Xi Wu[1,*]

[1]College of Computer Science, Chengdu University of Information Technology, Chengdu, 610225, China
[2]Xihua University, Chengdu, 610039, China
[3]University of Agriculture Faisalabad, Pakistan
*Corresponding Author: Xi Wu. Email: xi.wu@cuit.edu.cn
Received: 27 August 2021; Accepted: 11 November 2021

Abstract: Recently, deep learning-based image outpainting has made greatly notable improvements in computer vision field. However, due to the lack of fully extracting image information, the existing methods often generate unnatural and blurry outpainting results in most cases. To solve this issue, we propose a perceptual image outpainting method, which effectively takes the advantage of low-level feature fusion and multi-patch discriminator. Specifically, we first fuse the texture information in the low-level feature map of encoder, and simultaneously incorporate these aggregated features reusability with semantic (or structural) information of deep feature map such that we could utilize more sophisticated texture information to generate more authentic outpainting images. Then we also introduce a multi-patch discriminator to enhance the generated texture, which effectively judges the generated image from the different level features and concurrently impels our network to produce more natural and clearer outpainting results. Moreover, we further introduce perceptual loss and style loss to effectively improve the texture and style of outpainting images. Compared with the existing methods, our method could produce finer outpainting results. Experimental results on Places2 and Paris StreetView datasets illustrated the effectiveness of our method for image outpainting.

Keywords: Deep learning; image outpainting; low-level feature fusion; multi-patch discriminator

1 Introduction

Nowadays, artificial intelligence (AI) has ushered in a new big data era. The improvement of AI also promotes deep learning technology to be widely employed in many fields [1–4], especially in image processing field. Combining deep learning technology with image processing method, AI system could acquire more available environmental information to make correct decisions. For example, applying

deep learning-based image processing to the pattern recognition and automatic control field for the efficient analysis and real-time response, which is considered as a promising prospect.

Recently, deep learning-based methods [5–8] have been widely applied to image inpainting task and have made remarkable achievements. Image inpainting, as a common image editing task, aims to restore damaged images and remove objects. Existing image inpainting methods can mainly be divided into two groups: non-learning methods and learning-based methods. The former group is composed of diffusion-based [9,10] and distribution-based approaches [11,12]. Concretely, the diffusion-based approaches use the texture synthesis to fill the unknown parts, and search or collect the suitable pixels of known regions to diffuse into the unknown regions. These methods can generate meaningful textures for the missing regions. However, they generate inpainting results often with blurry and distorted contents when meet a big hole or sophisticated textures, because they fail to capture the semantic information of images. On the other hand, the distribution-based approaches utilize the whole dataset to obtain the data distribution information, and finally generate inpainting images. Similarly, due to only extracting low-level pixel information, they can't produce a fine texture. By contrary, learning methods [13–15] generally use convolutional neural networks to extract the semantic information of images such that they could realize a natural, realistic and plausible inpainting result.

Compared with image inpainting, image outpainting is studied relatively fewer in the image processing field. It uses the known parts of images to recursively extrapolate a complete picture. Moreover, image outpainting faces a greater challenge because of the less neighboring pixel information. Furthermore, the outpainting model must produce plausible contents and vivid textures for the missing regions. In practice, image outpainting can be applied in panorama synthesis, texture synthesis and so on. The generative adversarial network (GAN) [16,17] is commonly employed in image outpainting, and it is suitable for unsupervised learning on complicated distribution. GAN, as a generative model, aims to train jointly its generator and discriminator for an adversarial idea. Specifically, the generator minimizes the loss function, and the discriminator maximizes the loss function. Since the adversarial training promotes the generator to capture the real data distribution, the network can generate fine and reasonable images.

Existing image outpainting methods generally fail to effectively extract image information (such as structure and texture information), resulting in the unclarity and unnaturalness of outpainting results. To generate more semantically reasonable and visually natural outpainting results, we present a perceptual image outpainting method assisted by low-level feature fusion and multi-patch discriminator (LM). It is known that the low-level features map with higher resolution could acquire plentiful detail information (such as location information and texture information). However, it contains less semantic information. The high-level feature map could acquire more semantic information, and it perceives the less detail information. Therefore, we first fuse the texture information in the low-level feature map of encoder, and simultaneously incorporate these aggregated features reusability with semantic (structural) information of deep feature map by element-wise adding such that we could utilize more sophisticated texture information to generate more authentic outpainting images. Moreover, we introduce a multi-patch discriminator to enhance the generated texture information and comprehensively judge the reality of outpainting images. We design its outputs as a $n \times n$ tensor equal to judge the number of patches of an image, which could perceive the relatively bigger receptive field. Therefore, our multi-patch discriminator further effectively judges the generated image from the different level features and indirectly promotes the generator to grasp the real distribution of input data. This could impel our network to produce more natural and clearer outpainting images.

Furthermore, we employ perceptual loss [18] to extract the high-level feature information of both generated images and ground truths. Therefore, our network could restrain the texture generation of outpainting regions. Meanwhile, style loss [19] is employed to estimate the relevance of different features extracted by pre-trained Visual Geometry Group 19 (VGG19) network [20], and we further compute a Gram matrix to obtain the global style of outpainting images. In this way, our model can generate real and consistent outpainting results.

In general, our contributions are as follows:

(1) We effectively fuse and reuse the texture information of the low-level feature map of encoder and simultaneously incorporate these aggregated features reusability with semantic (structural) information of deep feature map in the decoder, which could utilize more sophisticated texture information to generate more authentic outpainting results with finer texture.
(2) We propose two multi-patch discriminators to comprehensively judge the generated images from the different level features, which further enlarges receptive field of discriminator network and finally improves the clarity and naturalness of outpainting results.

The rest part of paper is organized as follows: Section 2 presents related image outpainting works. The detail theory of our proposed method is illustrated in Section 3. Section 4 introduces our experimental results which include qualitative and quantitative comparisons with existing methods. In the last section, we present conclusions and future works.

2 Related Work

In the early time, image inpainting fills the missing areas through non-learning methods, including patch-based [21–23] and diffusion-based methods [24–26]. Caspi et al. [27] use bidirectional spatial similarity to maintain the information of input data, which can be applied in retargeting or image inpainting. Nonetheless, the spatial similarity estimation costs a large number of computation resources. Barnes et al. [28] propose a PatchMatch method, using a fast nearest neighbor estimation to match reasonable patches. Therefore, PatchMatch could save expensive computation cost. These methods all assume that the missing contents come from the known regions, thus they search and copy the patches of known areas to fill unknown areas. By this way, they can generally produce meaningful contents for the missing regions. However, they often exhibit badly for complicated structures or bigger holes, due to they only gain low-level image information such as the non-learning statistics information and simple pixel information of images.

Context Encoder (CE) [29] firstly applies the deep learning-based and GAN-based method to image inpainting task. It presents a new unsupervised learning method which is based on contextual pixel prediction. CE can be used to generate realistic contents according to known pixel information. Its overall network is an encoder-decoder architecture. The encoder maps the missing image into the latent space, and then the decoder utilizes these features of latent space to generate missing contents. A channel-wise fully-connected layer is introduced to connected encoder and decoder. In addition, both reconstruction loss and adversarial loss are used to train the CE model for realizing a sharp inpainting result. In this way, CE could simultaneously obtain both structure representation and semantic information of images. However, owing to the limitation of the fully-connected layer in the network, it fails to produce clear inpainting results.

Chao et al. [30] propose a multi-scale neural patch synthesis algorithm, which is composed of content network and texture network. It can generate fine content and texture through training jointly the two networks. The content network is used to fill contents for the missing areas, while

the texture network is used to further improve texture of output results generated by content network. Furthermore, in the texture network, a pre-trained VGG network is employed to force the patches in the inpainting regions to be perceptually similar to the patches in the known regions. Since they fully take the texture of missing regions into account, the network performs well for producing fine structures. However, due to the multi-scale learning which costs a lot of computation resources, this method has significant limitations.

Then, Iizuka et al. [31] present a novel image inpainting method which guarantees the inpainting images with both local and global consistency. More specifically, it uses a local discriminator and a global discriminator to realize fine inpainting results. The local discriminator judges the inpainting areas to achieve local detail consistency, while the global discriminator judges the whole image to ensure the consistent overall structure. Thanks to ensuring the consistency of local and global details, the model could produce much finer inpainting results. Moreover, it also achieves a more flexible inpainting without the limitation of image resolution and missing shape.

To get over the influence of subordinate pixels in the missing regions, Liu et al. [32] create a partial convolution for irregular image inpainting. In the method, they use the masked and renormalized convolution to force the network to focus on the valid pixels of input images. Moreover, they also present a method to automatically update the mask value for the next convolutional layer. By this way, the influence of subordinate information can be reduced in some degree, which promotes the network to process the input image more effectively. Ultimately, they realize natural and clear inpainting results.

Zheng et al. [33] propose a pluralistic image inpainting method (PICnet), which could produce multiple output for one input image. The most image inpainting methods only output one result, due to the limitation of one instance label provided by the ground truth. To let the model output diverse inpainting results, they invent a novel probabilistic theory to settle the problem. In addition, their network architecture contains two parallel paths, which are composed of the reconstructive path and the generated path. Concretely, the reconstructive path is used to obtain the distribution information of missing regions, and finally reconstructing a complete image. On the other hand, the generated path utilizes the distribution information of reconstructive path to guide the generation of missing images. By sampling from the variational auto-encoder (VAE) (another generative model), the network can produce pluralistic inpainting images. Owing to the considering of prior distribution of missing regions, they not only generate high-quality results but also create the diversity of images.

Mark et al. [34] recently apply GAN to the image outpainting for painting outside the box (IOGnet). They employ the deep learning-based GAN approach to outpaint the panorama contents for the sides of missing images, and finally recursively expand the parts beyond the border. Furthermore, they adopt a three-stage strategy to stabilize the training process. In the first stage, the generator is trained by the L2 distance between the generated images and the ground truths. In the second stage, the discriminator is trained alone according to the adversarial loss. In the last stage, the generator and discriminator are trained jointly through the adversarial loss. Finally, the model could even generate a five-time outpainting result than the original input. However, the obscure contents appear in the outpainting parts. As a result, the work needs to be improved in some aspects.

3 Perceptual Image Outpainting Assisted by Low-Level Feature Fusion and Multi-Patch Discriminator (LM)

To produce high-quality outpainting results, we present a simple perceptual image outpainting method assisted by low-level feature fusion and multi-patch discriminator. Moreover, we simultaneously employ both perceptual loss and style loss to improve the texture and style of outpainting images.

Network architecture will be introduced in Subsection 3.1, and the rest of subsections are used to introduce the principle of our method.

3.1 Network Architecture

As shown in Fig. 1, a simple GAN-based network, mainly consisting of the generator and discriminator, is used in our network. Firstly, our encoder in generator maps input images (both I_m and I_c) into a latent feature space. We first fuse the texture information in the low-level feature map of encoder, and simultaneously incorporate these aggregated features reusability with semantic (structural) information of deep feature map by element-wise adding in decoder. This could utilize more sophisticated texture information to generate more authentic outpainting images. Furthermore, the inference module (yellow block) connects encoder with decoder for utilizing the latent feature more effectively. In fact, the inference module is equal to the function of VAE [35], which computes the mean and variance of latent features to sample useful features. Finally, to generate more realistic results, we inject outpainting image into the pre-trained VGG [36] network for obtaining the feature information, which will be used to compute perceptual loss and style loss. In addition, we use the Least Squares Generative Adversarial Network (LSGAN) loss [37] to stabilize the training of our model. Then we present a multi-patch discriminator to enhance the generated texture information, which effectively judges the generated image from the different level features and impels our network to produce more natural and clearer outpainting images.

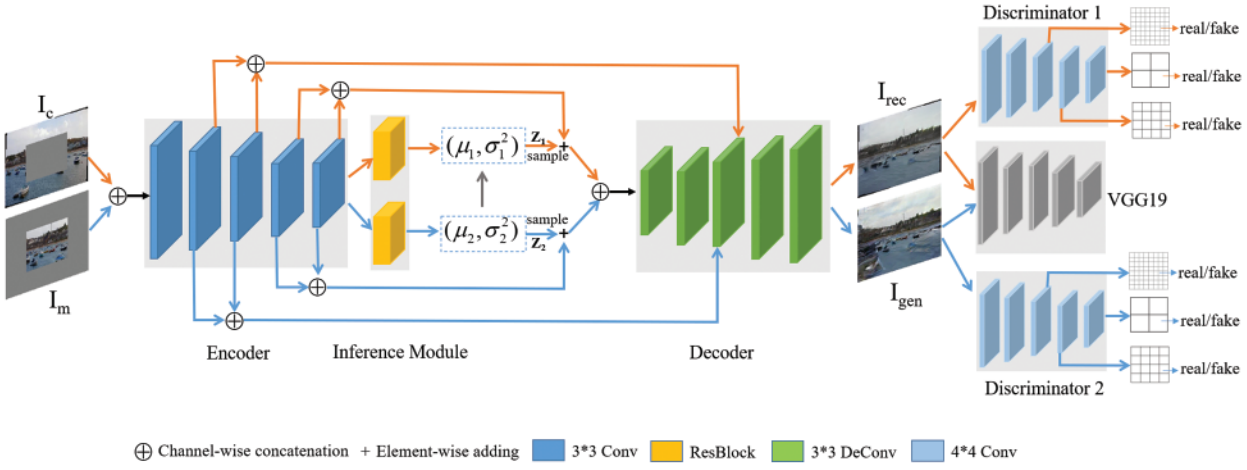

Figure 1: Overview of our network architecture

3.1.1 Generator

Fig. 1 shows that our network structure consists of two paths: yellow path in the top and blue path in the bottom. Note that the former path aims to reconstruct inpainting images and the latter path aims to generate outpainting results. In the training, both masked images I_m and I_c (complement of I_m) are concatenated by the channel-wise operation such that both can be simultaneously processed. Then we detach output features into both different inference modules (yellow block) to compute their latent features' mean and variance, which will be used to sample latent features. To simultaneously deal with both latent features, we concatenate both sampling features and feed them into the decoder. To easily grasp more sophisticated texture information and generate more authentic outpainting images, when decoder processes the latent features we fuse the texture information in the low-level feature

map of encoder, and simultaneously incorporate these aggregated features reusability with semantic (structural) information of deep feature map by element-wise adding in decoder. It is formally defined as:

$$F_i = CD(E_{i-1}(I)) \oplus E_i(I) \tag{1}$$

where F_i is i-th layer's aggregated features, I denotes input image, E_i is i-th layer in the encoder, $\oplus$ denotes channel-wise concatenation, and CD is down-sampling operation. Namely, we first down-sample (i-1)-th layer's features, and concatenate the down-sampling features with the i-th layer's features by channel-wise concatenation. Therefore, F_i contains (i-1)-th and i-th layers' aggregated feature information (see Eq. (1)). Then, we could pass aggregated features F_i into the decoder via element-wise adding. Therefore, the network could generate more sophisticated texture for the generated images. Finally, we produce both reconstructive image I_{rec} and generated image I_{gen}.

3.1.2 Discriminator

We design multi-patch discriminators (both Discriminator 1 and Discriminator 2) to enhance the generated texture information, which effectively judges the generated image I_{gen} and I_{rec} from the different level features and impels our network to produce more natural and clearer outpainting images. Formally, it is defined as:

$$L_{ad}^{g} = \sum_{i} [D_i(I_{gen}) - 1]^2 \tag{2}$$

where L_{ad}^{g} is the generator's adversarial loss, D_i is the i-th layer of discriminator, and I_{gen} is the generated image. Specifically, we judge the output patches in the last three layers of discriminator are real or fake. From the multi-patch information, the discriminator could effectively reinforce the ability of judgement for the output patch of discriminator (see Eq. (2)). Therefore, the discriminator could comprehensively judge an input image is real or fake. Finally, the real distribution of data is grasped by the generator, and the model could produce finer outpainting results.

3.2 Perceptual Loss and Style Loss

To further improve the texture and style of outpainting images and generate more realistic result, we simultaneously introduce both perceptual loss and style loss. Perceptual loss aims to extract semantic (structure) feature information via the pre-trained VGG19 network. By constraining the L_1 distance of these features, it can force outpainting results perceptually close to ground truths. Formally, the perceptual loss is defined as:

$$L_{perceptual} = E\left[\sum_{i} ||\Phi_i(I_{gen}) - \Phi_i(I_{gt})||_1\right] \tag{3}$$

where I_{gen} is the generated image and I_{gt} is the ground truth. $\Phi_i(\cdot)$ denotes the i-th layer features map of VGG. Actually, the perceptual loss is used to measure the difference of corresponding features extracted by VGG. The features in the convolutional neural network generally represent the semantic information of images such as the low-level textures or high-level attributes. Through penalizing these features dissimilar to the feature labels in the VGG, the outpainting parts can be improved in some degree. Thanks to the applying of perceptual loss in the training of GANs, the generator could be gradually tuned to produce a finer output result.

Style loss aims to extract the general style of generated images and ground truth. Concretely, to capture the overall style, we calculate the Gram matrix of their features extracted by VGG network. As a result of the L_1 norm constraint on the corresponding Gram matrices, the outpainting images will approach the realistic style by degrees. Analogously, the style loss is defined as follow:

$$L_{style} = E\left[\sum_i ||\mathrm{G}_{\Phi_i}(I_{gen}) - \mathrm{G}_{\Phi_i}(I_{gt})||_1\right] \tag{4}$$

where $\mathrm{G}_{\Phi_i}(\cdot)$ denotes the Gram matrix of i-th layer's feature extracted by VGG network. In fact, Gram matrix is the covariance matrix of eigenvectors in the Euclidean space, and it estimates the correlation of pair eigenvectors. Convolutional Neural Network (CNN) extracts the low-level texture information of images in the shallow layer, while in the deeper layer it obtains the high-level semantic information. The genuine attribute of an image is up to the combination of low-level and high-level information. Therefore, it can be used to measure the correlation of different features, including the important essence of images. Since we force the style of outpainting images to be similar to the style of ground truths, our model can produce the outpainting results with natural and authentic appearance.

3.3 Other Loss

Moreover, we apply the loss from PIC. Formally,

$$L_{pic} = \alpha_{KL}(L^r_{KL} + L^g_{KL}) + \alpha_{app}(L^r_{app} + L^g_{app}) + \alpha_{ad}(L^r_{ad} + L^g_{ad}) \tag{5}$$

where the subscript r denotes the reconstructive path (see yellow path in Fig. 1), and g denotes the generated path (see blue path in Fig. 1). L_{KL} is the KL loss for restraining the distribution of both reconstructive images and generated images. L_{app} is the reconstruction loss, and L_{ad} is the adversarial loss for GAN.

In our model, the total loss is defined as follow:

$$L = L_{pic} + \lambda_1 L_{perceptual} + \lambda_2 L_{style} \tag{6}$$

where $\lambda_1 = 0.1$, $\lambda_2 = 250.0$ in our experiments.

4 Experimental Results

4.1 Dataset

We evaluate our method on both Places2 [38] and Paris StreetView [39] datasets. Places2 dataset is a natural scene dataset which is widely used in image outpainting. We divided Places2 into training set 308,500 and test set 20,000. Paris dataset is building view dataset, and we divided Paris into training set 14,900 and test set 100. All of images are resized to 128 × 128 and normalized to [0,1]. These normalized inputs of [0,1] can accelerate the training of model, and it also summarizes the statistical distribution of uniform samples.

4.2 Experimental Setup

All experiments are implemented on Pytorch framework with Ubuntu 16.04, Python 3.6.9, PyTorch 1.2.0, and RTX 2080TI GPU. Moreover, we set a batch size of 64, and use Adam optimizer to train our network with an initial learning-rate of 0.00001, and the orthogonal method is used to initialize the parameters of model. Although the network consists of two paths, it is trained in an end-to-end style. We also employ a LSGAN loss to make the training stable. In the training procedure, we update the discriminator once and update the generator once to complete the adversarial training.

The test input is the masked image with missing center regular holes or long strips. Note that, during test, we only use the bottom blue path to output final results. During training time, our model spent 6 days and 5 days on Places2 and Paris datasets respectively, while PICnet spent 7 days and 6 days on Places2 and Paris datasets respectively. Therefore, it proves that our method is more efficient for training times.

4.3 Evaluation Metrics

We compare our method (PICnet-SP-LM) with PICnet and its variants (PICnet-S (PICnet with style loss) and PICnet-SP (PICnet with style loss and perceptual loss)) in terms of qualitative and quantitative aspects. In the qualitative aspect, we can visually judge whether the outpainting parts are fine or bad. In the quantitative aspect, six types of metrics are used to measure the performance of different methods:

(1) Inception Score (IS) [40] is a common quantitative metric which is used to judge the quality of generated images. GANs, which can generate clear and diverse images, are considered as good generated models. *IS* can be used to measure the clarity and diversity of images. Formally, *IS* is defined as follow:

$$IS = \exp(E_{y\sim P_g} D_{KL}(p(z|y)||p(z))) \tag{7}$$

where g is the generator, y denotes the generated image, and z is the label predicted by the pre-trained Inception V3 model. The higher IS score signifies that the generated images are clearer and more diverse.

(2) Another metric usually used to measure the quality of GAN is Frechet Inception Distance (FID) [41]. FID aims to estimate the distance between the feature vectors of generated image and ground truth in a same domain. Formally:

$$FID = ||\mu_x - \mu_y||_2^2 + Tr\left(\Sigma_x + \Sigma_y - 2(\Sigma_x\Sigma_y)^{\frac{1}{2}}\right) \tag{8}$$

where x denotes the ground truth and y denotes the generated image. μ is the mean value of eigenvectors, and Σ is the covariance matrix of eigenvectors. The lower FID score also means that the generated images are higher-quality for clarity and diversity.

(3) Structural similarity (SSIM) aims to evaluate the quality of image based on the luminance, contract and structure of two images. Formally,

$$SSIM = \frac{(2\mu_x\mu_y + c_1)(2\sigma_{xy} + c_2)}{(\mu_x^2 + \mu_y^2 + c_1)(\sigma_x^2 + \sigma_y^2 + c_2)} \tag{9}$$

where x, y denote ground truth and generated image respectively, μ_x is the mean value of x, σ_x denotes the variance of x, and σ_{xy} denotes the covariance of x and y. The higher SSIM means the generated images possess finer luminance, contract and structure.

(4) Peak signal-to-noise ratio (PSNR) is a full reference estimation metric, and it is used to measure the degree of image distortion. Formally,

$$PSNR = 10 \cdot log_{10}\left(\frac{MAX_I^2}{MSE}\right) \tag{10}$$

where MAX_I^2 denotes the max pixel value in an image, and MSE is the abbreviation of mean square error. A higher PSNR score signifies the generated images are more natural.

(5) L_1 loss measures the pixel-wise difference by computing the L1 distance. Formally,

$$L_1\ loss = \frac{1}{m}\sum_{(i,j)} |x^{i,j} - y^{i,j}| \tag{11}$$

where x, y denote ground truth and generated image respectively, (i, j) denotes the position in the image, and m signifies the number of total elements. The lower L_1 loss means generated images are closer to ground truths for pixel-wise difference.

(6) RMSE is used to measure the deviation between generated image and ground truth. Formally,

$$RMSE = \sqrt{\frac{1}{m}\sum_{(i,j)} (x^{i,j} - y^{i,j})^2} \tag{12}$$

where x, y denote ground truth and generated image respectively, (i, j) denotes the position in the image, and m signifies the number of total elements. Similarly, the lower RMSE means generated images are closer to ground truths.

4.4 Qualitative Results

Figs. 2 and 3 illustrate the qualitative results of different methods with 64 × 64 valid pixels input on the different datasets. It is easy to see that the original PICnet generated blurry textures and distorted structures in the outpainting areas (see Fig. 2c). To solve the existing problems, we first introduce perceptual loss and style loss. For style loss, PICnet-S (PICnet with style loss) could improve the existing distorted structures, and these coarse results become much smoother (see Fig. 2d). Furthermore, we used both style loss and perceptual loss in PICnet (denoted as PICnet-SP) (PICnet with style loss and perceptual loss) to improve the outpainting results. We can see the details from Fig. 2e. Compared with the results of PICnet-S, PICnet-SP exhibits better on the Places2. For instance, with style loss and perceptual loss, the results are more realistic and more natural in general. To further improve the quality of outpainting images, we fuse the texture information in the low-level feature map of encoder, and simultaneously incorporate these aggregated features reusability with semantic (structural) information of deep feature map by element-wise adding in decoder. Simultaneously, we designed multi-patch discriminator into the network. This could utilize more sophisticated texture information to generate more authentic outpainting images (see Fig. 2f). We can see that our PICnet-SP-LM achieved a more authentic outpainting result. Moreover, we also find a similar effect on the Paris dataset. In the Fig. 3c, the vanilla PICnet method produces poor results which are filled with fuzzy contents and shadows. However, the outpainting parts are improved a lot when we add style loss alone or both style loss and perceptual loss (see Figs. 3d and 3e). Specifically, these shadows disappear in some degree and the blurry textures become clearer. Fig. 3f with the low-level feature fusion and the multi-patch discriminator exhibits better than the former methods. This proves that low-level feature fusion and multi-patch discriminator could promote the network to generate higher-quality outpainting images.

To further evaluate the effectiveness of our method, we set 128 × 64 valid pixels as the input of network (see Figs. 4b and 5b). Figs. 4 and 5 show the qualitative results of different methods on Places2 and Paris, respectively. From Figs. 4c and 5c, the original PICnet produces poor outpainting results with apparent boundaries and warped structures. Nonetheless, these situations are greatly improved when we use perceptual loss and style loss. In the Figs. 4 and 5, the structures become more natural and clearer (see Figs. 4d, 4e, 5d and 5e). Moreover, Figs. 4f and 5f, generated by our PICnet-SP-LM, reach the higher effect than the others. Thus, these results once again demonstrate that both

low-level feature fusion and the multi-patch discriminator are instrumental for network to improve the quality of outpainting images.

Figure 2: Qualitative results of different methods with 64 × 64 valid pixels' input on the Places2 dataset

4.5 Quantitative Results

The qualitative results of different methods on both Paris and Places2 datasets with different inputs are shown in Tabs. 1–4. The quantitative results with 64 × 64 valid pixels' input on Paris and Places2 are shown in the Tabs. 1 and 2. In the Tab. 1, we exhibit the quantitative metrics of 20,000 test images on the Places2. In the experiments, our method with low-level feature fusion and the multi-patch discriminator also achieves better metrics. Specially, our PICnet-SP-LM method achives the lower 30.81 for FID, signifying our model can realize clearer and more diverse outpainting results. The higher PSNR of 13.72 and SSIM of 0.4261, proving our results have a better image structure. Besides, we also obtain lower L1 loss of 34.47 and RMSE of 64.76, which indicates our results are closer to ground truths for pixel difference. Tab. 2 shows the quantitative metrics on the Paris. As result of the limitation of the 100 test images of Paris, we only measure the metrics SSIM and RMSE. From the quantitative results, low-level feature fusion and multi-patch discriminator again improve the results generated by the vanilla PICnet.

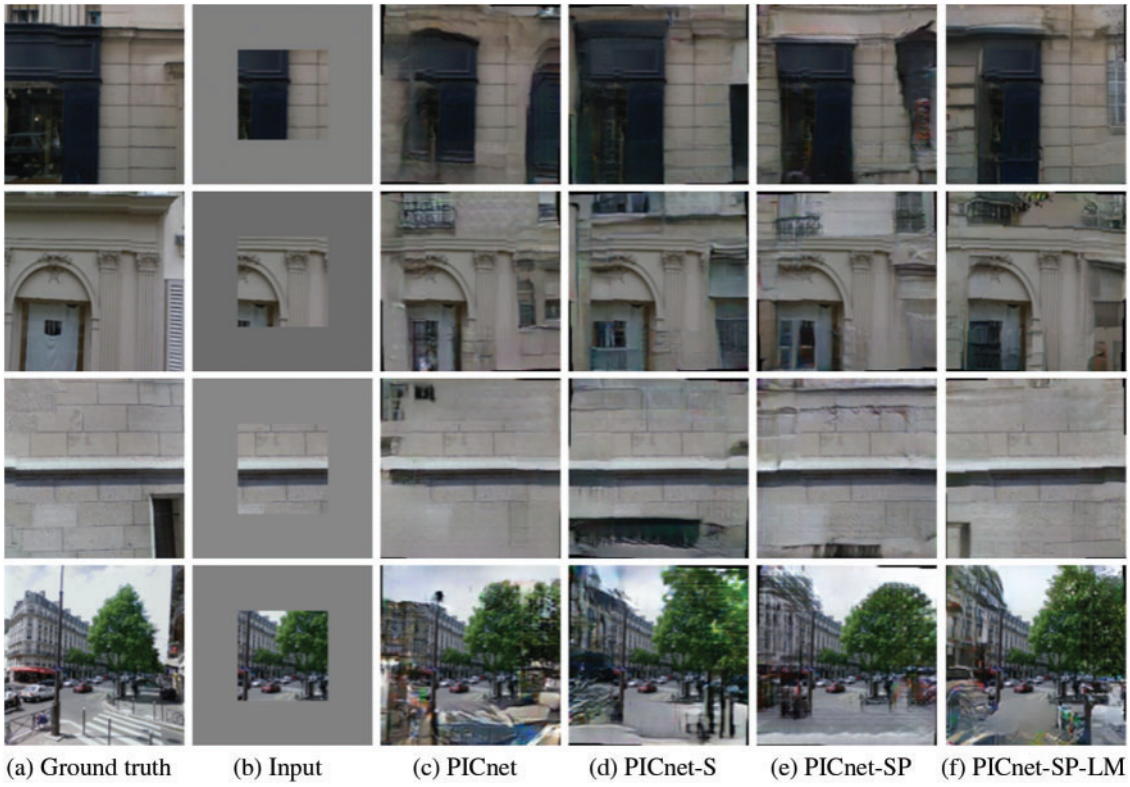

Figure 3: Qualitative results of different methods with 64 × 64 valid pixels' input on the Paris StreetView dataset

Furthermore, Tabs. 3 and 4 show the quantitative results of different methods with 128 × 64 valid pixels' input on Places2 and Paris. The effect of low-level feature fusion and the multi-patch discriminator once presents in the tables. Vanilla PICnet method produces the poor results which have lower-quality quantitative metrics. Contrarily, the quantitative metrics of outpainting results produced by PICnet-SP-LM can realize a better degree. Specially, with the effect of the low-level feature fusion and the multi-patch discriminator, PICnet-SP-LM achieves higher PSNR of 16.78 and SSIM of 0.6452 on the Places2 dataset. Meanwhile, PICnet-SP-LM also realizes the lower FID of 9.99 and L1 loss of 19.25. In addition, on the Paris dataset, PICnet-SP-LM also exhibits better for SSIM and RMSE. All the experiments demonstrate that both low-level feature fusion and the multi-patch discriminator are beneficial for outpainting network to improve the quality of outpainting images.

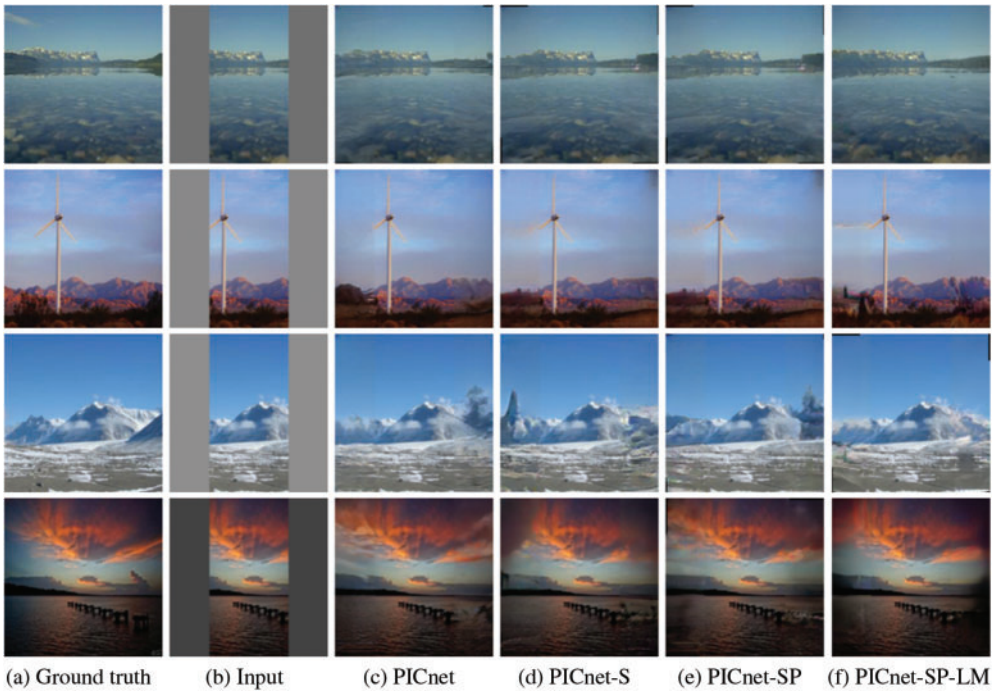

Figure 4: Qualitative results of different methods with 128 × 64 valid pixels' input on the Places2 dataset

Figure 5: Qualitative results of different methods with 128 × 64 valid pixels' input on the Paris StreetView dataset

Table 1: Quantitative results of different methods with 64×64 valid pixels' input on the Places2 dataset

Method	IS	FID	PSNR	L_1 loss	SSIM	RMSE
PICnet	5.60	34.81	12.95	38.29	0.4116	70.21
PICnet-S	5.49	31.87	12.92	38.21	0.4109	70.54
PICnet-SP	**5.69**	32.39	13.11	37.64	0.4140	68.76
PICnet-SP-LM	5.59	**30.81**	**13.72**	**34.47**	**0.4261**	**64.76**

Table 2: Quantitative results of different methods with 64×64 valid pixels' input on Paris StreetView. Because the limitation of the 100 test images of Paris StreetView, we only evaluate the SSIM and RMSE

Method	SSIM	RMSE
PICnet	0.4248	55.63
PICnet-S	0.4441	55.58
PICnet-SP	0.4367	55.89
PICnet-SP-LM	**0.4495**	**52.42**

Table 3: Quantitative results of different methods with 128×64 valid pixels' input on the Places2 dataset

Method	IS	FID	PSNR	L_1 loss	SSIM	RMSE
PICnet	4.05	13.94	16.38	20.62	0.6407	57.71
PICnet-S	4.11	12.78	16.38	20.66	0.6442	57.89
PICnet-SP	4.09	11.78	16.42	20.45	0.6438	57.49
PICnet-SP-LM	**4.13**	**9.99**	**16.78**	**19.25**	**0.6452**	**55.13**

Table 4: Quantitative results of different methods with 128×64 valid pixels' input on Paris StreetView. Because the limitation of the 100 test images of Paris StreetView, we only evaluate the SSIM and RMSE

Method	SSIM	RMSE
PICnet	0.6503	49.35
PICnet-S	0.6602	47.39
PICnet-SP	0.6654	45.34
PICnet-SP-LM	**0.6663**	**45.10**

4.6 Ablation Study

In addition, we also implement other experiments for further selecting the better PICnet-SP-LM method. Tab. 5 is the quantitative results of implemental experiments on the Places2 dataset. Specifically, PICnet-SP-LM-1 and PICnet-SP-LM-2 are the different hyper parameters for reconstruction loss and KL loss, respectively. (PICnet-SP-LM-1 with hyper parameter 20 for reconstruction loss and

hyper parameter 20 for KL loss, and PICnet-SP-LM-2 with hyper parameter 20 for reconstruction loss and hyper parameter 40 for KL loss.) From the experimental results, PICnet-SP-LM-1 achieves a better degree. Thus, PICnet-SP-LM-3 and PICnet-SP-LM-4 adopt the hyper parameters of PICnet-SP-LM-1. PICnet-SP-LM-3 utilizes one layer's aggregated features, and PICnet-SP-LM-4 utilizes two layers' aggregated features. Apparently, PICnet-SP-LM-4 utilizing more aggregated features achieves a better effect. Therefore, PICnet-SP-LM-4 is an optimal experimental setup, which could generate more natural and more realistic outpainting results. Moreover, for the qualitative aspect, the results generated by PICnet-SP-LM-4 are also clearer and more authentic than other methods. In the Fig. 6, we also select some outpainting results with borders in baseline model. Then we relieve or eliminate these borders through gradually adding our core blocks, which could present the obvious effect of these core blocks.

Table 5: Quantitative results of ablation study with 64 × 64 valid pixels' input on the Places2 dataset

Method	IS	FID	PSNR	L_1 loss	SSIM	RMSE
PICnet-SP-LM-1	5.13	38.11	13.70	34.59	0.4165	**63.04**
PICnet-SP-LM-2	5.09	34.99	13.67	34.54	0.4246	65.12
PICnet-SP-LM-3	5.06	40.33	13.56	34.89	0.4161	65.71
PICnet-SP-LM-4	**5.59**	**30.81**	**13.72**	**34.47**	**0.4261**	64.76

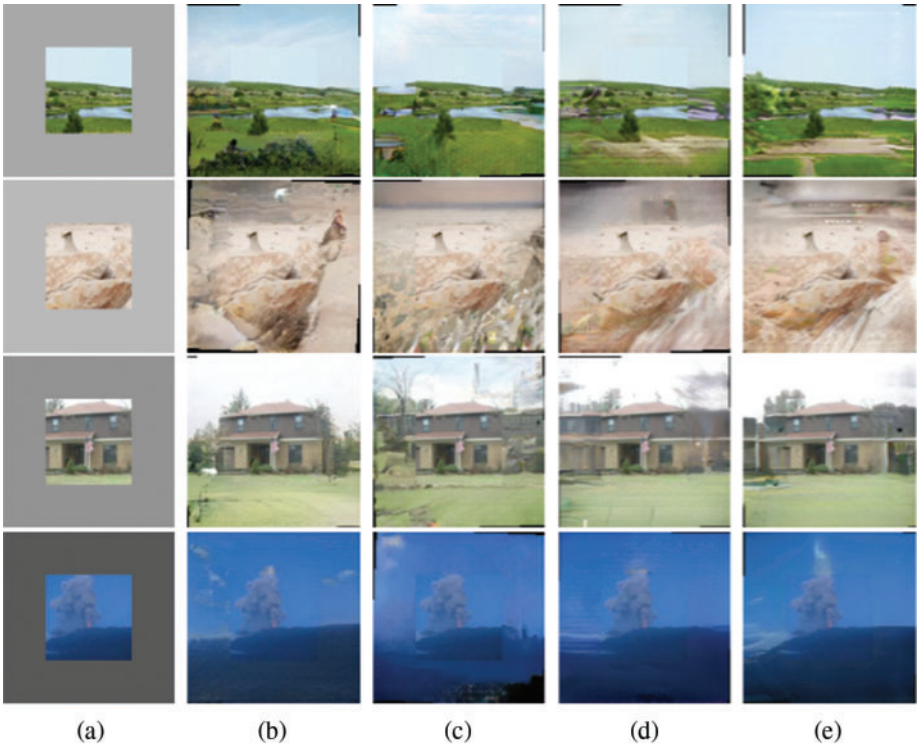

Figure 6: Qualitative results of ablation study on the Places2 dataset. (a) Input, (b) PICnet-SP-LM-1, (c) PICnet-SP-LM-2, (d) PICnet-SP-LM-3, (e) PICnet-SP-LM-4

5 Conclusion

In fact, image outpainting plays an important role in image processing field, and it can be also used to promote the image inpainting. In this paper, we present a perceptual image outpainting method, which is assisted by low-level feature fusion and multi-patch discriminator. In details, we first fuse the low-level texture information in the encoder, and simultaneously incorporate these fused features with semantic (or structural) information of deep feature map, which could promote the network to generate finer outpainting results. At the same time, we also present a multi-patch discriminator to enhance the generated image texture, which effectively judges the generated image from the different level features and impels our network to produce more natural and clearer outpainting results. To fully evaluate our model, we implement experiments on Places2 and Paris dataset. Finally, the experimental results show that our method is better than PICnet for qualitative effects and quantitative metrics, which proves the effectiveness and efficiency of our method for image outpainting task. In the future, we will further study more challenging image outpainting field, such as the input images with bigger missing regions. We also try to realize higher-quality outpainting results.

Acknowledgement: I would like to thank those who helped me generously in this research.

Funding Statement: This work was supported by the Sichuan Science and Technology program (2019JDJQ0002, 2019YFG0496, 2021016, 2020JDTD0020), and partially supported by National Science Foundation of China 42075142.

Conflicts of Interest: The authors declare that they have no conflicts of interest to report regarding the present study.

References

[1] Z. Li, J. Zhang, K. Zhang and Z. Li, "Visual tracking with weighted adaptive local sparse appearance model via spatio-temporal context learning," *IEEE Transactions on Image Processing*, vol. 27, no. 9, pp. 4478–4489, 2018.

[2] Z. Li, W. Wei, T. Zhang, M. Wang, S. Hou *et al.,* "Online multi-expert learning for visual tracking," *IEEE Transactions on Image Processing*, vol. 29, pp. 934–946, 2020.

[3] Y. Song, J. Sohl-Dickstein and D. P. Kingma, "Score-based generative modeling through stochastic differential equations," arXiv preprint arXiv:2011.13456, 2020.

[4] S. Zhao, J. Cui and Y. Sheng, "Large scale image completion via co-modulated generative adversarial networks," arXiv preprint arXiv:2103.10428, 2021.

[5] Y. Zeng, J. Fu and H. Chao, "Aggregated contextual transformations for high-resolution image inpainting," arXiv preprint arXiv:2104.01431, 2021.

[6] A. Gumaei, M. Al-Rakhami and H. AlSalman, "Dl-har: Deep learning-based human activity recognition framework for edge computing," *Computers, Materials & Continua*, vol. 65, no. 2, pp. 1033–1057, 2020.

[7] B. Hu and J. Wang, "Deep learning for distinguishing computer generated images and natural images: A survey," *Journal of Information Hiding and Privacy Protection*, vol. 2, no. 2, pp. 37–47, 2020.

[8] F. Li, J. Zhang, E. Szczerbicki, J. Song, R. Li *et al.,* "Deep learning-based intrusion system for vehicular ad hoc networks," *Computers, Materials & Continua*, vol. 65, no. 1, pp. 653–681, 2020.

[9] B. Coloma, B. Marcelo, C. Vient, S. Guillermo and V. Joan, "Filling-in by joint interpolation of vector fields and gray levels," *IEEE Transactions on Image Processing*, vol. 10, no. 8, pp. 1200–1211, 2001.

[10] J. Sun, L. Yuan, J. Jia and H. Shum, "Image inpainting with structure propagation," *ACM SIGGRAPH*, vol. 25, no. 8, pp. 861–868, 2005.

[11] M. Bertalmio, G. Sapiro, V. Caselles and C. Ballester, "Image inpainting," in *Proc. 27th Annual Conference on Computer Graphics and Interactive Techniques*, New York, NY, USA, pp. 417–424, 2000.

[12] H. James and E. Alexei, "Scene inpainting using millions of photographs," *ACM Transactions on Graphics*, vol. 26, no. 3, pp. 4–10, 2007.
[13] K. Sohn, H. Lee and X. Yan, "Learning structured output representation using deep conditional generated models," *Advances in Neural Information Processing Systems*, vol. 28, pp. 3483–3491, 2016.
[14] Y. Li, S. Liu, J. Yang and M. H. Yang, "Generated face inpainting," in *Proc. CVPR*, Honolulu, HI, USA, pp. 3911–3919, 2017.
[15] R. Yeh, C. Chen, T. Y. Lim, M. Hasegawa-Johnson and M. N. Do, "Semantic image inpainting with perceptual and contextual losses," arXiv preprint arXiv:1607.07539 2.3, 2016.
[16] I. Goodfellow, J. Pouget-Abadie and J. Mirza, "Generative adversarial nets," *Advances in Neural Information Processing Systems*, vol. 27, pp. 654–656, 2014.
[17] T. Karras, T. Aila, S. Laine and J. Lehtinen, "Progressive growing of gans for improved quality, stability, and variation," arXiv preprint arXiv:1710.10196, 2018.
[18] J. Johnson, A. Alahi and L. Fei-Fei, "Perceptual losses for real-time style transfer and super-resolution," in *Proc. ECCV*, Amsterdam, The Netherlands, pp. 694–711, 2016.
[19] L. A. Gatys, A. S. Ecker and M. Bethge, "Image style transfer using convolutional neural networks," in *Proc. CVPR*, Las Vegas, NV, USA, pp. 2414–2423, 2016.
[20] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," arXiv preprint arXiv:1409.1556, 2014.
[21] S. Darabi, E. Shechtman, C. Barnes, D. B. GolLMan and P. Sen, "Image melding: Combining inconsistent images using patch-based synthesis," *ACM Transactions on Graphics*, vol. 31, no. 4, pp. 1–10, 2012.
[22] J. B. Huang, S. B. Kang, N. Ahuja and J. Kopf, "Image inpainting using planar structure guidance," *ACM Transactions on Graphics*, vol. 33, no. 4, pp. 1–10, 2014.
[23] M. Wilczkowiak, G. J. Brostow, B. Tordoff and R. Cipolla, "Hole filling through photomontage," in *Proc. BMVC*, Glasgow, UK, 2005.
[24] S. Esedoglu and J. Shen, "Digital inpainting based on the mumford-shaheuler image model," *European Journal of Applied Mathematics*, vol. 13, no. 4, pp. 353–370, 2002.
[25] D. Liu, X. Sun, F. Wu, S. Li and Y. Q. Zhang, "Image compression with edge-based inpainting," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 17, no. 10, pp. 1273–1287, 2007.
[26] A. A. Efros and W. T. Freeman, "Image quilting for texture synthesis and transfer," in *Proc. 28th Annual Conference on Computer Graphics and Interactive Techniques*, New York, NY, USA, pp. 341–346, 2001.
[27] D. Simakov, Y. Caspi, E. Shechtman and M. Irani, "Summarizing visual data using bidirectional similarity," in *Proc. CVPR*, Anchorage, AK, USA, pp. 1–8, 2008.
[28] C. Barnes, E. Shechtman, A. Finkelstein and D. B. Gollman, "Patchmatch: A randomized correspondence algorithm for structural image editing," *ACM Trans. Graph.*, vol. 28, no. 3, pp. 24, 2009.
[29] D. Pathak, P. Krahenbuhl, J. Donahue, T. Darrell, and A. A. Efros, "Context encoders: Feature learning by inpainting," in *Proc. CVPR*, Honolulu, HI, USA, pp. 1457–1460, 2017.
[30] Y. Chao, L. Xin and L. Zhe, "High-resolution image inpainting using multi-scale neural patch synthesis," in *Proc. CVPR*, Honolulu, HI, USA, pp. 1457–1460, 2017.
[31] S. Iizuka, E. Simo-Serra and H. Ishikawa, "Globally and locally consistent image inpainting," *ACM Transactions on Graphics*, vol. 36, no. 4, pp. 1–14, 2017.
[32] G. Liu, F. A. Reda, K. J. Shih, T. -C. Wang, A. Tao *et al.,* "Image inpainting for irregular holes using partial convolutions," in *Proc. ECCV*, Munich, Germany, pp. 85–100, 2018.
[33] C. Zheng, C. Tat-Jen and C. Jianfei, "Pluralistic image inpainting," in *Proc. CVPR*, Long Beach, CA, USA, pp. 1438–1447, 2019.
[34] S. Mark and R. Gili, "Painting outside the box: Image outpainting with GANs," arXiv preprint arXiv:1808.08483, 2018.
[35] D. P. Kingma and M. Welling, "Auto-encoding variational Bayes," arXiv preprint arXiv:1312.6114, 2013.
[36] A. Sengupta, Y. Ye and R. Wang, "Going deeper in spiking neural networks: VGG and residual architectures," *Frontiers in Neuroscience*, vol. 13, no. 6, pp. 95, 2019.

[37] X. Mao, Q. Li, H. Xie, R. YK Lau, Z. Wang *et al.,* "Least squares generative adversarial networks," in *Proc. ICCV*, Venice, Italy, pp. 2813–2821, 2017.
[38] B. Zhou, A. Lapedriza, A. Khosla, A. Oliva and A. Torralba, "Places: A 10 million image data-base for scene recognition," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 40, no. 6, pp. 1452–1464, 2017.
[39] C. Doersch, S. Singh, A. Gupta, J. Sivic and A. Efros, "What makes Paris look like Paris," *ACM Transactions on Graphics,*, vol. 31, no. 4, pp. 14, 2012.
[40] T. Salimans, I. Goodfellow, W. Zaremba, V. Cheung, A. Radford *et al.,* "Improved techniques for training gans," *Advances in Neural Information Processing Systems*, vol. 29, pp. 2234–2242, 2016.
[41] M. Heusel, H. Ramsauer, T. Unterthiner, B. Nessler and S. Hochreiter, "Gans trained by a two time-scale update rule converge to a local nash equilibrium," *Advances in Neural Information Processing Systems*, vol. 30, pp. 25–34, 2017.