

# Double Global and Local Information-based Image Inpainting

Shibin Wang<sup>1</sup>(✉), Wenjie Guo<sup>1</sup>, Shiying Zhang<sup>1</sup>, Xuening Guo<sup>1</sup>, and Jiayi Guo<sup>1</sup>

<sup>1</sup> School of Computer and Information Engineering, Henan Normal University, No. 46, Construction Road, Xinxiang, 453007, China  
wangshibin@htu.edu.cn

**Abstract.** With the development of deep learning, significant progress has been made in image inpainting. Deep learning-based image inpainting methods can generate visually plausible inpainting results. However, the inpainting images may include the distortions or artifacts, especially at boundaries and high-texture regions. To address these issues, we propose an improved two-stage inpainting model with double local and global information. In the first stage, a Local Binary Pattern (LBP) learning network based on the U-Net architecture is employed to accurately predict the semantic and structural information of the missing regions. In the second stage, the double local and global network based on spatial attention module and Double-PatchGAN Discriminator (DPD) are proposed for further refinement. Aim to achieve the accurate, realistic, and high-quality inpainting results, the Multiple Loss Functions (MLF) is designed to strengthen the information at different levels. Extensive experiments conducted on public datasets, including CelebA-HQ, Places2 and Paris StreetView, demonstrate that our model outperforms several existing methods in terms of image inpainting.

**Keywords:** Deep learning, Image inpainting, Local Binary Pattern, Double-PatchGAN Discriminator, Multiple Loss Functions.

## 1 Introduction

Image inpainting is one of the fundamental tasks in computer vision. High-quality image inpainting has been applied in various real-world domains, such as digital artifact inpainting, image editing, medical image processing, and even in criminal investigations to enhance damaged images for better identification of suspects. However, achieving satisfied results becomes challenging when the large missing regions contain the complex texture content.

Recently, the learning-based methods treat image inpainting as an end-to-end mapping from the damaged images to restored complete images, which can generate new content by training on large-scale datasets. Some methods employ recurrent networks, applying the generator in an iterative cyclic manner [1, 2]. Two-stage networks [3-5] generate coarse structural information or edge details in the first stage, and then refine the damaged image to complete the image inpainting in the second stage. However, these methods often overlook the semantic coherence and feature continuity of the

generated content. The artifacts may still occur when they deal with complex texture in images like human eyes. Wu et al. [5] proposed a two-stage network, where the first stage predicts structural information for the missing regions, so as to guide the second-stage image inpainting network to better fill in the missing pixels. Moreover, the utilization of the designed spatial attention layer in the second stage significantly enhances the semantic consistency of the entire image and improvements in the inpainting of human eyes. However, there is still room for improvement in areas such as hair, boundaries, the authenticity of the inpainting results, and addressing large-scale missing regions with complex structures.

To address above issues, we propose a two-stage network image inpainting model, which fuses the multiple-level local and global information. The first network extracts the great amount of structural information to guide the painting task of the second network. As confirmed in [5, 6], visually close-to-original images can be reconstructed solely from their Local Binary Pattern (LBP) features. We choose the LBP learning network to recover the missing region. Compared with previous papers [5], in the second network, we have designed Double-PatchGAN Discriminator (DPD) to establish local and global correlations in the encoder and discriminator. Additionally, we design the Multiple Loss Functions (MLF) by considering different aspects of image quality metrics comprehensively. Compared with several state-of-art methods, the experimental results demonstrate that our model can generate better details and consistent style.

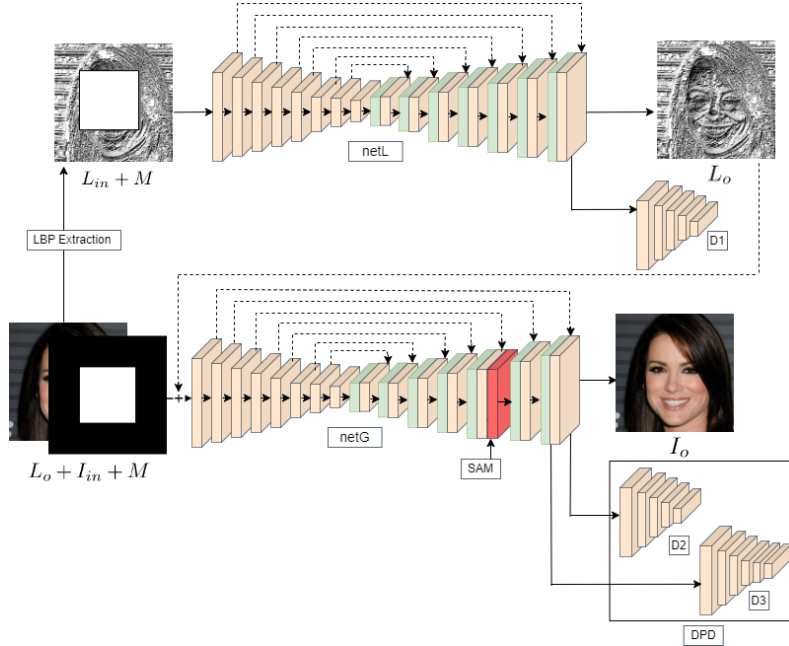
## 2 Related Works

In the early works, Pathak et al. [7] proposed a contextual encoder-decoder architecture trained by using pixel-level reconstruction loss and adversarial loss. Liu et al. [8] designed the partial convolution operations, which consider only valid pixels using masks and normalize the output accordingly. The partial convolutions treat all input pixels as valid and the mask updating strategy is manually crafted and non-learnable. To address above issues, Yu et al. [9] proposed the gated convolutions to replace the assumption that all elements are valid pixels. They extend the partial convolutions by a learnable dynamic feature selection mechanism for each channel at every spatial position across all layers. Zeng et al. [10] introduced a model, called the Pyramid Context Encoder Network, which includes a multi-scale decoder with a deep-supervised pyramid loss and adversarial loss. Wang et al. [11] introduced a special multi-stage attention module to extend the partial convolutions [8] by considering the consistency of structure and the fineness of details. Li et al. [12] designed a cyclic feature inference network, which consists of a plug-and-play cyclic feature inference module and a knowledge-consistent attention module. It propagates the confidence regions from the boundaries in the feature space to the center. However, the high recursive computational costs limit the efficiency in practical applications. Additionally, the recursive feature mapping-based methods may lead to structural discontinuities in the output images. Wu et al. [5] obtained the texture and structure of an image by learning its local binary patterns and introduce a spatial attention mechanism to focus on contextual information in the

missing region. Although these existing methods have made improvements in image processing, they can only capture limited relationships between textures and edges, failing to fully express the complex structures and features in the image. In order to improve the quality of image inpainting, we propose a novel network architectures from local and global perspective, which will be described in detail in Section 3.

### 3 Method

In this section, we will introduce our two-stage network architecture, see Fig. 1.  $I_g$  represents the ground-truth,  $L_g$  be the ground-truth LBP.  $I_{in}$  represents the damaged image with the missing regions filled with white pixels, whilst  $L_{in}$  denotes the LBP structural information extracted from the damaged image  $I_{in}$  in the grayscale channel.  $M$  represents a binary mask, where 1 represents the missing regions and 0 represents the known regions. During training, in the first stage,  $(L_{in}, M)$  is input into the generator  $G_1$  and discriminator  $D_1$ , working together to generate the completed LBP structural information  $L_{out}$  and  $L_o$ . The output of the first stage is inpainted image  $L_{out}$ , where  $L_o$  is merged image. In the second stage,  $(L_o, I_{in}, M)$  as input and outputs  $I_{out}$  and  $I_o$ . Similarly,  $I_{out}$  is inpainted image,  $I_o$  is merged image. During testing, the process is similar, but the discriminator is not used. The details can be described as follows.



**Fig. 1.** The network architecture of our proposed method.

### 3.1 LBP Learning Network

LBP [13] can obtain a significant amount of structural information, the fewer parameters. Here, we choose LBP as the first network in our model. LBP network consists of a generator  $G_l$  and a discriminator  $D_l$ . The generator follows a U-Net architecture [14], which consists of an encoder and a decoder. In the encoder, each layer consists of a  $4 \times 4$  convolution, a LeakyReLU [15] with  $\alpha=0.2$  and an InstanceNorm2d [16]. The encoder and decoder stages have symmetric structures, where the convolution and LeakyReLU in the encoder are replaced by transposed convolution and ReLU [17] in the decoder respectively. Skip connections are used to connect the corresponding layers of the encoder and decoder. The discriminator adopts the PatchGAN approach [18].

To ensure stable training of the model, the choice of loss functions plays a crucial role. Reconstruction loss then can be defined as follows:

$$L_r = \|L_o - L_g\|_2 \quad (1)$$

$$L_o = L_{in} \odot (1 - M) + L_{out} \odot M \quad (2)$$

Additionally, the second loss function, adversarial loss [19] is defined as follows:

$$\mathcal{T}_{adv1} = \min_{G_1} \max_{D_1} E_{L_g}[\log D_1(L_g)] + E_{L_{in}}[\log(1 - D_1(G_1(L_{in}, M)))] \quad (3)$$

We propose to add the weighted  $L_1$  loss for pixel-wise reconstruction [20] and total variation (TV) loss [8] to the original LBP network.

$$L_{valid} = \frac{1}{\text{Sum}(1-M)} \|(L_{out} - L_g) \odot (1 - M)\| \quad (4)$$

$$L_{hole} = \frac{1}{\text{Sum}(M)} \|(L_{out} - L_g) \odot M\| \quad (5)$$

$$L_{pwr} = L_{valid} + \lambda_h L_{hole} \quad (6)$$

where  $\odot$  denotes the element-wise product operation,  $M$  indicates the number of non-zero elements in  $M$ , and  $\lambda_h$  is a balancing factor. Additionally, TV loss provides a regularization mechanism to suppress high-frequency noise and preserve details in the image, striking a balance between smoothness and detail preservation to achieve more natural and visually pleasing inpainting results.

$$L_{tv} = \|L_o(i, j + 1) - L_o(i, j)\| + \|L_o(i + 1, j) - L_o(i, j)\| \quad (7)$$

Finally, the loss function for the LBP learning network is defined by combining the aforementioned four types of loss.

$$L_{total} = \lambda_r L_r + \lambda_a \mathcal{T}_{adv1} + L_{pwr} + L_{tv} \quad (8)$$

and the parameters trading off different terms are set to be  $\lambda_h = 6$ ,  $\lambda_r = 10$  and  $\lambda_a = 0.2$  for the best performance in LBP learning network.

### 3.2 Image Inpainting Refinement Network

The architecture of our proposed image inpainting refinement network is similar to our enhanced LBP learning network, except that newly designed DPD and MLF, and spatial attention module [13] used in the fifth layer of the decoder.

**Double-PatchGAN Discriminator** PatchGAN [18] penalizes structure only at the patch scale. Instead of outputting a single true or false vector for the entire image, its discriminator attempts to classify the real or fake for each  $N$  patch in the image. However, it lacks an understanding of the global structure, leading to significant artifacts and distortions in the inpainting results. SN-PatchGAN discriminator [9] utilizes the spectral normalization technique to normalize the weight matrix. It improves the stability and convergence of the discriminator and generates recovered images with realistic texture details.

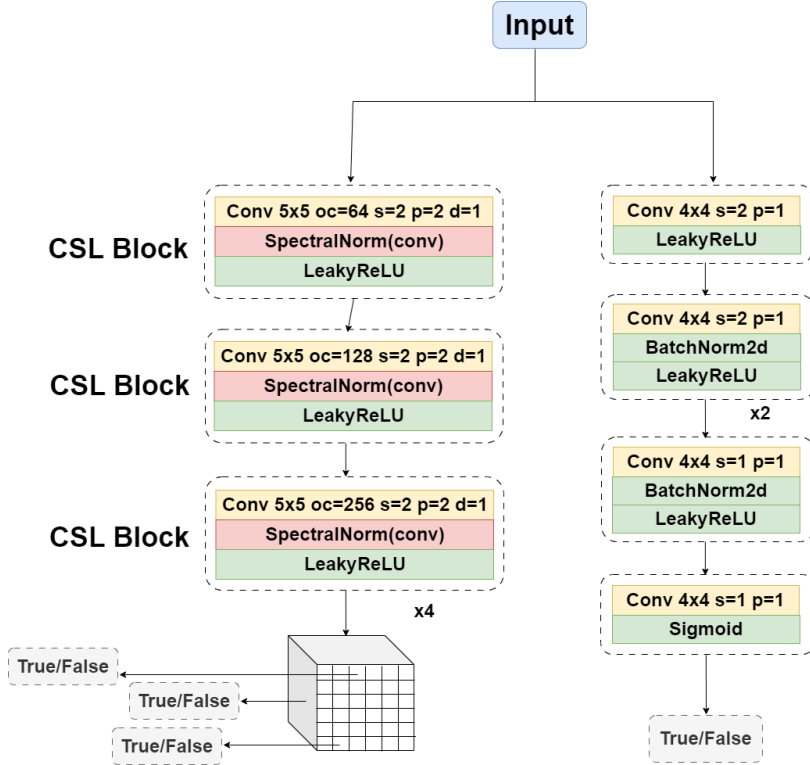


Fig. 2. Proposed DPD to introduce efficient local and global consistencies.

Inspired by the success of PatchGAN and SN-PatchGAN for image inpainting, we propose the DPD framework to effectively fill the damaged area, as shown in Fig. 2. The left branch is a global discriminator to focus on the spatial correlation between the damaged and undamaged regions. Its input consists of the image and mask and the

output is a 3D feature. Each CSL block consists of a  $5 \times 5$  convolution, SpectralNorm [21] and LeakyReLU with  $\alpha=0.2$ . The number of convolutional output channels in the first two CSL blocks are 64 and 128, respectively, while the number in the last four CSL blocks is 256. The right branch is a local discriminator to focus on the patch local information of the image, which consists of five  $4 \times 4$  convolutions. The first four layers use the LeakyReLU with  $\alpha=0.2$ , the Sigmoid for the last layer and the BatchNorm2d for normalization in the middle three layers.

Our objective function for local discriminator can be formulated as:

$$\mathcal{T}_{adv2} = \min_{G_2} \max_{D_2} E_{I_g}[\log D_2(I_g)] + E_{I_{in}}[\log(1 - D_2(G_2(I_{in}, M)))] \quad (9)$$

The generator and discriminator of the global discriminator can be formulated as:

$$\mathcal{T}_{adv3} = -E_{I_{in} \sim P_{I_{in}}(I_{in})}[D_3(G_2(I_{in}))] \quad (10)$$

$$\mathcal{T}_{D_3} = E_{I_g \sim P_{data}(I_g)}[ReLU(1 - D_3(I_g))] + E_{I_{in} \sim P_{I_{in}}(I_{in})}[ReLU(1 + D_3(G_2(I_{in})))] \quad (11)$$

where  $G_2$  represents image inpainting refinement network,  $D_2$  and  $D_3$  respectively represent the right and left branches of DPD.

**Multiple Loss Functions** To further enhance the quality of the inpainting results, we propose the MLF as the optimization objective. The multi-scale loss functions enable the comparison of differences between the inpainted results and the ground truth image in different scales. It can capture details and structural information in various levels within the image. Additionally, it effectively propagates stability during training. The multi-level loss function is defined as:

$$\mathcal{T}_m = \sum_{h \in d} \|\Phi_h(I_o) - \Phi_h(I_g)\|_2 \quad (12)$$

$$I_o = I_{in} \odot (1 - M) + I_{out} \odot M \quad (13)$$

Let  $I_g$  be the ground-truth, and its corresponding high-level features be  $\Phi_h(I_g)$ , where  $h$  is the layer index within  $G_2$ . As the training direction,  $\Phi_h(I_g)$  can optimize high-level features of the  $G_2$  globally, where  $d$  accommodates the indexes of all the convolution and deconvolution layers in  $G_2$ .

Building upon the multi-level loss functions, the addition of pixel-level loss yields more pronounced effects. We employ weighted  $L_1$  loss for pixel-level reconstruction, as  $L_1$  loss is more robust to outliers compared to  $L_2$  loss. Furthermore,  $L_1$  loss exhibits more stable gradient computation during training. The pixel-wise reconstruction loss [20] can be formulated as:

$$I_{valid} = \frac{1}{\text{Sum}(1-M)} \|(I_{out} - I_g) \odot (1 - M)\| \quad (14)$$

$$I_{hole} = \frac{1}{\text{Sum}(M)} \|(I_{out} - I_g) \odot M\| \quad (15)$$

$$I_{pwr} = I_{valid} + \lambda_h I_{hole} \quad (16)$$

The training objective  $I_{pwr}$  of Net<sub>G</sub> is similar with  $L_{pwr}$  of Net<sub>L</sub>, only replacing  $L_{out}$  with  $I_{out}$ ,  $L_g$  with  $I_g$  in the corresponding locations of  $L_{pwr}$ .

In our work, the reconstruction loss is defined as follow:

$$I_r = \|I_o - I_g\|_2 \quad (17)$$

Moreover, we introduce the perceptual loss [22] and style loss [23]. Perceptual loss be formulated as:

$$I_{per} = \sum_i \|\Psi_i(I_o) - \Psi_i(I_g)\| \quad (18)$$

where  $\Psi_i$  is the feature map of  $i$ -th layer in ImageNet-pretrained VGG-16 network.

Similarly, the style loss can be formulated as:

$$I_{style} = \sum_i \|\mathfrak{f}_i(I_o) - \mathfrak{f}_i(I_g)\| \quad (19)$$

Where  $\mathfrak{f}_i(\cdot) = \Psi_i(\cdot)\Psi_i(\cdot)^T$  is from [24].

Additionally, the introduction of Total Variation (TV) [8] loss promotes inpainted results with a smoother and more continuous appearance, reducing noise and discontinuities in the repaired image. The TV loss can be formulated as:

$$I_{tv} = \|I_o(i, j + 1) - I_o(i, j)\| + \|I_o(i + 1, j) - I_o(i, j)\| \quad (20)$$

The double adversarial loss is the same as Eqs. (9)(10)(11)

To this end, the total loss for Net<sub>G</sub> is:

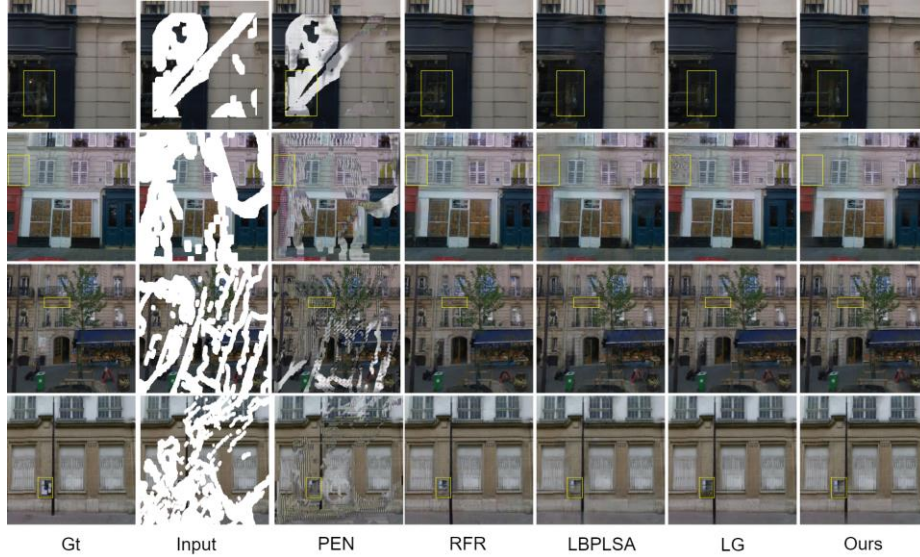
$$I_{total} = \lambda_m \mathcal{T}_m + I_{pwr} + \lambda_r I_r + I_{per} + \lambda_s I_{style} + I_{tv} + \lambda_{a1} \mathcal{T}_{adv2} + \mathcal{T}_{adv3} \quad (21)$$

The parameters trading off different terms are set to be  $\lambda_h = 6$ ,  $\lambda_r = 10$ ,  $\lambda_m = 0.01$ ,  $\lambda_s = 10$  and  $\lambda_{a1} = 0.2$  in image inpainting refinement network.

In conclusion, MLF as the optimization objective can significantly enhance the quality of inpainting results. By considering various aspects of loss functions, strengthening information at different levels, and providing diverse training signals, we can achieve more precise, realistic, and high-quality inpainting outcomes.

## 4 Experiments

We implement the proposed network architecture in PyTorch and train the network on NVIDIA A100 GPU. The parameters are optimized by Adam optimizer with a learning rate of 0.0002,  $\beta_1 = 0.5$ , and  $\beta_2 = 0.999$  for the generator and discriminator, separately. The difference is that the learning rate is 0.0001 for the global discriminator. In order to provide a fair comparison, we retrain the benchmark method based on our customized dataset partitioning approach, enabling us to derive both qualitative and quantitative results. Furthermore, an ablation study is undertaken to demonstrate the efficacy of the proposed DPD and MLF components.



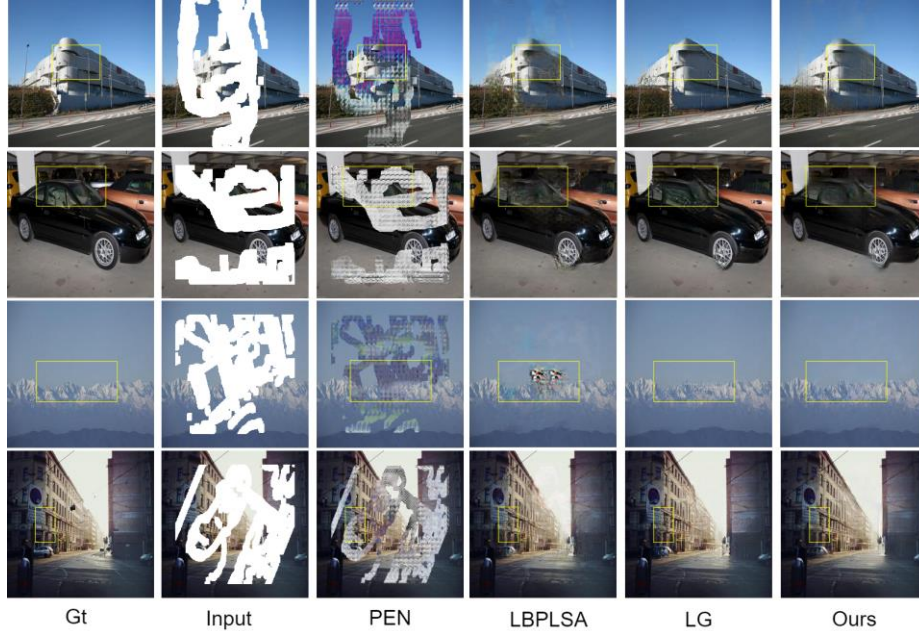
**Fig. 3.** Comparison of qualitative results between the proposed method and existing approaches on the Paris Street View dataset. Our proposed method generates more consistent structures. And we illustrate the inpainted results with different masks just for diverse comparisons like other existing inpainting methods PEN [10], RFR [12], LBPLSA [5], LG [20].

## 5 Results and Discussion

### 5.1 Datasets

In the experimental analysis, we adopt three publicly available datasets. The Paris StreetView [25] consists of 14,900 training images and 100 validation images captured from the streets of Paris. The CelebA-HQ [26] is a high-quality celebrity face dataset, containing a total of 30,000 images, with 28,000 for training and 2,000 for validation. The Places2 [27] comprises around 1.8M images from 365 distinct scene categories. In our experiments, we select 30,000 images for training and 2,000 images for testing. The mask comes from the the random mask NVIDIA dataset [8]. The masks with or without boundary constraints produce the different performance [8].





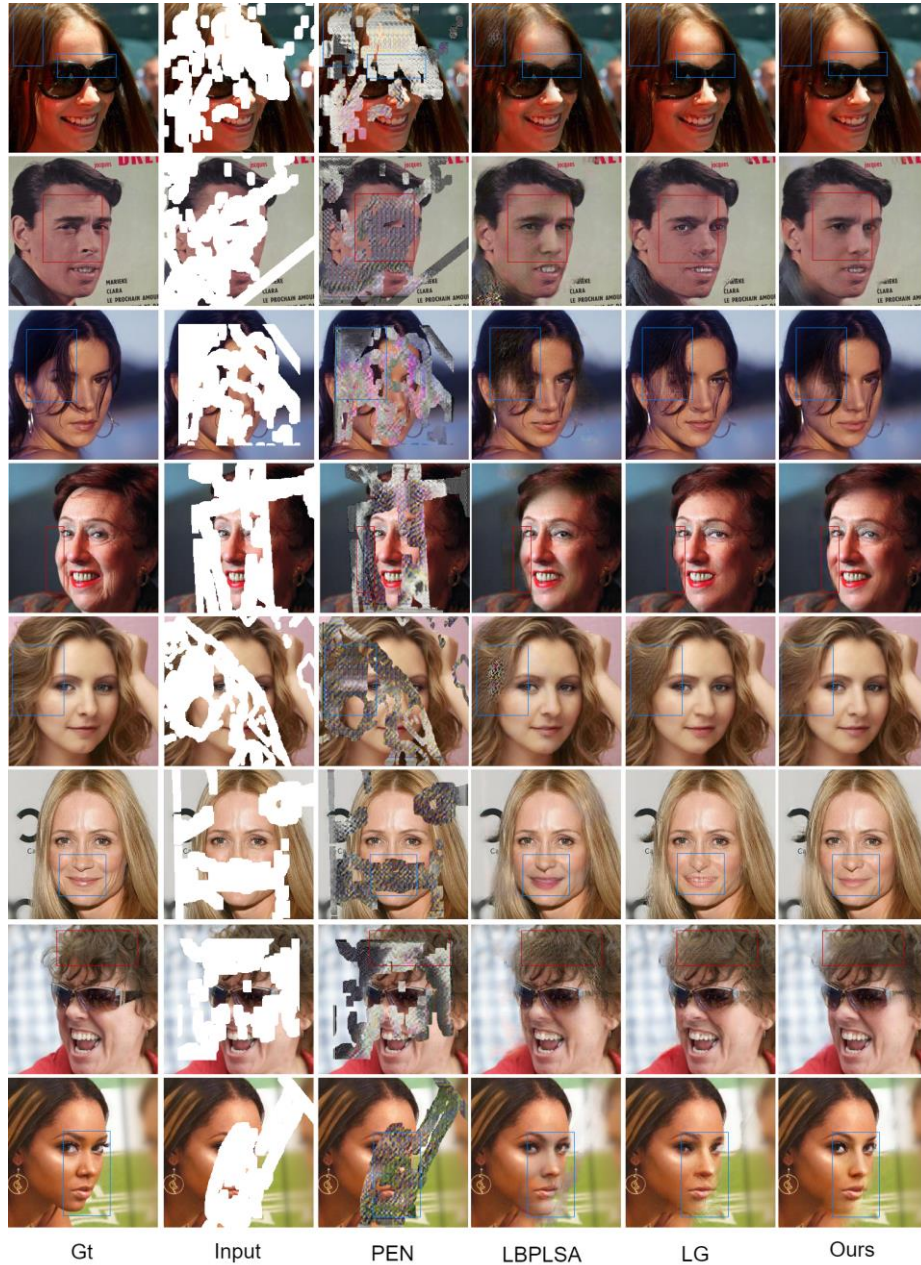
**Fig. 4.** Comparison of qualitative results between the proposed method and existing approaches on the Places2 dataset. Our proposed method generates more effective structural and texture information. And we illustrate the inpainted results with different masks just for diverse comparisons like other existing inpainting methods PEN [10], LBPLSA [5], LG [20].

## 5.2 Quantitative result analysis

We quantitatively compare our method with four representative advanced image inpainting methods:

- PEN [10]: Learning Pyramid-Context Encoder Network for High-Quality Image Inpainting.
- RFR [12]: Recurrent Feature Reasoning for Image Inpainting (CVPR 2020).
- LBPLSA [5]: Deep Generative Model for Image Inpainting with Local Binary Pattern Learning and Spatial Attention (CVPR 2020).
- LG [20]: Image Inpainting with Local and Global Refinement (IEEE Transactions on Image Processing, vol. 31, 2022).

We employ the Peak Signal-to-Noise Ratio (PSNR), the Structure Similarity Index Measure (SSIM) and the  $L_1$  loss as the joint metrics to assess the discrepancy between the original images and the inpainted images. The experiments with different mask



**Fig. 5.** Comparison of qualitative results between the proposed method and existing approaches on the CelebA-HQ dataset. Our proposed method demonstrates more effective inpainting results in various aspects including front-facing, side-facing, hair, facial features, and more. And we illustrate the inpainted results with different masks just for diverse comparisons like other existing inpainting methods PEN [10], LBPLSA [5], LG [20].

ratios (0.1-0.2, 0.2-0.3, 0.3-0.4, 0.4-0.5) are conducted for both boundary-constrained and unconstrained cases.

Table 1 presents the quantitative results for the CelebA-HQ dataset. Our experimental results rank first in SSIM, PSNR, and L1 on irregular masks. On rectangular masks, although LG method is slightly superior, our framework is lighter and runs faster. One possible reason is that LG scheme achieves better inpainting results only in regions with fewer structural textures. In highly textured regions, it achieves high metrics by generating blurred content.

The quantitative results on Paris StreetView are presented in Table 2. It can be observed that our method outperforms others in terms of PSNR and  $L_1$ . The SSIM performance is similar with the LG scheme across each mask group.

**Table 1.** Comparison between the proposed method and state-of-the-art methods on the CelebA-HQ dataset (+ indicates higher is better, - indicates lower is better).

	mask	10-20%		20-30%		30-40%		40-50%		rect
	border	N	B	N	B	N	B	N	B	120*120
PSNR <sup>+</sup>	PEN	18.01	16.82	15.57	14.43	14.00	13.00	12.84	11.92	23.42
	LBPLSA	33.33	33.00	29.80	29.22	27.10	26.40	24.92	24.23	25.48
	LG	33.74	33.55	30.06	29.80	27.44	27.10	25.34	24.92	<b>27.59</b>
	Ours	<b>34.39</b>	<b>34.22</b>	<b>30.77</b>	<b>30.50</b>	<b>28.31</b>	<b>27.89</b>	<b>26.33</b>	<b>25.89</b>	27.28
SSIM <sup>+</sup>	PEN	0.7840	0.7766	0.6677	0.6611	0.5706	0.5636	0.4748	0.4697	0.8614
	LBPLSA	0.9622	0.9595	0.9286	0.9204	0.8885	0.8697	0.8408	0.8118	0.9005
	LG	0.9678	0.9651	0.9361	0.9307	0.8983	0.8874	0.8539	0.8364	<b>0.9176</b>
	Ours	<b>0.9694</b>	<b>0.9676</b>	<b>0.9409</b>	<b>0.9367</b>	<b>0.9081</b>	<b>0.8982</b>	<b>0.8697</b>	<b>0.8544</b>	0.9112
L1 <sup>-</sup>	PEN	0.1047	0.1238	0.1792	0.2094	0.2522	0.2915	0.3284	0.3739	0.0609
	LBPLSA	0.0140	0.0141	0.0266	0.0278	0.0416	0.0456	0.0605	0.0670	0.0438
	LG	0.0125	0.0129	0.0240	0.0249	0.0374	0.0399	0.0540	0.0583	<b>0.0340</b>
	Ours	<b>0.0117</b>	<b>0.0119</b>	<b>0.0222</b>	<b>0.0229</b>	<b>0.0342</b>	<b>0.0364</b>	<b>0.0488</b>	<b>0.0526</b>	0.0360

Table 3 presents the quantitative results for the Places2 dataset, demonstrate that our method outperforms several methods when dealing with large holes in complex images. In the PSNR and  $L_1$ , the LG method is slightly higher than our method only when the mask ratio of irregular masks is (0.1-0.2), while our method is the best in rectangular masks and irregular masks with mask ratios of (0.2-0.3, 0.3-0.4, 0.4-0.5). In the SSIM, our method performs best in rectangular masks, while the LG method slightly outperforms our method in irregular masks. One possible reason is that the LG network has a three-layer network architecture, resulting in longer training time and a larger number of parameters, especially by generating fuzzy content to achieve high SSIM metrics, as shown in Fig. 4.

**Table 2.** Comparison between the proposed method and state-of-the-art methods on the Paris StreetView dataset (+ indicates higher is better, - indicates lower is better, -- indicates that the corresponding metric is not provided in the RFR paper).

mask		10-20%		20-30%		30-40%		40-50%		rect
border		N	B	N	B	N	B	N	B	120*120
PSNR <sup>+</sup>	PEN	20.26	17.46	17.32	14.71	15.65	13.00	14.61	12.00	22.75
	RFR	31.71		--		26.44		--		--
	LBPLSA	33.33	32.70	29.31	29.25	26.57	26.76	24.84	24.74	25.11
	LG	32.77		29.38		27.01		25.15		<b>25.39</b>
	Ours	<b>34.09</b>	<b>33.26</b>	<b>30.07</b>	<b>29.79</b>	<b>27.40</b>	<b>27.41</b>	<b>25.77</b>	<b>25.37</b>	25.27
SSIM <sup>+</sup>	PEN	0.8210	0.7962	0.7096	0.6825	0.6209	0.5970	0.5249	0.5040	0.8234
	RFR	0.9540		--		0.8620		--		--
	LBPLSA	0.9582	0.9513	0.9121	0.9085	0.8600	0.8526	0.8028	0.7858	0.8466
	LG	<b>0.9710</b>		<b>0.9400</b>		<b>0.9000</b>		<b>0.8510</b>		<b>0.8555</b>
	Ours	0.9645	0.9565	0.9238	0.9176	0.8787	0.8669	0.8308	0.8050	0.8544
L1 <sup>-</sup>	PEN	0.0941	0.1455	0.1704	0.2619	0.2508	0.3699	0.3136	0.4745	0.0727
	RFR	0.0110		--		0.0275		--		--
	LBPLSA	0.0164	0.0176	0.0336	0.0327	0.0530	0.0509	0.0740	0.0745	0.0550
	LG	0.0159		0.0297		<b>0.0457</b>		0.0644		<b>0.0515</b>
	Ours	<b>0.0142</b>	<b>0.0160</b>	<b>0.0293</b>	<b>0.0296</b>	0.0146	0.0459	<b>0.0628</b>	0.0672	0.0525

**Table 3.** Comparison between the proposed method and state-of-the-art methods on the Places2 dataset (+ indicates higher is better, - indicates lower is better).

mask		10-20%		20-30%		30-40%		40-50%		rect
border		N	B	N	B	N	B	N	B	120*120
PSNR <sup>+</sup>	PEN	19.21	17.62	16.70	15.19	15.19	13.68	13.93	12.62	19.26
	LBPLSA	29.48	28.87	26.10	25.36	25.36	22.70	21.88	20.78	21.35
	LG	<b>30.29</b>	<b>29.92</b>	26.77	26.27	26.27	23.56	22.40	21.56	21.32
	Ours	30.13	29.62	<b>27.09</b>	<b>26.52</b>	<b>26.52</b>	<b>24.08</b>	<b>23.02</b>	<b>22.17</b>	<b>22.01</b>
	PEN	0.8070	0.7962	0.6962	0.6848	0.5997	0.5870	0.5040	0.4912	0.8064
SSIM <sup>+</sup>	LBPLSA	0.9426	0.9355	0.8903	0.8750	0.8273	0.8020	0.7569	0.7203	0.8209
	LG	<b>0.9530</b>	<b>0.9488</b>	<b>0.9085</b>	<b>0.8981</b>	<b>0.8548</b>	<b>0.8357</b>	<b>0.7950</b>	<b>0.7642</b>	0.8288
	Ours	0.9490	0.9446	0.9051	0.8945	0.8519	0.8325	0.7918	0.7617	<b>0.8290</b>
L1 <sup>-</sup>	PEN	0.0919	0.1246	0.1574	0.2133	0.2218	0.2987	0.2889	0.3856	0.1031
	LBPLSA	0.0229	0.0248	0.0427	0.0474	0.0664	0.0755	0.0942	0.1085	0.0804
	LG	<b>0.0198</b>	<b>0.0206</b>	0.0371	0.0398	0.0577	0.0637	0.0815	0.0920	0.0766
	Ours	0.0206	0.0214	<b>0.0371</b>	<b>0.0396</b>	<b>0.0571</b>	<b>0.0624</b>	<b>0.0806</b>	<b>0.0897</b>	<b>0.0740</b>
	PEN	0.0919	0.1246	0.1574	0.2133	0.2218	0.2987	0.2889	0.3856	0.1031

### 5.3 Qualitative Result Analysis

To provide a more intuitive comparison between our proposed method and competitive techniques, we display the results of various experiments in Fig. 3, Fig. 4, Fig. 5. In the case of CelebA-HQ (Fig. 5), we produce the clearer results in various aspects such as hair (e.g., the first row), nose (e.g., the second row), facial contour edges (e.g., the fourth row) and mouth (e.g., the sixth row) than others.

The same level of performance can also be observed on the Paris StreetView and Places2 datasets. For the Paris StreetView dataset (Fig. 3), RFR can restore the basic textures, but it exhibits structural distortions and lacks realism (e.g., the first row).

Compared with RFR, LBPLSA can produce structures that are more reasonable. However, it shows blurry artifacts for images with larger missing areas (e.g., the fourth row). LG reduces the artifacts and preserve the overall style of the image. When the images have highly textured structures, it does not achieve the fine detail inpainting (e.g., the second row). For the Places2 dataset (Fig. 4), our reconstructed scenes are closer to reality. In summary, our method is capable of generating reasonable structures and fine details for missing areas, reducing the occurrence of artifacts and distorted structures, enhancing realism and reinforcing the generation of spatially consistent structural information in the restored images. We attribute these improvements in structural and spatial generation to our proposed Double PatchGAN Discriminator. Likewise, we credit the generation and enhancement of fine edge details in the restored images and the improved visual realism to our proposed multi-fusion loss function.

#### 5.4 Ablation Analysis

Here, we conduct some experiments to investigate the effectiveness of DPD and MLF in our method. The settings for the three experiments are as follows:

- Experiment I (NO): without (w/o) DPD and w/o MLF.
- Experiment II (DPD): with (w/i) DPD and w/o MLF.
- Experiment III (DPD + MLF): w/i DPD and w/i MLF.

The quantitative results of Experiments I-III are displayed in Table 4, Table 5 and the corresponding qualitative results are shown in Fig. 6. The outcomes further demonstrate the effectiveness of each proposed module in our network.

**Table 4.** Quantitative results of ablation study on Paris StreetView dataset. (+ indicates higher is better, - indicates lower is better).

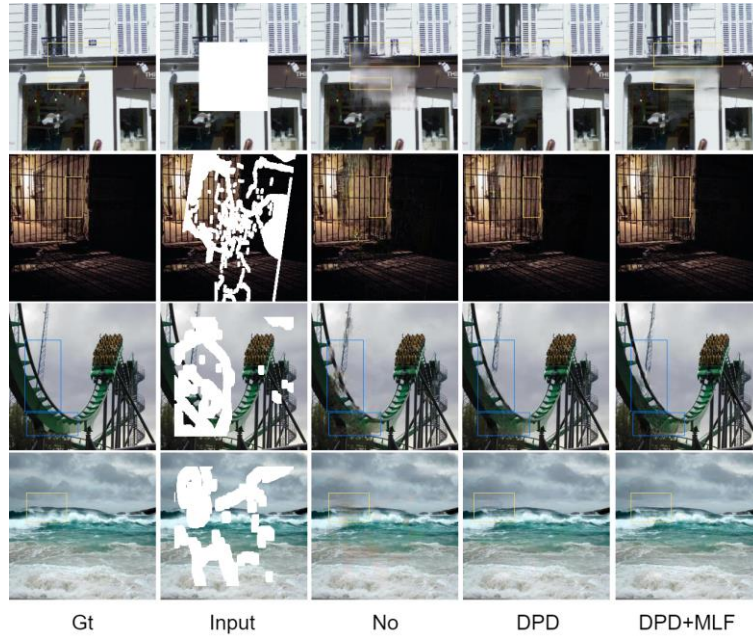
mask		10-20%		20-30%		30-40%		40-50%		rect
border		N	B	N	B	N	B	N	B	120
PSNR <sup>+</sup>	NO	33.33	32.70	29.31	29.25	26.57	26.76	24.84	24.74	25.11
	DPD	32.93	32.89	29.43	29.49	26.73	27.23	25.18	25.17	25.11
	DPD+MLF	<b>34.09</b>	<b>33.26</b>	<b>30.07</b>	<b>29.79</b>	<b>27.40</b>	<b>27.41</b>	<b>25.77</b>	<b>25.37</b>	<b>25.27</b>
SSIM <sup>+</sup>	NO	0.9582	0.9513	0.9121	0.9085	0.8600	0.8526	0.8028	0.7858	0.8466
	DPD	0.9567	0.9541	0.9134	0.9136	0.8630	0.8619	0.8100	0.7985	0.8530
	DPD+MLF	<b>0.9645</b>	<b>0.9565</b>	<b>0.9238</b>	<b>0.9176</b>	<b>0.8787</b>	<b>0.8669</b>	<b>0.8308</b>	<b>0.8050</b>	<b>0.8544</b>
L1 <sup>-</sup>	NO	0.0164	0.0176	0.0336	0.0327	0.0530	0.0509	0.0740	0.0745	0.0550
	DPD	0.0167	0.0168	0.0330	0.0310	0.0520	0.0475	0.0711	0.0698	0.0538
	DPD+MLF	<b>0.0142</b>	<b>0.0160</b>	<b>0.0293</b>	<b>0.0296</b>	<b>0.0461</b>	<b>0.0459</b>	<b>0.0628</b>	<b>0.0672</b>	<b>0.0525</b>

**Effect of Double-PatchGAN Discriminator** To demonstrate the effectiveness of our proposed DPD, the qualitative and quantitative results of adding and removing DPD modules in the network are used to visually demonstrate the its role on the Paris StreetView and Places2 datasets. From the Table 4, Table 5 and Fig. 6, we can clearly observe that the proposed DPD generate relatively clear structural textures and significantly visual effects in image inpainting.

**Table 5.** Quantitative results of ablation study on Places2 dataset. (+ indicates higher is better, - indicates lower is better).

	mask	10-20%		20-30%		30-40%		40-50%		rect
	border	N	B	N	B	N	B	N	B	120
PSNR <sup>+</sup>	NO	29.48	28.87	26.10	25.36	23.72	22.70	21.88	20.78	21.35
	DPD	29.92	29.40	26.79	26.26	24.68	23.86	22.93	22.00	21.14
	DPD+MLF	<b>30.13</b>	<b>29.62</b>	<b>27.09</b>	<b>26.52</b>	<b>24.86</b>	<b>24.08</b>	<b>23.02</b>	<b>22.17</b>	<b>22.01</b>
SSIM <sup>+</sup>	NO	0.9426	0.9355	0.8903	0.8750	0.8273	0.8020	0.7569	0.7203	0.8209
	DPD	0.9488	0.9420	0.9042	0.8878	0.8512	0.8234	0.7910	0.7527	0.8229
	DPD+MLF	<b>0.9490</b>	<b>0.9446</b>	<b>0.9051</b>	<b>0.8945</b>	<b>0.8519</b>	<b>0.8325</b>	<b>0.7918</b>	<b>0.7617</b>	<b>0.8290</b>
L1 <sup>-</sup>	NO	0.0229	0.0248	0.0427	0.0474	0.0664	0.0755	0.0942	0.1085	0.0804
	DPD	0.0211	0.0223	0.0380	0.0415	0.0577	0.0649	0.0809	0.0927	0.0800
	DPD+MLF	<b>0.0206</b>	<b>0.0214</b>	<b>0.0371</b>	<b>0.0396</b>	<b>0.0571</b>	<b>0.0624</b>	<b>0.0806</b>	<b>0.0897</b>	<b>0.0740</b>

**Effect of Multiple Loss Functions** In this study, we investigate the impact of incorporating MLF in training the proposed network for image inpainting. These analyses are carried out on the Paris StreetView and Places2 datasets. Following the addition of MLF based on Experiment II, a noteworthy improvement is observed across multiple indicators, as evident in Table 4, Table 5. From the qualitative results (Fig. 6), it can be seen that artifacts are reduced and the textures are clearer.

**Fig. 6.** Qualitative results comparison of ablation study.

## 6 Conclusion

We propose a two-stage image inpainting network, wherein the first stage adopts an enhanced LBP-guided image inpainting approach. In the second stage, the integration of DPD and a spatial attention module allows for more precise control over details during the inpainting process. This two-stage architecture not only ensures overall image consistency but also captures the subtle differences, details and textures of local regions. Moreover, MLF is designed to reconstruct information from different hierarchies throughout the inpainting process. Extensive research results provide evidence for the superiority of our proposed approach. However, with the increase of the incomplete area, the likelihood of artifacts or inaccuracies in semantics persists. In future, we intend to enhance the network by incorporating techniques, such as gated convolutions and transformers, or to replace LBP by introducing alternative guidance methods. Additionally, the paper will also try to apply model migration to oracle bone restoration work.

**Acknowledgments.** This study was funded by the scientific and technological project in Henan Province in 2022 (Grant No. 222102210187), the Key Research Project for Higher Education Institutions in Henan Province (Grant No. 24A520018).

**Disclosure of Interests.** The authors have no competing interests to declare that are relevant to the content of this article.

## References

1. Zhang, H., Hu, Z., Luo, C., Zuo, W., Wang, M.: Semantic image inpainting with progressive generative networks. In: Proceedings of the 26th ACM international conference on Multimedia, pp. 1939-1947.
2. Guo, Z., Chen, Z., Yu, T., Chen, J., Liu, S.: Progressive image inpainting with full-resolution residual network. In: Proceedings of the 27th acm international conference on multimedia, pp. 2496-2504.
3. Nazeri, K., Ng, E., Joseph, T., Qureshi, F.Z., Ebrahimi, M.: Edgeconnect: Generative image inpainting with adversarial edge learning. arXiv preprint arXiv:1901.00212 (2019)
4. Yi, Z., Tang, Q., Azizi, S., Jang, D., Xu, Z.: Contextual residual aggregation for ultra high-resolution image inpainting. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp. 7508-7517.
5. Wu, H., Zhou, J., Li, Y.: Deep generative model for image inpainting with local binary pattern learning and spatial attention. *IEEE Transactions on Multimedia* 24, 4016-4027 (2021)
6. Waller, B.M., Nixon, M.S., Carter, J.N.: Image reconstruction from local binary patterns. In: 2013 International Conference on Signal-Image Technology & Internet-Based Systems, pp. 118-123. IEEE.
7. Pathak, D., Krahenbuhl, P., Donahue, J., Darrell, T., Efros, A.A.: Context encoders: Feature learning by inpainting. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 2536-2544.
8. Liu, G., Reda, F.A., Shih, K.J., Wang, T.-C., Tao, A., Catanzaro, B.: Image inpainting for irregular holes using partial convolutions. In: Proceedings of the European conference on computer vision (ECCV), pp. 85-100.

9. Yu, J., Lin, Z., Yang, J., Shen, X., Lu, X., Huang, T.S.: Free-form image inpainting with gated convolution. In: Proceedings of the IEEE/CVF international conference on computer vision, pp. 4471-4480.
10. Zeng, Y., Fu, J., Chao, H., Guo, B.: Learning pyramid-context encoder network for high-quality image inpainting. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 1486-1494.
11. Wang, N., Ma, S., Li, J., Zhang, Y., Zhang, L.: Multistage attention network for image inpainting. *Pattern Recognition* 106, 107448 (2020)
12. Li, J., Wang, N., Zhang, L., Du, B., Tao, D.: Recurrent feature reasoning for image inpainting. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 7760-7768.
13. Ojala, T., Pietikäinen, M., Harwood, D.: A comparative study of texture measures with classification based on featured distributions. *Pattern recognition* 29, 51-59 (1996)
14. Ronneberger, O., Fischer, P., Brox, T.: U-net: Convolutional networks for biomedical image segmentation. In: Medical Image Computing and Computer-Assisted Intervention—MICCAI 2015: 18th International Conference, Munich, Germany, October 5-9, 2015, Proceedings, Part III 18, pp. 234-241. Springer.
15. Xu, B., Wang, N., Chen, T., Li, M.: Empirical evaluation of rectified activations in convolutional network. *arXiv preprint arXiv:1505.00853* (2015)
16. Ulyanov, D., Vedaldi, A., Lempitsky, V.: Instance normalization: The missing ingredient for fast stylization. *arXiv preprint arXiv:1607.08022* (2016)
17. Nair, V., Hinton, G.E.: Rectified linear units improve restricted boltzmann machines. In: Proceedings of the 27th international conference on machine learning (ICML-10), pp. 807-814.
18. Isola, P., Zhu, J.-Y., Zhou, T., Efros, A.A.: Image-to-image translation with conditional adversarial networks. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 1125-1134.
19. Yan, Z., Li, X., Li, M., Zuo, W., Shan, S.: Shift-net: Image inpainting via deep feature rearrangement. In: Proceedings of the European conference on computer vision (ECCV), pp. 1-17.
20. Quan, W., Zhang, R., Zhang, Y., Li, Z., Wang, J., Yan, D.-M.: Image inpainting with local and global refinement. *IEEE Transactions on Image Processing* 31, 2405-2420 (2022)
21. Miyato, T., Kataoka, T., Koyama, M., Yoshida, Y.: Spectral normalization for generative adversarial networks. *arXiv preprint arXiv:1802.05957* (2018)
22. Johnson, J., Alahi, A., Fei-Fei, L.: Perceptual losses for real-time style transfer and super-resolution. In: Computer Vision—ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part II 14, pp. 694-711. Springer.
23. 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.
24. Buades, A., Coll, B., Morel, J.-M.: A non-local algorithm for image denoising. In: 2005 IEEE computer society conference on computer vision and pattern recognition (CVPR'05), pp. 60-65. Ieee.
25. Doersch, C., Singh, S., Gupta, A., Sivic, J., Efros, A.: What makes paris look like paris? *ACM Transactions on Graphics* 31, (2012)
26. Karras, T., Aila, T., Laine, S., Lehtinen, J.: Progressive growing of gans for improved quality, stability, and variation. *arXiv preprint arXiv:1710.10196* (2017)



27. Zhou, B., Lapedriza, A., Khosla, A., Oliva, A., Torralba, A.: Places: A 10 million image database for scene recognition. *IEEE transactions on pattern analysis and machine intelligence* 40, 1452-1464 (2017)