MR-InpaintNet: Toward Deep Multi-Resolution Learning for Progressive Image Inpainting

Huan Zheng, Zhao Zhang, *Senior Member, IEEE*, Haijun Zhang, *Senior Member, IEEE*, Yi Yang, *Senior Member, IEEE*, Shuicheng Yan, *Fellow, IEEE*, and Meng Wang, *Fellow, IEEE*

Abstract—Deep learning based image inpainting methods have improved the performance greatly due to powerful representation ability of deep learning. However, current deep inpainting methods still tend to produce unreasonable structure and blurry texture, implying that image inpainting is still a challenging topic due to the ill-posed property of the task. To address these issues, we propose a novel deep multi-resolution learning-based progressive image inpainting method, termed MR-InpaintNet, which takes the damaged images of different resolutions as input and then fuses the multi-resolution features for repairing the damaged images. The idea is motivated by the fact that images of different resolutions can provide different levels of feature information. Specifically, the low-resolution image provides strong semantic information and the high-resolution image offers detailed texture information. The middle-resolution image can be used to reduce the gap between low-resolution and high-resolution images, which can further refine the inpainting result. To fuse and improve the multi-resolution features, a novel multi-resolution feature learning (MRFL) process is designed, which is consisted of a multi-resolution feature fusion (MRFF) module, an adaptive feature enhancement (AFE) module and a memory enhanced mechanism (MEM) module for information preservation. Then, the refined multi-resolution features contain both rich semantic information and detailed texture information from multiple resolutions. We further handle the refined multi-resolution features by the decoder to obtain the recovered image. Extensive experiments on the Paris Street View, Places2 and CelebA-HQ datasets demonstrate that our proposed MR-InpaintNet can effectively recover the textures and structures, and performs favorably against state-of-the-art methods.

Index Terms—Image inpainting, deep multi-resolution learning, multi-resolution neural network, multi-resolution feature fusion and enhancement.

I. INTRODUCTION

IMAGE inpainting is a task of generating visually plausible contents to fill in the missing regions in a damaged image or repair the deteriorated images. Due to the broad practical application areas, e.g., human face editing [3], privacy protection [4], photo disocclusion [5] and object removal [6],

H. Zheng, Z. Zhang, and M. Wang are now with the School of Computer Science and Information Engineering; also with the Key Laboratory of Knowledge Engineering with Big Data (Ministry of Education); also with the Intelligent Interconnected Systems Laboratory of Anhui Province, Hefei University of Technology, Hefei 230601, China. E-mail: huanzheng1998@gmail.com, cszzhang@gmail.com, eric.mengwang@gmail.com.

H. Zhang is with the Department of Computer Science, Harbin Institute of Technology (Shenzhen), Xili University Town, Shenzhen, China. E-mail: hjzhang@hit.edu.cn.

Y. Yang is with the Centre for Artificial Intelligence, University of Technology Sydney, Sydney, NSW, Australia. E-mail: Yi.Yang@uts.edu.au.

S. Yan is with Sea AI Lab (SAIL), Singapore; also with the National University of Singapore, Singapore 117583. E-mail: shuicheng.yan@gmail.com.

Fig. 1. Comparison of image inpainting results, where (a) denotes the input image, (b) denotes the recovered result of PC [1], (c) denotes the inpainting result of FE [2] and (d) denotes the restored image of our MR-InpaintNet.

image inpainting has been emerging as an important topic in the areas of image processing and computer vision. However, image inpainting task is still facing great challenges, since the restored global structures and local textures in the synthesized image still cannot keep consistent with the known region.

For a meaningful image, it must contain redundancy and non-local similarities. As such, traditional inpainting methods assume that the contents of the missing region have similar information to the known region [7]–[13]. By seeking the similarity between different patches of an image, one can choose the most similar patches to fill in the missing region. Note that although coping the known patches can well produce realistic local textures, it cannot capture high-level features and understand global contextual information. These issues will potentially result in inaccurate and even wrong structures in the repaired image. Besides, the above assumption is unreasonable in some cases, since there is no guarantee that all missing contents can be found similarly in the known area. As a result, simply coping the known patches will prevent the method from generating new contents, so that the image inpainting results will be awful for the cases of large masked region.

With the rapid development of deep neural networks, there is growing attention to designing advanced deep models for

Fig. 2. The framework of MR-InpaintNet. The generator has three major components: multi-resolution encoder (MR-Encoder), multi-resolution feature learning (MRFL) and decoder. Discriminator network is used to enhance the inpainting results, and a pretrained and fixed VGG-Net is used for feature matching.

different computer vision tasks [14]–[17], such as deep image inpainting methods that have obtained impressive results compared to the traditional methods. Most existing deep inpainting models employ the pipeline of encoder and decoder. To be specific, given a corrupted image and a corresponding mask, deep inpainting methods map the input to a feature space, and decode the features to obtain a repaired image. Certain carefully-designed modules and losses have also been used to enhance the global semantic information flow and capture local textural features. Attributed to these new modules, deep inpainting methods have been reported to be able to generate visually plausible inpainting results. However, for a high-resolution image, convolution neural networks (CNNs)-based deep models may be ineffective to build long-term relationship between the distant regions. They progressively dilate the receptive field, which makes the network learn less contextual information. As a result, the repaired image may contain artifacts, inaccurate structures and blurry textures, which are inconsistent with the known region, as shown in Figure 1.

In this paper, we therefore present a novel deep inpainting network via deep multi-resolution learning, as illustrated in Figure 2, so that both the global structure information and detailed texture information can be discovered jointly. The main contributions of this paper are summarized as follows:

- Technically, a new and effective deep multi-resolution learning-based progressive image inpainting network, termed MR-InpaintNet, is proposed. Different from current deep inpainting models, MR-InpaintNet handles the task based on multiple damaged images and multiple masks from different resolutions, so that different levels of feature information can be learnt to enhance the performance. To be specific, high-resolution images support exhaustive textural information to recover the texture details, and low-resolution images provide rich semantic

information for better understanding the global context. As for middle-resolution images, it can be applied to alleviate the gap between high-resolution and low-resolution images. To the best of our knowledge, this work is one of few existing deep inpainting methods to jointly recover the structure and texture information of the missing regions based on multiple resolution images.

- To build the relationship between distant regions in high-resolution image and obtain strong feature representation containing rich textural and semantic information, a novel module called multi-resolution feature learning (MRFL) is designed to aggregate the information from different resolutions, which can jointly fuse and enhance the multi-resolution features for image inpainting.
- For deep neural networks, some useful information from shadow layers are usually degraded with the deepening of the neural networks. Therefore, we also propose a memory enhanced mechanism (MEM) to solve this issue, which can prevent our inpainting model from information degradation, with the increase of layers. The features from those shadow layers can also be reused in the deeper layers for information preservation.
- Extensive experiments on Paris Street View, Places2 and CelebA-HQ datasets show that our inpainting method can generate more reasonable structures and realistic textures, than current state-of-the-art deep inpainting methods.

This paper is outlined as follows. A brief review of the related inpainting methods is shown in Section 2. In Section 3, we propose MR-InpaintNet. The experiments and analysis on several real-world datasets are shown in Section 4. Finally, conclusion with some future works is offered in Section 5.

II. RELATED WORK

In this section, we briefly introduce some image inpainting algorithms, which are related to our proposed method.

A. Traditional Image Inpainting

For an image to be repaired, traditional methods try to find the similar components in the background to fill in the missing regions. The basic assumption of these traditional methods is that there are always redundant information and non-local similarities in the image. Roughly, traditional methods include two types, i.e., patch-based methods [7]–[10] and diffusion-based methods [12], [13].

Patch-based methods. Patch-based methods simply copy the known patches instead of the missing patches, because the prerequisite is that the information of the missing parts can be found in the known area. As a result, these methods cannot generate new contents to fill in the missing region. PatchMatch [7], as one of the most classic patch-based methods, is an interactive image inpainting and editing method by driving a novel randomized algorithm to find the nearest neighbor matches between different patches. Ghorai et al. [10] also presented a multi-pyramid-based inpainting method to preserve the texture consistency and structure continuity in image based on the local patch statistics and sparse representation.

Diffusion-based methods. A robust inpainting method was proposed in [13], which uses the fractional-order nonlinear diffusion with difference curvature, and obtains a good balance between edge preservation and elimination of image staircase and speckle artifacts. For non-textured-based inpainting, total variation is a classic method. Li et al. [12] proposed an improved total variation method that computes the diffusion coefficient based on the damaged pixel and its neighbors.

Remarks. Traditional methods may produce realistic textures in some cases, since they simply copy the information from the known regions. However, they cannot capture high-level features, extract contextual information, generate new contents and retain the semantic consistency, especially for large masks, so these methods may cause inconsistent structures in the recovered images.

B. Deep Learning-based Image Inpainting

Deep inpainting methods designs novel encoder-decoder neural networks by feature learning for image inpainting. They encode the damaged images to obtain feature representations and then decode the feature maps to receive the repaired images. Deep image inpainting methods can be roughly divided into single-solution inpainting methods [1], [2], [18]–[38] and multi-solution inpainting methods [39], [40].

Single-solution inpainting methods. Most of current image inpainting methods belong to this category, which obtains one output image during repairing a masked image. Context encoders (CE) [18] is the first deep image inpainting method that takes an encoder-decoder pipeline as the main framework. For a damaged image, the useless information in the damaged region may greatly affect the recovered result. To this end, a new partial convolution (PC) [1] was proposed for image inpainting. Yu et al. [41] presented a region normalization (RN)

Fig. 3. The motivation of the proposed framework, where we show the process of building the relationship between distant regions in high-resolution images by multi-resolution learning. The red region denotes a patch of the image.

to replace the batch normalization in deep neural networks according to the masked region. A mutual encoder-decoder network [2] was also proposed for image inpainting, which contains a carefully-designed feature equalization (FE) module to help the decoder obtain better results.

Multi-solution inpainting methods. For image inpainting, we only need to obtain an output image with reasonable structures and textures, so there is no definite solution. Therefore, researchers also tried to obtain multiple solutions, such as PIC [39] and PD-GAN [2]. To be specific, PIC can synthesize multiple image inpainting results by using a probabilistically principled framework, which is the first method to explore diverse inpainting. PD-GAN proposed a novel normalization method to dynamically balance the realism and diversity inside the hole region. A perceptual diversity loss was also presented for further promoting diverse content generation.

Remarks. Compared with those traditional methods, deep inpainting methods can obtain high-level features to learn better structures. However, current deep inpainting methods still suffer from unreasonable structures, blurry textures and artifacts. As such, how to restore the textures and structures of the missing area better, and ensure the consistency with the known area for deep inpainting is still a difficult issue.

III. MR-INPAINTNET: DEEP MULTI-RESOLUTION LEARNING FOR IMAGE INPAINTING

In this section, we introduce the architecture and modules of MR-InpaintNet. The main motivation and network architecture are firstly introduced. Based on the deep multi-resolution learning, a novel multi-resolution feature learning (MRFL) process is proposed, which aggregates the information of multiple resolutions and can adaptively enhance the multi-resolution features. Finally, loss functions are detailed.

A. Motivation

For the task of deep image inpainting, similar textural and structural information exist in the images, which can be seen in Figure 3. Therefore, we can apply the information of the

Fig. 4. The detailed structure of the discriminator, which contain two branches, where the local branch is used for local feature extraction and the other one is used for global feature extraction. The inputs of the discriminator are composed of an input image and a corresponding cropped patch.

known region to repair the unknown region. That is, we need to establish the connections among similar regions in the images. However, the distance between those similar patches may be far away in images, making the deep CNNs inept or ineffective in connecting the distant regions, especially for those high-resolution image. To address this issue, we propose to map the distant areas in high-resolution image to a new data space to make the distance between similar regions smaller. A simple idea is to down-sample the high-resolution image to obtain the corresponding lower-resolution images. Then, we can use the smaller patches in lower-resolution images to represent the corresponding patches in high-resolution images. As shown in Figure 3, we illustrate the detailed process of building the relationship between distant patches in high-resolution image via multi-resolution learning. For lower-resolution image, CNN can quickly expand the receptive field to the entire picture as the number of network layers increases. As a result, CNN can potentially model the relationship between distant regions easily for lower-resolution images. According to the relationship between high-resolution and lower-resolution images, the modeling of remote regions in high-resolution images can be indirectly built.

B. Network Architecture

The framework of MR-InpaintNet is shown in Figure 2. As can be seen, MR-InpaintNet has three subnetworks, i.e., a generator to produce the inpainting results, a discriminator network to conduct adversarial training with generator, and a pretrained VGG-Net to compute the relative loss functions. Only the generator network is used in test phase.

The input of MR-InpaintNet includes three kinds of images from different resolutions, i.e., I_{in} , I'_{in} and I''_{in} , and three corresponding masks M , M' and M'' , where I_{in} and M denote the high-resolution input image and corresponding mask, I'_{in} and M' denote the middle-resolution input image and corresponding mask, I''_{in} and M'' denote the low-resolution image and corresponding mask. Given a ground-truth image

Fig. 5. The detailed process of MRFL that includes two main components: multi-resolution feature fusion (MRFF) and adaptive feature enhancement (AFE). The original input multi-resolution features are refined by cascading multiple MRFFs and AFEs. Besides, the memory enhanced mechanism (MEM) are embedded for feature preservation.

I_{gt} and a corresponding mask M (0 for known region and 1 for unknown region), the input image I_{in} can be defined as

$$I_{in} = I_{gt} \odot (1 - M), \quad (1)$$

where $\odot$ represents the element-wise multiplication. As such, the low-resolution and middle-resolution images I'_{in} and I''_{in} can be obtained by down-sampling I_{gt} and M .

The generator uses the classic U-Net [42] as basic model, which contains three main components, i.e., a multi-resolution encoder (MR-Encoder), a multi-resolution feature learning (MRFL) module and a decoder. Specifically, the MR-Encoder adopts multiple encoders to handle the multi-resolution inputs. As a result, the low-resolution inputs can provide strong semantic information to promote learning semantic information by MR-InpaintNet. The high-resolution inputs are encoded as features to supply better texture information that will be beneficial for the consistency in details. The middle-resolution inputs can be used to balance and reduce the gap between semantic information and texture information in the feature space to obtain more realistic repaired results. The structure of the discriminator is shown in Figure 4, which can distinguish the real pictures and synthesized pictures. The input of the discriminator has two elements, i.e., an image and a random patch of this image, where the entire image is adopted to keep the global consistency and the random patch is utilized to keep the local consistency. As such, the discriminator will contain two branches. The first one is for handling the entire image and the other one is for dealing with the patch. Further, the extracted features from these two branches will be concatenated for distinguishing the real and generated pictures.

C. Multi-Resolution Feature Learning (MRFL)

Given the encoded features from multiple resolutions, it is important to merge, refine and enhance the multi-resolution features to obtain stronger feature representations. We therefore propose MRFL to perform better multi-resolution representation learning. The detailed structure of MRFL is given in Figure 5, which contains two main units, i.e., a multi-resolution feature fusion (MRFF) module and an adaptive

Fig. 6. The detailed illustration of MRFF, where the input multi-resolution features are fused. There are two primary operations, i.e., channel fusion and spatial fusion.

feature enhancement (AFE) module. Specifically, MRFF fuses the multi-resolution features in both spatial and channel dimensions to construct the fused features, and AFE can adaptively embed the fused features into the multi-resolution features so that the multi-resolution features can be refined. By cascading these two modules, MRFL will be able to merge, refine and enhance the multi-resolution features progressively. In addition, we also present a novel memory enhanced mechanism (MEM) module in MRFL for information preservation.

Multi-resolution feature fusion (MRFF). The structure of MRFF is shown in Figure 6. Firstly, MRFF tries to fuse the multi-resolution features from two aspects, i.e., channel fusion and spatial fusion. For the inputted multi-resolution features F_i , F'_i and F''_i in the i -th stage, we use a global average pooling method to make the input tensors become vectors respectively, which is described as follows:

$$(v_i, v'_i, v''_i) = \frac{1}{H \times W} \sum_{j=1}^H \sum_{k=1}^W (F_i(j, k), F'_i(j, k), F''_i(j, k)), \quad (2)$$

where H and W denote the height and width of input multi-resolution features F_i , v_i , v'_i and v''_i are the corresponding vectors in the i -th stage. For enhancing the information flow, v_i , v'_i and v''_i are processed by a fully-connected layer respectively. As a result, information in the same resolution can be fully exchanged. Furthermore, to fuse these vectors from fully connected layer (FC) and aggregate richer information from different resolutions, we sum these vectors directly as follows:

$$C_i^{fuse} = fc(v_i) + fc(v'_i) + fc(v''_i), \quad (3)$$

where C_i^{fuse} denotes the result of channel fusion in the i -th stage and $fc(\cdot)$ denotes the operation of FC.

For the spatial fusion, the input multi-resolution features F_i , F'_i and F''_i in the i -th stage are firstly handled by a convolution layer and a residual learning process to obtain the intermediate

features D_i , D'_i and D''_i in the i -th stage as follows:

$$\begin{cases} D_i = conv(F_i) + F_i, \\ D'_i = conv(F'_i) + F'_i, \\ D''_i = conv(F''_i) + F''_i, \end{cases} \quad (4)$$

where $conv(\cdot)$ represents the convolution layer to promote the feature transformation. The residual learning process can reuse the previous information and prevent the network from gradient vanishing. Then, D_i , D'_i and D''_i are concatenated so that the information from different resolution images can be merged. Since the concatenation operation increases the number of channels, we use a convolution layer to reduce the channels and then fuse them, which are performed as follows:

$$S_i^{fuse} = conv([D_i, D'_i, D''_i]), \quad (5)$$

where S_i^{fuse} denotes the spatial fusion result in the i -th stage.

To further fuse the features from multiple resolutions, we merge the results of both the channel fusion and spatial fusion. Specifically, we perform the element-wise multiplication between C_i^{fuse} and S_i^{fuse} as follows:

$$F_i^{fuse} = C_i^{fuse} \odot S_i^{fuse}, \quad (6)$$

where F_i^{fuse} indicates the final fused features of MRFF in the i -th stage and $\odot$ denotes element-wise multiplication. C_i^{fuse} aggregates the information from multiple resolutions in channel dimension, while S_i^{fuse} can mix the information from multiple resolutions in spatial dimension. As a result, F_i^{fuse} will be able to merge the information from multiple resolutions in both channel and spatial dimensions, which means that it can be a better representation of context.

Memory enhanced mechanism (MEM). It is known that when the number of neural network layers increases, some useful information in shallow layers, which are helpful for the model to restore the damaged images, may be lost. MRFL may also suffer from this issue, since it cascades the MRFFs and AFEs. We therefore design a new memory enhanced mechanism (MEM) module to preserve the previous information and further enhance the representation ability. Given the fused feature maps F_i^{fuse} ($i = 1, 2, \dots, N$) in the i -th stage of MRFL, the process of MEM can be described as follows:

$$F_i^{mem} = \sum_{j=1}^i F_j^{fuse}, \quad (7)$$

where F_i^{mem} denotes the summation of all previous fused feature maps in i -th stage. Thus, the features from the shallow layers will be retained. Simultaneously, the information from different layers can be aggregated together, which will be helpful to address the long-term dependency problem.

Adaptive feature enhancement (AFE). Although the fused features by MRFF contain rich information from different resolutions, some important information from the original resolution may be dropped during the process of feature fusion. Toward this issue, we would like to take the features of different resolutions as the main component and then use the fused features to enhance them. To be specific, a new module AFE is designed to enhance the features adaptively. The feature enhancement strategy of AFE is to compute the

Fig. 7. The whole process of AFE. The fused features are used to adaptively enhance the multi-resolution features.

weighted summation of fused features and the input multi-resolution features. The structure of AFE can be seen in Figure 7. Since the input multi-resolution features are the primary part, we can simply set the corresponding weights to ones. As for the weights of the fused features, ResBlock and sigmoid function are used to learn the weights of the fused features adaptively. Finally, the weights θ , θ' and θ'' in the i -th stage can be obtained as follows:

$$\begin{cases} \theta_i = \text{sigmoid}(RB(F_i)), \\ \theta'_i = \text{sigmoid}(RB(F'_i)), \\ \theta''_i = \text{sigmoid}(RB(F''_i)), \end{cases} \quad (8)$$

where $RB(\cdot)$ represents the ResBlock and $\text{sigmoid}(\cdot)$ denotes the sigmoid function. Specifically, ResBlock can map the input features into the parameter space, and the sigmoid function can further constrain the values to (0, 1). Note that the values less than 1 ensures the importance of multi-resolution features. Further, the refined multi-resolution features can be performed by the following weighted summation operation:

$$\begin{cases} F_{i+1} = F_i^{men} \odot \theta_i + F_i, \\ F'_{i+1} = F_i^{men} \odot \theta'_i + F'_i, \\ F''_{i+1} = F_i^{men} \odot \theta''_i + F''_i, \end{cases} \quad (9)$$

where F_{i+1} , F'_{i+1} and F''_{i+1} denote the refined features, which are set as the input of the $(i+1)$ -th stage. As a result, the final multi-resolution features will contain the information from different resolutions and different layers, which can enable MR-InpaintNet to better understand the semantics and texture details, and further obtain more realistic inpainting results.

D. Loss Functions

Given an input image I_{in} and a mask M , the predicted image by our MR-InpaintNet is represented as I_{pre} . The final output image I_{out} is the merged result of the predicted image I_{pre} and the input image I_{in} , because there are known regions in the input image. This process can be performed as follows:

$$I_{out} = M \odot I_{pre} + I_{in}, \quad (10)$$

To better retain the consistency in both texture and structure, several loss functions are involved for training, i.e., reconstruction loss, perceptual loss, style loss and adversarial loss. Next, we will briefly introduce them one by one.

Reconstruction loss. To measure the difference in pixel level, we use two reconstruction losses from two aspects including the known region and unknown region:

$$L_{known} = \frac{1}{N} \|(1 - M) \odot (I_{pre} - I_{gt})\|_1, \quad (11)$$

$$L_{unknown} = \frac{1}{N} \|M \odot (I_{pre} - I_{gt})\|_1, \quad (12)$$

where N denotes the total number of elements in I_{pre} . The reconstruction loss can guide our inpainting model to restore the damaged images pixel by pixel.

Perceptual loss. We then introduce the VGG-16 based perceptual loss for capturing the high-level semantic information and simulating human perception of images quality:

$$L_{perceptual} = \sum_{i=1}^N \frac{1}{H_i W_i C_i} \|\varphi_i^{gt} - \varphi_i^{out}\|_1, \quad (13)$$

where C_i , H_i and W_i denote the channel size, height and weight of i -th feature map in a pretrained and fixed VGG-16. φ_i^{out} and φ_i^{gt} denote the corresponding feature maps of the i -th pooling layer in the pretrained and fixed VGG-16, respectively. The perceptual loss can instruct our MR-InpaintNet to learn semantic information, and further keep the semantic coherence.

Style loss. Similar with the perceptual loss, VGG-16 based style loss is also adopted for further ensuring the semantic consistency, which is defined as follows:

$$L_{style} = \sum_{i=1}^N \frac{1}{C_i \times C_i} \left\| \frac{1}{H_i W_i C_i} (\varphi_i^{gt} \varphi_i^{gtT} - \varphi_i^{out} \varphi_i^{outT}) \right\|_1, \quad (14)$$

Note that the style loss employs the Gram matrix of the selected feature maps to measure the difference between the recovered image and ground-truth. The style loss can therefore enable the content consistency of the synthesized image.

Adversarial loss. The adversarial loss is used to alternatively train the generator and discriminator:

$$L_{adv} = -\mathbb{E}_{m \sim p_{data}} \log(D(m)) - \mathbb{E}_{z \sim p_z} (1 - \log(D(G(z)))) , \quad (15)$$

where G and D denote the generator and discriminator, respectively. This loss can make the restored image close to the real one, i.e., making the generated image look more realistic.

Total loss. Based on the above several losses, the objective function of our MR-InpaintNet can be described as follows:

$$L = \lambda_{known} L_{known} + \lambda_{unknown} L_{unknown} + \lambda_{style} L_{style} + \lambda_{perceptual} L_{perceptual} + \lambda_{adv} L_{adv}, \quad (16)$$

where λ_{known} , $\lambda_{unknown}$, λ_{style} , $\lambda_{perceptual}$ and λ_{adv} are the tradeoff parameters. The reconstruction losses L_{known} and $L_{unknown}$ can reduce the difference between the recovered image and the ground-truth image in Euclidean space. The style loss and perceptual loss measures the difference between the recovered image and the ground-truth image in semantic space. The adversarial loss can make the generated image closer to the real image from both local and global perspectives.

Fig. 8. Visual comparison of the inpainting results of PC [1], PIC [39], RN [41] and FE [2] and our MR-InpaintNet on Places2 dataset.

TABLE I
NUMERICAL COMPARISON OF PC [1], PIC [39], RN [41] AND FE [2] AND OUR MR-INPAINTNET ON THE BUTTE, SNOWFIELD, CANYON CAN FARM SCENES OF PLACES2 DATASET, WHERE THE BOLD DENOTES THE BEST ONE.

Scen		Canyon					Snowfield					Butte					Farm				
Metric	Mask	PIC	PC	FE	RN	Ours	PIC	PC	FE	RN	Ours	PIC	PC	FE	RN	Ours	PIC	PC	FE	RN	Ours
$l_1(\%)$	0 – 10%	2.15	0.41	1.62	0.49	0.45	0.97	0.37	0.50	0.39	0.35	0.82	0.38	0.54	0.44	0.39	0.94	0.41	0.56	0.44	0.40
	10 – 20%	3.49	1.10	2.64	1.15	1.11	1.76	1.00	1.22	0.93	0.88	1.62	0.98	1.29	0.97	0.94	2.04	1.11	1.43	1.04	1.03
	20 – 30%	4.76	2.02	3.81	1.98	1.94	2.84	1.97	2.03	1.60	1.52	2.71	1.79	2.25	1.59	1.60	3.37	2.00	2.71	1.80	1.79
	30 – 40%	6.30	3.02	5.28	2.87	2.81	4.03	2.82	2.88	2.32	2.22	3.77	2.62	3.44	2.25	2.27	4.54	2.86	3.95	2.55	2.52
	40 – 50%	8.17	3.55	6.96	3.87	3.79	5.18	3.87	4.11	3.20	3.02	5.31	3.57	4.60	3.08	3.04	6.36	3.99	5.70	3.59	3.52
	50 – 60%	10.73	6.01	10.27	5.59	5.25	7.33	5.55	6.18	5.11	4.27	7.64	5.23	7.08	4.67	4.25	8.59	5.67	8.61	5.29	4.82
PSNR	0 – 10%	28.47	33.78	33.21	32.58	33.05	30.27	35.13	35.24	34.75	35.38	32.67	34.66	35.19	33.44	34.26	30.82	33.55	33.74	33.34	33.68
	10 – 20%	23.90	28.17	27.17	28.03	28.19	25.73	29.19	28.74	29.45	29.88	27.73	29.19	29.30	29.16	29.43	25.69	28.21	27.64	28.49	28.54
	20 – 30%	21.85	25.03	23.77	25.22	25.38	22.93	25.91	25.29	26.38	26.72	24.77	26.16	26.15	26.73	26.82	22.76	25.05	24.04	25.54	25.63
	30 – 40%	19.88	22.85	21.48	23.22	23.36	20.76	23.62	23.02	24.34	24.69	22.77	24.06	23.60	24.79	24.78	20.88	23.28	21.70	23.84	23.92
	40 – 50%	18.18	21.26	19.58	21.68	21.84	19.30	22.08	21.21	22.69	23.08	20.81	22.32	21.71	22.96	23.12	18.96	21.55	19.66	22.07	22.21
	50 – 60%	16.54	19.03	16.95	19.42	19.87	17.11	19.95	18.54	19.77	21.04	18.47	19.95	19.04	20.39	21.03	17.02	19.40	17.03	19.62	20.29
SSIM	0 – 10%	0.931	0.948	0.957	0.956	0.960	0.947	0.939	0.971	0.965	0.971	0.952	0.944	0.970	0.958	0.963	0.945	0.942	0.964	0.959	0.961
	10 – 20%	0.855	0.871	0.900	0.902	0.903	0.882	0.856	0.920	0.919	0.926	0.887	0.864	0.915	0.908	0.912	0.869	0.860	0.903	0.904	0.905
	20 – 30%	0.772	0.778	0.817	0.832	0.833	0.813	0.761	0.857	0.869	0.878	0.810	0.768	0.843	0.847	0.850	0.781	0.764	0.825	0.837	0.838
	30 – 40%	0.692	0.692	0.737	0.764	0.766	0.745	0.677	0.793	0.820	0.829	0.737	0.683	0.769	0.793	0.795	0.700	0.677	0.746	0.775	0.776
	40 – 50%	0.596	0.594	0.641	0.680	0.685	0.660	0.580	0.715	0.757	0.770	0.645	0.587	0.685	0.718	0.726	0.597	0.578	0.650	0.691	0.695
	50 – 60%	0.501	0.491	0.529	0.580	0.598	0.562	0.487	0.626	0.678	0.706	0.549	0.490	0.583	0.630	0.656	0.499	0.480	0.536	0.598	0.615

IV. EXPERIMENTS

In this section, we conduct extensive experiments to evaluate the effectiveness of our method, along with illustrating the comparison results with other related methods. We perform all experiments with the Pytorch framework [43] on a NVIDIA GeForce RTX 2080i GPU. All images in both training and testing are resized with 256×256 pixels for each method for fair comparison. We adopt the Adam optimizer [44] with initial learning of 0.0001 to train our model. Additionally, the learning rate will degrade 10% per epoch before it is

smaller than 0.00001. For the tradeoff parameters in the objective function of our MR-InpaintNet, we empirically set $\lambda_{known} = 1$, $\lambda_{unknown} = 4$, $\lambda_{style} = 250$, $\lambda_{perceptual} = 0.1$ and $\lambda_{adv} = 0.2$. In what follows, we introduce the used datasets and evaluation metrics in our experiments.

A. Datasets and Evaluations

Datasets. In this study, we use four widely-used datasets for evaluating the image inpainting tasks, which includes three image datasets (i.e., Places2 [45], Paris Street View [46] and

Fig. 9. Visual comparison of the inpainting results of PC [1], PIC [39], RN [41] and FE [2] and our MR-InpaintNet on Paris Street View dataset.

TABLE II
NUMERICAL COMPARISON OF PC [1], PIC [39], RN [41] AND FE [2] AND OUR MR-INPAINTNET ON THE PARIS STREET VIEW DATASET, WHERE THE BOLD DENOTES THE BEST ONE.

Datasets		Paris Street View					CelebA-HQ				
Metric	Mask	PIC	PC	FE	RN	Ours	PIC	PC	FE	RN	Ours
l_1 (%)	0 – 10%	2.08	0.32	1.55	0.35	0.28	1.77	0.27	1.61	0.30	0.24
	10 – 20%	2.78	0.91	2.27	0.86	0.76	2.49	0.76	2.17	0.71	0.62
	20 – 30%	3.86	1.64	3.35	1.44	1.31	3.49	1.39	3.00	1.24	1.11
	30 – 40%	4.75	2.47	4.45	2.11	1.94	4.66	2.10	4.07	1.90	1.68
	40 – 50%	6.66	3.55	6.08	3.01	2.72	6.14	2.91	5.32	2.70	2.34
	50 – 60%	9.99	5.27	8.86	4.60	4.05	8.68	4.45	8.17	4.40	3.63
PSNR	0 – 10%	30.46	35.62	33.95	35.11	36.55	31.48	36.79	33.38	36.46	37.75
	10 – 20%	25.72	29.99	28.82	30.63	31.38	27.41	30.85	29.32	31.34	32.17
	20 – 30%	22.90	26.79	25.79	27.98	28.49	24.57	27.64	26.17	28.34	29.09
	30 – 40%	21.02	24.72	23.59	25.90	26.42	22.37	25.37	23.69	25.92	26.80
	40 – 50%	19.66	22.78	21.60	24.00	24.72	20.48	23.60	21.72	23.92	24.93
	50 – 60%	18.71	20.43	18.90	21.25	22.16	18.13	20.99	18.62	20.87	22.26
SSIM	0 – 10%	0.961	0.955	0.970	0.969	0.977	0.942	0.956	0.954	0.977	0.982
	10 – 20%	0.906	0.886	0.922	0.928	0.938	0.892	0.890	0.915	0.947	0.954
	20 – 30%	0.836	0.802	0.858	0.881	0.893	0.831	0.812	0.864	0.901	0.919
	30 – 40%	0.768	0.718	0.789	0.824	0.840	0.762	0.735	0.806	0.862	0.878
	40 – 50%	0.677	0.621	0.708	0.755	0.777	0.683	0.654	0.742	0.806	0.831
	50 – 60%	0.551	0.510	0.589	0.648	0.686	0.583	0.556	0.643	0.710	0.754

CelebA-HQ [22]) and one mask dataset. For the mask dataset, we use the free-form masks provided in [1].

- **Places2** dataset is a large-scale image dataset containing more than 1.8m images from 365 scenes. Each scene has 5k training images and 100 test images. In this study, we pick four scenes, i.e., butte, snowfield, canyon and farm, for our experiments.
- **Paris Street View** dataset is composed of street view images in Paris. It contains 15k images, where 1.49k images for training and 100 images for testing.

- **CelebA-HQ** dataset is a human face dataset that is a high-quality version from the CelebA dataset [47]. There are 30k face images in CelebA-HQ dataset. We randomly select 27k images for training and 3k images for testing.
- **Mask** dataset [1] contains 12k masks and all the masks are irregular. Those masks are divided into six types according to the size of the masked region.

Compared methods. Four start-of-the-art image inpainting methods are included for comparison, including PC [1], PIC [39], RN [41] and FE [2]:

Fig. 10. Visual comparison of the inpainting results of PC [1], PIC [39], RN [41] and FE [2] and our MR-InpaintNet on CelebA-HQ dataset.

- **PC**: a novel convolution layer termed partial convolution for image inpainting. The overall pipeline adopts a basic U-Net [42] architecture that all the convolutions are replaced with partial convolution.
- **PIC**: a new and probabilistically principled framework with two parallel paths, which is the first image inpainting method to obtain diverse inpainting results.
- **RN**: two novel normalization methods called RN-B and RN-L, which can make the whole model avoid the mean and variance shifts for image inpainting problems.
- **FE**: a new module feature equalization that can retain the consistency between structure and texture. A mutual encoder-decoder is used as its main framework.

For PIC, RN and FE methods, we directly use the source codes shared by authors. For PC, we reimplement the code because the authors did not provide the source codes.

Evaluation metrics. To evaluate the inpainting performance of each method, three commonly-used evaluation metrics are utilized following [41], which includes the l_1 loss, peak signal-to-noise ratio (PSNR), and the structural similarity (SSIM) with window size being 11. Specifically, l_1 loss can evaluate the difference in pixel level. Thus, the smaller l_1 loss, the better the inpainting result. In contrast, the larger the PSNR and SSIM values, the more realistic the restored images.

B. Image Inpainting Results

To fully compare the performance of each method, we provide both quantitative and qualitative inpainting results. For

TABLE III
INPAINTING RESULTS WITH DIFFERENT NETWORK ARCHITECTURES ON PARIS STREET VIEW DATASET AND CELEBA-HQ DATASET, WHERE THE BOLD DENOTES THE BEST RECORD.

Datasets		Paris Street View				CelebA-HQ			
Metric	Mask	High	High +Mid	High +Low	Ours	High	High +Mid	High +Low	Ours
l_1 (%)	0 – 10%	0.29	0.29	0.29	0.28	0.25	0.25	0.25	0.24
	10 – 20%	0.81	0.78	0.80	0.76	0.65	0.65	0.65	0.62
	20 – 30%	1.38	1.34	1.37	1.31	0.65	0.65	0.65	0.62
	30 – 40%	2.05	1.98	2.02	1.94	1.74	1.73	1.73	1.68
	40 – 50%	2.86	2.78	2.84	2.72	2.43	2.41	2.42	2.34
	50 – 60%	4.20	4.05	4.15	4.05	3.75	3.71	3.74	3.63
PSNR	0 – 10%	36.11	36.38	36.27	36.55	37.32	37.44	37.43	37.75
	10 – 20%	30.87	31.22	31.05	31.38	31.85	31.88	31.92	32.17
	20 – 30%	28.08	28.41	28.17	28.49	28.79	28.80	28.84	29.09
	30 – 40%	26.01	26.31	26.12	26.42	26.49	26.51	26.54	26.80
	40 – 50%	24.30	24.66	24.39	24.72	24.63	24.67	24.67	24.93
	50 – 60%	21.92	22.12	22.02	22.16	22.01	22.03	22.01	22.26
SSIM	0 – 10%	0.974	0.975	0.975	0.977	0.980	0.980	0.980	0.982
	10 – 20%	0.932	0.936	0.934	0.938	0.951	0.952	0.952	0.954
	20 – 30%	0.885	0.890	0.886	0.893	0.915	0.915	0.915	0.919
	30 – 40%	0.829	0.836	0.832	0.840	0.873	0.873	0.873	0.878
	40 – 50%	0.764	0.771	0.766	0.777	0.824	0.825	0.825	0.831
	50 – 60%	0.674	0.682	0.677	0.686	0.748	0.749	0.747	0.754

fair comparison on each method, the output image is directly used as the restored images without any postprocessing.

1) Results on Places2 dataset: We evaluate each method for repairing the corrupted images of the Places2 dataset. We first visualize some image inpainting results in Figure 8. As can be seen, when the mask is small, all inpainting methods can obtain plausible visual results. We also notice that there are some artifacts in the restored images of other compared methods. For example, in the first row of Figure 8, PIC method produces blurry textures; the recovered image

of PC method also contains the shadow of the mask; the FE method fails to maintain the consistency of light intensity and synthesizes blurry textures; there is also the shallow shadow for RN method. In contrast, our MR-InpaintNet generates the most realistic and accurate images.

Table I describes the numerical results on the Places2 dataset with different levels of masks. We can find that: (1) the overall performance of our MR-InpaintNet is better than other competitors in most cases; (2) when the mask is relatively small, PC and FE methods obtain the best results based on some metrics, and our MR-InpaintNet also delivers highly competitive results; (3) as the masked area increases, the performance of each inpainting method degrades, which indicates that it is more difficult to fill in the large missing region, since less known information can be used for prediction. This is also consistent with the visual results.

2) **Results on Paris Street View dataset:** We first visualize some inpainting results in Figure 9, from which we can see that: (1) each inpainting method can produce reasonable structures, which means that current methods can well understand the semantics of the damaged image; (2) when the scene is complex, both RN and PIC methods synthesize unrealistic details and blurry textures. In contrast, our MR-InpaintNet obtains better texture details and more consistent structures.

The numerical comparison results are described in Table II. We see that: (1) our method obtains the best results for six kinds of mask, compared with competitors; (2) the inpainting performance of each method also degrades with the increase of masked regions. And the performance degrading speed of our MR-InpaintNet is slowest, which means that our method has stronger learning ability and robustness to the corruptions.

3) **Results on CelebA-HQ dataset:** We further evaluate each method on CelebA-HQ dataset to show the generalization ability. Some examples of image inpainting results are shown in Figure 10. We see that all image inpainting methods can well restore the damaged images with small masks. We also observe that: (1) PIC method cannot recover the plausible shape of human face, as shown in the second row of Figure 10. FE method suffers from the brown spots on the human face. RN method cannot obtain realistic eyes when the eyes are masked. The restored images of PC method contain some artifacts; (2) our MR-InpaintNet obtains more realistic human face images. Note that the hair in the third row of Figure 10 is difficult to be recovered, although our method generates the most reasonable hair. This is because the inpainting network confuses the black hat and the hair. In such cases, when the network tries to recover the black hat and hair at the same time, a blurry repairing image will be generated.

The numerical inpainting results on CelebA-HQ dataset are described in Table II. We can find that: (1) compared with the scene images in Places2 dataset and the street view images in Paris Street View dataset, there is an obvious performance improvement on CelebA-HQ. Because human face contains more prior knowledge, and the structures and textures of human face are easier to be handled; (2) our MR-InpaintNet obtains the highest PSNR and SSIM, and the lowest l_1 loss, which verifies its strong generalization ability on human faces.

TABLE IV
NUMERICAL COMPARISON OF PC [1], PIC [39], RN [41] AND FE [2] AND OUR MR-INPAINTNET ON THE PARIS STREET VIEW DATASET, WHERE THE BOLD DENOTES THE BEST ONE.

Metric	Mask	Baseline	W/O MRFF	W/O AFE	W/O MEM	Ours
l_1 (%)	0 – 10%	0.30	0.30	0.29	0.29	0.28
	10 – 20%	0.82	0.82	0.79	0.79	0.76
	20 – 30%	1.41	1.40	1.36	1.35	1.31
	30 – 40%	2.11	2.05	2.01	2.00	1.94
	40 – 50%	3.00	2.87	2.80	2.78	2.72
	50 – 60%	4.54	4.17	4.16	4.09	4.05
PSNR	0 – 10%	35.94	35.96	36.17	36.33	36.55
	10 – 20%	30.83	30.85	31.06	31.06	31.38
	20 – 30%	27.95	28.12	28.25	28.29	28.49
	30 – 40%	25.82	26.15	26.18	26.21	26.42
	40 – 50%	23.97	24.45	24.49	24.58	24.72
	50 – 60%	21.36	22.08	22.01	22.11	22.16
SSIM	0 – 10%	0.974	0.974	0.975	0.976	0.977
	10 – 20%	0.932	0.932	0.935	0.935	0.938
	20 – 30%	0.885	0.885	0.888	0.889	0.893
	30 – 40%	0.828	0.831	0.834	0.835	0.840
	40 – 50%	0.759	0.766	0.769	0.771	0.777
	50 – 60%	0.658	0.678	0.677	0.681	0.686

C. Ablation Study

We conduct experiments to explore the effect of different modules on the performance of our MR-InpaintNet. To be specific, we evaluate the effectiveness of the deep multi-resolution learning, and the modules MRFF, AFE and MEM.

1) **Effect of deep multi-resolution learning:** To verify the effectiveness of multi-resolution learning on image inpainting, we conduct an ablation study by using the images from different resolutions as the input. In this study, four settings are evaluated, including (1) **High**: using only the high-resolution images as input; (2) **High + Mid**: using the high-resolution and middle-resolution images as input, which contains two encoders to handle them respectively; (3) **High + Low**: using the high-resolution and low-resolution images as input, which contains two encoders to handle them respectively; (4) using the images from multiple resolutions as input, i.e., our MR-InpaintNet, denotes as **Ours**. As described in Table III, we can find that **High** setting obtains the worst performance among all the models, while our MR-InpaintNet obtains the best results. That is, setting the images from multiple resolutions as input is effective to handle the issue of building the relationship of distant regions in high-resolution images.

We further visualize the feature maps of different encoders over multiple resolution images in Figure 11. We see that there are many detailed textures in the feature maps from the encoder handling high-resolution images. As for the feature maps from the encoder handling low-resolution images, we cannot see many details directly. In addition, the feature maps from the encoder handling middle-resolution images is more concrete compared to the feature maps from the encoder with low-resolution images. As mentioned earlier, it is an intermediate variable to adjust and reduce the gap between high resolution and low resolution, which can further improve the image inpainting results.

2) **Effect of MRFF:** To evaluate the effectiveness of the module MRFF, we remove it from the framework of our MR-InpaintNet for comparison, denoted as **W/O MRFF**. That is, **W/O MRFF** uses three dependent paths to deal with the input images of different resolutions. While the extracted multi-resolution features are not fused, and are directly concatenated

Fig. 11. The visual results of the feature maps obtained by different encoders.

and feed into the decoder. From Table IV, we see that there is an obvious degradation in terms of l_1 loss, PSNR and SSIM. This implies that the module MRFF is indeed effective for the inpainting task, because MRFF can fuse the features from different resolutions and understand the information in different levels so that better feature representations can be obtained to enhance the image inpainting task.

3) **Effect of AFE**: To examine the effect of the module AFE, we replace the weighted summation operation in AFE by the sum of the multi-resolution features and fused features, denoted as **W/O AFE**. Different from MR-InpaintNet that adaptively learns the weights of fused features, the weights of the fused features in **W/O AFE** are set to ones, which is equal to the weight of multi-resolution features. As can be seen in Table IV, our method is still superior to **W/O AFE**, which can also prove the effectiveness of the module AFE.

4) **Effect of MEM**: Finally, we evaluate the effect of the module MEM. Similarly, we remove it from MR-InpaintNet to construct a simple model, denoted as **W/O MEM**. As can be seen in Table IV, **W/O MEM** obtains poor results com-

Fig. 12. Visual results of the feature maps from different layers in decoder.

pared with that of MR-InpaintNet. This means that removing the module MEM from MR-InpaintNet indeed degrades the results. This is because removing MEM will make the deep neural network lose some information from the shallow layers when continuing to propagate forward. As such, our MR-InpaintNet equipped with MEM can well resolve this issue, since it can preserve the previous information and directly reuse the information from shadow layers in deeper layers.

5) **Visual analysis of deep layers in decoder**: The idea of deep neural networks comes from the human brain neurons, which always imitates human beings to perform specific tasks. Therefore, we explore how our MR-InpaintNet repairs the damaged images in this part. When humans want to reconstruct an image, they usually draw the outline firstly, and then add the details. In other words, we first construct the overall structures, and then restore the detailed texture. In Figure 12, we show the visual results of the feature maps obtained from different layers in the decoder of our method. We can see that there are more and more detail information from the shadow to deep layers. Specifically, in the first layer of the decoder, we can only see the blurred outline. In the second and third layers, the feature maps contain some obvious structures. In the fourth layer, it is easy to find that the feature maps contain rich structural and textural information. As a result, in term of repairing a damaged image, the learning process of our MR-InpaintNet is similar with that of humans according to Figure 12. To a certain extent, these observations demonstrate that the design of our MR-InpaintNet is reasonable.

V. CONCLUSION

We presented deep multi-resolution learning for handling the image inpainting task, and proposed a novel deep inpainting network termed MR-InpaintNet. Based on the images from different resolutions, MR-InpaintNet can build the relationship between distant regions and obtain rich information in different levels, including both texture information and meaningful semantic information. To propagate the relationship and further improve the features, we have designed two new modules, i.e., a multi-resolution feature fusion module for fusing the features from different resolutions and an adaptive feature enhancement for automatically refining the multi-resolution features via embedding the fused features. Furthermore, to avoid the information loss in deeper layers and prevent MR-InpaintNet from the performance degradation, we also present

a memory enhanced mechanism to retain the previous information in each layer. Extensive visual and numerical results have demonstrated the effectiveness of our MR-IpaintNet. In future, we hope that the image inpainting results can be controlled artificially in the real-world applications. Therefore, designing an image inpainting algorithm that can be flexibly intervened by humans is an important problem to be solved. Besides, evaluating the deep multi-resolution learning for other computer vision tasks is also an interesting future work.

ACKNOWLEDGMENT

This work is partially supported by National Natural Science Foundation of China (62072151, 62020106007, 61972112 and 61832004), Anhui Provincial Natural Science Fund for Distinguished Young Scholars (2008085J30), Guangdong Basic and Applied Basic Research Foundation under Grant no. 2021B1515020088, and the Fundamental Research Funds for Central Universities of China (JZ2019HGPA0102). Zhao Zhang is the corresponding author of this paper.

REFERENCES

- [1] G. Liu, F. A. Reda, K. J. Shih, T.-C. Wang, A. Tao, and B. Catanzaro, "Image inpainting for irregular holes using partial convolutions," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018, pp. 85–100.
- [2] H. Liu, B. Jiang, Y. Song, W. Huang, and C. Yang, "Rethinking image inpainting via a mutual encoder-decoder with feature equalizations," in *Computer Vision—ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part II 16*. Springer, 2020, pp. 725–741.
- [3] L. Song, J. Cao, L. Song, Y. Hu, and R. He, "Geometry-aware face completion and editing," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 33, no. 01, 2019, pp. 2506–2513.
- [4] Q. Sun, L. Ma, S. J. Oh, L. Van Gool, B. Schiele, and M. Fritz, "Natural and effective obfuscation by head inpainting," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 5050–5059.
- [5] P. Buysens, O. Le Meur, M. Daisy, D. Tschumperlé, and O. Lézoray, "Depth-guided disocclusion inpainting of synthesized rgb-d images," *IEEE Transactions on Image Processing*, vol. 26, no. 2, pp. 525–538, 2016.
- [6] A. Criminisi, P. Pérez, and K. Toyama, "Region filling and object removal by exemplar-based image inpainting," *IEEE Transactions on image processing*, vol. 13, no. 9, pp. 1200–1212, 2004.
- [7] C. Barnes, E. Shechtman, A. Finkelstein, and D. B. Goldman, "Patch-match: A randomized correspondence algorithm for structural image editing," *ACM Trans. Graph.*, vol. 28, no. 3, p. 24, 2009.
- [8] K. He and J. Sun, "Statistics of patch offsets for image completion," in *European conference on computer vision*. Springer, 2012, pp. 16–29.
- [9] —, "Image completion approaches using the statistics of similar patches," *IEEE transactions on pattern analysis and machine intelligence*, vol. 36, no. 12, pp. 2423–2435, 2014.
- [10] M. Ghorai, S. Samanta, S. Mandal, and B. Chanda, "Multiple pyramids based image inpainting using local patch statistics and steering kernel feature," *IEEE Transactions on Image Processing*, vol. 28, no. 11, pp. 5495–5509, 2019.
- [11] H. Li, W. Luo, and J. Huang, "Localization of diffusion-based inpainting in digital images," *IEEE transactions on information forensics and security*, vol. 12, no. 12, pp. 3050–3064, 2017.
- [12] K. Li, Y. Wei, Z. Yang, and W. Wei, "Image inpainting algorithm based on tv model and evolutionary algorithm," *Soft Computing*, vol. 20, no. 3, pp. 885–893, 2016.
- [13] G. Sridevi and S. S. Kumar, "Image inpainting based on fractional-order nonlinear diffusion for image reconstruction," *Circuits, Systems, and Signal Processing*, vol. 38, no. 8, pp. 3802–3817, 2019.
- [14] Y. Wei, Z. Zhang, Y. Wang, M. Xu, Y. Yang, S. Yan, and M. Wang, "Deraincyclegan: Rain attentive cyclegan for single image deraining and rainmaking," *IEEE Transactions on Image Processing*, vol. 30, pp. 4788–4801, 2021.
- [15] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.
- [16] S. Ren, K. He, R. Girshick, and J. Sun, "Faster r-cnn: Towards real-time object detection with region proposal networks," *Advances in neural information processing systems*, vol. 28, pp. 91–99, 2015.
- [17] L. Liu, H. Zhang, X. Xu, Z. Zhang, and S. Yan, "Collocating clothes with generative adversarial networks cosupervised by categories and attributes: a multidiscriminator framework," *IEEE transactions on neural networks and learning systems*, vol. 31, no. 9, pp. 3540–3554, 2019.
- [18] D. Pathak, P. Krahenbuhl, J. Donahue, T. Darrell, and A. A. Efros, "Context encoders: Feature learning by inpainting," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 2536–2544.
- [19] S. Iizuka, E. Simo-Serra, and H. Ishikawa, "Globally and locally consistent image completion," *ACM Transactions on Graphics (ToG)*, vol. 36, no. 4, pp. 1–14, 2017.
- [20] R. A. Yeh, C. Chen, T. Yian Lim, A. G. Schwing, M. Hasegawa-Johnson, and M. N. Do, "Semantic image inpainting with deep generative models," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 5485–5493.
- [21] C. Yang, X. Lu, Z. Lin, E. Shechtman, O. Wang, and H. Li, "High-resolution image inpainting using multi-scale neural patch synthesis," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 6721–6729.
- [22] J. Yu, Z. Lin, J. Yang, X. Shen, X. Lu, and T. S. Huang, "Free-form image inpainting with gated convolution," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2019, pp. 4471–4480.
- [23] K. Nazeri, E. Ng, T. Joseph, F. Z. Qureshi, and M. Ebrahimi, "Edge-connect: Generative image inpainting with adversarial edge learning," *arXiv preprint arXiv:1901.00212*, 2019.
- [24] Y.-G. Shin, M.-C. Sagong, Y.-J. Yeo, S.-W. Kim, and S.-J. Ko, "Pepsi++: Fast and lightweight network for image inpainting," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 32, no. 1, pp. 252–265, 2020.
- [25] J. Yu, Z. Lin, J. Yang, X. Shen, X. Lu, and T. S. Huang, "Generative image inpainting with contextual attention," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 5505–5514.
- [26] W. Xiong, J. Yu, Z. Lin, J. Yang, X. Lu, C. Barnes, and J. Luo, "Foreground-aware image inpainting," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 5840–5848.
- [27] H. Liu, B. Jiang, Y. Xiao, and C. Yang, "Coherent semantic attention for image inpainting," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2019, pp. 4170–4179.
- [28] Y. Zeng, J. Fu, H. Chao, and B. Guo, "Learning pyramid-context encoder network for high-quality image inpainting," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 1486–1494.
- [29] Y. Ren, X. Yu, R. Zhang, T. H. Li, S. Liu, and G. Li, "Structureflow: Image inpainting via structure-aware appearance flow," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2019, pp. 181–190.
- [30] C. Xie, S. Liu, C. Li, M.-M. Cheng, W. Zuo, X. Liu, S. Wen, and E. Ding, "Image inpainting with learnable bidirectional attention maps," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2019, pp. 8858–8867.
- [31] L. Shen, R. Hong, H. Zhang, H. Zhang, and M. Wang, "Single-shot semantic image inpainting with densely connected generative networks," in *Proceedings of the 27th ACM International Conference on Multimedia*, 2019, pp. 1861–1869.
- [32] X. Hong, P. Xiong, R. Ji, and H. Fan, "Deep fusion network for image completion," in *Proceedings of the 27th ACM International Conference on Multimedia*, 2019, pp. 2033–2042.
- [33] N. Wang, J. Li, L. Zhang, and B. Du, "Musical: Multi-scale image contextual attention learning for inpainting," in *IJCAI*, 2019, pp. 3748–3754.
- [34] Z. Guo, Z. Chen, T. Yu, J. Chen, and S. Liu, "Progressive image inpainting with full-resolution residual network," in *Proceedings of the 27th acm international conference on multimedia*, 2019, pp. 2496–2504.
- [35] J. Li, F. He, L. Zhang, B. Du, and D. Tao, "Progressive reconstruction of visual structure for image inpainting," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2019, pp. 5962–5971.
- [36] J. Li, N. Wang, L. Zhang, B. Du, and D. Tao, "Recurrent feature reasoning for image inpainting," in *Proceedings of the IEEE/CVF*

Conference on Computer Vision and Pattern Recognition, 2020, pp. 7760–7768.

- [37] Z. Yi, Q. Tang, S. Azizi, D. Jang, and Z. Xu, “Contextual residual aggregation for ultra high-resolution image inpainting,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 7508–7517.
- [38] L. Liao, J. Xiao, Z. Wang, C.-W. Lin, and S. Satoh, “Guidance and evaluation: Semantic-aware image inpainting for mixed scenes,” in *Computer Vision—ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXVII 16*. Springer, 2020, pp. 683–700.
- [39] C. Zheng, T.-J. Cham, and J. Cai, “Pluralistic image completion,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 1438–1447.
- [40] H. Liu, Z. Wan, W. Huang, Y. Song, X. Han, and J. Liao, “Pd-gan: Probabilistic diverse gan for image inpainting,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 9371–9381.
- [41] T. Yu, Z. Guo, X. Jin, S. Wu, Z. Chen, W. Li, Z. Zhang, and S. Liu, “Region normalization for image inpainting,” in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 34, no. 07, 2020, pp. 12 733–12 740.
- [42] O. Ronneberger, P. Fischer, and T. Brox, “U-net: Convolutional networks for biomedical image segmentation,” in *International Conference on Medical image computing and computer-assisted intervention*. Springer, 2015, pp. 234–241.
- [43] A. Paszke, S. Gross, S. Chintala, G. Chanan, E. Yang, Z. DeVito, Z. Lin, A. Desmaison, L. Antiga, and A. Lerer, “Automatic differentiation in pytorch,” 2017.
- [44] D. P. Kingma and J. Ba, “Adam: A method for stochastic optimization,” *arXiv preprint arXiv:1412.6980*, 2014.
- [45] B. Zhou, A. Lapedriza, A. Khosla, A. Oliva, and A. Torralba, “Places: A 10 million image database for scene recognition,” *IEEE transactions on pattern analysis and machine intelligence*, vol. 40, no. 6, pp. 1452–1464, 2017.
- [46] 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, 2012.
- [47] Y. Zhang, Z. Yin, Y. Li, G. Yin, J. Yan, J. Shao, and Z. Liu, “Celebapspoo: Large-scale face anti-spoofing dataset with rich annotations,” in *European Conference on Computer Vision*. Springer, 2020, pp. 70–85.

Huan Zheng (Student Member, IEEE) received the bachelor degree from Northwest Agriculture & Forestry University. He is currently working toward the research degree at School of Computer Science and Information Engineering, Hefei University of Technology, China. His current research interests include deep learning, data mining, and their applications. He is interested in designing advanced algorithms for image representation and restoration.

Zhao Zhang (Senior Member, IEEE) received the Ph.D. degree from City University of Hong Kong in 2013. He is now a Professor at School of Computer Science, Hefei University of Technology, China. He was a Visiting Research Engineer at National University of Singapore, from Feb to May 2012. He then visited the National Laboratory of Pattern Recognition at Chinese Academy of Sciences, from Sep to Dec 2012. His research interests include Multimedia, Machine Learning and Computer Vision. He has authored/co-authored over 80 technical papers

published at prestigious journals and conferences, including 36 IEEE/ACM Transactions papers, with Google Scholar citations over 2,900 times and H-index 31. He serves as an Associate Editor (AE) of IEEE Transactions on Image Processing (IEEE TIP) and Pattern Recognition (PRJ), etc. Besides, he has been acting as a SPC member/Area Chair of ACM MM, IJCAI, SDM, ECAI, BMVC and PAKDD. He is now a Senior Member of the IEEE.

puting and Applications, and Pattern Analysis and Applications.

Haijun Zhang (Senior Member, IEEE) received the Ph.D. degree from City University of Hong Kong, Hong Kong in 2010. He was a Post-Doctoral Research Fellow with the University of Windsor, Windsor, Canada, from 2010 to 2011. Since 2012, he has been with Harbin Institute of Technology (Shenzhen), China, where he is currently a full Professor of Computer Science. His current research interests include multimedia data mining, machine learning, and computational advertising. He is currently an Associate Editor of *Neurocomputing*, *Neural Computing and Applications*, and *Pattern Analysis and Applications*.

Yi Yang (Senior Member, IEEE) received the Ph.D. degree in computer science from Zhejiang University, Hangzhou, China, in 2010. He is currently a Professor with the University of Technology Sydney, Ultimo, NSW, Australia. He was a Post-Doctoral Researcher with the School of Computer Science, Carnegie Mellon University, Pittsburgh, PA, USA. His current research interests include machine learning and its applications to multimedia content analysis and computer vision.

Intelligent Systems and Technology (TIST), and Journal of Computer Vision and Image Understanding. He is a Fellow of Academy of Engineering, Singapore, ACM Fellow, IEEE Fellow, IAPR Fellow.

Shuicheng Yan (Fellow, IEEE) received the Ph.D. degree from the Peking University in 2004. He is currently director of Sea AI Lab (SAIL) and group chief scientist of Sea. His research areas include computer vision, machine learning and multimedia analysis. Till now, he has published over 600 papers in top international journals and conferences, with H-index 117. He is/has been an Associate Editor of IEEE Trans. on Knowledge and Data Engineering (TKDE), IEEE Trans. on Circuits and Systems for Video Technology (TCSVT), ACM Transactions on

Engineering, IEEE Trans. on Neural Networks and Learning Systems, etc. He is now a Fellow of the IEEE.

Meng Wang (Fellow, IEEE) is a professor at Hefei University of Technology, China. He worked as an associate researcher at Microsoft Research Asia and a senior research fellow at National University of Singapore. His current research interests include multimedia content analysis and computer vision. He has authored or co-authored over 200 book chapters, journal and conference papers. He is or has been an editorial board member of IEEE Trans. on Circuits and Systems for Video Technology, IEEE Trans. on Multimedia, IEEE Trans. on Knowledge and Data Engineering, IEEE Trans. on Neural Networks and Learning Systems, etc.