IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received March 5, 2020, accepted March 20, 2020, date of publication March 31, 2020, date of current version April 14, 2020.

Digital Object Identifier 10.1109/ACCESS.2020.2984587

Image Inpainting Based on Inside–Outside Attention and Wavelet Decomposition

XINGCHEN HE[1,2], **XUDONG CUI**[1], **AND QILONG LI**[2]
[1]Institute of Chemical Materials, China Academy of Engineering Physics, Mianyang 621900, China
[2]School of Information Science and Technology, Chengdu University of Technology, Chengdu 610051, China

Corresponding author: Xudong Cui (xudcui@163.com)

ABSTRACT Recently developed deep learning-based image inpainting methods have suggested their potential applications in filling large missing regions displaying plausible content. Nevertheless, most existing studies either merely considered the external information of the missing regions or relied on the region context to yield semantically coherent patches while overlooking the semantic relevance and feature continuity exhibited by missing regions; these demerits are likely to cause a discontinuous contextual structure and blurry texture details. In this work, a novel inside-outside attention layer (IOA) was proposed, capable of exploiting unmasked image features as references as well as learning the affinity between hole features to predict more consistent semantic information. The adversarial loss attributed exclusively on the natural image level cannot adequately generate a sharp texture detail. To address this problem, a texture component discriminator was introduced via wavelet decomposition to enhance the specific performance. Several experiments were performed on the CelebA and Places2 datasets. As revealed from the results, in contrast to the existing research, the proposed method is capable of restoring images with complex structures and significantly enhancing plausible structure and visual quality.

INDEX TERMS Image inpainting, attention layer, wavelet decomposition, texture component discriminator.

I. INTRODUCTION

Image inpainting refers to an image processing technique, employing the undamaged information of the image to repair missing information or remove unwanted image fractions while maintaining the quality and natural structure of the image. In other words, the repair work faces the primary challenge of synthesizing visually realistic and semantically reasonable pixels for the missing regions and simultaneously maintaining consistency with the existing pixels. Given the wide existence of images, image inpainting can be broadly applied in the protection of artwork, which is reflected by the repairment of lost and broken information in old photos, the hiding of errors in pictures and videos, the removal of unwanted fractions of images, as well as image-based rendering and computer photography. Accordingly, the relevant task has aroused huge attention in recent years.

On the whole, conventional image inpainting methods are split into two major types in accordance with the sizes of the damaged regions. 1) Images with small damage scales (e.g., scratches) can be fixed using two methods. First, the partial differential equation (PDE) in physics is adopted to propagate the known information to the missing areas to achieve image inpainting [4]–[6]. The other complies with the variational principle; the inpainting problem is converted into a variational problem of the extremum by building a priori model and a data model of the image. 2) In terms of images exhibiting large damage scales, a sample patches-based texture synthesis algorithm is proposed, selecting the appropriate sample patch size in missing areas and then substituting it with the most similar patch from non-missing areas [1], [7]. However, conventional methods often fail to generate semantically reasonable results since they fill regions at the image level and lack a high-level understanding of the image. In contrast, early deep learning methods [3], [8], [9], [10] are capable of learning semantic priors and data distributions of the original image with generative adversarial networks, which can enhance the visual authenticity of the image. Nevertheless, these CNN-based methods cannot effectively exploit contextual semantic information; they often generate boundary artifacts and distorted structures discontinuous with unmasked regions.

The associate editor coordinating the review of this manuscript and approving it for publication was Hengyong Yu.

Several recent studies focused on exploiting the semantic information of non-missing regions to guide high-level features restoring in missing regions [11]–[13]. The mentioned methods are capable of maintaining the semantic continuity of the yielded image well. Yet they are limited to rectangular shape masks and merely consider the similarity between the pixels to be generated in missing regions and the known pixels outside missing regions, thereby overlooking the correlation between the generated pixels. Thus, they will restore blurry texture details and pixel-discontinuous results inside masked areas.

In the present study, to ensure high global structure consistency, an attention mechanism was used to fill missing regions at the high-feature level more effectively. The entire image inpainting process was split into two stages. The first stage aimed to assess a rough result similar to the missing part, while the second stage complied with a U-net [14] structure contributing to image denoising and extracting high-level semantic expressions to refine the sketchy result output from the first stage. To simultaneously preserve boundary pixel continuity and mask-inner semantic correlation, an inside-outside attention transfer layer (IOA) was embedded in this stage. The IOA first learned the affinity between the patches outside/inside unknown areas and then regarded it as outside attention. Second, the inside attention was calculated, complying with the similarity between adjacent patches inside missing regions. Third, the combination of outside and inside attention was transferred to missing regions to deal with generated patches. The IOA layer embedded here made the contextual semantics more coherent and achieved more effective inpainting in restoring a natural structure. Moreover, to eliminate boundary artifacts and enhance the texture details of images generated, a texture component discriminator was proposed, transferring images to the wavelet domain, learning the distribution of detail wavelet coefficients achieved by wavelet decomposition, as well as representing the texture part of the image to sharpen the images generated. The whole network was trained end to end with reconstruction loss, multiscale reconstruction loss [13], natural image adversarial loss, as well as texture component adversarial loss. The two adversarial losses primarily originated from the wavelet domain and natural domain of the missing regions. Experiments on multiple datasets (e.g., faces and natural images) proved that the proposed method generates higher-quality inpainting results than existing works. Our contributions are concluded as follows:

1) An inside-outside attention transfer layer was proposed to learn the affinity between high-level feature patches to more effectively reconstruct each missing pixel. The proposed method is capable of achieving satisfactory results for any shape of the missing regions (rectangular shape or irregular shape).

2) To enhance the texture details performance, a wavelet transform was introduced into the framework here, and clearer details were restored by learning the high-frequency coefficient distribution of the ground truth.

3) The proposed generative image inpainting system is capable of achieving high-quality and plausible results even if the images exhibit complex structures. Besides, the proposed method achieved more visual authentic inpainting results than existing methods on challenging datasets Place2 [15] and CelebA [16].

II. RELATED WORK

A. TRADITIONAL METHOD

Early image inpainting primarily represents diffusion-based or patch-based methods exhibiting low-level features. With diffusion equations along the mask boundary, Bertalmio *et al*. [17] iteratively transferred low-level features of known regions to unknown regions. Though the method performs well in repairs, it can be limited to handling small and evenly distributed areas. Bertalmio *et al.* [18] further optimized repair results by introducing texture synthesis. By learning the a priori of the image block, Chan and Shen [19] recovered images with missing pixels. Recently developed TV-based methods [20] have begun to consider the smoothness of the image, which is feasible to repair small missing areas and eliminate noise. Nevertheless, with the enhancement of missing areas, such method can cause a blurred appearance for the pixels inside known regions diffused to the unknown regions. Hay and Efros [21] attempted to employ the available parts of the image to find similar patches for its high-quality and high-efficiency results, which had been adopted as one of the most effective inpainting methods in a short time; however, as they assumed that missing patches can be identified somewhere in known regions, matching patches are hard to find for images with more complex textures. Moreover, these methods exhibit poor performance in extracting high-level semantics.

B. DEEP LEARNING METHOD

To restore large and irregular missing regions, more learning-based methods have been developed. These learning-based methods primarily use GAN to identify the latent distribution behaviors of missing regions and enhance the performance in semantic image inpainting compared with conventional methods. Larsen *et al.* [22] optimized the VAE [23] by adding an adversarial training discriminator capable of automatically encoding the input of an incomplete picture into a vector before decoding the vector into a complete image; they demonstrated that a more realistic image could be generated. Inspired by this work, the context encoder model was built [3], employing an encoder to combine learning visual representation with image inpainting and train deep neural networks with pixelwise reconstruction loss and generative adversarial loss as the objective function, whereas the restoring regions were obviously inconsistent in some cases. To solve the problems within the context encoder model, Iizuka *et al.* [8] extended the design to two discriminators and adopted trained global and local context discriminators to distinguish real and images generated,

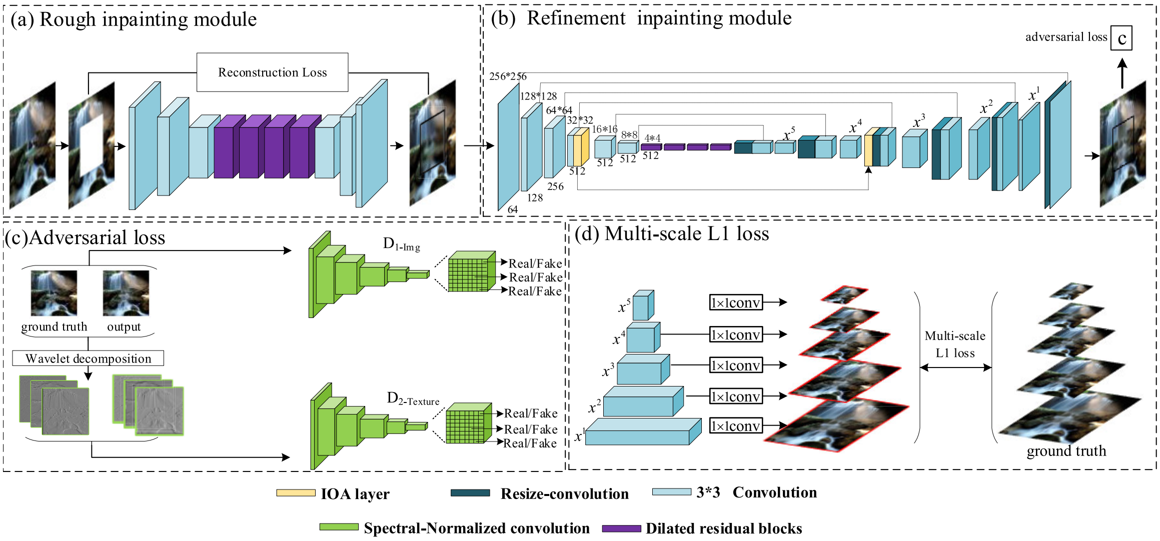


FIGURE 1. Our architecture build on U-net structure and it consists of two module: (a) rough inpainting module and (b) refinement inpainting module. Refinement inpainting module is optimize by minimizing (c) adversarial loss and (d) multi-scale L1 loss to predict better results.

respectively, thereby enabling the network to generate locally and globally consistent images. EdgeConnect [24] built a two-stage adversarial model with an edge generator under an image completion network. The edge generator can visualize edges of the missing regions of the image, and the image completion network can fill in the missing regions with visualized edges as a priori. This method, however, relies on the edge information prediction, and the existing methods fail to effectively restore reasonable edges. Such inpainting approach may lead to a confusing detailed texture. The partial convolutions [25] make and renormalize the convolution to be conditioned on only valid pixels. Though it can achieve sharp results on small and irregular missing regions, this method is likely to form a destroy structure when missing regions turn out to be continuous and larger. Shift-Net [12] introduces a shift-connection to shift the encoder feature of known regions to estimate missing parts. Contextual attention [11] designs a coarse-to-fine framework to guide image restoration; it first generates a rough estimate of the masked regions, and subsequently an attention mechanism is proposed and embedded in a refinement network to synthesize semantically relevant structures using high-level features in known regions as a reference.

III. APPROACH

In the present section, an image inpainting model is first proposed based on attention layers (IOA), successfully generating missing regions obscured by free-form masks as well as maintaining high contextual semantic consistency and visual authenticity. Subsequently, the operating mechanism of the IOA layer is described, followed by the training loss functions.

A. NETWORK ARCHITECTURE

Inspired by recent image completion studies [11], the image inpainting network here consisted of two modules, namely, a rough inpainting module and a refinement inpainting module (Fig. 1). The rough inpainting module initialized a coarse image to stabilize the training process and enlarge the receptive fields. Besides, the coarse prediction presented prior knowledge and smoother pixels for IOA in missing regions, thereby enhancing the calculation efficiency of the attention score. Context semantics and texture details of images were primarily reconstructed by refinement networks. The details of the two modules are elucidated as follows.

1) ROUGH INPAINTING MODULE

In this module, masked image Im and center or irregular mask M were combined as the input, and a coarse prediction Irough was outputted by the rough network. The architecture of the generators complies with the method proposed by [26], which has proven feasible in style transfer, super-resolution [27], as well as image-to-image translation [28]. A 3 × 3 sized kernel and local signal normalization [29] was employed across all layers of the network except for input and output layers. The tensor size of input was 256 × 256 × 4, covering an incomplete image and a binary mask. The encoder of the generator covered two downsampling operations. Each time, the image size was reduced by one-half, four dilated residual blocks [30] with a dilation factor of two were followed before decoding. To prevent the occurrence of checker-board artifacts [31], our decoder adopted the resize convolution [32] (bilinear interpolation was followed to resize the image) for upsampling. Reconstruction loss was used to optimize the rough network parameters explicitly.

2) REFINEMENT INPAINTING MODULE

At this stage, the output results from rough inpainting module Irough conditioned on binary mask M were adopted as input. To extract the feature information inside missing regions at different levels, the generator exploited a U-net structure capable of encoding a masked image into tight latent features and reusing the low-level feature information of the previous layer in the decoding step. The encoder of the generator downsampled the input to a 4 × 4 size with 6 convolution operations, followed by four dilated residual blocks. Consistent with the rough inpainting module, all convolution layers adopted a 3 × 3 sized kernel. The IOA was located in a 32 × 32 size downsampling layer of the encoder, and the output of the IOA was embedded in the corresponding layer of the decoder.

Our discriminator displayed an SN-PatchGAN [33] structure, and the receptive fields of each point in the output map could cover the entire input image in our training setting continuously. Thus, a global/local discriminator design would be not required in SN-PatchGAN. The discriminator forced the data distribution of reconstruction images consistent with that of real images to enhance the visual authenticity of the reconstructed images, whereas it only exploited adversarial loss built at the entire natural image level, so a clear and sharp texture details were not achieved in the results. To address this problem, a texture component discriminator was proposed to learn the texture component distribution of real images and enhance the visual quality of inpainting from the details. Multiscale reconstruction loss and natural image adversarial loss are also indispensable for training, and the details of the loss function are presented in section C.

B. INSIDE-OUTSIDE ATTENTION TRANSFER LAYER

In the image inpainting task, features in missing regions were sometimes more tightly correlated with those in distant spatial locations, whereas convolutional neural networks could not borrow features from such a long distance with local convolution kernels. Reference [11] proposed an attention model adopting the features of known regions to guide feature generation in missing region $\bar{M}$, whereas this method ignored the correlation between generated patches and might cause a semantically incoherent inpainting result. To address this problem, an IOA layer was proposed, which not only considered patches in M but also focused on the correlation between the neighboring generated patches in the missing areas. It covers three parts, namely, reconstruction, relevance calculation and generation. The procedure is illustrated in Fig 2.

1) RECONSTRUCTION

Following the state-of-the-art approaches [11], we first extract patches (3 × 3) in known regions and reshape them as convolutional filters to measure the affinity between patches inside and outside M, then calculate with normalized inner product (cosine similarity). Note that the values of the patches inside M are initialized by a rough inpainting module.

$$S_{i,j} = \frac{< p_i,\bar{p}_j >}{||p_i|| \cdot ||\bar{p}_j||}. \tag{1}$$

where p_i represents the i-th patch extracted from M and $i \in (1,n)$, n is the number of patches in M. $\bar{p}_j$ denotes the j-th patch extracted from contextual regions $\bar{M}$, $j \in (1,m)$. m is the number of patches in $\bar{M}$. $S_{i,j}$ represents similarity between p_i and $\bar{p}_j$

Then softmax is applied to compute the attention score $\lambda_{i,j}$ for each patch:

$$\lambda_{i,j} = \frac{\exp(S_{i,j})}{\sum_{j=1}^{m} \exp(S_{i,j})}. \tag{2}$$

After obtaining the attention score, the outside attention patch is defined as:

$$p_i{}' = \sum_{j=1}^{m} \lambda_{i,j} \cdot \bar{p}_j. \tag{3}$$

where $p_i{}'$ represents the reconstructed outside attention patch, which corresponds to the i-th position in M

2) RELEVANCE CALCULATION

In this part, we focus on the semantic relevance R^i_{inside} between two adjacent patches inside M, and set the previous patch p_{i-1} as an important reference to reconstruct p_i. To maintain the continuity and integrity of the reconstruction results, feature information outside M also need to be considered. so $p_i{}'$ is set to another reference during the process of reconstruction, and R^i_{inside} and $R^i_{outside}$ are denoted as:

$$R^i_{inside} = \frac{< p_i,p_{i-1} >}{||p_i|| \cdot ||p_{i-1}||}. \tag{4}$$

$$R^i_{outside} = \frac{< p_i,p_i{}' >}{||p_i|| \cdot ||p_i{}'||}. \tag{5}$$

3) GENERATION

We reconstruct the patches inside M in order from left to right and top to bottom. p_i^{New} stands for the reconstruction result of p_i. Note that p_1 has no previous patch, consequently p_1^{New} is directly replaced by outside attention patch $p_1{}'$. Thus the first reconstructed patch $p_1^{New} = p_1{}'$. After obtaining the value of p_1^{New}, we can utilize R^i_{inside} and $R^i_{outside}$ to calculate p_2^{New}. In summary,the reconstruction process can be denoted as:

$$p_i^{New} = \frac{R^i_{inside}}{R^i_{inside} + R^i_{outside}} \times p_{i-1}^{New} + \frac{R^i_{outside}}{R^i_{inside} + R^i_{outside}} \times p_i{}'. \quad (i \in (2,n)) \tag{6}$$

Note that the generate operation is an iterative process, which means all previous generated patches (p_1^{New} to p_{i-1}^{New}) and outside attention patch $p_i{}'$ are contribute to the reconstruction of p_i^{New}, thus more contextual semantic information can fill in missing regions. Finally, we reuse extracted patches ($\bar{p}_1$ to $\bar{p}_m$) in $\bar{M}$ as deconvolutional filters to reconstruct M.

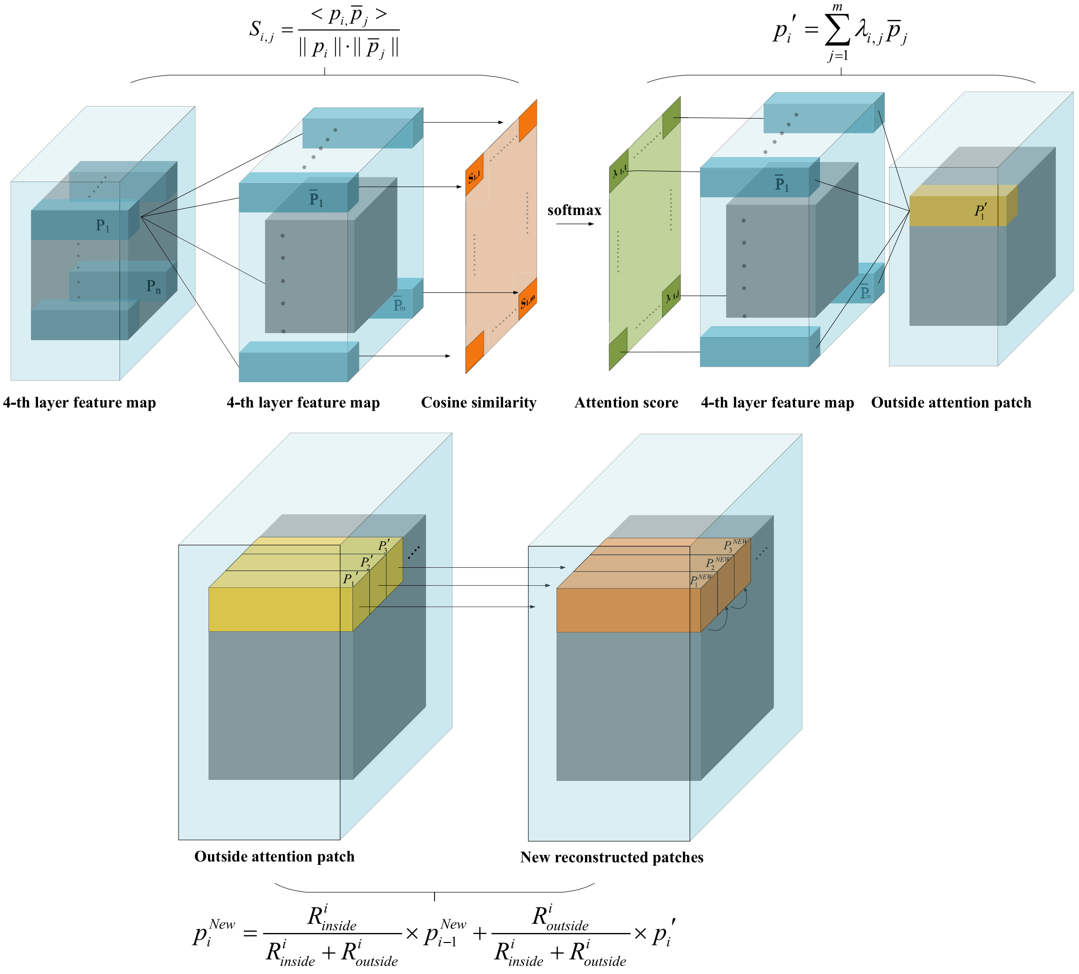

$S_{i,j} = \frac{< p_i, \bar{p}_j >}{\| p_i \| \cdot \| \bar{p}_j \|}$
$p_i' = \sum_{j=1}^{m} \lambda_{i,j} \bar{p}_j$
P1
Pn
$\bar{P}_1$
$\bar{P}_m$
$S_{1,1}$
$S_{1,m}$
softmax
$\lambda_{1,1}$
$\lambda_{1,m}$
P_1'
4-th layer feature map
4-th layer feature map
Cosine similarity
Attention score
4-th layer feature map
Outside attention patch
P_1'
P_2'
P_3'
P_1^{NEW}
P_2^{NEW}
P_3^{NEW}
Outside attention patch
New reconstructed patches
$p_i^{New} = \frac{R^i_{inside}}{R^i_{inside} + R^i_{outside}} \times p_{i-1}^{New} + \frac{R^i_{outside}}{R^i_{inside} + R^i_{outside}} \times p_i'$

FIGURE 2. Structural overview of Inside-outside attention.

C. LOSS FUNCTION

We want to obtain a approximately the same result as the real image in the rough inpainting module, by which we can provide an initial value for the pixels in M to calculate the attention score. By minimizing the reconstruction loss, the ability of the generated model to learn the overall structure is optimized. In the refinement inpainting module,we use multiscale reconstruction loss [13] to restore the generated image at different scales. Considering that images generated with only reconstruction losses usually obtain blurred results, adversarial loss is used to force the data distribution of generated images consistent with that of real images, which greatly improves the visual quality of generated images. Furthermore,the normal discriminator can only learn the statistics of natural images, and it can hardly capture high frequency texture details to produce satisfactory perceptual results. As high frequency wavelet coefficients can characterize the texture details of images, we proposed a texture component adversarial loss which is based on wavelet decomposition to help texture reconstruction. The details of the multi-scale reconstruction loss and adversarial loss are described below.

1) RECONSTRUCTION LOSS

In the rough inpainting module, we use L^1 distance as the reconstruction loss to constrain the difference between the rough result I_{rough} and the ground-truth I_{gt}.

$$L_r = \frac{1}{N_{I_{gt}}} ||M \odot (I_{rough} - I_{gt})||_1 + \alpha \frac{1}{N_{I_{gt}}} ||\bar{M} \odot (I_{rough} - I_{gt})||_1. \quad (7)$$

where N_a denotes the total number of pixels in a.

2) MULTISCALE RECONSTRUCTION LOSS

In the refinement inpainting module, we follow [13] to use multiscale reconstruction losses to refine the constraints on the predicted values at each scale. the groud truth is downsampled to different scales that correspond to the feature maps from each layer in the decoder. Then $L1$ loss is applied between the feature maps of prediction and the ground truth

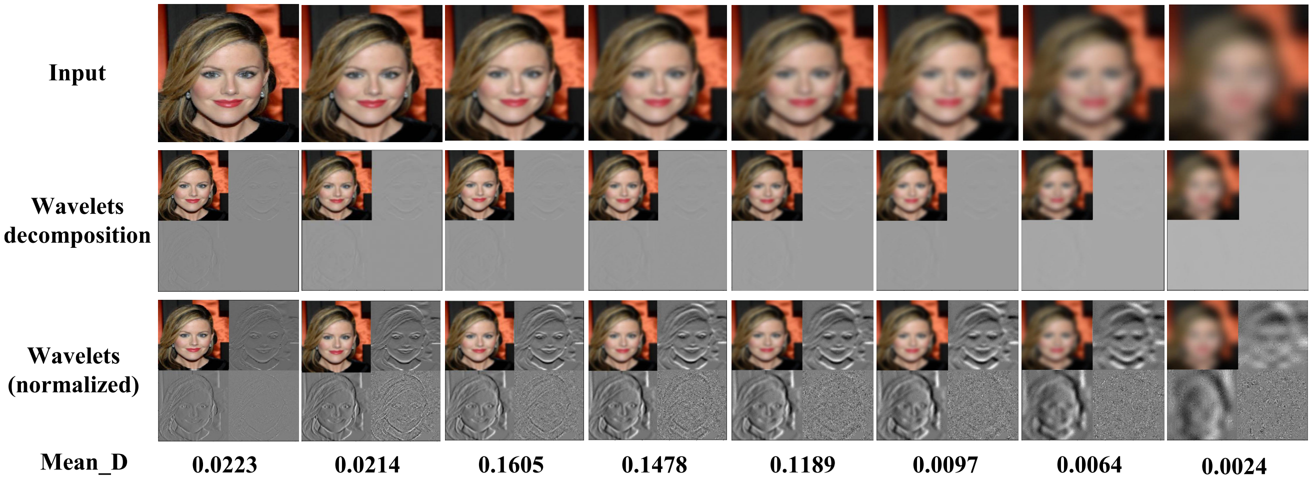


FIGURE 3. The correlation between detail wavelet coefficients and the image blur level. The first row are input images with different blur level. The second row are the results of wavelet decomposition, For the convenience of observation, the third row shows the normalized wavelet.

on each specific scale:

$$L_{ms-r} = \sum_{l=1}^{L-1} \left(\frac{1}{N_{I_{gt}^l}} ||M \odot (I_{gt}^l - f(x^l))||_1 + \alpha \frac{1}{N_{I_{gt}^l}} ||\bar{M} \odot (I_{gt}^l - f(x^l))||_1\right). \tag{8}$$

where x^l denotes the feature map of the l-th layer in the decoder, f is a 1×1 convolution that decodes x^l into an RGB image with corresponding resolution. I_{gt}^l denotes the ground truth downsampled to the same resolution as x^l

3) TEXTURE COMPONENT ADVERSARIAL LOSS

The wavelet transform has been proved an efficient and highly intuitive tool for representing and storing multiresolution images. It is capable of depicting the contextual and textural information of an image at various levels [34], inspiring this study to introduce wavelet transform to a deep image inpainting system. Our texture component adversarial loss complied with a 2-D discrete wavelet transform (DWT), decomposing the image into a sequence of wavelet coefficients of the identical resolution. Besides, a 1-level Haar wavelet was taken to decompose the images generated and the real images to calculate their approximate coefficients and detail coefficients. Approximate coefficients indicated the smooth transition of the brightness of the image color, which could merely reconstruct the overall outline of the object, rather than the fine details. In contrast, specific coefficients represent the high-frequency details of the image [35]. Thus, by learning the detailed coefficient distribution of real images in the wavelet domain, the texture component adversarial loss here attempted to optimize the texture details performance of the images generated. To elucidate the correlation between the detailed wavelet coefficients and the texture details of the natural image, 10,000 aligned face images were taken from CelebA randomly [15] and then resized to 256 x 256 pixels. Subsequently, bicubic interpolation was used to downsample these images to a range of scales, followed by upsampling using the identical method, through which several images with different blurring levels could be obtained. The average absolute value (MEAN_D) of the detailed wavelet coefficients for each blurring level was computed for intuitive comparison. In Fig. 3, the decomposition process is visualized; with the rise in the image blurring level, both visual and quantitative results of detail wavelet coefficients faded away, suggesting that it is essential to recover the high-frequency wavelets to enhance the texture details of the images generated.

We first perform a wavelet transform on the grayscale image to obtain the detailed coefficients in three directions (horizontal I_H, vertical I_V and diagonal I_D). After decomposition, the length and width of the image are reduced by half, and then the three high-frequency wavelet images are combined as input of the discriminator. We use the SN-PatchGAN [33] structure to apply GAN loss for each point in the output feature map, The output feature map of SN-PatchGAN is a 3-D feature of shape $\mathbb{R}^{h \times w \times c}$, where h, w, and c represents the height, width and number of channels respectively. To discriminate if the input is real or fake, our texture component adversarial loss for the discriminator is defined as:

$$L_D^{texture} = E_{I_{gt}^{hd}}[\log(1 - D^{sn}(I_{gt}^{hd})] + E_{I_{gen}^{hd}}[\log(1 + D^{sn}(I_{gen}^{hd}))]. \tag{9}$$

$$I^{hd} = I_H \oplus I_V \oplus I_D. \tag{10}$$

where D^{sn} represents the spectral-normalized discriminator, I_{gen}^{hd} represents the high-frequency detail part of the generated image, I_{gt}^{hd} denotes the high-frequency detail part of the ground truth. $\oplus$ is a concatenating operation. And texture component adversarial loss for the generator can be defined as:

$$L_G^{texture} = -E_{I_{gen}^{hd}}[D^{sn}(I_{gen}^{hd})]. \tag{11}$$

4) FINAL OBJECTIVE

With reconstruction loss, multiscale reconstruction loss, and adversarial loss, our overall loss function is defined as:

$$L = \lambda_r L_r + \lambda_{ms-r} L_{ms-r} + \lambda_t L_G^{texture} + \lambda_i L_G^{image}. \tag{12}$$

where λ_r, λ_{ms-r}, λ_t and λ_i are regularization parameters, specially, L_G^{image} is a adversarial loss on image level,it is defined similar to Eq 9 and Eq 11:

$$\begin{aligned} L_D^{image} &= E_{I_{gt}}[\log(1 - D^{sn}(I_{gt})] \\ &\quad + E_{I_{gen}}[\log(1 + D^{sn}(I_{gen})). \end{aligned} \tag{13}$$

$$L_G^{image} = -E_{I_{gen}}[D^{sn}(I_{gen})]. \tag{14}$$

where I_{gt} denotes the ground truth RGB images, and I_{gen} denotes the prediction results of the refinement inpainting module.

D. TRAINING DETAILS

Our model are implemented on TensorFlow, CUDNN v7.1, CUDA v9.2, and run on hardware with GPU TITAN V. The inpainting network G is trained using 256 × 256 images with a batchsize of eight. Our model is optimized by the Adam algorithm [22] with a learning rate of 1×10^{-4} and $\beta_1 = 0.5$, $\beta_2 = 0.999$. The parameters are set as $\lambda_r = 1$, $\lambda_{ms-r} = 1$, $\lambda_t = 1 \times 10^{-4}$, $\lambda_i = 1 \times 10^{-3}$. And it takes 0.54 seconds on GPU to complete a prediction for our full model. Training procedure is shown in Algorithm 1.

Algorithm 1 Training Procedure of our proposed framework

```
while Iteration times t < $T_{Train}$ do
    \\ Start the training of D.
    for i = 1,2 . . . , 5 do
        Sample a batch of images $I_{gt}$ from training data.
        Get a batch of binary masks M.
        Construct inputs $I_m = M \odot I_{gt}$.
        Obtain output $I_{gen} = I_m + G(I_m, M) \odot (1 - M)$.
        Transform $I_{gen}$ and $I_{gt}$ to wavelet domain.
        Compute the High frequency components $I_{gen}^{hd}$ and $I_{gt}^{hd}$ respectively.
        Update two critics with $I_{gen}^{hd}$. $I_{gt}^{hd}$ , $I_{gen}$ and $I_{gt}$.
    end for
    \\ Start the training of G
    Sample a batch of images $I_{gt}$ from training data.
    Get a batch of binary masks M.
    Update generator with reconstruction loss, multi-scale reconstruction loss and two adversarial critic losses.
end while
    return result
```

IV. EXPERIMENT RESULTS

Centering (128 × 128) and irregular masks were adopted for training and assessing the proposed network on two datasets: Places2 [15] and CelebA [16]. The results were qualitatively and quantitatively compared with the current state-of-the-art methods. To fairly evaluate, several experiments were performed as well to prove that our attention layers (IOA) and texture component adversarial loss are conducive to inpainting results. More prediction results of our experiment are presented in the supplementary material.

A. DATASETS

The Places2 [15] dataset covered images of 365 different types of scene environments collected from the natural world, and each class consisted of 5,000 pictures for training and 900 images per category in the testing set. Considering the computational cost, 40 categories totaling 200,000 images were selected for training and selected 1,200 images from the testing set (each category randomly chose 30 images) for testing. CelebA[16] refers to a large-scale face attributes dataset with 202,599 celebrity images, of which the images cover large pose variations and background clutter. 500 cropped and aligned images were randomly sampled for testing and the left images for training. The proposed method was compared with four recent works: CA: contextual attention [11], SH: shift-net [12], GL: globally and locally consistent image completion [8], and pconv: image inpainting for irregular holes using partial convolutions [25]. To achieve more convincing experimental results, the network structure in the above papers was not adopted, whereas their core algorithms were transplanted to the framework here for comparison with our proposed method.

B. QUALITATIVE COMPARISON

To illustrate the visual and semantic coherence of damaged regions, a center mask model was first train to compare with several similar methods. As shown in Fig. 4, the inpainting results of CA [11] and SH [12] exhibited poor performance in terms of the internal continuity of the restore areas, and the color was evidently different from the real image since these methods only focused on the attention score of known areas, overlooking the correlation between adjacent pixels within the missing areas. For irregular masks, the work of [25] was referenced, and their method was employed to train irregular masks. Irregular masks were augmented by performing random dilation, rotation and cropping, and then split into four categories in line with their sizes. Fig. 5 draws the comparison of our inpainting results with GL [8] and PConv [25]. When images were being restored with complex backgrounds, Gl [8] could not develop reasonable structures, and the results were often prone to color discrepancies. PConv [25] presented vague texture details, though it could generate semantically coherent results in some sense, while ours achieved obvious visual enhancement to plausible image structures and crisp textures.

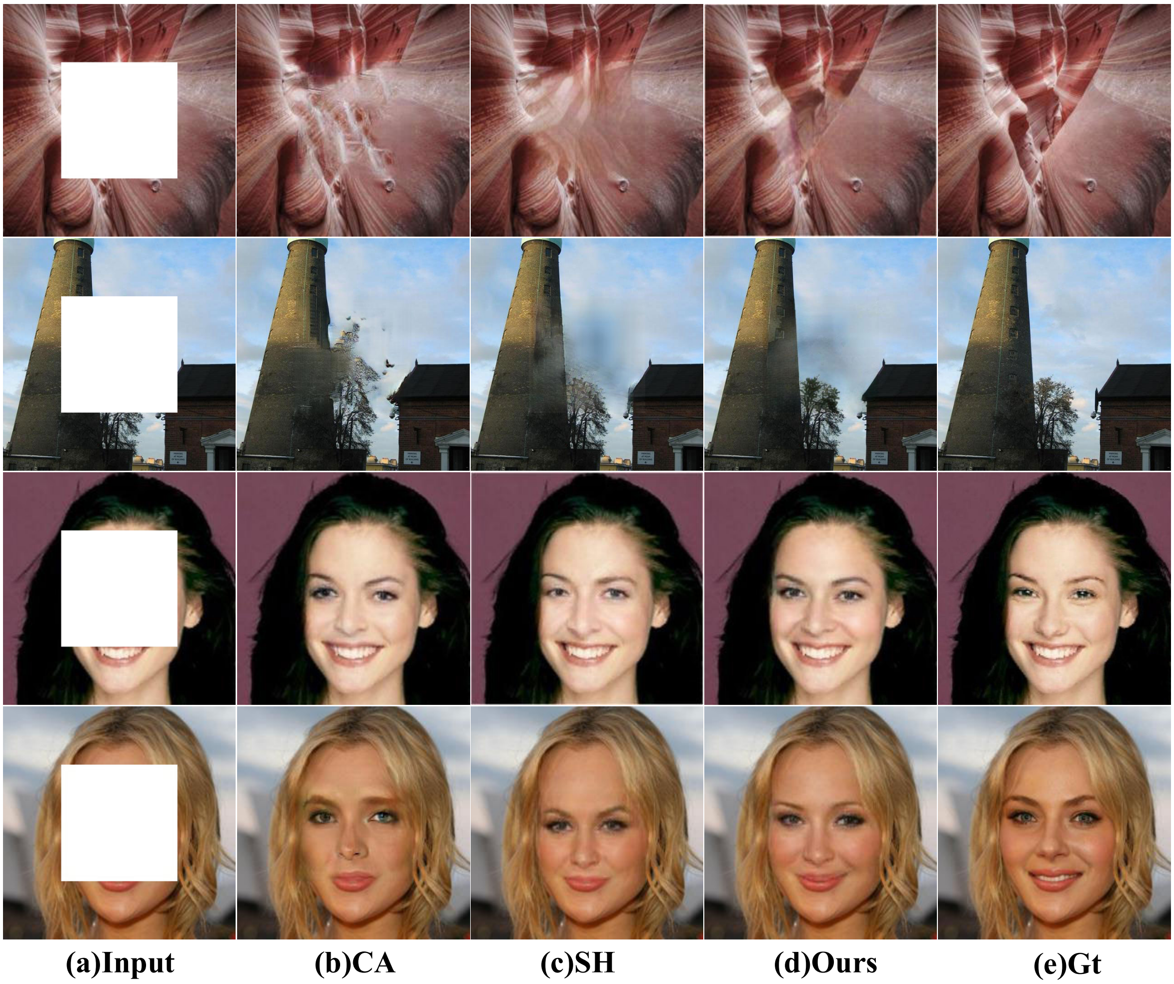


FIGURE 4. Qualitative comparisons in centering masks with other methods on test datasets with different characteristics. In each row, the first image is the input with a large mask in the center (128 × 128), and the left images from left to right are, the results generated by CA [11], SH [9], our model and ground truth, respectively. [Best viewed zoomed-in].

C. QUANTITATIVE COMPARISONS

Lacking effective quantitative evaluation metrics has always been a problem frequently facing by image generation tasks (e.g., image inpainting). The evaluation metrics of the GAN models do not apply to image inpainting since they place the primary emphasis on the semantic continuity between masked areas and known areas, rather than the ability to generate different classes of objects. Though some common quantitative evaluation metrics are not necessarily effective methods for image inpainting, for performing a fair quantitative comparison with existing methods, we still reported our evaluation results in terms of mean L_1 error, peak signal-to-noise ratio (PSNR) and structural similarity (SSIM) on the two testing datasets with both center rectangle masks and irregular masks for reference in Table 1, 2 and 3.

D. USER STUDY

Besides quantitative and qualitative comparisons, user studies were conducted as well on two types of experiments, and 20 volunteers were recruited to assess the performance of the proposed method on celebA testing images. They were not informed of any experimental information.

TABLE 1. Comparison results over Place2 and CelebA with centering masks between CA [11], SH [12], and Ours.− Lower is better.+ Higher is better.

Dataset	Method	$L^{1-}(\%)$	$PSNR^{+}$	$SSIM^{+}$
Place2	CA	8.76	20.01	0.735
	SH	7.71	21.56	0.746
	Ours	**7.68**	**22.34**	**0.781**
CelebA	CA	2.36	24.01	0.895
	SH	**1.67**	26.56	0.916
	Ours	1.73	**26.84**	**0.921**

During the first experiment, the main objective was to assess the naturalness and visual authenticity of the images generated reconstructed by the proposed method. 100 samples were randomly selected from the test images, inpainting work was performed after masking them, and then the completed image was mixed with the other 400 real images. Volunteers were only presented either prediction results or

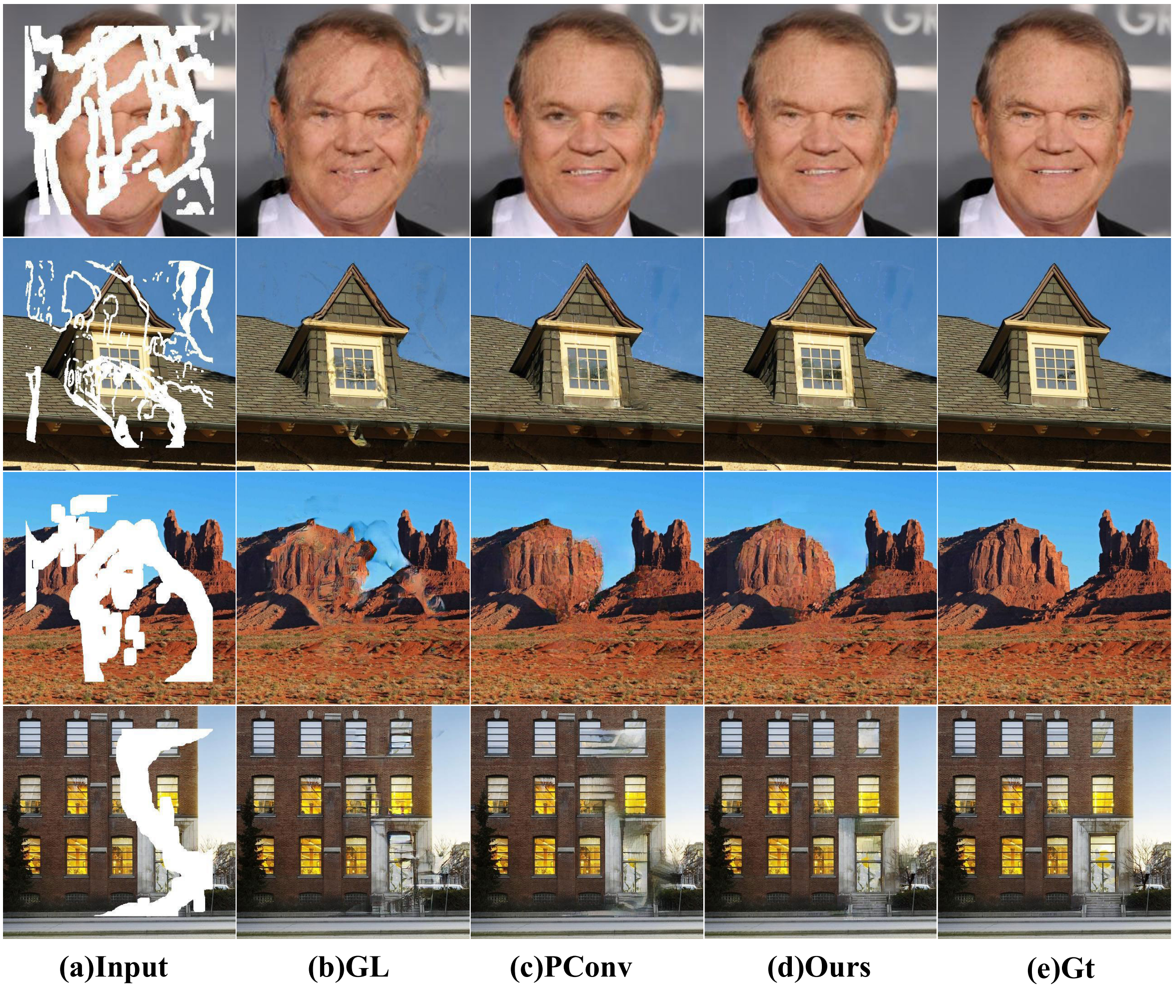


FIGURE 5. **Qualitative comparisons in irregular masks with other methods on test datasets with different characteristics. In each row, the first image is the input with a random irregular mask, and the left images from left to right are, the results generated by GL [8], SH [25], our model and ground truth, respectively. [Best viewed zoomed-in].**

TABLE 2. **Comparison results over Place2 with irregular masks between GL [8], Pcov [25], and Ours.– Lower is better.+ Higher is better.**

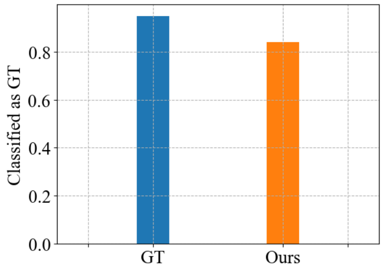
	Mask	GL	Pconv	Ours
$L^{1-}(\%)$	$10-20\%$	2.36	**1.09**	1.23
	$20-30\%$	4.13	**1.78**	2.01
	$30-40\%$	5.89	**3.74**	3.84
	$40-50\%$	7.7	**5.11**	5.36
$PSNR^{+}$	$10-20\%$	23.6	27.56	**28.63**
	$20-30\%$	20.59	24.88	**25.21**
	$30-40\%$	18.32	22.56	**23.11**
	$40-50\%$	17.23	20.45	**22.76**
$SSIM^{+}$	$10-20\%$	0.872	0.866	**0.913**
	$20-30\%$	0.775	0.784	**0.852**
	$30-40\%$	0.684	0.696	**0.785**
	$40-50\%$	0.615	0.581	**0.721**

TABLE 3. **Comparison results over CelebA with irregular mask between GL [8], Pcov [25], and Ours.– Lower is better.+ Higher is better.**

	Mask	GL	Pconv	Ours
$L^{1-}(\%)$	$10-20\%$	1.04	0.97	**0.85**
	$20-30\%$	1.65	1.23	**1.15**
	$30-40\%$	2.77	**2.32**	2.44
	$40-50\%$	4.11	**3.13**	3.26
$PSNR^{+}$	$10-20\%$	32.24	32.88	**34.11**
	$20-30\%$	30.14	31.51	**32.45**
	$30-40\%$	24.42	25.76	**26.13**
	$40-50\%$	22.93	23.89	**24.36**
$SSIM^{+}$	$10-20\%$	**0.972**	0.966	0.970
	$20-30\%$	0.945	0.954	**0.962**
	$30-40\%$	0.890	0.906	**0.913**
	$40-50\%$	0.815	0.861	**0.874**

the ground truth from the test dataset and given one second to determine whether the sample is a real image or completed image. The result revealed that 95% of the real images were correctly classified, and 84% of the completed images by the proposed method were categorized as real (Fig. 6), thereby thoroughly demonstrating that our results exhibit good visual authenticity.

During the second experiment, a horizontal comparison was drawn with other methods (Table 4). Each time, a pair of images (two images in respective pair) completed from



Method	Classified as GT
GT	95%
Ours	84%

FIGURE 6. Result of user study experiment 1 on our CelebA test images. The numbers represent the percentage of the images that are Classified as ground truth.

TABLE 4. Result of user study experiment 2 on our CelebA test images. Each entry is the percentage of cases in which the results of our approach were judged to be more realistic than those of another solution.

	Ours>CA	Ours>SH	Ours>GL	Ours>PConv
Percentage	90.2%	88.4%	93.2%	86.8%

the identical corrupted input by two different methods were presented to the volunteers without any information. The volunteers were supposed to choose the more natural and realistic image in each pair. This rule was followed to compare our model with CA [11], SH [12], GL [8] and Pconv [25], and all the completed results displayed the identical size of 256 × 256. By the entire experiment, all the images were shuffled to ensure unbiased comparison, and the observation time was not limited to free volunteers to spend much more time making rational judgments.

E. OBJECT REMOVAL

To evaluate our method is effective in practical application, we use our model trained on Place2 and CelebA to remove unwanted part of real images outside of these two datasets. the examples at first row in Fig.7 are from the scenes in life, we restore them with the model trained on Place2, we can see the inpainting regions integrate with the surrounding environment and has a coherent contextual semantics. The images at second row are human's face, we restore them with the model trained on CelebA. As is known to all, it is difficult to identify people when something on their face (like glasses and beard),thus we use our method to remove unwanted object on human's face and successfully restore a natural face which is similar to the original one.

F. ABLATION STUDY

To expound how IOA layer and texture components adversarial loss work for image inpainting, we study the effects of different parts in the image inpainting. Fig. 8 respectively shows the inpainting results obtained by our full framework, removing IOA layer, removing wavelet decomposition discriminator and the coarse-to-fine model without IOA and wavelet decomposition discriminator($D_{texture}$). From the results, we can see that without IOA and $D_{texture}$, the framework can hardly infer the consistent information and texture details with existing regions, and the generated part is illogical. Furthermore, if IOA layer is ignored, the effect of $D_{texture}$ is not obvious, only some unreasonable and disordered details are generated. As shown in the fourth column, if our model only contains IOA but not $D_{texture}$, the results still with boundary artifacts and unclear texture details although it shows coherent contextual semantics. With the help of both IOA and $D_{texture}$, our full model enjoys strong power to generate visually and semantically close images to the ground truth.

1) IOA LAYER VS. CONTEXTUAL ATTENTION LAYER [11]

To assess the effect of the IOA layer, the IOA layer was substituted with the contextual attention layer and a conventional 3 × 3 layer, and then these layers were trained on the identical framework to compare the differences between the three methods. Fig. 9 (b)reveals that when only the U-net structure was used without any attention layer, the masked part restored unnatural and blurry results accompanied by artifacts. Fig. 9 (c) presents the contextual attention layer inpainting results, suggesting that the semantic information in missing regions was inconsistent with the known regions, though it indeed enhanced the quality of the image. Compared with them, the proposed method exhibited better performance (Fig. 9 (d)).

2) EFFECT OF IOA LAYER AT DIFFERENT LAYERS

The comparison of the IOA layer embedded here and other methods demonstrated the superiority of our attention model. Fig. 9 shows that it achieved an outstanding result in preserving semantics. Note that the IOA layer appeared to be embedded in different upsampling layers of the decoder. The shallower the embedding position, the more the information can be obtained, whereas the much more the computational time will be required; if the embedding position is overly deep, some specific information conducive to inpainting may be lost. Given this, a reasonable embedding position should be selected to ensure that our model considers both the calculation costs and the inpainting performance. Fig 10 shows the result of the IOA layer at a range of upsampling layers. When the IOA was placed at a 64 × 64 size upsampling layer, our network yielded visually natural results (Fig. 10 (b)), whereas it took too much time on the calculation (i.e., 1.34 seconds per image). When the IOA was placed at a 16 × 16 size upsampling layer, our model exhibited high efficiency (i.e., 0.12 seconds per image), whereas it remarkably discounted the restoration results (Fig. 10 (c)). By performing the IOA layer in the 32 × 32 size upsampling layer, the efficiency (i.e., 0.54 seconds per image) and performance can be balanced at a high level by our model (Fig. 10 (d)).

3) WITH AND WITHOUT TEXTURE COMPONENT ADVERSARIAL LOSS

A complete model that does not involve texture component adversarial loss was first trained as shown in Fig. 11 (b). blurry images were generated without high-frequency information. Subsequently, the feature GAN loss [36] was added to

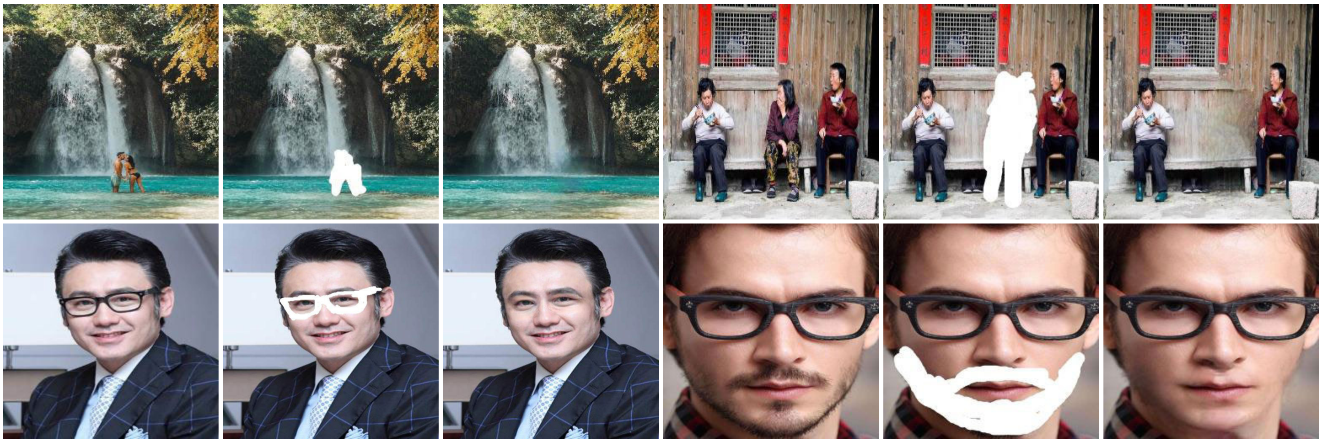


FIGURE 7. Comparison results of object removal case on real world images. Each example from left to right: ground truth, input mask, our result.

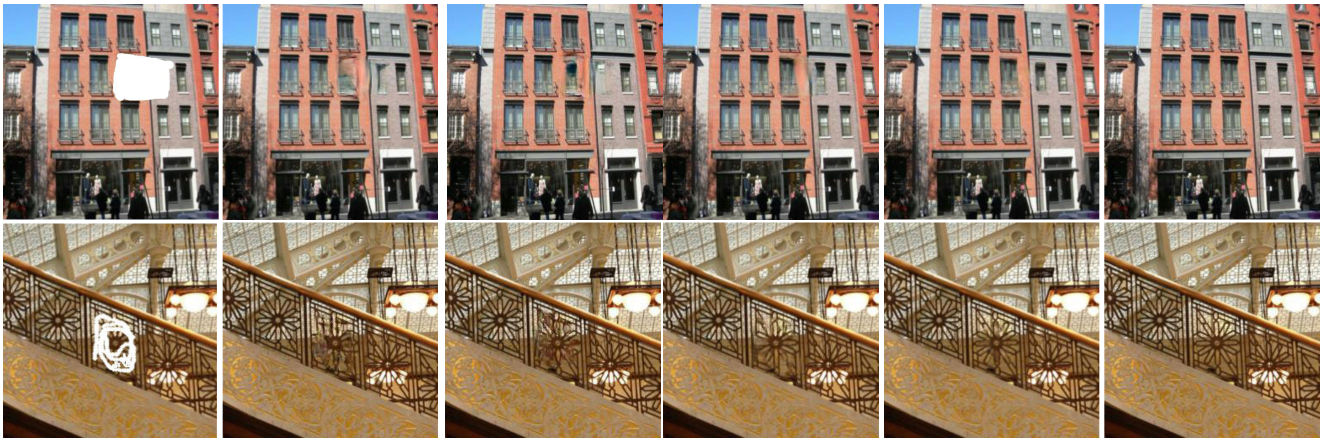


FIGURE 8. The effect of different components in our model. Each example from left to right: the input incomplete images, the results wihout IOA and $D_{texture}$, the results without IOA, the results without $D_{texture}$, the results with our full model and the ground truth.

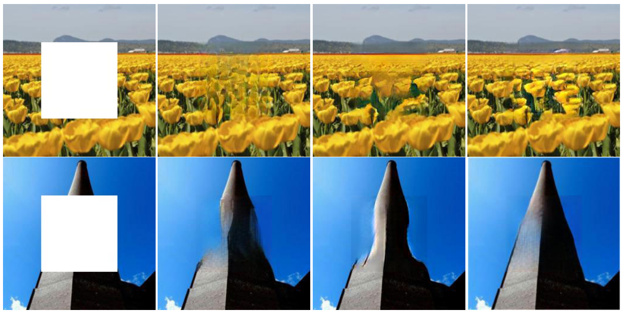


FIGURE 9. The comparison result of our IOA attention model with a non-attention model and contextual attention [11]. Each example from left to right: input image, non-attention model, Contextual attention, IOA attention.

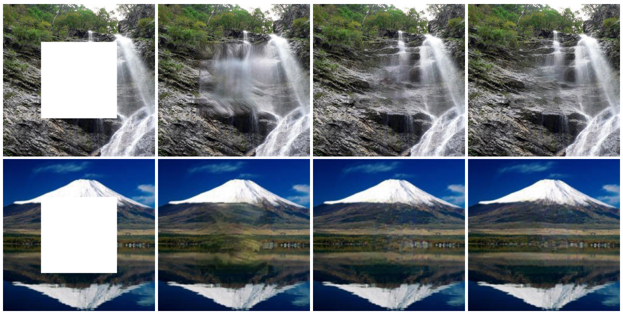


FIGURE 10. The effect of IOA layer on different layers.Each example from left to right: input image, 16 × 16 size upsampling layer, 32 × 32 size upsampling layer, 64 × 64 size upsampling layer.

highlight the high-frequency information, whereas the results still did not show sufficiently clear details (Fig. 11 (c)). Finally, feature GAN loss was substituted with texture component adversarial loss (Fig. 11 (d)), suggesting that the stone in the rectangle box displayed distinct edges and corners. It is therefore proved that our texture component adversarial loss successfully suppressed noisy high frequencies while generating perceptually plausible structured textures. Moreover, our results showed naturally synthesized sharp details without blurriness or high-frequency artifacts.

4) TWO-STAGE DEEP LEARNING METHOD

Two-stage network architecture is similar in spirits to residual learning [30] or deep supervision [38]. The rough inpainting

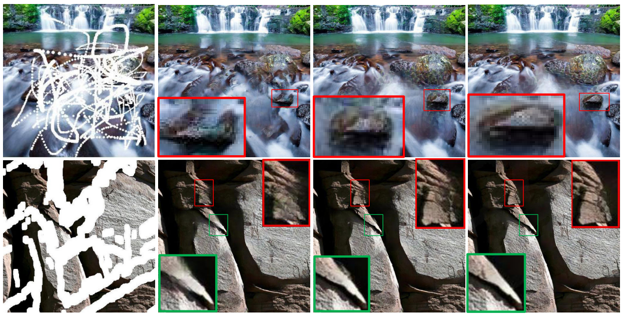


FIGURE 11. The comparison result of our texture component adversarial loss with non-texture component adversarial loss model and feature GAN loss. Each example from left to right: input image, non-texture component adversarial loss, feature GAN loss, with-texture component adversarial loss.

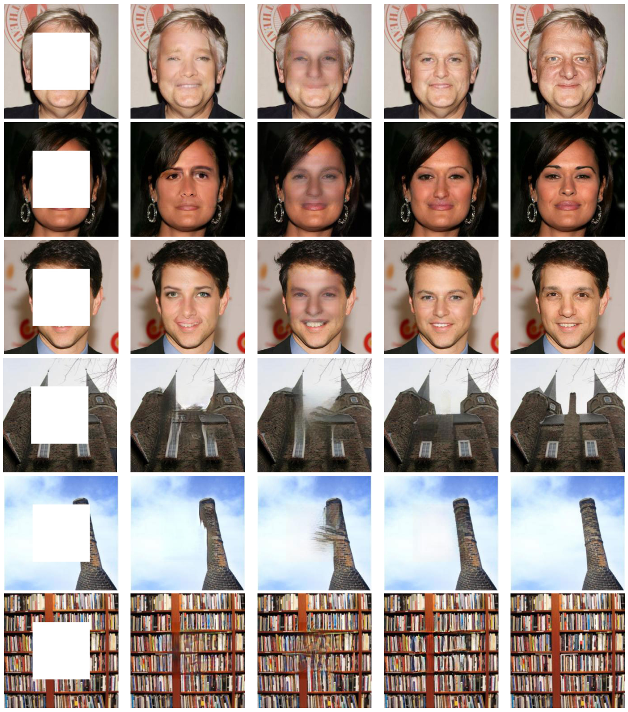


FIGURE 12. The comparison result of our method and previous two-stage deep learning methods. Each example from left to right: input image, CA[11], EC[24], our full model and ground truth.

module enlarge the receptive fields and provide prior knowledge for missing regions. Intuitively, the refinement network sees a more complete scene than the original image with missing regions, so its encoder can learn better feature representation than the rough network. Similar to CA [11] and EdgeConnect (EC) [24], our approach also uses a two-stage strategy. We conduct further experiment to evaluate the superiority of our proposed two-stage network. As shown in Fig. 12, when compared with previous two-stage deep learning methods, our approach performs better and generates visually pleasing results.

V. CONCLUSION

In this study, a novel inside-outside attention layer was presented via a coarse-to-fine network structure capable of achieving semantically-reasonable and visually-realistic results for image inpainting, since it can not only learn the relevance between generated and existing pixels, but also enhance the continuity among generated features in missing regions. Moreover, the texture component adversarial loss was introduced to learn the distribution of images in the wavelet domain to restore sharp and fine-detailed images. Experimental results verified the effectiveness and superiority of our proposed methods. In the subsequent work, we plan to enhance the effect of texture details reconstruction from two aspects. From a horizontal perspective, different wavelet bases exhibit unique features, thereby leading to a variety of wavelet transform results. We will explore the contribution of other types of wavelet bases (e.g., daubechies wavelet and symlets wavelet). Furthermore, from a vertical perspective, hopefully, we can introduce multilevel wavelet decomposition to determine whether it is conducive to enhancing the performance of the texture component discriminator.

REFERENCES

[1] A. Criminisi, P. Perez, and K. Toyama, "Region filling and object removal by exemplar-based image inpainting," *IEEE Trans. Image Process.*, vol. 13, no. 9, pp. 1200–1212, Sep. 2004.

[2] A. Levin, A. Zomet, S. Peleg, and Y. Weiss, "Seamless image stitching in the gradient domain," in *Proc. Eur. Conf. Comput. Vis (ECCV)*, Sep. 2004, pp. 377–389.

[3] D. Pathak, P. Krahenbuhl, J. Donahue, T. Darrell, and A. A. Efros, "Context encoders: Featurelearning by inpainting," in *Proc. IEEE Conf. Comput. Vis Pattern Recognit (CVPR)*, Jun. 2016, pp. 2536–2544.

[4] A. A. Efros and T. K. Leung, "Texture synthesis by non-parametric sampling," in *Proc. 7th IEEE Int. Conf. Comput. Vis. (ICCV)*, Sep. 1999, pp. 1033–1038.

[5] Z. Xu and J. Sun, "Image inpainting by patch propagation using patch sparsity," *IEEE Trans. Image Process.*, vol. 19, no. 5, pp. 1153–1165, May 2010.

[6] C. Barnes, E. Shechtman, A. Finkelstein, and D. Goldman, "PatchMatch:A randomized correspondence algorithm for structural image editing," *ACM Trans. Graph.*, vol. 28, no. 3, p. 24, 2009.

[7] J. Sun, L. Yuan, J. Jia, and H.-Y. Shum, "Image completion with structure propagation," *ACM Trans. Graph.*, vol. 24, no. 3, pp. 861–868, Jul. 2005.

[8] S. Iizuka, E. Simo-Serra, and H. Ishikawa, "Globally and locally consistent image completion," *ACM Trans. Graph.*, vol. 36, no. 4, pp. 1–14, Jul. 2017.

[9] Y. Li, S. Liu, J. Yang, and M.-H. Yang, "Generative face completion," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 236–244.

[10] R. Yeh, C. Chen, T. Y. Lim, M. Hasegawa-Johnson, and M. N. Do, "Semantic image inpainting with perceptual andcontextual losses," 2016, *arXiv:1607.07539*. [Online]. Available: https://arxiv.org/abs/1607.07539

[11] J. Yu, Z. Lin, J. Yang, X. Shen, X. Lu, and T. S. Huang, "Generative image inpainting with contextual attention," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 5505–5514.

[12] Z. Yan, X. Li, M. Li, W. Zuo, and S. Shan, "Shift-Net: Image inpaintingvia deep feature rearrangement," in *Proc. Eur. Conf. Comput. Vis (ECCV).*, Sep. 2018, pp. 377–389

[13] Y. Zeng, J. Fu, H. Fu, and B. Guo, "Learning pyramid-context encoder network for high-quality image inpainting," in *Proc. IEEE Conf. Comput. Vis Pattern Recognit (CVPR)*, Jun. 2019, pp. 1486–1494.

[14] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convo-lutional networks for biomedical image segmentation," in *Proc. Conf. Int. Conf. Med. Image Comput. Comput.-Assist. Intervent*. Springer, Oct. 2015, pp. 234–241.

[15] B. Zhou, A. Lapedriza, A. Khosla, A. Oliva, and A. Torralba, "Places: A 10million image database for scene recognition," *IEEE Trans. Pattern Anal.Mach. Intell.*, vol. 40, no. 6, pp. 1452–1464, Jun. 2018.

[16] Z. Liu, P. Luo, X. Wang, and X. Tang, "Deep learning face attributesin the wild," in *Proc. IEEE Int. Conf. Comput. Vis. (CVPR)*, Dec. 2015, pp. 3730–3738.

[17] M. Bertalmio, G. Sapiro, V. Caselles, and C. Ballester, "Image inpainting," in *Proc. 27th Annu. Conf. Comput. Graph. Interact. Techn.*, 2000, pp. 417–424.

[18] M. Bertalmio, L. Vese, G. Sapiro, and S. Osher, "Simultaneous structure and texture image inpainting," *IEEE Trans. Image Process.*, vol. 12, no. 8, pp. 882–889, Aug. 2003.

[19] T. F. Chan and J. Shen, "Nontexture inpainting by curvature-driven diffusions," *J. Vis. Commun. Image Represent*, vol. 12, no. 4, pp. 436–449, Dec. 2001.

[20] D. Zoran and Y. Weiss, "From learning models of natural image patches to whole image restoration," in *Proc. Int. Conf. Comput. Vis.*, Nov. 2011, pp. 479–486.

[21] J. Hays and A. A. Efros, "Scene completion using millions of photographs," *ACM Trans. Graph.*, vol. 26, no. 3, pp. 786–794, Jul. 2007.

[22] A. B. L. Larsen, S. K. Sonderby, H. Larochelle, and O. Winther, "Autoencoding beyond pixels using a learned similarity metric," in *Proc. 33rd Int. Conf. Int. Conf. Mach. Learn*, 2016, pp. 1558–1566.

[23] D. P. Kingma and M. Welling, "Auto-encoding variational Bayes," 2013, *arXiv:1312.6114*. [Online]. Available: http://arxiv.org/abs/1312.6114

[24] N. Kamyar, E. Ng, T. Joseph, F. Z. Qureshi, and M. Ebrahimi, "EdgeConnect: Generative image inpainting with adversarial edge learning," 2018, *arXiv:1901.00212*. [Online]. Available: https://arxiv.org/abs/1901.00212

[25] G. Liu, F. A. Reda, K. J. Shih, T. Wang, A. Tao, and B. Catanzaro, "Image inpainting for irregular holes using partial convolutions," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, Sep. 2018, pp. 85–100.

[26] J. Johnson, A. Alahi, and L. Fei-Fei, "Perceptual losses for real-time style transfer and super-resolution," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, Oct. 2016, pp. 694–711.

[27] M. S. M. Sajjadi, B. Scholkopf, and M. Hirsch, "EnhanceNet: Single image super-resolution through automated texture synthesis," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 4501–4510.

[28] J.-Y. Zhu, T. Park, P. Isola, and A. A. Efros, "Unpaired Image-to-Image translation using cycle-consistent adversarial networks," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 410–4001.

[29] T. Karras, T. Aila, S. Laine, and J. Lehtinen, "Progressive growing of gans for improved quality, stability, and variation," 2017, *arXiv:1710.10196*. [Online]. Available: https://arxiv.org/abs/1710.10196

[30] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 770–778.

[31] J. You, J. Wang, and Z. Liang, "Range condition and ML-EM checkerboard artifacts," *IEEE Trans. Nucl. Sci.*, vol. 54, no. 5, pp. 1696–1702, Oct. 2007.

[32] C. Dong, C. C. Loy, K. He, and X. Tang, "Image super-resolution using deep convolutional networks," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 38, no. 2, pp. 295–307, Feb. 2016.

[33] J. Yu, Z. Lin, J. Yang, X. Shen, X. Lu, and T. Huang, "Free-form image inpainting with gated convolution," 2018, *arXiv:1806.03589*. [Online]. Available: http://arxiv.org/abs/1806.03589

[34] S. Mallat, "Wavelets for a vision," *Proc. IEEE*, vol. 84, no. 4, pp. 604–614, Apr. 1996.

[35] H. Demirel and G. Anbarjafari, "IMAGE resolution enhancement by using discrete and stationary wavelet decomposition," *IEEE Trans. Image Process.*, vol. 20, no. 5, pp. 1458–1460, May 2011.

[36] S. Park, H. S. Chao, K. Hong, and S. Lee, "Srfeat: Single image super-resolution with feature discrimination," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, Sep. 2018, pp. 455–471.

[37] D. P. Kingma and J. L. Ba, "Adam: A method for stochastic optimization," 2014, *arXiv:1412.6980*. [Online]. Available: https://arxiv.org/abs/1412.6980

[38] C.-Y. Lee, S. Xie, P. Gallagher, Z. Zhang, and Z. Tu, "Deeply-supervised nets," 2014, *arXiv:1409.5185*. [Online]. Available: https://arxiv.org/abs/1409.5185

XINGCHEN HE received the B.S. degree from the School of Information Science and Technology, Chengdu University of Technology, Chengdu, China, in 2016, where he is currently pursuing the M.S. degree with the School of Information Science and Technology. His current research interests include computer vision and image processing.

XUDONG CUI received the B.S. degree from the Beijing University of Technology, Beijing, China, the M.S. degree from the China Academy of Engineering Physics, Mianyang, China, and the Ph.D. degree from the Swiss Federal Institute of Technology, Zurich, Switzerland. He is currently a Researcher with the China Academy of Engineering Physics. His current research interests include image processing, nano-optics, electromagnetic fields, and material physics and chemistry.

QILONG LI received the B.S. degree from the School of Information Science and Technology, Chengdu University of Technology, Chengdu, China, in 2016, where he is currently pursuing the M.S. degree with the School of Information Science and Technology. His current research interests include computer vision and machine learning.