J. Vis. Commun. Image R. 91 (2023) 103777

Contents lists available at ScienceDirect

Journal of Visual Communication and Image Representation

journal homepage: www.elsevier.com/locate/jvci

Full Length Article

Semantic prior-driven fused contextual transformation network for image inpainting☆

Haiyan Li [a], Yingqing Song [a], Haijiang Li [b,*], Zhengyu Wang [a]

[a] School of Information, Yunnan University, Kunming 650050, Yunnan, China
[b] Yunnan Communications Investment and Construction Group Co, Ltd., Kunming 650050, Yunnan, China

ARTICLE INFO

Keywords:
Image inpainting
Semantic prior generator
Fused contextual transformation
Aggregated semantic attention-aware
Discriminator

ABSTRACT

Recent advances in image inpainting have achieved impressive performance for generating plausible visual details on small regular image defects or simple backgrounds. However, current solution suffers from the lack of semantic priors for the image and the inability to deduce the image content from distant background, leading to distorted structures and artifacts in the results when inpainting large random irregular complicated images. To address these problems, a semantic prior-driven fused contextual transformation network for image inpainting is proposed as a promise solution. First, the semantic prior generator is put forward to map the semantic features of ground truth images and the low-level features of broken images to semantic priors. Subsequently, an image split-transform-aggregated strategy, named fusion context transformation block, is presented to infer rich multi-scale remote texture features and thus to improve the restored image finesse. Thereafter, an aggregated semantic attention-aware module, consisting of spatially adaptive normalization and enhanced spatial attention is designed to aggregate semantic priors and multi-scale texture features into the decoder to restore reasonable structure. Finally, the mask guided discriminator is developed to effectively discriminate between real and false pixels in the output image to improve the capability of the discriminator and hence to reduce the probability of artifacts containing in the output image. Comprehensive experimental results on CelebA-HQ, Paris Street View, and Places2 datasets demonstrate the superiority of the proposed network over the state-of-the-arts, whose PSNR, SSIM and MAE are improved about 20 %, 12.6 %, and 42 % gains, respectively.

1. Introduction

Image inpainting [1] aims to reconstruct the missing regions of damaged images with realistic contents, which is an important task in computer vision with many practical applications, such as old photo restoration [2], photo editing [3] and de-captioning [4]. Despite the significant advances of image inpainting made in recent years, restoring images with reasonable content and clear textures remains challenging especially when images have intricate background and large random-form corrupted regions. Traditional inpainting algorithms try to hunt for patches from the background region to recovery the missing regions [4,5]. Although these traditional approaches perform well on simple cases, they are not adequate for painting complicated scenes due to a lack of semantic understanding of the image.

For cases involving complicated or non-repetitive structures, how to make a full use of information in known regions to infer the semantic features of damaged regions is a key issue, and therefore research efforts in recent years have shifted from traditional approaches to information-driven deep CNN-based methods. These methods can be roughly classified into two categories. The first class of methods utilize the constraints on the semantic consistency of generated content with contextual information [6–11]. However, as the corrupted regions become large and the distances between unknown and known pixels increase, the constraints for the hole center loosen and these correlations are weakened because only pixels within related regions have strong associations. The second type is to predict the corrupted regions condition from known regions in the original image, thereafter, the detailed texture information is then inferred under the guidance of the contextual

☆ This paper has been recommended for acceptance by Chuan Qin.Peer review under responsibility of All the manuscript should have the following footnote: This paper has been recommended for acceptance by Chuan Qin. Note: Associate editor information can be fetched from EES system. They will have the role "Editorial Board".

* Corresponding author.
E-mail address: li_cannie@163.com (H. Li).

https://doi.org/10.1016/j.jvcir.2023.103777
Received 18 October 2022; Received in revised form 8 December 2022; Accepted 27 January 2023
Available online 1 February 2023

information [6,12].These methods can handle irregular holes appropriately, but the generated contents still encounter problems of boundary artifacts and semantic fault for large missing areas. In summary, the above methods cannot apply image semantic priors for inpainting tasks, especially for images with complex backgrounds or non-repetitive patterns.

In order to obtain fine restored contents for large and free-form corrupted regions, it is essential to make reasonable content inferences about the missing areas. Contextual attention modules [9,13] were proposed to establish the connection between aperture regions and image contexts by patch-style matching. However, patch-style matching often leads to structural distortion due to the non-repetitive patterns [14]. Another line of works stacked serialized dilated convolution layers to capture remote contents. However, as discussed extensively in earlier works [9,15], serialized dilated convolutions preferred to encode features of predefined latticed patterns rather than capturing rich patterns of interest for content reasoning.

To reduce the presence of artifacts in the output image and to facilitate texture fine-grained synthesis, a joint training consisting of global and local discriminators was proposed by Iizuka et al [30]. The global discriminator focused on the whole image and the local discriminator concentrated on the local consistency. However, the local discriminator could only handle the missing regions of fixed shapes due to the fact that a fully connected layer was utilized in the network. To solve this issue, Yu et al. [31] inherited the discriminator from PatchGAN due to its tremendous success in image translation [7,9,11], which The discriminator was designed to discriminate patches of ground truth images from those of restored results. In addition, they applied spectral normalization to the discriminator for each layer to stabilize the training of GAN [32]. However, Patch GAN-based models usually overlooked the notion that patches outside the missing regions were common to both ground truth images and restored images, and pushing the discriminator blindly distinguish these similar patches could cripple the discriminator.

In summary, most of the existing image inpainting methods face the following problems. Firstly, most CNN-based information-driven approaches cannot acquire valid semantic priors for images, which leads to the inability to generate realistic structural and semantically sound images. Secondly, large damaged areas cannot be inferred from distant known backgrounds to enrich the image content, and relatively few complete pixels exist in the image, thus making it difficult to be repaired and reducing the clarity of the restored image. Finally, the current common discriminator predicts all pixels of an image as false, ignoring the fact that the known pixels are true, which results in a poorly detailed restored image. To address these issues, a semantic prior-based fusion context transformation network for image inpainting is proposed, which can effectively withdraw the structural features and multi-scale texture features from images and tackle the problems of structural distortion and artifacts present in the restored results for large irregularly corrupted regions. The main contributions of the network are summarized as follows:

- The semantic prior generator maps the semantic features and the low-level features into semantic priors, which is processed by residual blocks to obtain structural features, and thus the output images contains abundant reasonable details.
- A fused contextual transformation block (FCTB) is conceived to capture the multi-scale texture features of images. The input image is separated-transformed-aggregated by the FCTB to carry out regional affinity and learn to gain image multi-scale contextual information.
- A spatially adaptive normalization and enhanced spatial attention semantic aggregation structure, called the aggregate semantic attention-aware (ASA) module, is proposed to adaptively incorporate the structural features and the multiscale texture features to handle free-form corruption and generate coherent structures as well as sharp textures.
- A mask guided discriminator is designed to separate the restored pixels from the known area to improve the power of the discriminator, while decreasing the possibility of the existence of artifacts.

The rest of this paper is organized as follows. Section 2 outlines the work related to image inpainting and content reasoning. The proposed method is described in detail in Section 3. A series of ablation experiments are conducted and the proposed model is compared with state-of-the-art to evaluate the inpainting performance of our approach in Section4. The conclusion is drawn in Section 5.

2. Related works

2.1. Image inpainting

Traditional inpainting approaches can be grouped into two main categories, diffusion-based and patch-based. Diffusion-based methods modeled diffusion processes with different operators to diffuse background information into the corrupted regions by using the information available in the original image [20]. Therefore, it cannot obtain a reasonable structure or recover large corrupted regions. Barnes et al. [4] proposed the Patch-Match algorithm, which iteratively searched for the most appropriate patch to synthetize the content of the missing part from the boundary of the corrupted area. However, the approach did not capture high-level semantic information properly, thus leading to the results of ambiguous details. Thereafter, Criminisi et al. [21] proposed a patch priority method to populate the structure, however, it had troubles of disillusioning plausible content in semantic restoration due to the lack of strong reasoning for missing contents and textures in sophisticated scenes.

Deep convolutional networks [22] have been frequently employed, especially generative adversarial networks (GAN) [23] for semantic inpainting algorithms. Context Encoder [24] employed conditional GAN [25] for the first time for image inpainting to generate the content of the damaged by making reasonable assumptions about the missing regions, and showing promising for image inpainting tasks. However, the semantic priors were not learned, and the assumptions on the missing regions could not handle random irregular holes. To acquire semantic priors, Liu et al. [6] proposed partial convolution (Pconv) to overcome the shortcomings that standard convolution could not capture the semantic features or modify the mask renewal mechanism to gradually learn the semantic information to repair irregularly broken regions. Nevertheless, the extracted semantic prior is so sparse as to be invalid, and thus the generated images have low global network consistency. To exclude the effect of invalid semantic priors during the feature extraction process, Yu et al. [25] considered that the features of known regions and the inferred regions should first be region normalized (RN) separately before learning the semantic priors separately. Although RN is put forward to learn image features, it usually fails to recover reasonable content in complex scenarios due to the problem of neglecting semantic priors, leading to inadequate access to image contextual structure. Zhang et al.[16] proposed the SPL framework, which learned the semantics priors using image encoders and multiclassification models to improve inpainting performance. However, SPL still suffered from the problem that the semantic features failed to interact with the image structure. Therefore, it could not generate structurally coherent images with complex contents and could be implemented only for small corrupted regions. To address the problem of structural distortion, Cao et al. [45] introduced a pre-trained mask-based autoencoder into the inpainting model, which was able to obtain rich information prior, enhance the inpainting performance and improve the connectivity of the network from local features to overall consistency. However, this method extracted insufficient low-level features, could not reason about the long-range action relationship between known and missing regions. Therefore, the restored images did not satisfy human visual realism.

Previous studies have shown that GAN-based networks are effective

for image inpainting. However, it is still challenging because the generators of most models cannot extract enough valid semantic features and apply them to effectively solve the structure distortion issues. To solve this issue, the semantic prior generator is proposed to map and act on the semantic features and low-level features extracted from different images as structural features, enabling the restored structure to be more coherent and closer to human vision.

2.2. Content reasoning

Inspired by recent works of style transfer, a growing number of deep inpainting models exploit a perceptual loss [28] and a style loss [29] for the synthesis of fine-grained textures. Typically, the joint optimization of a perceptual loss and a style loss aims at minimizing the perceptual distance between the deep features of the generated results and the initial images. Although promising results have emerged, it still needs to reason about the abundant image content from distant backgrounds in order to improve texture granularity [6]. Recently, several methods have been introduced for image inpainting by reasoning about distant scenes to obtain image content. Dolhansky et al. [26] proved the importance of paradigm information for inpainting. Although the method is able to achieve both obvious and accurate inpainting, however, it exclusively fills in the missing eye region in human frontal images, and thus the paradigm information cannot be inferred at a remote distance, causing it to fail in practice for complex scenes. AOT-Net [27] is designed to aggregate contextual transform modules for content inference from distant backgrounds, which incorporates multi-scale information features to extract image textures for a rational structure. However, this solution has two restrictions: 1) it lacks of sufficient prior information to assist the inference coming from the content to make the generated image texture complete, and 2) this approach cannot handle disrupted regions with arbitrary shapes due to the loss of structural information caused by focusing too much on ture information during the reasoning process. To reason about the long-term inferred content-semantic relationship of images, Zheng et al. [46] proposed cascaded modulated GAN (CM-GAN) network, containing an encoder with Fourier convolution blocks and a dual-stream decoder, which the encoder extracted multi-scale feature representations from the damaged image and cascaded global spatial modulation blocks at each scale. In each decoding block, coarse semantic-aware structure synthesis was first performed using global modulation, followed by spatial modulation to further adjust the feature map in a spatially adaptive manner. CM-GAN is capable of generating rich texture details, but suffers from structural incoherence when dealing with large scale images of arbitrary shapes.

In summary, most of the above methods cannot reason about the unknown content from distant backgrounds and do not combine enough semantic priors. In this paper, FCTB is designed to acquire the multi-scale texture features from images and combine them with semantic priors in order to handle randomly shaped broken regions and output images with clear textures and rich details. In the following, the proposed approach is compared with our baseline network SPL and the SPN [41], which SPN was proposed by the authors of SPL as further research.

First, SPL learns the complete semantic prior of the damaged visual elements using a multi-label classification model pre-trained on an open image dataset [34]. SPL constructs a semantic learner to establish the mapping relationship between visual regions and semantic priors. Based on this, the SPL constructs the improved model SPN[41], which transports the learned semantic priors to a multiscale feature pyramid to obtain the multiscale semantic priors and designs an image generator consisting of a variational inference module to adaptively and incrementally improve low-level visual representations at multiple scales using a (stochastic) prior pyramid. The proposed semantic prior generator is different from that of the SPL on three aspects. First, the obtained semantic priors are processed by the residual blocks to generate structural features. Second, the proposed network integrates enhanced spatial attention (ESA) and fused content transformation blocks (FCTB) into the encoder to obtain the multi-scale texture features of images. In addition, SPN feeds the SPADE-Relu-Conv unit in parallel with the SPADE module, while the ASA module of the proposed network not only implement this operation but also combines ESA to adaptively fuse structural features and multi-scale texture features. Finally, the discriminator employed in both SPN and SPL networks is a Patch discriminator, while the proposed network utilizes the proposed mask guided discriminator, which can better distinguish between true and false pixels in the output image and improve the capability of the discriminator, thus reducing the probability of containing artifacts in the output image. With these difference from SPN, the proposed network demonstrates a significant improvement in structural coherence and artifacts decrease in the inpainting results compared to SPL.

3. Approach

The overall architecture of the proposed inpainting system is illustrated in Fig. 1. The semantic prior generator leverages the image low-level features F_1 obtained from the incomplete image I_{in}, the corresponding mask M and the semantic features F_p extracted from the multi-classification labeling model K to get the structural information F_s. On the other side, the FCTB modules split, transform, and aggregate image features to capture the rich remote multi-scale texture features. Subsequently, the ASA module fuses the structural features and the multi-scale texture features to provide to the decoder for generating the restored results I_{output}. Finally, the mask guided discriminator helps generators to determine whether the output image I_{output} is true or false.

3.1. Semantic prior generator

The objective of the semantic prior generator, shown in Fig. 1, is to learn the full semantic prior of the damaged visual components under the supervision of the pretrained deep neural network. To make the encoder aware of the global semantic feature of the corrupted images, a multi-label classification model K with features F_p is leveraged as a supervision for learning the semantic prior. The parameters of this model are trained on an open image dataset [34] and symmetric loss (ASL) [42] is utilized as supervision of model K.

Firstly, the input image is up-sampled to acquire rich image features. In particular, the ground truth images are up sampled $I_{gt} \in \mathbb{R}^{3\times H\times W}$ to $I_{s1} \in \mathbb{R}^{3\times 2H\times 2W}$.The incomplete images I_{in} and the counterpart mask M are unsampled in the same way to form the inputs I_{n1} and M_1 for the encoder E_1.After that, I_{s1} is served as the input of the multi classification labeling model K, and the feature mapping $F_p \in \mathbb{R}^{d\times\frac{H}{4}\times\frac{W}{4}}$ is withdrawn as the supervision of the semantic prior:

$$F_p = K(I_{s1}) \tag{1}$$

Furthermore, the encoder E_1 containing four down sampling layers produces the low-level features F_1 by utilizing the enlarged image and the mask as input:

$$F_1 = E_1(I_{n1}, M_1) \tag{2}$$

where $I_{n1} \in \mathbb{R}^{3\times 2H\times 2W}$, $M_1 \in \mathbb{R}^{1\times 2H\times 2W}$, $F_1 \in \mathbb{R}^{c\times\frac{H}{4}\times\frac{W}{4}}$.Thereafter, we exploit the 1×1 convolutional layer to allow F_1 to adapt the representation of the pretext task:

$F_1^{'} = conv_{1\times 1}(F_1)$(3).

where $F_1^{'} \in R^{d\times\frac{H}{4}\times\frac{W}{4}}$. Next, a semantic learner [16] for l_1 reconstruction loss with missing region constraints is combined to obtain the semantic prior information of the damaged image:

$$F_{prior} = \|(F_p - F_1^{'}) \odot (1 + \alpha Ms)\|_1 \tag{4}$$

where $\odot$ denotes the Hadamard product operator. α indicates the additional weight of the missing region. Ms indicates the adjustment mask with the same space size as F_p. Finally, the semantic prior F_{prior} is

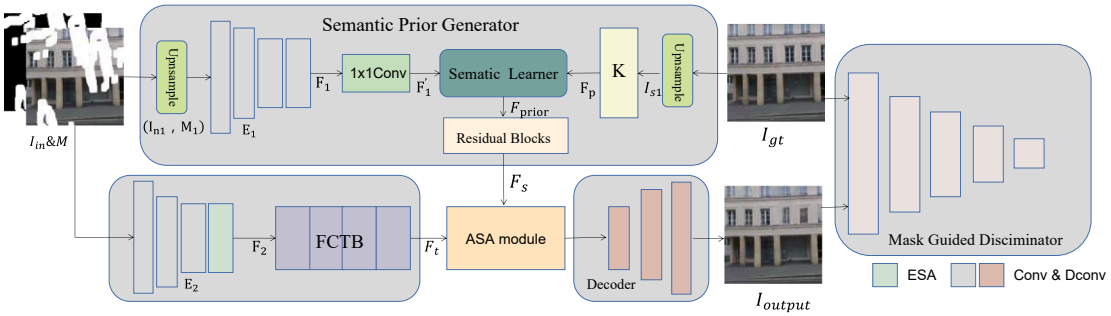

Fig. 1. The overall architecture of semantic prior-based fused contextual transformation network.

fed into the Residual module to gain structural features F_s. By doing the above, the proposed network can retrieve beneficial information for inpainting large random damage and filter out the task-irrelevant components in F_p.

3.2. Fused contextual transformation block

For the lower branch shown in Fig. 1, we intend to extract rich texture features from the undamaged content to form the conditional input, which the features play an essential role in recovering distinctive local textures. Thereafter, the network leverages the encoder E_2, the enhanced spatial attention (ESA) [36] and the fused content transformation blocks (FCTB) extract image content from the corrupted input image and perform the transformation fusion to obtain the multi-scale texture features F_t.

FCTB is presented as a simple, highly modular module. As shown in Fig. 2, FCTB transforms the input features as two components, which the first component computes the mapping output x_{o1} of x_i using standard convolution and sigmoid operations:

$$x_{o1} = sigmoid(conv_{3\times3}(x_i)) \tag{5}$$

The second branch adopts the split-transformation-merge strategy by three steps [27] and the input part x_i is divided into four subsets, denoted by x_j. Different transformations of the input features x_j are applied for each subset by employing different dilation rates. Using a larger dilation rate allows the sub-kernel to "see" a larger region of the input image, while a sub-kernel with a small dilation rate concentrates on the local patterns of a smaller receptive domain. The contextual transformations from different receptive fields are ultimately integrated by concatenation, and thereafter feature fusion is performed using standard convolution to yield the output feature x_{o2}:

$$x_{02} = conv_3(\sum_{j=1}^{4} conv(x_i)) \tag{6}$$

Subsequently, FCTB weights and sums the output features x_{o1} and x_{o2} to express the final output x_o as:

$$x_o = x_{o1} \odot x_{o2} + x_i \odot (1 - x_{o1}) \tag{7}$$

In particular, feature reasoning for a large free-form missing region is one of the grand challenges for image inpainting [11,33]. For one thing, to ensure the coherent structure and clear texture with surrounding contexts, inpainting models need to propagate information from distant contexts to the missing regions. For another, as the objects in complex scenes have various scales and angles of view, capturing as rich as possible patterns of interest is important for feature reasoning. Therefore, the FCTB is designed to capture both informative distant image contents and rich patterns of interest for feature reasoning in our proposed method.

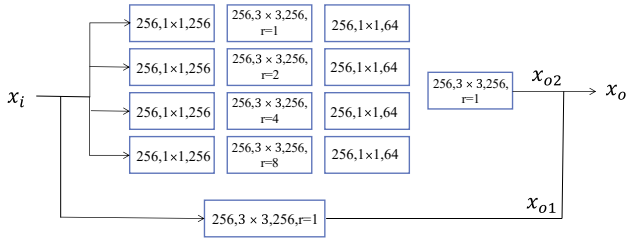

Fig. 2. The description of fused contextual transformation block.

3.3. Aggregated semantic attention-aware module

Since multiscale texture features F_t and structural features F_s have different focus on the image, thus it is not possible to fuse the two features directly. In order to effectively fuse the multiscale texture features and the structural features, the aggregated semantic attention-aware (ASA) module is designed, which focuses on ESA and SPADE [35] components to spatially feed structural features into the decoder and be capable of adaptively modifying the multi-scale texture features. The structure of ASA module is illustrated in Fig. 3, where the multiscale texture features F_t are first normalized with non-parametric instance normalization (IN). Thereafter, two distinct sets of parameters are learned from the structural features F_s to perform the spatial pixel affine transformation on the multiscale texture feature F_t:

$$[\gamma, \beta] = ASA[F_s], \tag{8}$$

$$F_t^{'} = r \cdot IN(F_t) + \beta, \tag{9}$$

where $\gamma, \beta, F_t^{'} \in \mathbb{R}^{c\times\frac{H}{4}\times\frac{W}{4}}$. Finally, the output feature $F_t^{'}$ is fed into the decoder to yield the output image I_{output}:

$$I_{output} = D(F_t^{'}). \tag{10}$$

The ASA residual block is designed as parallel branches by integrating SPADE and ESA modules to adaptively fuse the structural

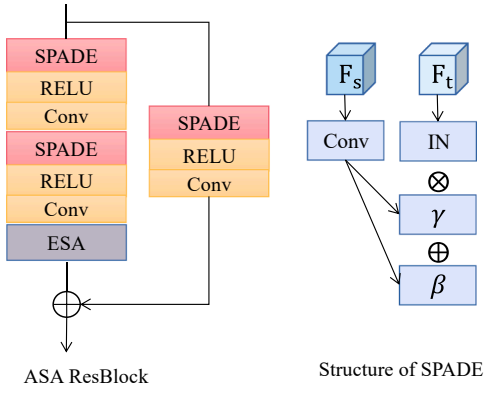

Fig. 3. The illustration of ASA Module.

features and the multi-scale texture features, which is a novel structure that has not been studied in depth in previous work.

3.4. Mask guided discriminator

The ordinary discriminator blindly predicts all patches in the inpainted image as false, and ignores that the patches outside the incomplete region are the truth from the ground truth image, resulting in the generated image containing blurred textures. To address this issue, the proposed network distinguishes the synthetic patches in missing regions from the real patches in the context using a mask-guided discriminator [27], which helps the generator synthesize sharper textures, shown in Fig. 4.

The structure of the discriminator network is shown in Fig. 1, consisting of 4 layers of standard convolution, each of which reduces the space size of the feature map by a half. It takes the restored result and the ground truth as input and outputs the predicted map. Each pixel of the predicted map indicates whether the prediction of the N × N patch in the input image is true or false. The ground truth image is denoted as I_{gt}, the corresponding mask as M (with known pixel value of 0 and missing region of 1), and $\odot$ is the pixel-level multiplication, and G is the generator, so the restored result can be expressed as:

$$z = I_{gt} \odot (1-m) + G\big(I_{gt} \odot (1-m), m\big) \odot m, \tag{11}$$

The discriminator is trained by employing the mask obtained by Gaussian filtering, and the adversarial loss of the discriminator is represented as

$$L_{adv}^{D} = E_{z\sim pz}\big[(D(z) - \theta(1-m))^2\big] + E_{x\sim pdata}\big[(D(x)-1)^2\big], \tag{12}$$

where θ is the composite function of down sampling and Gaussian filtering. Then, the adversarial loss of the generator is denoted as:

$$L_{adv}^{G} = E_{z\sim pz}\big[(D(z)-1)^2 \odot m\big], \tag{13}$$

Through the joint optimization of the discriminator and the generator, the discriminator segments the synthetic patches of the missing regions from the known regions of the context outside the missing regions, facilitating the generation of highly realistic textures by the generator.

3.5. Overall optimization

To ensure the per-pixel reconstruction precision and visual fidelity of the image, the proposed network is trained by employing reconstruction loss and adversarial loss of the mask-guided discriminator. The l_1 loss is applied to the reconstructed image, paying attention to the content of the missing regions to ensure pixel-level reconstruction accuracy and produce images more realistic:

$$L_{rec} = \|I_{gt} - G(x \odot (1-m), m)\|_1 \tag{14}$$

In summary, the whole loss function of the proposed network can be described as:

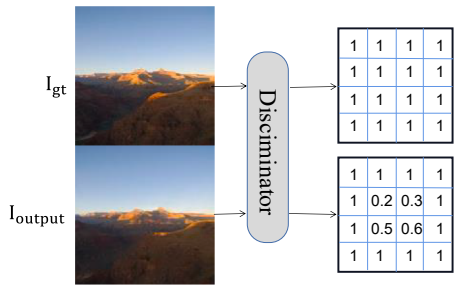

Fig. 4. Description of the training discriminator.

$$L = \lambda_1 L_{rec} + \lambda_2 L_{adv}^{G} + \lambda_3 L_{prior}. \tag{15}$$

where L_{prior} includes ASL and l_1 reconstruction loss in the semantic prior generator. The hyperparameters are optimized by grid search and set as $\lambda_1 = 10$, λ_2=$\lambda_3 = 1$.

4. Experiments

Extensive experiments are performed on three public datasets for subjective and objective evaluations. Ablation studies are also conducted to validate the architectures and modules specifically proposed for each purpose.

4.1. Experimental settings

The network is evaluated on the CelebA-HQ256, the Paris Street View and the Places2 datasets. The CelebA-HQ256 contains 30,000 face images, in which 1000 images are randomly selected as validation and testing dataset. The remained 29,000 images are utilized as training dataset. The Paris Street View is collected from street views of Paris, containing 14,900 training samples and 100 testing samples. The Places2 consists of 2,000,000 images from 365 scenes. Four complete categories are selected to obtain 160,000 images, in which 3,000 images from each category are randomly selected as the test set. The remaining 157,000 images are employed as the training set. The irregular masks are obtained from ref.[6] and categorized according to their hole size relative to 10 % increments of the whole image.

The proposed approach is compared with the following state-of-the-art methods, Pluralistic Image Completion (PIC) [39], Recurrent Feature Reasoning (RFR) [8] and Learned Semantic Priors(SPL) [16].PIC proposed a probabilistic principle-based image complementation framework and a network structure with two parallel training paths, which obtained multiple seemingly reasonable results with large diversity for the same mask. RFR proposed a recurrent feature inference (RFR) module that significantly improved network performance and also bypassed some of the limitations of the asymptotic approach. SPL suggested a new context-aware image restoration model that adaptively merged learned semantic prior information and local features of images into a unified generative network.

In the experiments, three common criteria and two perceptual metrics are applied as quantitative measures to quantify the quality of the restored images. 1) the peak signal-to-noise ratio (PSNR), 2) the structural similarity index (SSIM) [37], 3) the mean l_1 error (MAE) [38], the fréchet inception distance (FID) [43], and the learned perceptual image patch similarity (LPIPS) [44]. In addition, the recovery results are qualitatively evaluated by visually comparing various models for randomly selected test samples with various degrees of corruption. As a complement to the standard assessment metrics, we conduct further user studies to verify our results against the state-of-the-art in human assessment.

The model is implemented in pytorch. Training is initiated on a single RTX 2080 GPU (8 GB) with the batch size of 2 and optimized using the Adam optimizer [40] with $\beta_1 = 0.0$ and$\beta_2 = 0.9$. The initial learning rate is 1e−4 for all experiments and is decayed to 1e−5 for different periods on different datasets. For Places2 and CelebA-HQ, the learning rate is initially decayed for 40 epochs and further fine-tuned the model for another 10 epochs. For Paris Street View, the learning rate is decayed in 60 epochs and fine-tune the model for 20 epochs.

4.2. Qualitative comparisons

Qualitative comparisons of irregular hole filling results by CelebA-HQ[17], Paris Street View[18] and Places2 [19]datasets are shown in Fig. 5, Fig. 6, and Fig. 7, respectively. For a fair comparison, the inpainted results of the compared algorithms are implemented by

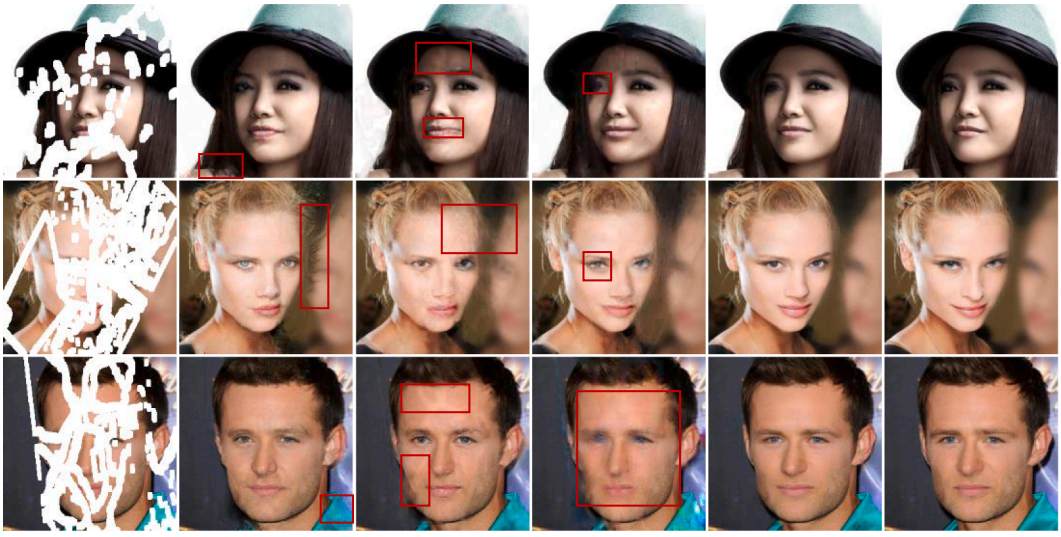

Fig. 5. Qualitative comparisons of our approach with PIC, RFR and SPL on CelebA-HQ.

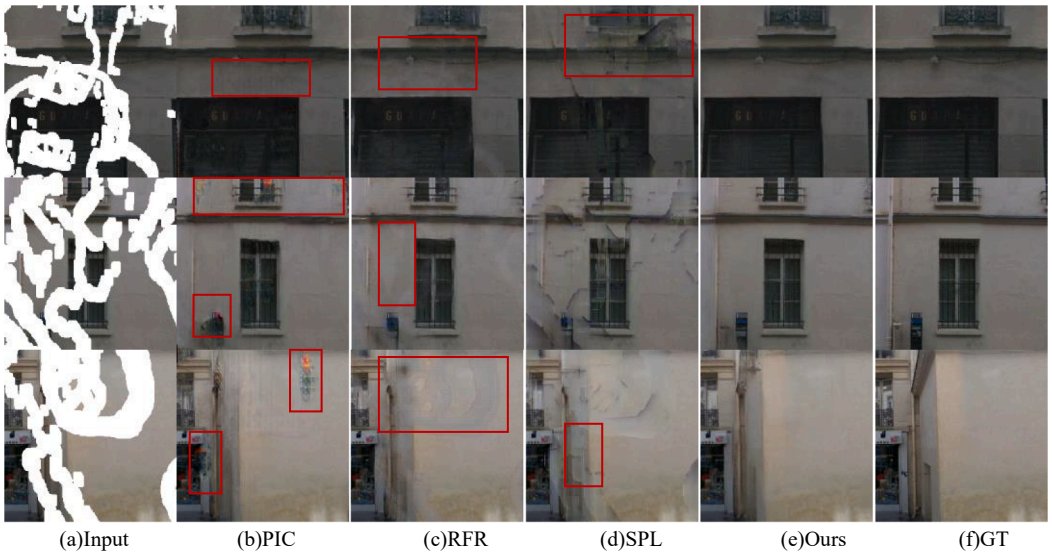

Fig. 6. Qualitative comparisons of our approach with PIC, RFR and SPL on Paris Street View.

retraining their source code. All models are trained and tested in the same way. The same mask is applied for each test image to evaluate the results of different methods.

The images restored from the CelebA-HQ dataset are shown in Fig. 5 (a). The first image has a mask of about 30 % with a sparse distribution, while the second and third images have a mask of about 50 %, occupying most of the face. The inpainted results of the PIC are displayed in Fig. 5 (b). The upper image is well recovered, but the lips are not realistic enough compared with the original image, which the hair with unclear texture appears in the lower left corner, the hair on the right side of the middle image contains unreasonable structure, and the collar of the lower image contains mask traces. The restored results of RFR are presented in Fig. 5(c), which the forehead in the upper image generates extra details not appearing in the original image, and the restored results of the middle and the lower row contain artifacts. For the output image of SPL in Fig. 5(d), the left eye of the first image cannot be restored, the left eye of the second image meets unclear textures, and the remaining images appears poor overall results and contains heavy mask traces. The experimental results of the proposed network are presented in Fig. 5(e), and the output images of the three images contain richer semantics and more detailed textures compared with other methods.

A comparison of the restored results for about 50 % of the irregular disruptions on the Paris street dataset is demonstrated in Fig. 6. In the first row, the overall color is rather dark. The restored effect from PIC occurs with continuous detail, but contains light artifacts in the red box. The output image of RFR presents a small amount of mask traces, while the result of SPL contains heavy mask traces. In contrast, the proposed approach yields result that is closer to the ground truth images. The image in the second row displays concentrated mask distribution. Therefore, the result of PIC emerges structures that are not presented in the original image. The result obtained by RFR demonstrates light artifacts. The output image of SPL exhibits distorted structures, while the proposed network improves the overall visual effect. The overall structures of the last row are relatively simple. Therefore, PIC does not show much mask traces but illogical structures are created. RFR produces unclear textures. The result of SPL contains poor structural similarity of the content. In contrast, the proposed network is able to generate more reasonable semantic information without structural distortion.

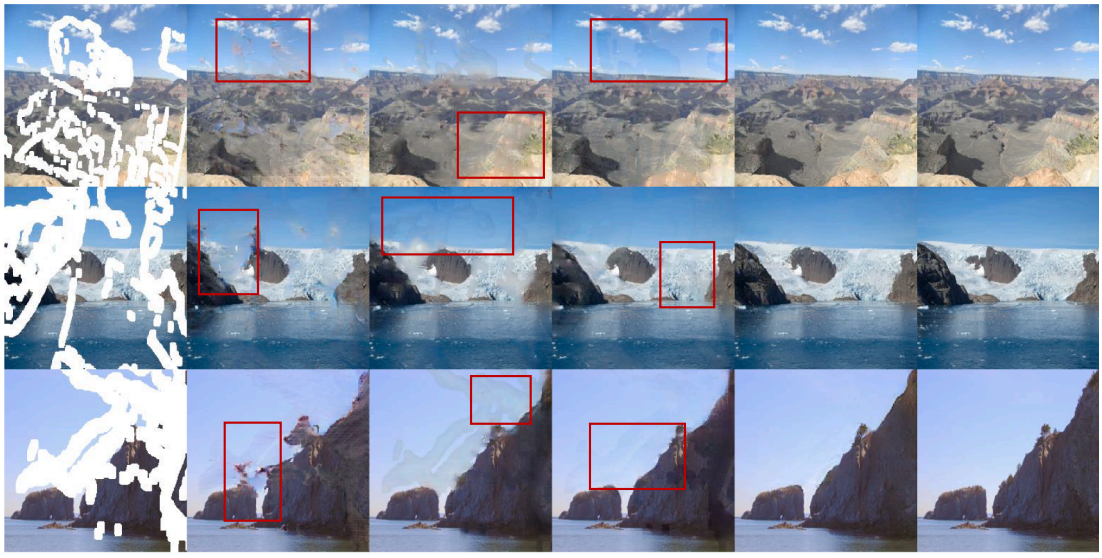

Fig. 7. Qualitative comparisons of our approach with PIC, RFR and SPL on Places2.

Places2 dataset is the most tough dataset because it contains a variety of complex scenes and various objects. As is apparent from Fig. 7, PIC generates distorted structure and cannot extract the image content effectively, RFR and SPL show blurred textures and mask traces. In contrast, the proposed method generates images with clear textures and coherent structures that satisfy human visual perception.

4.3. Quantitative comparisons

In order to quantify the model performance, large irregular damage masks are performed for testing, reclassified into different mask ratios, i.e. (10 %, 20 %), (20 %, 30 %), (30 %, 40 %), (40 %, 50 %) and (50 %, 60 %). As presented in Table 1, the quantitative results of the proposed network are compared with state-of-the-art methods on three datasets of various scale of irregular restoration. In the Paris Street View dataset, the proposed network far outperforms other competing methods on three metrics. As shown in Fig. 8, the proposed network achieves significant improvements on each mask rate. For the mask ratio of (30 %, 40 %), the network improves the PSNR and SSIM by 10.7 % and 7.4 %, respectively, and meanwhile reduces the MAE, FID and LPIPS by 33.3 %, 35.8 % and 51.6 %, respectively, compared to the SPL. For the CelebA-HQ dataset, despite the relatively fixed facial structure, the proposed method still improves the PSNR gain by more than 2.24 dB over other state-of-the-art methods at a large hole-to-image mask ratio of (50 %,60 %). For the Places2 dataset with the mask ratios of (40 %, 50 %), our method increases the PSNR and SSIM by 1.3 % and 1.2 %,respectively, meanwhile decreases the MAE, FID and LPIPS by 10 %, 44.1 % and 22.2 %, respectively, compared to SPL. It can be concluded that the proposed method learns the semantic prior knowledge of images and fuses the image features with multi-scale information to generate high-quality

Table 1
Quantitative comparison of the images produced by our method with state-of-the-art inpainting models on three image inpainting datasets.

Dataset		CelebA-HQ				Paris Street View				Places2			
Method		PIC	RFR	SPL	Ours	PIC	RFR	SPL	Ours	PIC	RFR	SPL	Ours
MAE	10 %–20 %	0.010	0.007	0.006	**0.004**	0.013	0.014	0.009	**0.006**	0.015	0.010	**0.008**	**0.008**
	20 %–30 %	0.017	0.014	0.012	**0.009**	0.020	0.020	0.018	**0.012**	0.025	0.018	0.015	**0.014**
	30 %–40 %	0.024	0.019	0.018	**0.012**	0.030	0.028	0.027	**0.018**	0.035	0.025	0.022	**0.020**
	40 %–50 %	0.033	0.028	0.025	**0.019**	0.040	0.036	0.038	**0.025**	0.048	0.033	0.030	**0.027**
	50 %–60 %	0.054	0.045	0.043	**0.026**	0.057	0.047	0.058	**0.039**	0.068	0.048	0.046	**0.040**
PSNR	10 %–20 %	30.40	31.54	33.35	**33.43**	27.96	30.11	29.97	**33.63**	26.90	29.76	30.50	**31.65**
	20 %–30 %	26.59	29.36	29.27	**30.58**	25.34	27.26	26.55	**29.72**	24.18	26.7	27.51	**27.96**
	30 %–40 %	24.78	26.94	26.91	**29.12**	23.4	25.08	24.39	**27.02**	22.13	24.71	25.58	**25.94**
	40 %–50 %	22.88	25.32	25.08	**27.51**	21.9	23.73	22.63	**25.38**	20.43	23.24	22.88	**23.19**
	50 %–60 %	19.6	21.81	21.52	**23.76**	19.88	21.51	20.34	**22.62**	18.45	21.2	21.42	**21.95**
SSIM	10 %–20 %	0.94	0.97	**0.98**	**0.98**	0.91	0.92	0.93	**0.96**	0.88	0.92	0.93	**0.94**
	20 %–30 %	0.89	0.93	0.92	**0.94**	0.85	0.87	0.86	**0.92**	0.81	0.87	**0.88**	**0.89**
	30 %–40 %	0.85	0.89	0.88	**0.92**	0.79	0.82	0.81	**0.87**	0.75	0.82	0.83	**0.84**
	40 %–50 %	0.79	0.84	0.84	**0.89**	0.72	0.76	0.74	**0.82**	0.68	0.76	0.78	**0.78**
	50 %–60 %	0.70	0.77	0.76	**0.81**	0.64	0.68	0.65	**0.74**	0.60	0.69	0.70	**0.71**
FID	10 %–20 %	28.42	26.2	22.25	**7.58**	43.47	37.08	30.73	**14.7**	46.37	37.65	26.96	**14.05**
	20 %–30 %	42.4	43.71	35.63	**16.93**	38.65	33.98	35.01	**25.36**	61.99	59.16	43.23	**28.14**
	30 %–40 %	47.29	46.15	43.25	**24.17**	53.74	57.73	60.32	**38.68**	77.4	80.6	66.01	**42.2**
	40 %–50 %	63.4	59.17	58.27	**31.9**	62.98	63.14	70.27	**50.81**	89.45	91.97	76.62	**52.11**
	50 %–60 %	73.2	73.10	88.56	**46.21**	85.51	90.34	95.38	**78.2**	95.12	99.46	96.29	**75**
LPIPS	10 %–20 %	0.05	0.04	0.04	**0.01**	0.08	0.07	0.07	**0.02**	0.15	0.07	0.06	**0.03**
	20 %–30 %	0.07	0.07	0.06	**0.04**	0.12	0.16	0.14	**0.05**	0.17	0.13	0.10	**0.07**
	30 %–40 %	0.10	0.11	0.09	**0.07**	0.17	0.15	0.19	**0.09**	0.22	0.17	0.14	**0.11**
	40 %–50 %	0.15	0.13	0.12	**0.09**	0.22	0.2	0.25	**0.13**	0.28	0.22	0.20	**0.15**
	50 %–60 %	0.21	0.17	0.19	**0.13**	0.30	0.29	0.34	**0.21**	0.35	0.3	0.29	**0.22**

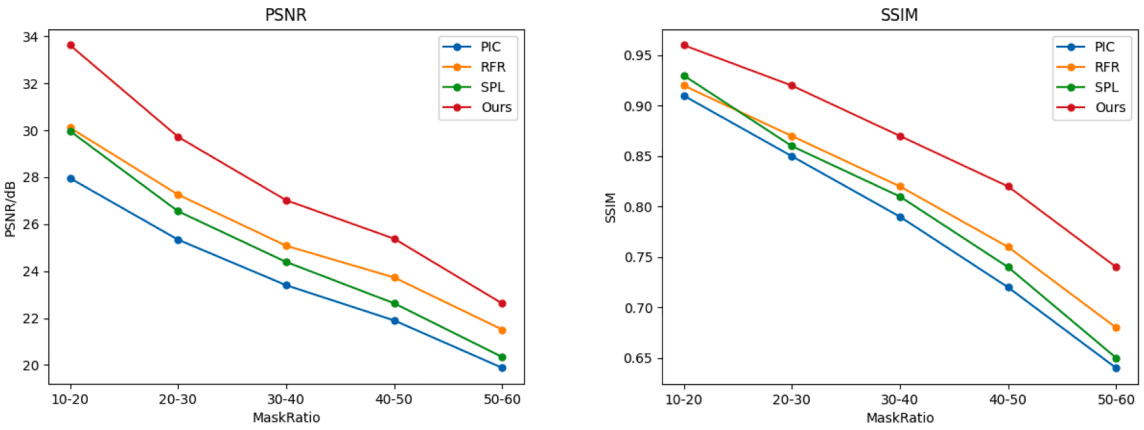

Fig. 8. Quantitative comparisons of different mask ratios are performed on Paris Street View based on PSNR and SSIM.

content with similar semantic distributions.

4.4. User study

A subjective user study is further conducted to evaluate the performance of the proposed method that 20 participants with specialized knowledge in image processing contribute to this evaluation. Participants, who do not know the position of the missing areas, are required to choose the most reasonable image by the proposed approach and the representative state-of-the-art approaches. More specifically, there are 10 questions requested for each participant, which are randomly selected from the Paris Street View and Places2 dataset. The ballots are enumerated and the statistical data is displayed in Table 2. Compared with other approaches, our approach shows more advantages and thus clearly validates its efficiency.

4.5. Ablation studies

To prove the usefulness of three important components of the proposed network in image inpainting, three sets of ablation experiments are performed, i.e., the semantic prior generator, the fused contextual transformation block (FCTB), and the aggregated semantic attention (ASA). To ensure experimental fairness, all models are trained as well as tested using the same settings. The qualitative results of the proposed model compared with the selected ablation model for the restoration of the Places2 dataset are shown in Fig. 9. The quantitative comparison of the ablation experiments is given in Table 3.

4.5.1. Semantic prior generator

The semantic prior generator is studied to extract the structural features to generate well-structured and semantically rich images. The output image without adding the semantic prior is given in Fig. 9(b), containing distorted structure in the top image and obvious artifacts in the bottom image. The quantitative results in Table 3 also indicate that images without the semantic prior have worse metric scores.

4.5.2. Fused contextual transformation block

The fused contextual transformation block (FCTB) is engineered to acquire low-level multiscale texture features of the image. As shown in Fig. 9(c), the top image contains restored traces at the upper left corner, the bottom image includes indistinct textures, and pixel loss occurs in the red box. The quantitative results in Table 3 also demonstrate that on 20 %–40 % irregularly masked images, the PSNR of the generated images without FCTB is reduced by 3.0 %, FID is improved by 30.2 %, and LPIPS is improved by 30.2 % compared to the proposed network. As a result, the generated results without FCTB contain texture information loss and present lower quality.

Table 2
User study for PIC, RFR, and SPL on Paris Street View and Places2.

Method	PIC	RFR	SPL	Ours
Paris Street View	18.60 %	5.10 %	5.35 %	70.91 %
Places2	15.84 %	8.30 %	7.32 %	68.54 %

4.5.3. Aggregated semantic attention-aware module

The ASA module is developed to spatially aggregate low-level multiscale texture features and structural features to the decoder. The restored results without the ASA module are shown in Fig. 9(d), where there are unnatural artifacts in the top image and unexpected noise of the missing regions can be noticed. To make the comparison more concrete, Table 3 presents the quantitative results proving that ASA module contributes to the efficiency gain.

4.6. Limitations and failure cases

The Places2 dataset is one of the most difficult datasets to handle in image inpainting task. As shown in Fig. 10, when the proposed method restores large broken and arbitrarily shaped images, there are still some cases of blurred details and structural distortion in the missing area. In response, we will design a novel Transformer in future work to obtain the interaction between image patches and expand the image perceptual field.

5. Conclusion

Since the current mainstream inpainting methods cannot obtain sufficient semantic priors or reason out the content of unknown regions from the distant background in an image, there are structural distortion and artifact problems in restoring large irregular broken images under complex background. To address these shortcomings, an improved algorithm based on GAN networks is proposed to obtain better inpainting results by end-to-end training. The proposed network contains the semantic prior generator and the fused contextual transformation block, which effectively solves the problems of structural incoherence and texture indistinctness. A novel aggregation semantic module is proposed to fuse structural features and multi-scale texture features into the decoder, which makes it possible to promote texture fine-grained synthesis effectively and solve the problem of irregular breakage. In addition, the mask-guided discriminator is put forward to help the proposed network to distinguish the pixel authenticity of the obtained image, which enhances the competence of the discriminator and makes the generate results free from artifacts effects. Extensive experiments show that the proposed model is capable of handling the problem of blurred details as well as distorted structure and outperforms the state-of-the-art. In the future, we intend to integrate our algorithm into a more

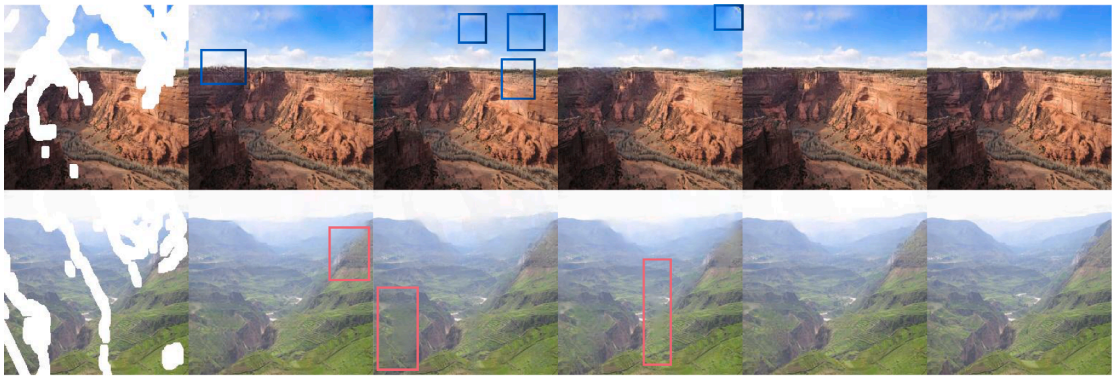

Fig. 9. Visualization of the effect of the network architecture and individual modules on Places2.

Table 3
Quantitative ablation study on Places2.

Metric	PSNR			FID			LPIPS		
Mask Ratio	0–30 %	30 %–50 %	50 %–60 %	0–30 %	30 %–50 %	50 %–60 %	0–30 %	30 %–50 %	50 %–60 %
w/o Semantic Prior	29.23	24.83	21.71	35.45	50.23	60.78	0.10	0.24	0.38
w/o FCTB	29.32	24.62	21.50	40.32	54.65	67.21	0.07	0.19	0.33
w/o ASA	28.23	24.30	21.1	32.33	49.75	58.99	0.06	0.17	0.29
Ours	**29.81**	**25.07**	**21.95**	**28.12**	**47.15**	**55**	**0.05**	**0.13**	**0.22**

Fig. 10. Failure cases of our method.

advanced transformer architecture to design a network with better robustness, which can restore high-resolution images and can perform target removal tasks.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper. This work is supported by the National Natural Science Foundation of China (No. 62266049, No. 62066047 and No. 61861045).

Data availability

Data will be made available on request.

References

[1] M. Bertalmio, G. Sapiro, V. Caselles, et al., Image inpainting, in: Proceedings of the 27th annual conference on Computer graphics and interactive techniques, 2000, pp. 417–424.
[2] Z. Wan, B. Zhang, D. Chen, et al., Bringing old photos back to life, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2020, pp. 2747–2757.
[3] R. Shetty, M. Fritz, B. Schiele, Adversarial scene editing: automatic object removal from weak supervision, arXiv preprint arXiv:1806.01911, 2018.
[4] C. Barnes, E. Shechtman, A. Finkelstein, et al., PatchMatch: a randomized correspondence algorithm for structural image editing, ACM Trans. Graph. 28 (3) (2009) 24.
[5] M. Bertalmio, L. Vese, G. Sapiro, et al., Simultaneous structure and texture image inpainting, IEEE Trans. Image Process. 12 (8) (2003) 882–889.
[6] G. Liu, F.A. Reda, K.J. Shih, et al., Image inpainting for irregular holes using partial convolutions, in: Proceedings of the European Conference on Computer Vision (ECCV), 2018, pp. 85–100.
[7] J. Yu, Z. Lin, J. Yang, et al., Free-form image inpainting with gated convolution, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2019, pp. 4471–4480.
[8] J. Li, N. Wang, L. Zhang, et al., Recurrent feature reasoning for image inpainting, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2020, pp. 7760–7768.
[9] J. Yu, Z. Lin, J. Yang, et al. Generative image inpainting with contextual attention, in: Proceedings of the IEEE conference on computer vision and pattern recognition, 2018, pp. 5505–5514.
[10] Z. Yan, X. Li, M. Li, et al., Shift-net: Image inpainting via deep feature rearrangement, in: Proceedings of the European conference on computer vision (ECCV), 2018, pp. 1–17.
[11] Y. Song, C. Yang, Z. Lin, et al. Contextual-based image inpainting: infer, match, and translate, in: Proceedings of the European Conference on Computer Vision (ECCV), 2018, pp. 3–19.
[12] K. Nazeri, E. Ng, T. Joseph, et al., Edgeconnect: structure guided image inpainting using edge prediction, in: Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops, 2019.
[13] Y. Zeng, J. Fu, H. Chao, et al., 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.
[14] H. Liu, B. Jiang, Y. Xiao, et al., Coherent semantic attention for image inpainting, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2019, pp. 4170–4179.

[15] P. Wang, P. Chen, Y. Yuan, et al., Understanding convolution for semantic segmentation, in: 2018 IEEE winter conference on applications of computer vision (WACV), IEEE, 2018, pp. 1451–1460.
[16] W. Zhang, J. Zhu, Y. Tai, Y. Wang, W. Chu, B. Ni, C. Wang, X. Yang, Context-aware image inpainting with learned semantic priors, in: Z.-H. Zhou (Ed.), Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence, IJCAI-21, International Joint Conferences on Artificial Intelligence Organization, 2021, pp. 1323–1329.
[17] Z. Liu, P. Luo, X. Wang, et al., Deep learning face attributes in the wild, in: Proceedings of the IEEE international conference on computer vision, 2015, pp. 3730–3738.
[18] C. Doersch, S. Singh, A. Gupta, et al., What makes Paris look like Paris? ACM Trans. Graphics 31 (4) (2012).
[19] B. Zhou, A. Lapedriza, A. Khosla, et al., Places: a 10 million image database for scene recognition, IEEE Trans. Pattern Anal. Mach. Intell. 40 (6) (2017) 1452–1464.
[20] C. Ballester, M. Bertalmio, V. Caselles, et al., Filling-in by joint interpolation of vector regions and gray levels, IEEE Trans. Image Process. 10 (8) (2001) 1200–1211.
[21] A. Criminisi, P. Perez, K. Toyama, Object removal by exemplar-based inpainting, in: 2003 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2003, Proceedings, IEEE, 2003, 2: II-II.0.
[22] R.A. Yeh, C. Chen, T. Yian Lim, et al., Semantic image inpainting with deep generative models, in: Proceedings of the IEEE conference on computer vision and pattern recognition, 2017, pp. 5485–5493.
[23] I. Goodfellow, J. Pouget-Abadie, M. Mirza, et al., Generative adversarial nets, Adv. Neural Inf. Process. Syst. 27 (2014).
[24] M. Mirza, S. Osindero, Conditional generative adversarial nets, arXiv preprint arXiv:1411.1784, 2014.
[25] T. Yu, Z. Guo, X. Jin, et al., Region normalization for image inpainting, in: Proceedings of the AAAI Conference on Artificial Intelligence, 2020, Vol. 34(07), pp. 12733–12740.
[26] B. Dolhansky, C.C. Ferrer, Eye in-painting with exemplar generative adversarial networks, in: Proceedings of the IEEE conference on computer vision and pattern recognition, 2018, pp. 7902–7911.
[27] Y. Zeng, J. Fu, H. Chao, et al., Aggregated contextual transformations for high-resolution image inpainting, IEEE Trans. Visual. Comput. Graphics (2022).
[28] J. Johnson, A. Alahi, L. Fei-Fei, Perceptual losses for real-time style transfer and super-resolution, in: European conference on computer vision, Springer, Cham, 2016, pp. 694–711.
[29] L.A. Gatys, A.S. Ecker, M. Bethge, Image style transfer using convolutional neural networks, in: Proceedings of the IEEE conference on computer vision and pattern recognition, 2016, pp. 2414–2423.
[30] S. Iizuka, E. Simo-Serra, H. Ishikawa, Globally and locally consistent image completion, ACM Trans. Graphics (ToG) 36 (4) (2017) 1–14.
[31] P. Isola, J.Y. Zhu, T. Zhou, et al., Image-to-image translation with conditional adversarial networks, in: Proceedings of the IEEE conference on computer vision and pattern recognition, 2017, pp. 1125–1134.
[32] T. Miyato, T. Kataoka, M. Koyama, et al., Spectral normalization for generative adversarial networks, arXiv preprint arXiv:1802.05957, 2018.
[33] Y. Ren, X. Yu, R. Zhang, et al., Structureflow: Image inpainting via structure-aware appearance flow, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2019, pp. 181–190.
[34] A. Kuznetsova, H. Rom, N. Alldrin, et al., The open images dataset v4, Int. J. Computer Vis. 128 (7) (2020) 1956–1981.
[35] T. Park, M.Y. Liu, T.C. Wang, et al., Semantic image synthesis with spatially-adaptive normalization, in: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 2019, pp. 2337–2346.
[36] F. Kong, M. Li, S. Liu, et al., Residual local feature network for efficient super-resolution, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2022, pp. 766–776.
[37] Z. Wang, A.C. Bovik, H.R. Sheikh, et al., Image quality assessment: from error visibility to structural similarity, IEEE Trans. Image Process. 13 (4) (2004) 600–612.
[38] R. Grosse, M.K. Johnson, E.H. Adelson, et al., Ground truth dataset and baseline evaluations for intrinsic image algorithms, in: 2009 IEEE 12th International Conference on Computer Vision, IEEE, 2009, pp. 2335–2342.
[39] C. Zheng, T.J. Cham, J. Cai, Pluralistic image completion, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2019, pp. 1438–1447.
[40] D.P. Kingma, J. Ba, Adam: a method for stochastic optimization, arXiv preprint arXiv:1412.6980, 2014.
[41] W. Zhang, Y. Wang, J. Zhu, et al., Fully context-aware image inpainting with a learned semantic pyramid, arXiv preprint arXiv:2112.04107, 2021.
[42] E. Ben-Baruch, T. Ridnik, N. Zamir, et al., Asymmetric loss for multi-label classification, arXiv preprint arXiv:2009.14119, 2020.
[43] M. Heusel, H. Ramsauer, T. Unterthiner, et al., Gans trained by a two time-scale update rule converge to a local nash equilibrium, Adv. Neural Inf. Process. Syst. 30 (2017).
[44] R. Zhang, P. Isola, A.A. Efros, et al., The unreasonable effectiveness of deep features as a perceptual metric, in: Proceedings of the IEEE conference on computer vision and pattern recognition, 2018, pp. 586–595.
[45] C. Cao, Q. Dong, Y. Fu, Learning prior feature and attention enhanced image inpainting, in: Computer Vision – ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XV. Springer-Verlag, Berlin, Heidelberg, 2022, pp. 306–322, doi: 10.1007/978-3-031-19784-0_18.
[46] H. Zheng et al., Image inpainting with cascaded modulation GAN and object-aware training, in: S. Avidan, G. Brostow, M. Cissé, G.M. Farinella, T. Hassner (Eds.), Computer Vision – ECCV 2022, ECCV 2022, Lecture Notes in Computer Science, Vol. 13676, Springer, Cham, 2022, doi: 10.1007/978-3-031-19787-1_16.