IEEE Access
Multidisciplinary : Rapid Review : Open Access Journal

Received 29 May 2022, accepted 13 June 2022, date of publication 24 June 2022, date of current version 30 June 2022.

Digital Object Identifier 10.1109/ACCESS.2022.3186009

Image Inpainting Based on Interactive Separation Network and Progressive Reconstruction Algorithm

JUN GONG[1,2], SIYUAN LI[2], SHIYU CHEN[2], LIANG NIE[2], XIN CHENG[3], (Graduate Student Member, IEEE), ZHIQIANG ZHANG[3], AND WENXIN YU[2], (Member, IEEE)

[1]Information System and Security and Countermeasures Experimental Center, Beijing Institute of Technology, Beijing 100081, China
[2]School of Computer Science and Technology, Southwest University of Science and Technology, Mianyang 621010, China
[3]Graduate School of Science and Engineering, Hosei University, Tokyo 184-8584, Japan

Corresponding author: Xin Cheng (xin.cheng.5x@stu.hosei.ac.jp)

This work was supported in part by Ministry of Science and Technology Project under Grant 2021ZD0110600, in part by the Ministry of Industry and Information Technology Project of the People's Republic of China under Grant 2020-0103-1-2, and in part by the Mianyang Science and Technology Program under Grant 2020YFZJ016.

ABSTRACT Recently, learning-based image inpainting has gained much attention. It widely utilizes an auto-encoder structure and can obtain compact feature representation in the encoder to achieve high-quality image inpainting. Although this approach has achieved encouraging inpainting results, it inevitably reduces the high-resolution representation due to interval downsampling. In order to solve this problem and achieve an excellent image inpainting effect, this paper proposes a brand-new generative network, Interactive Separation Network, which retains the high-resolution information and extracts the semantic features from corrupted images. Furthermore, this paper also discusses network designs with different complexity in different application scenarios. Finally, to improve the effectiveness and robustness of our proposal to large corrupted regions in the inpainting image, we further propose a flexible and highly reusable reconstruction scheme to complete the inpainting in the prediction process gradually. Experiments show that our proposed generation network and reconstruction scheme can significantly improve the quality of repaired images. The proposed method significantly outperforms the state-of-the-art image inpainting approaches in image quality.

INDEX TERMS Image inpainting, image completion, feature fusion, reconstruction algorithms.

I. INTRODUCTION

Image inpainting, a.k.a. image completion, refers to the process of filling in missing content in damaged images. It plays a critical role in addressing the various issue of computer vision, such as object or artifact removal, 3D reconstruction, and depth-image-based rendering (DIBR) technology. Due to the complexity of the context in different scenes, image inpainting becomes one of the challenging problems in imaging tasks.

The last decade has seen a growing trend toward convolution neural networks (CNN) [1], [2] and generative adversarial networks (GAN) [3], so they have obtained much attention. CNN can learn the high-level representation of images in deep learning, so it delivers marvelous success in a variety of computer vision tasks (e.g., image classification [4] and object recognization [5]). Besides, based on GAN, the content produced by the generative network is more realistic and closely conforms to the human visual patterns. Benefiting from these technics, many approaches [6]–[12] have the ability to recover the damaged image with strong prior knowledge for the semantic understanding of the scenes. They have achieved encouraging results, which have driven the rapid development of image inpainting in recent years.

The CNN-based encoder-decoder architecture [6] has been widely employed in image inpainting, and most of them are similar to Unet-style [13] structures. Although the pooling layer in the Unet-style network can compress the features into a compact representation, they inevitably discard numerous high-resolution signals in spatial dimension due to the interval downsampling, as observed by Figure 1. The problem is not having an efficient one-stage model that also puts enough attention on the textured pattern and semantic analysis of images in a learning-based fashion.

The associate editor coordinating the review of this manuscript and approving it for publication was Chun-Wei Tsai.

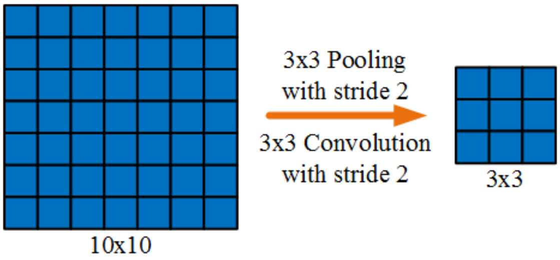

FIGURE 1. The stride convolution or stride pooling reduces the spatial representation of features by interval downsampling. Although these operations compress the data efficiently, they inevitably throw away some of the original information from the data.

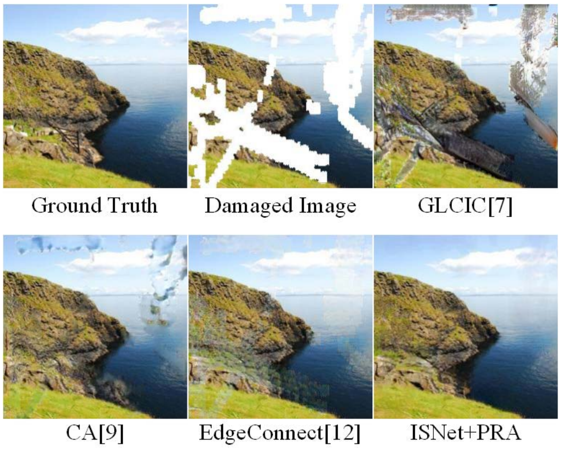

FIGURE 2. Visual comparison of inpainting performance with official pre-trained models: GLCIC [7] CA [9], EdgeConnect [12], ISNet+PRA (Proposed approach). The result of GLCIC [7] has no post processing added.

There are serval attempts [7], [9] to perform semantic inpainting in a learning-based fashion. Nevertheless, these approaches do not maintain the trade-off between texture content and structural semantics in images. For example, as shown in Figure 2, the image recovered by GLCIC [7] has a large region of texture artifacts, and the Contextual Attention (CA) [9] crafts well-structured massif but fails in the structures of clouds. To acquire a better inpainting effect, researchers have also begun to investigate the legible inpainting with the perceived semantics of images. One of the common ways is building an additional network [12] to further refine the general results. The result of EdgeConnector [12] in Figure 2 displays pretty general structures of objects, the completed textures have subtle flaws. However, there are many regions in the result of EdgeConnector that have subtle flaws, which make the overall inpainting effect mediocre. On the other hand, these kinds of approaches need a large amount of computation and will increase space complexity because they commonly require two large networks.

In order to solve the above problems and achieve a better image inpainting effect, this paper proposes a brand-new generative network, named Interactive Separation Network (ISNet), which maintains the balance between textured pattern and semantic context through two well-designed network branches. For the convenience of interpretation, this paper definite three main operations — Inpainting, Interaction, and Aggregation (the definition is similar to the literature [14]), which manipulate the feature representation and form each independent stage in ISNet, and the step-by-step connection of the formed stages constitute the main body of ISNet. A brief overview of the ISNet framework is shown in Figure 3. In order to solve the inherent problem of current image inpainting approaches, that is, lack of robustness to a big damaged region, we further propose an efficient, straightforward, and highly reusable algorithm in the prediction process to progressively image completion. The proposed method is demonstrated that achieve the state-of-the-art inpainting results. The pre-trained models and code for network structure, progressive inpainting, and ablation research can be accessed at https://github.com/GuardSkill/Large-Scale-Feature-Inpainting/tree/journal.

In summary, the main contributions of this article are as follows:

- We propose an efficient completion network (ISNet) that can both understand the scene and recognize the texture pattern of images. Experimental results show that it achieves excellent image inpainting performance;
- We propose an efficient and highly reusable completion scheme that can progressively complete the images in the prediction stage to improve the robustness of the proposed network structure;
- In order to verify our proposed method that can be widely applied to different situations, the experiments study the efficiency of ISNet in diverse network structure settings, which can leave a valuable experience for researchers to design their own models.

II. RELATED WORK

The rise of deep learning has made great progress in the image processing field, such as image caption [15], [16], image generation [17], [18], image reconstruction [19], [20], and image inpainting [6], [21]. Among them, image inpainting is one of the most challenging and widely used research.

In the past few years, a variety of learning-based approaches have flourished in the field of image inpainting. Benefiting from CNN, well-trained generative networks can perform the high-level recognition of diverse scenes. However, even though there are several earlier CNN-based approaches [21], [22] that were designed to restore the corrupted image or text-covered image, these early learning-based inpainting approaches lack the versatility for irregular masks of different sizes because they only handle very small and thin holes. At that time, the generalization of inpainting models still needed to be improved. In order to achieve higher quality image inpainting results, some improved methods are proposed. Specifically, there are three types: learning-based image inpainting approaches, image inpainting approaches by introducing perceptual loss, and progressive image inpainting approaches.

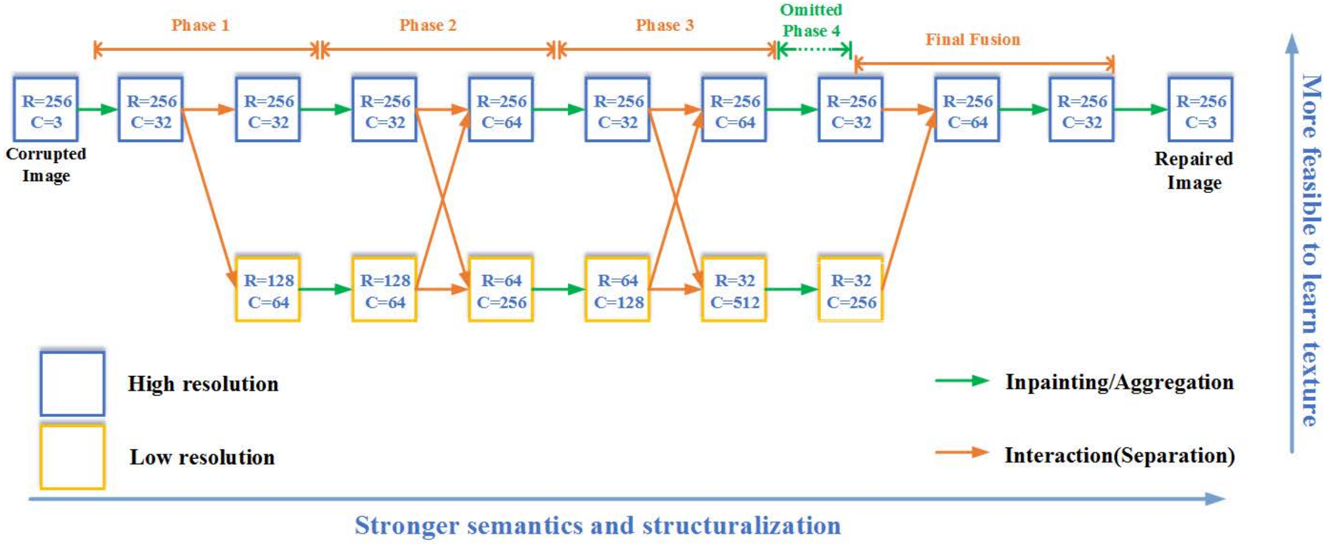

FIGURE 3. The generative model in the GAN-based architecture of Interactive Separation Network, where the *R* (in the figure) refers to the resolution of the current feature block and the *C* represents the channel number. (The 4th phase and discriminator of ISNet are omitted for clarity.) The forwarding path traversing the blue rectangular feature is the high-resolution branch, which handles a series of features with a fixed full resolution. While the low-resolution branch (the yellow rectangular) scales down the propagating features as it progresses.

A. LEARNING-BASED INPAINTING APPROACHES

Goodfellow proposed Generative Adversarial Network (GAN) [3] that can calculate the adversarial loss via a primary network called generator and an auxiliary network called discriminator. Using this kind of adversarial learning, GAN has attracted major research interest from various fields. One of the typical and classic learning-based inpainting approaches is the Context Encoder [6], in which the encoder embeds the input image into the high-level feature maps with low spatial dimension, then the decoder exploits the compact features to reconstruct the original image. This approach has shown that learning-based architecture has the extraordinary ability to understand image context and hallucinate realistic objects in the completed region. Influenced by this exploratory contribution, the Unet-style networks [13] have been widely used as a generative model among the learning-based approaches in the field of image inpainting. However, due to the highly efficient compression scheme and sequential structure of Context Encoder, the generated content in the resulting images are excessively smooth and visually obscured.

Following the Context Encoder, the other well-known learning-based inpainting approach was proposed by Iizuka [7]. Iizuka built two discriminators to respectively verify the authenticity of general images and inpainted regions. Afterward, the discriminators feed realistic scores back to generative models to recover images with comprehensive coherence. However, there still exist style inconsistencies between the completed region and the existing region, which make their results greatly dependent on the post-processing. Due to the lack of optimization of discriminators, the intricate training procedure of Iizuka work is time-consuming and unstable.

However, previous methods often lead to issues such as inter-frequency collisions and repair impairments since they all simply apply together different losses that focus on synthesizing content at different frequencies. Therefore, Yu *et al.* [23] proposed a wavelet-based inpainting network, which effectively alleviates inter-frequency conflicts and fills the missing regions in each frequency band by decomposing the image into multiple frequency bands and applying L1 reconstruction loss in low-frequency bands and adversarial loss in high-frequency bands.

B. PERCEPTUAL LOSS IN IMAGE INPAINTING

Liu *et al.* [8] introduced the high-weighted style loss term to produce structural content, which can carry out a high-level recognition of content and style. Although their result is visually plausible, some of their inpainted images contain excess smoothly content in the filled region, and some still exist the checkboard artifacts in the unofficial reproduction of their work.

In the mode of Partial Convolution Network (PConv) [8], the perceptual loss and style loss [24], [25] are considered to be two of the objective function terms that need to be minimized. For the perceptual loss, it encourages the model learning to generate images that have a similar high-level representation as to the original image. The perceptual loss can be expressed as Eq.(1).

$$\mathcal{L}_{perc} = \sum_{i}^{L} \frac{1}{N_i} \left\| \phi_i\left(\mathbf{I}_{gt}\right) - \phi_i\left(\mathbf{I}_{pred}\right)\right\|_1 \tag{1}$$

where $\mathbf{I}_{pred}$ and $\mathbf{I}_{gt}$ are repaired image by generative model and corresponding ground truth image, $\phi_i(x)$ is an image feature extractor, which uses VGG [4] to extract the corresponding features of the image. Here, ϕ_i corresponds to the values in the feature maps from pool1,pool2, and pool3, so the L is equal to 3 in this paper. N_i represents the number of elements in ϕ_i, and this formula can be understood as the Mean Absolute Error [26] of the higher-level feature spaces. For the style loss, it can penalize the difference between the predicted image and ground truth in terms of style and general tone by the correlation of feature maps. The style loss can be formalized as Eq.(2).

$$\mathcal{L}_{style} = \sum_{j}^{L} \frac{1}{C_j C_j} \left\| G\left(\phi_i\left(\mathbf{I}_{gt}\right)\right) - G\left(\left(\mathbf{I}_{pred}\right)\right) \right\|_1 \quad (2)$$

G is a $C_j \times C_j$ nomarlized Gram matrix. C_jC_j refer to normalization factor.

C. PROGRESSIVE INPAINTING APPROACHES

Progressive image inpainting, including the structure-to-texture approach and the boundary-to-center approach, has recently been investigated. The structure-to-text approach usually employs a two-stage network structure, of which each stage respectively generates the structure/edge and texture features. Yu *et al.* [9] proposed a contextual attention mechanism and a two-stage inpainting scheme. They built two networks to complete high-quality inpainting. The first network in their approach is designed to infer the coarse content, indicating the rough structure information of the missing region, and then the second one aims to refine the produced coarse results. Afterward, their following work introduced Gated Convolution and user-guided information [11] into the two-staged approach and achieved further expansion. The other promising two-stage inpainting approach proposed by Kamyar *et al.* [12]. This work firstly restores the general contour of the image from the corrupted image and the corresponding edge map and then takes the filled edge map as prior structure information to guide colorization.

Although these methods attempted to solve inpainting tasks by adding structural constraints, they still have some problems. Firstly, due to the use of a series-coupled architecture, it is easily subjected to the adverse effects of unreasonable structure preconditions during the inference time. For this issue, Liu *et al.* [27] recovered structures and textures via feeding a structure branch and a texture branch with the deep and shallow features and concatenating and equalizing the features they output. Guo *et al.* [28] proposed a two-stream coupled network for image inpainting, which uses a structure-constrained texture synthesis stream and a texture-guided structure reconstruction stream to better utilize each other for a more rational generation. Also, Bi-directional Gated Feature Fusion (Bi-GFF) module is proposed to combine the results of the two streams. Secondly, it still lacks information for restoring deeper pixels in holes for their backbone when the corrupted region is relatively large. To solve this problem, Li *et al.* [29] devised a Recurrent Feature Reasoning (RFR) module which recurrently infers the hole boundaries of the convolutional feature maps and then uses them as clues for further inference, progressively strengthening the constraints for the hole center and inpainting the image.

Although these approaches [9], [11], [12], [28], [29] can generate well-defined content by a two-stage inpainting strategy, which means that they need to separately build two different networks in two stages, it consumes intensive computation. Besides, the performance of their second network will suffer if the first stage network has poor inpainting prediction. By contrast, this paper proposes a network that can efficiently capture both structural and texture information in a one-stage network.

III. APPROACH

This paper proposes a GAN-based neural network, called the Interactive Separation Network (ISNet), which is trained to perform image inpainting tasks. For better comparison, we employ the discriminator and objective function similar to the EdgeConnect [12] in the early stages of the experiment. Instead, the novel generative network in ISNet has two branches designed to maintain low-resolution representations and high-level information, which can be seen in Figure 3. Blue and yellow rectangular represent high resolution and low resolution branches, respectively. In order to concisely describe the proposed model, this paper divides the forward process of the network into 4 consecutive segments (called phases), which have similar ways to manipulate the features of the two branches. In this section, we state the internal structure inside each phase, the generator's overall structure, and the objective function used in the proposed approach.

A. THREE OPERATIONS

Each phase of ISnet is composed of three defined operations — Inpainting, Interaction, and Aggregation. Except for the first phase, all the phases process feature representations separately at two different resolutions. Taking the first two phases as an example, Figure 4 describes the detailed propagation of these phases. It is worth mentioning that the internal structure of phases 3 and 4 in Figure 3 is consistent with that of phase 2 in Figure 4. During the Inpainting process of the second phase, two different ResBlocks [30] are adopted separately in two branches to handle two types of features that are produced by the previous phase. In the first phase, only one branch is equipped with one ResBlocks to process the feature.

The Interaction operation aims to interact the information between the two branches and further downsample the low-resolution branch with a higher channel dimension. It employs the convolution with stride 2 and the double number of filters to downsample the feature resolution to half of the previous state. Meanwhile, repeat n (where n is equal to phase index) stride convolution to progressive scale down the first-branch feature into the resolution same as the second branch. Then outputs are concatenated together into the second branch. To propagate the high-resolution branch, the previous second branch features are put into $(n-1)$ Sub-Pixel [31] convolutional layers and then stacked with

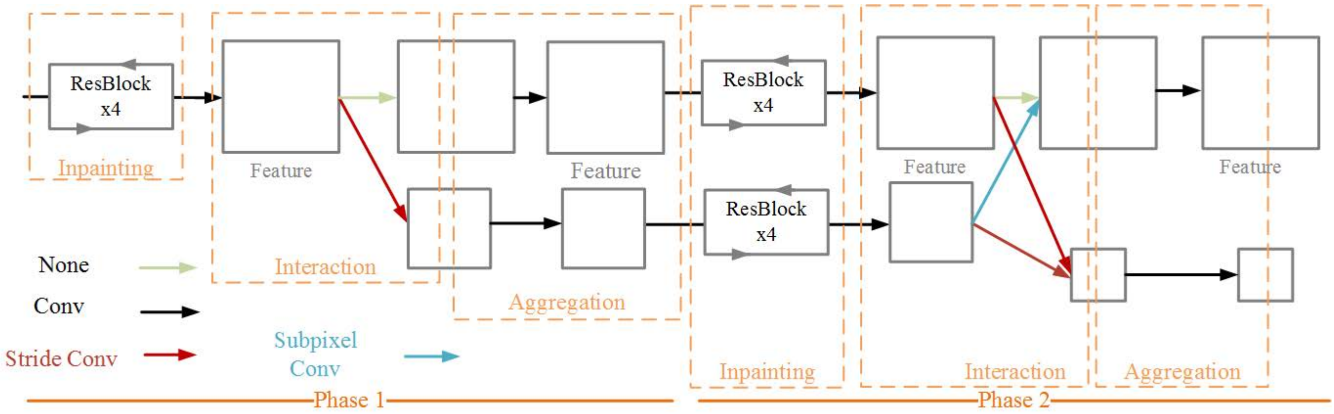

FIGURE 4. The internal structure of the previous two phases in the Interactive Separation Network is shown above. 'None' refers to there is no operation for feature processing (in phase 1), or just concatenate input with other processed features (in phase 2).

previous features of the high-resolution branch, which adaptively learn an upsampling scheme using CNN. As same to the Inpainting operation, the Interaction process only inputs the high-resolution branch in the first phase because there only exists a larger-resolution feature branch.

The Aggregation operation is designed to fuse the information that is produced by the Interaction operation. It exploits convolution with a 3 × 3 kernel in both branches to merge and compress the channels into specific numbers, where the resolution of the current feature determines the compressed channel number. In the proposed ISNet, the channel numbers are distributed in 32, 64, 128, 256, and 512 as the resolution of the feature decreases.

B. NETWORK DESIGN

Let's look back to Figure 3 again. Before inputting the features into the 4 sequential phases, ISNet firstly embeds the masked RGB images into 32-dimensional feature maps with the resolution of 256 × 256. Because there are two branches to process data, the resolutions of outputs produced by the last phase (4th phase) are 256 × 256 and 16 × 16, respectively. At the end of the generative model, the Final Fusion (FF) process is applied to fuse these output features into a high-resolution feature block through stacked Sub-Pixel layers and concatenation. Finally, the features are decoded into a repaired image using a 3 × 3 convolution. To verify the effectiveness of the final fusion process, we try to drop out of the FF process and decode the high-resolution branch directly into 3-channel images. Section IV demonstrates that this simplified design results in worse performance than the network equipped with this process. It is noticeable that the final 3-channel features are input into $tanh(x)$ function and mapped to the range between 0 and 1. The mapping can be formalized as Eq.(1). In all intermediate processing of the generator, the $tanh(x)$ activation function is adopted to deliver feature values, and the zero padding is used to control the variation of resolution.

In the discriminator of ISNet, the high-level features are collected using 3 continuous vanilla convolution layers with stride 2 and then using the 2 convolution layers with the same padding. We are able to manipulate the number of feature channels and reduce the number to 1. In addition, the discriminator in this work is based on the PatchGAN [11], [32]. Therefore, the final output of the discriminator is a single feature map, in which each pixel will judge the part of the region related to it. In order to represent the probability of whether the receptive region of the neural unit is generated by networks or real, the sigmoid activation function is used in each layer of the discriminator.

As mentioned in [33], the advantage of using spectral normalization is that it can stabilize the training process. Spectral normalization suppresses the weight matrixes in each layer by utilizing the maximum singular value of the weight matrixes, which limits the Lipschitz constant of the network to 1. The spectral normalization is originally used only in the discriminator, however, Odena [34] has recently demonstrated that it can refrain generators away from dramatic changes in parameters and gradient. As a result, spectral normalization is applied to both the generators and discriminators in ISNet.

C. LOSS FUNCTION AND EXPLORATION

In this paper, the mapping function of the generator and discriminator of ISNet are respectively denoted as $G(x)$ and $D(x)$ for short. Following the aforementioned symbols, $\mathbf{M}$ refers to the mask that only includes binary values, $\mathbf{I}_{gt}$ and $\mathbf{I}_{pred}$ respectively represent the ground truth image and the image inferred by the generative model. Using $\mathbf{I}_{gt} \odot M$ represent damaged image, the image generation process can be expressed as $\mathbf{I}_{pred} = G\left(\mathbf{I}_{gt} \odot \mathbf{M}\right)$, where $\odot$ is Hadamard product. To train the discriminator, the real image and repaired image are sent to the discriminator in a one-to-one ratio for discrimination. The hinge loss is adopted as the objective function of the discriminator, which can maximize the margin between positive and negative samples. By minimizing the objective function, the discriminator can better distinguish whether an image is repaired by the generative

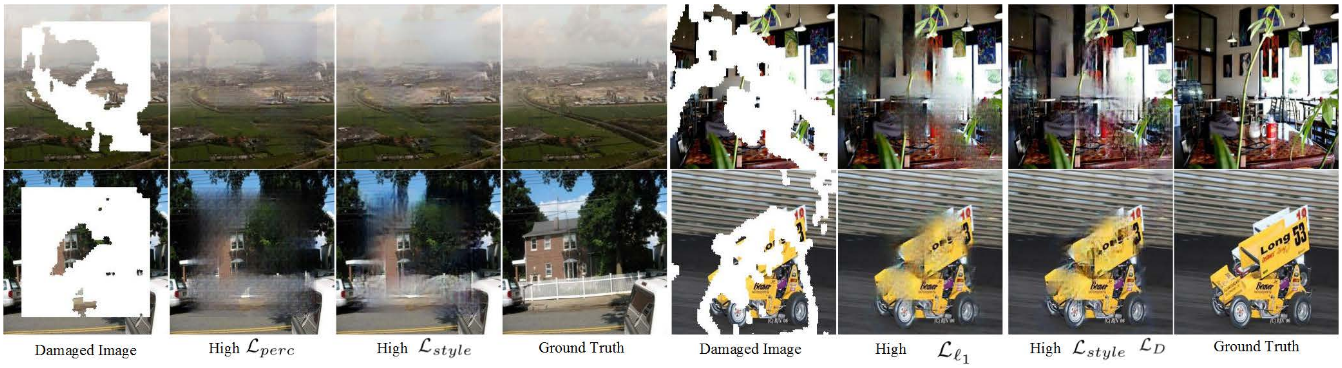
Damaged Image
High $\mathcal{L}_{perc}$
High $\mathcal{L}_{style}$
Ground Truth
Damaged Image
High $\mathcal{L}_{\ell_1}$
High $\mathcal{L}_{style}$ $\mathcal{L}_D$
Ground Truth

FIGURE 5. The image inpainting results of the models trained by different loss functions are shown above. The column of 'High $\mathcal{L}_{perc}$', 'High $\mathcal{L}_{style}$', 'High $\mathcal{L}_{\ell_1}$' and 'High $\mathcal{L}_{style}$ $\mathcal{L}_D$' is the inpainting results of the model in the fourth, second, first, and eighth rows of Table 1, respectively. It's worth noting that all these models are trained 3200 times and might are not the optimal models.

model. The hinge loss can be described as Eq.(3)

$$\mathcal{L}_D = \mathbb{E}_{gt}\left[m(1 - D\left(\mathbf{I}_{gt}\right))\right] + \mathbb{E}_{pred}\left[\psi\left(1 + D\left(\mathbf{I}_{pred}\right)\right)\right] \tag{3}$$

where m refers to the parameter of margin, ψ represents the ReLU function that is used to filter the negative values.

In the early stages of the trial, the generative model is trained on a joint loss function that is similar to EdgeConnect [12], the objective function that expects to minimize can be described as Eq.(4).

$$\mathcal{L}_G = \mathcal{L}_{\ell_1} + 0.1(-\mathcal{L}_D) + 0.1\mathcal{L}_{perc} + 250\mathcal{L}_{style} + 10\mathcal{L}_{FM} \tag{4}$$

For the objective function (Eq.(4)) of the generator, we further investigate the effects of different components of the loss function and analysis the principal components of the loss function.

Firstly, we conduct a random hyperparameter search on the weights of each loss term, and we train the dozens of models for different weight combinations, and each model is iterated nearly 3200 times with nearly 256,000 samples (batch size is set to 8). In order to widely search the weight space, all the parameter magnitudes are randomly chosen and the ranges of magnitudes are empirically selected. Some of the experimental results are described in Table 1.

However, it's very time-consuming work, and we find that quantitative score is not necessarily proportional to qualitative effect. The models achieving high quantitative scores may generate extremely unpleasant images. For example, the objective function with high-weighted $\mathcal{L}_{\ell_1}$ and $\mathcal{L}_{perc}$ term can achieve a higher quantitative score, but the checkboard artifacts and blurry contents may exist in the generated image. These phenomenons can be observed in Figure 5. According to this study, the high-weighted $\mathcal{L}_{\ell_1}$, $\mathcal{L}_{perc}$, and $\mathcal{L}_{style}$ terms can greatly improve the quantization score of the repaired images than other components. By paying more attention to these key components, we find that $\mathcal{L}_{perc}$ loss term can produce more structural inpainting, and the $\mathcal{L}_D$ term has the ability to reduce the checkboard artifacts produced by $\mathcal{L}_{\ell_1}$ and $\mathcal{L}_{perc}$, and thus make the recovered image more realistic, which also can be observed in Figure 5.

Based on the aforementioned experience, we further conduct a series of small-scale ablation experiments on loss function, which can be observed in Table 2 (each experiment trains the model 255,000 times). In the first 4 rows, it can be clearly seen that $\mathcal{L}_{\ell_1}$, $\mathcal{L}_D$, $\mathcal{L}_{style}$ and $\mathcal{L}_{perc}$ loss term benefit the quantitative score. However, the simple combination of these loss terms produces the unpleased inpainting results (Figure 5). After weighting the $\mathcal{L}_{style}$ and $\mathcal{L}_D$ to improve the visual effect as well as maintain high quantitive scores, the final objective function is defined as Eq.(5).

$$\mathcal{L}_G = \mathcal{L}_{\ell_1} + 3(-\mathcal{L}_D) + \mathcal{L}_{perc} + 3\mathcal{L}_{style} \tag{5}$$

D. PROGRESSIVE RECONSTRUCTION ALGORITHM

In order to comply with the habit of human painting and improve the robustness of repaired images to large damaged regions, this paper proposes a progressive inpainting strategy that sequentially inputs images and masks into the generative model multiple times during the prediction stage. Specifically, suppose there is a rough recovered image $\mathbf{I}_{comp} = \mathbf{I}_{pred} \odot (1 - M) + \mathbf{I}_{gt} \odot M$ that is obtained from the first inpainting process, then the mask $\mathbf{M}$ is dilated to $\mathbf{M}'$, it means that the size of the damaged area is reduced to a certain extent (it depends on the size of the convolution kernel). Afterward, the data of $\mathbf{I}_{comp} \odot \mathbf{M}'$ is inputted into the models to produce a new recovered image $\mathbf{I}'_{comp}$. By repeating the above process, the final recovered image can be considered as our final output. The whole inpainting process can be defined as Algorithm 1.

$sum(x)$ in Algorithm 1 calculates the sum of the values of the elements in the matrix x, and the $numel(x)$ counts the number of elements in the matrix x.

Intuitively, Figure 6 displays the variations of masks in each dilation operation. It can be observed that the valid region becomes larger as the algorithm processing.

Because the progressive reconstruction scheme takes place in the prediction stage, it improves the results while only slightly increasing the time-consuming. The comparison of prediction latency can be observed in Table 3. In this experiment, the kernel size of each dilation step is set to 15, and the

TABLE 1. **The part of results from hyperparameters search experiments is shown below. Each row represents one conditional model, the $\mathcal{L}_{\ell_1}$, $\mathcal{L}_D$, $\mathcal{L}_{style}$, $\mathcal{L}_{perc}$, $\mathcal{L}_{FM}$ columns respectively refer to the weights of each loss term, PSNR and MAE refer to the quantitative score produced by corresponding conditional models and 3000 test samples.**

$\mathcal{L}_{\ell_1}$	$\mathcal{L}_D$	$\mathcal{L}_{style}$	$\mathcal{L}_{perc}$	$\mathcal{L}_{FM}$	PSNR ↑	MAE(%) ↓
48.081212	5.412144	15.96878	40.149662	0.118075	24.06	4.445
0.604847	0.202847	44.101921	0.387274	0.488302	23.68	4.583
24.159013	66.923958	0.103212	6.038855	206.568582	22.65	5.265
13.484269	30.525771	1.034876	155.314466	32.331865	23.13	5.269
7.700032	3.438013	1.986786	22.0823	0.31213	23.60	4.937
3.229381	12.299439	6.084766	18.260382	14.062116	23.45	4.880
63.075208	0.206318	144.371788	1.288317	6.51374	23.88	4.501
1.110633	15.054847	76.107149	0.145567	0.24907	21.99	5.673
8.017214	1.670086	0.491773	95.361352	70.59277	22.92	5.290
1.74567	49.057089	1.153775	1.412379	0.189158	23.01	4.974

Algorithm 1 The Progressive Reconstruction Algorithm

```
procedure Inpainting($\mathbf{I}_{gt}$, $\mathbf{M}$)
    ISNet and Variables Initialization
    Network Trainning                              ▷ Until convergence
    $i = 0$, $ratio = sum(\mathbf{M} \neq 1)/numel(\mathbf{M})$
    $\mathbf{I} = \mathbf{I}_{gt}$
    while $ratio \geq 0.2 \wedge i \leq 10$ do
        $\mathbf{I}_{pred} = G(\mathbf{I} \odot \mathbf{M})$         ▷ Genrator Prediction
        $i{+}{+}$
        $\mathbf{I}_{comp} = \mathbf{I}_{pred} \odot (1 - \mathbf{M}) + \mathbf{I}_{gt} \odot \mathbf{M}$
        $\mathbf{M} = Dilation(\mathbf{M})$
        $\mathbf{I} = \mathbf{I}_{comp}$
    end while
    return $\mathbf{I}$
end procedure
```

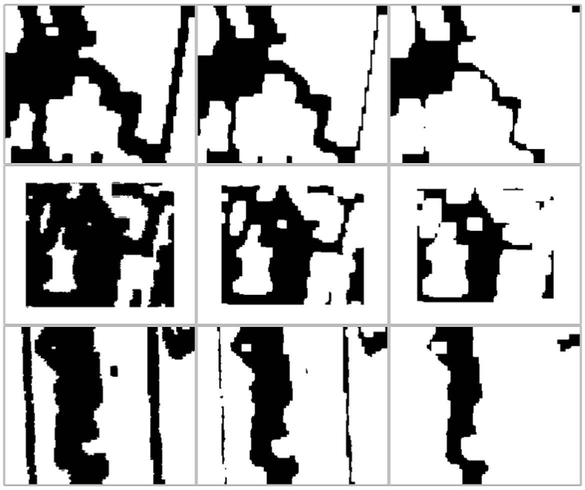

FIGURE 6. **The visualization of mask updating is shown above. The kernel size of dilation is set to 9 in these examples below. The white region represents the valid region, and the black area means the corrupted region.**

prediction latency only includes the time for prediction, not the time to initialize the model, calculate scores, etc. It can be observed that the prediction latencies per image of proposals are less than 0.1 seconds.

TABLE 2. **The ablation study and exploration of the loss function are shown below. Each row represents one conditional model, the $\mathcal{L}_{\ell_1}$, $\mathcal{L}_D$, $\mathcal{L}_{style}$, $\mathcal{L}_{perc}$, $\mathcal{L}_{FM}$ columns respectively refer to the weights of each loss term, PSNR and MAE refer to the quantitative score produced by corresponding conditional models and 5000 inpainting samples.**

$\mathcal{L}_{\ell_1}$	$\mathcal{L}_D$	$\mathcal{L}_{style}$	$\mathcal{L}_{perc}$	$\mathcal{L}_{FM}$	PSNR ↑	MAE(%) ↓
1	1	0	0	0	24.04	0.109
1	1	1	0	0	24.14	0.102
1	1	1	1	0	24.45	0.101
1	1	1	1	1	24.28	0.113
1	3	1	1	0	24.30	0.105
1	10	1	1	0	24.06	0.108
1	3	3	1	0	24.31	0.101

TABLE 3. **The average prediction latency of one-shot prediction and progressive inpainting on the Interactive Separation Network is shown below. These experiments are conducted on 10,000 samples from Places2 [35] test dataset and one 1080TI NVIDIA GPU.**

	ISNet	ISNet+PRA
Latency(ms/image)	52.69	85.46

In order to validate the improvement of the progressive reconstruction algorithm (PRA), we also perform a comparative experiment on the model that is optimized by the objective function of the 5th row in Table 2, which is not too robust to the large damaged region. According to the experiment, it is observed that PRA can visually improve the inpainting performance in the case of a large damaged region. Some of the inpainting results with large holes are selected to display in Figure 7.

IV. EXPERIMENT

A. IMPLEMENTATION AND TRAINING SETUP

The proposed architecture and all supplemental experiments are implemented by Pytorch [36]. All the images in experiments are from two public datasets — Places2 [35] and CelebA [37]. The details of both datasets are shown in Table 4. Therefore these images have a uniform resolution of 256 × 256. All mask maps that marked damaged areas with values 0 are sampled from the NVIDIA public dataset [8] and are resized into the resolution of 256 × 256. And the generative loss and adversarial loss are both optimized by

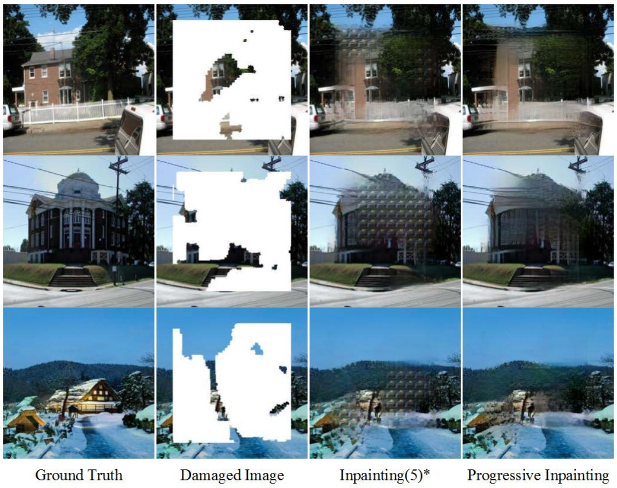

FIGURE 7. The figure shows the effect of the progressive reconstruction algorithm. The last two columns are the results produced by the model that is optimized by the loss component of the 5th row in Table 2, yet the last column is the result using the progressive reconstruction algorithm.

one well-known stochastic descent method — Adam optimizer [38]. The learning rate of the generator is 10^{-4}, and the learning rate of the discriminator is set to one-tenth of the learning rate of the generator, the models in all experiments are trained until their generator convergence. It is worth mentioning that the evaluations in this section only conditional measure the difference between the real testing image $\mathbf{I}_{gt}$ and the combined image $\mathbf{I}_{comp} = \mathbf{I}_{pred} \odot (1 - M) + \mathbf{I}_{gt} \odot M$.

B. QUANTITATIVE COMPARISON AND ANALYSIS

Furthermore, in order to study the effect of different sizes of damaged areas on the inpainting performance, we evaluate the performance of the models under the different sizes of the damaged region using the three following indicators to conduct the quantitative evaluation: the Peak Signal to Noise Ratio (PSNR), Structural Similarity Index (SSIM) [41], and Mean Absolute Error (L1 distances) [26].

Given the ground-truth image and inpainted image, MAE measures the total absolute difference between the pixel values of the ground-truth and the inpainted image. A low MAE computed as Eq. (6) indicates that the quality of the reconstructed image is good.

$$MAE(x, y) = \frac{1}{N}\sum_{i=1}^{N}|x_i - y_i| \tag{6}$$

PSNR is the ratio of the maximum possible value (power signal) to the power of distortion noise that affects the quality representation based on two homogeneous images (reconstructed/original). The higher the PSNR value computed as Eq. (7), the better the quality of the inpainted image, where MAX_I is the maximum fluctuation in the input image data type.

$$MSE(x, y) = \frac{1}{N}\sum_{i=1}^{n}(x_i - y_i)^2$$

TABLE 4. The detailed information of Places2 and CelebA datasets is shown below.

Datasets	Places2	CelebA
Objects	real-world occurrences	celebrity facial
training_set val_set test_set	1,803,460 - 36,000	162,770 19,867 19,962
note	comprising a large and diverse list of the types of environments encountered in the world.	comprising human face attributes including different genders, skin tones, hair colors, and w/o sunglasses, etc.

$$PSNR = 20 \cdot \log_{10}(\frac{MAX_I}{\sqrt{MSE}}) \tag{7}$$

The SSIM models three factors of two images, namely correlation loss, luminance distortion, and contrast distortion. Given the input signals (x,y), SSIM computes the combination of luminance, contrast, and structure to output a similarity measure expressed in Eq. (8). The higher the SSIM value, the better the quality of the predicted image.

$$SSIM(x, y) = \frac{(2\mu_x\mu_y + C_1)(2\sigma_{xy} + C_2)}{(\mu_x^2 + \mu_y^2 + C_1)(\sigma_x)^2 + \sigma_y)^2 + C_2} \tag{8}$$

where C_1 and C_2 are constants, (μ_x, μ_y), (σ_x, σ_y), and σ_{xy} represent mean, standard deviation, and covariance operations.

The quantitative comparison on the Places2 [35] and CelebA [37] test sets are shown on Tables 5 and 6, respectively.

The comparison results show that ISNet can achieve better quantitative performance in all aspects of the indicators regardless of the proportion of damaged areas. It proves the efficiency of the two-branch network in image inpainting tasks and also demonstrates that high-resolution components play an important role in addressing the issue of texture inconsistency in image inpainting. According to the column ‘ISNet(Eq.(4))’ and column ‘ISNet(Eq.(5))’ of Table 5, the proposed loss function (Equation 5) achieves better inpainting performance than using Equation 4 as the objective function. Besides, the results in the column ‘ISNet(Eq.(5)’ and column ‘ISNet(Eq.(5))+PRA’ of Tables 5 and 6 demonstrate the effectiveness of the proposed progressive reconstruction algorithm. Due to the PRA can improve the visual performance of inpainting (see in Figure 7) with slight degeneration of quantitative score, we consider the ISNet+PRA as our main proposal.

C. QUALITATIVE COMPARISON AND OBSERVATION

Figure 8 shows some selected inpainting results from the Places2 test dataset. It can be observed that the proposed approach (ISNet+PRA) produces more visually plausible results and remove the style (color) difference between the filled and non-filled areas. From the yellow box of the second row, the proposed method fills these holes with more reasonable textures, and it makes the inpainted region consistent with the surrounding in terms of color and semantics. For

TABLE 5. The inpainting performance of ISNet on the Places2 test dataset [35], the existing recorded data are taken from the literatures [8], [12] [39], [40], and the records of our proposal are estimated over 10,000 test samples from Places2 dataset, the mask dataset in training and test phase are provided by Liu [8].

Mask Ratio		CA [9]	GLCIC [7]	PConv [8]	EdgeConnect [12]	MED [27]	RFR [29]	WaveFil [23]	CTSDG [28]	ISNet(Eq.(4)	ISNet(Eq.(5))	ISNet(Eq.(5))+PRA
10-20%	PSNR	24.36	23.49	28.02	27.95	28.48	27.75	28.93	30.02	31.41	**31.95**	**31.95**
	SSIM	0.893	0.862	0.869	0.920	0.934	0.929	0.948	0.950	0.960	**0.962**	**0.962**
	MAE(%)	2.05	2.66	1.14	1.31	1.22	1.27	1.19	1.12	0.63	**0.60**	**0.60**
20-30%	PSNR	21.19	20.45	24.9	24.92	25.26	25.12	26.28	26.67	26.88	**27.51**	27.49
	SSIM	0.815	0.771	0.777	0.861	0.902	0.878	0.907	0.911	0.912	**0.918**	**0.918**
	MAE(%)	3.52	4.7	1.98	2.26	2.12	2.19	2.01	1.88	1.38	**1.30**	**1.30**
30-40%	PSNR	19.13	18.5	22.45	22.84	23.15	23.13	23.67	23.77	23.90	**24.59**	24.55
	SSIM	0.739	0.686	0.685	0.799	0.803	0.809	0.837	0.853	0.854	0.864	**0.865**
	MAE(%)	5.07	6.78	3.02	3.25	3.13	3.11	2.65	2.42	2.34	**2.18**	2.19
40-50%	PSNR	17.75	17.17	20.86	21.16	21.17	21.22	21.29	21.31	21.34	**22.01**	21.92
	SSIM	0.662	0.603	0.589	0.731	0.749	0.750	0.763	0.775	0.778	0.794	**0.796**
	MAE(%)	6.62	8.85	4.11	4.39	3.97	3.89	3.83	3.75	3.74	**3.49**	3.54

TABLE 6. The inpainting performance of ISNet over CelebA dataset [37], the records of our proposal are estimated by 5,000 samples from CelebA test dataset, the mask dataset in training and test phase are provided by Liu [8].

Mask Ratio		CA [9]	GLCIC [7]	EdgeConnect [12]	ISNet(Eq.(4))	ISNet(Eq.(5))+PRA
10-20%	PSNR	25.32	24.09	33.51	37.31	**38.95**
	SSIM	0.888	0.865	0.961	0.980	**0.984**
	MAE(%)	2.48	2.53	0.76	0.33	**0.26**
20-30%	PSNR	22.09	20.71	30.02	32.13	**33.39**
	SSIM	0.819	0.773	0.928	0.954	**0.963**
	MAE(%)	3.98	4.67	1.38	0.75	**0.63**
30-40%	PSNR	19.94	18.50	27.39	28.69	**29.75**
	SSIM	0.750	0.689	0.890	0.921	**0.934**
	MAE(%)	5.64	6.95	2.13	1.3	**1.12**
40-50%	PSNR	18.41	17.09	25.28	25.37	**26.15**
	SSIM	0.678	0.609	0.846	0.869	**0.888**
	L1(%)	7.35	9.18	3.03	2.26	**2.04**

the inpainting of the large damaged region, the proposed approach output a more realistic inpainted scene without serious artifacts.

Figure 9 displays the inpainting results of facial reconstruction from the CelebA test dataset. What stands out in the figure is the authenticity of the inpainting results of the proposed method. The proposed approach has the ability to reconstruct high-resolution faces while maintaining the consistency of the color of skin and the authenticity of decoration objects on the head.

D. ABLATION STUDY

Places2 is a public dataset that collected various natural scenes. It consists of more than 1,000,000 training samples and more than 300,000 images for testing. Considering the abundant training data and testing data in the dataset, it's very suitable to conduct the ablation experiments on it. We quantitatively evaluate the inpainted result of the test dataset in three kinds of measurement — PSNR, SSIM, and MAE.

We initially build a simple model prototype without the sub-pixel layers and the design of final fusion (FF). Hence, we consider the simple model as a base model and compare our approach in the cases of whether equipped these modules/local architecture are. It is worth mentioning that the adding of final fusion will increase the learnable weights of the generator model. Aiming to demonstrate that improvement does not just benefit from the increase in weight number, we decrease the channel number of some intermediate layers of the base model when conducting the experiment that adds the final fusion operation. Because the sub-pixel layer is a technic that allows the network adaptively learn upsampling means, it's inevitable to add more learnable weight to the model. Hence, the experiment no more reduces the number of channels when adding the sub-pixel layers. In order to demonstrate the efficiency of ISNets-like generative models, we also introduced a UNet-like [13] generator model from the work [12], and adjusted the number of parameters close to ISNet via adding additional layers to the generator.

TABLE 7. The evaluation of models with different network designs, the ISNet in the third row means the combined design of the base model, final fusion (FF), and Sub-pixel. PSNR, SSIM, and MAE are the three aforementioned indicators. (The 'M' in 'Parameters' column equal to 2^{20}.)

Models	PSNR ↑	SSIM ↑	MAE(%) ↓	Parameters
Base Model	27.91	0.880	2.12	15.75M
Base Model+FF	28.28	**0.886**	2.06	13.93M
ISNet	**28.53**	0.885	**1.91**	23.63M
UNet-like Model	25.99	0.851	2.51	23.27M

As shown in Table 7, it shows the performance of trained models with different network designs on the Places2 test dataset. It's worth noting that all the environments and training setup are constant (e.g., batch size, number of iterations). According to the table, it can be observed that all the techniques used in ISNet can improve the quantitative score. Regardless of the number of parameters, it can be concluded that the ISNets-style models have better quantitative performance than the UNet-like model.

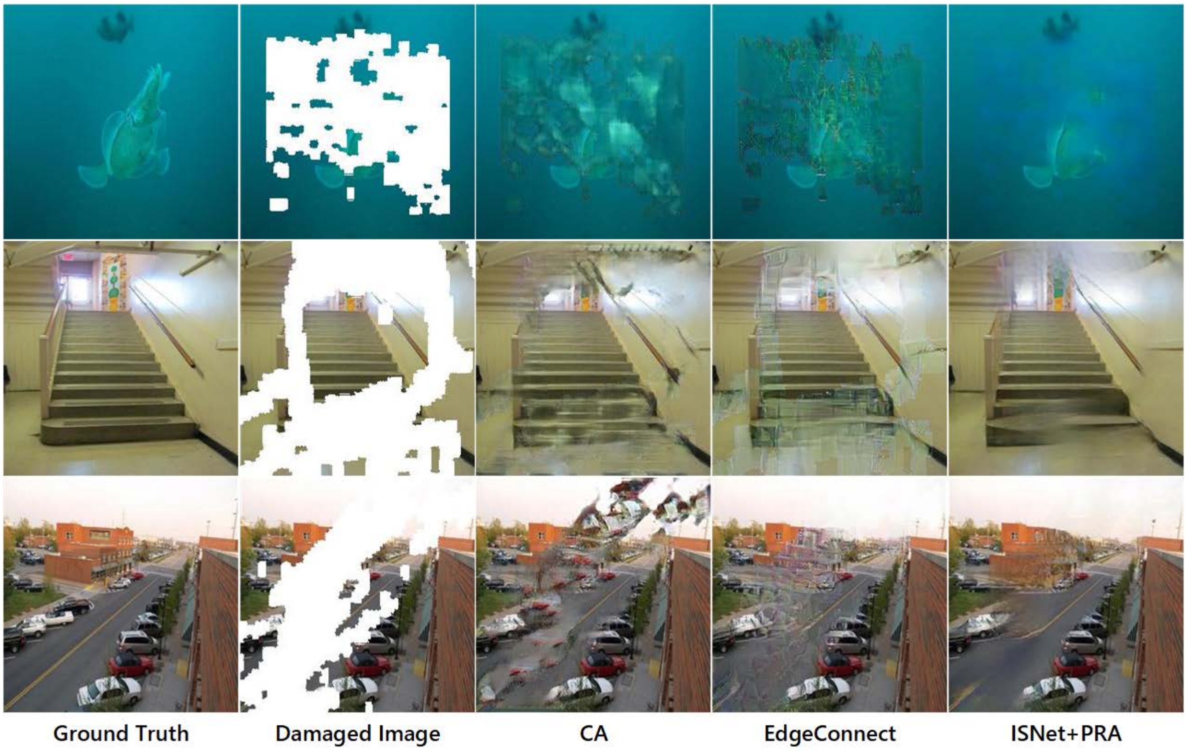

FIGURE 8. Visual comparison of inpainting performance over the Places2 dataset between our method and methods are shown above. In order to observe the generalization of the model, the ground truth images in the figure were sampled from the Places2 test set, and most of the damaged images have a large corrupted area.

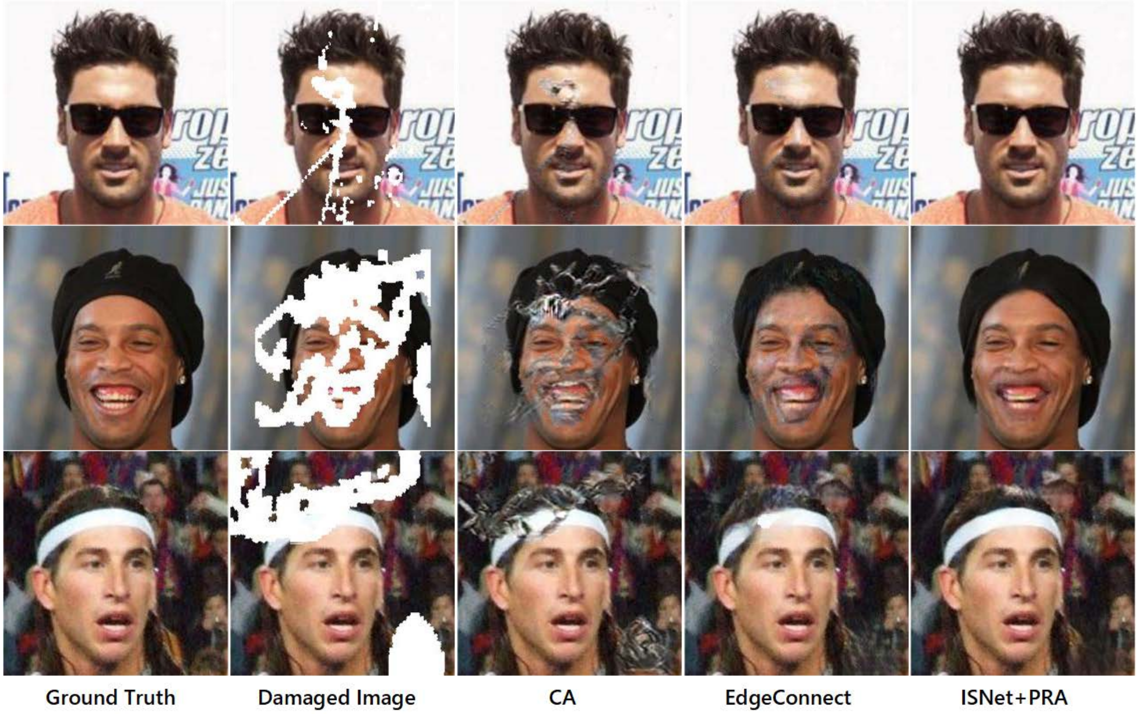

FIGURE 9. This figure shows the qualitative comparison between our method and other image inpainting methods on the CelebA dataset.

Besides structural design, this paper also explores the network performance under different numbers of residual blocks added in each Inpainting process. As shown in Table 8, we evaluate the inpainting performances of ISNet with the different number of residual blocks. The maximum block number is set to 4 due to the limitation of our 11GB GPU

TABLE 8. The table shows the evaluation of ISNet under increasing numbers setting of residual blocks. The 'Blocks' column refers to the numbers of stacked residual blocks in each network phase.

Blocks	PSNR ↑	SSIM ↑	MAE(%) ↓	Parameters
1	28.02	0.884	2.05	18.99M
2	28.47	0.884	1.97	20.53M
3	28.47	0.884	1.97	22.08M
4	**28.53**	**0.885**	**1.91**	23.63M

device. Apart from validating the effectiveness of applying residual blocks, these results also provide a reference for the network deployments requiring diverse RAM resources.

V. CONCLUSION

Due to two-branch generative networks having pivotal roles in the maintenance of low-resolution components and high-resolution information, the proposed method (ISNet) is suitable for the inpainting task and obtains excellent inpainting performance in terms of visually and quantitatively. ISNet performs better than other state-of-the-art approaches on the two public image datasets. On the other hand, this paper explores the trade-off of quantitative performance and visual results in different combinations, the proposed composition of loss function greatly improves the inpainting performance of ISNet. Furthermore, the proposed progressive reconstruction algorithm improves the visual robustness of ISNet for the large-damaged-region inpainting (Figure 7).

ACKNOWLEDGMENT

(Jun Gong and Siyuan Li contributed equally to this work.)

REFERENCES

[1] Y. LeCun, B. Boser, J. S. Denker, D. Henderson, R. E. Howard, W. Hubbard, and L. D. Jackel, "Backpropagation applied to handwritten zip code recognition," *Neural Comput.*, vol. 1, no. 4, pp. 541–551, 1989.

[2] K. Fukushima, "Neocognitron: A self-organizing neural network model for a mechanism of pattern recognition unaffected by shift in position," *Biol. Cybern.*, vol. 36, no. 4, pp. 193–202, Apr. 1980.

[3] I. J. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. C. Courville, and Y. Bengio, "Generative adversarial nets," in *Proc. Adv. Neural Inf. Process. Syst.*, 2014, pp. 2672–2680.

[4] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," in *Proc. Int. Conf. Learn. Represent.*, Y. Bengio and Y. LeCun, Eds., 2015.

[5] W. Liu, D. Anguelov, D. Erhan, C. Szegedy, S. Reed, C.-Y. Fu, and A. C. Berg, "SSD: Single shot MultiBox detector," in *Proc. Eur. Conf. Comput. Vis.*, 2016, pp. 21–37.

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

[7] S. Iizuka, E. Simo-Serra, and H. Ishikawa, "Globally and locally consistent image completion," *ACM Trans. Graph.*, vol. 36, no. 4, p. 107, 2017.

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

[9] 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.

[10] H. V. Vo, N. Q. K. Duong, and P. Pérez, "Structural inpainting," in *Proc. 26th ACM Int. Conf. Multimedia*, Oct. 2018, pp. 1948–1956.

[11] J. Yu, Z. Lin, J. Yang, X. Shen, X. Lu, and T. Huang, "Free-form image inpainting with gated convolution," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 4471–4480.

[12] K. Nazeri, E. Ng, T. Joseph, F. Z. Qureshi, and M. Ebrahimi, "EdgeConnect: Generative image inpainting with adversarial edge learning," *CoRR*, vol. abs/1901.00212, Jan. 2019.

[13] O. Ronneberger, P. Fischer, and T. Brox, "U-Net: Convolutional networks for biomedical image segmentation," in *Proc. Int. Conf. Med. Image Comput. Comput.-Assist. Intervent.*, 2015, pp. 234–241.

[14] K. Sun, B. Xiao, D. Liu, and J. Wang, "Deep high-resolution representation learning for human pose estimation," 2019, *arXiv:1902.09212*.

[15] J. Mao, W. Xu, Y. Yang, J. Wang, and A. L. Yuille, "Deep captioning with multimodal recurrent neural networks (m-RNN)," in *Proc. Int. Conf. Learn. Represent.*, 2015.

[16] O. Vinyals, A. Toshev, S. Bengio, and D. Erhan, "Show and tell: A neural image caption generator," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2015, pp. 3156–3164.

[17] Z. Zhang, W. Yu, J. Zhou, X. Zhang, N. Jiang, G. He, and Z. Yang, "Customizable GAN: A method for image synthesis of human controllable," *IEEE Access*, vol. 8, pp. 108004–108017, 2020.

[18] Y. Taigman, A. Polyak, and L. Wolf, "Unsupervised cross-domain image generation," in *Proc. Int. Conf. Learn. Represent.*, 2017.

[19] N. Pakkaranang, P. Kumam, Y. I. Suleiman, and B. Ali, "Bounded perturbation resilience of viscosity proximal algorithm for solving split variational inclusion problems with applications to compressed sensing and image recovery," *Math. Methods Appl. Sci.*, vol. 45, no. 8, pp. 4085–4107, May 2020.

[20] 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.

[21] J. Xie, L. Xu, and E. Chen, "Image denoising and inpainting with deep neural networks," in *Proc. Adv. Neural Inf. Process. Syst.*, 2012, pp. 341–349.

[22] R. Köhler, C. J. Schuler, B. Schölkopf, and S. Harmeling, "Mask-specific inpainting with deep neural networks," in *Proc. German Conf. Pattern Recognit.*, X. Jiang, J. Hornegger, and R. Koch, Eds., vol. 8753, 2014, pp. 523–534.

[23] Y. Yu, F. Zhan, S. Lu, J. Pan, F. Ma, X. Xie, and C. Miao, "WaveFill: A wavelet-based generation network for image inpainting," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 14114–14123.

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

[25] F. Qi, D. Zhao, and W. Gao, "Reduced reference stereoscopic image quality assessment based on binocular perceptual information," *IEEE Trans. Multimedia*, vol. 17, no. 12, pp. 2338–2344, Dec. 2015.

[26] C. J. Willmott and K. Matsuura, "Advantages of the mean absolute error (MAE) over the root mean square error (RMSE) in assessing average model performance," *Climate Res.*, vol. 30, no. 1, pp. 79–82, Dec. 2005.

[27] H. Liu, B. Jiang, Y. Song, W. Huang, and C. Yang, "Rethinking image inpainting via a mutual encoder-decoder with feature equalizations," in *Proc. Eur. Conf. Comput. Vis.*, 2020, pp. 725–741.

[28] X. Guo, H. Yang, and D. Huang, "Image inpainting via conditional texture and structure dual generation," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 14134–14143.

[29] J. Li, N. Wang, L. Zhang, B. Du, and D. Tao, "Recurrent feature reasoning for image inpainting," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 7760–7768.

[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] W. Shi, J. Caballero, F. Huszár, J. Totz, A. P. Aitken, R. Bishop, D. Rueckert, and Z. Wang, "Real-time single image and video super-resolution using an efficient sub-pixel convolutional neural network," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 1874–1883.

[32] P. Isola, J.-Y. Zhu, T. Zhou, and A. A. Efros, "Image-to-image translation with conditional adversarial networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 1125–1134.

[33] T. Miyato, T. Kataoka, M. Koyama, and Y. Yoshida, "Spectral normalization for generative adversarial networks," in *Proc. Int. Conf. Learn. Represent.*, 2018.

[34] A. Odena, J. Buckman, C. Olsson, T. B. Brown, C. Olah, C. Raffel, and I. J. Goodfellow, "Is generator conditioning causally related to GAN performance?" in *Proc. Int. Conf. Mach. Learn.*, J. G. Dy and A. Krause, Eds., vol. 80, 2018, pp. 3846–3855.

[35] B. Zhou, A. Lapedriza, A. Khosla, A. Oliva, and A. Torralba, "Places: A 10 million image database for scene recognition," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 40, no. 6, pp. 1452–1464, Jun. 2017.

[36] A. Paszke, S. Gross, S. Chintala, G. Chanan, E. Yang, Z. DeVito, Z. Lin, A. Desmaison, L. Antiga, and A. Lerer, "Automatic differentiation in pyTorch," in *Proc. NIPS*, 2017. [Online]. Available:https://openreview.net/pdf?id=BJJsrmfCZ

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

[38] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," in *Proc. Int. Conf. Learn. Represent.*, Y. Bengio and Y. LeCun, Eds., 2015.

[39] R. Zhang, Y. Ren, J. Qiu, and G. Li, "Base-detail image inpainting," in *Proc. Brit. Mach. Vis. Conf.*, 2019, p. 195.

[40] Y. Ma, X. Liu, S. Bai, L. Wang, D. He, and A. Liu, "Coarse-to-fine image inpainting via region-wise convolutions and non-local correlation," in *Proc. 28th Int. Joint Conf. Artif. Intell.*, Aug. 2019, pp. 3123–3129.

[41] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli, "Image quality assessment: From error visibility to structural similarity," *IEEE Trans. Image Process.*, vol. 13, no. 4, pp. 600–612, Apr. 2004.

JUN GONG received the B.E. degree from Tongji University and the M.E. degree from the University of Electronic Science and Technology of China. He is currently pursuing the Ph.D. degree with the Information System and Security & Countermeasures Experimental Center, Beijing Institute of Technology. His current research interests include artificial intelligence, intrusion detection, program analysis, and information security.

SIYUAN LI received the B.S. degree from the Department of Information Science and Engineering, Chengdu University, Chengdu, China, in 2018. He is currently pursuing the M.S. degree with the School of Computer Science and Technology, Southwest University of Science and Technology. He has authored over two papers about image processing in international conferences. His research interests include image inpainting, view synthesis, object recognition, and deep learning. He is a Reviewer of the journal of the IEEE Transactions on Image Processing.

SHIYU CHEN was born in Taizhou, China, in 1997. He received the B.S. degree from the Nanjing Institute of Technology, Nanjing, China, in 2020. He is currently pursuing the M.S. degree in electrical and information engineering with the Southwest University of Science and Technology, Mianyang, China. His research interests include image processing, machine learning, and deep learning.

LIANG NIE received the B.S. degree from the Weifang University of Science and Technology, Weifang, China, in 2019. He is currently pursuing the M.S. degree with the School of Computer Science and Technology, Southwest University of Science and Technology. He has authored over two papers about image processing in international conferences. His research interests include image inpainting, view synthesis, object recognition, and deep learning. He is a Reviewer of the Conference of International Conference on Neural Information Processing.

XIN CHENG (Graduate Student Member, IEEE) was born in Ziyang, Sichuan, China, in 1996. He received the B.S. and M.S. degrees from the Southwest University of Science and Technology, Mianyang, China, in 2016 and 2021, respectively. He is currently pursuing the Ph.D. degree with Hosei University, Tokyo, Japan. His research interests include analytic optimization of placement and other physical design automation.

ZHIQIANG ZHANG received the B.S. and M.S. degrees from the Southwest University of Science and Technology, Mianyang, China, in 2017 and 2020, respectively. He is currently pursuing the Ph.D. degree with Hosei University, Tokyo, Japan. His research interests include image synthesis, multi-modal information transformation and fusion, game theory, computer vision, and deep learning.

WENXIN YU (Member, IEEE) received the B.S. degree from Shanghai Jiaotong University, Shanghai, China, in 2006, and the M.S. and Ph.D. degrees in system LSI from Waseda University, Kitakyushu, Japan, in 2010 and 2014, respectively. He currently works as the Vice Dean of the School of Computer Science and Technology, Southwest University of Science and Technology. He has authored or coauthored over 40 papers in international journals and conferences. His current research interests include 3D multi-view synthesis, image stereo matching, neural networks, pattern recognition, video decoding algorithm, and image error concealment.