Free-Form Image Inpainting via Contrastive Attention Network

Xin Ma[1,2,3], Xiaoqiang Zhou[2,3,4], Huaibo Huang[1,3], Zhenhua Chai[2], Xiaolin Wei[2], Ran He*[1,3]

[1]School of Artificial Intelligence, University of Chinese Academy of Sciences
[2]Vision Intelligence Center, AI Platform, Meituandianping Group
[3]NLPR & CEBSIT, CASIA   [4]University of Science and Technology of China
{xin.ma, huaibo.huang}@cripac.ia.ac.cn, xq525@mail.ustc.edu.cn
{chaizhenhua, weixiaolin02}@meituan.com, rhe@nlpr.ia.ac.cn

Abstract—Most deep learning based image inpainting approaches adopt autoencoder or its variants to fill missing regions in images. Encoders are usually utilized to learn powerful representational spaces, which are important for dealing with sophisticated learning tasks. Specifically, in image inpainting tasks, masks with any shapes can appear anywhere in images (i.e., free-form masks) which form complex patterns. It is difficult for encoders to capture such powerful representations under this complex situation. To tackle this problem, we propose a self-supervised Siamese inference network to improve the robustness and generalization. It can encode contextual semantics from full resolution images and obtain more discriminative representations. we further propose a multi-scale decoder with a novel dual attention fusion module (DAF), which can combine both the restored and known regions in a smooth way. This multi-scale architecture is benefit for decoding discriminative representations learned by encoders into images layer by layer. In this way, unknown regions will be filled naturally from outside to inside. Qualitative and quantitative experiments on multiple datasets, including facial and natural datasets (i.e., Celeb-HQ, Pairs Street View, Places2 and ImageNet), demonstrate that our proposed method outperforms state-of-the-art methods in generating high-quality inpainting results.
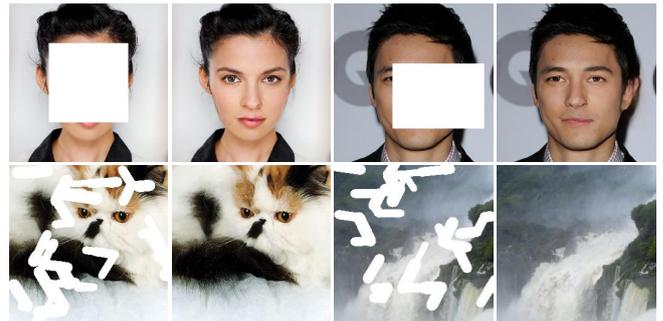
Fig. 1. Images with free-form masks including rectangular and irregular masks and corresponding inpaining results by using our proposed method.

I. INTRODUCTION

Image inpainting (a.k.a image completion or image hole-filling) aims at filling missing regions of an image with plausible contents [1]. It is a fundamental low-level vision task and can be applied to many real-world applications such as photo editing, distracting object removal, image-based rendering, etc [2]–[4]. The core goal of image inpainting is producing semantically meaningful contents in unknown areas, which can incorporate smoothly with known areas.

Traditional exemplar-based methods [3], [5], [6] (e.g., PatchMatch) can gradually synthesize plausible stationary contents by copying and pasting similar patches from known areas. The performances of them are satisfying when dealing with background inpainting tasks. But non-repetitive and complicated scenes, such as faces and objects are the Waterloo of these traditional methods because of limited ability of capturing high-level semantics.

Recently, deep convolutional neural networks (CNNs) have made great progress in many computer vision tasks [7]–[10]. Benefiting from the powerful ability of representation learning of CNNs, many deep learning based methods have been proposed. These approaches adopt autoencoder or its variants architecture jointly trained with generative adversarial networks (GANs) to hallucinate semantically plausible contents in unknown regions [2], [11], [12].

Specifically, masks can be of intricate and irregular patterns and can appear anywhere in images, which greatly increase the difficulty of image inpainting. Previous image inpainting approaches jointly train an encoder and a decoder by some common loss functions (e.g., reconstruct loss, style loss, etc). It is difficult for the encoder to learn powerful representational spaces form images with free-form masks. As a result, these CNN-based approaches will produce depressing results with obvious color contrasts and artifacts especially in boundary areas. A naive way is to design a very deep network to obtain a larger model capacity. However, it heavily increases the computational cost and may not learn accurate latent representational spaces.

For handling this limitation, we propose a self-supervised Siamese inference network with contrastive learning. As explored in [13], [14], contrastive learning trains an encoder to perform a dictionary $look - up$ task. A 'query' encoded by one of the encoders should be similar with its corresponding 'key' (token) that is sampled from data (e.g., image or patch) and is usually represented by another encoder. We assume that two identical images with different masks consist a positive pair while two different images form a negative pair. In order

* indicates the correspondence author

to acquire large and consistent dictionaries that are benefit for representation learning, we follow MoCo [13] to design a queue dictionary and a momentum-update based key encoder. In this way, the robustness and the accuracy (i.e., not producing a noisy representation of the input image) of the encoder can be improved.

Many previous approaches consider image inpainting as a conditional image generation task. The roles of the encoder and decoder are recognizing high-level semantic information and synthesizing low-level pixels [15], respectively. These approaches, e.g., PConv [12], LBAM [11] and Yu's method [2], focus more on missing areas and synthesize realistic alternative contents by a well-designed attention architecture or some effective loss functions. However, due to less attention to structural consistency, there are obvious color contrasts, or artificial edge responses, especially in boundaries of results produced by these methods. In fact, the progress of biology inspired us that the human visual system is sensitive to disharmonious transition regions. Therefore, we pay more attention to the structural continuity of restored images surrounding holes while generating texture-rich images.

To properly suppress color discrepancy and artifacts in boundaries, we propose a novel and independent dual attention fusion module (DAF) to synthesize pixel-wise smooth contents, which can be inserted into autoencoder architectures in a plug-and-play way. The core idea of the fusion module is to calculate the similarity between the synthesized content and the known region. Some methods are proposed to address this problem such as DFNet [16] and Perez's method [17]. But DFNet lacks flexibility in handing different information types (e.g., different semantics), hindering more discriminative representations. Our proposed DAF is developed to adaptively recalibrate channel-wise feature by taking interdependencies between channels into account while force CNNs focusing more on unknown regions. DAF will provide a combining map to blend restored contents and original images in a smooth way.

Qualitative and quantitative experiments on multiple datasets including facial and natural datasets (i.e., Celeb-HQ, Paris SteetView, Places2 and ImageNet) are utilized to evaluate our proposed method. The experimental results demonstrate that our proposed method outperforms state-of-the-art methods in generating high-quality inpainting results. To sum up, the main contributions of this paper are as follows:

- We propose a Siamese inference network based on contrastive learning for free-form image inpainting. It helps to improve the robustness and accuracy of representation learning for complex mask patterns.
- We propose a novel independent dual attention fusion module that can explore feature interdependencies in spatial and channel dimensions and provide a combine map. Smooth contents with rich texture information can be naturally synthesized from outside to inside.
- Our proposed method achieves smooth inpainting results with richer texture information on four standard datasets against state-of-the-art image inpainting methods.

II. RELATED WORK

A. Image Inpainting

Traditional image inpainting methods are mainly diffusion-based [1] or patch-based [3]. Bertalmio *et al.* [1] proposed a algorithm to propagate appearance information from the neighboring region to the unknown regions. In PatchMatch [3], a fast nearest neighbor searching algorithm is proposed to search and paste the most similar image patches from the known regions. These methods utilize low-level image features to guide the feature propagation from known image background or image datasets to corrupted regions. These methods work well when holes are small and narrow, or there are plausible matching patches in known region. However, when suffering from complicated scenes, it is difficult for these approaches to produce semantically plausible solutions, due to a lack of semantic understanding of images.

To accurately recover damaged images, many methods [18]–[22] adopt deep convolutional neural networks (CNNs) [23], especially generative adversarial networks (GANs) [24] in image inpainting. Context Encoder [18] formulates image inpainting as a conditional image generation problem. Global-Local [25] utilizes two discriminators to improve the quality of the generated images at different scales, facilitating both globally and locally consistent image completion. Some works [20], [26], [27] design a coarse-to-fine framework to solve sub-problem of image inpainting in different stages, e.g., Edge connect [26] firstly recovers the edge map of the corrupted image and generates image textures in the second stage.

B. Self-Supervised Representation Learning

Self-supervised representation learning has shown great potential recently in many research works [13], [28], [29]. Compared with supervised learning, self-supervised learning utilizes unlabeled data to learn representations. In MoCo [13] and SimCLR [29], good representations are learned by contrasting positive pairs against negative pairs. In MoCo [13], the authors use a self-supervised learning strategy and achieve comparable performance with supervised methods. Self-supervised learning strategy is also used in many other vision tasks recently. Mustikovela *et al.* [30] used the self-supervised learning to viewpoint learning by taking use of generative consistency and symmetry constraint. Zhan *et al.* [31] utilized a mask completion network to predict occlusion ordering, which is trained with self-supervised learning strategy.

C. Attention Mechanism

Attention mechanism is a hot topic in computer vision and has been widely investigated in many works [32]–[35]. The ways to utilize attention mechanism can be coarsely divided into two categories: spatial attention [32] and channel attention [35]. Yu *et al.* [20] proposed a contextual attention to calculate the spatial attention scores between pixels in corrupted region and known region. DFNet [16] utilizes a spatial alpha composition map to combine features in the corrupted region and known region. In this paper, we investigate both spatial
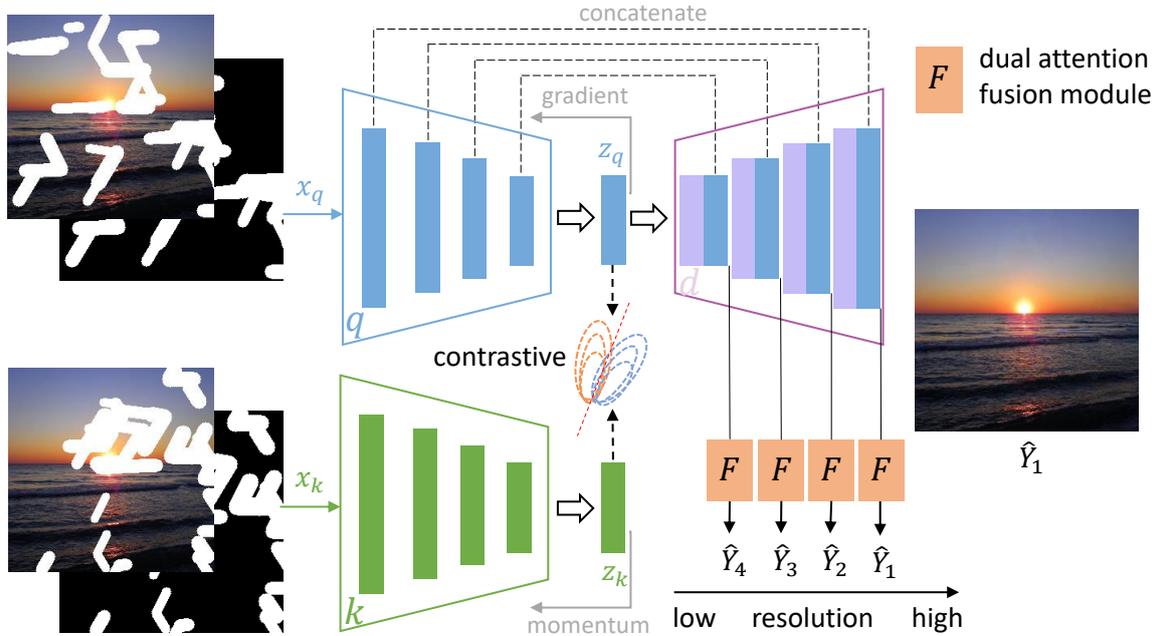
9243

Fig. 2. The network architecture of our method. The self-supervised Siamese inference network consists of encoder E_q and E_k with contrastive learning. This inference network encodes the new key representations on-the-fly by using the momentum-updated encoder E_k. We insert the dual attention fusion module into several decoder layers, thus forming a multi-scale decoder. The inference network is firstly trained on ImageNet with contrastive learning. Then the pre-trained encoder E_q and the decoder are jointly trained with the fusion module.

attention and channel attention mechanism to further improve the performance of image inpainting.

III. METHODOLOGY

In this section, we first present our self-supervised Siamese inference network. Subsequently, the details of the dual attention fusion (DAF) module and learning objectives in our method are provided. The overall framework of our image inpainting method is shown in Fig. 2.

A. Self-supervised Siamese inference network

Our proposed self-supervised Siamese inference network consists of two identical encoders but not sharing parameter weights [13], [14], [36], noted as E_q and E_k, respectively. The proposed inference network is trained by contrastive learning, which can be seen as training an encoder to perform a dictionary look-up task: a 'query' encoded by E_q should be similar with its corresponding 'key' (i.e., positive key) represented by another encoder E_k and dissimilar to others (i.e., negative keys). Two identical images with different masks are required for the proposed inference network, named as x_q and x_k, respectively. Thus, we can obtain a query representation $z_q = E_q(x_q)$ and a key representation $z_k = E_k(x_k)$, respectively. Followed many previous self-supervised works [37], [38], the contrastive loss is utilized as self-supervised objective function for training the proposed inference network and can be written as:

$$\mathcal{L} = -log \frac{exp(z_q.z_k^+/\tau)}{\sum_{i=0}^{k} exp(z_q.z_{k_i}/\tau)}, \qquad (1)$$

where τ is the temperature hyper-parameter and it degrades into the original $softmax$ when τ is equal to 1. The output will be less sparse with τ increasing [39]. The τ is set as 0.07 for efficient training process in this work. Specially, this loss, also known as InfoNCE loss [13], [14], tries to classify z_q as z_k^+.

High-dimensional continuous images can be projected into a discrete dictionary by contrastive learning. There are three general mechanisms for implementing contrastive learning (i.e., end-to-end [14], memory bank [40] and momentum updating [13]), whose main differences are how to maintain keys and how to update the key encoder. Considering GPU memory size and powerful feature learning, we follow MoCo [13] to design a consistent dictionary implemented by $queue$. Thus, the key representations of the current batch data are enqueued into the dictionary while the older representations are dequeued progressively. The length of the queue is under control, which enable the dictionary to contain a large number of negative image pairs. Specially, representation learning can be beneficial from a dictionary with a large scale negative pairs. We set the length of the queue as 65536 in this work.

It is worth noting that the encoder E_k is updated by a momentum update strategy instead of direct back-propagation. The main reason is that it's difficult to propagate the gradient to all keys in the queue. The updating process of E_k can be formulated as follows:

$$\theta_k \leftarrow m\theta_k + (1-m)\theta_q, \qquad (2)$$

where θ_q and θ_k are denoted as the parameters of E_q and E_k, respectively. θ_q is updated by back-propagation. $m \in [0,1)$ is

the momentum coefficient hyper-parameter and set as 0.9 in this paper. In this way, the encoder E_k can update smoothly relative to E_q, resulting in a more consistent discrete dictionary.

B. Dual Attention Fusion Module

We now give more details about our proposed dual attention fusion module (see Fig. 3), which contains a channel attention mechanism and a spatial attention mechanism. This fusion module are embedded to the last few layers of the decoder and output inpainting results with multi-scale resolutions [41], [42]. Thus, constraints can be imposed on multi-scale outputs for high-quality results.

Previous CNN-based image inpainting approaches treat channel-wise features equally, thus hindering the ability of the representation learning of the network. Meanwhile, high-level and interrelated channel features can be considered as specific class responses. For more discriminative representations, we first build a channel attention module in our proposed fusion module.
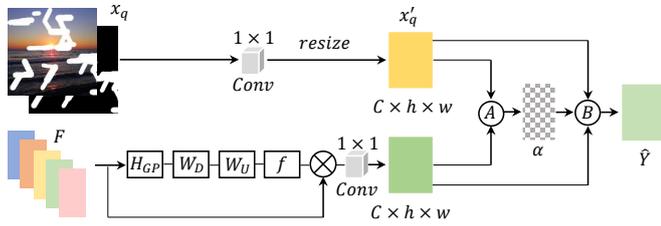

Fig. 3. The network architecture of the dual attention fusion module. It first predicts an α combine map with the learnable transformation function $\mathcal{A}$. Then we can obtain final smooth inpainting results with rich texture by the combine function $\mathcal{B}$.

As shown in Fig. 3, let a feature map $F = [f_1, \cdots, f_c, \cdots, f_C]$ be one of the inputs of the fusion module, whose channel number is c and size is $h \times w$. The channel descriptor can be acquired from the channel-wise global spatial information by global averaging pooling. Then we can obtain the channel-wise statistics $z_c \in \mathbb{R}^c$ by shrinking F:

$$z_c = H_{GP}(x_c) = \frac{1}{h \times w} \sum_{i=1}^{h} \sum_{j=1}^{w} f_c(i, j), \qquad (3)$$

where $z_c(i, j)$ is the c-th element of z. H_{GP} means the global pooling function.

In order to fully explore channel-wise dependencies from the aggregated information by using global average pooling, we introduce a gating mechanism. As illustrated in [35], [43], the sigmoid function can be used as a gating function:

$$\omega = f(W_U \delta(W_D z)), \qquad (4)$$

where $f(\cdot)$ and $\delta(\cdot)$ are the sigmoid gating and ReLU functions, respectively. W_D and W_U are the weight sets of $Conv$ layers. They set channel number as C/r and C, respectively. Finally, the channel statistics ω is acquired and used to rescale the input f_c:

$$\hat{f}_c = w_c \cdot f_c, \qquad (5)$$

where w_c and f_c are the scaling factor and feature map of the c-th channel, respectively.

The long-range contextual information is essential for discriminant feature representations. We propose a spatial attention module that forms the final part of the proposed fusion module. Given an input image with a mask x_q, we first get x_q' that matches the size of the re-scaled feature map $\hat{F} \in \mathbb{R}^{c \times h \times w}$,

$$x_q' = (W_C x_q) \downarrow, \qquad (6)$$

where W_C and $\downarrow$ are the weight set of a 1×1 convolutional layer and down-scaled module, respectively.

Then the combining map $\alpha \in \mathbb{R}^{C \times h \times w}$ is given by,

$$\alpha = f(\mathcal{A}(W_D \hat{F}, x_q')), \qquad (7)$$

where W_D is the weight set of a 1×1 convolutional layer. It sets channel number of $\hat{F}$ to be same with x_q'. $\mathcal{A}$ is a learnable transformation function implemented by three 3×3 convolutional layers. $W_D \hat{F}$ and x_q' are first concatenated and then fed into convolutional layers. $f(\cdot)$ is the sigmoid function that can make α an attention map to some extent.

The final inpainting result $\hat{Y}$ is obtained by,

$$\hat{Y} = \mathcal{B}(\alpha, W_D \hat{F}, x_q') = \alpha \odot W_D \hat{F} + (1 - \alpha) \odot x_q', \quad (8)$$

where $\odot$ and $\mathcal{B}$ denote the Hadamard product and combine function, respectively. In this way, we can eliminate obvious color contrasts and artifacts especially in boundary areas and get a smoother inpainting results with richer texture.

C. Loss Functions

As explored in [11], [16], [22], for synthesizing richer texture details and correct semantics, the element-wise reconstruction loss, the perceptual loss [44], the style loss, the total variation loss and the adversarial loss are used to train our proposed method.

Reconstruction Loss. It is calculated as $\mathcal{L}_1$-norm between the inpainting result $\hat{Y}$ and the target image Y,

$$\mathcal{L}_{rec} = ||Y - \hat{Y}||_1. \qquad (9)$$

Perceptual Loss. Due to lacking high-level semantics by only using the element-wise loss, we introduce the perceptual loss,

$$\mathcal{L}_{per} = \frac{1}{N} \sum_{i=1}^{N} ||\Phi^i(Y) - \Phi^i(\hat{Y})||_1, \qquad (10)$$

where Φ is the VGG-16 network that pre-trained on ImageNet [45]. $\Phi^i(\cdot)$ outputs feature maps of the i-th pooling layer. We select $pool-1$, $pool-2$ and $pool-3$ layers of the pre-trained VGG-16 in this work.

Style Loss. For getting richer textures, we also adopt the style loss defined on the feature maps produced by the pre-trained VGG-16. Followed [11], [12], the style loss can be calculated as $\mathcal{L}_1$-norm between the Gram matrices of the feature maps,

$$\mathcal{L}_{style} = \frac{1}{N} \sum_{i=1}^{N} \frac{1}{C_i \cdot C_i} ||\Phi^i(Y)(\Phi^i(Y))^T - \Phi^i(\hat{Y})(\Phi^i(\hat{Y}))^T||_1,$$

$$(11)$$

9245

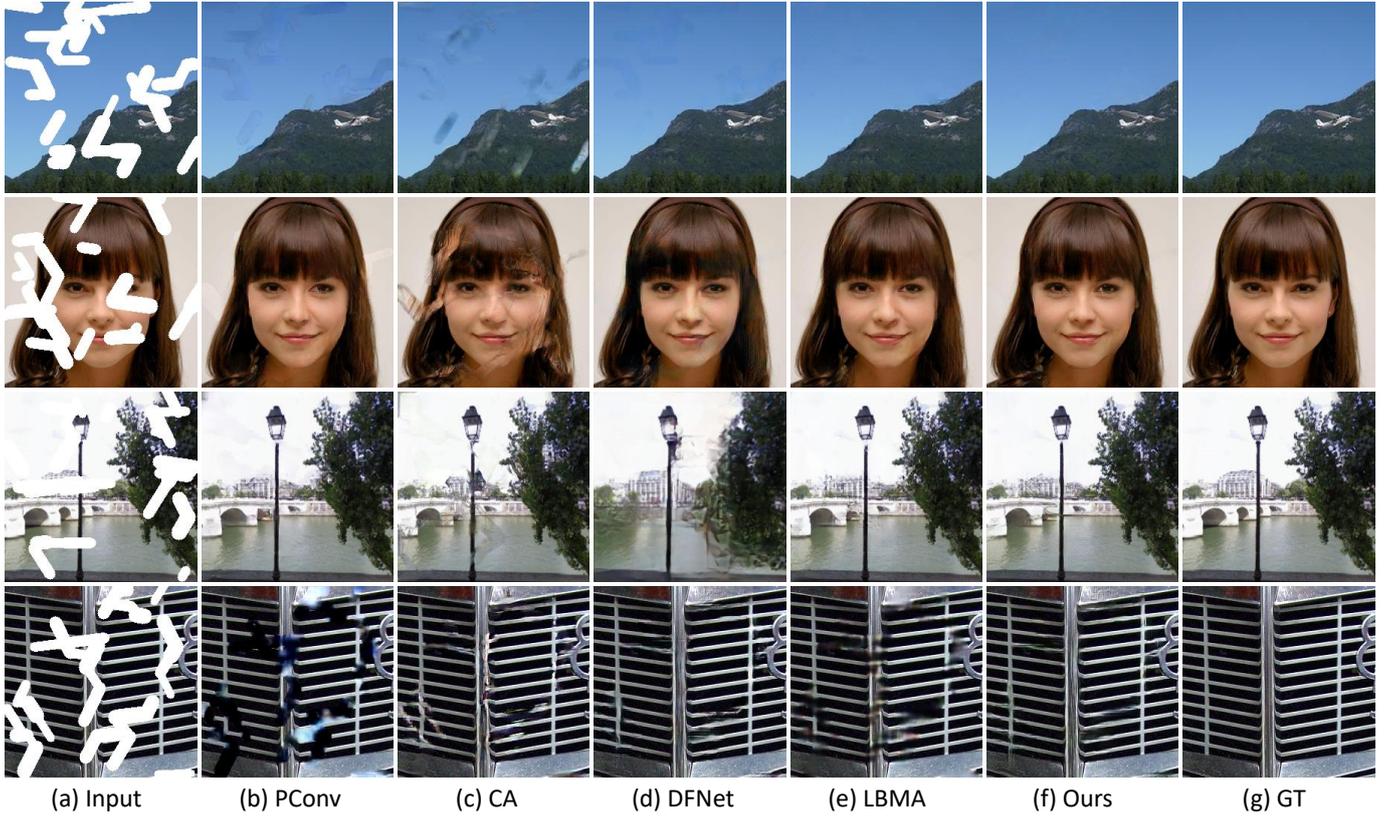| (a) Input | (b) PConv | (c) CA | (d) DFNet | (e) LBMA | (f) Ours | (g) GT |

Fig. 4. Qualitative experiments compared with state-of-the-arts on four datasets with free-form masks. (a) is the input with irregular mask. (b), (c), (d), (e) and (f) are the results generated by PConv [12], CA [15], DFNet [16], LBMA [11] and ours method respectively from left to right. (g) is the ground truth.

where C_i denotes the channel number of the feature map at i-th layer in the pre-trained VGG-16.

Total Variation Loss. It is a smoothing penalty term and calculated on the region of 1-pixel dilation of the unknown regions.

$$\mathcal{L}_{tv} = \frac{1}{N} \sum_{(i,j),(i,j+1)\in\Omega} ||\hat{Y}^{i,j+1} - \hat{Y}^{i,j}||_1 \\ + \frac{1}{N} \sum_{(i,j),(i+1,j)\in\Omega} ||\hat{Y}^{i+1,j} - \hat{Y}^{i,j}||_1, \quad (12)$$

where Ω means the unknown regions.

Adversarial Loss. We introduce the adversarial loss for improving the visual quality of the generated inpainting image. Followed [46], it can be formulated as,

$$\mathcal{L}_{adv} = \min_G \max_D \mathbb{E}_{Y\sim P_Y} D(Y) - \mathbb{E}_{\hat{Y}\sim P_{\hat{Y}}} D(\hat{Y}) \\ + \lambda \mathbb{E}_{Y'\sim P_{Y'}}((||\nabla_{Y'} D(Y')||)^2) - 1)^2, \quad (13)$$

where $D(\cdot)$ means the discriminator. Y' is the resized version that sampled from $\hat{Y}$ and Y by interpolation with a random scale factor. We set λ as 10 in this work.

Model Objective. Taking the above loss functions and the multi-scale network architecture into account, we group them into two categories: $Structure\ Loss$ and $Texture\ Loss$, respectively.

$$\mathcal{L}_{struct}^k = \lambda_{rec}\mathcal{L}_{rec}^k, \quad (14)$$

where λ_{rec} means the weight factor and set as 6. $\mathcal{L}_{struct}^k$ is calculated as $\mathcal{L}_{rec}$ at the k-th layer of the decoder (1_{st} means the last decoder layer).

The $Texture\ Loss$ is given by,

$$\mathcal{L}_{text}^k = \lambda_{per}\mathcal{L}_{per}^k + \lambda_{style}\mathcal{L}_{style}^k + \lambda_{tv}\mathcal{L}_{tv}^k + \lambda_{adv}\mathcal{L}_{adv}^k, \quad (15)$$

where λ_{per}, λ_{style}, λ_{tv}, λ_{adv} are trade-off factors and set as 0.1, 240, 0.1 and 0.001 respectively in this work.

Finally, the total model objective is given by,

$$\mathcal{L}_{total} = \frac{1}{|P|} \sum_{k\in P} \mathcal{L}_{struct}^k + \frac{1}{|Q|} \sum_{k\in Q} \mathcal{L}_{text}^k, \quad (16)$$

where both P and Q are the selected decoder layer sets that imposed constraints. We select P as $\{1, 2, 3, 4, 5, 6\}$ and Q as $\{1, 2, 3\}$ respectively for better quality inpainting results. Note that 1 represents the outermost layer.

IV. EXPERIMENTS

To demonstrate the superiority of our approach against state-of-the-art image inpainting methods, both quantitative and qualitative experiments are conducted. In this section, we will introduce the details of our experimental settings and the experimental results one by one.

A. Experiments Settings

Datasets. We conduct experiments on several public datasets including:

- CelebA-HQ [41], a dataset that contains 30,000 high-quality face images.
- Places2 [47], a dataset that contains over 8,000,000 images from over 365 scenes collected from the natural world.
- Paris Street View [48], a dataset which consists of 15,000 images collected from street views of Paris
- ImageNet [45], a large visual database containing more than 14 million images of the world.

For Places2, Pairs Street View and ImageNet, we use their origin splits for validation and testing. As for CelebA-HQ, we randomly select 28000 images for training and the rest for testing.
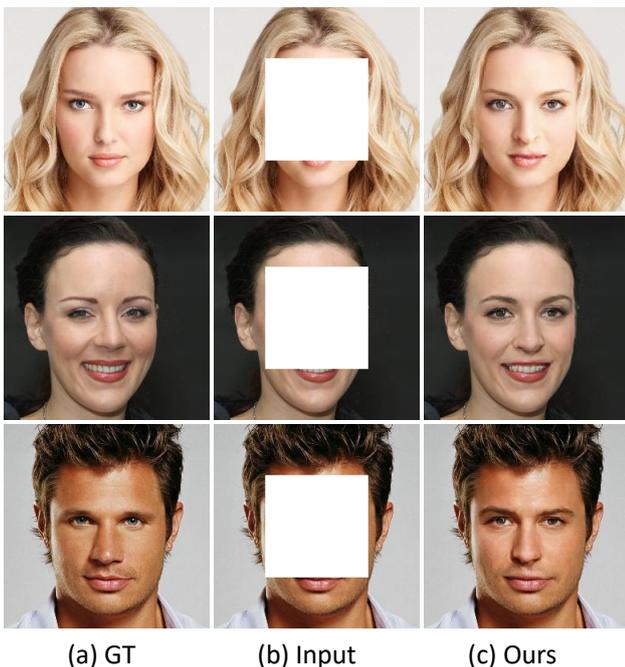

| (a) GT | (b) Input | (c) Ours |

Fig. 5. Example results predicted by our proposed method on CelebA-HQ.

Implementation Details. All the images are resized to 256×256 during training and testing. Data augmentation is utilized during training such as flipping. Masks are free-form including rectangular and irregular and automatically generated on-the-fly during training [22]. Taken as a whole, our proposed method can be broken down into two stages. In the first stage, the inference network is trained on ImageNet through contrastive learning until convergence. And in the next stage, the pre-trained encoder and the decoder are jointly trained with the fusion module. We use the SGD optimizer with the learning rate as 0.015 for training the Siamese inference network while use the Adam optimizer with the learning rate as 10^{-4} for jointly training the encoder and decoder. All the results are reported directly from the trained model without any additional post-processing.

Method	Mask Type	$\mathcal{L}_1$ error†	PSNR‡	SSIM‡	FID†
PConv [12]		0.0357	22.83	0.7737	26.45
CA [15]		0.0372	22.58	0.7800	21.78
DFNet [16]	rectangular	0.0349	22.95	0.7789	26.73
LBMA [11]		0.0350	23.08	0.7817	24.01
Ours		**0.0332**	**23.52**	**0.7928**	**21.42**
PConv [12]		0.0202	26.65	0.9224	12.03
CA [15]		0.0384	22.08	0.8182	28.69
DFNet [16]	irregular	0.0180	27.40	0.9343	10.12
LBMA [11]		0.0179	27.43	0.9351	10.56
Ours		**0.0178**	**27.54**	**0.9350**	**9.15**

TABLE I
QUANTITATIVE COMPARISON ON VALIDATION IMAGES OF PLACES2 WITH BOTH RECTANGLE AND IRREGULAR MASKS. †LOWER IS BETTER. ‡HIGHER IS BETTER.

B. Quantitative Results

We conduct quantitative experiments on the Places2 dataset with free-form masks. The $L_1 \ loss$, peak signal-to-noise ratio (PSNR), structural similarity index (SSIM) and Fré'chet Inception Distance (FID) are used for evaluation metrics.

The ability of the model to recover the original content in the hole can be roughly reflected by $\mathcal{L}_1 \ error$. PSNR and SSIM measure the similarity between the inpainting result and the target image. As for FID, it can measure the Wasserstein-2 distance between real and inpainting images through the pre-trained Inception-V3.

Table I shows the performance of our proposed method against other state-of-the-art methods. Our method outperforms all the other compared methods in four metrics. The main reasons are that 1): images with free-form masks dramatically increase the difficulty of image inpainting, thus hindering the ability of the representation learning of the encoder; 2): exiting methods firstly take generating realistic images into account but ignore the structural consistency of the generated image.

C. Qualitative Results

We compare our proposed method with state-of-the-art methods in term of visual and semantic coherence. We conduct qualitative experiments on the test set of four datasets with free-form masks. As shown in Fig. 4, we mask the test images with irregular masks. It can be seen that PConv, CA, DFNet and LBMA tend to synthesise blurred and unsmooth final results. Our proposed method can generate smoother inpainting results with reasonable semantics and richer textures with the help of the self-supervised Siamese inference network and the DAF module. It demonstrates that our proposed method is superior to the comparison methods in terms of consistent structures and colors. Furthermore, as shown in Fig. 5, we also conduct experiments on the test images of CelebA-HQ with typical rectangular squares to evaluate the inpainting ability of our proposed method. Our method can generate face images with consistent colors and structures.

D. Ablation Study

The multi-scale decoder can progressively refine the inpainting results at each scale. The experiments are conducted on test

contrastive learning	$\times$	$\checkmark$	$\times$	$\checkmark$
DAF	$\times$	$\times$	$\checkmark$	$\checkmark$
$\mathcal{L}_1$ error†	0.0192	0.0190	0.0186	**0.0178**
PSNR‡	26.98	27.13	27.21	**27.54**
SSIM‡	0.9282	0.9302	0.9315	**0.9351**
FID†	12.62	12.56	11.14	**9.15**

TABLE II

ABLATION STUDY EXPERIMENTS ON ON VALIDATION IMAGES OF PLACES2. †LOWER IS BETTER. ‡HIGHER IS BETTER.

images of CelebA-HQ. Then we visualize the images predicted by decoder at several scales (i.e., from 1_{st} to 5_{th}). As shown in Fig. 6, it demonstrates that this multi-scale architecture is benefit for decoding learned representations into generated images layer by layer.

We also investigate the effectiveness of different components of the proposed method. We train several variants of the proposed method: remove the self-supervised Siamese inference network (denote as contrastive learning) and/or the DAF module. As shown in Table II, it clearly demonstrates that the inference network and the DAF module play important roles in determining the performance. As shown in Fig. 7, the uncompleted models usually generate obvious artifacts, especially in boundaries while our full model can suppress color discrepancy and artifacts in boundaries and produce realistic inpainting results.
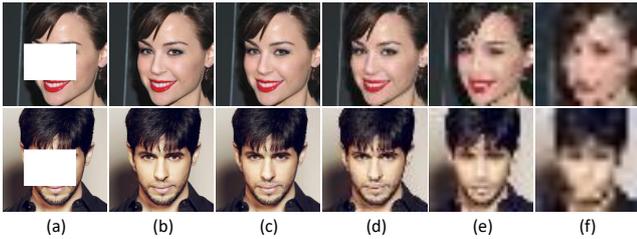

Fig. 6. Images produced by the multi-scale decoder. (a) is the input with a rectangle mask. (b) is the final inpainting result. (c), (d), (e) and (f) are outputs at multi-scales.
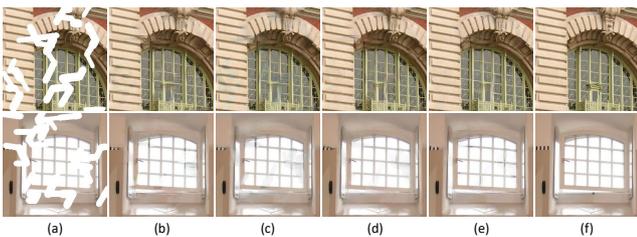

Fig. 7. Images produced by the variants of our proposed method. (a) is the input with irregular masks. (e) is the inpainting result generated by the full model. (b), (c) and (d) are generated by the model without DAF and contrastive learning, the model only with contrastive learning and the model only with DAF respectively. (f) is the ground truth.

V. CONCLUSION

In this paper, we propose a novel two-stage paradigm image inpainting method to generate smoother results with reasonable semantics and richer textures. Specifically, the proposed method boosts the ability of the representation learning of the inference network by using contrastive learning. We further design a novel dual attention fusion module to form a multi-scale decoder, which can be embedded into decoder layers in a plug-and-play way. Experiments on CelebA-HQ, Places2, Pairs Street View and ImageNet show the superiority of our proposed method in generating smoother, more coherent and fine-detailed results.

REFERENCES

[1] M. Bertalmio, G. Sapiro, V. Caselles, and C. Ballester, "Image inpainting," in *Proceedings of the 27th annual conference on Computer graphics and interactive techniques*, 2000, pp. 417–424.

[2] J. Yu, Z. Lin, J. Yang, X. Shen, X. Lu, and T. S. Huang, "Free-form image inpainting with gated convolution," in *Proceedings of the IEEE International Conference on Computer Vision*, 2019, pp. 4471–4480.

[3] C. Barnes, E. Shechtman, A. Finkelstein, and D. B. Goldman, "Patchmatch: A randomized correspondence algorithm for structural image editing," *ACM Transactions on Graphics*, vol. 28, no. 3, pp. 24–33, 2009.

[4] D. Simakov, Y. Caspi, E. Shechtman, and M. Irani, "Summarizing visual data using bidirectional similarity," in *IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, 2008, pp. 1–8.

[5] A. A. Efros and W. T. Freeman, "Image quilting for texture synthesis and transfer," in *Proceedings of the 28th annual conference on Computer graphics and interactive techniques*, 2001, pp. 341–346.

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

[7] C. Finn, P. Abbeel, and S. Levine, "Model-agnostic meta-learning for fast adaptation of deep networks," *arXiv preprint arXiv:1703.03400*, 2017.

[8] H. Liu, K. Simonyan, and Y. Yang, "Darts: Differentiable architecture search," *arXiv preprint arXiv:1806.09055*, 2018.

[9] J. Cao, Y. Hu, H. Zhang, R. He, and Z. Sun, "Learning a high fidelity pose invariant model for high-resolution face frontalization," in *Advances in neural information processing systems*, 2018, pp. 2867–2877.

[10] H. Huang, R. He, Z. Sun, and T. Tan, "Wavelet domain generative adversarial network for multi-scale face hallucination," *International Journal of Computer Vision*, vol. 127, no. 6-7, pp. 763–784, 2019.

[11] C. Xie, S. Liu, C. Li, M.-M. Cheng, W. Zuo, X. Liu, S. Wen, and E. Ding, "Image inpainting with learnable bidirectional attention maps," in *Proceedings of the IEEE International Conference on Computer Vision*, 2019, pp. 8858–8867.

[12] G. Liu, F. A. Reda, K. J. Shih, T.-C. Wang, A. Tao, and B. Catanzaro, "Image inpainting for irregular holes using partial convolutions," in *Proceedings of the European Conference on Computer Vision*, 2018, pp. 85–100.

[13] K. He, H. Fan, Y. Wu, S. Xie, and R. Girshick, "Momentum contrast for unsupervised visual representation learning," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2020, pp. 9729–9738.

[14] R. Hadsell, S. Chopra, and Y. LeCun, "Dimensionality reduction by learning an invariant mapping," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, vol. 2. IEEE, 2006, pp. 1735–1742.

[15] J. Yu, Z. Lin, J. Yang, X. Shen, X. Lu, and T. S. Huang, "Generative image inpainting with contextual attention," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 5505–5514.

[16] X. Hong, P. Xiong, R. Ji, and H. Fan, "Deep fusion network for image completion," in *Proceedings of the 27th ACM International Conference on Multimedia*, 2019, pp. 2033–2042.

[17] P. Pérez, M. Gangnet, and A. Blake, "Poisson image editing," in *ACM SIGGRAPH 2003 Papers*, 2003, pp. 313–318.

[18] D. Pathak, P. Krahenbuhl, J. Donahue, T. Darrell, and A. A. Efros, "Context encoders: Feature learning by inpainting," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 2536–2544.

[19] S. Zhang, R. He, Z. Sun, and T. Tan, "Demeshnet: Blind face inpainting for deep meshface verification," *IEEE Transactions on Information Forensics and Security*, vol. 13, no. 3, pp. 637–647, 2017.

[20] J. Yu, Z. Lin, J. Yang, X. Shen, X. Lu, and T. S. Huang, "Generative image inpainting with contextual attention," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 5505–5514.

[21] Y. Wang, X. Tao, X. Qi, X. Shen, and J. Jia, "Image inpainting via generative multi-column convolutional neural networks," in *Advances in Neural Information Processing Systems*, 2018, pp. 331–340.

[22] J. Yu, Z. Lin, J. Yang, X. Shen, X. Lu, and T. S. Huang, "Free-form image inpainting with gated convolution," in *Proceedings of the IEEE International Conference on Computer Vision*, 2019, pp. 4471–4480.

[23] T.-C. Wang, M.-Y. Liu, J.-Y. Zhu, A. Tao, J. Kautz, and B. Catanzaro, "High-resolution image synthesis and semantic manipulation with conditional gans," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018.

[24] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial nets," in *Advances in Neural Information Processing Systems*, 2014, pp. 2672–2680.

[25] S. Iizuka, E. Simo-Serra, and H. Ishikawa, "Globally and locally consistent image completion," *ACM Transactions on Graphics*, vol. 36, no. 4, pp. 1–14, 2017.

[26] K. Nazeri, E. Ng, T. Joseph, F. Z. Qureshi, and M. Ebrahimi, "Edge-connect: Generative image inpainting with adversarial edge learning," *arXiv preprint arXiv:1901.00212*, 2019.

[27] Y. Ren, X. Yu, R. Zhang, T. H. Li, S. Liu, and G. Li, "Structureflow: Image inpainting via structure-aware appearance flow," in *Proceedings of the IEEE International Conference on Computer Vision*, 2019, pp. 181–190.

[28] X. Zhan, X. Pan, B. Dai, Z. Liu, D. Lin, and C. C. Loy, "Self-supervised scene de-occlusion," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, June 2020.

[29] T. Chen, S. Kornblith, M. Norouzi, and G. E. Hinton, "A simple framework for contrastive learning of visual representations," *arXiv: Learning*, 2020.

[30] S. K. Mustikovela, V. Jampani, S. D. Mello, S. Liu, U. Iqbal, C. Rother, and J. Kautz, "Self-supervised viewpoint learning from image collections," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2020, pp. 3971–3981.

[31] X. Zhan, X. Pan, B. Dai, Z. Liu, D. Lin, and C. C. Loy, "Self-supervised scene de-occlusion," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2020, pp. 3784–3792.

[32] X. Wang, R. Girshick, A. Gupta, and K. He, "Non-local neural networks," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 7794–7803.

[33] T. Dai, J. Cai, Y. Zhang, S.-T. Xia, and L. Zhang, "Second-order attention network for single image super-resolution," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2019, pp. 11 065–11 074.

[34] J. Fu, J. Liu, H. Tian, Y. Li, Y. Bao, Z. Fang, and H. Lu, "Dual attention network for scene segmentation," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2019, pp. 3146–3154.

[35] J. Hu, L. Shen, and G. Sun, "Squeeze-and-excitation networks," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 7132–7141.

[36] X. Wang and A. Gupta, "Unsupervised learning of visual representations using videos," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2015, pp. 2794–2802.

[37] P. Bachman, R. D. Hjelm, and W. Buchwalter, "Learning representations by maximizing mutual information across views," in *Advances in Neural Information Processing Systems*, 2019, pp. 15 535–15 545.

[38] C. Zhuang, A. L. Zhai, and D. Yamins, "Local aggregation for unsupervised learning of visual embeddings," in *Proceedings of the IEEE International Conference on Computer Vision*, 2019, pp. 6002–6012.

[39] Y. Chen, X. Dai, M. Liu, D. Chen, L. Yuan, and Z. Liu, "Dynamic convolution: Attention over convolution kernels," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2020, pp. 11 030–11 039.

[40] Z. Wu, Y. Xiong, S. X. Yu, and D. Lin, "Unsupervised feature learning via non-parametric instance discrimination," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 3733–3742.

[41] T. Karras, T. Aila, S. Laine, and J. Lehtinen, "Progressive growing of gans for improved quality, stability, and variation," *arXiv preprint arXiv:1710.10196*, 2017.

[42] A. Brock, J. Donahue, and K. Simonyan, "Large scale gan training for high fidelity natural image synthesis," *arXiv preprint arXiv:1809.11096*, 2018.

[43] Y. Zhang, K. Li, K. Li, L. Wang, B. Zhong, and Y. Fu, "Image super-resolution using very deep residual channel attention networks," in *Proceedings of the European Conference on Computer Vision*, 2018, pp. 286–301.

[44] J. Johnson, A. Alahi, and L. Fei-Fei, "Perceptual losses for real-time style transfer and super-resolution," in *Proceedings of the European Conference on Computer Vision*. Springer, 2016, pp. 694–711.

[45] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "Imagenet: A large-scale hierarchical image database," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. Ieee, 2009, pp. 248–255.

[46] I. Gulrajani, F. Ahmed, M. Arjovsky, V. Dumoulin, and A. C. Courville, "Improved training of wasserstein gans," in *Advances in Neural Information Processing Systems*, 2017, pp. 5767–5777.

[47] B. Zhou, A. Lapedriza, A. Khosla, A. Oliva, and A. Torralba, "Places: A 10 million image database for scene recognition," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 40, no. 6, pp. 1452–1464, 2017.

[48] C. Doersch, S. Singh, A. Gupta, J. Sivic, and A. Efros, "What makes paris look like paris?" *Communications of the ACM*, pp. 103–110, 2015.