Aggregated Contextual Transformations for High-Resolution Image Inpainting

Yanhong Zeng, *Student Member, IEEE*, Jianlong Fu, *Senior Member, IEEE*, Hongyang Chao, *Member, IEEE*, and Baining Guo, *Fellow, IEEE*

Abstract—Image inpainting that completes large free-form missing regions in images is a promising yet challenging task. State-of-the-art approaches have achieved significant progress by taking advantage of generative adversarial networks (GAN). However, these approaches can suffer from generating distorted structures and blurry textures in high-resolution images (e.g., 512×512). The challenges mainly drive from (1) image content reasoning from distant contexts, and (2) fine-grained texture synthesis for a large missing region. To overcome these two challenges, we propose an enhanced GAN-based model, named **A**ggregated C**O**ntextual-**T**ransformation GAN (**AOT-GAN**), for high-resolution image inpainting. Specifically, to enhance context reasoning, we construct the generator of AOT-GAN by stacking multiple layers of a proposed AOT block. The AOT blocks aggregate contextual transformations from various receptive fields, allowing to capture both informative distant image contexts and rich patterns of interest for context reasoning. For improving texture synthesis, we enhance the discriminator of AOT-GAN by training it with a tailored mask-prediction task. Such a training objective forces the discriminator to distinguish the detailed appearances of real and synthesized patches, and in turn facilitates the generator to synthesize clear textures. Extensive comparisons on Places2, the most challenging benchmark with 1.8 million high-resolution images of 365 complex scenes, show that our model outperforms the state-of-the-art. A user study including more than **30 subjects** further validates the superiority of AOT-GAN. We further evaluate the proposed AOT-GAN in practical applications, e.g., logo removal, face editing, and object removal. Results show that our model achieves promising completions in the real world. We release codes and models in https://github.com/researchmm/AOT-GAN-for-Inpainting.

Index Terms—Image synthesis, image inpainting, object removal, generative adversarial networks (GAN)

1 Introduction

Image inpainting aims at filling missing regions in images with plausible contents [1]. Image Inpainting plays an important role in numerous applications, such as image editing and diminished reality [2], [3], [4], [5], [6]. Despite the great benefits of this technology, image inpainting meets grand challenges in simultaneously recovering reasonable contents and clear textures for large free-form missing regions in high-resolution images.

Early works solve the problem by either diffusion-based or patch-based algorithms [1], [2], [3], [9], [10], [11]. A diffusion-based algorithm propagates contextual information from boundaries into holes in the isophotes direction [1], [9]. A patch-based algorithm synthesizes missing contents by copying and pasting similar patches from known areas or external database [2], [3], [4], [12]. These approaches can work well in completing narrow holes in stationary backgrounds. However, they often fail to hallucinate plausible contents in semantic inpainting due to the lack of powerful reasoning for missing contents in complex scenes [13], [14].

Significant progress has been made by the emergence of deep feature learning [15] and adversarial training [16] for image inpainting [13], [14], [17], [18]. These deep image inpainting models are typically built upon generative adversarial networks (GAN) [16]. Through the joint optimization of reconstruction losses and adversarial losses [7], [13], existing deep models have shown promising results in semantic inpainting. However, directly applying these deep models in higher-resolution images (e.g., 512×512) often results in distorted structures and blurry textures [19], [20]. Such a problem hinders these algorithms from practical applications, where users' photos are often larger than 256×256. The difficulties mainly drive from simultaneously inferring reasonable contents from distant image contexts and generating fine-grained image presentations for each pixel over a large region. We discuss these two challenges respectively as below.

Challenge 1: Context Reasoning. To infer reasonable contents for missing regions, deep image inpainting models leverage features from distant image contexts. For example, contextual attention modules are proposed to model relations between hole regions and image contexts by patch-wise matching [13], [21]. However, the patch-wise matching often results in distorted structures due to the issue of repeating patterns [22]. The other kind of works stack serialized dilated convolution layers to capture distant contexts. However, as widely discussed in previous works [13], [23], serialized dilated

- *Yanhong Zeng and Hongyang Chao are with the School of Computer Science and Engineering, Sun Yat-sen University, Guangzhou 510275, China, and also with Key Laboratory of Machine Intelligence and Advanced Computing, Ministry of Education, Guangzhou 510006, China. E-mail: zengyh7@mail2.sysu.edu.cn, isschhy@mail.sysu.edu.cn.*
- *Jianlong Fu and Baining Guo are with Microsoft Research, Redmond, WA 98052 USA. E-mail: {jianf, bainguo}@microsoft.com.*

Manuscript received 31 Mar. 2021; revised 17 Feb. 2022; accepted 28 Feb. 2022. Date of publication 7 Mar. 2022; date of current version 30 May 2023.
This work was supported in part by NSF of China under Grants 61672548 and U1611461.
(Corresponding authors: Jianlong Fu and Hongyang Chao.)
Recommended for acceptance by T. T. Wong.
Digital Object Identifier no. 10.1109/TVCG.2022.3156949

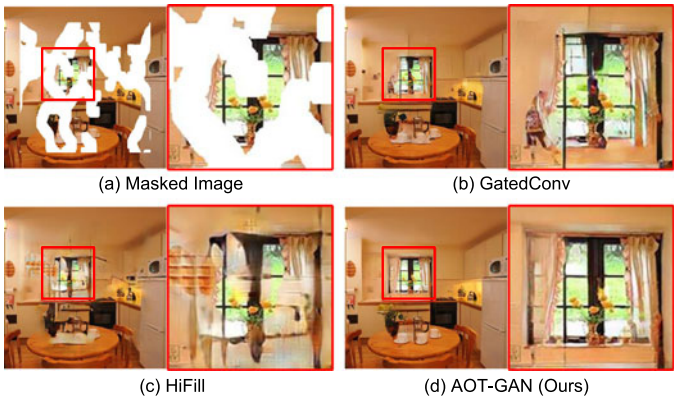

Fig. 1. An illustration of image inpainting. Given an image (a) (resolution: 512×512) with large irregular holes, our model is able to reconstruct more plausible structures and clearer textures of the window in (d) compared with GatedConv [7] (b) and HiFill [8] (c). We present enlarged patches by red boxes next to the images.

convolutions tend to encode features of predefined gridding patterns rather than capturing rich patterns of interest for context reasoning. As shown in Fig. 1, the state-of-the-art approaches, GatedConv [7] (b) and HiFill [8] (c), fail in completing the structures of the window in results.

Challenge 2: Fine-Grained Texture Synthesis. To synthesize fine-grained textures, state-of-the-art models utilize adversarial training by designing a game-theoretical min-max game between a generator and a discriminator. The majority of them inherit the discriminator from PatchGAN [24] due to its great success in image translation [17], [24], [25]. The discriminator of PatchGAN aims to distinguish the patches of a real image from those of an inpainted result. However, they ignore the fact that the patches outside missing regions are shared by both real and inpainted images and blindly push the discriminator to distinguish these identical patches. Such a problematic learning objective weakens the discriminator, and in turn hinders the generator from synthesizing realistic fine-grained textures. As shown in Fig. 1, GatedConv [7] (b) and HiFill [8] (c) tend to generate unrealistic textures in the missing regions.

To overcome the above two challenges, we propose a novel GAN-based model for high-resolution image inpainting, named *A*ggregated C*O*ntextual-*T*ransformation GAN (AOT-GAN). It consists of (1) a generator that exploits aggregated contextual transformations of distant image contexts for enhancing context reasoning, and (2) a discriminator trained by a tailored mask-prediction task to facilitate the synthesis of fine-grained textures.

To enhance context reasoning for large missing regions, we build the generator of AOT-GAN by repeating a carefully-designed building block, namely AOT block. Specifically, the AOT blocks adopt the *split-transform-merge* strategy. First, an AOT block splits the kernel of a standard convolution into multiple sub-kernels. Second, it applies each sub-kernel on the incoming features with different dilation rates. Finally, the AOT block aggregates different transformations from all the sub-kernels as output (Fig. 3b). Through these three steps, informative distant image contexts are leveraged by using various dilation rates, while rich patterns of interest are captured by aggregating multiple transformation results, leading to better context reasoning for missing regions. To the best of our knowledge, AOT block is the first highly-modularized building block tailored for high-resolution image inpainting.

To facilitate the synthesis of fine-grained textures, we design a novel mask-prediction task to train the discriminator. Most existing approaches follow PatchGAN [11] to force the discriminator to predict all patches in inpainted images as fake while ignoring the fact that those outside missing regions indeed come from real images. Therefore, these approaches may struggle to generate realistic fine-grained textures. To overcome the above limitation, we instead train the discriminator to predict a downsampled patch-level inpainting mask. In other words, for an inpainted image, the discriminator is expected to segment the synthesized patches from real ones. Such a learning objective leads to a stronger discriminator and in turn improves the generator to synthesize fine-grained realistic textures.

We summarize our contributions as follows:

- We propose to learn aggregated contextual transformations for high-resolution image inpainting, which allows capturing both informative distant contexts and rich patterns of interest for context reasoning.
- We design a novel mask-prediction task to train the discriminator tailored for image inpainting. Such a design forces the discriminator to distinguish the detailed appearances of real and synthesized patches, which in turn facilitates the generator to synthesize fine-grained textures.
- We conduct extensive evaluations on the most challenging benchmark, Places2 [26]. Results show that our model has achieved state-of-the-art performance in terms of both reconstruction accuracy and high-fidelity visual results.

2 Related Work

Image inpainting aims at filling missing regions in images with plausible contents [1]. Due to its significant practical values for image editing applications (e.g., object removal and photograph restoration), image inpainting has become an active research topic for decades [2], [3], [13], [14], [17], [21]. Existing approaches can fall into two categories, i.e., non-learning based or learning-based methods.

2.1 Non-Learning Based Image Inpainting

Prior to the prevalence of deep learning, approaches for image inpainting are typically divided into two categories, diffusion-based and patch-based algorithms [1], [2], [3], [9], [12]. We introduce both of them in this section.

A diffusion-based algorithm propagates contextual pixels from boundaries to the holes along the isophotes direction [1], [9], [29], [30]. Specifically, many boundary conditions are imposed by the use of Partial Differential Equations during the pixel propagation. Through leveraging the prior knowledge of these boundary conditions, diffusion-based approaches have shown promising results in filling in narrow sketches. However, these approaches typically introduce diffusion-related blurs thus can fail in completing large missing regions [1], [9], [29], [30].

Inspired by patch-based methods for texture synthesis [31], [32], more and more patch-based filling approaches are

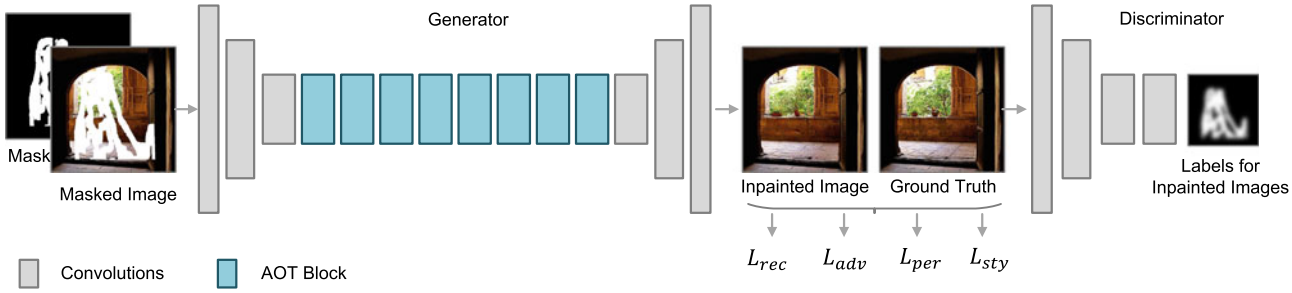

Fig. 2. The overview of the **A**ggregated C**O**ntextual-**T**ransformation GAN (AOT-GAN). For enhancing context reasoning, the generator of AOT-GAN is highly modularized by stacking multiple layers of a carefully-designed block, i.e., AOT block. To improve the training of the discriminator, we train the discriminator by a tailored mask-prediction task, which aims at predicting a patch-level inpainting mask for the input. Our model is jointly optimized by a reconstruction loss, an adversarial loss [16], a perceptual loss [27] and a style loss [28]. We use white color in the mask, the masked image and the labels of inpainted images to denote missing regions and black color in the mask for known regions. See more details in Section 3.

proposed for image inpainting [3], [33]. A patch-based approach typically synthesizes missing contents by copying and pasting similar patches from known image contexts or external databases. To improve the patch-based algorithm, many efforts have been made to find the optimal patch size, patch offset, filling orders, and matching algorithms [2], [3], [4], [12]. For example, Criminisi et al. employ an exemplar-based texture synthesis technique and use calculated confidence values together with image isophotes for determining the filling order of the target inpainting region [3]. Sun et al. propose a curve-based interactive approach to complete important structures before remaining unknown regions [12]. Barnes et al. propose a new randomized algorithm, namely PatchMatch, for quickly finding approximate nearest neighbor matches between image patches [2]. These approaches can work well especially in the stationary background inpainting with repeating patterns. However, these approaches can fall short in completing large missing regions of complex scenes for semantic inpainting. This is because patch-based approaches heavily rely on patch-wise matching by low-level features (e.g., isophotes). Such a technique is unable to synthesize contents for which similar patches do not exist in known image contexts.

2.2 Learning-Based Image Inpainting

Significant progress has been made by the emergence of deep feature learning [15], [34], [35] and adversarial training [14], [16], [36], [37] for image synthesis. Compared with non-learning-based approaches, deep image inpainting models are able to synthesize more plausible contents for complex scenes.

To infer missing contents for a large missing region, deep image inpainting models take advantage of features from distant image contexts. Context Encoder [14], a seminal deep inpainting model, adopts a fully-connected layer in a compact latent feature space for global image contexts encoding. Early works that follow the architecture of Context Encoder have shown promising results on images of human faces, street views, etc. [19], [38], [39]. Iizuka et al. propose to build models upon fully convolutional networks (FCN) for various image resolutions [40]. To capture distant contexts based on FCN, a contextual attention module is proposed to find patches of interest from contexts by patch-wise matching [13], [22], [41]. Zeng et al. further adopt a non-local module named attention transfer network to fill missing regions in feature pyramids [21]. Despite the effective long-range relation modeling, non-local modules tend to repeat patterns in missing regions and distort structures [22], [41], [42], [43]. The other kind of works stack serialized dilated convolution layers to capture distant contexts [13], [20], [44]. Dilated convolutions use kernels that are spread out, allowing to compute each output pixel with a much larger receptive field than standard convolutions [45], [46]. By stacking the dilated convolutions, the model can effectively "see" a larger area of the input image for context reasoning. However, as widely acknowledged, serialized dilated convolutions tend to encode features of predefined gridding patterns rather than patterns of interest for context reasoning [13], [23], [47].

Inspired by recent works of style transfer [27], [28], [48], [49], increasing numbers of deep inpainting models exploit a style loss [28] and a perceptual loss [27] for the synthesis of fine-grained textures. Specifically, the joint optimization of a style loss and a perceptual loss aims at minimizing the perceptual distance between the deep features of inpainted results and the original images. Although promising results have been shown, it remains challenging for these models to generate clear textures [50]. Recent works often adopt the generative adversarial networks (GAN) to improve image fidelity [13], [17], [21]. Specifically, the discriminator of a GAN is trained to distinguish inpainted images from real ones, while the generator is optimized to synthesize realistic images to fool the discriminator. Through the game-theoretical min-max game between the generator and the discriminator, GAN-based inpainting models are able to generate clearer textures [16]. Iizuka et al. propose the joint training of a global and a local discriminator [40]. Yu et al. inherit the discriminator from PatchGAN [24] due to its great success in image translation [7], [13], [17], [21], [25], [51]. Specifically, the discriminator of PatchGAN aims to distinguish patches of real images from those of inpainted results. Moreover, they apply spectral normalization to each layer of the discriminator to stabilize the training of GAN [52].

3 Approach

Inferring reasonable contents and generating clear textures for a large free-form missing region are vital yet challenging to high-resolution image inpainting. In this section, we present the details of the proposed *A*ggregated C*O*ntextual-*T*ransformation GAN (AOT-GAN), which enhances the context reasoning and texture synthesis for high-resolution image inpainting. As shown in Fig. 2, AOT-GAN is built upon a generative adversarial network (GAN), which consists of a

generator network and a discriminator network. Specifically, the generator is constructed by stacking multiple layers of a carefully-designed building block, i.e. AOT block. The AOT block is able to capture both informative distant contexts and rich patterns of interest for enhancing context reasoning. The discriminator is improved by a tailored mask-prediction task, which can in turn facilitate the generator to synthesize clear textures. AOT-GAN is trained by the joint optimization of a reconstruction loss, an adversarial loss, a perceptual loss and a style loss. Through such joint optimization, AOT-GAN is able to synthesize reasonable contents and clear textures for missing regions of high-resolution images.

3.1 Aggregated Contextual Transformations

In this section, we present a simple and highly-modularized generator network, which is enhanced for context reasoning. In particular, context reasoning for a large free-form missing region is one of the grand challenges to high-resolution image inpainting [17], [51]. For one thing, to ensure a coherent structure with surrounding image contexts, deep image inpainting models need to propagate information from distant contexts to missing regions. For another, as objects in complex scenes have various scales and angles of view, capturing as rich as possible patterns of interest is important for context reasoning. The AOT block is designed to capture both informative distant image contexts and rich patterns of interest for context reasoning in high-resolution image inpainting.

The overview of the generator network of AOT-GAN is depicted in Fig. 2. Specifically, the generator consists of an encoder, a stack of the building blocks, i.e. AOT blocks, and a decoder. The generator takes as inputs a masked image and a binary mask indicating missing regions, and it outputs a completed image. First, we stack several layers of standard convolutions as the encoder for feature encoding from the input. Second, features from the encoder are processed by the stack of AOT blocks. Through the backbone built by AOT blocks, the generator is able to capture informative distant image contexts as well as rich patterns of interest for filling missing regions. Finally, we use multiple layers of transposed convolutions to decode features from AOT blocks to a completed image as the final result. We discuss more details of the design of AOT blocks as below.

AOT blocks adopt the *split-transformation-merge* strategy by three steps [53]. *(i) Splitting*: as shown in Fig. 3b, an AOT block splits the kernel of a standard convolution into multiple sub-kernels, each of which has fewer output channels. For example, splitting a kernel with 256 output channels into four sub-kernels makes each sub-kernel has 64 output channels. *(ii) Transforming*: each sub-kernel performs a different transformation of the input feature x_1 by using a different dilation rate. Specifically, a dilated kernel is spread out by introducing zeros between consecutive positions, which has achieved a great success in the field of semantic segmentation [45]. Using a larger dilation rate enables the sub-kernel to "see" a larger area of the input image, while a sub-kernel that uses a small dilation rate focuses on the local patterns of a smaller receptive field. *(iii) Aggregating*: the contextual transformations from different receptive fields are finally aggregated by a concatenation followed by a standard convolution for feature fusion. Such a design allows the AOT block to predict each output pixel through different views. Through the above three steps, an AOT block is able to aggregate multiple contextual transformations for enhancing context reasoning.

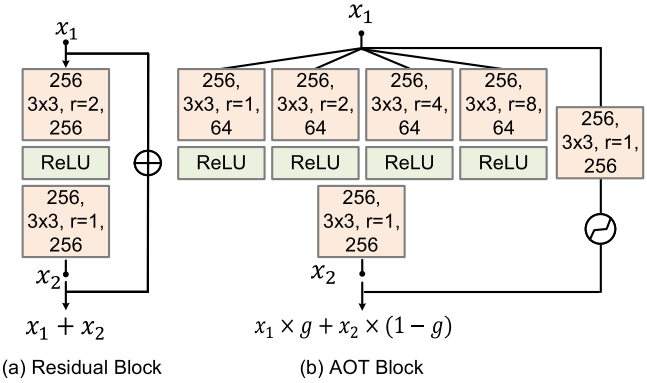

Fig. 3. An illustration of the Residual block (a) used in the state-of-the-art deep inpainting models [17], [42], [51] and the proposed AOT block (b). The numbers inside orange blocks denote (#input channels, filter sizes, dilation rates, #output channels).

Furthermore, the stack of AOT blocks largely enriches the combinations of different pathways in the generator network, allowing the generator to capture as many as possible patterns of interest for context reasoning. It is worth noting that, compared with a standard residual block used in the state-of-the-art deep inpainting models (Fig. 3a), AOT blocks do not introduce extra model parameters and computation costs in these three steps.

As widely discussed by previous works, using the same convolution filter on both valid pixels and the substitute values in the missing regions often leads to blur and color discrepancy [7], [50]. To ease the discrepancy issue, recent works propose to integrate features inside and outside missing regions in a spatially-variant way (e.g., gated convolutions [7] and partial convolutions [50]). However, as shown in Fig. 3a, a vanilla residual block usually adopts an identical residual connection to aggregate the input feature x_1 and the learned residual feature x_2 by element-wise summing in a spatially-invariant way. Such a connection ignores the discrepancies between the values of input pixels inside and outside missing regions and suffers from the well-known discrepancy issue. To alleviate this problem, we propose to use a novel gated residual connection in the building block. As shown in Fig. 3b, a gated residual connection first calculates the spatially-variant gate value $g \in \mathbb{R}^{c \times h \times w}$ from $x_1 \in \mathbb{R}^{c \times h \times w}$ by a standard convolution and a sigmoid operation, where c is the number of feature channels and $h \times w$ is the spatial size of feature maps. Then the AOT block aggregates the input feature x_1 and the learned residual feature $x_2 \in \mathbb{R}^{c \times h \times w}$ by a weighted sum with g, which is denoted as

$$x_3 = x_1 \times g + x_2 \times (1 - g). \tag{1}$$

Such a feature aggregation is able to update features inside missing regions while retaining rich details outside missing regions in a spatially-variant way. We show both quantitative and qualitative experiments to verify the effectiveness of the gated residual connection in Section 4.7.2.

Relation to residual blocks [15]. To enlarge the receptive fields, the state-of-the-art inpainting models also improve the residual block by using dilated convolutions as their building blocks (Fig. 3a) [17], [42], [51]. Specifically, they use the same dilation rate for the kernel inside one building block. Although distant image contexts can be leveraged by the stack of such an improved residual block, it often suffers from encoding features of predefined gridding patterns instead of patterns of interest for context reasoning [13], [23], [47]. In comparison, the proposed AOT block is able to capture rich patterns of interest by learning aggregated contextual transformations with various dilation rates.

Relation to multi-scale dilated convolution modules [45], [54], [55]. Atrous Spatial Pyramid Pool (ASPP) is a seminal work that shares a similar multi-scale topology of the proposed AOT block [45]. However, ASPP is proposed for the task of semantic segmentation and it is used at the end of networks for high-level recognition, while AOT block is a basic building block repeating in inpainting models for low-level tasks. Besides, we noticed that recent image inpainting approaches also adopt similar multi-scale designs and have achieved promising results, e.g., Spatial Pyramid Dilation (SPD) in DeepGIN [54] and multi-scale attention unit (MSAU) [55]. We highlight two key differences between our work and them. *(a) Different ways of feature fusion.* SPD adopts vanilla residual connections and MSAU uses spatial and channel attention for boosting the multi-scale features, while both of them ignore the issue of the discrepancy inside/outside missing regions. Our AOT block incorporates a carefully-designed gated residual connection that largely eases the discrepancy issue for image inpainting. We have verified the effectiveness of the AOT block in Section 4.7.1 and 4.7.2. *(b) Different target resolution.* Both SPD and MSAU are designed for relatively low-resolution images, while the AOT is designed for high-resolution image inpainting. For example, as introduced by Li et al. [54], DeepGIN needs to divide images larger than 256×256 into multiple sub-images and regroup the results for high-resolution image inpainting, while our model is able to apply to high-resolution images without any pre-/post-processing.

3.2 Soft Mask-Guided PatchGAN (SM-PatchGAN)

In this section, we present a novel mask-prediction task for enhancing the training of the discriminator of AOT-GAN, which can in turn facilitate the generator to synthesize clear textures for high-resolution image inpainting.

Synthesizing clear fine-grained textures for a large missing region is another challenge for high-resolution image inpainting. This is because there are various possible results for filling a missing region, whereas deep inpainting models are trained to reconstruct the original images based on reconstruction losses (e.g., L_1 loss) [13]. With such an optimization objective, deep inpainting models tend to generate an average of all the possible solutions, resulting in blurry textures. To overcome this challenge, we build our model upon a generative adversarial network (GAN) [16]. Through the joint optimization of reconstruction losses and an adversarial loss, our model, AOT-GAN, is able to generate clearer results that fall near the manifold of natural images.

In practice, we denote a corresponding binary inpainting mask as m, where value 1 for known pixels (black regions in Fig. 2) and value 0 for missing regions (white regions in Fig. 2). We denote a real image as I, $\odot$ as pixel-wise multiplication, G as the generator, and thus the inpainted results can be denoted as

$$z = I \odot m + G(I \odot m, m) \odot (1 - m). \tag{2}$$

We inherit the network architecture of the discriminator from PatchGAN [24] due to its great success in the field of image translation. As shown in Fig. 2, the discriminator network consists of several layers of standard convolutions, each of which reduces the spatial size of feature maps by two. The discriminator takes as input an image from inpainted results or real data, and outputs a prediction map. In particular, each pixel of the prediction map indicates the prediction for a $N \times N$ patch in the input image as real or fake.

To facilitate the synthesis of fine-grained textures, we design a mask-prediction task for the training of the discriminator, which we term a Soft Mask-guided PatchGAN (SM-PatchGAN). Specifically, we downsample the inpainting masks as the ground truths of the mask-prediction task. To improve the training of the discriminator and force it to focus more on the central parts of the missing regions instead of the boundaries, we use a soft patch-level mask for training as shown in Fig. 4. The soft mask is obtained by Gaussian filtering. We denote the adversarial loss [56] for the discriminator as

$$\begin{aligned} L_{adv}^{D} = \mathbb{E}_{z \sim p_z}\Big[(D(z) - \sigma(m))^2\Big] \\ + \mathbb{E}_{x \sim p_{data}}\Big[(D(x) - 1)^2\Big], \end{aligned} \tag{3}$$

where σ denotes the processing (i.e., down-sampling and Gaussian filtering) of the masks used for training. Correspondingly, the adversarial loss [56] for the generator is denoted as

$$L_{adv}^{G} = \mathbb{E}_{z \sim p_z}\Big[(D(z) - 1)^2\Big]. \tag{4}$$

Through such an optimization, the discriminator is encouraged to segment synthesized patches of missing regions from real ones of contexts outside missing regions. The enhanced discriminator can in turn facilitate the generator to synthesize more realistic textures.

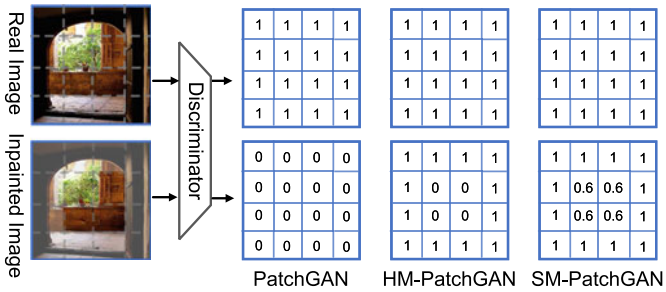

Fig. 4. An illustration of different tasks for the training of the discriminator. PatchGAN aims at distinguishing patches of inpainted images from those of real images, while HM-PatchGAN and SM-PatchGAN aim to segment synthesized patches of missing regions from real ones of contexts according to inpainting masks.

Relation to PatchGAN [24]. As shown in Fig. 4, the PatchGAN discriminator used in previous inpainting models aims at distinguishing patches of inpainted images from those of real ones. In this way, they blindly push the discriminator to predict all patches in inpainted images as fake, while ignoring that patches outside missing regions indeed come from real images. In contrast, the proposed SM-PatchGAN aims at distinguishing synthesized patches of missing regions from real ones of contexts, which can lead to a stronger discriminator.

Relation to HM-PatchGAN. We consider another possible design of the mask-prediction task, which we term a Hard Mask-guided PatchGAN (HM-PatchGAN). As shown in Fig. 4, the HM-PatchGAN enhances the PatchGAN discriminator by training by hard binary inpainting masks without Gaussian filtering. The HM-PatchGAN ignores the fact that a patch in an inpainted image around boundaries may contain both real and synthesized pixels, while the proposed SM-PatchGAN is able to improve the training of the discriminator by adopting soft labels and forces the model to focus more on the central part of the missing regions rather than the boundaries. We conduct extensive ablation studies to show the superiority of SM-PatchGAN in Section 4.7.3.

3.3 Overall Optimization

The principle of choosing optimization objectives for image inpainting is to ensure both per-pixel reconstruction accuracy and visual fidelity of inpainted images. To this end, we carefully select four optimization objectives, i.e., a reconstruction loss, a style loss [28], a perceptual loss [27], and an adversarial loss of the SM-PatchGAN, for AOT-GAN following the majority of existing deep inpainting models [13], [17], [41], [50]. First, we aim to minimize the L_1 distance to ensure the reconstruction accuracy on pixel-level

$$L_{rec} = \|x - G(x \odot (1-m), m)\|_1. \tag{5}$$

Since the effectiveness of the perceptual loss [27] and the style loss [28] for image inpainting have been widely verified [17], [50], we include them to improve the accuracy of perceptual reconstruction. Specifically, a perceptual loss aims at minimizing the L_1 distance between the activation maps of inpainted and real images

$$L_{per} = \sum_i \frac{\|\phi_i(x) - \phi_i(z)\|_1}{N_i}, \tag{6}$$

where ϕ_i is the activation map from the ith layers of a pre-trained network (e.g., VGG19 [57]), N_i is the number of elements in ϕ_i. Similarly, the Style loss is defined as the L_1 distance between the Gram matrices of deep features of inpainted and real images

$$L_{sty} = \mathbb{E}_i \Big[\|\phi_i(x)^T \phi_i(x) - \phi_i(z)^T \phi_i(z)\|_1 \Big]. \tag{7}$$

Finally, we include the adversarial loss of SM-PatchGAN described in Eq. 4 to improve the visual fidelity of inpainted images. The whole AOT-GAN is trained by the joint optimization of these four objectives, and we conclude the overall optimization objective as below

$$L = \lambda_{adv} L^G_{adv} + \lambda_{rec} L_{rec} + \lambda_{per} L_{per} + \lambda_{sty} L_{sty}. \tag{8}$$

For our experiments, we empirically choose $\lambda_{adv} = 0.01$, $\lambda_{rec} = 1$, $\lambda_{per} = 0.1$ and $\lambda_{sty} = 250$ for training.

3.4 Implementation Details

To build the generator network of the proposed AOT-GAN, we stack three layers of convolutions as the encoder, which reduces the spatial size of images by four times in total. Correspondingly, we use three layers of transposed convolutions for the upsampling in the decoder. We stack eight layers of AOT blocks to build the backbone of the generator. For building the discriminator network, we stack four layers of convolutions with stride as two following 70×70 PatchGAN [24]. Accordingly, for the processing of Gaussian filtering in SM-PatchGAN, we set the kernel size of the Gaussian kernel as 70×70 to simulate the pixel propagation by convolutions in the discriminator network. In particular, to avoid the issue of color shift caused by normalization layers, we remove all normalization layers in the generator network following previous works [7], [13].

For the model training, eight images are randomly sampled and corresponding masks are randomly created as pairs in each mini-batch. We use a fixed learning rate at $1e-4$ for both the training of the discriminator and the generator. We use Adam optimizer with $\beta_1 = 0$ and $\beta_2 = 0.9$ for training. In practice, we use VGG19 [57] pre-trained on ImageNet [58] as our pre-trained network for calculating the style loss and the perceptual loss. Specifically, we use the activation maps from layers relu1_1, relu2_1, relu3_1, relu4_1 and relu5_1 of the VGG19 network to calculate the style loss and relu2_2, relu3_4, relu4_4, relu5_2 to calculate the perceptual loss, following previous works [17]. More implementation details can be found in our released codes and models in https://github.com/researchmm/AOT-GAN-for-Inpainting.

4 Experiments

4.1 Datasets

We introduce three datasets with their brief introductions below:

- *Places2* [26] contains over 1.8 million images from 365 kinds of scenes. It is one of the most challenging datasets for image inpainting due to its complex scenes. We use the splits of training/testing (i.e., 1.8 million/36,500) following the settings used by most inpainting models [13], [17], [21].
- *CELEBA-HQ* [61] is a high-quality dataset of human faces. The high-frequency details of hairs and skins can help us to evaluate the fine-grained texture synthesis of models. We use 28,000 images for training and 2,000 for testing following a common setting [13], [17].
- *QMUL-OpenLogo* [62] contains 27,083 images from 352 logo classes. Each image is annotated by fine-grained bounding box annotations of logos. We use 15,975 training images for training and 2,777 validation images for testing.

Specifically, all the images are resized and cropped to 512×512 for both training and testing following a standard

high-resolution setting [19], [26], [43]. Since the free-form mask dataset proposed by Liu *et al.* [50] has been verified to be effective for improving the training of inpainting models and it has been widely adopted by recent inpainting approaches [7], [8], [22], [50], we adopt the free-form mask dataset for both training and testing in our experiments. Specifically, the free-form mask dataset consists of six categories of masks with different hole-to-image area ratios: (0.01, 0.1], (0.1, 0.2], (0.2, 0.3], (0.3, 0.4], (0.4, 0.5], (0.5, 0.6]. In particular, the mask dataset does not include mask portions that are larger than these area ratios because extremely large masks are not commonly-used in usual inpainting applications (e.g., object removal and restoration) [50].

4.2 Baselines

We list all the state-of-the-art models used for comparisons with their abbreviations and brief introductions as below:

- *CA* [13] is a two-stage coarse-to-fine model. It deploys a patch-based non-local module, i.e., contextual attention module, to effectively model long-range correlations.
- *PEN-Net* [21] is built upon a U-Net [63] backbone. It proposes to fill missing regions from deep to shallow by a cross-layer non-local module.
- *PConv* [50] adopts a proposed partial convolution layer instead of the standard convolution to deal with the issue of color discrepancy inside and outside holes.
- *EdgeConnect* [17] model consists of two stages. The first stage completes the edges of corrupted images, and the second stage completes the color images with the guidance of the completed edges from the first stage.
- *GatedConv* [7] is a two-stage model that incorporates a gated convolution and SN-PatchGAN for image inpainting.
- *HiFill* [8] proposes a light-weight model with an efficient Contextual Residual Aggregation mechanism that enables ultra super-resolution image inpainting.
- *MNPS* [19] is a multi-scale neural patch synthesis approach based on the joint optimization of content and texture.
- *DeepGIN* [54] proposes a multi-scale module and an attention mechanism to enhance inpainting results.
- *ICT* [59] combines transformer architectures and CNNs to achieve both high-fidelity and diversified image completion.
- *CoModGAN* [60] leverages both image-conditional and modulated unconditional generative ability to achieve both diverse and high-quality image inpainting.

4.3 Evaluation Metrics

We list all the objective metrics used for quantitative comparisons and the reasons for using them as below:

- L_1 *error* is widely used in previous works [13], [19], [21]. It calculates the mean absolute error between inpainted and original images to evaluate per-pixel reconstruction accuracy.
- *Peak Signal-to-Noise Ratio (PSNR)* is one of the classical image quality assessments that are used by many inpainting approaches [7], [13].
- *Structural Similarity Index (SSIM)* [64] compares inpainted results with the original images from the aspects of luminance, contrast and structure.
- *Fréchet Inception Distance (FID)* [65] is a popular deep metric for perceptual rationality [24], [52]. FID measures the distance between the distributions of real and fake image features. It is worth noting that, deep metrics have been proved to be closer to human perception [66].

Although the above four objective metrics can reflect the performance of inpainting models partly, a subjective assessment is always the best way to evaluate inpainting models. Therefore, we conduct extensive qualitative evaluations and a user study for a comprehensive comparison.

4.4 Quantitative Comparisons

To evaluate the AOT-GAN on benchmarks, we conduct a quantitative comparison on the most challenging benchmark, Places2 [26]. Specifically, we use all the images of the evaluation set of Places2 following the common setting used in many deep inpainting approaches [17], [50], [51]. Each kind of scene in Places2 contains 100 evaluation images. For each testing image, we randomly sample a free-form mask as the testing mask for fair comparisons. The free-form masks are provided by Liu et al. [50]. We use the same image-mask pairs for all approaches for fair comparisons.

The quantitative comparisons in Table 1 show that the AOT-GAN has achieved state-of-the-art performance in both reconstruction accuracy and visual fidelity. Specifically, with the enhanced generator architecture, AOT-GAN shows the best reconstruction accuracy in terms of L_1 error, PSNR and SSIM. Recent SOTA approaches aim at achieving diverse image inpainting by Transformer-based [59] or StyleGAN-based [60] generative models. Specifically, CoModGAN is designed by training with only adversarial losses, showing impressive performance in terms of FID in completing large masks (e.g., (0.5, 0.6] mask area ratio). However, CoModGAN is hard to reconstruct inpainting results that are consistent with contexts. In comparison, AOT-GAN is able to yield both high reconstruction accuracy and high-fidelity visual results.

MNPS [19] is a multi-stage model designed for high-resolution image inpainting (512×512). Since only the parameters of the MNPS model trained on ImageNet [58] is available, we compare MNPS and AOT-GAN on ImageNet. For fair comparisons, the AOT-GAN is trained on ImageNet by center square inpainting masks, following the same training setting used by MNPS. Images are resized to 512×512 for training. It takes MNPS a long time to fill a missing region in a 512×512 image even on GPUs, due to the hundreds of iterations for post-processing. Therefore, we randomly sample 2,000 images from the testing set of ImageNet for testing and we use a center 224×224 square as inpainting mask, following a common setting [19], [43]. The results in Table 2 show that our single-stage model is able to gain improvements in terms of all the metrics compared with MNPS.

TABLE 1
Quantitative Comparisons of the AOT-GAN With State-of-the-Art Deep Inpainting Models, i.e., CA [13], PEN-Net (PEN.) [21], PConv [50], EdgeConnect (Edge.) [17], GatedConv (Gated.) [7], HiFill [8], DeepGIN [54], ICT [59] and CoModGAN (CoMod.) [60] on Places2 [26]

Mask		CA [13]	PEN. [21]	PConv [50]	Edge. [17]	Gated. [7]	HiFill [8]	DeepGIN[54]	ICT[59]	CoMod.[60]	AOT-GAN
$L_1(10^{-2})\downarrow$	1-10%	0.89	0.69	0.68	0.55	0.66	0.67	**0.40**	0.81	0.47	0.55
	10-20%	2.07	1.58	1.28	1.28	1.55	1.81	1.21	2.21	1.31	**1.19**
	20-30%	3.54	2.72	2.42	2.42	2.73	3.31	2.22	3.99	2.46	**2.11**
	30-40%	5.19	3.97	3.43	3.47	4.08	5.02	3.31	5.90	3.79	**3.20**
	40-50%	7.07	5.40	4.62	4.92	5.68	7.12	4.53	8.07	5.38	**4.51**
	50-60%	10.11	7.76	7.74	7.83	8.09	10.47	7.38	10.83	8.37	**7.07**
PSNR↑	1-10%	31.07	32.73	34.04	33.98	32.95	30.97	**35.11**	29.25	33.86	34.79
	10-20%	25.81	27.17	28.75	28.53	27.05	25.36	29.16	23.65	27.68	**29.49**
	20-30%	22.93	24.18	25.59	25.22	23.81	22.35	25.94	20.81	24.29	**26.03**
	30-40%	20.89	22.19	23.40	22.88	21.55	20.21	23.43	18.96	21.93	**23.58**
	40-50%	19.23	20.60	21.56	21.00	19.75	18.35	21.63	17.46	20.04	**21.65**
	50-60%	17.10	18.62	18.75	18.22	16.94	16.07	18.82	16.06	17.48	**19.01**
SSIM↑	1-10%	0.961	0.967	0.971	0.974	0.970	0.963	**0.976**	0.962	0.974	**0.976**
	10-20%	0.906	0.917	0.928	0.931	0.921	0.905	0.936	0.902	0.929	**0.940**
	20-30%	0.844	0.856	0.875	0.876	0.860	0.835	0.885	0.833	0.872	**0.890**
	30-40%	0.783	0.795	0.820	0.816	0.796	0.762	0.833	0.767	0.812	**0.835**
	40-50%	0.720	0.733	0.762	0.751	0.727	0.680	0.767	0.699	0.747	**0.773**
	50-60%	0.648	0.664	0.677	0.660	0.626	0.588	0.678	0.633	0.658	**0.682**
FID↓	1-10%	1.30	0.42	0.36	0.25	0.21	0.53	0.23	1.06	0.19	**0.20**
	10-20%	6.33	2.14	1.85	0.97	0.85	2.52	1.02	4.87	0.65	**0.61**
	20-30%	17.36	6.66	5.83	2.83	2.40	7.60	3.11	12.92	**0.93**	1.57
	30-40%	34.26	14.46	12.96	6.52	5.33	17.18	6.94	24.68	**1.62**	3.38
	40-50%	56.89	27.04	24.63	13.44	10.66	36.23	13.67	42.29	**2.68**	6.89
	50-60%	82.67	45.27	47.09	33.82	32.90	72.03	33.54	61.04	**4.73**	20.20

↓ *Lower is better.* ↑ *Higher is better. The best and the second best results are* ***highlighted*** *and* underlined.

4.5 Qualitative Comparisons

For fair qualitative comparison, we randomly sampled the results of different models and show them in Fig. 5. Specifically, we compare AOT-GAN with the state-of-the-art deep inpainting models, i.e., CA [13], PEN-Net [21], PConv [50], EdgeConnect [17], GatedConv [7] and HiFill [8] on Places2 [26]. We show the results of various scenes, including *valley*, *street*, *field wild* and *alcove*, with their enlarged patches in Fig. 5.

As shown in Fig. 5, most objects are occluded by a large missing region. For example, a large portion of the window in the street case are occluded and existing approaches are hard to reconstruct it. Specifically, CA [13] , PEN-Net [21] and PConv [50] tend to generate blurry textures. EdgeConnect [17] fills in the holes with checkerboard artifacts. GatedConv [7] and HiFill [8] distort the structures in results by unreasonable repeating patterns. With the enhanced generator and discriminator networks, our model is able to reconstruct more plausible contextual structures and generate clearer textures in various kinds of complex scenes. For example, in the results of *street* in Fig. 5, AOT-GAN is able to reconstruct the structure of the window. In the results of *valley*, AOT-GAN is able to hallucinate plausible structures and synthesize fine-grained textures of the valley even with the large missing regions.

TABLE 2
Quantitative Comparison Results With MNPS [19] on ImageNet [58]

	$L_1(10^{-2})\downarrow$	PSNR↑	SSIM↑	FID↓
MNPS[19]	3.37	21.36	0.821	43.87
Ours	**3.25**	**22.10**	**0.840**	**41.35**

↓ *Lower is better.* ↑ *Higher is better.*

4.6 User Study

PConv [50] and GatedConv [7] are two competitive and representative deep image inpainting models that have shown state-of-the-art performance in plenty of images. Therefore, we conduct pair-wise comparisons on real data, our model and the results of PConv [50] and GatedConv [7] in our user study. In each trial, two images are shown to the volunteers in an anonymous way. Specifically, one inpainted image is sampled from our model and the other is from other corresponding results (i.e., real data, or PConv [50], or GatedConv [7]). We randomly sample 400 cases for the experiments on Places2 and CELEBA-HQ, respectively. Participants are not informed of the locations of missing regions and they are asked to select images with better structures and more realistic textures. *More than 30 participants* are invited in the user study and we have collected *951 valid votes*. The results of the user study can be found in Table 3. We can see that the AOT-GAN outperforms other methods in most cases. Compared with real data, the results of AOT-GAN can surprisingly win 22.47% cases on CELEBA-HQ.

4.7 Ablation Studies

We show inpainting results of masks with different hole-to-image area ratios in Fig. 6 to demonstrate the capacity of AOT-GAN. As shown in Fig. 6, when dealing with a small

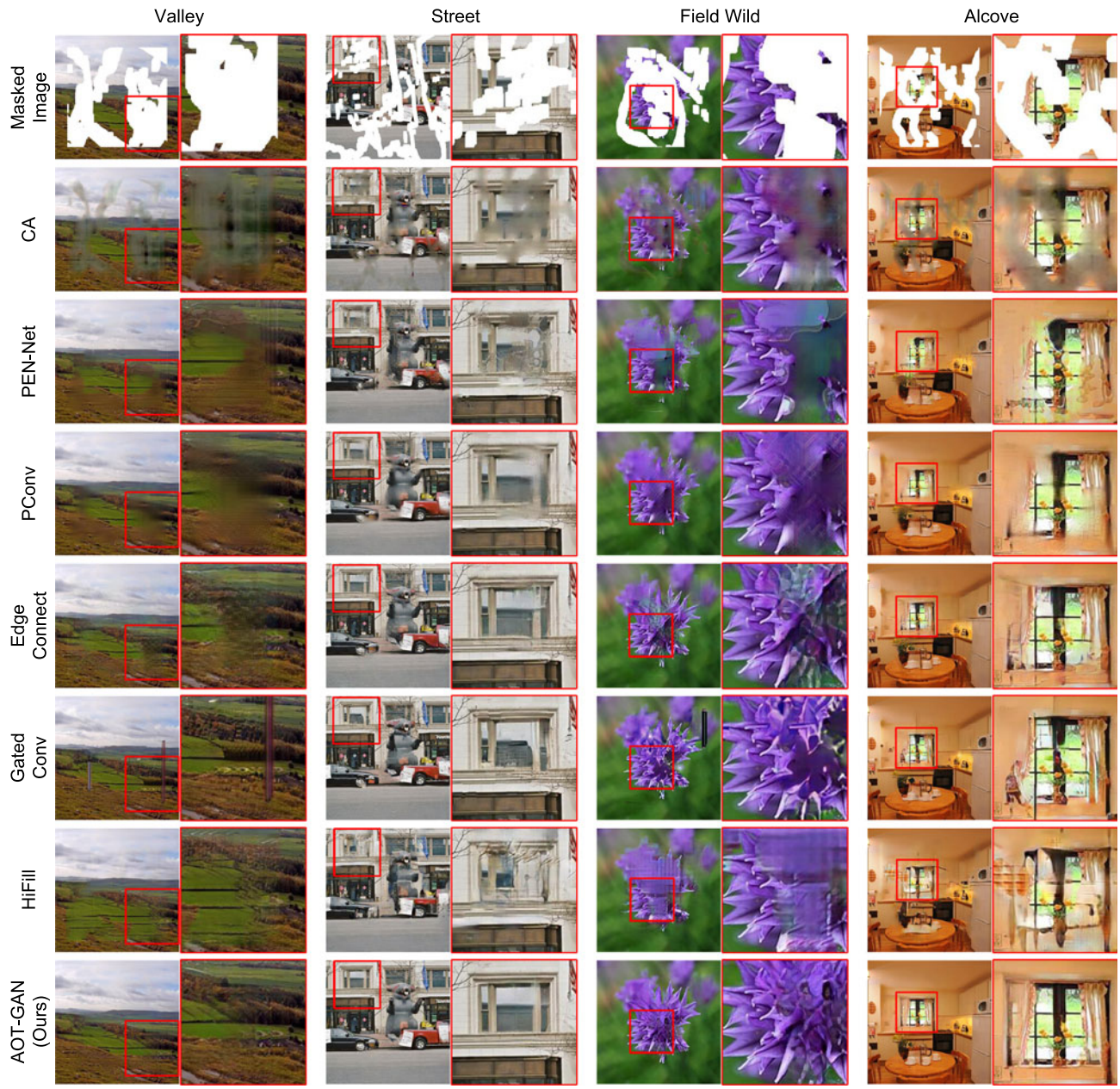

Fig. 5. Qualitative comparisons of AOT-GAN with CA [13], PEN-Net [21], PConv [50], EdgeConnect [17], GatedConv [7] and HiFill [8] on Places2 [26]. Each column shows the results and their enlarged patches marked by red boxes next to them. As shown in these cases, our model is able to reconstruct more plausible structures and clearer textures of various scenes, including *valley*, *street*, *field wild* and *alcove*. All the images are center-cropped and resized to 512×512. See analysis in Section 4.5. [Best viewed with zooming-in].

mask, AOT-GAN is able to generate images that are indistinguishable from real images. When dealing with a larger mask, inpainting becomes more challenging due to limited contexts. Compared with EdgeConnect [17] and GatedConv [7], AOT-GAN is able to generate high-fidelity images without a catastrophic performance drop. For example, in the mask of (0.5, 0.6] area ratio in Fig. 6, AOT-GAN is able to reconstruct the smiling face even though almost the whole face is occluded.

TABLE 3
User Study for Pair-Wise Comparisons With PConv [50], GatedConv [7], and Real Data on CELEBA-HQ [61] and Places2 [26]

Percentage	CELEBA-HQ	Places2
AOT-GAN > PConv [50]	95.65%	97.67%
AOT-GAN > GatedConv [7]	86.62%	95.08%
AOT-GAN > Real	22.47%	17.39%

"Percentage" means the percentage of cases where our result is better than the other.

In this section, we conduct three sets of ablation studies to verify the effectiveness of three components of AOT-GAN, i.e., aggregated contextual transformations, the gated residual connections and the SM-PatchGAN discriminator. As shown in Table 4, each set of ablation studies is conducted based on the previous components. Since different models have similar results of small masks and the comparison becomes not obvious enough, we use a larger mask (i.e., (0.4, 0.5] mask area ratio) for ablation study for clear comparisons.

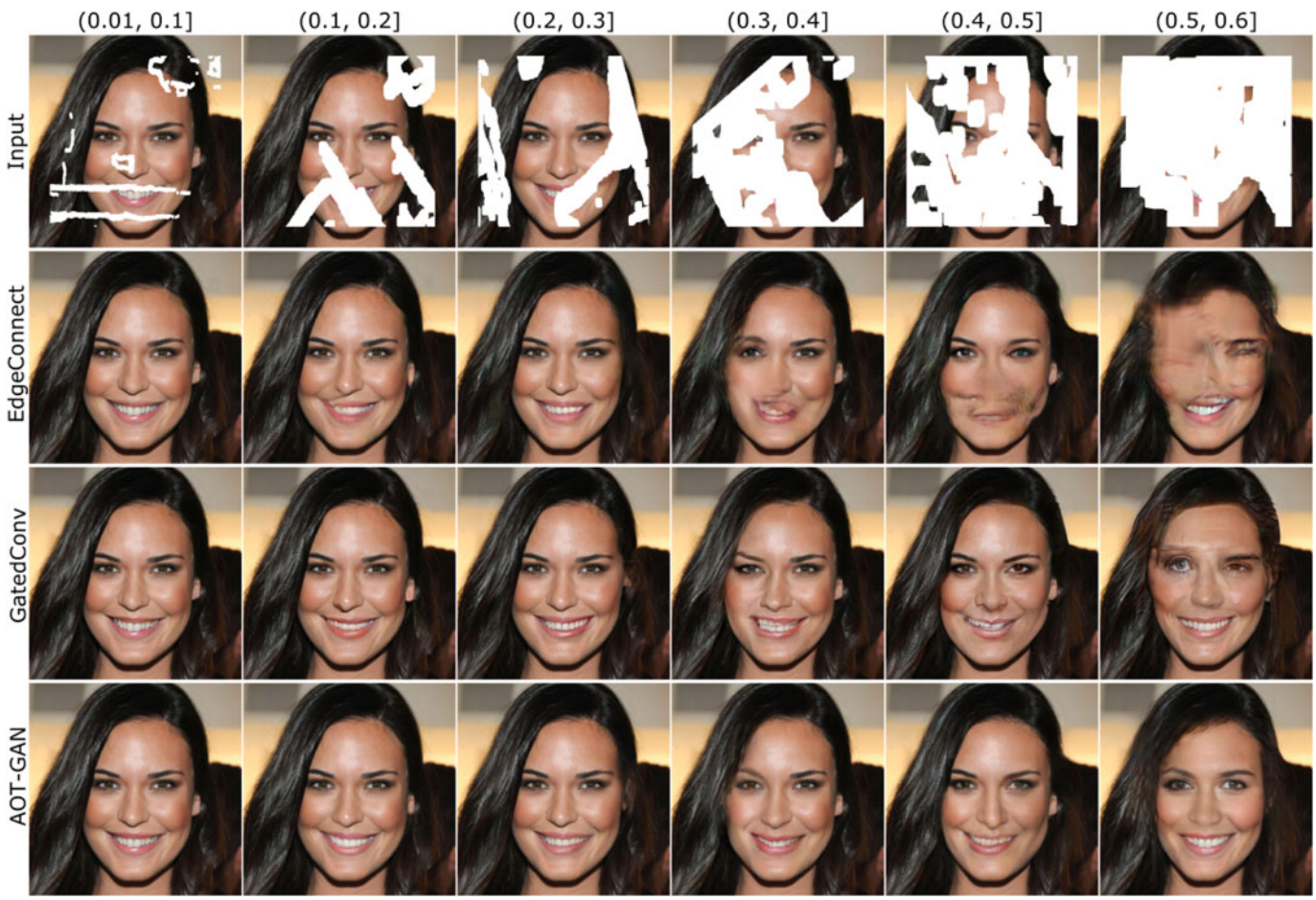

Fig. 6. Results of masks with different hole-to-image area ratios. The first row shows the input images with masks of different hole-to-image area ratios, following the output by EdgeConnect [17], GatedConv [7] and AOT-GAN. With a small mask (e.g., (0.01, 0.1]), different models have similar outputs. With a larger mask (e.g., (0.4, 0.5]), inpainting becomes challenging. Compared with EdgeConnect and GatedConv, AOT-GAN is able to generate high-fidelity results without a catastrophic performance drop.

TABLE 4
Ablation Studies for Each Component of the Proposed AOT-GAN

Model	AOT block	Gated residual connections	SM-PatchGAN	$L_1(10^{-2})\downarrow$	PSNR↑	SSIM↑	FID↓
1	✓			3.28	24.06	0.834	9.86
2	✓	✓		3.06	24.47	0.849	8.02
3 (AOT-GAN)	✓	✓	✓	**2.98**	**24.65**	**0.852**	**7.37**

The full model (AOT-GAN) that exploits aggregated contextual transformations, gated residual connections and SM-PatchGAN performs the best. ↓ Lower is better. ↑Higher is better.

4.7.1 Aggregated Contextual Transformations

To verify the effectiveness of the aggregated contextual transformations in AOT blocks, we use a model of the same network depth as AOT-GAN as a baseline. In particular, we remove the components of gated residual connections and SM-PatchGAN. The baseline uses standard residual connections in its building blocks and it is trained by a standard PatchGAN, so that it can be improved only by the aggregated contextual transformations. We conduct extensive comparisons for the design of aggregated contextual transformations by using different numbers of branches and dilation rates based on the baseline.

The quantitative comparison in Table 5 shows that, compared with a single-branch module, the enhanced models with aggregated contextual transformations gain significant improvements in terms of L_1 error, PSNR, SSIM, and FID. We can also observe that a larger number of branches with more diverse dilation rates is able to achieve larger improvements. We find that more branches with more diverse dilation rates can largely enrich the patterns captured by AOT blocks, which is important to high-resolution image inpainting, especially in filling large missing regions.

According to the quantitative comparisons in Table 5, we use four branches and the dilation rate of each branch is 1, 2, 4, 8, respectively in our final model. We compare this setting and the single-branch module with $rate = 2$ (i.e., the residual block used in EdgeConnect [17]) by qualitative comparisons in Fig. 7. We can see that our model is able to complete the large missing region of the human face with better structures, as well as clearer textures of hairs. We can draw a conclusion that the design of aggregated contextual

TABLE 5
Quantitative Comparisons on CELEBA-HQ [61] for Using Different Numbers of Branches and Dilation Rates in AOT Blocks

#branch	rates	$L_1(10^{-2})\downarrow$	PSNR↑	SSIM↑	FID↓
1	1	3.65	23.16	0.807	14.21
1	2	3.59	23.29	0.810	12.28
2	1,2	3.58	23.25	0.812	11.54
2	2,8	3.36	23.83	0.830	10.56
3	1,2,8	3.36	23.91	0.831	10.06
4	1,2,4,8	**3.28**	**24.06**	**0.834**	**9.86**

↓ *Lower is better.* ↑*Higher is better.*

Fig. 7. Visual comparisons on images (512×512) of CELEBA-HQ [61] to verify the effectiveness of the aggregated contextual transformations in AOT block. [Best viewed with zoom-in].

transformations is able to enhance context reasoning and result in better structures and textures for high-resolution image inpainting.

4.7.2 Gated Residual Connections

The gated residual connections are used to aggregate the input features and the learned residual features in AOT blocks. To verify the effectiveness of the gated residual connection, we compare it with several counterparts for aggregating residual features, i.e., identical residual connections [15], GatedConvolution [7], and feature concatenation with a 1×1 convolution in AOT blocks. Specifically, GatedConvolution generalizes partial convolution [50] by providing a learnable dynamic feature selection for each channel at each spatial location. We compare these methods by replacing the gated residual connections in AOT blocks.

The quantitative comparison in Table 6 shows that our gated residual connection outperforms others in terms of all the metrics. The visual comparison in Fig. 8 shows that leveraging the gated residual connections, our model is able to generate more realistic facial textures than other residual fusion methods. This is because the proposed gated residual connection provides a learnable spatially-variant connection between the input features and the learned residual features. Such a gated connection is able to update features inside missing regions while retaining low-level details of valid pixels outside holes.

4.7.3 Soft Mask-Guided PatchGAN (SM-PatchGAN)

We compare different discriminators in AOT-GAN, i.e., PatchGAN [24], HM-PatchGAN and SM-PatchGAN. Specifically, the generator is constructed by the proposed full AOT blocks. We follow common practices and adopt the concatenation of images and masks as the input of the PatchGAN to improve its performance [13], [17]. We also compare SM-PatchGAN using different kernel sizes of the Gaussian filter for generating the labels for the discriminator. Specifically, we set $\sigma = \frac{1}{3}k$ according to the three-sigma rule [67] in our experiments, where σ is the standard deviation of the Gaussian distribution and k is the kernel size of the Gaussian filter.

TABLE 6
Quantitative Comparisons on CELEBA-HQ [61] to Verify the Effectiveness of the Proposed Gated Residual Connection in AOT Blocks

	$L_1(10^{-2})\downarrow$	PSNR↑	SSIM↑	FID↓
Identical ResBlock	3.28	24.06	0.834	9.86
GatedConv. [7]	3.28	23.98	0.835	10.22
1×1 conv.	3.14	24.46	0.846	9.06
Ours	**3.06**	**24.47**	**0.849**	**8.02**

↓ *Lower is better.* ↑ *Higher is better.*

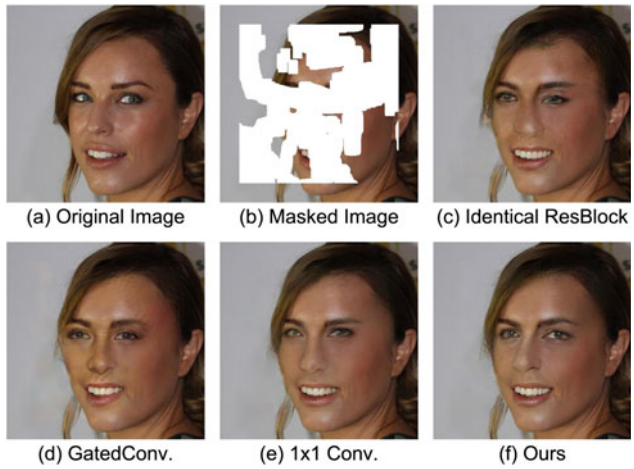

Fig. 8. Visual comparisons on images (512×512) of CELEBA-HQ [61] to verify the effectiveness of the gated residual connection in AOT block. [Best viewed with zoom-in].

The quantitative comparison in Table 7 shows that both the proposed HM-PatchGAN and SM-PatchGAN outperform PathGAN. Specifically, SM-PatchGAN (70×70) performs the best especially in terms of FID. As shown in the qualitative results in Fig. 9, all the models are able to reconstruct the plausible structures (e.g., the eyes) in the large missing regions thanks to the multi-scale generator. However, PatchGAN tends to generate blur textures for the eyes while SM-PatchGANs are able to generate clear textures (e.g., the eyelashes). Specifically, SM-PatchGAN using 70×70 Gaussian kernel size shows better results. Therefore, we adopted the SM-PatchGAN (70×70) as our full model.

4.8 Real Applications

In this section, we apply AOT-GAN to three practical applications (i.e., logo removal, face editing and object removal), to verify the effectiveness of AOT-GAN in the real world.

4.8.1 Logo Removal

The automatic logo removal technique is very useful in logo design, media content creation, and so on [68], [69]. It requires removing logos from images and filling in the missing regions seamlessly. However, since the scales of logos vary largely and the backgrounds are typically complex, it remains challenging for existing inpainting approaches to do well in logo removal.

TABLE 7
Quantitative Comparisons on CELEBA-HQ [61] to Verify the Effectiveness of the SM-PatchGAN Discriminator of AOT-GAN

	$L_1(10^{-2})\downarrow$	PSNR↑	SSIM↑	FID↓
PatchGAN [24]	3.06	24.47	0.849	8.02
HM-PatchGAN	2.99	24.63	**0.853**	7.58
SM-PatchGAN(16×16)	3.01	24.63	0.850	7.42
SM-PatchGAN(70×70)	**2.98**	**24.65**	0.852	**7.37**
SM-PatchGAN(128×128)	3.03	24.50	0.849	7.60

↓ *Lower is better.* ↑ *Higher is better.* 70×70 *denotes the kernel size used for generating soft masks.*

To evaluate the performance of AOT-GAN on logo removal, we use the images of QMUL-OpenLogo [62] for both training and testing. For training, images are center-cropped and resized to 512×512 and we use rectangles of arbitrary sizes as training masks. For testing, images are resized to keep the shortest side as 512 and we use the bounding boxes of logos as inpainting masks.

We show the results of AOT-GAN on QMUL-OpenLogo in Fig. 10. We can find that the AOT-GAN is able to remove logos from images in an undetectable way and fill in missing regions with plausible contents. We attribute the promising results to the enhanced context reasoning and texture synthesis by AOT-GAN. Specifically, both rich patterns of interest and informative distant contexts can be captured by learning aggregated contextual transformations in the generator of AOT-GAN, and the synthesis of clear textures is ensured by the use of SM-PatchGAN.

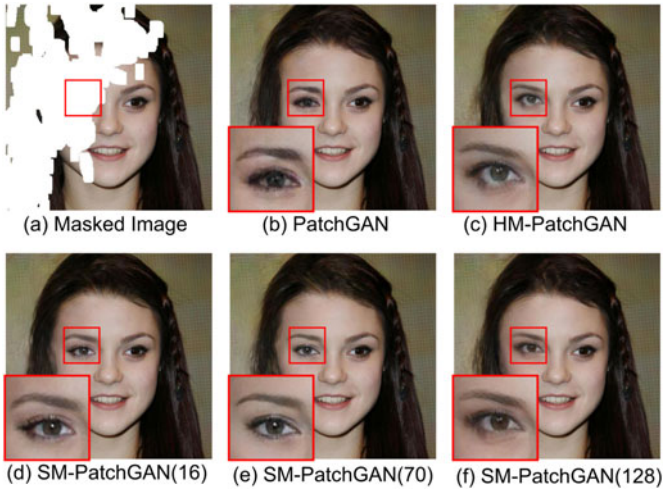

Fig. 9. Visual comparisons on images (512×512) of CELEBA-HQ [61] to verify the effectiveness of the SM-PatchGAN (70) discriminator, where 70 denotes the Gaussian kernel size used for generating soft masks for training. [Best viewed with zoom-in].

4.8.2 Face Editing

The application of face editing is more and more popular on phones. Users usually use face editing to remove wrinkles, scars, or beards to make their face photos more attractive. To evaluate the performance of AOT-GAN on face editing, we use the images of CELEBA-HQ [61] and all images are resized to 512×512 for both training and testing. We show the visual results of face editing by the proposed AOT-GAN in Fig. 11. Results show that, our model is able to complete coherent structures of faces or mouths and generate

Fig. 10. Visual results of the proposed AOT-GAN in the application of logo removal. The AOT-GAN is able to remove logos from images in an undetectable way and fill in missing regions with plausible contents. All these images are from QMUL-OpenLogo [62] and are resized to 832×512. We highlight the inpainting masks covering logos with blue bounding boxes. [Best viewed with zoom-in].

Fig. 11. Visual results of AOT-GAN in the application of face editing. The proposed AOT-GAN is able to generate plausible facial structures and realistic textures for high-resolution images of faces (i.e., 512×512). [Best viewed with zoom-in].

Fig. 12. Visual results of the proposed AOT-GAN in the application of object removal. The proposed AOT-GAN shows promising results in filling irregular holes that are provided by users in high-resolution images (i.e., 512×512). [Best viewed with zoom-in].

clear facial textures. For example, in the third case in Fig. 11, AOT-GAN is able to reconstruct the missing structures of the face and the mouth of the woman face, while synthesizing fine-grained textures of the teeth.

4.8.3 Object Removal

Object removal is widely used in many image editing applications. It aims at removing unwanted objects from images and filling in missing regions with plausible backgrounds. To evaluate the performance of AOT-GAN for object removal, we train the model on Places2 [26] with irregular masks [50]. For testing, we use the masks provided by users as inpainting masks. All the images are resized to 512×512 for both training and testing. We show the results of object removal by the proposed AOT-GAN in Fig. 12. The visual results show that the AOT-GAN is able to remove unwanted objects in different kinds of complex scenes with the real masks provided by users. For example, in the third case in Fig. 12, AOT-GAN removes the man, reconnects the railings and synthesizes clear textures in the wall.

4.9 Limitations

We have shown that the AOT-GAN has achieved state-of-the-art performance with the enhanced generator and discriminator networks. In this section, we discuss two limitations of AOT-GAN as below.

Customized AOT block. In practice, the number of branches and dilation rates in AOT block are empirically studied and set. When the image size changes, the optimal setting may need to search again. We plan to investigate adaptive mechanisms to improve AOT block in the future [70], [71].

Unnecessary boundary propagation. In this paper, we use user-specified masks (e.g., Fig. 12) or masks generated by models (e.g., Fig. 10) as input. However, interactive or automatic object segmentation remains a challenging problem [72]. As shown in the failure case in Fig. 13, the inpainting mask cannot cover the logo well and it leaves some boundaries outside the mask. The remaining boundaries are unobtrusive (as shown in Fig. 13b), however, AOT-GAN tends to propagate the boundaries and results in noticeable artifacts. Therefore, multiple times of interactive operations and refinements are required to include all affected pixels in the mask. We plan to design a semantic-aware propagation mechanism in our future work so that the inpainting model is able to propagate meaningful boundaries while stopping propagating the unobtrusive and useless boundaries.

5 Conclusion

In this paper, we propose to learn aggregated contextual transformations and an enhanced discriminator for high-

Fig. 13. A failure case. The inpainting mask cannot cover the logo well and it leaves some boundaries outside the mask. The remaining boundaries are unobtrusive (as shown in (b)), however, AOT-GAN tends to perform unnecessary boundary propagation and results in noticeable artifacts. [Best viewed with zoom-in].

resolution image inpainting. In comparison to previous approaches, the aggregated contextual transformations show impressive improvements by allowing to capture both informative distant contexts and rich patterns of interest for extremely large missing regions. Besides, the discriminator is improved by a novel mask prediction task, which in turn facilitates the generator to synthesize more realistic textures. We conduct extensive evaluations including a user study to show that AOT-GAN outperforms the state-of-the-art approaches by a significant margin. Besides, we conduct ablation studies to analyze each component of AOT-GAN. We also show the results of AOT-GAN on three practical applications to verify the effectiveness of AOT-GAN in the real-world. As our future work, we plan to extend AOT-GAN to other low-level vision tasks, such as single image super-resolution, image denoising, and image-to-image translation.

ACKNOWLEDGMENTS

This work is conducted when Yanhong Zeng was a research intern in Microsoft Research.

REFERENCES

[1] M. Bertalmio, G. Sapiro, V. Caselles, and C. Ballester, "Image inpainting," in *Proc. Annu. Conf. Comput. Graph. Interactive Techn.*, 2000, pp. 417–424.
[2] C. Barnes, E. Shechtman, A. Finkelstein, and D. B. Goldman, "PatchMatch: A randomized correspondence algorithm for structural image editing," *ACM Trans. Graph.*, vol. 28, no. 3, pp. 24:1–24:11, 2009.
[3] A. Criminisi, P. Pérez, and K. Toyama, "Region filling and object removal by exemplar-based image inpainting," *IEEE Trans. Image Process.*, vol. 13, no. 9, pp. 1200–1212, Sep. 2004.
[4] Y. Wexler, E. Shechtman, and M. Irani, "Space-time completion of video," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 29, no. 3, pp. 463–476, Mar. 2007.
[5] N. Kawai, T. Sato, and N. Yokoya, "Diminished reality based on image inpainting considering background geometry," *IEEE Trans. Vis. Comput. Graphics*, vol. 22, no. 3, pp. 1236–1247, Mar. 2016.
[6] J. Herling and W. Broll, "High-quality real-time video inpainting with PixMix," *IEEE Trans. Vis. Comput. Graphics*, vol. 20, no. 6, pp. 866–879, Jun. 2014.
[7] J. Yu, Z. Lin, J. Yang, X. Shen, X. Lu, and T. S. Huang, "Free-form image inpainting with gated convolution," in *Proc. Int. Conf. Comput. Vis.*, 2019, pp. 4471–4480.
[8] Z. Yi, Q. Tang, S. Azizi, D. Jang, and Z. Xu, "Contextual residual aggregation for ultra high-resolution image inpainting," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2020, pp. 7505–7514.
[9] M. Bertalmio, L. Vese, G. Sapiro, and S. Osher, "Simultaneous structure and texture image inpainting," *IEEE Trans. Image Process.*, vol. 12, no. 8, pp. 882–889, Aug. 2003.
[10] A. Gilbert, M. Trumble, A. Hilton, and J. Collomosse, "Inpainting of wide-baseline multiple viewpoint video," *IEEE Trans. Vis. Comput. Graphics*, vol. 26, no. 7, pp. 2417–2428, Jul. 2020.
[11] Q. Guo, S. Gao, X. Zhang, Y. Yin, and C. Zhang, "Patch-based image inpainting via two-stage low rank approximation," *IEEE Trans. Vis. Comput. Graphics*, vol. 24, no. 6, pp. 2023–2036, Jun. 2018.
[12] J. Sun, L. Yuan, J. Jia, and H.-Y. Shum, "Image completion with structure propagation," *ACM Trans. Graph.*, vol. 24, no. 3, 2005, pp. 861–868.
[13] J. Yu, Z. Lin, J. Yang, X. Shen, X. Lu, and T. S. Huang, "Generative image inpainting with contextual attention," in *Proc. Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 5505–5514.
[14] 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.*, 2016, pp. 2536–2544.
[15] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2016, pp. 770–778.
[16] I. Goodfellow *et al.*, "Generative adversarial nets," in *Proc. 27th Int. Conf. Neural Informat. Process. Syst.*, 2014, pp. 2672–2680.
[17] K. Nazeri, E. Ng, T. Joseph, F. Qureshi, and M. Ebrahimi, "EdgeConnect: Generative image inpainting with adversarial edge learning," in *Proc. Int. Conf. Comput. Vis.*, 2019.
[18] Y. Zeng, J. Fu, and H. Chao, "Learning joint spatial-temporal transformations for video inpainting," in *Proc. Eur. Conf. Comput. Vis.*, 2020, pp. 528–543.
[19] C. Yang, X. Lu, Z. Lin, E. Shechtman, O. Wang, and H. Li, "High-resolution image inpainting using multi-scale neural patch synthesis," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 4076–4084.
[20] Y. Wang, X. Tao, X. Qi, X. Shen, and J. Jia, "Image inpainting via generative multi-column convolutional neural networks," in *Proc. Int. Conf. Neural Informat. Process. Syst.*, 2018, pp. 331–340.
[21] Y. Zeng, J. Fu, H. Chao, and B. Guo, "Learning pyramid-context encoder network for high-quality image inpainting," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2019, pp. 1486–1494.
[22] H. Liu, B. Jiang, Y. Xiao, and C. Yang, "Coherent semantic attention for image inpainting," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2019, pp. 4169–4178.
[23] P. Wang *et al.*, "Understanding convolution for semantic segmentation," in *Proc. IEEE Winter Conf. Appl. Comput. Vis.*, 2018, pp. 1451–1460.
[24] 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.*, 2017, pp. 5967–5976.
[25] W. Xiong *et al.*, "Foreground-aware image inpainting," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2019, pp. 5833–5841.
[26] 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. 2018.
[27] 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.
[28] L. A. Gatys, A. S. Ecker, and M. Bethge, "Image style transfer using convolutional neural networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2016, pp. 2414–2423.
[29] C. Ballester, V. Caselles, J. Verdera, M. Bertalmio, and G. Sapiro, "A variational model for filling-in gray level and color images," in *Proc. Int. Conf. Comput. Vis.*, 2001, pp. 10–16.
[30] M. Bertalmio, A. L. Bertozzi, and G. Sapiro, "Navier-stokes, fluid dynamics, and image and video inpainting," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, 2001, p. I.
[31] L. Liang, C. Liu, Y.-Q. Xu, B. Guo, and H.-Y. Shum, "Real-time texture synthesis by patch-based sampling," *ACM Trans. Graph.*, vol. 20, no. 3, pp. 127–150, 2001.
[32] A. A. Efros and T. K. Leung, "Texture synthesis by non-parametric sampling," in *Proc. Int. Conf. Comput. Vis.*, 1999, pp. 1033–1038.
[33] A. A. Efros and W. T. Freeman, "Image quilting for texture synthesis and transfer," in *Proc. Annu. Conf. Comput. Graph. Interactive Techn.*, 2001, pp. 341–346.
[34] H. Xue, B. Liu, H. Yang, J. Fu, H. Li, and J. Luo, "Learning fine-grained motion embedding for landscape animation," in *Proc. 29th ACM Int. Conf. Multimedia*, 2021, pp. 291–299.
[35] H. Zheng, H. Yang, J. Fu, Z.-J. Zha, and J. Luo, "Learning conditional knowledge distillation for degraded-reference image quality assessment," in *Proc. Int. Conf. Comput. Vis.*, 2021, pp. 10242–10251.
[36] F. Yang, H. Yang, J. Fu, H. Lu, and B. Guo, "Learning texture transformer network for image super-resolution," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2020, pp. 5790–5799.
[37] Y. Zeng, H. Yang, H. Chao, J. Wang, and J. Fu, "Improving visual quality of image synthesis by a token-based generator with transformers," in *Proc. Int. Conf. Neural Informat. Process. Syst.*, 2021.
[38] Y. Li, S. Liu, J. Yang, and M.-H. Yang, "Generative face completion," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 3911–3919.
[39] R. A. Yeh, C. Chen, T. Yian Lim , A. G. Schwing, M. Hasegawa-Johnson , and M. N. Do, "Semantic image inpainting with deep generative models," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 6882–6890.
[40] S. Iizuka, E. Simo-Serra , and H. Ishikawa, "Globally and locally consistent image completion," *ACM Trans. Graph.*, vol. 36, no. 4, 2017, Art. no. 107.
[41] Z. Yan, X. Li, M. Li, W. Zuo, and S. Shan, "Shift-net: Image inpainting via deep feature rearrangement," in *Proc. Eur. Conf. Comput. Vis.*, 2018, pp. 1–17.
[42] C. Xie *et al.*, "Image inpainting with learnable bidirectional attention maps," in *Proc. Int. Conf. Comput. Vis.*, 2019, pp. 8858–8867.

[43] Y. Song *et al.*, "Contextual-based image inpainting: Infer, match, and translate," in *Proc. Eur. Conf. Comput. Vis.*, 2018, pp. 3–19.
[44] M.-C. Sagong, Y.-G. Shin, S.-W. Kim, S. Park, and S.-J. Ko, "PEPSI: Fast image inpainting with parallel decoding network," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2019, pp. 11352–11360.
[45] F. Yu and V. Koltun, "Multi-scale context aggregation by dilated convolutions," in *Proc. Int. Conf. Learn. Representations*, 2016.
[46] L.-C. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and A. L. Yuille, "DeepLab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected CRFs," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 40, no. 4, pp. 834–848, Apr. 2018.
[47] Z. Wang and S. Ji, "Smoothed dilated convolutions for improved dense prediction," in *Proc. 24th ACM SIGKDD Int. Conf. Knowl. Discov. Data Mining*, 2018, pp. 2486–2495.
[48] Z. Hu, J. Jia, B. Liu, Y. Bu, and J. Fu, "Aesthetic-aware image style transfer," in *Proc. 29th ACM Int. Conf. Multimedia*, 2020, pp. 3320–3329.
[49] K. Hong, S. Jeon, H. Yang, J. Fu, and H. Byun, "Domain-aware universal style transfer," in *Proc. Int. Conf. Comput. Vis.*, 2021, pp. 14609–14617.
[50] 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.
[51] Y. Ren, X. Yu, R. Zhang, T. H. Li, S. Liu, and G. Li, "Structureflow: Image inpainting via structure-aware appearance flow," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2019, pp. 181–190.
[52] T. Miyato, T. Kataoka, M. Koyama, and Y. Yoshida, "Spectral normalization for generative adversarial networks," in *Proc. Int. Conf. Learn. Representations*, 2018.
[53] S. Xie, R. Girshick, P. Dollár, Z. Tu, and K. He, "Aggregated residual transformations for deep neural networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 5987–5995.
[54] C.-T. Li, W.-C. Siu, Z.-S. Liu, L.-W. Wang, and D. P.-K. Lun, "DeepGIN: Deep generative inpainting network for extreme image inpainting," in *Proc. Eur. Conf. Comput. Vis. Workshop*, 2020, pp. 5–22.
[55] J. Qin, H. Bai, and Y. Zhao, "Multi-scale attention network for image inpainting," *Comput. Vis. Image Understanding*, vol. 204, 2021, Art. no. 103155.
[56] X. Mao, Q. Li, H. Xie, R. Y. Lau, Z. Wang, and S. Paul Smolley , "Least squares generative adversarial networks," in *Proc. Int. Conf. Comput. Vis.*, 2017, pp. 2794–2802.
[57] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," 2014, *arXiv:1409.1556*.
[58] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei , "ImageNet: A large-scale hierarchical image database," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2009, pp. 248–255.
[59] Z. Wan, J. Zhang, D. Chen, and J. Liao, "High-fidelity pluralistic image completion with transformers," in *Proc. Int. Conf. Comput. Vis.*, 2021, pp. 4692–4701.
[60] S. Zhao *et al.*, "Large scale image completion via co-modulated generative adversarial networks," in *Proc. Int. Conf. Learn. Representations*, 2020.
[61] T. Karras, T. Aila, S. Laine, and J. Lehtinen, "Progressive growing of GANs for improved quality, stability, and variation," in *Proc. Int. Conf. Learn. Representations*, 2018.
[62] H. Su, X. Zhu, and S. Gong, "Open logo detection challenge," in *Proc. Brit. Mach. Vis. Conf.*, 2018, Art. no. 16.
[63] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in *Proc. Int. Conf. Med. Image Comput. Comput.-Assist. Interv.*, 2015, pp. 234–241.
[64] Z. Wang, E. P. Simoncelli, and A. C. Bovik, "Multiscale structural similarity for image quality assessment," in *Proc. 37th Asilomar Conf. Signals, Syst. Comput.*, 2003, pp. 1398–1402.
[65] M. Heusel, H. Ramsauer, T. Unterthiner, B. Nessler, and S. Hochreiter, "GANs trained by a two time-scale update rule converge to a local nash equilibrium," in *Proc. 31st Int. Conf. Neural Informat. Process. Syst.*, 2017, pp. 6626–6637.
[66] R. Zhang, P. Isola, A. A. Efros, E. Shechtman, and O. Wang, "The unreasonable effectiveness of deep features as a perceptual metric," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 586–595.
[67] F. Pukelsheim, "The three sigma rule," *The Amer. Statistician*, vol. 48, no. 2, pp. 88–91, 1994.
[68] W.-Q. Yan, J. Wang, and M. S. Kankanhalli, "Automatic video logo detection and removal," *Multimedia Syst.*, vol. 10, no. 5, pp. 379–391, 2005.
[69] C. Qin, Z. He, H. Yao, F. Cao, and L. Gao, "Visible watermark removal scheme based on reversible data hiding and image inpainting," *Signal Process. Image Commun.*, vol. 60, pp. 160–172, 2018.
[70] J. Dai, H. Qi, Y. Xiong, Y. Li, G. Zhang, H. Hu, and Y. Wei, "Deformable convolutional networks," in *Proc. Int. Conf. Comput. Vis.*, 2017, pp. 764–773.
[71] X. Zhu, H. Hu, S. Lin, and J. Dai, "Deformable ConvNets v2: More deformable, better results," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2019, pp. 9308–9316.
[72] J.-B. Huang, S. B. Kang, N. Ahuja, and J. Kopf, "Temporally coherent completion of dynamic video," *ACM Trans. Graph.*, vol. 35, no. 6, pp. 1–11, 2016.

Yanhong Zeng (Student Member IEEE) received the BS degree from the School of Data and Computer Science, Sun Yat-sen University, Guangzhou, China, in 2017, where she is currently working toward the PhD degree. Her research interests include image processing, deep learning, and computer vision.

Jianlong Fu (Senior Member, IEEE) received the PhD degree in pattern recognition and intelligent system from the Institute of Automation, Chinese Academy of Science in 2015. He is currently a lead researcher with the Multimedia Search and Mining Group, Microsoft Research Asia (MGPA). His current research interests include computer vision, computational photography, vision and language. He has authored or coauthored more than 40 papers in journals and conferences, and one book chapter. He serves as a lead organizer and guest editor of *IEEE Trans. Pattern Analysis and Machine Intelligence Special Issue on Fine-grained Categorization*. He is an area chair of ACM Multimedia 2018, ICME 2019. He received the Best Paper Award from ACM Multimedia 2018, and has shipped core technologies to a number of Microsoft products, including Windows, Office, Bing Multimedia Search, Azure Media Service and XiaoIce.

Hongyang Chao (Member IEEE) received the BS and PhD degrees in computational mathematics from the Sun Yet-sen University, Guangzhou, China. In 1988, she joined the Department of Computer Science, Sun Yet-sen University, where she was initially an assistant professor and later became an associate professor. She is currently a full professor with the School of Computer Science. She has published extensively in the area of image/video processing and holds three U.S. patents and four Chinese patents in the related area. Her current research interests include image and video processing, image and video compression, massive multimedia data analysis, and content-based image (video) retrieval.

Baining Guo (Fellow IEEE) received the BS degree from the Beijing University, and the MS and PhD degrees from the Cornell University. He is currently the assistant managing director with the Microsoft Research Asia, where he also serves as the head of the Graphics Lab. Prior to joining Microsoft in 1999, he was a senior staff researcher with the Microcomputer Research Labs of Intel Corporation in Santa Clara, California. His research interests include computer graphics and visualization, in the areas of texture and reflectance modeling, texture mapping, translucent surface appearance, real-time rendering, and geometry modeling.