A Joint Guidance-Enhanced Perceptual Encoder and Atrous Separable Pyramid-Convolutions for Image Inpainting

Yongle Zhang*, Yingyu Wang*, Junyu Dong*¶, Lin Qi*, Hao Fan*, Xinghui Dong†, Muwei Jian‡§ and Hui Yu§
*Department of Computer Science and Technology, Ocean University of China, Qingdao, China
Email:{zhangyongle, wangyingyu}@stu.ouc.edu.cn, {dongjunyu, qilin, fanhao}@ouc.edu.cn
†University of Manchester, Manchester, UK, Email: dongxinghui@gmail.com
‡Shandong University of Finance and Economics, Jinan, China, Email: 20173016@sdufe.edu.cn
§University of Portsmouth, Portsmouth, UK, Email: hui.yu@port.ac.uk
¶Institute for Advanced Ocean Study

***Abstract*—Satisfactory image inpainting requires visually-exquisite details and semantically-plausible structures, where encoder-decoder networks have shown their potentials but bear undesired local and global inconsistencies, such as blurry textures. To address this issue, we incorporate a perception operation in the encoder, which extracts features from known areas of the input image, to improve textured details in missing areas. We also propose an iterative guidance loss for the perception operation to guide perceptual encoding features approaching to ground-truth encoding features. The guidance-enhanced perceptual encoding features are transferred to the decoder through skip connections, mutually reinforcing the entire encoder-decoder performance. Since the inpainting task involves different levels of feature representations, we further apply atrous separable parallel-convolutions (i.e. atrous separable pyramid-convolutions or ASPC) with different receptive fields in the last guidance-enhanced perceptual encoding feature, which is used to learn high-level semantic features with multi-scale information. Experiments on public databases show that the proposed method achieves promising results in terms of visual details and semantic structures.**

***Index Terms*—Image Inpainting, Perceptual Encoder-Decoder, Generative Adversarial Networks**

I. Introduction

Image restoration aims to restore missing pixels in damaged images, making observers cannot discern these images are restored [1]. Image inpainting has attracted much attention for decades [2]–[5], and has been applied in many practical scenarios, such as photo editing, image generation, object removal and hole-filling [6]–[10]. The main challenge of image restoration is to generate a plausible result with global semantic structures and visually consistent texture details. Early researches usually used texture synthesis [2] and patch-based inpainting methods [4], [11]. For example, Barnes *et al.* [4] used a randomized nearest neighbor algorithm to search the most similar patch in known regions of the input image to fill missing parts. These methods well performed in synthesizing fine textures, but are usually incompetent to retain global structures and reasonable semantics.

Recently, researchers trained Deep Convolutional Neural Networks (CNNs) with a large number of samples and predicted plausibly semantic missing parts of an image [5], [12], [13]. Yu *et al.* found that general CNN-based methods are ineffective in establishing far enough relationships between the missing region and context information, resulting in boundary artifacts and blurry textures [13]. They proposed that traditional texture and patch synthesis methods are reliable when it needs to exploit surrounding texture information. To take advantages of traditional inpainting methods and CNNs, they proposed a contextual attention layer to effectively reconstruct missing patches using features of known background patches. However, the contextual attention independently existed in the second stage of the two stacked encoder-decoder network. To further ensure that both fine-detail and global-semantic consistency can be satisfied, we propose an enhanced contextual attention operation and integrate it into multiple convolutional layers for a compact encoder-decoder network. The multiple attention layers progressively learn similar region relationships by known background attention from the shallow feature map to the high-level semantic feature map in the encoder, and these learned attention features are then transferred to the decoder.

In our work, we first employ an encoder-decoder network [14] with skips as the backbone. In previous image restoration studies [5], [13], [15], [16], researchers applied encoder-decoder architectures, which produces remarkable performance. The encoder encodes the full resolution input into the high-level feature space, and the decoder gradually restores the spatial features. Meanwhile, we used the skip to connect encoding features of our guidance-enhanced perceptual encoder into the decoder. This two basic design is significant to the image restoration task, and the skeleton architecture is shown in Fig. 1.

Moreover, we design a perception encoder by adding the

Junyu Dong and Lin Qi are co-corresponding authors

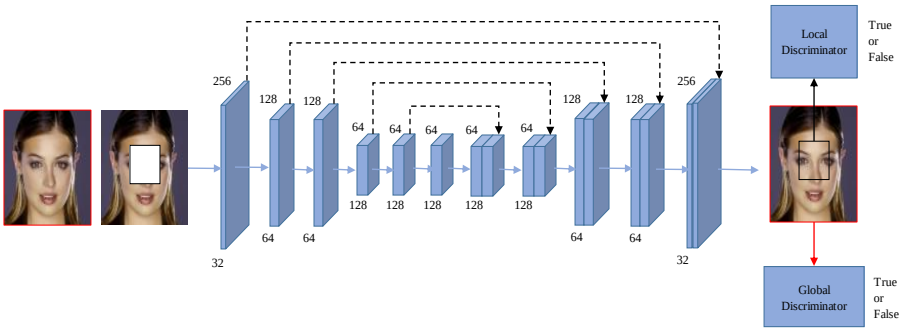


Fig. 1. The architecture of the skeleton network, which is referred to as S-Net. This network is trained using the reconstruction loss of spatial-variant weights, and the global and local adversarial losses.

perception operation (i.e., contextual attention) to the last three layers of the encoder. This not only iteratively improves the network performance on learning perceptual correlations (i.e., similarity correlations) between the known and unknown feature patches of the masked image, but also generates missing feature patches using the weighted feature patches sampled from known regions. On top of the last three perception encoding layers, we propose an iterative guidance loss (L1 Loss) function to minimize the distance between the perceptual encoding features and the ground-truth encoding features, which can enhance the effectiveness of the perception operation.

To fulfill the purpose of image restoration with different levels of feature representations, we use the atrous separable pyramid-convolution (ASPC) to extract multi-scale features with high-level semantics. Particularly, the low-level features (with more details) extracted at the first two layers of the encoder and the guidance-enhanced perceptual features are transferred to the decoder through the skips. As a result, the compact-designed network can recover the resolution information with fine details and semantic structures. The proposed network (see Fig. 4) can be end-to-end trained and is optimized using the iterative guidance loss in the encoder, the reconstruction loss of the spatial-variant weight in the generative network, and the local and global adversarial losses. In our network, we did not add perception operation to the shallow layers of the encoder in order to connect the low-level features with rich details to the decoder. Besides, it can simplify the parameter calculation of the generative network.

The contributions of this study can be summarized as follows.

- We propose an iterative guidance loss (IGL) to guide the encoding direction of the perception layer, which makes the encoder to map the input into high-level semantic feature space under accurate guidance. Meanwhile, the skip connection propagates the guidance-enhanced perceptual encoding features to the decoder layer, which enables the decoder to generate high-quality inpainting.
- We employ the atrous separable pyramid-convolutions (ASPC) to capture multi-scale context information in the guidance-enhanced perceptual encoding feature space. These semantic features with different scale representations are also delivered to the decoder.
- The network is trained using reconstruction loss with spatial-variant weights and an adversarial loss. These losses enable the generative network to generate visually detailed and semantically plausible results on multiple data sets, including the face, texture and natural scene data sets.

II. Related Work

In this section, we first investigate the related work on image restoration, including traditional methods and generative neural networks. Then, the most relevant encoder-decoder networks are reviewed.

A. Traditional image restoration methods

Traditional diffusion-based and patch-based approaches [1], [2], [4], [11] normally use some distance field metrics [1], [17] to propagate the surface information from neighboring pixels to the missing regions.

The diffusion-based method iteratively spreads the information around the holes to fill missing areas [1]. This method produces promising results with small and narrow holes but fails to restore the large and complex hole region such as face image. In contrast, patch-based methods [2], [4], [11] perform well in filling relatively large holes, which find matching patches from non-missing areas of the image dataset and paste them to the missing area. Barnes *et al.* [4] proposed a fast randomized nearest neighbor algorithm, namely, PatchMatch, which has produced satisfactory results when the image is self-similar.

To summarize, the above diffusion-based and patch-based methods can normally be used to synthesize clear texture structures when the missing hole is small. Nevertheless, it is difficult to use these methods to fill the large hole with the global semantic rationality because they do not capture the high-level semantics or the global structure of an image.

B. Generative image restoration methods

As the deep convolutional neural networks become prevalent in the computer vision community [18]–[21], the genera-

tive network based on the adversarial model leads to a guiding direction for achieving the plausible semantic results.

Pathak *et al.* [5] used a deep generative model, called context encoders, along with the adversarial network to fill the center region of 64×64. Although semantically reasonable results were produced, the details were lost. Based on context encoders, Lizuka *et al.* [22] proposed the global and local discriminators in order to improve the adversarial network module of the original GANs [23]. This adaption improved the coherence of the local texture and the global semantics. The global and local discriminators have also been used in image or face completion [13], [15], [16], [24].

Based on the work conducted in [22], Yu *et al.* [13] designed a two-stage coarse-to-fine generative network with a contextual attention strategy using the nearest neighbor patch matching in feature spaces. This network shows a remarkable performance, but the error of the coarse restoration result is also brought into the fine network. Wang *et al.* [16] observed that the one-stream encoder-decoder model can achieve satisfactory results but it only uses the same size of receptive fields to transform the image into a common feature space. However, image restoration involves different levels of feature representations, so they designed three parallel encoder-decoder branches with different receptive fields to extract different scales of features. Compared with other reconstruction losses used in image restoration [5], [13], [22], [25], Wang *et al.* [16] also proposed a confidence-driven reconstruction loss, which explores the spatial location and relative priority by propagating the confidence of known pixels to unknown pixels.

The generative model was also applied to face completion fields. Li *et al.* [15] found that semantic retrieval [26] or semantic segmentation helps repair certain object (e.g., faces). Therefore, they proposed an additional semantic parsing network for normalizing the generative network. Song *et al.* [24] designed a facial geometry network and used it to learn the face landmark heatmaps and parsing maps. The predicted facial geometry images were transmitted to the generative network for producing a completion result. Yeh *et al.* [25] proposed a generative model with the context and prior losses. This network was used to find the closest encoding for the corrupted image in the latent picture manifold to enhance the semantic rationality of the inpainting results. Then, the closest encoding was input in a generative model to infer the missing content.

C. The Encoder-decoder with the skip connection

The encoder-decoder network [14] has been widely used in many computer vision tasks and has achieved state-of-the-art results. In semantic segmentation fields [27]–[29], Chen *et al.* [28] introduced that the encoder gradually captures the higher-level semantic information from the low-level input image, and the corresponding decoder gradually restores the spatial information. Furthermore, image inpainting also involves the plausible semantic requirement.

To ensure this requirement, many image restoration studies used the encode-decoder architecture [15], [22], [30]. In [15], it was demonstrated that the encoder has the ability to encode the masked image into hidden representations (with higher-level semantic information). These representations capture the interrelationship between the unknown and known region and were further fed into the decoder to generate the restoration results. In image translation tasks [31], some low-level information is shared between the input and output, and it is desirable to pass the low-level information directly through the network. For image restoration, the low-level information also played an important role in local structure details. The U-Net [32] allows the low-level information to be transmitted quickly across the network. Isola *et al.* [31] observed that the encoder-decoder networks with disconnecting the skip connection in the U-Net could not produce realistic image results. It was also shown that the combination of the skip connections in the U-Net and the encoder-decoder together achieved the best result.

TABLE I
THE DETAIL ARCHITECTURE OF THE LOCAL DISCRIMINATOR.

Type	Kernel	Dilation	Stride	Outputs	Activation
L_conv1	5×5	1	2×2	64	Leaky ReLU
L_conv2	5×5	1	2×2	128	Leaky ReLU
L_conv3	5×5	1	2×2	256	Leaky ReLU
L_conv4	5×5	1	2×2	512	Leaky ReLU
L_FC	1×1	1	1×1	1	Leaky ReLU

TABLE II
THE DETAIL ARCHITECTURE OF THE GLOBAL DISCRIMINATOR.

Type	Kernel	Dilation	Stride	Outputs	Activation
G_conv1	5×5	1	2×2	64	Leaky ReLU
G_conv2	5×5	1	2×2	128	Leaky ReLU
G_conv3	5×5	1	2×2	256	Leaky ReLU
G_conv4	5×5	1	2×2	256	Leaky ReLU
G_FC	1×1	1	1×1	1	Leaky ReLU

III. METHODOLOGY

Our end-to-end image inpainting network consists of an encoder-decoder combined with the skip connection. We refer to this network as the skeleton network. Fig. 1 shows the architecture of this network. In particular, we introduce a guidance-enhanced perceptual encoding layer and utilize atrous separable pyramid-convolutions for extracting multi-scale features. We will provide the details of the two modules in Sections III-A and III-B respectively. Besides, the reconstruction loss of spatial-variant weights and the adversarial loss used for our network are introduced in Sections III-C and III-D respectively.

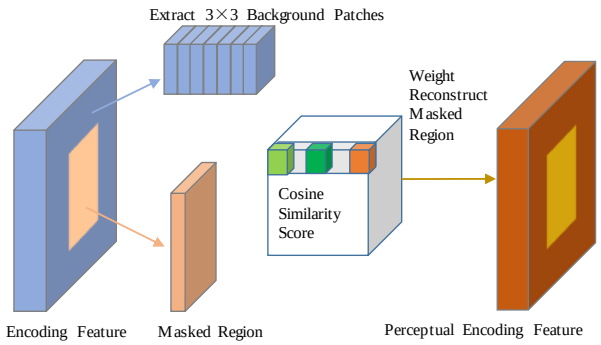


Fig. 2. The diagram of the contextual attention model [13].

A. The guidance-enhanced perceptual encoding layer

Yu *et al.* [13] proposed a contextual attention model (see Fig. 2 for its architecture). The main idea behind this model is to extract 3×3 background patches $b_{i,j}$ from the background outside the masked region at an encoding feature. The similarity between the masked region patches m_{i^*,j^*} and the background region patches $b_{i,j}$ is calculated using the cosine similarity (i.e., the normalized inner product), as expressed below:

$$S_{(i^*,j^*)(i,j)} = \left\langle \frac{m_{i^*,j^*}}{\|m_{i^*,j^*}\|}, \frac{b_{i,j}}{\|b_{i,j}\|} \right\rangle . \tag{1}$$

The cosine similarity score of each background patch is obtained by the *softmax* with a scale of ω in the similarity. The weighted cosine similarity score is used to reconstruct the masked region. This is depicted as:

$$W_{S_{(i^*,j^*)(i,j)}} = softmax\left(\omega S_{(i^*,j^*)(i,j)}\right). \tag{2}$$

The above-mentioned perceptual operation is applied to the last three layers of the encoder.

The iterative guidance loss (see Fig. 3) is used to guide the accurate encoding direction of the perceptual encoding feature, forming the guidance-enhanced perceptual encoding feature. The perceptual feature map $x3\_p$ is derived using the perception operation on the encoding features $x3$ extracted at the third layer. The convolution operation is performed on the perceptual feature map $x3\_p$ to generate the encoding feature map $x4$. This operation is conducted until the fifth-layer perceptual feature map $x5\_p$ is obtained.

Furthermore, we perform the iterative guidance loss (Equation 3) on the three-layer perceptual encoding features $x3\_p$, $x4\_p$ and $x5\_p$ to minimize the distance between these features and the corresponding ground-truth encoding features. The loss is expressed as:

$$L_{IGL} = \sum_{i=3}^{n} L_i. \tag{3}$$

Specifically, the losses applied to the three layers are defined as:

$$L_3 = \|p\left(E_3\left(I\right)\right) - E_3\left(I_{gt}\right)\|_1, \tag{4}$$

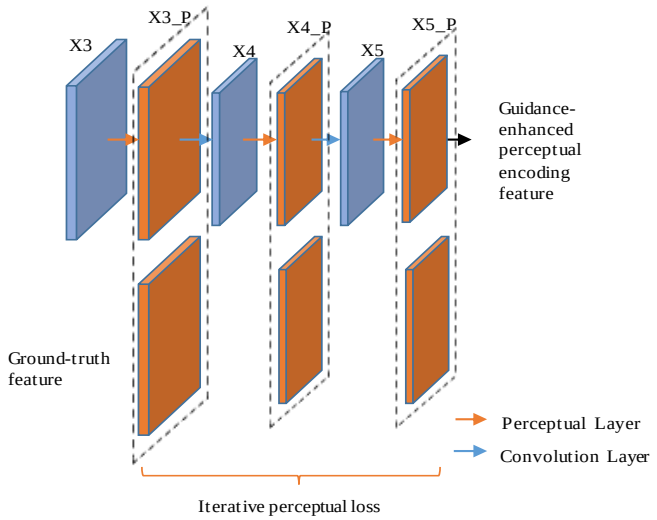


Fig. 3. The diagram of the guidance-enhanced perceptual encoding layer.

$$L_4 = \|p\left(E_4\left(p\left(E_3\left(I\right)\right)\right)\right) - E_4\left(I_{gt}\right)\|_1, \tag{5}$$

and

$$L_5 = \|p\left(E_5\left(p\left(E_4\left(p\left(E_3\left(I\right)\right)\right)\right)\right)\right) - E_5\left(I_{gt}\right)\|_1 \tag{6}$$

in turn, where I represents a masked input image, I_{gt} denotes a ground-truth image, E_i stands for the encoding operation of the i-th layer, and p represents the perception operation. For example, $E_3\left(I\right)$ extracts the features represented by $x3$ as shown in Fig. 3, and $p\left(E_3\left(I\right)\right)$ calculates the perceptual encoding features $x3\_p$ as shown in Fig. 3. The use of the iterative guidance loss further guides the encoding direction of the perception operation to the global correctness.

We name the combination of the perceptual layer operation and the iterative guidance loss as the guidance-enhanced perceptual encoding layer, which produces higher-level semantic features with fine details in the encoder. These features are propagated through the skips to the corresponding decoding layer. This design also improves the performance of the decoder.

B. The atrous separable pyramid-convolution

Multiple atrous separable convolution layers with different rates have been used to capture the multi-scale information in many semantic segmentation tasks [28], [33], [34]. Comparably, the multi-scale features were also applied to the restoration task [16] because they help generate semantically plausible results and novel contents. Inspired by these studies, we use the atrous separable convolution in parallel at the proposed network.

To be specific, the multi-scale information is captured by processing input feature maps using filters with different dilated rates. Benefiting from the encoder-decoder model with the skip connection and the guidance-enhanced perceptual layer used in our network, the final encoding features processed by the atrous separable pyramid-convolution (ASPC)

also possess the higher-level semantics and the fine-detailed information. The structure of the ASPC is shown in Fig. 4.

1) The atrous convolution: The atrous convolution [33], [35] is a powerful tool to adjust the field-of-view of the filter and control the resolution of deep features. By changing the dilated rate, the receptive field is expanded to capture multi-scale feature information. (When the dilated rate is 1, the atrous convolution becomes a standard convolution). This convolution operation has achieved success in solving multi-scale problems of semantic segmentation [28], [33]. It is also used in visual processing tasks, such as image restoration [13], [16], [22]. In [16], it has been demonstrated that the encoder-decoder network with different sizes of receptive fields improves experiment results.

2) The depthwise separable convolution: The depthwise separable convolution [36], [37] factorizes a standard convolution operation into two steps. The first step is a depthwise convolution, which uses the same filter to convolve with each input channel. The second step is a pointwise convolution which uses a 1×1 convolution to collapse the output of the depthwise convolution operation across different channels. Compared to the standard convolution operation, the depthwise separable convolution can reduce the number of network parameters while it can achieve the equivalent (or even better) performance. For example, there is a 3×3 convolutional layer with an input channel of 16 and an output channel of 32. The standard convolution operation uses 32 3×3 convolution kernels to convolve with the input image. In this case, each convolution kernel requires 3×3×16 parameters while the resulting output has only one channel. The 32 convolution kernels need a total of $(3 \times 3 \times 16) \times 32 = 4608$ parameters. In contrast, 16 feature maps are obtained by traversing 16 input channels using 16 3×3 convolution kernels when the depthwise separable convolution is utilized. Before the fusion operation is performed, the 16 feature maps are traversed with 32 1×1 convolution kernels. This process uses $16 \times 3 \times 3 + 16 \times 32 \times 1 \times 1 = 656$ parameters, which are less than the 4608 parameters required by the standard convolution.

C. The reconstruction loss of spatial-variant weights

The reconstruction loss of spatial-variant weights (which is also known as the confidence-driven reconstruction loss) was originally introduced by Wang et al. [16]. Inpainting tasks involve hallucination of plausible pixels to fill the missing region. L1 loss is prone to smooth pixels and produces blurry results on the restoration region. The reconstruction loss of spatial-variant weights considers spatial locations and relative order by propagating confidence from known pixels to unknown pixels. Given that the confidence level of known pixels is set to 1, the confidence of unknown pixels in the masked region is related to the distance between the pixel and the boundary. In this context, a 64×64 Gaussian filter G is used to convolve with $\bar{M}$ in order to create a weighted mask M_w:

$$M_w^i = \left(G * \bar{M}^i\right) \odot M, \tag{7}$$

where $\bar{M}^i = 1 - M + M_w^{i-1}$, $M_w^0 = 0$, M represents a binary mask (1 indicates unknown pixels (i.e., masked area) while 0 suggests otherwise) which has the same resolution as that of the ground-truth image, and $i = 1, ..., 7$. $\odot$ is an element-wise multiplication. When $i = 1$, the Gaussian filter G is used to convolve with $1 - M$. The confidence of the known pixels is propagated into the masked area to obtain M_w^1. Then, the $\bar{M}^2$ is updated. After seven iterations, M_w is obtained.

We define the reconstruction loss of the spatial-variable weights as:

$$L_{svw} = \left\| \left(I_{gt} - G\left(I\right)\right) \odot M_w \right\|_1, \tag{8}$$

where I_{gt} denotes the ground-truth image, $G\left(I\right)$ represents the output of the generative network, I is the masked image. In [16], it was shown that this loss was superior to other reconstruction losses for image restoration.

D. The adversarial loss

The adversarial training usually involves a generative network G and a discriminator network D, which are comprised of a Generative Adversarial Network (GAN) [23], [38]. The goal of the generative network G is to generate as realistic images as possible which can deceive the discriminator network D. On the other hand, the goal of D is to distinguish the image generated by G from the ground-truth image. Thus, G and D constitute a dynamic "game process". The original GANs sometimes suffer from vanishing or exploding gradients, which causes the network to produce ambiguous results. Therefore, we use the improved Wasserstein GANs [39] together with both the local and global discriminators [22].

The improved Wasserstein GANs add the gradient penalty into the discriminator loss to solve the problems of the weak modeling and vanishing or exploding gradients caused by the weight clipping of WGANs [40]. The local discriminator focuses on the filling area to enforce local details, while the global discriminator assesses if the inpainting is coherent as a whole. The adversarial loss for the generator is defined as:

$$L_{adv} = -E_{z \sim p_g}\left[D\left(z\right)\right] + \theta_{gp} E_{\hat{z} \sim p_{\hat{z}}}\left[\left(\left\| \nabla_{\hat{z}} D\left(\hat{z}\right) \odot M_w \right\|_2 - 1\right)^2\right], \tag{9}$$

where p_g is the distribution of the generator model defined by $z = G\left(I\right)$ and I is the input of the generative network. $p_{\hat{z}}$ is uniformly sampled along the straight lines between pairs of points sampled from the real data distribution p_r and the generator distribution p_g, and satisfies $\hat{z} = \varepsilon \tilde{z} + \left(1 - \varepsilon\right) z$, $\varepsilon \in [0, 1]$, $\tilde{z}$ belongs to real data distribution p_r. θ_{gp} is set to 10 in our experiments.

Finally, the entire objective function is defined as:

$$L = \gamma_{IGL} L_{IGL} + \gamma_{adv} L_{adv} + \gamma_{svw} L_{svw}, \tag{10}$$

where γ_{IGL}, γ_{adv}, γ_{svw} are used to weigh the iterative guidance loss, the adversarial training loss, and the reconstruction loss respectively.

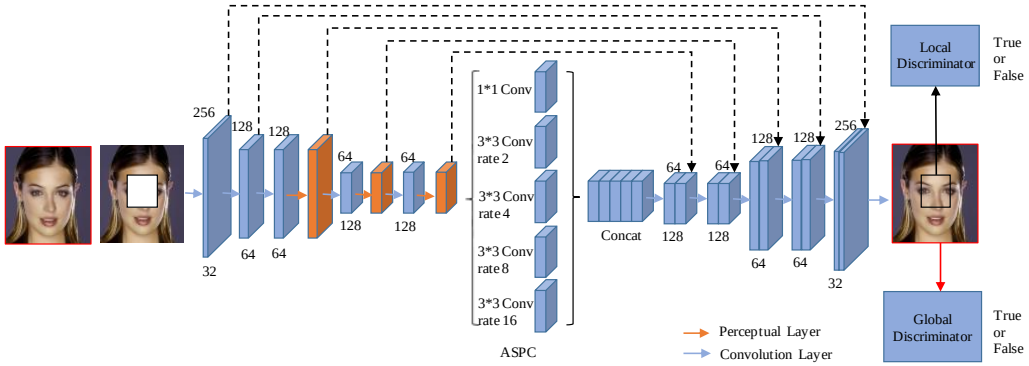


Fig. 4. The full model of the proposed guidance-enhanced perceptual encoder network, which is referred to as GEPE-Net. It consists of the perceptual layer enhanced by IGL and the ASPC operation in the encoder.

(a) Input (b) CA (c) GMCNN (d) PE-Net (e) GEPE-Net(Our) (f) GT

Fig. 5. Qualitative Evaluations of the side face. From the left to the right: (a) Masked input (b) Results of CA [13] (c) Results of GMCNN [16] (d) Results of our PE-Net (e) Results of our full GEPE-Net (f) Ground-truth.

IV. Experiments

In this section, we first introduce the experimental setup. Then, we report the results obtained in our experiments.

A. Experimental Setup

1) Training Data Sets: We conduct our experiments on three data sets, including the CelebA-HQ data set [12], the DTD texture data set [41] and the canyon scene subset of the Places2 data set [42]. For the CelebA-HQ data set, we randomly select 27,000 images and 3000 images as the training and test sets respectively. The DTD data set contains 5,640 texture images of 47 different categories. We randomly select 20 images from each category as the test set and use the remaining 100 images as the training set. The canyon scene data set contains 4,700 training images and 300 test images. All images used in our experiments are resized to the resolution of 256×256 pixels.

2) Training Process: The full network model is shown in Fig. 4. Specifically, the encoder integrates the three perceptual layers enhanced by the iterative guidance loss, as well as the atrous separable pyramid-convolution. The network details of each part are shown in Tables IV, V, I, and II. Given a raw 256×256 image I_{gt} and a binary mask M of the same size, the network input $I = I_{gt} \odot (1 - M)$. Our network is optimized using the Adam algorithm [43] with a learning rate of 0.0001, β_1 =0.5 and β_2 =0.9. The trade-off parameters γ_{IGL}, γ_{adv} and γ_{svw} are set to 0.08, 0.01, and 1.0 respectively. Our network was implemented using Tensorflow v1.6. All experimental results have been derived without the post-processing.

TABLE III
Quantitative results on CelebA-HQ and DTD texture test set.

Method	CelebA-HQ		
	L1 Loss	PSNR	SSIM
CA [13]	6.020	23.663	0.860
GMCNN [16]	**4.916**	**25.161**	0.876
S-Net	8.222	21.002	0.857
PE-Net	5.970	23.688	0.870
GEPE-Net	5.190	24.901	**0.877**
Method	DTD textures		
	L1 Loss	PSNR	SSIM
CA [13]	7.347	23.623	0.813
GEPE-Net	**7.021**	**24.049**	**0.816**

(a) Input (b) S-Net (c) CA (d) GMCNN (e) PE-Net (f) GEPE-Net(Our)

Fig. 6. Qualitative Evaluations of our method and other generative methods on CelebA-HQ datasets. From the left to the right: (a)Masked input (b) Results of our S-Net (our skeleton network) (c) Results of CA [13] (d) Results of GMCNN [16] (e) Results of our PE-Net (f) Results of our full GEPE-Net.

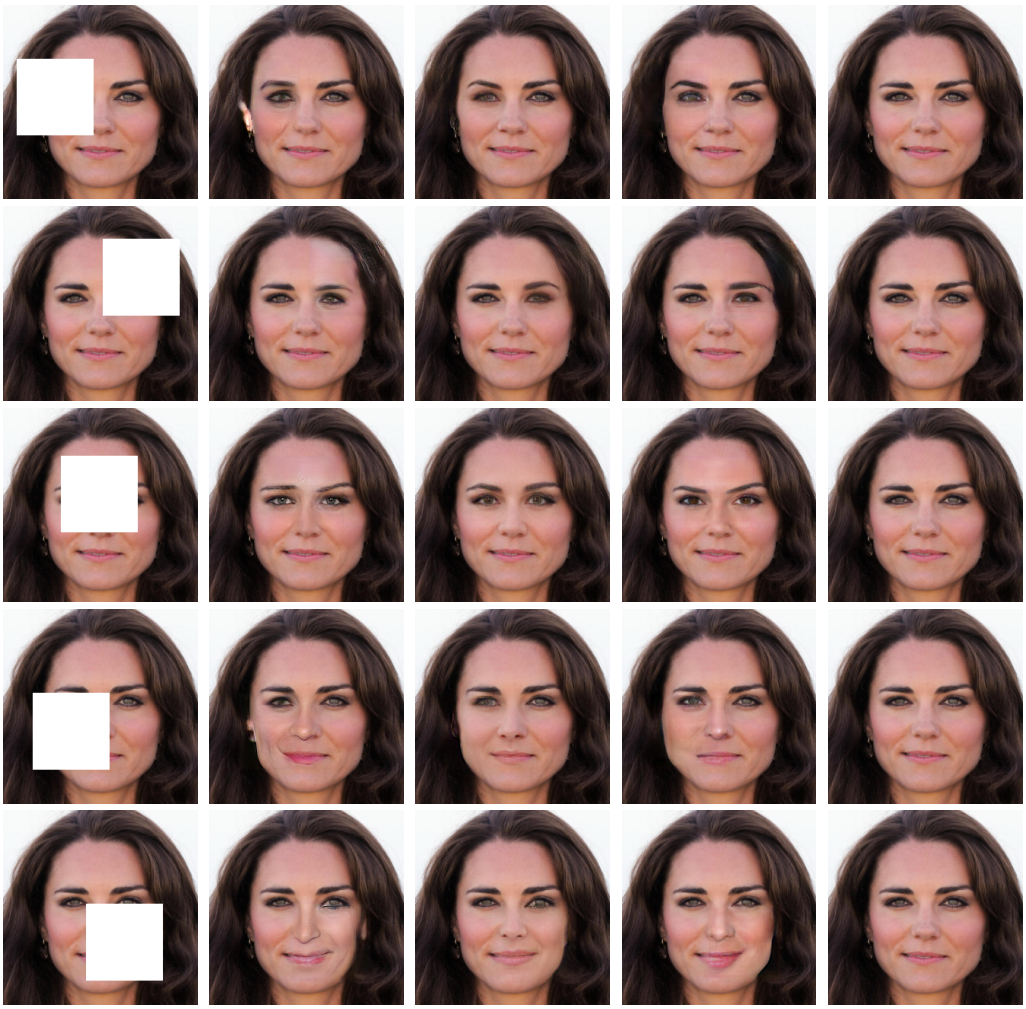
B. Experimental Results

1) Quantitative Evaluation: We perform quantitative evaluations on the CelebA-HQ and DTD texture test set using three different performance metrics, including the L1 loss, the peak signal-to-noise ratio (PSNR) and the structural similarity index (SSIM) [44]. The L1 loss can roughly measure the ability of models to restore the raw image pixels. The PSNR measures the difference in pixel values between the generated image and the original image. The SSIM estimates the holistic similarity between two images. We compared with two state-of-the-art methods: CA [13] and GMCNN [16]. Our skeleton network (i.e., S-Net, also see Fig. 1), full network (i.e., GEPE-Net, see Fig. 4) and the PE-Net that removes the iterative guidance loss (IGL) from the GEPE-Net are tested. All test images are masked with a 128×128 blank square which has been randomly placed.

The quantitative evaluation results are listed in Table III. For the celebA-HQ test set, our GEPE-Net produces better results than the CA model [13] and is comparable to GMCNN [16]. However, the performance of our network declines when we removed the iterative guidance loss. This suggests that the proposed IGL plays a positive guiding role. Since the GMCNN model [16] doesn't provide results on the DTD training set, we only compared with the CA [13] model with our GEPE-Net on this data set. It can be seen that the full model outperforms its counterparts in terms of all three metrics.

2) Qualitative Evaluation: Qualitative comparisons were conducted on the CelebA-HQ, DTD and canyon scene data sets. As shown in Fig. 6, both CA [13] and GMCNN [16] produce blurred artifacts or distorted structures in the masked region, especially in the eye area. The results generated by these methods are lack of consistency with the surrounding areas. In contrast, PE-Net generates more ambiguous results when the proposed IGL has been removed from the GEPE-Net. The S-Net (which have been trained using the same parameters as those used for training the GEPE-Net) produced blurred, almost invisible results in the filling area. With the help of guidance-enhanced perceptual encoder and ASCP mechanisms, the proposed GEPE-Net can generate fine local details and reasonably semantic structures. We also explore the effect of different mask regions on inpainting results of an image. The 100×100 mask is positioned on the left eye, right eye, two eyes, left side and right side of a face image. As shown in Fig. 7, our GEPE-Net achieves better inpainting results than those of CA [13] and comparable performance than GMCNN [16] model in five different mask regions. Meanwhile, as shown in Fig. 5, GEPE-Net is also superior to the CA [13] and GMCNN [16] methods for repairing the side face images.

Qualitative evaluation results for the canyon and texture data



(a) Input (b) CA (c) GMCNN (d) GEPE-Net (e) GT

Fig. 7. Qualitative Evaluations of different mask regions between our method and other generative methods on CelebA-HQ datasets. From the left to the right: (a)Masked input (b) Results of CA [13] (c) Results of GMCNN [16] (d) Results of our full GEPE-Net (e) Ground-truth.

sets are shown in Fig. 8 and Fig. 9 respectively, our GEPE-Net shows competitive inpainting results with better details and consistent structures with context. As shown in the third row of Fig. 8, our model produces the global consistency result, while CA [13] model produces color artifacts in the masked region.

V. Ablation Studies

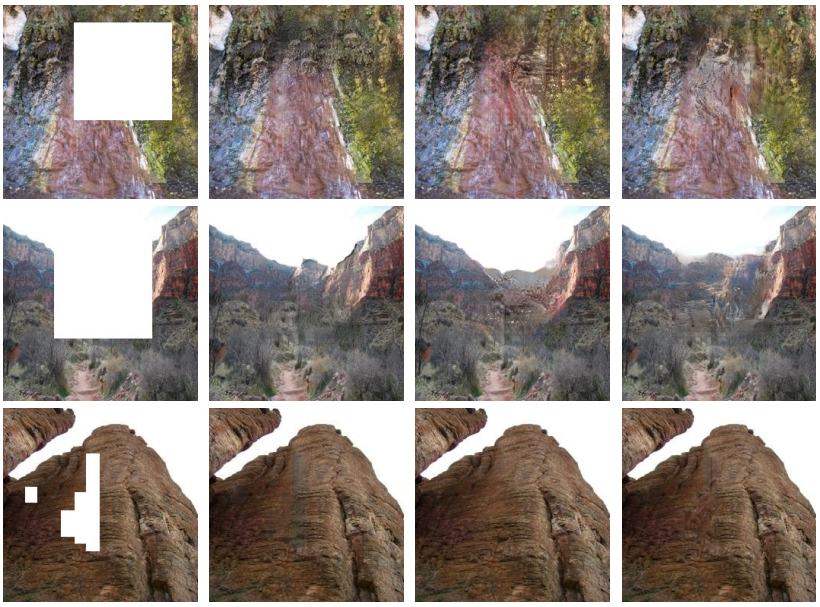
In this section, we report the results obtained by tuning the parameters of the proposed network.

A. Effect of the iterative guidance loss

The proposed iterative guidance loss is iteratively performed in the multi-layer perception operation to minimize the distance between the perceptual encoding feature and the ground-truth encoding feature. Thus, an appropriate trade-off parameter γ_{IGL} should be considered. We follow the similar rule presented in [45] to select our weighting parameters. We test three different values: 0.8, 0.08 and 0.008 for the parameter. For each value, the experiment is run 10 epochs on the DTD data set. It can be seen from Fig. 10 that the fuzzy texture is easily generated in the masked area when the value of γ_{IGL} is large (e.g., =0.8). When the γ_{IGL} value becomes small (e.g., =0.008), the global structure is not as good as that produced using the value of 0.08. Therefore, we empirically set the value of γ_{IGL} to 0.08 in our experiments.

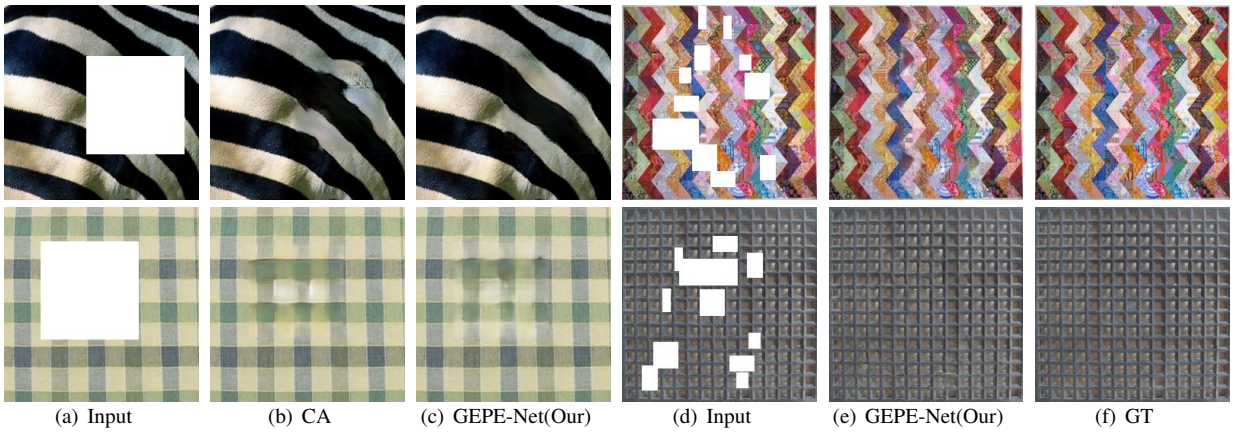
B. Effect of the atrous separable pyramid-convolution

We use the atrous separable pyramid-convolution to extract multi-scale features from the last guidance-enhanced encoding feature. To verify the effectiveness of the iterative guidance loss (i.e., IGL) and atrous separable pyramid-convolutions (i.e., ASPC). We test the GEPE-Net on the DTD data set and visualize the IGL-enhanced (i.e., enhanced by IGL operation) encoding feature maps and the fused feature map through ASPC. In the first row of Fig. 11, encoding feature maps become clearer benefited from the effect of the iterative



(a) Input (b) CA (c) GMCNN (d) GEPE-Net(Our)

Fig. 8. Qualitative Evaluations on Places2. From the left to the right: (a) Masked input (b) Results of CA [13] (c) Results of GMCNN [16] (d) Results of our full GEPE-Net. [Best viewed with zoom-in]



(a) Input (b) CA (c) GEPE-Net(Our) (d) Input (e) GEPE-Net(Our) (f) GT

Fig. 9. Qualitative Evaluations on DTD datasets. From the left to the right: (a) Masked input (b) Results of CA [13] (c) Results of our GEPE-Net (d) Multi-masked input (e) Results of our full GEPE-Net (f) Ground-truth image. [Best viewed with zoom-in]

guidance loss. After the feature map has been processed by the ASPC, it becomes clearer in the masked region.

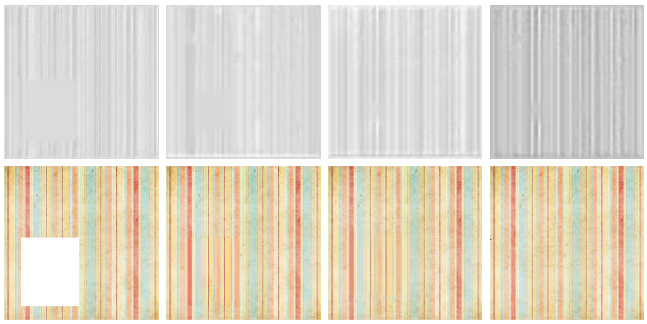
We also conduct a comparative experiment using our full GEPE-Net and the GEPE-Net without the ASPC. As shown in the second row of Fig. 11, the model without the ASPC produces an abrupt or non-matching color result in the masked area than that derived using the full model, while the full GEPE-Net with the ASPC can generate a visually promising result with consistent color with the whole image.

C. Comparison of different reconstruction losses

We compare the differences between the reconstruction loss of the spatial-variant weights that we use and the L1 loss for our full GEPE-Net model. The experiment is performed on the CelebA-HQ data set. The results are shown in Fig. 12 suggest that the L1 loss tends to produce blurry results in the masked region, while the reconstruction loss that we use can produce overall consistency results.

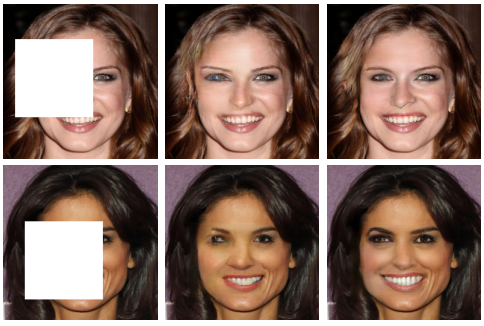
(a) $\gamma_{IGL}=0.8$ (b) $\gamma_{IGL}=0.08$ (c) $\gamma_{IGL}=0.008$

Fig. 10. Results of different tradeoff parameters γ_{IGL} of iterative guidance loss.



(a) (b) (c) (d)

Fig. 11. In the first row, from the left to the right: (a) IGL-enhanced feature map from the third encoder layer (b) IGL-enhanced feature map from the fourth encoder layer (c) IGL-enhanced feature map from the fifth encoder layer (d) Fused feature map through ASPC. In the second row, from the left to the right: (a) Masked input (b) The result of GEPE-Net without ASPC (c) The result of GEPE-Net with ASPC (d) Raw image. [Best viewed with zoom-in in the first row, Best viewed in color difference of the second row.]



(a) (b) (c)

Fig. 12. Results of different reconstruction losses. (a) Masked input (b) Results of L1 loss (c)Results of the spatial-variant weight loss(Our).

VI. Conclusion and Future Work

In this paper, we presented a guidance-enhanced perceptual encoder network, in which iterative guidance loss progressively guides perception operation in an accurate encoding direction, and atrous separable pyramid-convolutions help extract multi-scale features. In addition, the enhanced perceptual encoding features are transferred to the decoder through the skip connection. This also boosts the decoding effectiveness of our GEPE-Net. Experimental results showed that our GEPE-Net generates visually-pleasing and semantically-plausible results in multiple data sets.

In future work, an end-to-end semantic segmentation network can be investigated in order to further improve the restoration results on complex texture images.

Acknowledgment

This work was supported in part by the National Natural Science Foundation of China under Grant 41576011, L1824025, 61501417 and 61976123.

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*. ACM Press/Addison-Wesley Publishing Co., 2000, pp. 417–424.

[2] 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*. ACM, 2001, pp. 341–346.

[3] H. Noori, S. Saryazdi, and H. Nezamabadi-Pour, "A convolution based image inpainting," in *1st International conference on Communications Engineering*, 2010, pp. 130–134.

[4] C. Barnes, E. Shechtman, A. Finkelstein, and D. B. Goldman, "Patchmatch: A randomized correspondence algorithm for structural image editing," in *ACM Transactions on Graphics (ToG)*, vol. 28, no. 3. ACM, 2009, p. 24.

[5] 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.

[6] A. Newson, A. Almansa, M. Fradet, Y. Gousseau, and P. Pérez, "Video inpainting of complex scenes," *SIAM Journal on Imaging Sciences*, vol. 7, no. 4, pp. 1993–2019, 2014.

[7] A. Criminisi, P. Pérez, and K. Toyama, "Region filling and object removal by exemplar-based image inpainting," *IEEE Transactions on image processing*, vol. 13, no. 9, pp. 1200–1212, 2004.

[8] J. Dong, S. Ma, L. Li, and Z. Yu, "Hole filling on three-dimensional surface texture," in *2007 IEEE International Conference on Multimedia and Expo*. IEEE, 2007, pp. 1299–1302.

[9] X. Dong, J. Dong, G. Sun, Y. Duan, L. Qi, and H. Yu, "Learning-based texture synthesis and automatic inpainting using support vector machines," *IEEE Transactions on Industrial Electronics*, vol. 66, no. 6, pp. 4777–4787, June 2019.

[10] J. Dong, L. Wang, J. Liu, Y. Gao, L. Qi, and X. Sun, "A procedural texture generation framework based on semantic descriptions," *Knowledge-Based Systems*, vol. 163, pp. 898–906, 2019.

[11] V. Kwatra, I. Essa, A. Bobick, and N. Kwatra, "Texture optimization for example-based synthesis," in *ACM Transactions on Graphics (ToG)*, vol. 24, no. 3. ACM, 2005, pp. 795–802.

[12] 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.

[13] 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.

[14] G. E. Hinton and R. R. Salakhutdinov, "Reducing the dimensionality of data with neural networks," *science*, vol. 313, no. 5786, pp. 504–507, 2006.

[15] Y. Li, S. Liu, J. Yang, and M.-H. Yang, "Generative face completion," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 3911–3919.

[16] 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.

[17] C. Ballester, M. Bertalmio, V. Caselles, G. Sapiro, and J. Verdera, "Filling-in by joint interpolation of vector fields and gray levels," 2000.

[18] M. Jian, C. Cui, X. Nie, H. Zhang, L. Nie, and Y. Yin, "Multi-view face hallucination using svd and a mapping model," *Information Sciences*, vol. 488, pp. 181–189, 2019.

[19] N. Zeng, Z. Wang, H. Zhang, K.-E. Kim, Y. Li, and X. Liu, "An improved particle filter with a novel hybrid proposal distribution for quantitative analysis of gold immunochromatographic strips," *IEEE Transactions on Nanotechnology*, vol. 18, pp. 819–829, 2019.

[20] N. Zeng, Z. Wang, H. Zhang, W. Liu, and F. E. Alsaadi, "Deep belief networks for quantitative analysis of a gold immunochromatographic strip," *Cognitive Computation*, vol. 8, no. 4, pp. 684–692, 2016.

[21] N. Zeng, H. Zhang, B. Song, W. Liu, Y. Li, and A. M. Dobaie, "Facial expression recognition via learning deep sparse autoencoders," *Neurocomputing*, vol. 273, pp. 643–649, 2018.

[22] S. Iizuka, E. Simo-Serra, and H. Ishikawa, "Globally and locally consistent image completion," *ACM Transactions on Graphics (ToG)*, vol. 36, no. 4, p. 107, 2017.

[23] 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.

TABLE IV
THE DETAIL ARCHITECTURE OF THE GUIDANCE-ENHANCED PERCEPTUAL ENCODER. CONCATENATION IN THE COLUMN OF TYPE DENOTES ASPC FEATURE CONCATENATION. CONV3_PERCEPTUAL REFERS TO THE PERCEPTION OPERATION ENHANCED BY IGL.

Type	Kernel	Dilation	Stride	Outputs	Activation
Conv1	5×5	1	1×1	32	ELUs
Conv2	3×3	1	2×2	64	ELUs
Conv3	3×3	1	1×1	64	ELUs
Conv3_perceptual					
Conv4	3×3	1	2×2	128	ELUs
Conv4_perceptual					
Conv5	3×3	1	1×1	128	ELUs
Conv5_perceptual					
Conv6_1×1	1×1	1	1×1	128	ELUs
Conv7_atrous_sep	3×3	2	1×1	128	ELUs
Conv8_atrous_sep	3×3	4	1×1	128	ELUs
Conv9_atrous_sep	3×3	8	1×1	128	ELUs
Conv10_atrous_sep	3×3	16	1×1	128	ELUs
Concatenation					

TABLE V
THE DETAIL ARCHITECTURE OF THE DECODER. ⊙ DENOTES FEATURE CONCATENATION.

Type	Kernel	Dilation	Stride	Outputs	Activation	Skips
Conv11	3×3	1	1×1	128	ELUs	⊙ Conv5_perceptual
Conv12	3×3	1	1×1	128	ELUs	⊙ Conv4_perceptual
Deconv13 (Nearest Neighbor ×2↑)	3×3	1	1×1	64		⊙ Conv3_perceptual
Conv14	3×3	1	1×1	64	ELUs	⊙ Conv2
Deconv15 (Nearest Neighbor ×2↑)	3×3	1	1×1	32		⊙ Conv1
Conv16	3×3	1	1×1	16	ELUs	
Conv17	3×3	1	1×1	3		

[24] L. Song, J. Cao, L. Song, Y. Hu, and R. He, "Geometry-aware face completion and editing," *arXiv preprint arXiv:1809.02967*, 2018.

[25] 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 *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 5485–5493.

[26] M. Jian and K.-M. Lam, "Face-image retrieval based on singular values and potential-field representation," *Signal processing*, vol. 100, pp. 9–15, 2014.

[27] J. Fu, J. Liu, Y. Wang, J. Zhou, C. Wang, and H. Lu, "Stacked deconvolutional network for semantic segmentation," *IEEE Transactions on Image Processing*, 2019.

[28] L.-C. Chen, Y. Zhu, G. Papandreou, F. Schroff, and H. Adam, "Encoder-decoder with atrous separable convolution for semantic image segmentation," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018, pp. 801–818.

[29] V. Badrinarayanan, A. Kendall, and R. Cipolla, "Segnet: A deep convolutional encoder-decoder architecture for image segmentation," *IEEE transactions on pattern analysis and machine intelligence*, vol. 39, no. 12, pp. 2481–2495, 2017.

[30] C. Yang, X. Lu, Z. Lin, E. Shechtman, O. Wang, and H. Li, "High-resolution image inpainting using multi-scale neural patch synthesis," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 6721–6729.

[31] P. Isola, J.-Y. Zhu, T. Zhou, and A. A. Efros, "Image-to-image translation with conditional adversarial networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 1125–1134.

[32] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in *International Conference on Medical image computing and computer-assisted intervention*. Springer, 2015, pp. 234–241.

[33] 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 transactions on pattern analysis and machine intelligence*, vol. 40, no. 4, pp. 834–848, 2017.

[34] L.-C. Chen, G. Papandreou, F. Schroff, and H. Adam, "Rethinking atrous convolution for semantic image segmentation," *arXiv preprint arXiv:1706.05587*, 2017.

[35] F. Yu and V. Koltun, "Multi-scale context aggregation by dilated convolutions," *arXiv preprint arXiv:1511.07122*, 2015.

[36] F. Chollet, "Xception: Deep learning with depthwise separable convolutions," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 1251–1258.

[37] M. Wang, B. Liu, and H. Foroosh, "Design of efficient convolutional layers using single intra-channel convolution, topological subdivisioning

and spatial" bottleneck" structure," *arXiv preprint arXiv:1608.04337*, 2016.

[38] A. Radford, L. Metz, and S. Chintala, "Unsupervised representation learning with deep convolutional generative adversarial networks," *arXiv preprint arXiv:1511.06434*, 2015.

[39] 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.

[40] M. Arjovsky, S. Chintala, and L. Bottou, "Wasserstein gan," *arXiv preprint arXiv:1701.07875*, 2017.

[41] M. Cimpoi, S. Maji, I. Kokkinos, S. Mohamed, and A. Vedaldi, "Describing textures in the wild," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2014, pp. 3606–3613.

[42] 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.

[43] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.

[44] Z. Wang, A. C. Bovik, H. R. Sheikh, E. P. Simoncelli *et al.*, "Image quality assessment: from error visibility to structural similarity," *IEEE transactions on image processing*, vol. 13, no. 4, pp. 600–612, 2004.

[45] Z. Yan, X. Li, M. Li, W. Zuo, and S. Shan, "Shift-net: Image inpainting via deep feature rearrangement," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018, pp. 1–17.