

PEPSI++: Fast and Lightweight Network for Image Inpainting

Yong-Goo Shin[✉], Min-Cheol Sagong, Yoon-Jae Yeo[✉], Seung-Wook Kim, and Sung-Jea Ko[✉], *Fellow, IEEE*

Abstract—Among the various generative adversarial network (GAN)-based image inpainting methods, a coarse-to-fine network with a contextual attention module (CAM) has shown remarkable performance. However, due to two stacked generative networks, the coarse-to-fine network needs numerous computational resources, such as convolution operations and network parameters, which result in low speed. To address this problem, we propose a novel network architecture called parallel extended-decoder path for semantic inpainting (PEPSI) network, which aims at reducing the hardware costs and improving the inpainting performance. PEPSI consists of a single shared encoding network and parallel decoding networks called coarse and inpainting paths. The coarse path produces a preliminary inpainting result to train the encoding network for the prediction of features for the CAM. Simultaneously, the inpainting path generates higher inpainting quality using the refined features reconstructed via the CAM. In addition, we propose Diet-PEPSI that significantly reduces the network parameters while maintaining the performance. In Diet-PEPSI, to capture the global contextual information with low hardware costs, we propose novel rate-adaptive dilated convolutional layers that employ the common weights but produce dynamic features depending on the given dilation rates. Extensive experiments comparing the performance with state-of-the-art image inpainting methods demonstrate that both PEPSI and Diet-PEPSI improve the qualitative scores, i.e., the peak signal-to-noise ratio (PSNR) and structural similarity (SSIM), as well as significantly reduce hardware costs, such as computational time and the number of network parameters.

Index Terms—Deep learning, generative adversarial network (GAN), image inpainting.

I. INTRODUCTION

IMAGE inpainting techniques that attempt to remove an unwanted object or synthesize missing parts of an image have attracted widespread interest in computer vision and graphics communities [1]–[17]. Recent studies used the generative adversarial network (GAN) to produce appropriate structures for the missing regions, i.e., hole regions [8], [9], [18]. Among the recent state-of-the-art inpainting methods,

Manuscript received May 21, 2019; revised October 8, 2019 and December 11, 2019; accepted March 2, 2020. Date of publication March 19, 2020; date of current version January 5, 2021. This work was supported by the Institute for Information and communications Technology Promotion (IITP) grant funded by the Korea Government under Grant 2019-0-00268, Development of SW technology for recognition, judgment and path control algorithm verification simulation and dataset generation. (*Corresponding author: Sung-Jea Ko*)

The authors are with the School of Electrical Engineering Department, Korea University, Seoul 136-713, South Korea (e-mail: ygshin@dali.korea.ac.kr; mcsagong@dali.korea.ac.kr; yjyeo@dali.korea.ac.kr; swkim@dali.korea.ac.kr; sjko@korea.ac.kr).

Color versions of one or more of the figures in this article are available online at <https://ieeexplore.ieee.org>.

Digital Object Identifier 10.1109/TNNLS.2020.2978501

the coarse-to-fine network with a contextual attention module (CAM) has shown remarkable performance [4], [19]. This network is composed of two stacked generative networks, including the coarse network and refinement one. The coarse network roughly fills the hole regions using a simple dilated convolutional network trained with reconstruction loss. The refinement network improves the quality of the roughly completed image by using the CAM that generates feature patches of the hole regions by borrowing information from distant spatial locations. Despite the promising results, the coarse-to-fine network requires high computational resources and consumes considerable memories.

In the previous work [20], we introduced a novel network structure called parallel extended-decoder path for semantic inpainting (PEPSI), which aims at reducing the number of convolution operations as well as improving the inpainting performance. PEPSI is composed of a single encoding network and parallel decoding networks that consist of coarse and inpainting paths. The coarse path generates a preliminary inpainting result to train the encoding network for the prediction of features for the CAM. At the same time, the inpainting path produces an image with high quality using the refined features reconstructed via the CAM. To make a single encoding network handle two different tasks, which are feature extraction for both roughly completed and high-quality results, we propose a joint learning technique that jointly optimizes two different paths. This learning scheme facilitates the generation of high-quality inpainting images without the stacked generative networks, i.e., the coarse-to-fine network.

Although PEPSI exhibits faster operation speed compared with conventional methods, it still needs substantial memory owing to a series of dilated convolutional layers in the encoding network, which occupies nearly 67% of network parameters. The intuitive way to save memory consumption is to prune channels in the dilated convolutional layers; however, it often results in inferior results. To address this challenge, this article presents an extended version of PEPSI, called Diet-PEPSI, which significantly reduces the network parameters while retaining the inpainting performance. In Diet-PEPSI, we propose novel rate-adaptive dilated convolutional layers that require low hardware costs by sharing the weights in every layer but generate dynamic features according to the given dilation rates. More specifically, to produce the rate-specific features, the rate-adaptive dilated convolutional layers modulate the shared weights by differently scaling and shifting according to the given dilation rates. Since the rate-adaptive dilated convolutional layers share the weights with each other, the number of network parameters can be significantly reduced

compared with multiple standard dilated convolutional layers. In this article, we apply the proposed rate-adaptive dilated convolutional layers to Diet-PEPSI using residual blocks [21] called Diet-PEPSI units (DPUs). By replacing the multiple dilated convolutional layers with DPUs, Diet-PEPSI covers the same size of the receptive field with a smaller number of parameters than PEPSI.

Furthermore, we investigate an obstacle with the discriminator in traditional GAN-based image inpainting methods [14], [22]. In general, conventional methods employ global and local discriminators trained with a combined loss, i.e., the L_2 pixelwise reconstruction loss and adversarial loss, which assists the networks in generating a more natural image by minimizing the difference between the reference and the inpainted images. More specifically, the global discriminator takes the whole image as input to recognize global consistency, whereas the local one only views a small region around the hole in order to judge the quality of a more detailed appearance. However, the local discriminator has a drawback that it can only deal with a single rectangular hole region. In other words, since the holes can appear with arbitrary shapes, sizes, and locations in real-world applications, the local discriminator is difficult to apply to the inpainting network for inpainting the holes with irregular shapes. To solve this problem, we propose a region ensemble discriminator (RED) that integrates the global and local discriminators. Since each pixel in the last layer has a different receptive field in the image domain, the RED adopts individual fully connected layers on each pixel in the last convolutional layer. By individually computing an adversarial loss in each pixel, the RED can deal with the various holes with arbitrary shapes.

In summary, this article has three major contributions: 1) we propose a novel network architecture called PEPSI that achieves superior performance compared with conventional methods as well as significantly reduces the operation time; 2) we propose Diet-PEPSI that applies novel rate-adaptive convolution layers to further reduce the hardware costs while maintaining the overall quality of the results, which makes the proposed method compatible with the hardware; and 3) a novel discriminator, called RED, is proposed to handle both squared and irregular hole regions for real applications. In the remainder of this article, we introduce the related work and preliminaries in Sections II and III, respectively. The PEPSI and Diet-PEPSI are discussed in Section IV. In Section V, extensive experimental results are presented to demonstrate that the proposed method outperforms conventional methods on various data sets, such as CelebA [23], [24], Place2 [25], and ImageNet [26]. Finally, the conclusion is provided in Section VI.

II. RELATED WORK

Existing image inpainting techniques can be divided into two groups [4]: traditional and deep learning-based methods. The traditional techniques include diffusion- and patch-based methods. The diffusion-based method fills the hole regions by propagating the local image appearance around the

holes [1], [2], [4], [5]. The diffusion-based method performs well on the small and narrow holes but often fails to fill complex hole regions, such as faces and objects, with non-repetitive structures. In contrast, the patch-based technique results in better performance in filling the complicated images with large hole regions [4], [27]. This method samples texture patches from the existing regions of an image, i.e., background regions, and pastes them into the hole region. Barnes *et al.* [3] introduced a fast approximate nearest neighbor patch search algorithm, called PatchMatch, which exhibited notable performance for image editing applications, such as the image inpainting. However, PatchMatch often fills the hole regions regardless of the visual semantics or the global structure of an image, which results in the resultant images with poor visual quality.

By using the convolutional neural network (CNN), the deep learning-based method learns how to extract semantic information for producing the structures of the hole regions [8], [9], [18]. The CNN-based image inpainting methods employing an encoder-decoder structure have shown superior performance on inpainting the complex hole region compared with the diffusion- or patch-based methods [8], [18]. However, these methods often generate an image with visual artifacts, such as boundary artifacts and blurry texture inconsistent with surrounding areas. To alleviate this problem, Pathak *et al.* [10] adopted the GAN [22] to enhance the coherence between the background and hole regions. They trained the entire network using a combined loss, i.e., the L_2 pixelwise reconstruction loss and adversarial loss, which drives the networks to minimize the difference between the reference and inpainted images as well as to produce plausible new contents in highly structured images, such as faces and scenes. However, this method has a limitation that it only can fill the square hole located at the center of an image.

To inpaint the images with a square hole in arbitrary locations, as shown in Fig. 1(a), Iizuka *et al.* [7] proposed an improved network structure that employs two sibling discriminators: global and local discriminators. More specifically, the local discriminator only considers the inpainted region to classify the local texture consistency, whereas the global discriminator inspects that the resultant image is consistent across the whole image. Recently, Yu *et al.* [4] have extended this work by using the coarse-to-fine network and the CAM. In particular, by computing the cosine similarity between the background and foreground feature patches, the CAM learns where to borrow the background features for the hole region. In order to collect the background features involved with the missing region, the CAM requires the features at the missing region encoded from roughly completed images. Thus, as shown in Fig. 1(b), this method employs two stacked generative networks (coarse and refinement networks) to generate an intermediate result, i.e. the coarse result, and an inpainting result refined through the refinement network having the CAM. This method achieved remarkable performance compared with the recent state-of-the-art inpainting methods; however, it requires considerable computational resources due to the two stacked generative networks.

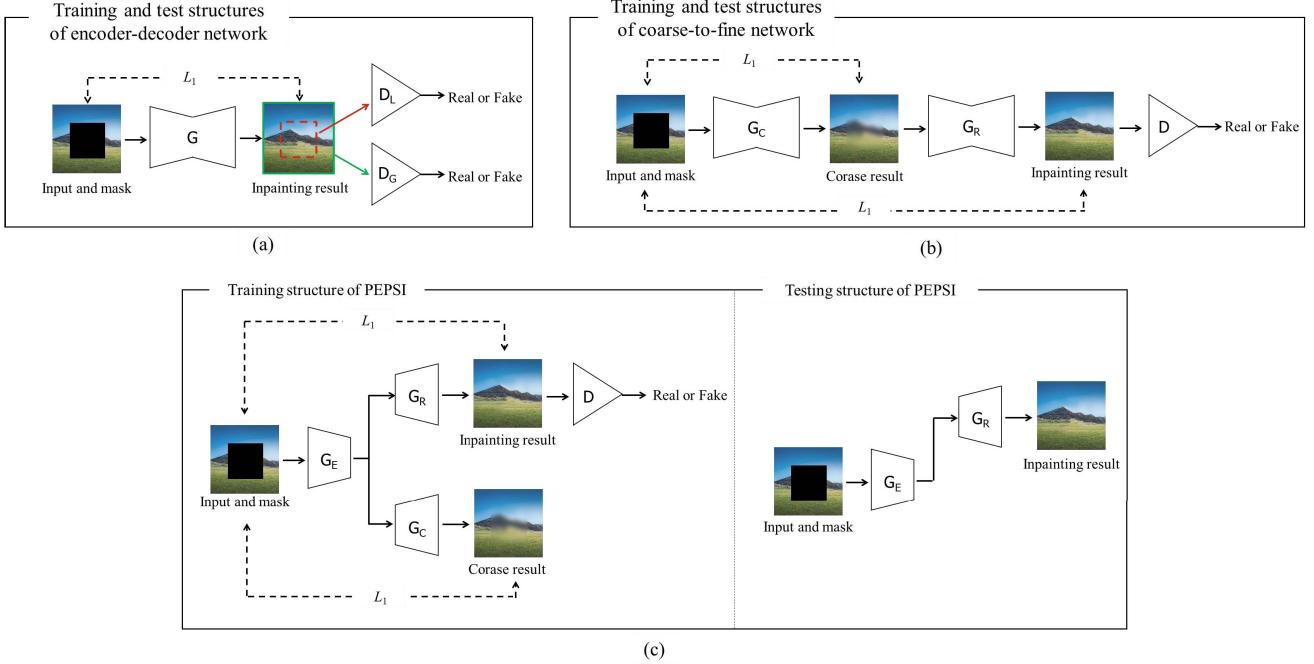


Fig. 1. Overview of the network architectures of the conventional and proposed methods, where D and G indicate the discriminator and generator, respectively. (a) Architecture of traditional encoder-decoder network [7]. (b) Architecture of coarse-to-fine network [4], [19]. (c) Architecture of PEPSI.

III. PRELIMINARIES

A. Generative Adversarial Networks

The GAN was first introduced by Goodfellow *et al.* [22] for image generation. In general, GAN consists of a generator G and a discriminator D that are trained with competing goals. The generator is trained to produce a new image, indistinguishable from real images, while the discriminator is optimized to differentiate between real and generated images. Formally, the G (D) tries to minimize (maximize) the loss function, i.e., adversarial loss, as follows:

$$\min_G \max_D E_{x \sim P_{\text{data}}(x)}[\log D(x)] + E_{z \sim P_z(z)}[\log(1 - D(G(z)))] \quad (1)$$

where z and x denote a random noise vector and a real image sampled from the noise $P_z(z)$ and real data distribution $P_{\text{data}}(x)$, respectively. Recently, the GAN has been applied to several semantic inpainting techniques [4], [7], [10] to fill the holes naturally.

B. Coarse-to-Fine Network

In [4] and [19], a two-stage network called a coarse-to-fine network, which separately conducts a couple of tasks, is proposed. The coarse-to-fine network first generates an initial coarse prediction, i.e., coarse result, using the coarse network, and then refines the results by encoding features from the coarse result with the refinement network. To refine the coarse prediction effectively, they introduced CAM that generates patches of the hole region using features from distant background patches. As depicted in Fig. 2, the CAM divides the input feature maps into a target foreground and

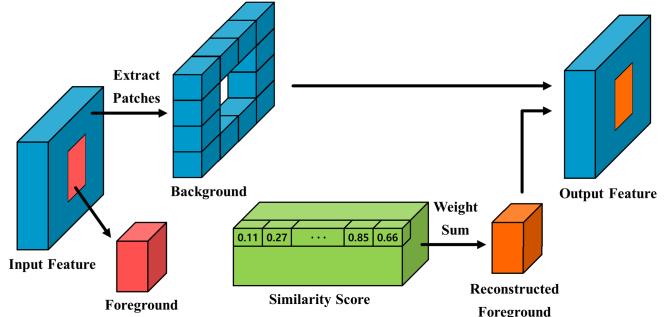


Fig. 2. Illustration of the CAM. The conventional CAM reconstructs foreground patches by measuring the cosine similarities with background patches. In contrast, the modified CAM uses the Euclidean distance to compute similarity scores.

its surrounding background and extracts 3×3 patches. Then, the similarity score $s_{(x,y),(x',y')}$ between the foreground patch at (x, y) , $f_{x,y}$ and the background patch at (x', y') , $b_{x',y'}$ is computed by using the normalized inner product (cosine similarity), which is expressed as follows:

$$s_{(x,y),(x',y')} = \left\langle \frac{f_{x,y}}{\|f_{x,y}\|}, \frac{b_{x',y'}}{\|b_{x',y'}\|} \right\rangle \quad (2)$$

$$s_{(x,y),(x',y')}^* = \text{softmax}(\lambda s_{(x,y),(x',y')}) \quad (3)$$

where λ is a hyperparameter for scaled softmax. By weighted sum of background patches using $s_{(x,y),(x',y')}^*$ as weights, the CAM rebuilds features of foreground regions, i.e., reconstructed feature. The CAM effectively learns where to borrow or copy the feature information from the background region for the unknown foreground regions, but it requires the coarse

TABLE I

EXPERIMENTAL RESULTS WITH GATEDCONV (GC) [19] USING DIFFERENT COARSE PATH. GC* INDICATES A MODEL TRAINED WITHOUT COARSE RESULTS, AND GC† INDICATES A MODEL TRAINED WITH SIMPLIFIED COARSE PATH

	Square mask		Free-form mask		Time
	PSNR	SSIM	PSNR	SSIM	
GC	24.67	0.8949	27.78	0.9252	21.39 ms
GC*	23.50	0.8822	26.35	0.9098	14.28 ms
GC†	23.71	0.8752	26.22	0.9026	13.32 ms

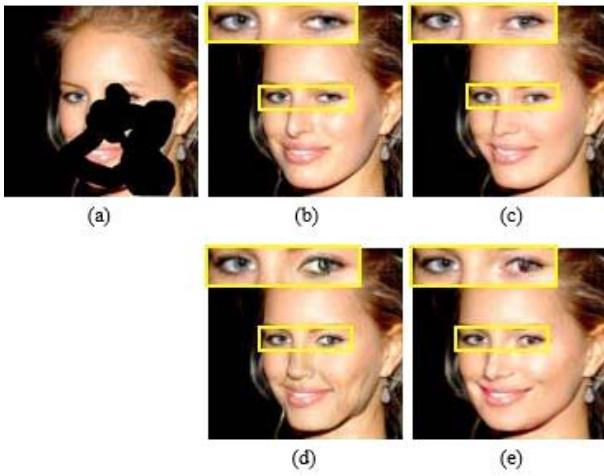


Fig. 3. Toy examples for the coarse network. (a) Masked input image. (b) Original image. (c) Result from the coarse-to-fine network. (d) Result without the coarse result. (e) Result with LR coarse path.

result to explicitly attend on related features at distant spatial locations.

To justify this assumption, we conducted experiments that measure the performance of the coarse-to-fine network with/without the coarse path. In our experiments, we trained the refinement network using raw masked images as an input. As shown in Table I and Fig. 3, the refinement network without the coarse result shows worse results than the full coarse-to-fine network. These results reveal that if the coarse feature of the hole region is not encoded well, the CAM reconstructs features using unrelated feature patches, resulting in inferior results. For instance, as shown in Fig. 3(d), the refinement network trained without the coarse result produces artifacts, such as a wrinkle on the cheek. In other words, the coarse-to-fine network must pass through a two-stage encoder-decoder network that requires massive computational resources. Furthermore, to reduce the operation time of the coarse-to-fine network with another way, we conducted an extra experiment by simplifying the coarse network. In our experiments, we generated the coarse result with low resolution (64×64) and fed it to the refinement network by resizing its resolution to the original size. However, as shown in Fig. 3(e) and Table I, the simplified coarse network exhibits worse performance. For instance, as shown in Fig. 3(e), the simplified coarse network results in asymmetric eyes with defects; the generated right eye has different colors as compared with the left one. These observations indicate that the simplified coarse network can

TABLE II
DETAILED ARCHITECTURE OF ENCODING NETWORK

Type	Kernel	Dilation	Stride	Outputs
Convolution	5×5	1	1×1	32
Convolution	3×3	1	2×2	64
Convolution	3×3	1	1×1	64
Convolution	3×3	1	2×2	128
Convolution	3×3	1	1×1	128
Convolution	3×3	1	2×2	256
Dilated convolution	3×3	2	1×1	256
Dilated convolution	3×3	4	1×1	256
Dilated convolution	3×3	8	1×1	256
Dilated convolution	3×3	1	1×1	256

TABLE III
DETAILED ARCHITECTURE OF THE DECODING NETWORK. THE OUTPUT LAYER CONSISTS OF A CONVOLUTION LAYER CLIPPED VALUE TO THE $[-1, 1]$

Type	Kernel	Dilation	Stride	Outputs
Convolution $\times 2$	3×3	1	1×1	128
Upsample ($\times 2 \uparrow$)	-	-	-	-
Convolution $\times 2$	3×3	1	1×1	64
Upsample ($\times 2 \uparrow$)	-	-	-	-
Convolution $\times 2$	3×3	1	1×1	32
Upsample ($\times 2 \uparrow$)	-	-	-	-
Convolution $\times 2$	3×3	1	1×1	16
Convolution (Output)	3×3	1	1×1	3

produce the roughly completed image with fast speed, but this completed image is not suitable for the refinement network.

IV. PROPOSED METHOD

A. Architecture of PEPSI

As shown in Fig. 4, PEPSI unifies the stacked networks of the coarse-to-fine network into a single generative network with a single shared encoding network and a parallel decoding network called the coarse and inpainting paths, respectively. The encoding network aims at jointly learning to extract the features from background regions as well as to complete the features of hole regions without the coarse results. As listed in Table II, the encoding network consists of a series of 3×3 convolutional layers, except for the first layer that uses a 5×5 convolutional layer. To enlarge the receptive field of the encoding network, we utilize multiple dilated convolutional layers with different dilation rates in the last four convolutional layers.

A parallel decoding network consists of coarse and inpainting paths that share the weight parameters with each other. A detailed architecture of the decoding network is described in Table III. The coarse path produces a roughly completed result from the feature maps obtained via the encoding network, whereas the inpainting path first reconstructs the encoded feature map by using the CAM and produces a higher-quality inpainting result by decoding the reconstructed features. Since two different paths use the same encoded feature maps as their input, this joint learning strategy encourages

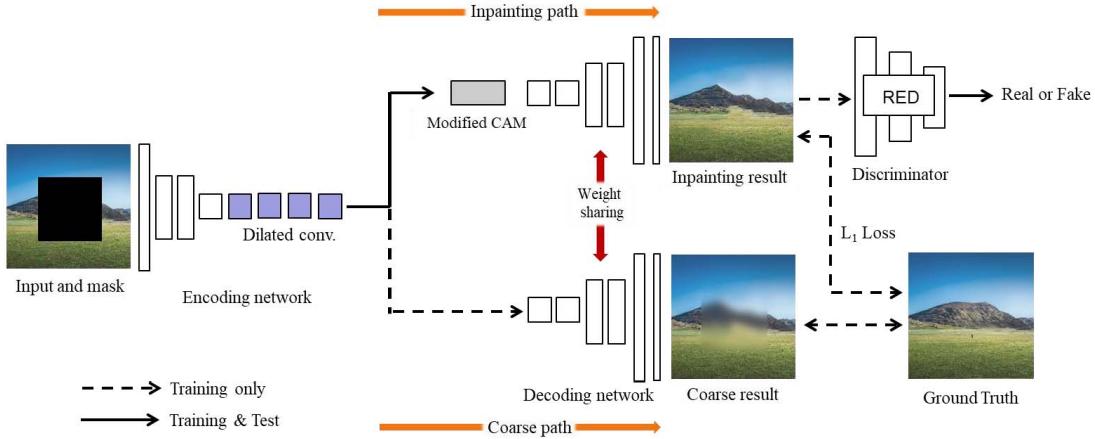


Fig. 4. Architecture of PEPSI. The coarse path and inpainting path share their weights to improve each other. The coarse path is trained only with the L_1 reconstruction loss, while the inpainting path is trained with both of L_1 and adversarial losses.

the encoding network to produce valuable features for two different image generation tasks. To jointly train both paths, we explicitly employ the reconstruction L_1 loss to the coarse path, whereas the inpainting path is trained by using both L_1 and the adversarial losses. Additional information about the joint learning scheme will be described in Section IV-E. It should be noted that we employ only the inpainting path during the tests, as shown in Fig. 1(c), which substantially reduces the computational complexity.

In terms of layer implementations in the encoding and decoding networks, PEPSI employs reflection padding for all convolutional layers and uses the exponential linear unit (ELU) [28] as an activation function, except for the last convolutional layer. In addition, $[-1, 1]$ normalized image with 256×256 pixels is employed as an input of PEPSI, and PEPSI produces the output image with the same resolution by clipping the output values into $[-1, 1]$ instead of using the tanh function.

B. Architecture of Diet-PEPSI

Although PEPSI effectively reduces the number of convolution operations, it still requires a similar number of network parameters as the coarse-to-fine network. As mentioned in Section IV-A, PEPSI aggregates the contextual information using a series of dilated convolutional layers, which requires numerous network parameters. The intuitive way to reduce hardware cost is to prune the channels of these layers, but it often yields inferior results in practice. To cope with this problem, we propose novel rate-adaptive dilated convolutional layers that utilize the shared weights but produce dynamic feature maps depending on the given dilation rates. More specifically, to produce rate-specific features, the rate-adaptive dilated convolutional layers alter the shared weights by scaling and shifting differently according to the given dilation rates. Since the rate-adaptive dilated convolutional layers share the weights in every layer, the number of network parameters can be significantly reduced compared with multiple standard dilated convolutional layers. In this section, we first introduce how the rate-adaptive dilated convolutional layers produce

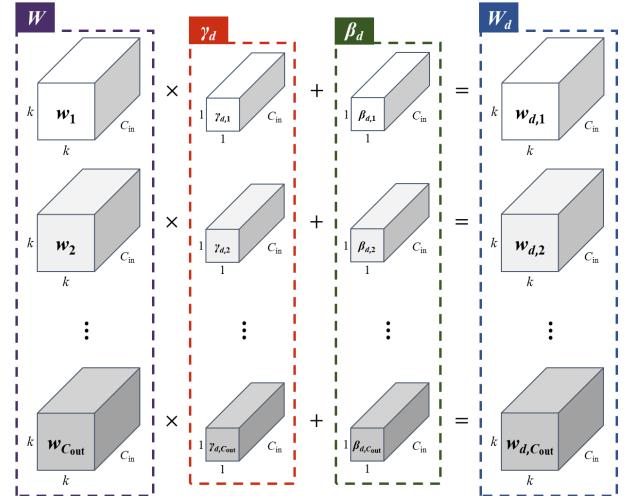


Fig. 5. Rate-adaptive scaling and shifting operations. β_d and γ_d have different values depending on the given rate. Tensor broadcasting is included in scaling and shifting operations.

different feature maps. Then, we explain how the rate-adaptive dilated convolutional layers are applied to PEPSI.

In general, the weights of the convolutional layer are considered as a 4-D tensor $W \in \mathbb{R}^{k \times k \times C_{in} \times C_{out}}$, where k is the kernel size, while C_{in} and C_{out} are the number of input and output channels, respectively. In other words, as shown in Fig. 5, the weights in each convolutional layer can be represented as C_{out} filters with C_{in} channels, i.e., w_i , $\{i = 1, \dots, C_{out}\}$. To produce different features according to the given dilation rates, we modulate W using the learned scale $\gamma_d \in \mathbb{R}^{1 \times 1 \times C_{in} \times C_{out}}$ and bias $\beta_d \in \mathbb{R}^{1 \times 1 \times C_{in} \times C_{out}}$ parameters, where d indicates the dilation rate and γ_d and β_d are learned separately depending on the given dilation rate. This modulating process can be expressed as follows:

$$W_d = \gamma_d \cdot W + \beta_d \quad (4)$$

where $W_d \in \mathbb{R}^{k \times k \times C_{in} \times C_{out}}$ represents the rate-adaptively modified weights. Note that tensor broadcasting is included

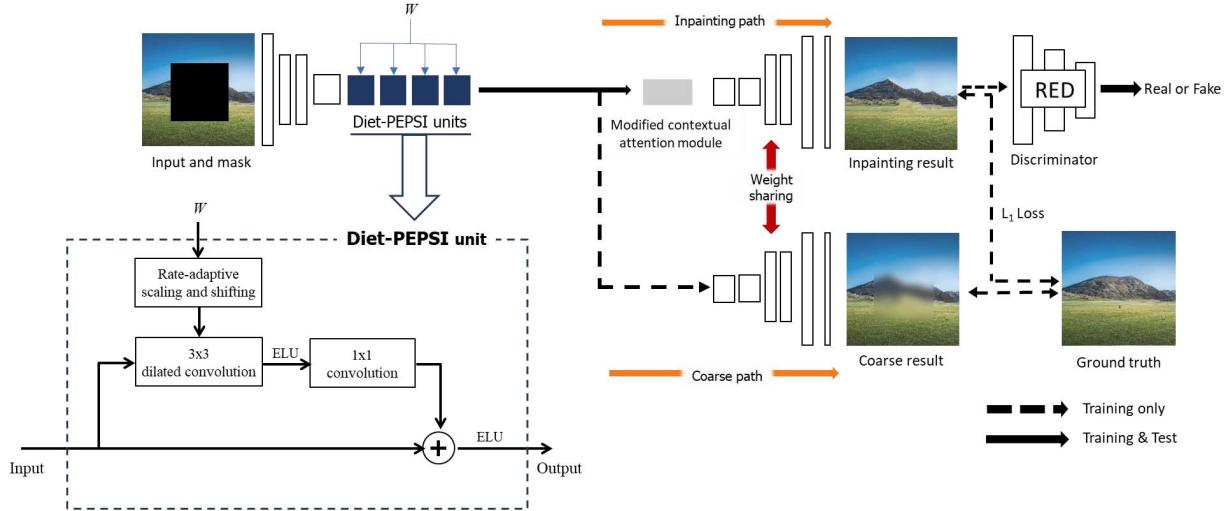


Fig. 6. Architecture of Diet-PEPSI. We replace the multiple dilated convolutional layers with DPUs. In the DPUs, rate-adaptive convolution layers share their weights, whereas the 1×1 standard convolutional layers do not share their weights.

in (4). Using these scaling and shifting processes, the common weights W can be specialized to the desired dilation rate using a small number of parameters.

To demonstrate how W_d can generate different feature maps depending on the given dilation rate, we analyze the computational process in the rate-adaptive dilated convolutional layer. The output of this convolutional layer y is formulated as follows:

$$y = x \otimes (\gamma_d W + \beta_d) = x \otimes \gamma_d W + x \otimes \beta_d \quad (5)$$

where x and \otimes indicate the input and convolution operation, respectively. The first term $x \otimes \gamma_d W$ represents a scaling process that produces the features that are scaled differently according to the given dilation rate, whereas the second term $x \otimes \beta_d$ is a projection process that derives the rate-specific features by projecting x into β_d . In other words, even though the same features are used as the input of the rate-adaptive convolutional layer, this layer can produce different features depending on the given dilation rates.

Using the rate-adaptive convolutional layers, in this article, we propose a novel lightweight model of PEPSI called Diet-PEPSI that significantly reduces the network parameters while preserving the inpainting performance. In Diet-PEPSI, as shown in Fig. 6, we replace the standard dilated convolution layers of PEPSI with residual blocks, i.e. DPUs, which consists of a 3×3 rate-adaptive dilated convolutional layer and a 1×1 standard convolutional layer. By increasing the dilation rate, the DPUs can cover the same size of the receptive field with PEPSI. While the standard dilated convolutional layers need $3 \times 3 \times C_{\text{in}} \times C_{\text{out}} \times n$ network parameters, the DPUs require $(9 + 3n) \times C_{\text{in}} \times C_{\text{out}}$ network parameters, where n indicates the number of DPUs or dilated convolutional layers. Thus, when n is larger than one, DPUs require fewer parameters than the multiple dilated convolutional layers. We will empirically demonstrate the validity of DPUs in Section V-B.

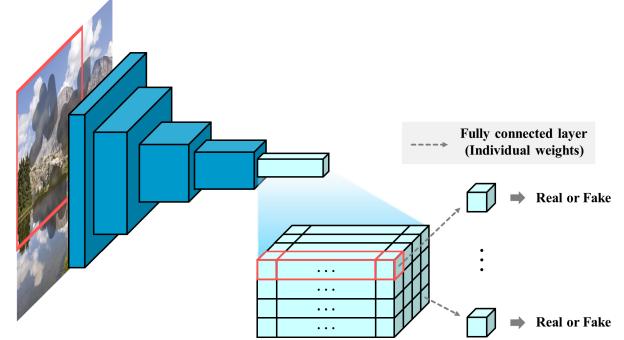


Fig. 7. Overview of the RED. In the last layer, each pixel employs the fully connected layer with different weights. It aims to classify hole regions that may appear in any region with any sizes in an image.

C. Region Ensemble Discriminator

Traditional image inpainting networks [4] utilized both global and local discriminators to determine whether an image has been completed consistently. However, the local discriminator can only handle the hole region with a fixed-size square shape. Thus, it is difficult to employ the local discriminator to train the inpainting network for irregular holes. To solve this problem, we propose a RED inspired by the region ensemble network [29] that detects a target object appearing anywhere in the image by individually handling multiple feature regions. As described in Fig. 7 and Table IV, six strided convolutions with a kernel size of 5×5 and stride 2 are stacked to captures the feature of the whole image. Then, we adopt an individual fully connected layer on each pixel in the last convolutional layer to individually differentiate that each block is real or fake. In other words, we conduct the 1×1 convolution operation on the last layer using pixelwise different weights. It is worth noting that the major difference between RED and existing discriminator [30], called PatchGAN-discriminator, is the last convolutional layer. The PatchGAN-discriminator uses the single regressor in the last convolutional layer,

TABLE IV

DETAILED ARCHITECTURE OF RED. AFTER EACH CONVOLUTION LAYER, EXCEPT LAST ONE, THERE IS A LEAKY-RELU AS THE ACTIVATION FUNCTION. EVERY LAYER IS NORMALIZED BY A SPECTRAL NORMALIZATION. FC* INDICATES THE FULLY CONNECTED LAYER THAT EMPLOYS PIXELWISE DIFFERENT WEIGHTS FOR CONVOLUTION OPERATION

Type	Kernel	Stride	Outputs
Convolution	5×5	2×2	64
Convolution	5×5	2×2	128
Convolution	5×5	2×2	256
Convolution	5×5	2×2	256
Convolution	5×5	2×2	256
Convolution	5×5	2×2	512
FC*	1×1	1×1	1

whereas the RED employs individual regressors in each pixel. This approach allows the RED to act as global and local discriminators simultaneously. The effectiveness of the RED will be revealed in Section V-B.

D. Modified CAM

As mentioned in III-B, the conventional CAM [4] uses the cosine similarity to measure the similarity scores between foreground and background feature patches. However, in 2, since the magnitudes of foreground and background patches, i.e., $f_{x,y}$ and $b_{x',y'}$, are ignored, this approach can result in the distortion of the semantic feature representation. To alleviate this problem, we propose a modified CAM that utilizes the Euclidean distance to measure the similarity scores ($d_{(x,y),(x',y')}$) without the normalization procedure. In the modified CAM, we apply the Euclidean distance instead of the cosine similarity to compute ($d_{(x,y),(x',y')}$). Since the Euclidean distance considers the angle between two vectors of feature patches and their magnitudes simultaneously, it is more appropriate for reconstructing the feature patch. However, since the range of the Euclidean distance is $[0, \infty)$, it is difficult to be directly applied to the softmax. To cope with this problem, we define the truncated distance similarity score $\tilde{d}_{(x,y),(x',y')}$ as follows:

$$\tilde{d}_{(x,y),(x',y')} = \tanh \left(- \left(\frac{d_{(x,y),(x',y')} - m(d_{(x,y),(x',y')})}{\sigma(d_{(x,y),(x',y')})} \right) \right) \quad (6)$$

where $d_{(x,y),(x',y')} = \|f_{x,y} - b_{x',y'}\|$. Since $\tilde{d}_{(x,y),(x',y')}$ has limited values within $[-1, 1]$, it operates like a threshold that sorts out the distance scores less than the mean value. In other words, $\tilde{d}_{(x,y),(x',y')}$ supports to divide the background patches into two groups that may or may not be related to the foreground patch. By using $\tilde{d}_{(x,y),(x',y')}$, the modified CAM weighs them via scaled softmax and reconstructs the foreground patch using the weighted sum of background ones at last like the conventional CAM. The superiority of the modified CAM will be explained in Section V-B.

E. Loss Function

To train PEPSI and Diet-PEPSI, we jointly optimize two different paths: the inpainting path and the coarse path.

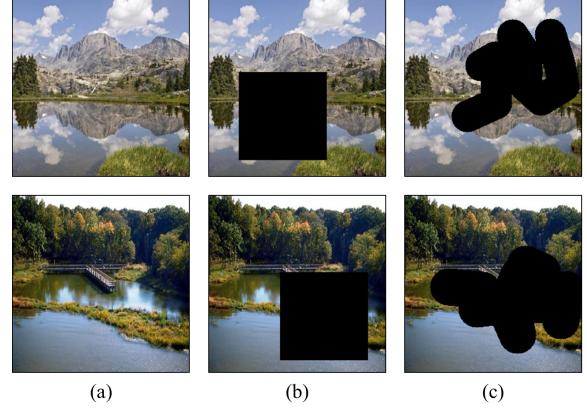


Fig. 8. Examples of masked image. (a) Original images. (b) Images with square mask. (c) Images with free-form mask.

For the inpainting path, we employ the GAN [22] optimization framework in 1, which is described in Section III-A. To avoid the gradient vanishing problem in the generator, inspired by [31], we employ the hinge version of the adversarial loss instead of 1, which is expressed as follows:

$$L_G = -E_{x \sim P_{X_i}}[D(x)] \quad (7)$$

$$L_D = E_{x \sim P_Y}[\max(0, 1 - D(x))] + E_{x \sim P_{X_i}}[\max(0, 1 + D(x))] \quad (8)$$

where P_{X_i} and P_Y denote the data distributions of inpainting results and input images, respectively. It is worth noting that we apply the spectral normalization [32] to all layers in the RED to further stabilize the training of GANs. Since the goal of the inpainting path is not to produce the hole regions naturally but also to recover the missing part of the original image accurately, we add a strong constraint using L_1 -norm to 7 as follows:

$$L_G = \frac{\lambda_i}{N} \sum_{n=1}^N \|X_i^{(n)} - Y^{(n)}\|_1 - \lambda_{\text{adv}} E_{x \sim P_{X_i}}[D(x)] \quad (9)$$

where $X_i^{(n)}$ and $Y^{(n)}$ represent the n th image pair of the generated image through the inpainting path and its corresponding original image in a minibatch, respectively, N is the number of image pairs in a minibatch, and λ_i and λ_{adv} are hyperparameters that control the relative importance of each loss term.

On the other hand, the coarse path is designed to accurately restore the missing features for the CAM. Therefore, we simply optimize the coarse path using an L_1 loss function that is defined as follows:

$$L_C = \frac{1}{N} \sum_{n=1}^N \|X_c^{(n)} - Y^{(n)}\|_1 \quad (10)$$

where $X_c^{(n)}$ are the n th image pair of the generated image via the coarse path in a minibatch. Finally, we define the total loss function of the generative network of PEPSI and Diet-PEPSI as follows:

$$L_{\text{total}} = L_G + \lambda_c \left(1 - \frac{k}{k_{\max}} \right) L_C \quad (11)$$

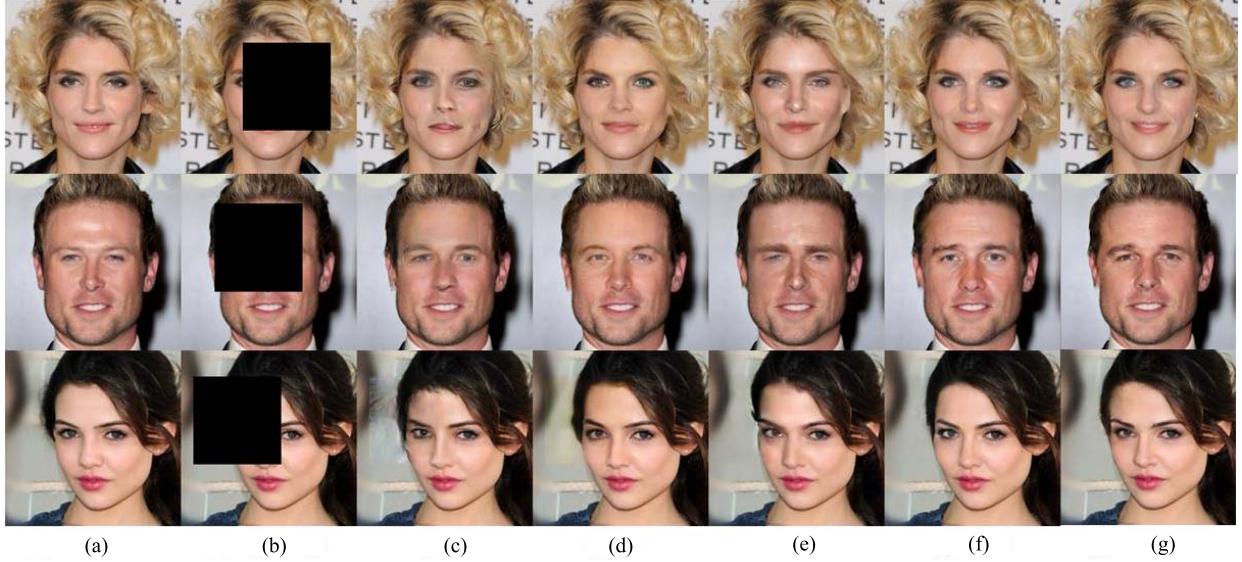


Fig. 9. Comparison of the proposed and conventional methods on randomly square masked CelebA-HQ data sets. (a) Ground truth. (b) Input image of the network. (c) Results of context encoder [10]. (d) Results of globally locally [7]. (e) Results of gated convolution [19]. (f) Results of PEPSI. (g) Results of Diet-PEPSI.

where λ_c is a hyperparameter controlling the contributions from each loss term, and k and k_{\max} represent the iteration of the learning procedure and the maximum number of iterations, respectively. As the training progresses, we gradually decrease the contribution of the L_C for the decoding network to focus on the inpainting path. More specifically, as the training progresses, $(1 - k/k_{\max})$ becomes zero, which results in reducing the contribution of L_C .

V. EXPERIMENTS

A. Implementation Details

1) *Free-Form Mask*: As shown in Fig. 8(b), existing image inpainting methods [4], [7], [10] usually adopt the regular mask, e.g., hole region with rectangular shape, which indicates the background regions during the training procedure. However, the networks trained with the regular mask often exhibit weak performance on inpainting the hole with irregular shape and result in visual artifacts, such as color discrepancy and blurriness. To address this problem, as shown in Fig. 8(c), Yu *et al.* [19] adopted the free-form mask algorithm during the training procedure, which automatically generates multiple random free-form holes with variable numbers, sizes, shapes, and locations randomly sampled at every iteration. More specifically, this algorithm first produces the free-form mask by drawing multiple different lines and erasing pixels closer than the arbitrary distance from these lines. For a fair comparison, in our experiments, we employed the same free-form mask generation algorithm for training PEPSI and Diet-PEPSI.

2) *Training Procedure*: PEPSI and Diet-PEPSI were trained for one million iterations using a batch size of eight in an end-to-end manner. Because the parameters in PEPSI and Diet-PEPSI can be differentiated, we performed an optimization using the Adam optimizer [33] and set the parameters of Adam optimizers β_1 and β_2 to 0.5 and 0.9, respectively.

Motivated by [34], we applied the two-timescale update rule (TTUR) where the learning rates of the discriminator and generator were 4×10^{-4} and 1×10^{-4} , respectively. In addition, we reduced the learning rate to 1/10 after 0.9 million iterations. The hyperparameters of the proposed method were set to $\lambda_i = 10$, $\lambda_c = 5$, and $\lambda_{\text{adv}} = 0.1$. Our experiments were conducted using CPU Intel Xeon CPU E3-1245 v5 and GPU TITAN X (Pascal), and implemented in TensorFlow v1.8.

For our experiments, we used the CelebA-HQ [23], [24], ImageNet [26], and Place2 [25] data sets. More specifically, in the CelebA-HQ data set, we randomly sampled the 27000 images as a training set and 3000 ones as a test set. We also trained the network with all images in the ImageNet data set and tested it on the Place2 data set to measure the performance of trained deep learning models on other data sets; these experiments were conducted to confirm the generalization ability of the proposed method. To demonstrate the superiority of PEPSI and Diet-PEPSI, in addition, we compared their qualitative, quantitative, operation speeds, and number of network parameters with those of the conventional generative methods: context encoders (CE) [10], globally and locally completion network (GL) [7], generator with contextual attention (GCA) [4], and generator with gated convolution (GatedConv) [19].

B. Performance Evaluation

1) *Qualitative Comparison*: To reveal the superiority of PEPSI and Diet-PEPSI, we compared the qualitative performance of the proposed methods with those of the conventional generative methods using images with the squared mask and free-form mask. In our experiments, we implemented conventional methods by following the training procedure in each study. The resultant images with squared mask and free-form mask are described in Figs. 9 and 10. As shown

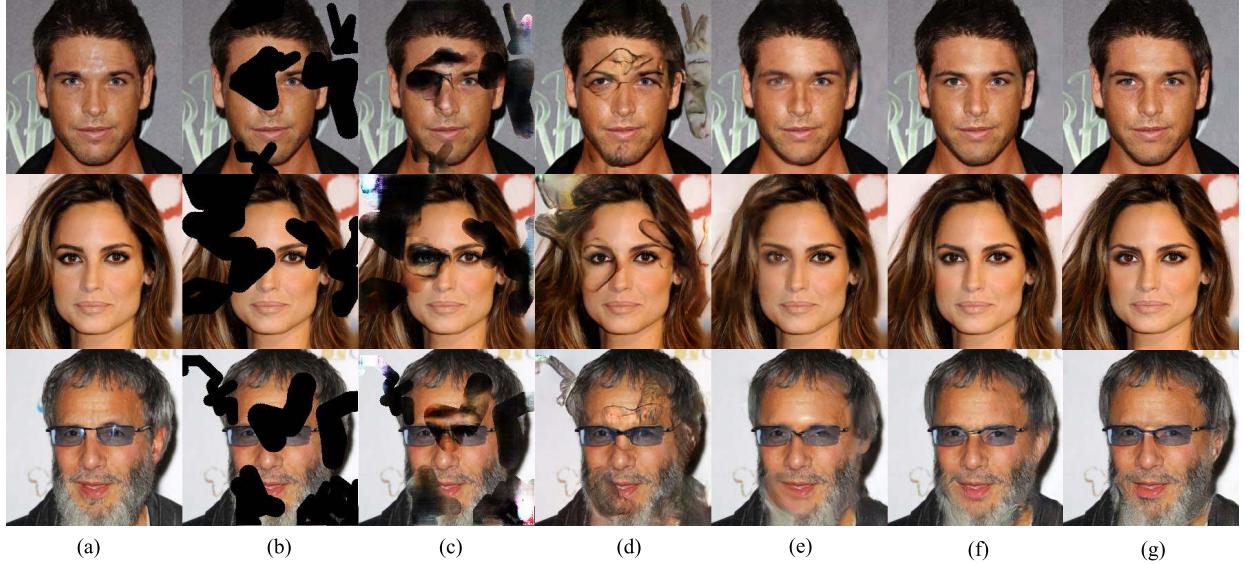


Fig. 10. Comparison of the proposed and conventional methods on free-form masked CelebA-HQ data sets. (a) Ground truth. (b) Input image of the network. (c) Results of context encoder [10]. (d) Results of globally locally [7]. (e) Results of gated convolution [19]. (f) Results of PEPSI. (g) Results of Diet-PEPSI.

in Figs. 9 and 10, CE [10] and GL [7] show obvious visual artifacts including blurred or distorted images in the masked region. In particular, these methods show inferior performance when inpainting the free-form mask, which indicates that CE [10] and GL [7] cannot be applied to real applications. GatedConv [19] exhibits fine performance compared with CE [10] and GL [7]; it still suffers from a lack of relevance between the hole and background regions, such as symmetry of eyes. Compared with the existing methods, PEPSI shows visually appealing results and high relevance between hole and background regions. In addition, as shown in Figs. 9(g) and 10(g), the output image produced via Diet-PEPSI was comparable to PEPSI while saving a significant number of network parameters. From these results, we confirmed that the proposed methods outperform compared with the conventional methods while significantly reducing the hardware costs.

Furthermore, we trained and tested PEPSI and Diet-PEPSI using the challenging data sets, i.e., ImageNet and Place2 data sets, to demonstrate that the proposed methods can be applied to real applications. In this article, we compared the performance of the proposed methods with that of the GatedConv and the nongenerative method, called PatchMatch [3], which is widely applied to image editing applications. We set the image resolution as 256×256 . Resultant images are depicted in Fig. 11. PatchMatch shows visually poor performance especially on the edge of images since it cannot consider the global contexts of the image for inpainting the hole region. GatedConv generates more realistic results without color discrepancy or edge distortion compared with the PatchMatch technique. However, it often produces the images with wrong textures, as shown in the first and third rows in Fig. 11. In contrast to the conventional methods, PEPSI and Diet-PEPSI generate the most natural images without artifacts or distortion on various contents and complex scenes. Thus, we confirmed that the proposed method can be applied to the real application for image inpainting.

2) *Quantitative Comparison:* In this study, we adopted the two different metrics for quantitative assessment: peak signal-to-noise ratio (PSNR) of the local and global regions, i.e., PSNR of the hole region and the whole image, and structural similarity (SSIM) [35] of the whole image. Table V provides the comprehensive performance benchmarks between the proposed methods and conventional ones [4], [7], [10], [19] in the CelebA-HQ data sets [23]. As shown in Table V, compared with the proposed methods, CE [10] and GCA [4] show worse performance on both square mask and free-form mask. GL [7] exhibits comparable performance with the proposed methods only in the square mask since it uses an image blending technique as postprocessing. However, it needs additional computation time due to postprocessing and still suffers from blurred images, as shown in Fig. 9. Also, like the CE and GCA, GL shows poor performance on the free-from mask. Because these methods, i.e., CE, GCA, and GL, designed for inpainting the rectangular mask, they could not cover the free-from mask; they could not generalize well on the free-form mask.

GatedConv [19] shows better performance in both square and free-form holes than other existing methods, but it needs some computation time due to the two stacked generative networks. Compared with conventional methods, PEPSI and Diet-PEPSI show fine performance in both square and free-from masks. In particular, compared with GatedConv, PEPSI and Diet-PEPSI not only exhibit better PSNR and SSIM performances but also require less computational time and hardware costs. In addition, Diet-PEPSI achieves comparable performance with PEPSI while reducing the network parameters almost by 30%. Consequently, these observations indicate that the proposed methods can successfully generate inpainting results with high quality and less hardware costs compared with the conventional inpainting techniques.

Moreover, to reveal the effectiveness of the coarse path, we conducted an extra experiment in which PEPSI was trained without using coarse path learning. The experimental results

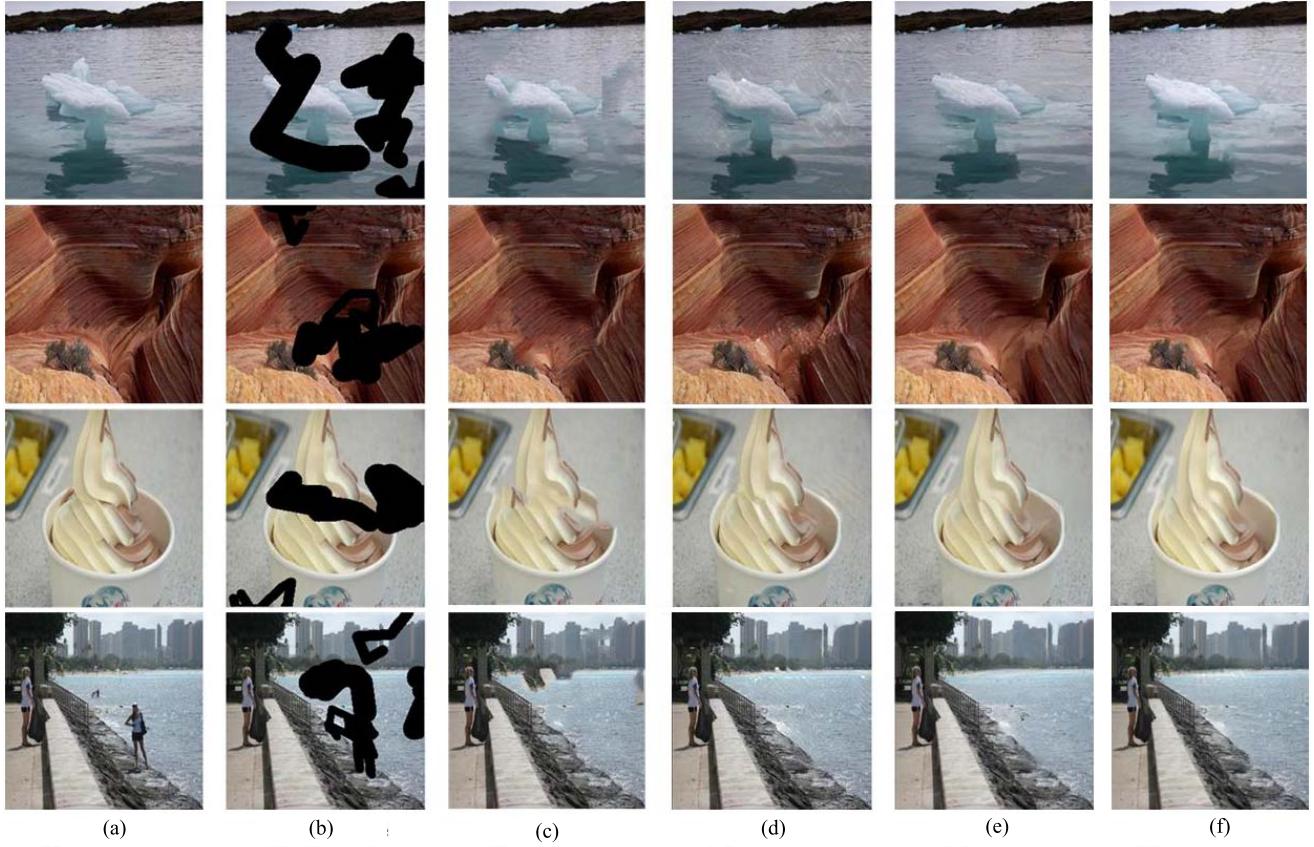


Fig. 11. Comparison of the proposed and conventional methods on Place2 data set. (a) Ground truth. (b) Input image of the network. (c) Results of the nongenerative method, PatchMatch [3]. (d) Results of GatedConv [19]. (e) Results of PEPSI. (f) Results of Diet-PEPSI.

TABLE V

RESULTS OF GLOBAL AND LOCAL PSNRs, SSIM, AND OPERATION TIME WITH SQUARE AND FREE-FORMED MASKS ON CELEBA-HQ DATA SET

Method	Square mask			Free-form mask			Time (ms)	Number of Network Parameters		
	PSNR		SSIM	PSNR		SSIM				
	Local	Global		Local	Global					
CE [10]	17.7	23.7	0.872	9.7	16.3	0.794	5.8	5.1M		
GL [7]	19.4	25.0	0.896	15.1	21.5	0.843	39.4	5.8M		
GCA [4]	19.0	24.9	0.898	12.4	18.9	0.798	22.5	2.9M		
GatedConv [19]	18.7	24.7	0.895	21.2	27.8	0.925	21.4	4.1M		
PEPSI	19.5	25.6	0.901	22.0	28.6	0.929	9.2	3.5M		
PEPSI w/o coarse path	19.2	25.2	0.894	21.6	28.2	0.923	—	—		
Diet-PEPSI	19.4	25.5	0.898	22.0	28.5	0.928	10.9	2.5M		

are described in Table V. PEPSI exhibits better performance than PEPSI trained without the coarse path in terms of all quantitative metrics. These results demonstrate that the coarse path drives the encoding network to produce missing features properly for the CAM. In other words, the single-stage network structure of PEPSI can overcome the limitation of the two-stage coarse-to-fine network through a joint learning scheme.

On the other hand, Diet-PEPSI retains the ability of PEPSI while significantly reducing the network parameters, as shown in Table V. These results reveal that the DPU with a rate-adaptive convolutional layer can replace the standard dilated convolutional layer with a small number of network parameters. To further reduce the hardware costs of

Diet-PEPSI, we conducted additional experiments that apply the group convolution technique [36] to the DPU. In our experiments, we trained Diet-PEPSI by employing the group convolution technique to both layers in the DPU. Note that we utilized the channel shuffling technique between the two convolutional layers of the DPU. As shown in Table VI, even though Diet-PEPSI utilizes a significantly less number of network parameters, it achieves competitive performance with PEPSI as well as shows superior performance compared with other conventional methods. These results confirm that Diet-PEPSI can generate high-quality images with low hardware costs.

To demonstrate the generalization ability of PEPSI and Diet-PEPSI, we conduct another experiment using the challenging

TABLE VI

EXPERIMENTAL RESULTS THAT FURTHER REDUCE THE NETWORK PARAMETERS USING THE GROUP CONVOLUTION TECHNIQUE

	Square mask		Free-form mask		Number of parameters
	PSNR	SSIM	PSNR	SSIM	
PEPSI	25.6	0.901	28.6	0.929	3.5M
Diet-PEPSI	25.5	0.898	28.5	0.928	2.5M
Diet-PEPSI ($g = 2$)	25.4	0.896	28.5	0.928	1.8M
Diet-PEPSI ($g = 4$)	25.2	0.894	28.4	0.926	1.5M

TABLE VII

RESULTS OF GLOBAL AND LOCAL PSNRs AND SSIM ON PLACE2 DATA SET

Mask	Method	PSNR		SSIM
		Local	Global	
Square	GatedConv [19]	14.2	20.3	0.818
	PEPSI	15.2	21.2	0.832
	Diet-PEPSI	15.5	21.5	0.840
Free-form	GatedConv [19]	17.4	24.0	0.875
	PEPSI	18.2	24.8	0.882
	Diet-PEPSI	18.7	25.2	0.889

data sets, i.e., ImageNet [26] and Place2 [25]. As mentioned in Section V-A, in our experiments, we trained the network using the ImageNet data set and tested the trained network on the Place2 data set. Among the various conventional methods, we selected the GatedConv [19], which exhibits superior performance compared with other conventional methods in the CelebA-HQ data set, as our comparison. As shown in Table VII, PEPSI achieves better performance than GatedConv in the Place2 data set. Furthermore, Diet-PEPSI exhibits superior performance compared with GatedConv and PEPSI. These results indicate that the proposed methods can consistently generate high-quality results using various contents and complex images.

3) *DPU Analysis*: To demonstrate the ability of the DPU, we conducted additional experiments that reduce the network parameters using different techniques. Fig. 12 shows the models used in our experiments. Fig. 12(a) and (b) illustrates the convolution layer with a pruning channel and the residual block with a dilated group convolutional layer, respectively, which are an intuitive approach to decrease the number of parameters. Note that we employed the residual block with the same architecture as the DPU for a fair comparison. In addition, we adjusted the pruning channel and the number of groups to make models using an almost similar number of parameters. In our experiments, we set the channels of pruned convolution layers to 113 and the group numbers of the residual block with dilated group convolutional layer to four. The number of groups in the DPU is set to two. As shown in Table VIII, the pruning strategy shows inferior quantitative scores in terms of PSNR in both square and free-form masks. Although the residual block with group dilated convolutional layer shows slightly better performance compared with the pruning strategy, it is still weak. Compared with these models, the DPU shows superior performance in both square and free-form masks. Therefore, these results confirm that the DPU is suitable to effectively aggregate the global contextual information with a small number of parameters.

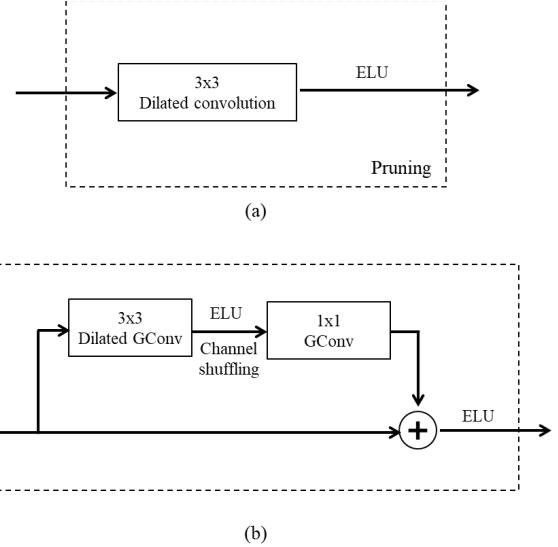


Fig. 12. Illustration of techniques to aggregate the global contextual information while reducing the number of parameters. (a) Dilated convolutional layer with pruning channel. (b) Residual block consisting of group convolutional layers.

TABLE VIII
EXPERIMENTAL RESULTS USING DIFFERENT LIGHTWEIGHT UNITS

	Square mask		Free-form mask	
	PSNR	SSIM	PSNR	SSIM
Pruning	25.21	0.8961	28.28	0.9270
DGC	25.28	0.8959	28.43	0.9270
DPU	25.38	0.8960	28.53	0.9278

4) *RED Analysis*: We demonstrated the superiority of RED by comparing its performance to that of the SNM-discriminator [19] (SNM-Disk), which is an extended version of the PatchGAN-discriminator for image inpainting with a free-form mask. For a fair comparison, we employed each discriminator on the same generator with PEPSI. As shown in Table IX, the SNM-Disk exhibits slightly better performance in terms of PSNR and SSIM compared with the RED. However, Fig. 13 shows that the SNM-Disk could not generate a visually plausible image despite having a high PSNR value; PEPSI trained with the SNM-Disk produced the results with visual artifacts, such as blurred or distorted images in the masked region. These results indicate that the SNM-Disk cannot effectively compete with the generative networks, which makes the generator mainly focus on minimizing the L_1 loss in the objective function of PEPSI. Therefore, even though PEPSI trained with the SNM-Disk exhibits good quantitative performance, it is difficult to apply to the image inpainting in practice.

On the other hand, we investigated the reason why RED could effectively drive the generator to produce visually pleasing inpainting results. The RED follows the inspiration of the region ensemble network [29] that classifies objects in any region of the image. Thus, in adversarial learning, the generator attempts to produce every region of the image to be indistinguishable from real images. This procedure further

TABLE IX
EXPERIMENTAL RESULTS USING DIFFERENT DISCRIMINATORS

	Square mask		Free-form mask	
	PSNR	SSIM	PSNR	SSIM
SNM-Disc [19]	25.68	0.901	28.71	0.932
RED	25.57	0.901	28.59	0.929



Fig. 13. Comparison of RED and SNM-Disk [19] on CelebA-HQ data sets. (a) Input image. (b) Results of PEPSI trained with RED. (c) Results of PEPSI trained with SNM-Disk.

improves the performance of the generator in free-form masks, including irregular holes. Thus, we expect that RED can be applied to various image inpainting networks for generating visually plausible images.

5) *Modified CAM Analysis:* To demonstrate the validity of modified CAM, we performed toy examples comparing the cosine similarity and truncated distance similarity. We reconstructed the hole region using the weighted sum of existing image patches, where the weights, i.e., similarity scores, are computed by using the cosine similarity or the truncated Euclidean distance. Fig. 14 shows comparisons of reconstructed images. As shown in Fig. 14(c) and (d), images reconstructed by applying the truncated distance similarity can collect more similar patches than the cosine similarity; these results indicate that the Euclidean distance is more suitable

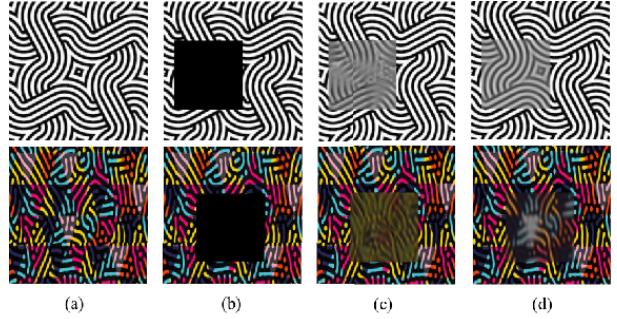


Fig. 14. Comparisons of image reconstruction between the cosine similarity and truncated distance similarity. (a) Original images. (b) Masked images. (c) Images reconstructed by using the cosine similarity. (d) Images reconstructed by using the truncated distance similarity.

TABLE X
PERFORMANCE COMPARISON BETWEEN COSINE SIMILARITY AND EUCLIDEAN DISTANCE APPLYING ON PEPSI

	Square mask		Free-form mask	
	PSNR	SSIM	PSNR	SSIM
Cosine similarity	25.16	0.8950	27.95	0.9218
Euclidean distance	25.57	0.9007	28.59	0.9293

to calculate the similarity score compared with the cosine similarity. To confirm the improvement of the modified CAM, moreover, we compared the quantitative performance of PEPSI with conventional and modified CAMs. As shown in Table X, the modified CAM enhances the performance compared with the conventional CAM, implying that the modified CAM is more appropriate to learn the relationship between background and hole regions.

VI. CONCLUSION

In this article, we have introduced a novel image inpainting model called PEPSI that overcomes the limitation of the two-stage coarse-to-fine network via the joint learning scheme. We provided qualitative and quantitative comparisons on CelebA-HQ and Place2 data sets. Experimental results revealed that PEPSI not only achieves superior performance as compared with conventional techniques but also significantly reduces the computational time via a parallel decoding path and an effective joint learning scheme. Furthermore, we have introduced Diet-PEPSI that utilizes novel rate-adaptive convolutional layers to aggregate the global contextual information with low hardware costs. Experimental results show that Diet-PEPSI preserves the performance of PEPSI while significantly reducing the hardware costs, which facilitates hardware implementation. Both networks are trained with the proposed RED and show visually plausible results in square holes as well as holes with an irregular shape. Therefore, it is expected that the proposed methods can be widely employed in various applications, including image generation, style transfer, and image editing.

REFERENCES

- [1] M. Bertalmio, G. Sapiro, V. Caselles, and C. Ballester, “Image inpainting,” in *Proc. 27th Annu. Conf. Comput. Graph. Interact. Techn.* Reading, MA, USA: Addison-Wesley, 2000, pp. 417–424.

- [2] A. A. Efros and W. T. Freeman, "Image quilting for texture synthesis and transfer," in *Proc. 28th Annu. Conf. Comput. Graph. Interact. Techn. (SIGGRAPH)*, 2001, pp. 341–346.
- [3] 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, p. 24, 2009.
- [4] J. Yu, Z. Lin, J. Yang, X. Shen, X. Lu, and T. S. Huang, "Generative image inpainting with contextual attention," 2018, *arXiv:1801.07892*. [Online]. Available: <http://arxiv.org/abs/1801.07892>
- [5] H. Noori, S. Saryazdi, and H. Nezamabadi-Pour, "A convolution based image inpainting," in *Proc. 1st Int. Conf. Commun. Eng.*, 2010, pp. 1–5.
- [6] H. Li, G. Li, L. Lin, and Y. Yu, "Context-aware semantic inpainting," 2017, *arXiv:1712.07778*. [Online]. Available: <http://arxiv.org/abs/1712.07778>
- [7] S. Iizuka, E. Simo-Serra, and H. Ishikawa, "Globally and locally consistent image completion," *ACM Trans. Graph.*, vol. 36, no. 4, pp. 1–14, Jul. 2017.
- [8] R. Köhler, C. Schuler, B. Schölkopf, and S. Harmeling, "Mask-specific inpainting with deep neural networks," in *Proc. German Conf. Pattern Recognit.* Cham, Switzerland: Springer, 2014, pp. 523–534.
- [9] C. Li and M. Wand, "Combining Markov random fields and convolutional neural networks for image synthesis," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 2479–2486.
- [10] D. Pathak, P. Krahenbuhl, J. Donahue, T. Darrell, and A. A. Efros, "Context encoders: Feature learning by inpainting," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 2536–2544.
- [11] 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. (CVPR)*, Jul. 2017, vol. 1, no. 2, pp. 4076–4084.
- [12] A. Fawzi, H. Samulowitz, D. Turaga, and P. Frossard, "Image inpainting through neural networks hallucinations," in *Proc. IEEE 12th Image, Video, Multidimensional Signal Process. Workshop (IVMSP)*, Jul. 2016, pp. 1–5.
- [13] N. Cai, Z. Su, Z. Lin, H. Wang, Z. Yang, and B. W.-K. Ling, "Blind inpainting using the fully convolutional neural network," *Vis. Comput.*, vol. 33, no. 2, pp. 249–261, 2017.
- [14] R. A. Yeh, C. Chen, T. Y. Lim, A. G. Schwing, M. Hasegawa-Johnson, and M. N. Do, "Semantic image inpainting with deep generative models," in *Proc. CVPR*, Jul. 2017, vol. 2, no. 3, pp. 6882–6890.
- [15] 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. (ECCV)*, Sep. 2018, pp. 85–100.
- [16] Y. Song *et al.*, "Contextual-based image inpainting: Infer, match, and translate," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2018, pp. 3–18.
- [17] Y. Wang, X. Tao, X. Qi, X. Shen, and J. Jia, "Image inpainting via generative multi-column convolutional neural networks," in *Proc. Adv. Neural Inf. Process. Syst.*, 2018, pp. 331–340.
- [18] L. Xu, J. S. J. Ren, C. Liu, and J. Jia, "Deep convolutional neural network for image deconvolution," in *Proc. Adv. Neural Inf. Process. Syst.*, 2014, pp. 1790–1798.
- [19] J. Yu, Z. Lin, J. Yang, X. Shen, X. Lu, and T. S. Huang, "Free-form image inpainting with gated convolution," 2018, *arXiv:1806.03589*. [Online]. Available: <https://arxiv.org/abs/1806.03589>
- [20] 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 Conf. Comput. Vis. Pattern Recognit.*, Jun. 2019, pp. 11352–11360.
- [21] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 770–778.
- [22] I. Goodfellow *et al.*, "Generative adversarial nets," in *Proc. Adv. Neural Inf. Process. Syst.*, 2014, pp. 2672–2680.
- [23] T. Karras, T. Aila, S. Laine, and J. Lehtinen, "Progressive growing of GANs for improved quality, stability, and variation," 2017, *arXiv:1710.10196*. [Online]. Available: <https://arxiv.org/abs/1710.10196>
- [24] Z. Liu, P. Luo, X. Wang, and X. Tang, "Deep learning face attributes in the wild," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Dec. 2015, pp. 3730–3738.
- [25] 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.
- [26] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet classification with deep convolutional neural networks," in *Proc. Adv. Neural Inf. Process. Syst.*, 2012, pp. 1097–1105.
- [27] D. Simakov, Y. Caspi, E. Shechtman, and M. Irani, "Summarizing visual data using bidirectional similarity," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2008, pp. 1–8.
- [28] D.-A. Clevert, T. Unterthiner, and S. Hochreiter, "Fast and accurate deep network learning by exponential linear units (ELUs)," 2015, *arXiv:1511.07289*. [Online]. Available: <https://arxiv.org/abs/1511.07289>
- [29] H. Guo, G. Wang, X. Chen, C. Zhang, F. Qiao, and H. Yang, "Region ensemble network: Improving convolutional network for hand pose estimation," in *Proc. IEEE Int. Conf. Image Process. (ICIP)*, Sep. 2017, pp. 4512–4516.
- [30] P. Isola, J.-Y. Zhu, T. Zhou, and A. A. Efros, "Image-to-image translation with conditional adversarial networks," 2016, *arXiv:1611.07004*. [Online]. Available: <http://arxiv.org/abs/1611.07004>
- [31] H. Zhang, I. Goodfellow, D. Metaxas, and A. Odena, "Self-attention generative adversarial networks," 2018, *arXiv:1805.08318*. [Online]. Available: <http://arxiv.org/abs/1805.08318>
- [32] T. Miyato, T. Kataoka, M. Koyama, and Y. Yoshida, "Spectral normalization for generative adversarial networks," 2018, *arXiv:1802.05957*. [Online]. Available: <http://arxiv.org/abs/1802.05957>
- [33] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," 2014, *arXiv:1412.6980*. [Online]. Available: <http://arxiv.org/abs/1412.6980>
- [34] 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. Adv. Neural Inf. Process. Syst.*, 2017, pp. 6626–6637.
- [35] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli, "Image quality assessment: From error visibility to structural similarity," *IEEE Trans. Image Process.*, vol. 13, no. 4, pp. 600–612, Apr. 2004.
- [36] X. Zhang, X. Zhou, M. Lin, and J. Sun, "ShuffleNet: An extremely efficient convolutional neural network for mobile devices," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 6848–6856.



Yong-Goo Shin received the B.S. and Ph.D. degrees in electronics engineering from Korea University, Seoul, South Korea, in 2014 and 2020, respectively.

He is currently a Research Professor with the Department of Electrical Engineering, Korea University. His research interests are in the areas of digital image processing, computer vision, and artificial intelligence.



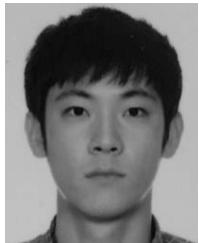
Min-Cheol Sagong received the B.S. degree in electrical engineering from Korea University, Seoul, South Korea, in 2018, where he is currently pursuing the M.S. degree in electrical engineering.

His research interests are in the areas of digital signal processing, computer vision, and artificial intelligence.



Yoon-Jae Yeo received the B.S. degree in electrical engineering from Korea University, Seoul, South Korea, in 2017, where he is currently pursuing the Ph.D. degree in electrical engineering.

His research interests are in the areas of image processing, computer vision, and deep learning.



Seung-Wook Kim received the B.S. and Ph.D. degrees in electronics engineering from Korea University, Seoul, South Korea, in 2012 and 2019, respectively.

He is currently a Research Professor with the Department of Electrical Engineering, Korea University. His research interests are in the areas of image processing and computer vision based on deep learning.



Sung-Jea Ko (Fellow, IEEE) received the B.S. degree in electronic engineering from Korea University, Seoul, South Korea, in 1980, and the M.S. and Ph.D. degrees in electrical and computer engineering from the State University of New York at Buffalo, Buffalo, NY, USA, in 1986 and 1988, respectively.

In 1992, he joined the Department of Electronic Engineering, Korea University, where he is currently a Professor. From 1988 to 1992, he was an Assistant Professor with the Department of Electrical and Computer Engineering, University of Michigan-Dearborn, Dearborn, MI, USA. He has authored or coauthored more than 210 international journal articles. He also holds over 60 registered patents in fields such as video signal processing and computer vision.

Dr. Ko is a member of the National Academy of Engineering of Korea and also a member of the Editorial Board of the IEEE TRANSACTIONS ON CONSUMER ELECTRONICS. He received the Best Paper Award from the IEEE Asia-Pacific Conference on Circuits and Systems in 1996, the LG Research Award in 1999, and both the Technical Achievement Award in 2012, and the Chester Sall Award from the IEEE Consumer Electronics Society in 2017. He was the President of the IEIE in 2013 and the Vice-President of the IEEE CE Society from 2013 to 2016.