IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received 30 December 2022, accepted 30 January 2023, date of publication 3 February 2023, date of current version 8 February 2023.

Digital Object Identifier 10.1109/ACCESS.2023.3242326

RESEARCH ARTICLE

Semantic-Aware Face Deblurring With Pixel-Wise Projection Discriminator

SUJY HAN[1], TAE BOK LEE[1], AND YONG SEOK HEO[1,2]
[1]Department of Artificial Intelligence, Ajou University, Suwon 16499, South Korea
[2]Department of Electrical and Computer Engineering, Ajou University, Suwon 16499, South Korea

Corresponding author: Yong Seok Heo (ysheo@ajou.ac.kr)

This work was supported in part by the Brain Korea 21 (BK21) FOUR Program of the National Research Foundation of Korea through the Ministry of Education under Grant NRF5199991014091; in part by the Basic Science Research Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Education under Grant 2022R1F1A1065702; and in part by the Ministry of Science and Information and Communications Technology (MSIT), South Korea, through the Information Technology Research Center (ITRC) Support Program, supervised by the Institute for Information and Communications Technology Promotion (IITP), under Grant IITP-2022-2018-0-01424.

ABSTRACT Most recent face deblurring methods have leveraged the distribution modeling ability of generative adversarial networks (GANs) to impose a constraint that the deblurred image should follow the distribution of sharp ground-truth images. However, generating sharp face images with high fidelity and realistic properties from a blurry face image remains challenging under the GAN framework. To this end, we focus on modeling the joint distribution of sharp face images and segmentation label maps for face image deblurring in a GAN framework. We propose a semantic-aware pixel-wise projection (SAPP) discriminator that models pixel-label matching with semantic label map information and generates a pixel-wise probability map of realness for the input image as well as a per-image probability. Moreover, we introduce a prediction-weighted (PW) loss to focus on erroneous pixels in the output of the decoder, using per-pixel real/fake probability map to re-weight the contribution of each pixel in the decoder. Furthermore, we present a coarse-to-fine training technique for the generator, which encourages the generator to focus on global consistency in the early training stages and local details in the later stages. Extensive experimental results show that our method outperforms existing methods both quantitatively and qualitatively in terms of perceptual image quality.

INDEX TERMS Face image deblurring, semantic-aware pixel-wise projection discriminator, prediction-weighted loss.

I. INTRODUCTION

Single face image deblurring (SFID) aims to restore a sharp face image from a single blurred face image. It is one of the significant but challenging research areas in computer vision because face analysis plays an important role for many applications including face detection [1], [2], [3], [4], face recognition [5], [6], [7], [8], and age prediction [9], [10], [11], [12]. SFID is a highly ill-posed problem that can have many possible sharp images for a given blurred image; therefore, recent SFID methods have typically leveraged face-specific priors, including face landmarks [13], face sketches [14], face 3D shape [15], face segmentation label maps [16], [17], [18], and deep features [19]. Despite these efforts, these methods [13], [15], [17] often suffer from over-smoothed and perceptually unnatural results.

Some SFID methods [14], [16], [18], [19] are effective at improving the perceptual qualities of deblurred images on the strength of generative adversarial networks (GANs) [20]. GANs have demonstrated an ability to generate realistic samples via a min-max game between a generator and a discriminator. The generator captures the training data distribution, and builds a mapping function from a prior noise distribution to generated data distribution. The discriminator guesses whether the input sample is from the training sharp image (real) or the generator (fake) [21] as shown in Fig. 1 (a). Several SFID methods [14], [16], [18], [19] leverage this distribution modeling

The associate editor coordinating the review of this manuscript and approving it for publication was Taous Meriem Laleg-Kirati.

ability to impose a constraint that the deblurred image should follow the distribution of sharp ground-truth (GT) images.

However, generating sharp face images with high fidelity and realistic properties from a blurry face image remains challenging under the GAN framework. One possible reason is attributed to the discriminator. In the case that additional information, such as semantic labels, exists, the discriminator typically estimates the data distribution of only the sharp face images, and it does not learn the joint distribution of the sharp images and semantic labels. Even though face images are highly structured with semantic components, the decisions of discriminator can be based on relatively unimportant details [22] and not based on the semantically structured features of the face [23]. In addition, the discriminators in the above methods are limited to estimating the global (per-image) real/fake decision without considering local (per-pixel) decisions [24]. Hence, these approaches lack pixel-level details of the generated image and do not guarantee that their generators can synthesize locally plausible images [24], [25], [26].

For the joint distribution modeling of data and additional information such as labels, conditional GANs (cGANs) are widely used [21], [28], [29], [30], [31], [32], [33], [34], [35]. By incorporating data and additional labels, the discriminators identify real images in a principled way, thereby resulting in generators that produce realistic images [36], [37]. Recently, projection GANs [28], [33], [34], [35] have successfully decomposed joint distributions into image distribution (marginal) and label matching distribution (conditional). Specifically, as shown in Fig. 1 (b), the projection discriminator utilizes a class embedding matrix, an image embedding network (encoder), and a linear layer. Despite their promising distribution modeling ability, projection GANs cannot model the pixel-wise joint distribution of the image and semantic label map because they assume that each pixel in the input image shares the same label information.

Meanwhile, U-Net GAN [24] has been proposed to synthesize locally plausible images. As shown in Fig. 1 (c), U-Net GAN utilizes a U-Net [38] structure-based discriminator, which consists of an encoder and a decoder acting as a classifier and a segmenter, respectively. The U-Net discriminator simultaneously outputs the probabilities of whether the input samples are real or fake in both the entire image and each pixel. This global and local feedback encourages the generator to improve the quality of synthesized samples. However, there exists a limitation in that this structure is not designed to take an additional label information as an input. Another limitation is that feedback to the generator can be overwhelmed by the dominant correct pixels of the decoder in the discriminator, resulting in inefficient training. The qualities of face images may be dependent on small components, such as the eyes, nose, and lips. Therefore, it is important to focus on erroneous pixels for generating high-quality details.

To address the limitations mentioned above, we propose a semantic-aware pixel-wise projection (SAPP) GAN with a SAPP discriminator for face image deblurring in a GAN framework. Our SAPP GAN exploits both the U-Net GAN [24] and projection GAN [28]. As shown in Fig. 1 (d), the SAPP discriminator models pixel-label matching with semantic label map information. Unlike projection GAN [28] that utilizes image-level label information as illustrated in Fig. 1 (b), the proposed discriminator models pixel-wise joint distribution of the images and pixel-wise label maps. Furthermore, unlike U-Net GAN [24] (Fig. 1(c)) that models data distribution of only the sharp images, our discriminator can capture joint the distribution of sharp images and corresponding segmentation label maps. By using a face segmentation map as condition information, the SAPP discriminator considers face components when it makes pixel-wise real/fake decisions during training. Empowered by semantic-aware pixel-by-pixel feedback, the generator can restore more accurate and detailed face image with high perceptual quality. Moreover, we propose a prediction-weighted (PW) loss to focus on erroneous pixels in the output of the decoder. The PW loss utilizes a per-pixel real/fake probability map to re-weight the contribution of each pixel in the decoder. Thus, the decoder can discriminate between the generator distribution and the target distribution more precisely, thereby enabling the generator to obtain more powerful and accurate feedback. Furthermore, based on the global and local feedback from the SAPP discriminator, we introduce a coarse-to-fine training technique for the generator, which encourages the generator to focus more on global consistency in the early training stages and local details in the later stages.

We validate the performance of our method on the MSPL dataset [18] and Real-Blur dataset [39] and compare its performance with those of the other SFID methods. Based on these extensive experiments, we show that our method outperforms existing methods quantitatively and qualitatively.

Our contributions can be summarized as follows:

- We present a semantic-aware pixel-wise projection discriminator that models the joint distribution of sharp face images and segmentation label maps.
- We introduce a prediction-weighted loss that gives a high penalty for incorrectly predicted pixels.
- We propose a coarse-to-fine generator training technique that enables the generator to focus on global consistency in the early stages and local details in later stages.
- Our method achieves state-of-the-art performance for single face image deblurring both quantitatively and qualitatively in terms of perceptual image quality.

II. RELATED WORK

Single image deblurring has been studied extensively over the past decades. In this section, we focus our discussion on recent deep learning (DL)-based deblurring methods, which can be divided into two categories: general image deblurring

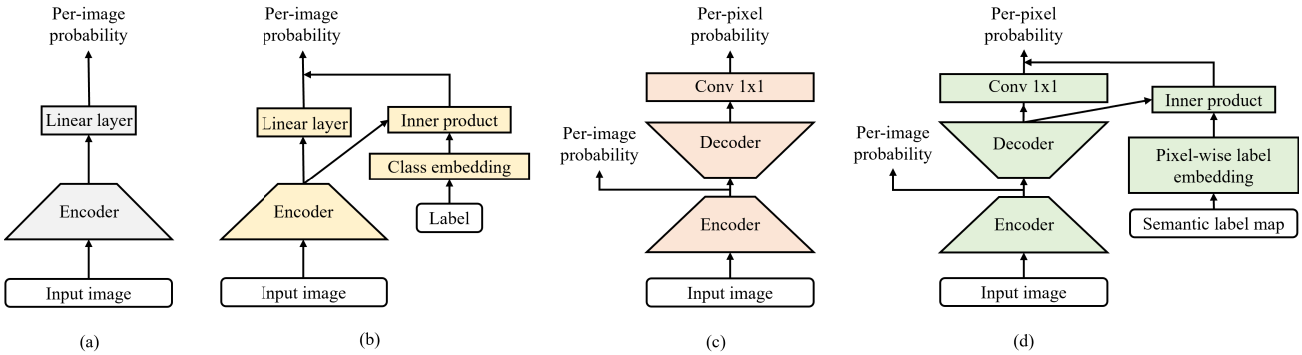

FIGURE 1. Comparison of discriminator architectures. (a) GAN-based SFID methods [16], [18], [19], [27], (b) Projection GAN [28], (c) U-Net GAN [24], (d) proposed SAPPGAN.

and face image deblurring. We also discuss projection-based conditional GANs and U-Net based GANs, which are highly relevant to the proposed method.

A. SINGLE IMAGE DEBLURRING

1) GENERAL DEBLURRING

General image deblurring restores a sharp image from a blurred image captured in a general (natural) scene. Image deblurring has been typically considered as an ill-posed problem with a large solution space [40]. To overcome this, various priors have been studied to regularize the solution space, such as Gaussian mixture [41], hyper-Laplacian [42], ℓ_1/ℓ_2-norms [43], ℓ_0-norms [44], [45], variational Bayes approximations [46], [47], adaptive sparse priors [48], patch priors [49], and dark channel priors [50]. Although these methods perform well in certain cases, they are not flexible for real-world examples, owing to their restrictive assumptions due to regularization [51], [52].

DL-based approaches have recently made significant advances in image deblurring. Early DL-based studies [51], [53], [54] combined convolutional neural networks (CNNs) with traditional optimization-based deconvolution algorithms. Most of these methods used CNNs for blur kernel estimation and then employed optimization-based methods to obtain sharp images. Hence, such methods relied on accurate kernel estimation step [55]. In contrast, Nah et al. [55] proposed a deep neural network that directly restored a sharp image from a blurry image without estimating the blur kernel. In particular, they built a multi-scale CNN, which consisted of multiple sub-networks at each sub-scale and predicted sharp images in a coarse-to-fine manner. Instead of stacking multiple sub-networks, multi-recurrent approaches [52], [56] have proposed to implement coarse-to-fine procedures with a single recurrent neural network (RNN). More recently, multi-patch hierarchy methods [57], [58], [59] have been proposed to restore sharp images progressively from non-overlapping patches. To effectively reduce the computational cost, Cho et al. [60] proposed a multi-input multi-output (MIMO) architecture that accepts multi-scale input images with a single encoder and outputs multiple scales of sharp images with a single decoder.

2) FACE DEBLURRING

While general deblurring models have been well generalized to capture the natural representation of images, they have not been specialized in specific domains, such as face and text images [16], [61]. In contrast, most face deblurring approaches primarily focus on facilitating face restoration by utilizing effective and powerful face prior information, *e.g.*, reference priors [62], [63], face landmarks [13], [64], face sketches [14], face 3D shapes [15], face semantic segmentation maps [16], [17], [18] and deep feature priors [19]. Reference-based approaches [62], [63], [65] use an additional sharp reference face as a guide for face deblurring. However, such methods require a time-consuming procedure to find an adequate reference image. Instead of searching for similar faces, recent face deblurring methods focused on estimating facial priors through deep neural networks (DNNs). Shen et al. [16] first proposed a DNN-based framework consisting of two sub-networks: a semantic face parsing network and a multi-scale deblurring network. The face parsing network first estimates the semantic segmentation maps from blurry face images. Then, the deblurring network performs restoration. Inspired by [16], Yasarla et al. [17] proposed an uncertainty-based multi-stream network (UMSN) that measures the uncertainty score to prevent the negative effects of inaccurate parsing maps. More recently, Lee et al. [18] proposed a multi-semantic progressive learning (MSPL) framework that progressively restores sharp faces component-by-component. Jung et al. [19] developed a deep feature prior-based method that extracts rich information of pre-trained face recognition network to utilize not only the shape prior of the face but also the texture prior.

However, most existing methods still yield blurry face images because they try to model the data distribution of only the sharp images. In contrast to existing methods, the proposed method focuses on modeling a joint distribution of sharp images and their segmentation label maps for semantic-aware restoration.

B. GENERATIVE ADVERSARIAL NETWORKS

GANs [20] have been widely known, showing ability that generates samples similar to a given data distribution via

a min-max game between a generator and a discriminator. The representations learned by GANs have been leveraged in various applications, including style transfer [66], [67], [68], [69], [70], super resolution [71], [72], [73], image generation [29], [33], [37], [74], [75], [76], [77], [78], [79], and hyperspectral image processing [80], [81].

1) CONDITIONAL GANs

For conditional image synthesis, cGANs [21], [28], [29], [30], [31], [32], [33], [34], [35] extend GANs to model the joint distribution of data and conditional information (*e.g.* class labels). cGANs can be categorized into classifier-based [29], [30], [31], [32] and projection-based [28], [33], [34], [35] depending on how the joint distribution is modeled. Classifier-based cGANs feed the class labels into the generator through an additional input layer, while the discriminator utilizes conditional information via an additional classifier. Meanwhile, projection-based cGANs decompose the joint distribution into image distribution and label distribution, utilizing a class embedding matrix and an image embedding network to project the condition information. These networks allows stable training by directly embedding the class label into a feature vector.

2) U-NET BASED GANs

One of the other development directions of GANs is to generate locally coherent images [24], [27], [77]. In particular, U-Net GAN [24] implements U-Net [38] based discriminator architecture, which consists of an encoder and a decoder, acting as a classifier and a segmenter, respectively. U-Net based discriminator outputs the probabilities of the input sample being real over the entire image and per-pixel through the encoder-decoder architecture. The per-pixel decision provides spatially coherent feedback to the generator while the per-image decision gives global coherent feedback. Encouraged by the feedback, the generator attempts to improve the quality of synthesized samples. Owing to its powerful data representation, U-Net GAN has been adopted in various studies [25], [26], [82].

Although the proposed SAPPGAN is highly inspired by the projection GAN [28] and U-Net GAN [24], there are some key differences. Unlike the projection GAN, which considers each pixel in the input image to share the same label information, our method has the ability to model the pixel-wise joint distribution of the image and semantic label map. Moreover, the U-Net GAN [24] has a limited ability to model the joint distribution of input data and external data because it has been developed under unconditional settings. In contrast, we condition the U-Net GAN [24] on additional information (segmentation maps) to construct the conditional framework.

III. PROPOSED METHOD

The proposed SAPPGAN consists of a deblurring network G and a discriminator network D. The overall architecture is shown in Fig. 2. G takes a blurred face image $I_{blur} \in \mathbb{R}^{H\times W\times 3}$ as its input, where H and W the represent height and width of image, respectively. Then, G outputs a deblurred face image $I_{deblur} \in \mathbb{R}^{H\times W\times 3}$ as follows:

$$I_{deblur} = G(I_{blur}). \tag{1}$$

D takes a sharp face image $x \in \mathbb{R}^{H\times W\times 3}$ and a segmentation map $y_s \in \mathbb{R}^{H\times W\times 1}$ as its input and models the joint distribution $p(x, y_s)$. Here, x can be a deblurred image I_{deblur} or a GT sharp image I_{GT}. y_s is used for the condition information of D. The proposed discriminator can be employed in several SFID methods that use the GAN framework. Thus, we adopt DFPGNet [19] as our generator. Similar to other GAN-based networks [20], we alternatively train our generator G and discriminator D.

In this section, we first introduce our semantic-aware pixel-wise projection (SAPP) discriminator, which utilizes segmentation maps for pixel-wise real/fake decisions. We then describe our discriminator training technique with the proposed prediction-weighted (PW) loss, which re-weights the contribution of each pixel in the decoder using the probability map output from the SAPP discriminator. We subsequently introduce the generator training technique based on the coarse-to-fine strategy.

A. SEMANTIC-AWARE PIXEL-WISE PROJECTION DISCRIMINATOR

We propose a SAPP discriminator that considers face component information y_s when it makes the pixel-wise real/fake decision. The encoder D_{enc} and decoder D_{dec} of our SAPP discriminator D is adopted from those of the U-Net discriminator [24]. As shown in Fig. 2, the body network of the encoder D_{enc}^{body} takes a face image x as its input and outputs a feature map $Z \in \mathbb{R}^{\frac{H}{32}\times\frac{W}{32}\times 1024}$, on which the head network of the encoder D_{enc}^{head} is applied to generate a probability of realness as follows:

$$\begin{aligned} Z &= D_{enc}^{body}(x), \\ p_{enc} &= D_{enc}^{head}(Z), \end{aligned} \tag{2}$$

where p_{enc} denotes a global probability of x being real. There are 5 downsampling stages in D_{enc}^{body}, where each stage is a series that includes 3×3 convolution layer, *ReLU* activation, 3×3 convolution layer, and 2×2 average pooling layer. D_{enc}^{head} is a series of a global sum pooling layer and fully connected layer.

With the Z from D_{enc}^{body}, D_{dec} outputs the per-pixel real/fake prediction. However, unlike the U-Net discriminator [24], the SAPP discriminator leverages face semantic maps y_s to further decide whether the input image matches the semantic label map condition. Thus, D_{dec} segments the input image as real or fake, conditioned on y_s, which results in $Q_{dec} \in \mathbb{R}^{H\times W\times 1}$.

As shown in Fig. 2, D_{dec} consists of body network D_{dec}^{body}, label embedding matrix V and head layer D_{dec}^{head} as in [28]. D_{dec} is connected with D_{enc} through skip connections that concatenate corresponding feature maps from the stages of

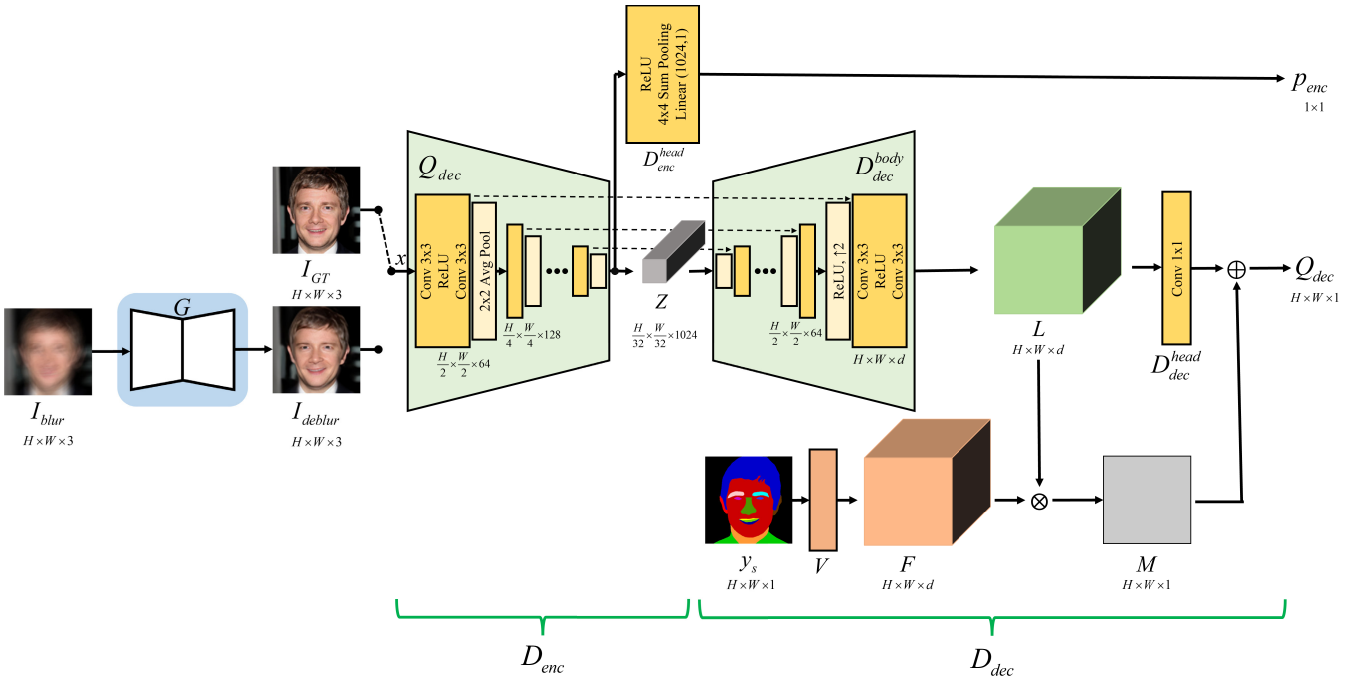

FIGURE 2. Overall architecture of the proposed face deblurring framework. The proposed SAPP discriminator includes the encoder and decoder that predict the real/fake decisions at the global image-level and local pixel-level, respectively. To estimate the joint distribution of sharp face images and the semantic structure of the face, our discriminator takes the inner product between the embedding of the face segmentation label map and the feature map of the input image.

D_{enc}^{body} and D_{dec}^{body}. There are 5 upsampling stages in D_{dec}^{body}, where each stage is a series that includes *ReLU* activation, $\uparrow$ 2 upsampling using nearest-neighbor interpolation, a 3×3 convolution layer, *ReLU* activation, and a 3×3 convolution layer. Thus, D_{dec}^{body} upsamples the input feature map as follows:

$$L = D_{dec}^{body}(Z), \tag{3}$$

where $L \in \mathbb{R}^{H \times W \times d}$ is the output feature map.

$V \in \mathbb{R}^{N \times d}$ contains a list of the d-dimensional row embedding vectors of the N class labels. Note that unlike [28], V embeds the segmentation label map y_s pixel-wise to feature map $F \in \mathbb{R}^{H \times W \times d}$. Thus, y_s is first one-hot encoded and then unrolled to $\hat{y}_s \in \mathbb{R}^{HW \times N}$. Then the embedded matrix $\hat{F} \in \mathbb{R}^{HW \times d}$ is obtained as:

$$\hat{F} = \hat{y}_s \odot V, \tag{4}$$

where $\odot$ represents matrix multiplication. Finally, $\hat{F}$ is rearranged to feature map F of dimension $H \times W \times d$.

By taking the inner product between L and F at the pixel level, we can obtain a per-pixel conditional probability map M as:

$$M = \{M_{i,j} | M_{i,j} = F_{i,j} \cdot L_{i,j}\}, \tag{5}$$

where $F_{i,j} \in \mathbb{R}^{1 \times 1 \times d}$ and $L_{i,j} \in \mathbb{R}^{1 \times 1 \times d}$ represent the vector element at location (i, j) of F and L, respectively. $M_{i,j} \in \mathbb{R}^{1 \times 1 \times 1}$ represents the degree of matching between the pixel of x and semantic label of y_s at location (i, j). Thus, M represents the conditional probabilities *i.e.* the image-label matching map.

D_{dec}^{head} is a 1×1 convolution layer that takes L as its input and outputs per-pixel marginal probabilities *i.e.* an image-based real/fake probability map. Therefore, the final decoder output Q_{dec} is calculated by the summation of the image-label matching map and image-based real/fake probability map as follows:

$$Q_{dec} = M \oplus D_{dec}^{head}(L), \tag{6}$$

where $\oplus$ denotes element-wise summation. Giving a condition y_s as a semantic face map enables D_{dec} to further make an accurate per-pixel decision. Thus, G can generate more accurate and realistic face details when driven by semantic-aware feedback.

B. DISCRIMINATOR TRAINING

We propose a prediction-weighted (PW) loss that utilizes the probability map from the decoder to re-weight the contribution of each pixel to the loss for the decoder [24]. For convenience in notation, we define $Q_{dec}^r \in \mathbb{R}^{H \times W \times 1}$ and $Q_{dec}^f \in \mathbb{R}^{H \times W \times 1}$ for decoder output Q_{dec} when the input is real and fake, respectively:

$$\begin{aligned} Q_{dec}^r &= D_{dec}(I_{GT}), y_s), \\ Q_{dec}^f &= D_{dec}(G(I_{blur}), y_s). \end{aligned} \tag{7}$$

Additionally, we define probability map $p_r \in \mathbb{R}^{H \times W \times 1}$ for the real input and $p_f \in \mathbb{R}^{H \times W \times 1}$ for the fake input. Note

that each pixel in both p_r and p_f represent the realness of the corresponding pixel, *i.e.* the probability of the real class when the input image is real and fake, respectively. Then, we can derive p_r and p_f from the original GAN loss [20] as follows:

$$\begin{aligned} -\log(p_r) &= A(-Q^r_{dec}) = \log(1+\exp(-Q^r_{dec})), \\ -\log(1-p_f) &= A(Q^f_{dec}) = \log(1+\exp(Q^f_{dec})), \end{aligned} \tag{8}$$

where $A(t) = log(1 + exp(t))$ refers to the *SoftPlus* function [28], [35]. By rearranging above equations, p_r and p_f are obtained as:

$$\begin{aligned} p_r &= \frac{1}{1+\exp(-Q^r_{dec})} = sigmoid(Q^r_{dec}), \\ p_f &= 1 - \frac{1}{1+\exp(Q^f_{dec})} = \frac{1}{1+\exp(-Q^f_{dec})} \\ &= sigmoid(Q^f_{dec}). \end{aligned} \tag{9}$$

Finally, our PW loss is defined as:

$$\begin{aligned} L_{D_{dec}} = &\sum_{i,j}^{W,H} [\xi(\mathbb{1} - p_r) \otimes A(-D_{dec}(I_{GT}, y_s))]_{i,j} \\ &+ \sum_{i,j}^{W,H} [\xi(p_f) \otimes A(D_{dec}(G(I_{blur}), y_s))]_{i,j}. \end{aligned} \tag{10}$$

Here, $\otimes$ refers to element-wise multiplication and $[\cdot]_{i,j}$ represents pixel location (i, j). $\xi(\cdot)$ is a normalization function as $[\xi(t)]_{i,j} = t_{i,j} / \sum_{i,j}^{W,H} t_{i,j}$, and $\mathbb{1} \in R^{H\times W\times 1}$ is a matrix filled with ones.

As PW loss aims to emphasize the erroneous prediction of the decoder, $\mathbb{1} - p_r$ and p_f are used as per-pixel weighting factors for the real and fake inputs, respectively. For example, when the discriminator incorrectly determines that a real pixel is fake, the value of that pixel in p_r becomes low, which highly affects the PW loss. Similarly, in the fake data, misjudged pixels have high p_f values; thus they have a large impact on PW loss and vice versa. As $\mathbb{1} - p_r$ and p_f have different values for each pixel, PW loss can highlight regions with wrong predictions, similar to [83].

Overall, our discriminator objective function L_D consists of an encoder loss $L_{D_{enc}}$ and decoder loss $L_{D_{dec}}$ as:

$$L_D = L_{D_{enc}} + L_{D_{dec}}, \tag{11}$$

where encoder loss is defined as follows [20]:

$$\begin{aligned} L_{D_{enc}} = &-\log D_{enc}(I_{GT}) \\ &-\log(1 - D_{enc}(G(I_{blur}))), \end{aligned} \tag{12}$$

and decoder loss $L_{D_{dec}}$ is defined as Eq. (10).

C. GENERATOR TRAINING

The generator objective function includes reconstruction loss L_{pixel}, prior feature loss L_{feat}, and adversarial loss L_{adv}. Reconstruction loss is defined as L_1 distance between the GT image I_{GT} and the deblurred image I_{deblur} in image domain as follows:

$$L_{pixel} = ||I_{GT} - I_{deblur}||_1. \tag{13}$$

Inspired by [19], we employ deep feature prior loss L_{feat} to utilize the rich information of deep features extracted from the well-trained VGGFace [84] network θ. Let $\theta(\cdot)_l$ be the intermediate output features of the l^{th} layer of θ. Then, L_{feat} minimizes the L_2 distance between the deep features obtained using I_{GT} and I_{deblur} as:

$$L_{feat} = \sum_{l=1}^{3} ||\theta(I_{GT})_l - \theta(I_{deblur})_l||_2. \tag{14}$$

We select the `relu1_2`, `relu2_2`, and `relu3_3` layers of VGGFace for $\theta(\cdot)_1, \theta(\cdot)_2$ and $\theta(\cdot)_3$, respectively following [19].

L_{adv} is an adversarial loss from D that encourages G to generate more realistic details as:

$$L_{adv} = \alpha L_{adv,enc} + (1-\alpha) L_{adv,dec}, \tag{15}$$

where $L_{adv,enc}$ and $L_{adv,dec}$ are defined as:

$$\begin{aligned} L_{adv,enc} &= -\log D_{enc}(G(I_{blur})), \\ L_{adv,dec} &= \sum_{i,j}^{W,H} [\xi(\mathbb{1} - p_f) \\ &\quad \otimes A(-D_{dec}(G(I_{blur}), y_s))]_{i,j}. \end{aligned} \tag{16}$$

Here, α is a balancing coefficient that makes the training process effective by enabling coarse-to-fine training. As the deblurring task is a challenging and extremely ill-posed problem, it is effective to decompose the deblurring task into smaller and easier sub-tasks. Thus, we divide the SFID task into two sub-tasks, which are to the learn global face image distribution and learn the structural features and detailed textures of the real face image. Thus, α is a scalar that decreases in proportion to the current epoch η_c:

$$\alpha = \frac{\eta_t - \eta_c}{\eta_t}, \tag{17}$$

where η_t refers to the total number of epochs. In the early stages, $L_{adv,enc}$ has a higher effect on L_{adv} than $L_{adv,dec}$, thus, the generator focuses more on global consistency than local consistency by balancing $L_{adv,enc}$ and $L_{adv,dec}$. As the training proceeds, *i.e.* for larger η_c, the effect of $L_{adv,dec}$ increases, so that the generator focuses on local details. By doing this, the generator can output detailed deblurred face images.

Finally, the final generator objective L_G becomes

$$L_G = \lambda_{pixel} L_{pixel} + \lambda_{feat} L_{feat} + \lambda_{adv} L_{adv}, \tag{18}$$

where λ_{pixel}, λ_{feat}, and λ_{adv} are hyperparameters that are empirically set as 1, 1, and 0.06, respectively.

TABLE 1. Quantitative Comparisons on MSPL testset. The best and the second best results are marked in bold and underline, respectively.

Method	MSPL-Center														
	CelebA					CelebA-HQ					FFHQ				
	$d_{VGG}\downarrow$	$d_{ARC}\downarrow$	LPIPS ↓	PSNR ↑	SSIM ↑	$d_{VGG}\downarrow$	$d_{ARC}\downarrow$	LPIPS ↓	PSNR ↑	SSIM ↑	$d_{VGG}\downarrow$	$d_{ARC}\downarrow$	LPIPS ↓	PSNR ↑	SSIM ↑
Shen *et al.* [16]	113.66	1.141	0.301	19.75	0.740	267.41	0.872	0.287	19.95	0.755	180.10	1.252	0.342	19.57	0.723
Lu *et al.* [61]	123.35	1.172	0.228	17.93	0.617	243.06	0.880	0.190	18.63	0.649	177.00	1.028	0.226	18.26	0.630
Zhang *et al.* [57]	117.68	-	0.314	20.40	0.744	239.04	-	0.295	20.90	0.764	170.41	-	0.343	20.64	0.743
*Zhang *et al.*	45.13	0.936	0.241	23.98	0.824	83.36	0.699	0.212	24.84	0.844	71.51	1.079	0.287	23.52	0.813
Xia *et al.* [85]	39.58	0.851	0.179	25.03	0.873	83.46	0.634	0.161	25.79	0.886	57.66	0.962	0.208	24.66	0.859
Yasarla *et al.* [17]	55.01	-	0.213	22.73	0.817	102.97	-	0.196	23.02	0.827	86.43	-	0.251	22.19	0.795
*Yasarla *et al.*	37.80	0.806	0.183	24.71	0.857	50.67	0.562	0.148	26.11	0.882	58.46	0.958	0.218	24.31	0.843
Lee *et al.* [18]	18.19	0.499	0.115	28.08	0.921	40.93	0.404	0.097	28.82	0.929	25.39	0.648	0.133	27.36	0.908
Jung *et al.* [19]	14.76	0.453	0.102	**29.06**	**0.933**	33.29	0.370	0.085	**29.86**	**0.940**	20.28	0.595	0.118	**28.76**	**0.921**
Ours	**12.33**	**0.208**	**0.071**	28.32	0.926	**32.90**	**0.246**	**0.082**	29.26	0.934	**17.56**	**0.309**	**0.086**	27.65	0.915
Method	MSPL-Random														
	CelebA					CelebA-HQ					FFHQ				
	$d_{VGG}\downarrow$	$d_{ARC}\downarrow$	LPIPS ↓	PSNR ↑	SSIM ↑	$d_{VGG}\downarrow$	$d_{ARC}\downarrow$	LPIPS ↓	PSNR ↑	SSIM ↑	$d_{VGG}\downarrow$	$d_{ARC}\downarrow$	LPIPS ↓	PSNR ↑	SSIM ↑
Shen *et al.* [16]	90.37	1.284	0.331	18.89	0.711	157.49	1.031	0.319	19.18	0.729	127.71	1.131	0.336	19.03	0.713
Lu *et al.* [61]	96.71	1.609	0.269	17.41	0.631	156.96	1.262	0.230	18.04	0.664	129.43	1.444	0.259	17.94	0.654
Zhang *et al.* [57]	86.77	-	0.328	19.36	0.702	144.74	-	0.311	19.85	0.726	122.07	-	0.333	19.77	0.715
*Zhang *et al.*	30.46	1.058	0.254	23.35	0.794	54.06	0.816	0.227	24.09	0.817	46.03	0.943	0.255	23.54	0.804
Xia *et al.* [85]	30.94	0.976	0.204	23.66	0.849	60.95	0.789	0.194	24.48	0.861	44.62	0.858	0.202	23.95	0.855
Yasarla *et al.* [17]	45.05	-	0.245	21.24	0.777	72.56	-	0.230	21.46	0.789	65.06	-	0.241	21.28	0.778
*Yasarla *et al.*	33.83	0.976	0.234	22.92	0.789	58.89	0.789	0.214	23.56	0.812	49.94	0.846	0.227	23.16	0.793
Lee *et al.* [18]	11.41	0.409	0.109	28.95	0.936	26.91	0.395	0.094	29.80	0.945	15.44	0.337	0.099	29.22	0.941
Jung *et al.* [19]	8.37	0.364	0.100	**29.96**	**0.945**	23.05	0.363	0.084	**30.76**	**0.953**	10.93	0.299	0.089	**30.29**	**0.951**
Ours	**7.72**	**0.165**	**0.062**	29.13	0.939	**22.48**	**0.255**	**0.073**	30.04	0.949	**10.60**	**0.148**	**0.058**	29.51	0.945

IV. EXPERIMENTS

A. EXPERIMENTAL DETAILS

1) DATASETS

The training and evaluation are conducted on the MSPL dataset [18], which has been used in recent SFID studies [18], [19]. The MSPL dataset consists of training set and a test set for face deblurring collected from various face images. The detailed description of the MSPL dataset is as follows.

- **The MSPL training set** consists of 24, 183 pairs of blurred face images and the corresponding sharp GT face images. The GT face images are collected from the CelebAMask-HQ dataset [86], which contains pairs of high-quality (1024 × 1024 resolution) face images and corresponding segmentation label maps. Each segmentation label map is precisely and manually annotated with 19 classes, including facial components and accessories, such as the eyes, eyebrows, nose, mouth, lips, ears, hair and skin. In practice, segmentation label maps for face image datasets can be obtained leveraging pre-trained face parsing networks [87], [88], [89]. In [18], 18000 motion blur kernels are synthesized from random 3D trajectories, where the size of blur kernel ranges from 13 × 13 to 27 × 27 including {13 × 13, 15 × 15, 17 × 17, 19 × 19, 21 × 21, 23 × 23, 25 × 25, 27 × 27}. Each blurred image is obtained by convolving the sharp image with one of blur kernels and adding Gaussian noise with standard deviation 0.015.
- **The MSPL testset** is further divided into the MSPL-Center test set and MSPL-Random test set. The former primarily consists of images with a frontal face at the center position. The latter provides images of randomly rotated or/and cropped versions of the former. Each of the MSPL-Center and MSPL-Random test sets contain 240 sharp-blurry face pairs collected from the CelebA [90], CelebAMASK-HQ [86] and FFHQ [91].

2) IMPLEMENTATION DETAILS

We implement our model using the PyTorch [92] and train it using two NVIDIA Titan Xp GPUs. During training, we adopt the Adam optimizer [93] with $\beta_1 = 0.9$, $\beta_2 = 0.999$. The learning rates of the generator and discriminator are initialized to 1×10^{-4} and decayed exponentially by 0.99 every epoch. For every training iteration, the pairs of GT images, blurry images and segmentation label maps are sampled with a batch size of 16. As in [17], [18], [19], random horizontal flips and random rotations are performed for data augmentation. The proposed network is trained for 300 epochs, which is sufficient for convergence.

3) EVALUATION METRICS

For the quantitative evaluation, we employ the perceptual image quality assessment metrics: the identity distance (d_{ARC}) [94] between the GT and restored face images using the pre-trained Arcface [95] embedding vector, feature distance (d_{VGG}) of the pre-trained VGGFace [84] to measure

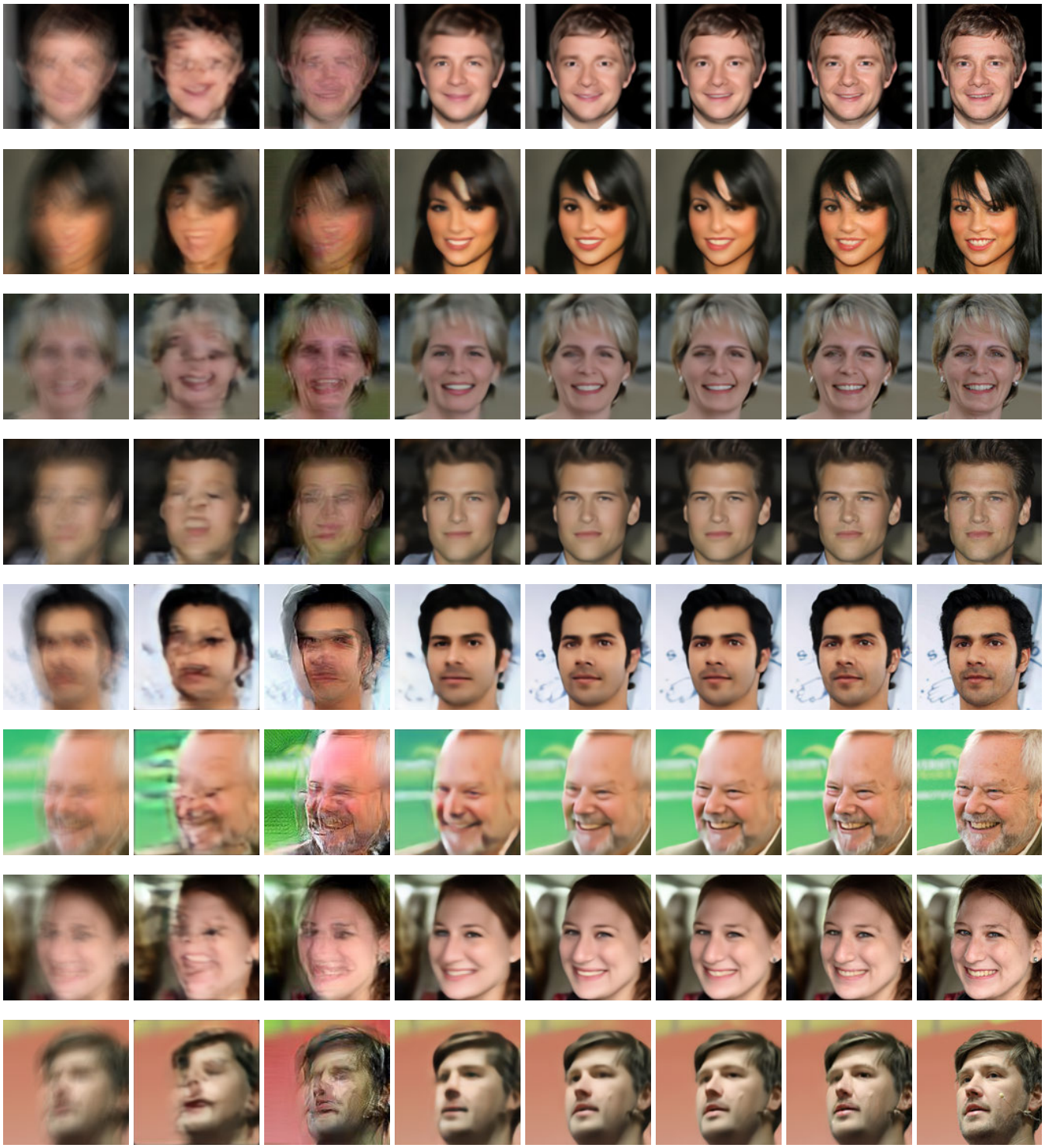

FIGURE 3. Qualitative comparison on MSPL-Center testset [18]. For a better comparison of visual quality, zooming-in is recommended.

the similarity of the facial identity between the GT and deblurred images, and learned perceptual image patch similarity (*LPIPS*) [96] for perceptual quality. Note that smaller values of d_{ARC}, d_{VGG}, and *LPIPS* indicate higher consistency with the GT face image. Moreover, we report widely-used image quality assessment metrics, which are the peak signal-to-noise ratio (PSNR) and structural similarity index map (SSIM) [97].

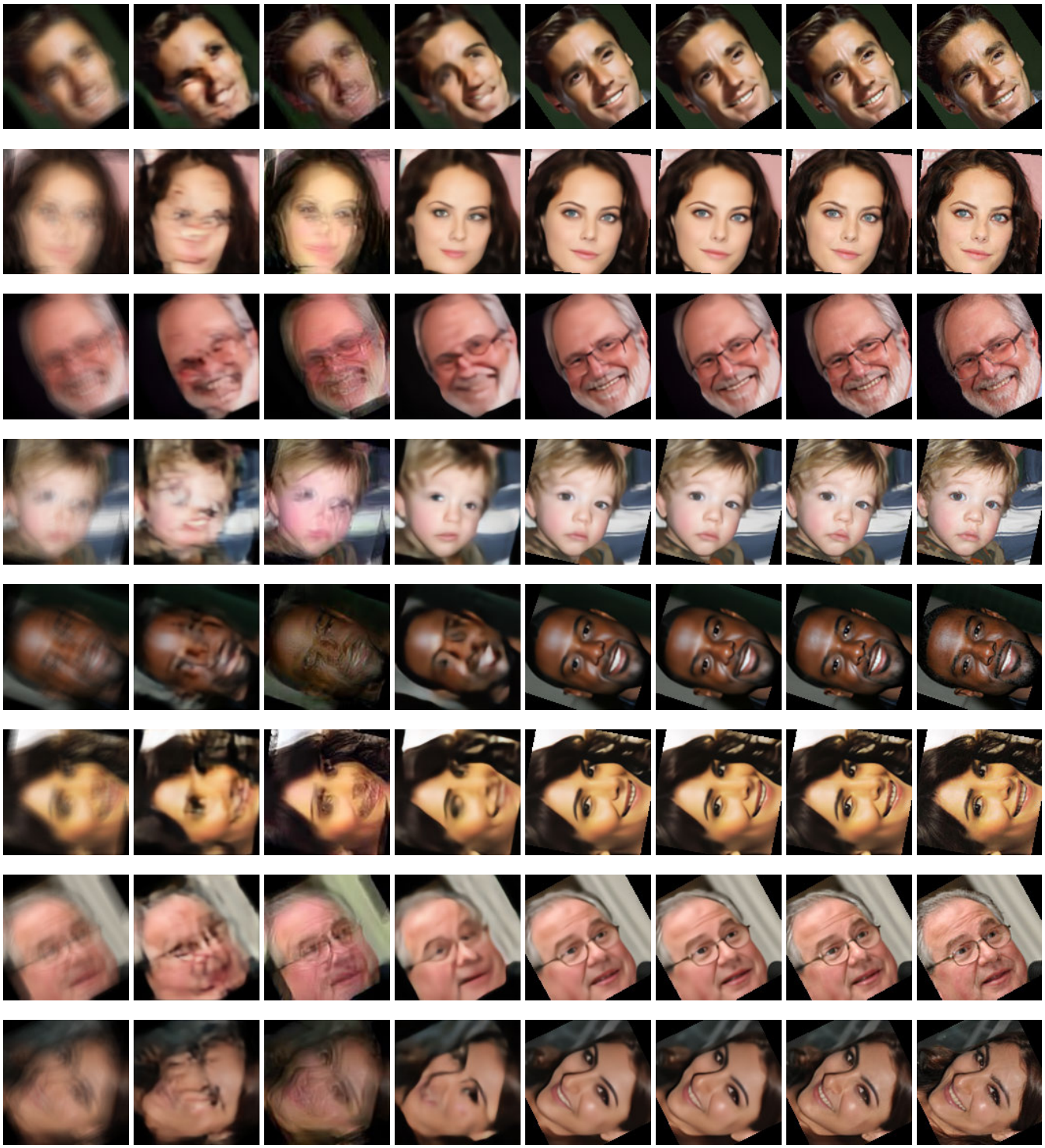

FIGURE 4. Qualitative comparison on MSPL-Random testset [18]. For a better comparison of visual quality, zooming-in is recommended.

B. COMPARISONS ON MSPL DATASET

We compare the proposed SAPPGAN with state-of-the-art deblurring models, including general models [57], [60] and face models [16], [17], [18], [19], [61], [85]. For general deblurring models [57], [60] which are originally trained on natural scenes, we report additional results using the retrained models on the MSPL training set. As existing face deblurring models [16], [17], [18], [19] are trained on different training set or/and synthetic blur kernels, we also retrained them on the MSPL training set for a fair comparison. Throughout

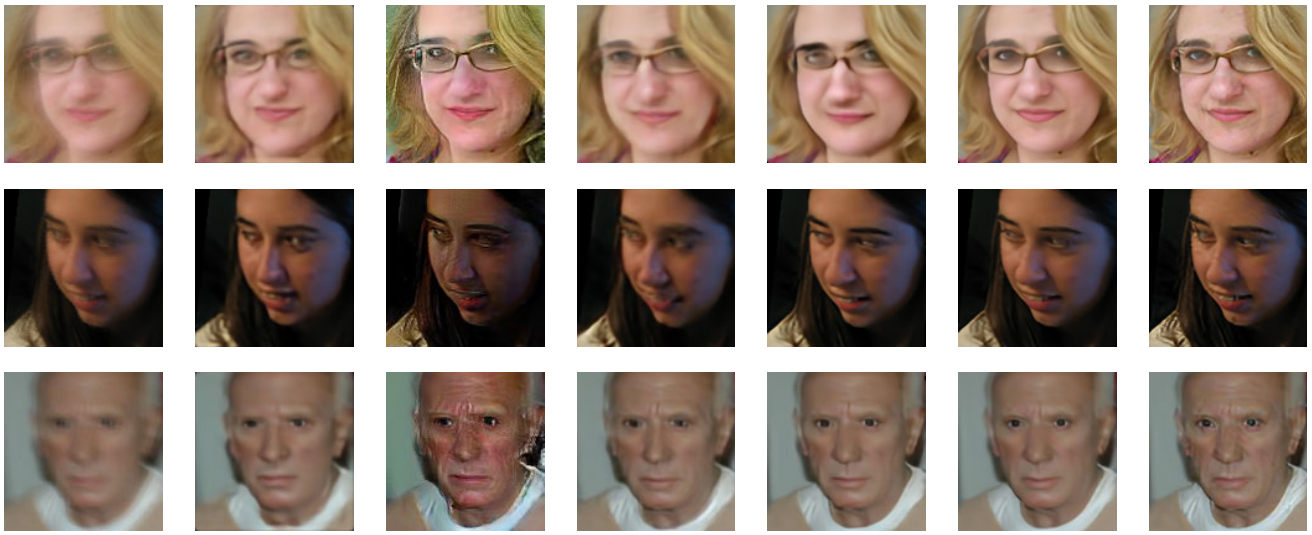

FIGURE 5. Qualitative comparison on Real-Blur test set [39].

TABLE 2. Comparison with recent SFID methods for average run time, model parameters and verification accuracy.

Method	Implementation	Run time (sec)	Parameters (M)	Acc (%) ↑
GT	-	-	-	93.47
Blurred	-	-	-	77.05
Shen *et al.* [16]	MATLAB(GPU)	0.05	14.8	87.03
Lu *et al.* [61]	Pytorch(GPU)	0.02	53.0	80.56
Xia *et al.* [85]	Tensorflow(GPU)	0.19	41.8	89.12
Yasarla *et al.* [17]	Pytorch(GPU)	0.16	14.4	87.84
Lee *et al.* [18]	Pytorch(GPU)	0.08	18.5	89.59
Jung *et al.* [19]	Pytorch(GPU)	0.05	44.7	89.87
Ours	Pytorch(GPU)	0.05	44.7	**90.64**

this experimental section, those retrained models using the MSPL training set are marked with *. The official models of [18], [19] are not retrained because they are trained on the MSPL training set. All experiments are conducted with official codes provided by the authors. Note that we did not re-implement and retrain the model in [16] because the official training codes have not yet been released.

Table 1 reports the quantitative evaluation results on the MSPL-Center and MSPL Random test sets. The proposed SAPPGAN outperforms the state-of-the-art methods in terms of perceptual metrics, such as LPIPS, d_{VGG}, and d_{ARC}. Importantly, our proposed SAPPGAN achieves significant improvements in perceptual metrics over recent GAN-based SFID methods [16], [18], [19] that were developed to restore perceptually satisfactory images. In contrast to GAN-based SFID methods whose objective function primarily focuses on making the global decision of sharp face images with the data distribution of only sharp face images, the proposed SAPPGAN estimates the joint probability of the sharp face images and semantic label map of the faces and is able to provide pixel-level and global feedback to the generator. With this powerful capability, the proposed SAPPGAN is able to restore images that are perceptually outstanding.

The perceptual improvement of SAPPGAN is also noticeable in visual comparisons on the MSPL-Center (see Fig. 3) and MSPL-Random test sets (see Fig. 4). The resulting images of [17], which are not based on GANs, appear overly smooth and lack sharp details. Moreover, GAN-based models (*i.e.* Lee et al. [18], Jung et al. [19], and our SAPPGAN) outperform other methods in restoring realistic facial details. Among them, the proposed SAPPGAN significantly improves image quality with fine details and realistic textures. Specifically, SAPPGAN restores the main components (*i.e.* the eyes, nose, mouth and ears) of the face with high-fidelity textures (see 3^{rd}, 4^{th}, 8^{th} rows in Fig. 3 and 1^{st}, 3^{rd}, 7^{th} rows in Fig. 4 for eyes, 1^{st}, 6^{th}, 7^{th} rows in Fig. 3 and 1^{st}, 2^{nd}, 3^{rd} rows in Fig. 4 for nose, 3^{rd}, 6^{th}, 7^{th} rows in Fig. 3 and 3^{rd}, 5^{th}, 8^{th} rows in Fig. 4 for mouth/teeth and 5^{th}, 7^{th}, 8^{th} rows in Fig. 3 and 4^{th}, 8^{th} rows in Fig. 4 for ears). Moreover, the proposed SAPPGAN can generate realistic skin textures *i.e.* wrinkle and beard (see 1^{st}, 3^{rd}, 5^{th}, 6^{th}, $7th$ rows in Fig. 3 and 3^{rd}, 7^{th}, 8^{th} rows in Fig. 4 for wrinkle and 1^{st}, 2^{nd}, 8^{th} rows in Fig. 3 and 5^{th} row in Fig. 4 for beard). These semantic-aware deblurred results are attributed to the powerful and detailed feedback from the SAPP discriminator.

C. COMPARISONS ON REAL BLURRED IMAGES

Most existing SFID methods [16], [17], [18], [19], [61], including the proposed SAPPGAN, are trained with datasets that are degraded by synthetic blur kernels. However, SFID on real-world scenarios must consider more complex degradation factors, such as motion blur, sensor saturation, lens distortion, nonlinear transform functions, noise, and compression [39]. We conduct experiments on the real-world blurred images provided by [16], [39] to demonstrate the generalization capability of our proposed method on real-world SFID task. Since GT images do not exist for the real-world

TABLE 3. Effectiveness of different components of SAPPGAN on the MSPL-Center testset.

Model	Method				MSPL-Center			
	UD	SAPP	PW	C2F	$d_{ARC}\downarrow$	LPIPS $\downarrow$	PSNR $\uparrow$	SSIM $\uparrow$
S0					0.4909	0.0950	**28.9177**	**0.9297**
S1	✓				0.2844	0.0879	28.3405	0.9201
S2	✓		✓		0.2645	0.0825	28.3836	0.9242
S3		✓			0.2613	0.0824	28.3586	0.9239
S4		✓	✓		<u>0.2561</u>	<u>0.0816</u>	28.3951	0.9247
S5		✓	✓	✓	**0.2543**	**0.0799**	<u>28.4088</u>	<u>0.9251</u>

blurred images, qualitative results with competitive SFID methods [16], [17], [18], [19], [61] are shown in Fig. 5. The results of [16], [17], [61] are relatively smooth, whereas those of [18], [19] reconstruct sharper images. Compared with [18], [19], our method improves the restoration of the fine details and rich textures of the face, because it benefits from the proposed SAPPGAN.

D. EXECUTION TIME AND FACE VERIFICATION

Considering that SFID can be used in the preprocessing step of high-level face-related vision tasks (*i.e.* face recognition [5], [6], [7], [8]), SFID methods must enable the accurate recovery of the identity of the GT face. Therefore, we report the performance of face verification using deblurred images on the CelebA test set provided by [16]. For a fair comparison, we follow the evaluation setting in [19] and measure the estimated mean accuracy (Acc) [98]. In addition, we compare the inference time and the number of model parameters of the existing methods and proposed model. Following [16], [18], [19], the average inference time for 10 images is reported using a single NVIDIA Titan XP GPU. The spatial size of each image is $128 \times 128 \times 3$.

The experimental results are shown in Table 2. When comparing the verification Acc on the GT images and blurred images, it can be observed that Acc is remarkably degraded from 93.47% to 77.05% by blur artifacts. The Acc of the deblurred images using our method is 90.64%, which is the most comparable to the Acc of the GT images. In addition, Table 2 shows that our method maintains the parameters and inference time of the original model from [19]. This demonstrates that our SAPP discriminator and training method can be easily applied to other GAN-based SFID models, and it improves the reconstruction quality of the deblurring network without the additional load of parameters and inference time.

E. ABLATION STUDY

In this section, we conduct an ablation study to verify the effect of each component in our approach. Table 3 shows the brief configurations of each experiment and its quantitative results. Specifically, the baseline model, termed as S0 in Table 3, is set to the generator architecture of [19] and is trained with the sum of L_{pixel} (Eq. (13)) and L_{feat} (Eq. (14)) without adversarial loss. S1 is a model trained using the adversarial loss of the original U-Net discriminator [24], [25] in addition to the loss function of the S0 model. Compared to S0, the performance of S1 is increased in terms of d_{ARC} and LPIPS by $|0.4909 - 0.2844|/0.4909 = 42.07\%$ and $|0.0950 - 0.0879|/0.0950 = 7.47\%$, respectively. This verifies the effectiveness of per-pixel adversarial loss in the image deblurring task. When the proposed PW loss is involved (S2), our discriminator is trained to rapidly focus on difficult and misclassified examples. This allows the discriminator to be trained more accurately. Thus, the discriminator can provide more accurate adversarial feedback to the generator during training. This boosts the performance of S2 in terms of d_{ARC} and LPIPS by $|0.2844 - 0.2645|/0.2844 = 8.12\%$ and $|0.0879 - 0.0825|/0.0879 = 6.14\%$, respectively compared to S1. The effectiveness of the incorporating the SAPP discriminator instead of the U-Net discriminator is shown in S3 in Table 3. The results of S3 show that d_{ARC} is enhanced by $|0.2844 - 0.2613|/0.2844 = 8.12\%$ and the LPIPS is enhanced by $|0.0879 - 0.0824|/0.0879 = 6.26\%$, compared to S1. These improvements demonstrate that our key concept, *i.e.* conditional image restoration by forcing the discriminator network to estimate pixel-wise semantic-aware probability, is effective in face deblurring tasks. The results of S4 in Table 3 indicate that using the PW and SAPP discriminator together enables a better performance than using them separately. The d_{ARC} and LPIPS values are improved by $|0.2613 - 0.2561|/0.2613 = 1.99\%$ and $|0.0824 - 0.0816|/0.824 = 0.97\%$, respectively, compared to those of S3. The final model of our method (S5 in Table 3) outperforms S4 by $|0.2561 - 0.2543|/0.2561 = 0.70\%$ and $|0.0816 - 0.0799|/0.0816 = 2.08\%$ in terms of d_{ARC} and LPIPS, respectively. These results demonstrate the effectiveness of the proposed coarse-to-fine training scheme (noted as C2F), which allows our generator to focus first on the global consistency of the restored image and then on the local consistency.

V. CONCLUSION

This paper presents a semantic-aware pixel-wise projection (SAPP) GAN, a novel GAN-based framework for single face image deblurring. The proposed SAPP discriminator

is designed to incorporate a label matching (conditional) distribution into an image (marginal) distribution using a pixel-wise projection technique. This approach enables our discriminator to focus on the realness of the restored face by taking into account semantically important information. Furthermore, our SAPP discriminator provides global (per-image) and local (per-pixel) feedback to the generator by adopting a U-Net-like architecture. In addition, our discriminator can be trained more accurately with the proposed PW loss, which dynamically weights the incorrect predictions of the discriminator on a pixel-by-pixel basis. The generator is effectively trained through the proposed coarse-to-fine training technique to balance adversarial feedback between the global and local decisions of the discriminator. Overall, the proposed SAPPGAN improves on recent face image deblurring methods in terms of image perceptual quality. We believe that our SAPPGAN framework can be applied to various fields of face image restoration.

REFERENCES

[1] J. Li, Y. Wang, C. Wang, Y. Tai, J. Qian, J. Yang, C. Wang, J. Li, and F. Huang, "DSFD: Dual shot face detector," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 5060–5069.

[2] C. Chi, S. Zhang, J. Xing, Z. Lei, S. Z. Li, and X. Zou, "Selective refinement network for high performance face detection," in *Proc. AAAI Conf. Artif. Intell.*, Jan. 2019, vol. 33, no. 1, pp. 8231–8238.

[3] X. Tang, D. K. Du, Z. He, and J. Liu, "PyramidBox: A context-assisted single shot face detector," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*. Cham, Switzerland: Springer, Sep. 2018, pp. 797–813.

[4] K. Zhang, Z. Zhang, Z. Li, and Y. Qiao, "Joint face detection and alignment using multitask cascaded convolutional networks," *IEEE Signal Process. Lett.*, vol. 23, no. 10, pp. 1499–1503, Oct. 2016.

[5] P. Terhörst, M. Ihlefeld, M. Huber, N. Damer, F. Kirchbuchner, K. Raja, and A. Kuijper, "QMagFace: Simple and accurate quality-aware face recognition," 2021, *arXiv:2111.13475*.

[6] F. Boutros, N. Damer, F. Kirchbuchner, and A. Kuijper, "ElasticFace: Elastic margin loss for deep face recognition," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Jun. 2022, pp. 1578–1587.

[7] I. Kim, S. Han, S.-J. Park, J.-W. Baek, J. Shin, J.-J. Han, and C. Choi, "DiscFace: Minimum discrepancy learning for deep face recognition," in *Proc. Asian Conf. Comput. Vis. (ACCV)*, Nov. 2020, pp. 1–17.

[8] B. Li, T. Xi, G. Zhang, H. Feng, J. Han, J. Liu, E. Ding, and W. Liu, "Dynamic class queue for large scale face recognition in the wild," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 3763–3772.

[9] S. Chen, C. Zhang, M. Dong, J. Le, and M. Rao, "Using ranking-CNN for age estimation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 5183–5192.

[10] J. Wan, Z. Tan, G. Guo, S. Z. Li, and Z. Lei, "Auxiliary demographic information assisted age estimation with cascaded structure," *IEEE Trans. Cybern.*, vol. 48, no. 9, pp. 2531–2541, Sep. 2018.

[11] A. S. Al-Shannaq and L. A. Elrefaei, "Comprehensive analysis of the literature for age estimation from facial images," *IEEE Access*, vol. 7, pp. 93229–93249, 2019.

[12] Y. Cao, D. Berend, P. Tolmach, G. Amit, M. Levy, Y. Liu, A. Shabtai, and Y. Elovici, "Fair and accurate age prediction using distribution aware data curation and augmentation," in *Proc. IEEE/CVF Winter Conf. Appl. Comput. Vis. (WACV)*, Jan. 2022, pp. 3551–3561.

[13] G. G. Chrysos, P. Favaro, and S. Zafeiriou, "Motion deblurring of faces," *Int. J. Comput. Vis.*, vol. 127, nos. 6–7, pp. 801–823, Jun. 2019.

[14] S. Lin, J. Zhang, J. Pan, Y. Liu, Y. Wang, J. Chen, and J. Ren, "Learning to deblur face images via sketch synthesis," in *Proc. AAAI Conf. Artif. Intell.*, vol. 34, Apr. 2020, pp. 11523–11530.

[15] W. Ren, J. Yang, S. Deng, D. Wipf, X. Cao, and X. Tong, "Face video deblurring using 3D facial priors," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 9388–9397.

[16] Z. Shen, W.-S. Lai, T. Xu, J. Kautz, and M.-H. Yang, "Deep semantic face deblurring," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 8260–8269.

[17] R. Yasarla, F. Perazzi, and V. M. Patel, "Deblurring face images using uncertainty guided multi-stream semantic networks," *IEEE Trans. Image Process.*, vol. 29, pp. 6251–6263, 2020.

[18] T. B. Lee, S. H. Jung, and Y. S. Heo, "Progressive semantic face deblurring," *IEEE Access*, vol. 8, pp. 223548–223561, 2020.

[19] S. H. Jung, T. Bok Lee, and Y. S. Heo, "Deep feature prior guided face deblurring," in *Proc. IEEE/CVF Winter Conf. Appl. Comput. Vis. (WACV)*, Jan. 2022, pp. 3531–3540.

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

[21] M. Mirza and S. Osindero, "Conditional generative adversarial nets," 2014, *arXiv:1411.1784*.

[22] J. T. Springenberg, "Unsupervised and semi-supervised learning with categorical generative adversarial networks," 2015, *arXiv:1511.06390*.

[23] V. Sushko, E. Schönfeld, D. Zhang, J. Gall, B. Schiele, and A. Khoreva, "You only need adversarial supervision for semantic image synthesis," 2020, *arXiv:2012.04781*.

[24] E. Schonfeld, B. Schiele, and A. Khoreva, "A U-Net based discriminator for generative adversarial networks," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 8207–8216.

[25] Y. Jo, S. Yang, and S. J. Kim, "Investigating loss functions for extreme super-resolution," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Jun. 2020, pp. 424–425.

[26] E. Schönfeld, V. Sushko, D. Zhang, J. Gall, B. Schiele, and A. Khoreva, "You only need adversarial supervision for semantic image synthesis," in *Proc. Int. Conf. Learn. Represent. (ICLR)*, May 2021, pp. 2903–2923.

[27] C. H. Lin, C.-C. Chang, Y.-S. Chen, D.-C. Juan, W. Wei, and H.-T. Chen, "COCO-GAN: Generation by parts via conditional coordinating," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 4512–4521.

[28] T. Miyato and M. Koyama, "CGANs with projection discriminator," 2018, *arXiv:1802.05637*.

[29] A. Odena, C. Olah, and J. Shlens, "Conditional image synthesis with auxiliary classifier GANs," in *Proc. Int. Conf. Mach. Learn.*, Jul. 2017, pp. 2642–2651.

[30] M. Gong, Y. Xu, C. Li, K. Zhang, and K. Batmanghelich, "Twin auxiliary classifiers GAN," in *Proc. Adv. Neural Inf. Process. Syst. (NIPS)*, vol. 32, Dec. 2019, p. 1328.

[31] M. Kang and J. Park, "ContraGAN: Contrastive learning for conditional image generation," in *Proc. Adv. Neural Inf. Process. Syst. (NIPS)*, Dec. 2020, pp. 21357–21369.

[32] M. Kang, W. Shim, M. Cho, and J. Park, "Rebooting ACGAN: Auxiliary classifier GANs with stable training," 2021, *arXiv:2111.01118*.

[33] T. Miyato, T. Kataoka, M. Koyama, and Y. Yoshida, "Spectral normalization for generative adversarial networks," 2018, *arXiv:1802.05957*.

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

[35] L. Han, M. R. Min, A. Stathopoulos, Y. Tian, R. Gao, A. Kadav, and D. Metaxas, "Dual projection generative adversarial networks for conditional image generation," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 14438–14447.

[36] I. Goodfellow, "NIPS 2016 tutorial: Generative adversarial networks," 2017, *arXiv:1701.00160*.

[37] T. Salimans, I. Goodfellow, W. Zaremba, V. Cheung, A. Radford, and X. Chen, "Improved techniques for training GANs," in *Proc. Adv. Neural Inf. Process. Syst. (NIPS)*, vol. 29, Dec. 2016, pp. 2234–2242.

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

[39] W.-S. Lai, J.-B. Huang, Z. Hu, N. Ahuja, and M.-H. Yang, "A comparative study for single image blind deblurring," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 1701–1709.

[40] S. Nah, S. Son, J. Lee, and K. M. Lee, "Clean images are hard to reblur: Exploiting the ill-posed inverse task for dynamic scene deblurring," in *Proc. Int. Conf. Learn. Represent.*, Apr. 2021, pp. 1–19.

[41] S. Cho and S. Lee, "Fast motion deblurring," *ACM Trans. Graph.*, vol. 28, no. 5, pp. 1–8, Dec. 2009.

[42] D. Krishnan and R. Fergus, "Fast image deconvolution using hyper-Laplacian priors," in *Proc. Adv. Neural Inf. Process. Syst.*, Jun. 2009, pp. 1033–1041.

[43] D. Krishnan, T. Tay, and R. Fergus, "Blind deconvolution using a normalized sparsity measure," in *Proc. IEEE CVPR*, Jun. 2011, pp. 233–240.

[44] L. Xu, S. Zheng, and J. Jia, "Unnatural L_0 sparse representation for natural image deblurring," in *Proc. IEEE CVPR*, Jun. 2013, pp. 1107–1114.

[45] J. Pan, Z. Hu, Z. Su, and M.-H. Yang, "Deblurring text images via L_0-regularized intensity and gradient prior," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Columbus, OH, USA, Jun. 2014, pp. 2901–2908.

[46] A. Levin, Y. Weiss, F. Durand, and W. T. Freeman, "Understanding and evaluating blind deconvolution algorithms," in *Proc. IEEE CVPR*, Jun. 2009, pp. 1964–1971.

[47] A. Levin, Y. Weiss, F. Durand, and W. T. Freeman, "Efficient marginal likelihood optimization in blind deconvolution," in *Proc. IEEE CVPR*, Jun. 2011, pp. 2657–2664.

[48] H. Zhang, D. Wipf, and Y. Zhang, "Multi-image blind deblurring using a coupled adaptive sparse prior," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2013, pp. 1051–1058.

[49] L. Sun, S. Cho, J. Wang, and J. Hays, "Edge-based blur kernel estimation using patch priors," in *Proc. IEEE ICCP*, Apr. 2013, pp. 1–8.

[50] J. Pan, D. Sun, M.-H. Yang, and H. Pfister, "Blind image deblurring using dark channel prior," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 1628–1636.

[51] D. Gong, J. Yang, L. Liu, Y. Zhang, I. Reid, C. Shen, A. Van Den Hengel, and Q. Shi, "From motion blur to motion flow: A deep learning solution for removing heterogeneous motion blur," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 2319–2328.

[52] X. Tao, H. Gao, X. Shen, J. Wang, and J. Jia, "Scale-recurrent network for deep image deblurring," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2018, pp. 8174–8182.

[53] A. Chakrabarti, "A neural approach to blind motion deblurring," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*. Cham, Switzerland: Springer, Sep. 2016, pp. 221–235.

[54] J. Sun, W. Cao, Z. Xu, and J. Ponce, "Learning a convolutional neural network for non-uniform motion blur removal," in *Proc. IEEE CVPR*, Jun. 2015, pp. 769–777.

[55] S. Nah, T. H. Kim, and K. M. Lee, "Deep multi-scale convolutional neural network for dynamic scene deblurring," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 3883–3891.

[56] D. Park, D. U. Kang, J. Kim, and S. Y. Chun, "Multi-temporal recurrent neural networks for progressive non-uniform single image deblurring with incremental temporal training," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*. Cham, Switzerland: Springer, Oct. 2020, pp. 327–343.

[57] H. Zhang, Y. Dai, H. Li, and P. Koniusz, "Deep stacked hierarchical multi-patch network for image deblurring," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 5978–5986.

[58] M. Suin, K. Purohit, and A. N. Rajagopalan, "Spatially-attentive patch-hierarchical network for adaptive motion deblurring," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 3606–3615.

[59] S. W. Zamir, A. Arora, S. Khan, M. Hayat, F. S. Khan, M.-H. Yang, and L. Shao, "Multi-stage progressive image restoration," 2021, *arXiv:2102.02808*.

[60] S.-J. Cho, S.-W. Ji, J.-P. Hong, S.-W. Jung, and S.-J. Ko, "Rethinking coarse-to-fine approach in single image deblurring," 2021, *arXiv:2108.05054*.

[61] B. Lu, J.-C. Chen, and R. Chellappa, "Unsupervised domain-specific deblurring via disentangled representations," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2019, pp. 10225–10234.

[62] J. Pan, Z. Hu, Z. Su, and M.-H. Yang, "Deblurring face images with exemplars," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*. Cham, Switzerland: Springer, Sep. 2014, pp. 47–62.

[63] K. Grm, W. J. Scheirer, and V. Struc, "Face hallucination using cascaded super-resolution and identity priors," *IEEE Trans. Image Process.*, vol. 29, pp. 2150–2165, 2020.

[64] G. G. Chrysos and S. Zafeiriou, "Deep face deblurring," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Jul. 2017, pp. 69–78.

[65] Y. Hacohen, E. Shechtman, and D. Lischinski, "Deblurring by example using dense correspondence," in *Proc. IEEE Int. Conf. Comput. Vis.*, Dec. 2013, pp. 2384–2391.

[66] J.-Y. Zhu, T. Park, P. Isola, and A. A. Efros, "Unpaired image-to-image translation using cycle-consistent adversarial networks," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 2223–2232.

[67] X. Huang and S. Belongie, "Arbitrary style transfer in real-time with adaptive instance normalization," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Dec. 2017, pp. 1501–1510.

[68] F. Luan, S. Paris, E. Shechtman, and K. Bala, "Deep photo style transfer," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 4990–4998.

[69] H. Zhang and K. Dana, "Multi-style generative network for real-time transfer," in *Proc. Eur. Conf. Comput. Vis. (ECCV) Workshops*, Sep. 2018, pp. 990–4998.

[70] H. Chen, L. Zhao, Z. Wang, H. Zhang, Z. Zuo, A. Li, W. Xing, and D. Lu, "DualAST: Dual style-learning networks for artistic style transfer," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 872–881.

[71] C. Ledig, L. Theis, F. Huszar, J. Caballero, A. Cunningham, A. Acosta, A. Aitken, A. Tejani, J. Totz, Z. Wang, and W. Shi, "Photo-realistic single image super-resolution using a generative adversarial network," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 4681–4690.

[72] X. Wang, K. Yu, S. Wu, J. Gu, Y. Liu, C. Dong, Y. Qiao, and C. C. Loy, "ESRGAN: Enhanced super-resolution generative adversarial networks," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, Oct. 2018, pp. 1–16.

[73] X. Wang, L. Xie, C. Dong, and Y. Shan, "Real-ESRGAN: Training real-world blind super-resolution with pure synthetic data," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. Workshops (ICCVW)*, Oct. 2021, pp. 1905–1914.

[74] A. Sauer, K. Chitta, J. Müller, and A. Geiger, "Projected GANs converge faster," in *Proc. Adv. Neural Inf. Process. Syst. (NIPS)*, vol. 34, Dec. 2021, pp. 17480–17492.

[75] J. Wu, Z. Huang, D. Acharya, W. Li, J. Thoma, D. P. Paudel, and L. V. Gool, "Sliced Wasserstein generative models," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 3713–3722.

[76] M. Arjovsky, S. Chintala, and L. Bottou, "Wasserstein generative adversarial networks," in *Proc. Int. Conf. Mach. Learn. (ICML)*, Aug. 2017, pp. 214–223.

[77] H. Zhang, I. Goodfellow, D. Metaxas, and A. Odena, "Self-attention generative adversarial networks," in *Proc. ICML*, Feb. 2019, pp. 7354–7363.

[78] T. R. Shaham, T. Dekel, and T. Michaeli, "SinGAN: Learning a generative model from a single natural image," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, Oct. 2019, pp. 4570–4580.

[79] X. Chen, Y. Duan, R. Houthooft, J. Schulman, I. Sutskever, and P. Abbeel, "InfoGAN: Interpretable representation learning by information maximizing generative adversarial nets," in *Proc. Adv. Neural Inf. Process. Syst. (NIPS)*, vol. 29, Dec. 2016, pp. 4570–4580.

[80] D. Hong, J. Yao, D. Meng, Z. Xu, and J. Chanussot, "Multimodal GANs: Toward crossmodal hyperspectral–multispectral image segmentation," *IEEE Trans. Geosci. Remote Sens.*, vol. 59, no. 6, pp. 5103–5113, Jun. 2020.

[81] J. Yao, D. Hong, J. Chanussot, D. Meng, X. Zhu, and Z. Xu, "Cross-attention in coupled unmixing nets for unsupervised hyperspectral super-resolution," in *Proc. Eur. Conf. Comput. Vis.* Cham, Switzerland: Springer, Aug. 2020, pp. 208–224.

[82] Z. Huang, J. Zhang, Y. Zhang, and H. Shan, "DU-GAN: Generative adversarial networks with dual-domain U-Net-based discriminators for low-dose CT denoising," *IEEE Trans. Instrum. Meas.*, vol. 71, pp. 1–12, 2022.

[83] T.-Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollár, "Focal loss for dense object detection," in *Proc. IEEE Int. Conf. Comput. Vis.*, Oct. 2017, pp. 2980–2988.

[84] O. M. Parkhi, A. Vedaldi, and A. Zisserman, "Deep face recognition," in *Proc. Brit. Mach. Vis. Conf.*, Swansea, U.K., Sep. 2015, pp. 1–12.

[85] Z. Xia and A. Chakrabarti, "Training image estimators without image ground truth," in *Proc. Adv. Neural Inf. Process. Syst. (NIPS)*, Jun. 2019, pp. 2436–2446.

[86] C.-H. Lee, Z. Liu, L. Wu, and P. Luo, "MaskGAN: Towards diverse and interactive facial image manipulation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 5549–5558.

[87] P. Luo, X. Wang, and X. Tang, "Hierarchical face parsing via deep learning," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2012, pp. 2480–2487.

[88] J. Lin, H. Yang, D. Chen, M. Zeng, F. Wen, and L. Yuan, "Face parsing with RoI Tanh-warping," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2019, pp. 5654–5663.
[89] Q. Zheng, J. Deng, Z. Zhu, Y. Li, and S. Zafeiriou, "Decoupled multi-task learning with cyclical self-regulation for face parsing," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2022, pp. 4156–4165.
[90] 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.
[91] T. Karras, S. Laine, and T. Aila, "A style-based generator architecture for generative adversarial networks," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 4401–4410.
[92] A. Paszke et al., "PyTorch: An imperative style, high-performance deep learning library," in *Proc. Adv. Neural Inf. Process. Syst.*, Jun. 2019, pp. 8026–8037.
[93] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," 2014, *arXiv:1412.6980*.
[94] X. Wang, Y. Li, H. Zhang, and Y. Shan, "Towards real-world blind face restoration with generative facial prior," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 9168–9178.
[95] J. Deng, J. Guo, N. Xue, and S. Zafeiriou, "ArcFace: Additive angular margin loss for deep face recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2019, pp. 4690–4699.
[96] R. Zhang, P. Isola, A. A. Efros, E. Shechtman, and O. Wang, "The unreasonable effectiveness of deep features as a perceptual metric," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 586–595.
[97] 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.
[98] G. B. Huang, M. Ramesh, T. Berg, and E. Learned-Miller, "Labeled faces in the wild: A database for studying face recognition in unconstrained environments," Univ. Massachusetts, Amherst, MA, USA, Tech. Rep., 07-49, Oct. 2007.

SUJY HAN received the B.S. degree from the Department of Electrical and Computer Engineering, Ajou University, Suwon, South Korea, in 2021, where she is currently pursuing the M.S. degree with the Department of Artificial Intelligence. Her research interests include computer vision, deep learning, image restoration, and image generation.

TAE BOK LEE received the B.S. degree in electrical and computer engineering from Ajou University, Suwon, South Korea, in 2018, where he is currently pursuing the integrated M.S. and Ph.D. degrees with the Department of Artificial Intelligence. His research interests include computer vision, deep learning, and image restoration.

YONG SEOK HEO received the B.S. degree in electrical engineering and the M.S. and Ph.D. degrees in electrical engineering and computer science from Seoul National University, South Korea, in 2005, 2007, and 2012, respectively. From 2012 to 2014, he was at the Digital Media and Communications Research and Development Center, Samsung Electronics. He is currently with the Department of Electrical and Computer Engineering and the Department of Artificial Intelligence, Ajou University, as an Associate Professor. His research interests include segmentation, stereo matching, 3D reconstruction, and computational photography.