Progressive Semantic Face Deblurring

TAE BOK LEE¹, SOO HYUN JUNG², AND YONG SEOK HEO², (Member, IEEE)

¹Department of Artificial Intelligence, Ajou University, Suwon 443-749, South Korea
²Department of Electrical and Computer Engineering, Ajou University, Suwon 443-749, South Korea

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

This work was supported by the Ministry of Science and ICT (MSIT), South Korea, under the Information Technology Research Center (ITRC) Support Program supervised by the Institute for Information and Communications Technology Promotion (IITP) under Grant IITP-2020-2018-0-01424.

ABSTRACT Previous face deblurring methods have utilized semantic segmentation maps as prior knowledge. Most of these methods generated the segmentation map from a blurred facial image, and restore it using the map in a sequential manner. However, the accuracy of the segmentation affects the restoration performance. Generally, it is difficult to obtain an accurate segmentation map from a blurred image. Instead of sequential methods, we propose an efficient method that learns the flows of facial component restoration without performing segmentation. To this end, we propose a multi-semantic progressive learning (MSPL) framework that progressively restores the entire face image starting from the facial components such as the skin, followed by the hair, and the inner parts (eyes, nose, and mouth). Furthermore, we propose a discriminator that observes the reconstruction-flow of the generator. In addition, we present new test datasets to facilitate the comparison of face deblurring methods. Various experiments demonstrate that the proposed MSPL framework achieves higher performance in facial image deblurring compared to the existing methods, both qualitatively and quantitatively. Our code, trained model and data are available at https://github.com/dolphin0104/MSPL-GAN.

INDEX TERMS Facial image deblurring, semantic mask, progressive learning, generative adversarial network, deep learning.

I. INTRODUCTION Facial image deblurring is to restore a blurred face image as a sharp face image. Although human faces are highly variable, they have hierarchical structures that comprise components, such as skin, hair, eyes, nose, and mouth. These facial components are the crucial elements that characterize a specific face; each element has inherent shapes and textures. Thus, most face deblurring methods utilize the prior knowledge (i.e., reference face [1], [2], 3D face [3], face landmark [4], [5], 2D face sketch [6] and semantic segmentation map [7], [8]) of face images to estimate a unique solution.

Recently, deep learning-based methods [7], [8] have achieved state-of-the-art performances in facial image deblurring by utilizing the semantic segmentation map as prior knowledge. These methods consist of a two-step process that generates the segmentation map from the blurred image, and restores the facial image using this segmentation map in a sequential manner. In these methods, the semantic segmentation map is employed to localize the position of each facial component and the boundaries between them for the deblurring process.

However, these methods have a limitation. Accuracy of the estimated segmentation map affects the restoration performance due to their sequential properties. In general, it is nontrivial to obtain an accurate segmentation map from a blurred image. Inaccurate segmentation results often lead to inaccurate localization of facial components, and consequently generate a deblurred result with distorted shapes and/or blurred textures [3], [9].

Specifically, small components of the face (that is, the eyebrows, eyes, nose, and lips) which characterize the faces, are significantly affected by inaccurate segmentation compared to the large components such as hair and skin. In general, each facial component is different in size, and has large deviations. For instance, the eyes have a very small number of pixels compared to the skin or hair regions. Moreover, these small components of the face are more likely to lose information due to noise and blur artifacts than the large components. Due to this degradation, it is difficult to obtain accurate semantic segmentation maps especially
for small components from the blurred images. Thus, small components require more attention than the large components for the restoration of their exact shapes and textures. This problem, called the class imbalance in Yasarla et al. [8], is one of the major challenges in the previous method [7], [8].

As investigated in [3], [8], Shen et al. [7] often fails to restore the small components of the face when the generated segmentation results are inaccurate. To address this problem, Yasarla et al. [8] proposed measuring the confidence score of the facial components from the generated segmentation. If the estimated segmentation maps have low confidence, their model reduces the impact of segmentation maps in the deblurring process. This method can effectively reduce the effects of inaccurate segmentation maps. However, their solution is suboptimal, because they do not provide how to utilize the semantic prior when the segmentation map is inaccurate due to severe blurs.

To deal with this problem, we propose a multi-semantic progressive learning (MSPL) framework based on the generative adversarial network (GAN) [10]. Our method leverages the semantic prior information of the face without performing segmentation to prevent the side effects of an inaccurate segmentation map. Furthermore, our method progressively restores the face within four steps, inspired by the success of progressive learning techniques [11], [12]. Conventional progressive learning [11], [12] does not consider the semantic context of the target object in an image. Therefore, we modified the concept of the conventional coarse-to-fine approach to understand the underlying semantic structure of the target object better.

Specifically, the proposed generator network has a cascaded architecture with sub-networks to restore the entire face progressively and incrementally, starting with the simpler facial components. Our network is trained to focus on low-frequency components first and then incrementally restore the smaller and high-frequency components in the face image, instead of learning all the components of the face image simultaneously. During training, each sub-network focuses on restoring both the shape and texture for their assigned class-specific facial components. This is achieved by minimizing the difference between the sharp and output facial components using masks obtained by precise ground-truth segmentation maps. In addition, the architecture of our generator mitigates the class-imbalance problem. The proposed generator consists of multiple sub-networks. Each sub-network is trained to focus on restoring the assigned facial key component. This simple method allows the proposed method to handle the class imbalance problem of face deblurring more effectively. Fig. 1 clearly shows the effects of the proposed framework. Our MPSL framework restores a sharper face with fine-detailed facial components compared to the previous methods [7], [8].

To generate facial images that are more photo-realistic, we propose a multi-semantic discriminator in our GAN framework. It is designed to handle all the intermediate outputs of the generator using a single classifier network by allowing the flow of gradients at all semantic components in the discriminator to the generator. Through this, our discriminator oversees the entire flow of reconstructions of the facial components.

To the best of our knowledge, there are only a few public test datasets available for facial image deblurring. Shen et al. [7] provided a pioneering test dataset for evaluation. However, most of the provided images are of low quality with unknown blocking artifacts. In addition, all of the test faces are well aligned and centered with the same facial key points. However, in the real-world scenarios, faces are captured in various shapes and poses. Thus, these images are not appropriate to fully evaluate various methods for a range of cases; in view of this, we provide new test datasets that are suitable for a more practical evaluation of face deblurring.

The contributions of our work are summarized as follows:

- We propose an MSPL network, which progressively learns the semantics of a human face for deblurring the facial images from complex motion blur. To the best of our knowledge, this is the first time that the idea of a semantic coarse-to-fine manner has been introduced in face deblurring.
- We propose the Multi-Semantic discriminator, which can observe all the outputs of a generator. From this,
our generator reconstructs more photo-realistic face components in addition to the entire face.

- To conduct a more accurate and practical evaluation of face deblurring, we suggest new test datasets with extensive and high-quality images. The experimental results show that the proposed model significantly outperforms the previous methods.

The remainder of this paper is organized as follows. Section II provides an overview of the previous works on image deblurring. In Section III, we present the details of our proposed framework. Section IV provides the quantitative and qualitative results of the proposed method compared to the existing methods. Finally, Section V presents the conclusions and discusses the future works.

II. RELATED WORK

Image deblurring has been studied for a long time in the field of image processing and computer vision. In this section, we briefly review the image deblurring methods and recent deep learning-based progressive learning approaches.

A. GENERIC IMAGE DEBLURRING

Single image deblurring is a highly ill-posed problem; thus, the traditional deblurring methods utilize natural image priors, such as sparsity priors [13], [14], $L_0$ gradient priors [15], low-rank priors [16], patch priors [17], [18] and channel priors [19], [20]. Although these handcrafted priors suffice for a small subset of the scene, it is difficult to apply them to real-world images with complex blurs.

After the advent of deep learning [21], various convolutional neural network (CNN) models have been proposed to estimate the complex blur kernels [22], [23]. Sun et al. [22] proposed to predict the probabilistic distribution of motion blur at patch level. Chakrabarti et al. [23] predicted the complex Fourier coefficients of a deconvolution filter and applied them to an input patch. These methods combine the CNNs and maximum a posteriori probability (MAP)-based algorithms. On the other hand, several CNN models directly restore the sharp image from blurred image in an end-to-end manner [24]–[29]. Nah et al. [24] proposed a multi-scale CNN model. They first extended the traditional coarse-to-fine pipeline to a CNN-based deblurring field and achieved impressive results. Tao et al. [25] investigated a multi-scale strategy for the recurrent neural network (RNN)-based multi-scale architecture. To restore the realistic images, Kupyn et al. [30] introduced a GAN-based deblurring model that exploited Wasserstein GAN with a gradient penalty and perceptual loss.

B. FACE IMAGE DEBLURRING

While the aforementioned methods perform well for natural image deblurring, they often do not perform satisfactorily on domain-specific images such as face images. Therefore, several studies have proposed estimating various types of prior facial knowledge such as the face alignment [4], face sketches [6], reference faces [2], [31], 3D face models [3], and face segmentation maps [7], [8]. The reference prior-based methods [2], [31] extract useful information to restore the face image from a sharp face similar to a degraded face image. However, these methods require a redundant collecting and matching computation process to utilize the exemplar face images. Ren et al. [3] proposed a video deblurring method for faces by generating 3D facial priors. They trained the 3D face reconstruction network to estimate more textured facial priors. Despite satisfactory result on video deblurring, their model was unable to perform on a single image. More recently, Shen et al. [7] and Yasarla et al. [8] proposed the use of semantic prior of the face for the deblurring process and achieved state-of-the-art results. These methods are two-step process. The steps include generating the semantic labels from the blurred face first, and then using them as strong prior knowledge for the deblurring process. However, extracting the segmentation map from the blurred face is difficult, and the erroneous prior information directly degrades the quality of the reconstructed face image. To reduce the side effects of inaccurate segmentation maps, Yasarla et al. [8] proposed measuring the confidence score of an estimated semantic map. However, they do not provide how to utilize the semantic prior when the segmentation map is inaccurate due to severe blurs.

Unlike previous works [7], [8], the proposed method exploits only the ground-truth segmentation maps for training purposes, instead of generating them from blurred images. Using this procedure, our method can be trained with accurate segmentation maps, regardless of the degree of blurs, and prevent the side effects of inaccurate segmentation maps. In addition, the architecture of the proposed generator allows the restoration of the small components of the face more effectively.

C. PROGRESSIVE LEARNING

Progressive learning is a training strategy that involves starting with an easy task and gradually refining the details. Most existing methods that use progressive learning are based on a multi-scale (coarse-to-fine) approach. Multi-scale frameworks have made significant progress in estimating complex motion blur kernels in single image deblurring [13], [24], [25], [32], [33]. In addition to the single image deblurring field, the multi-scale approach is widely used in other image processing fields such as depth map estimation [34]–[36], and video frame prediction [37]. In recent years, progressive learning [11], [12], [20], [38]–[42] has been actively applied to CNN-based image synthesis. In particular, Karras et al. [11] proposed a progressive growing technique that progressively increases the depth of the layer as well as the resolution of the generated image. Karnewar et al. [12] proposed the multi-scale gradient generative adversarial network (MSG-GAN) by allowing the flow of gradients from the discriminator to the generator at multi-scales. Meanwhile, Yang et al. [43] proposed a method to generate the background and foreground recursively and separately. In contrast
to the conventional methods, we suggest progressive learning techniques according to the semantic information of the face.

III. PROPOSED METHOD

As illustrated in Fig. 2, our MSPL framework is composed of a generator $(G)$ and a discriminator $(D)$. $G$ generates a sharp face image $\hat{I}_{\text{deblur}}$ from a blurred face image $I_{\text{blur}}$. Our proposed $G$ incrementally generates the facial components step-by-step in the order of skin, hair, inner parts (eyes, nose, and mouth), and then the entire face. Meanwhile, the proposed discriminator $(D)$ oversees all the generated face image components from the $G$. A detailed explanation of each network is provided in the following subsections.

A. SEMANTIC PROGRESSIVE GENERATOR

Following Yasarla et al. [8], we divide the ground-truth segmentation labels into three classes as follows: $M_1 = \{\text{skin}\}$, $M_2 = \{\text{hair}\}$, $M_3 = \{\text{nose, eyes, eyebrows, ears, mouth, lip}\}$, and $M_4 = \{\text{entire}\}$. Here, $M_i$ is a binary mask image with 1 for the assigned region, and 0 for other regions. Note that $M_4$ represents the entire area of the image, including the face and background. From Fig. 2, it can be observed that $G$ consists of multiple functional blocks as $g_i$ and $t$, where $1 \leq i \leq 4$. First, $h$ is an initial $1 \times 1$ convolution layer that converts an input RGB image $I_{\text{blur}}$ to a feature map $F_{\text{init}}$, for the following layer $g_1$. Thus, $h$ can be defined as $h : I_{\text{blur}} \rightarrow F_{\text{init}}$.

Second, $g_i$ is a function of our sub-network defined by

$$g_i : F_{i-1} \rightarrow F_i, \quad \text{where } 1 \leq i \leq 4, \quad F_0 = F_{\text{init}}. \quad (1)$$

Finally, $t$ is a $1 \times 1$ convolution layer that converts the output feature map $F_i$ generated from $g_i$ to an output RGB image $O_i$, as $t : F_i \rightarrow O_i$. Then, the $O_i$ can be defined as

$$O_i = t(g_i(F_{i-1})) = t(F_i), \quad \text{where } 1 \leq i \leq 4. \quad (2)$$

Thus, the entire network $G$ can be defined as a sequential composition of all sub-networks as

$$G : I_{\text{blur}} \mapsto \{O_1, O_2, O_3, O_4\}. \quad (3)$$

In our framework, $g_i$ is the key module of our network that focuses on restoring each semantic structure of the face. Each $g_i$ shares the same network architectures; however, their roles are different as each $g_i$ renders each facial component using the previous feature maps generated from the $g_{i-1}$ layer. For this, we design each $g_i$ as a fully convolutional U-shaped network, which consists of residual blocks [44]. As investigated in [45], [46], we remove the normalization layers from the standard residual blocks because the normalization layers get rid of flexibility from the network for low-level tasks. To extract more focused features, we apply a channel-attention mechanism [47], [48] to our residual blocks. The entire architecture of $g_i$ is shown in Table 1. In Table 1, each row of the “Kernel” column specifies the kernel size and, the number of filters and strides. For example, “$3 \times 3, 64, s1$” represents the 64 filters of size $3 \times 3$, with stride 1.

Our goal is to train each $g_i$ to reconstruct the assigned facial component perfectly. For this, we define the facial component loss ($L_i^c$), which is defined as $L_1$ distance between the facial components of the ground truth (GT) images and those generated from $g_i$ as

$$L_i^c = \|I_{\text{GT}} \odot M_i - (O_i \odot M_i)\|_1, \quad (4)$$

where $\odot$ represents the Hadamard product. To refine the entire face more naturally and restore the background of the face, we compare the last output image $O_4$, and the entire target image $I_{\text{GT}}$. This allows all sub-networks to share a common objective and provide stability during training.

FIGURE 2. Overview of the proposed face deblurring framework comprising of a generator and a discriminator. The generator reconstructs the image in four steps, and the discriminator observes all the output images from the generator.
The total facial component loss of our $G$ ($L_G$) can be formulated as follows:

$$L_G = \sum_{i=1}^{4} L_c^i. \quad (5)$$

Our proposed objective function in Eq. (5) allows a single $g_i$ to focus on the specified facial component using only ground-truth segmentation maps. Thus, our $G$ is able to reconstruct more precise shapes and finer details of the target face without suffering from the side effects of using an inaccurate segmentation map.

### B. MULTI-SEMANTIC DISCRIMINATOR

We also propose a multi-semantic discriminator $D$ in our method. Inspired by the MSG-GAN \[12\], our $D$ handles multiple outputs of $G$ which allows the restoration of more realistic facial components at all intermediate layers. As shown in Fig. 2, multiple intermediate images are fed to our single $D$. Thus, a single network $D$ is a function of multiple input images and predicts a final probability $p \in \{0, 1\}$ as

$$D : \{x_1, x_2, x_3, x_4\} \mapsto p, \quad (6)$$

where $x_j$ is a $j^{th}$ input RGB image of $D$. Let $d_j$ be the $j^{th}$ intermediate layer of $D$, and let $A_j$ be an output feature map of $d_j$. Then, $d_j$ can be defined as

$$d_j : (x_j, A_{j-1}) \mapsto A_j, \text{ where } 1 \leq j \leq 4, A_0 = \emptyset. \quad (7)$$

Each $d_j$ consists of a $3 \times 3$ convolutional layer $c$, and a single concatenation operation. Then, $A_j$ is formulated as

$$A_j = d_j(x_j, A_{j-1}) = c(c(x_j) \oplus A_{j-1}), \quad (8)$$

where $\oplus$ represents a channel-wise concatenation operation. $D$ is alternatively trained using either the ground-truth image or the result of $G$. Thus, when training $D$ with the outputs from $G$, $x_1$ becomes $O_4$, and $x_j$ is the generated facial component $(O_{4-j} \odot M_{j})$, where $2 \leq j \leq 4$. On the other hand, when $x_j$ is the ground-truth facial component $(I_{GT} \odot M_{j})$, $D$ learns the real face component, and $x_1$ is $I_{GT}$. Then, the last block $d_3$ acts as a classifier to estimate the probability of multiple input images being real or fake. Table 2 shows the detailed architecture of our discriminator. We applied spectral normalization \[49\] to all the convolutional layers to stabilize the training of discriminator.

Following Goodfellow et al. \[10\], we optimized $G$ and $D$ in an alternating manner to solve the following adversarial min-max function $V(G, D)$:

$$\min_{G} \max_{D} V(D, G) = \mathbb{E}_y[\log D(y)] + \mathbb{E}_{\text{blur}}[\log(1 - D(I_{\text{blur}}))]. \quad (9)$$

Here, $y$ is a set of images corresponding to the ground-truth components of faces as $y = \{y_i \mid y_i = (I_{GT} \odot M_i), 1 \leq i \leq 3, y_4 = I_{GT}\}$. Then, the adversarial loss ($L_{adv}$) is defined as follows:

$$L_{adv} = -\log(D(I_{\text{blur}})). \quad (10)$$

During training the generator, the error is backward-propagated to the intermediate layers of $G$ from the intermediate layers of the $D$ simultaneously. This provides stability in training, because the sub-networks of the generator can share the same goal. Meanwhile, the discriminator observes not only the final output of $G$, but also all the intermediate outputs of $G$.

Recently, perceptual loss \[50\] has been widely adopted for better visual quality. To take advantage of this, we employed a VGG-face loss $L_{vgg}$ which is defined by

$$L_{vgg} = \|\phi(I_{GT}) - \phi(I_{\text{deblur}})\|_1, \quad (11)$$

where $\phi(\cdot)$ represents the feature extracted from the Pool5 layer of the VGG-Face network \[51\].

The total loss of our MSPL framework $L_{total}$ is the combination of all the above loss functions discussed so far ($L_G$ in Eq. (5), $L_{adv}$ in Eq. (10), and $L_{vgg}$ in Eq. (11)) as

$$L_{total} = L_G + \lambda_1 L_{adv} + \lambda_2 L_{vgg}, \quad (12)$$

where $\lambda_1$ and $\lambda_2$ represent the weights used to balance the different loss terms. We empirically set the weights of our loss terms as $\lambda_1 = 0.05$ and $\lambda_2 = 0.05$. 

### Table 1. Architecture of the proposed sub-network ($g_i$). $F_{i-1}$ is an input feature of $i^{th}$ sub-network $g_i$. “W” and “H” represent the width and height of the feature, all of the “downconv” layers represent convolutional layer with stride 2 for the downsampling operation, and “upconv” layers represent the transpose convolution for upsampling, and “$+$” represents a channel-wise concatenation operation.

| Blocks       | In→Out     | Kernel | Output size |
|--------------|------------|--------|-------------|
| Resblock×2   | $F_{i-1}$→1| 3×3, 128, s1 | W×H         |
| downconv1    | r1→d1     | 4×4, 128, s2 | (W/2)×(H/2) |
| Resblock×2   | d1→r2     | 3×3, 128, s2 | (W/2)×(H/2) |
| downconv2    | r2→d2     | 4×4, 128, s2 | (W/4)×(H/4) |
| Resblock×4   | d2→r3     | 3×3, 128, s1 | (W/4)×(H/4) |
| upconv1      | r3→u1     | 4×4, 128, s2 | (W/2)×(H/2) |
| Resblock×2   | r2+u1→r3 | 3×3, 128, s1 | (W/2)×(H/2) |
| upconv2      | r3→u2     | 4×4, 128, s2 | W×H         |
| Resblock×2   | r1+u2→$F_i$ | 3×3, 128, s1 | W×H         |

### Table 2. Architecture of the proposed discriminator. All the multiple input images are denoted as $x_1, x_2, x_3$ and $x_4$. “ReLU” is a leaky ReLU, “SN” is a spectral normalization, and “$g$” represents a channel-wise concatenation operation.

| Blocks       | Operation     | In→Out     | Kernel |
|--------------|---------------|------------|--------|
| $d_1$        | Conv2d, ReLU, SN | $x_1$→$A_1$ | 3×3, 64, s1 |
| $d_2$        | Conv2d, ReLU, SN | $x_2$→$C_2$ | 3×3, 64, s1 |
| $d_3$        | Conv2d, ReLU, SN | $x_3$→$C_3$ | 3×3, 64, s1 |
| $d_4$        | Conv2d, ReLU, SN | $x_4$→$C_4$ | 3×3, 64, s1 |
| $d_5$        | Conv2d, ReLU, SN | $A_4$→$C_5$ | 4×4, 64, s2 |
| $d_6$        | Conv2d, ReLU, SN | $C_5$→$C_6$ | 4×4, 128, s2 |
| $d_7$        | Conv2d, ReLU, SN | $C_6$→$C_7$ | 4×4, 256, s2 |
| $d_8$        | Conv2d, ReLU, SN | $C_7$→$C_8$ | 4×4, 512, s2 |
| $d_9$        | Conv2d, ReLU, SN | $C_8$→$p$  | 4×4, 1, s2 |
IV. EXPERIMENTAL RESULTS

A. DATASETS

1) TRAINING DATA

For training, we used the CelebAMask-HQ dataset [57], which provides 30,000 high-quality (1024×1024 resolution) face images. Each image has 19 classes of segmentation labels such as skin, nose, eyes, eyebrows, ears, mouth, lips, hair, hat, eyeglass, earring, necklace, neck, and cloth. We regrouped these labels into three classes, which are the skin, hair and inner parts, as [8]. Following [58], we synthesized 18,000 motion blur kernels using the method of [23]. As trained in [7], [58], the size of the generated motion blur kernels ranges from 13×13 to 27×27. After applying the blur kernel to the image, we added Gaussian noise with \( \sigma = 0.03 \). The generated images were then split into two subsets: the training images (24,183 images), and the validation images (5,817 images).

2) TEST DATA

For face deblurring, Shen et al. [7] provide a pioneering testset to evaluate the motion-blurred faces. However, many GT images in the testset are of low quality with unknown block artifacts (see Fig. 3(a)). Because the purpose of image deblurring is to restore a sharp and high-quality image, these low-quality GT images are not suitable for evaluating performance. In addition, facial images in their testset are well aligned with the same facial key points [59]. However, the blurred face images are not aligned in the real world, because faces are usually captured under a wide range of conditions. Therefore, their testset does not consider the practical situations where these blurred face images occur. Therefore, we generated two types of test datasets called the MSPL-Center and MSPL-Random. We collected 240 sharp face images from three different datasets (i.e., 80 images each from CelebA [60], CelebAMASK-HQ (CelebA-HQ) [57] and Flickr-Faces-HQ thumbnails (FFHQ) [42]). Note that each dataset is aligned with different facial key points.

Subsequently, we synthesized 240 random motion blur kernels using the method presented in [23]. Following the protocol by Shen et al. [7], the size of blur kernels ranges from 13×13 to 27×27. As shown in Fig. 3(b), MSPL-Center contains high-quality and differently aligned face images. Meanwhile, MSPL-Random comprises images that are randomly augmented versions using MSPL-Center (samples are shown in Fig. 3(c)). To be specific, we conducted random crops, random rotation, and random horizontal flips to the MSPL-Center images and convolved the random blur kernels synthesized using [23].

3) IMAGE QUALITY COMPARISON

Following [61]–[63], we used four of the no-reference image quality assessment (NRIQA) metrics (i.e., NIQE [52], BRISQUE [53], NRQM [54], and PIQE [55]) to compare the quality of the images of our testset (MSPL-Center) and testset provided by Shen et al. [7]. The NIQE and BRISQUE metrics measure the image naturalness (or its lack thereof) based on their own natural scene statistics (NSS) model [52], [53]. NRQM provides the quality scores of the images based on extracted features from the trained CNN model [54]. PIQE is a perception-based image quality evaluation method that estimates the amount of distortion present in a given image [55]. We also employ the perception index (PI) metric [56], which is formulated as the adjusted mean value of NIQE and NRQM.

Shen et al. [7] and MSPL-Center testsets were compared because they both consist of centered facial images. As aforementioned, the face images in MSPL-Random are randomly transformed version of the MSPL-Center. Thus, we do not compare the MSPL-Random dataset with the Shen et al. [7] testset. All the results of image assessments are listed in Table 3. When comparing [7] and MSPL-Center, the values of NIQE are comparable. However, the MSPL-Center dataset achieved better BRISQUE, NRQM, and PIQE than the Shen et al. [7] testset.
For a fair comparison, we evaluated the image quality of the Shen et al. [7] and MSPL-Center subsets, both of which were synthesized using CelebA [60]. The results in Table 4 show that the test images in Shen et al. [7] are clearly degraded compared to those in MSPL-Center, even when considering the images selected from the same face dataset [60]. These assessment metrics quantify the noise, artifacts, sharpness and overall quality of the image. Therefore, the comparative results indicate that the Shen et al. [7] testset consist of low-quality GT images and the proposed MSPL testsets are more suitable for the evaluation of face deblurring performance.

### TABLE 3. Results of image quality assessment on face deblurring testsets. We compare the quality of GT images using the four image quality assessment metrics. Each result is an average of the number of GT images. Best results are highlighted as bold.

| Testset          | # GT images | NIQE [52] | BRISQUE [53] | NRQM [54] | PIQE [55] | PI [56] |
|------------------|-------------|-----------|--------------|-----------|-----------|--------|
| Shen et al. [7]  | 200         | 18.8742   | 34.5743      | 5.8856    | 48.2324   | 11.4944|
| MSPL-Center      | 240         | 18.8740   | 30.3057      | 6.6018    | 42.5591   | 11.1363|

### TABLE 4. Results of image quality assessment on CelebA images in testsets. Best results are highlighted in bold.

| CelebA           | # GT images | NIQE [52] | BRISQUE [53] | NRQM [54] | PIQE [55] | PI [56] |
|------------------|-------------|-----------|--------------|-----------|-----------|--------|
| Shen et al. [7]  | 100         | 18.8744   | 32.7266      | 6.0758    | 43.6343   | 11.3992|
| MSPL-Center      | 80          | 18.8743   | 30.7639      | 6.6247    | 42.6092   | 11.1265|

B. TRAINING DETAILS

To implement our models, we used Pytorch [64]. The generator and discriminator were trained using the Adam optimizer [65] with \( \beta_1 = 0.9, \beta_2 = 0.999 \). The learning rate was initialized as \( 1 \times 10^{-5} \) and decayed exponentially by a factor of 0.99 for every epoch. When training, we first resized the 1024 \( \times \) 1024 images to 512 \( \times \) 512 images using bilinear downsampling. Then, we randomly cropped the images to 448 \( \times \) 448 and resized them to 128 \( \times \) 128. We augment the resized images with random horizontal flips and random rotations in the range \([0^\circ, 90^\circ]\). We set the batch size as 16 and trained the model with a single NVIDIA TITAN-RTX GPU.

C. EVALUATION METRICS

To evaluate the performances of various methods, we used PSNR and SSIM [66], which are widely used in image restoration fields. The feature distance \( d_{VGG} \) of the VGG-Face network [67] was measured to compare the similarity in facial identity between the GT images and the deblurred face images. Following [68], we computed the \( L_2 \) distance using the output features from the Pool5 layer of the VGG-Face network [67]. Following the 2020 NTIRE challenge [63], we employed the LPIPS [69] distance using the output features from the learned CNN for computing human visual perception.

D. ABLATION STUDY

1) EFFECT OF THE SEMANTIC PROGRESSIVE GENERATOR

To investigate the impacts of the reconstruction procedure of the face components, we gradually modified the baseline model and compared the differences. We set the baseline model to the same architecture as \( G \), with the difference being that it is trained without facial component losses \( (L_1^c, L_2^c, \text{and } L_3^c) \) and discriminator. Thus, the baseline model generates only a single output image at the final image convert layer \( t(F_4) \) of \( G \). This baseline was trained with only \( L_4^c \) loss between the \( O_4 \) and \( I_{GT} \), which is commonly used in image restoration studies. From this baseline model, we added intermediate output layers and compared the performance of the models (denoted as MSPL) trained with different restoration procedures.

Table 5 shows the performances of all the conducted models trained with different reconstruction procedures and identical training settings and data. The \( M_i \) specified in Table 5 represents the semantic mask used to train the corresponding intermediate output of the MSPL model. MSPL_a is a model that is trained to restore a blurry input image in the order of entire image, skin, hair, and inner components. Comparing the results of MSPL_a and other models, it can be observed that restoring the entire face from the last module is crucial to the accuracy of the restoration process. MSPL_b is a trained model in reverse order of the proposed method except the entire component. This model restores the facial components starting from small and high-frequency components first (inner components), and restore the large and low-frequency components (hair, skin) later. The results of the MSPL_b also show that the wrong order of the face reconstruction degrades the deblurring performance. MSPL w/o GAN is a model trained to restore facial components with the proposed procedure using Eq. (5). The results in Table 5 demonstrate that the architecture trained with the proposed order improves the restoration performance compared to the other orders.

In addition, Table 6 quantitatively demonstrates the effect of our progressive generator for each facial component. When comparing the PSNRs of \( O_1, O_2, O_3 \), and \( O_4 \),...
we can confirm that the PSNR values of all the facial components gradually increase. These results indicate that the proposed method performs the deblurring incrementally. All the sub-networks enhance the quality of the image compared to the output image of the previous stage. Furthermore, we can observe that each sub-network is gradually improving not only the class-specific component, but also the entire face. This is because all the sub-networks share the goal of restoring the whole face. This lends stability to our progressive training method.

In Fig. 4, we can qualitatively observe the deblurring procedure in our proposed method that progressively restores the face image in the order of skin, hair, inner parts (eyes, eyebrows, nose, and mouth) and the entire face. From a blurred input face, the first sub-network generates the first output image $O_1$. At this stage, we can see that the overall shape of the facial skin is restored excluding the other facial components (see Fig. 4(b)). In Fig. 4(c), we confirm that the shape and texture of the hair in $O_2$ are restored from $O_1$. However, some blurred artifacts remain in the facial inner parts and background. In the third stage, the $O_3$ (Fig. 4(d)) shows that the inner parts of the face are significantly restored compared to $O_2$ from the previous stage. The final output image $O_4$ is shown in Fig. 4(e). The final result demonstrates that the final sub-network recovers the entire face and background. In particular, the facial components of $O_4$ are more natural compared to those of $O_3$.

2) EFFECT OF THE MULTI-SEMANTIC DISCRIMINATOR

In our MSPL framework, the multi-semantic discriminator is utilized to recover the faces that are more photo-realistic. To study the effects of this, we additionally trained our generator model with the proposed discriminator and trained with $L_{total}$ in Eq. (12). We denote our model as MSPL_GAN which are trained with the loss function $L_{total}$ in Eq. (12). The results listed in Table 5 indicate that the MSPL_GAN achieves slightly poor results compared to those archived by MSPL w/o GAN. However, the visual results presented in Fig. 5 show that our discriminator assists in reconstructing the faces that are more realistic. In Fig. 5, we compare the results of MSPL w/o GAN and MSPL_GAN (the odd rows and the even rows in Fig. 5, respectively) with the same blurred image as the input. It can be observed that the MSPL_GAN model restores the more realistic facial components than those restored by the MSPL w/o GAN model, especially for the nose, mouth, eyes, and texture of hair. For example, the output images in the second row contain more a realistic nose and mouth than the outputs in the first row. When comparing the images in the third and fourth rows, we can see a clear effect of GAN. This demonstrates the effect of our multi-semantic discriminator that allows the generator to restore the more realistic facial components. In addition, we confirm that our discriminator can affect on not only the final output image $O_4$, but also all the intermediate outputs from $O_1$ to $O_3$. In our experiments, the performance of MSPL_GAN model was slightly lower in the PSNR/SSIM compared to those of the MSPL w/o GAN model. However, MSPL_GAN model can restore the more visually plausible outputs that contain inner components, which are more natural, by using our multi-semantic discriminator.

### E. COMPARISONS WITH EXISTING METHODS

We compared the performance of our MSPL framework with recent methods based on CNN models [7], [8], [58], [68]. All the experiments were conducted using the official codes provided by the authors [7], [8], [58], [68]. For Xia and Chakrabarti [58], we used the model trained in a supervised manner, as this model has been reported as the best model in their studies.

1) CLASS IMBALANCE PROBLEM

As mentioned earlier, the class imbalance problem is an important and challenging issue for the existing face deblurring methods [7], [8]. To compare the restoring capability of restoring the small and thin components of the face (such as the eyes, lips, and eyebrows), we compared the PSNR value of each individual class in the face using a ground-truth segmentation map, following the experiment presented by Yasarla et al. [8].

As shown in Table 6, our model significantly outperforms the previous state-of-the-arts in restoring individual classes of the face, especially for the inner parts of the face which contain small and important features of the face. These results show that our model effectively restores the facial image by reducing the class-imbalance problem compared to the previous methods.

| Method                  | Facial Components | Skin Mean PSNR(%) | Hair Mean PSNR(%) | Inner Parts Mean PSNR(%) | Average Mean PSNR(%) |
|-------------------------|-------------------|-------------------|-------------------|--------------------------|----------------------|
| Shen et al. [7]         |                   | 19.75             | 19.76             | 16.16                    | 18.56                |
| Yasarla et al. [8]      |                   | 23.52             | 22.55             | 19.46                    | 21.85                |
| MSPL w/o GAN $O_1$      |                   | 26.77             | 24.40             | 19.10                    | 23.42                |
| MSPL w/o GAN $O_2$      |                   | 27.30             | 26.58             | 22.68                    | 25.52                |
| MSPL w/o GAN $O_3$      |                   | 28.30             | 27.22             | 24.84                    | 26.79                |
| MSPL w/o GAN $O_4$      |                   | 30.40             | 28.31             | 25.74                    | 28.15                |

**TABLE 6.** PSNR values for each facial component on CelebA-HQ in MSPL-Center Testset. We measured the metrics using individual classes of GT segmentation maps.
FIGURE 5. Comparisons of MSPL w/o GAN and MSPL_GAN. From left to right, input blurred faces, $O_1$, $O_2$, $O_3$, $O_4$, and the GT faces; rows 1 and 3 of the figure are the results of the MSPL w/o GAN, rows 2, and 4 are the output of MSPL_GAN.

TABLE 7. Quantitative comparisons on testset of Shen et al. [7]. Best results are highlighted as bold.

| Method       | Helen | CelebA |
|--------------|-------|--------|
|              | PSNR | SSIM   | $d_{VGG}$ | LPIPS | PSNR | SSIM | $d_{VGG}$ | LPIPS |
| Shen et al. [7] | 25.58 | 0.861 | 91.06 | 0.1527 | 24.34 | 0.860 | 117.50 | 0.1832 |
| Lu et al. [68] | 20.25 | 0.705 | 241.93 | 0.1654 | 19.96 | 0.742 | 305.96 | 0.1688 |
| Xia et al. [58] | 26.13 | 0.886 | 55.97 | 0.1052 | 25.18 | 0.892 | 68.05 | 0.1199 |
| Yasarla et al. [8] | 27.75 | 0.897 | 86.87 | 0.1086 | 26.62 | 0.908 | 66.33 | 0.1401 |
| MSPL_GAN      | 25.91 | 0.881 | **47.80** | **0.0828** | 24.91 | 0.885 | **57.54** | **0.0962** |

2) COMPARISONS USING SHEN ET AL. [7] TESTSET

We conducted experiments on the testset provided by Shen et al. [7]. In Table 7, it can be noted that Yasarla et al. [8], and Xia and Chakrabarti [58] performed the best PSNR and SSIM. However, as can be seen in Fig. 6, the results obtained using the previous methods [7], [8], [58] are overly smoothed images, their models obtain better results in PSNR and SSIM compared to our model. As mentioned before, the problems of low-quality images in the Shen et al. [7] testset are noteworthy. The results in Fig. 6 show this problem more clearly. First, we can observe that not a few GT images have severe blocking artifacts (for example, see the last column in Fig. 6). Second, our model restores the sharp images that are even better than GT images. When comparing the GT images with our images, our results have sharper boundaries at the border of the facial components without blocking artifacts. These observations support that the existing Shen et al. [7] testset has a limitation in providing accurate deblurring evaluation.

On the other hand, the proposed MSPL w/o GAN and MSPL_GAN achieved the best performance in $d_{VGG}$ and LPIPS with a huge margin, as listed in Table 7. The result of
$d_{VGG}$ indicates that our restored faces are similar to the GT images in terms of face identification. In addition, this shows that our model is the best model for higher vision task such as face recognition. LPIPS is the metric, which correlates better with human perceptual opinions [63]. The results of LPIPS show that the restored faces using our model are more visually plausible in terms of human vision.

### 3) COMPARISONS USING MSPL TESTSET

In extended experiments on MSPL-Center and MSPL-Random testsets, we observed that our proposed method achieved the best performance both quantitatively and qualitatively. The quantitative results are listed in Table 8. The result values of PSNR, SSIM, $d_{VGG}$, and LPIPS indicate that our framework significantly outperformed the existing methods. Additional visual results in Fig. 7 and Fig. 8 demonstrate that the proposed method restored sharper and more detailed face images than previous methods. In our experiments, we observed that the performance of Shen et al. [7] was sensitive to alignment and rotation. When a given blurred face image was aligned differently or rotated differently from the training face images, the restoration performance of Shen et al. [7] was severely degraded (refer to the second column images of Fig. 7). The results of Yasarla et al. [8] were visually plausible for all the test images. However, the restored small facial components (i.e., eyes, nose, mouth, and teeth) still lacked details and textures when the blurred artifacts were severe in the input image. Meanwhile, the proposed framework achieved superior performance compared to previous existing methods.

### 4) REAL-WORLD BLURRED FACIAL IMAGES

We conducted experiments on the twenty facial images distorted by real-world blur provided by [7], [70]. In the real-world, images are easily degraded with unknown complex factors such as the motion blur, lens distortion, sensor saturation, nonlinear transform functions, noise, and compression, in the camera pipeline [70]. However, all of these
factors are not considered when generating the synthetically blurred images. Therefore, this experiment allows us to confirm the more practical performances of face deblurring methods that cannot be evaluated using the synthetically generated blur testsets. The comparative results of the sample images for the real-world blur are shown in Fig. 9. As can be seen in Fig. 9, the results of the proposed method show the most sharp and natural face images compared to other face deblurring methods. These results indicate that our proposed method has the capability of reconstructing the highest-quality images from the facial images with a real-world blur.

5) INFERENCE TIME AND MODEL PARAMETERS
As shown in Table 9, we measured the inference time and the number of model parameters of the existing methods.
FIGURE 9. Qualitative comparison on real-world blurred facial images.

and the proposed model. The inference time is measured by averaging inference time of 10 images with a size of 128 × 128 on a single NVIDIA Titan Xp GPU.

When comparing the proposed model with the recent state-of-art model of Yasarla et al. [8], our model is slightly larger in number of parameters. However, it can be seen that our model is 2 times faster than Yasarla et al. [8]. This shows that the proposed method is more efficient than the two-stage deblurring method [8] consisting of segmentation and deblurring processes.

V. CONCLUSION

In this study, we propose a multi-semantic progressive learning framework for facial image deblurring. Our framework employs an effective GAN-based architecture to restore the semantic structures of the face progressively without performing semantic segmentation. To evaluate the more practical and accurate performances of face deblurring methods, we have provided additional new testsets. Overall, the proposed method outperforms the existing methods both qualitatively and quantitatively. To the best of our knowledge, this is the first study on facial image deblurring using a semantic in progressive approach. We believe that our framework provides a potential approach for numerous other facial image restoration fields.

ACKNOWLEDGMENT

(Tae Bok Lee and Soo Hyun Jung contributed equally to this work.)

REFERENCES

[1] Y. Hacohen, E. Shechtman, and D. Lischinski, “Deblurring by example using dense correspondence,” in Proc. IEEE Int. Conf. Comput. Vis. (ICCV), Dec. 2013, pp. 2384–2391.

[2] 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.

[3] 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.

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

[5] 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.

[6] S. Ren, 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.

[7] 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. (CVPR), Jun. 2018, pp. 8260–8269.

[8] 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.

[9] Y. Chen, Y. Tai, X. Liu, C. Shen, and J. Yang, “FSRNet: End-to-end learning face super-resolution with facial priors,” in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit., Jun. 2018, pp. 2492–2501.

[10] 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. (NIPS), Jun. 2014, pp. 2672–2680.

[11] T. Karras, T. Aila, S. Laine, and J. Lehtinen, “Progressive growing of GANs for improved quality, stability, and variation,” Oct. 2017, arXiv:1710.10196. [Online]. Available: http://arxiv.org/abs/1710.10196.

[12] A. Karnewar and O. Wang, “MSG-GAN: Multi-scale gradients for generative adversarial networks,” in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR), Jun. 2020, pp. 7799–7808.

[13] R. Fergus, B. Singh, A. Hertzmann, S. T. Roweis, and W. T. Freeman, “Removing camera shake from a single photograph,” in Proc. ACM SIGGRAPH Papers. New York, NY, USA: Association for Computing Machinery, Jul. 2006, pp. 787–794, doi: 10.1145/1179352.1141956.

[14] D. Krishnan, T. Tay, and R. Fergus, “Blind deconvolution using a normalized sparsity measure,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), Jun. 2011, pp. 233–240.

[15] L. Xu, S. Zheng, and J. Jia, “Unnatural L0 sparse representation for natural image deblurring,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), Jun. 2013, pp. 1107–1114.
[63] E. Agustsson and R. Timofte, “NTIRE 2017 challenge on single image super-resolution: Dataset and study,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW), Jul. 2017, pp. 126–135.

[64] 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.

[65] D. P. Kingma and J. Ba, “Adam: A method for stochastic optimization,” Dec. 2014, arXiv:1412.6980. [Online]. Available: http://arxiv.org/abs/1412.6980

[66] 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.

[67] O. M. Parkhi, A. Vedaldi, and A. Zisserman, “Deep Face Recognition,” in Proc. BMVC Sep. 2015, pp. 41.1–41.12, doi: 10.5244/C.29.41.

[68] B. Lu, J.-C. Chen, and R. Chellappa, “Unsupervised domain-specific deblurring via disentangled representations,” in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR), Jun. 2019, pp. 10225–10234.

[69] 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. (CVPR), Jun. 2018, pp. 586–595.

[70] 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.

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.

SOO HYUN JUNG received the B.S. degree in electronic and electrical engineering from Dongguk University, Seoul, South Korea, in 2019. She is currently pursuing the M.S. degree in electrical and computer engineering with Ajou University, Suwon, South Korea. Her research interests include computer vision, deep learning, image restoration, and visual reasoning.

YONG SEOK HEO (Member, IEEE) received the B.S. and M.S. degrees in electrical engineering, and the Ph.D. degree in computer science from Seoul National University, South Korea, in 2005, 2007, and 2012, respectively. From 2012 to 2014, he was with 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.

* * *