Received 21 June 2022, accepted 7 July 2022, date of publication 12 July 2022, date of current version 25 July 2022.

Digital Object Identifier 10.1109/ACCESS.2022.3190089

RESEARCH ARTICLE

Continuous Facial Motion Deblurring

TAE BOK LEE[1], **SUJY HAN**[1], **AND YONG SEOK HEO**[1,2]

[1]Department of Artificial Intelligence, Ajou University, Suwon-si 16499, South Korea
[2]Department of Electrical and Computer Engineering, Ajou University, Suwon-si 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 funded by the Ministry of Education under Grant NRF5199991014091; and in part 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 & Communications Technology Promotion (IITP) under Grant IITP-2022-2018-0-01424.

ABSTRACT We introduce a novel framework for continuous facial motion deblurring that restores the continuous sharp moment latent in a single motion-blurred face image via a moment control factor. Although a motion-blurred image is the accumulated signal of continuous sharp moments during the exposure time, most existing single image deblurring approaches aim to restore a fixed number of frames using multiple networks and training stages. To address this problem, we propose a continuous facial motion deblurring network based on GAN (CFMD-GAN), which is a novel framework for restoring the continuous moment latent in a single motion-blurred face image with a single network and a single training stage. To stabilize the network training, we train the generator to restore continuous moments in the order determined by our facial motion-based reordering process (FMR) utilizing domain-specific knowledge of the face. Moreover, we propose an auxiliary regressor that helps our generator produce more accurate images by estimating continuous sharp moments. Furthermore, we introduce a control-adaptive (ContAda) block that performs spatially deformable convolution and channel-wise attention as a function of the control factor. Extensive experiments on the 300VW datasets demonstrate that the proposed framework generates a various number of continuous output frames by varying the moment control factor. Compared with the recent single-to-single image deblurring networks trained with the same 300VW training set, the proposed method show the superior performance in restoring the central sharp frame in terms of perceptual metrics, including LPIPS, FID and Arcface identity distance. The proposed method outperforms the existing single-to-video deblurring method for both qualitative and quantitative comparisons. In our experiments on the 300VW test set, the proposed framework reached 33.14 dB and 0.93 for recovery of 7 sharp frames in PSNR and SSIM, respectively.

INDEX TERMS Continuous facial motion deblurring, AC-GAN, control-adaptive block.

I. INTRODUCTION

Facial motion deblurring for a single image is a specific but critical branches of image deblurring, aimed at restoring a sharp image latent in a motion-blurred face image. Besides being visually unpleasant, blurry face images also degrade the performance of many facial-related computer vision tasks such as face detection [6]–[8], face recognition [9], [10], facial emotion recognition [11], [12], and face medical image segmentation [13]. Therefore, face deblurring studies in computer vision and image processing have received much attention.

Recently, deep neural networks (DNNs) have become widespread in image restoration fields [14]–[17]. Among them, it has been achieved remarkable success in single image face deblurring [18]–[24]. Most of these methods recover only a single sharp image from a motion-blurred facial image. However, motion-blurred images are the integration of continuous sharp moments during the exposure time [16], [25]. Thus, recovering such aggregated sharp moments from the blurred image can be considered the ideal goal of single image deblurring.

The associate editor coordinating the review of this manuscript and approving it for publication was Anand Paul.

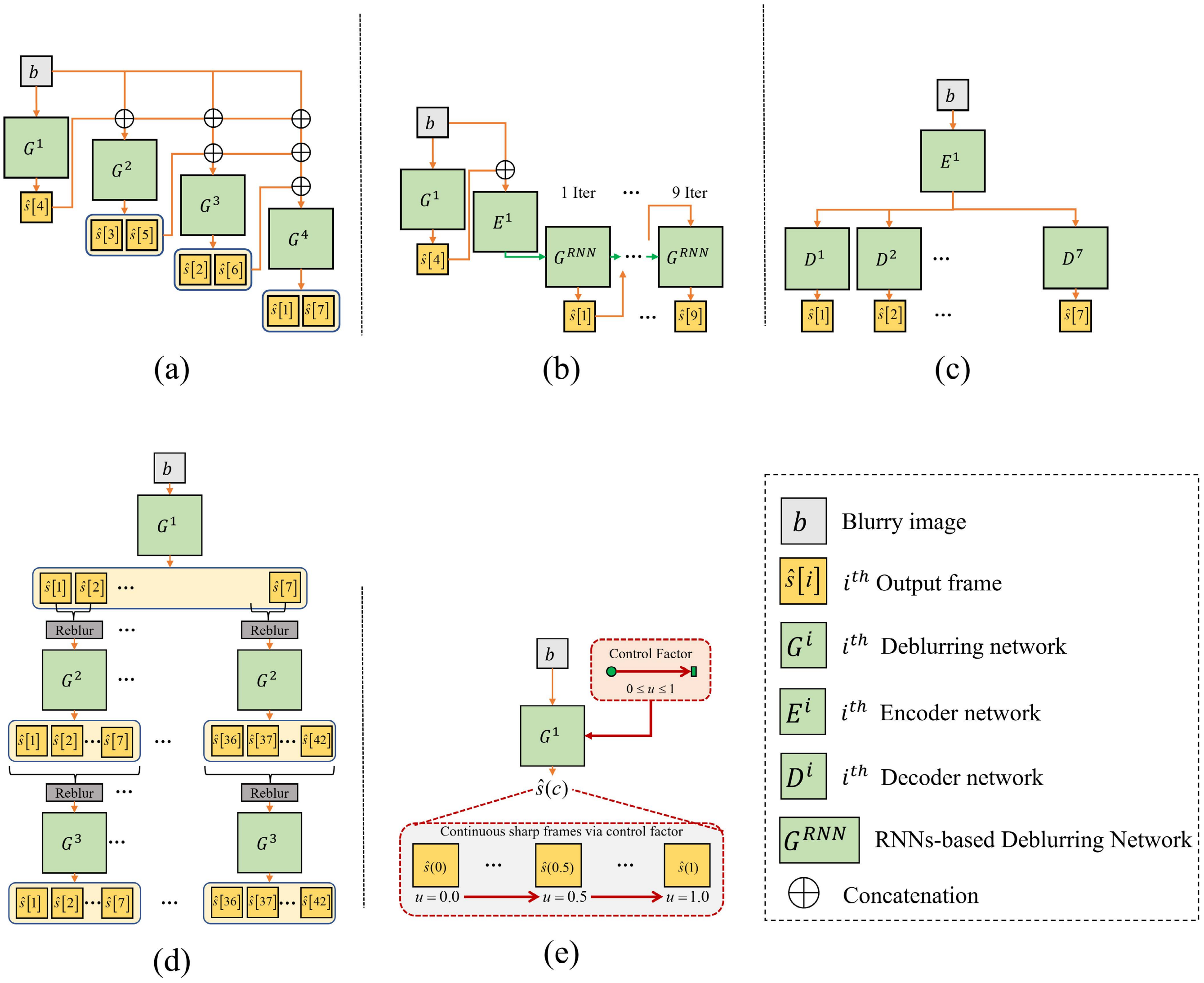
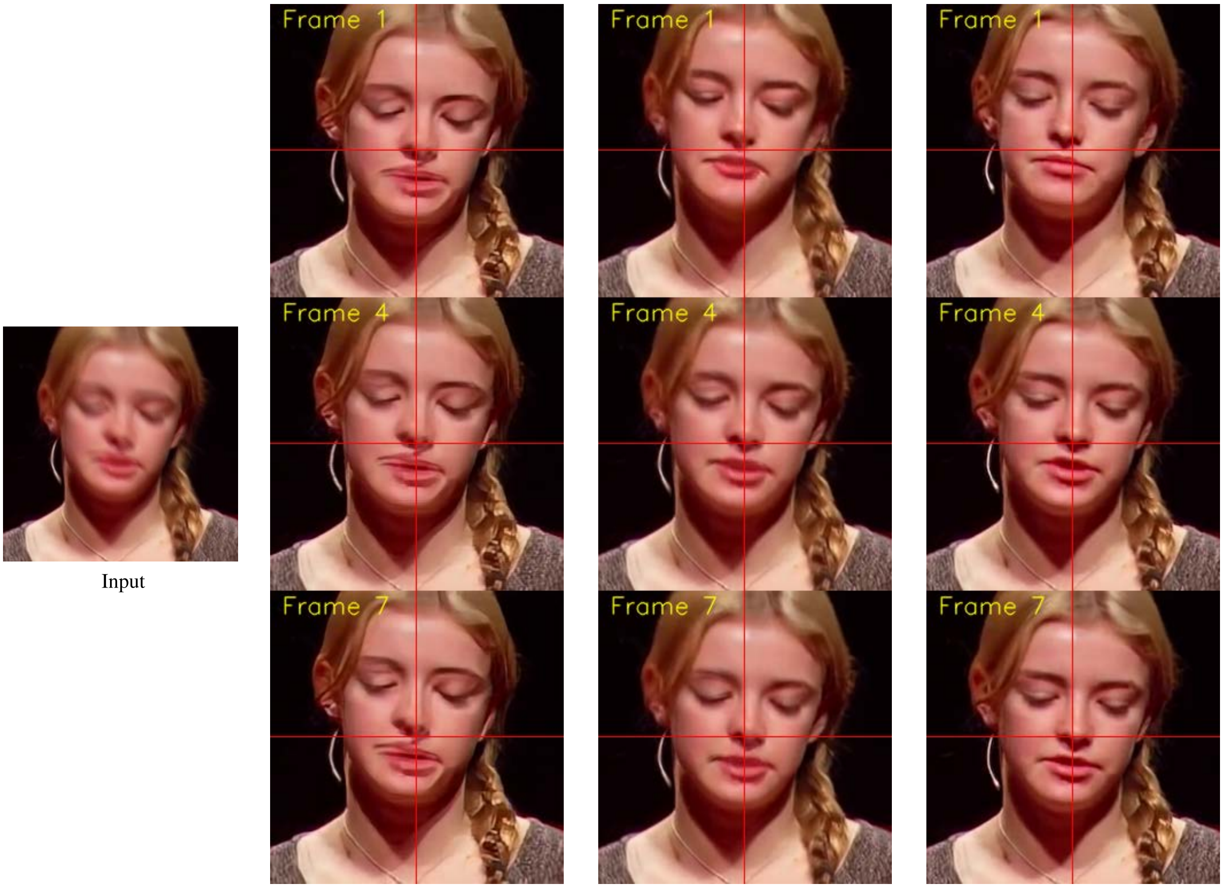


FIGURE 1. Comparison of single-to-video deblurring network architectures. The proposed method can restore continuous sharp motion of the face with a single network. (a) Jin *et al.* [1], (b) Purohit *et al.* [2], (c) Argaw *et al.* [3], (d) Zhang *et al.* [4], and (e) proposed CFMD.

Several methods [1]–[4] have been proposed to restore sharp sequences from a blurry image. However, most of these methods have several drawbacks. First, the temporal ordering problem is extremely challenging, because it is difficult to uniquely define the temporal order of the motion of an object in a blurry image [1]–[3]. For this reason, most existing methods fail to extract the accurate temporal order. This temporal ambiguity of the underlying motion in blurry images remains unsolved issue [3]. Second, as shown in Fig. 1, most existing models aim to only restore fixed frames, owing to architectural design or training strategies. Jin *et al.* [1] proposed a cascaded architecture consisting of four deblurring networks. As depicted in Fig. 1a, each network is assigned to restore neighboring frames using the outputs from the previous networks. Thus, this method requires a large number of networks according to the number of output frames to be extracted. Purohit *et al.* [2] proposed the using a recurrent neural network (RNN) so that they can handle various numbers of frames without architectural changes (Fig. 1b). They first extracted the middle frame using a pre-trained deblurring network and extracted nine frames using an RNN. However, their model is fixed to restore the entire sequence with nine frames, which is the predefined number of iterations of the RNN in the training phase. Argaw *et al.* [3] proposed a single encoder-multiple decoder architecture trained in a single training step. However, as shown in Fig. 1c, this architecture requires as many decoders as output frames. Recently, Zhang *et al.* [4] have shown promising results by restoring 42 frames from a blurry image. They trained three generative adversarial networks (GANs) by repeating the reblurring and deblurring processes (Fig. 1d). However, they restore a fixed number of frames and require multiple training steps.



Input

Jin *et al.* [1]

Ours

Ground truth

FIGURE 2. Exemplar deblurring results. "GT" denotes the ground-truth sharp frames in 300VW dataset [5]. "# Fr" in parentheses denotes the number of frames. The results in (e) and (f) denote the outputs of the same network. By adjusting the control factor value, our single network can restore any number of sharp movements from a given blurry face image. *This figure contains videos that are best viewed using Adobe Reader.*

To address the problems described above, as shown in Fig. 1e, we propose a facial motion-based reordering (FMR) process and a continuous facial motion deblurring network based on GAN (CFMD-GAN), a novel framework for restoring continuous moment latent in a single motion-blurred face image with a single training stage.

To alleviate the difficulty of resolving temporal ambiguity, we estimate the reordered frames instead of estimating the frames in the original temporal order. To this end, we apply a facial motion-based reordering (FMR) process, which reorders frames in the dataset based on the position of the left eye in the face (*e.g.* from top-left to right-bottom position) [26]. This reordering process helps the network stabilize training.

On the other hand, we introduce CFMD-GAN that restores sharp moments by varying the continuous moment control factor to estimate frames under continuous scenario. This approach is primarily inspired by conditional GANs (cGANs) [27]–[31], which are effective for training generators to synthesize diverse and realistic data conditioned on interpretable information, such as class labels. In our case, a single image deblurring network serves as the generator, and the conditional information for sharp image generation is the moment control factor. However, we have found that there are two main challenges in effectively incorporating cGANs into a single image deblurring framework. **First**, most existing cGANs are primarily developed for image synthesis conditioned on *discrete labels* (*e.g.* class labels) [32]. In contrast, we aim to restore the output images conditioned on the *continuous control factor*. Unlike most cGANs [28], [33]–[35] that use an auxiliary classifier for discrete class labels, we propose an auxiliary regressor to estimate the continuous control factor. It allows the proposed deblurring network to learn the image deblurring as a function of the continuous control factor. **Second**, an effective network module is required to apply the control factor into the deblurring network. Most existing single image deblurring approaches directly learn image-to-image mapping functions without the use of control factor. Recently, DNNs-based controllable image restoration models [36]–[38] have been extensively studied. Generally, these methods use a channel-wise attention module as a function of the control factor to resolve the Gaussian blurs and noise in static scenes. However, spatially-variant blurs with dynamic scenes must be considered. To this end, we present a control-adaptive (ContAda) block to effectively incorporate a control factor into recent deblurring architectures. The proposed block learns the modulation weights using a spatially deformable convolution and channel-wise attention as functions of the control factor.

Extensive experiments show that the proposed CFMD-GAN restores continuous sharp moments latent in a blurry face image using a single network and a single training process. Fig. 2 exemplifies our results, and compares our method with previous method [1].

The main contributions of this study are summarized as follows.

- We introduce the FMR process to stabilize the network training. It allows the network to utilize rich and accurate information of the ground-truth frames corresponding to the control factor during training.
- We propose a CFMD-GAN for continuous facial motion deblurring that restores continuous sharp frames latent in a single motion-blurred face image via a moment control factor.
- We present a ContAda block to learn the feature modulation weights of the deblurring network using spatially deformable convolution and channel-wise attention as functions of the control factor.

II. RELATED WORKS

In this section, we briefly review recent single image deblurring methods and conditional GANs, which are closely related to the present work.

A. SINGLE IMAGE DEBLURRING

Traditionally, the motion-blur process is formulated as the accumulation of continuous sharp moments that occur during exposure [16], [39]. By mimicking this, large-scale deblurring datasets [16], [40]–[42] have been proposed by synthesizing a blurry image by averaging consecutive sharp frames. By leveraging such datasets, DNNs-based methods have become widespread for single image deblurring. In the following, we introduce existing DNNs-based single image deblurring methods into three categories.

1) SINGE-TO-SINGLE, GENERAL DEBLURRING

Single-to-single image deblurring aims to restore a single sharp image when a blurry image in a general scene is given. Earlier studies [43]–[45] estimated the blur kernel using DNNs and obtained the resulting image using deconvolution methods. Chakrabarti *et al.* [43] proposed a network that predicts the complex Fourier coefficients of a deconvolution filter and applies the predicted deconvolution filter to the input patch. Sun *et al.* [44] proposed a deep learning approach that estimates motion blur kernels from local patches using a Markov random field model. Gong *et al.* [45] developed a DNN to predict the motion flow from blurred images, which was used to recover deblurred images. Without estimating the deconvolution kernel, Nah *et al.* [16] utilized a coarse-to-fine network to directly restore a sharp image using their synthesized large-scale dynamic scene blur dataset. Following the success of [16], variants of coarse-to-fine networks have been proposed, such as multi-recurrent networks [46], [47], multi-patch networks [48] and efficient multi-scale networks [49]. Concretely, Tao *et al.* [46] designed a scale-recurrent network that shares network parameters across scales. Zhang *et al.* [48] cascaded a multi-patch network to restore sharp images based on different patches. In addition, Cho *et al.* [49] reduced computational costs by utilizing a U-Net [50]-based architecture that exploits multi-scale features extracted from an input image and outputs.

2) SINGE-TO-SINGLE, FACE DEBLURRING

Face deblurring is a domain-specific task of single image deblurring that aims to obtain a sharp face from a blurry face image. Most existing methods have been studied in a manner that utilizes strong prior knowledge of the face, such as reference faces [51], [52], face landmark [19], [20], face sketches [53], multi-task embedding [21], 3D face models [54], facial parsing maps [18], [22], [23] and deep feature priors [24]. Specifically, Shen *et al.* [18] proposed to estimate the facial parsing map from the blurry face and then utilize it for restoring the sharp image. To avoid side effects caused by incorrect parsing maps, Yasarla *et al.* [22] utilized an uncertainty-based multi-stream architecture. Lee *et al.* [23] proposed restoring the face progressively from large components, such as skin, to small components, such as the eyes and nose. More recently, Jung *et al.* [24] utilized the rich information of feature maps extracted from a pre-trained deep neural network on the face.

However, all single-to-single deblurring methods, including the general and facial image domains, focus on restoring only one of the many moments accumulated in the blurred image. Unlike these methods, the proposed method restores various numbers of moments from a blurred image.

3) SINGE-TO-VIDEO, GENERAL DEBLURRING

Instead of restoring a single output image, single-to-video deblurring is to predict multiple sharp frames from a single blurred image. In the pioneering work of Jin *et al.* [1], a sequentially cascaded architecture consisting of multiple networks trained with the corresponding number of training steps was utilized. In their method, each network is assigned to predict pre-specified frames among all sharp frames. Thus, this method requires changing the number of networks based on the desired number of output frames and training them from scratch. Purohit *et al.* [2] proposed a recurrent neural networks (RNNs)-based method trained with two stages. In the first stage, they trained a video autoencoder to learn the motion and frame generation from sharp frames. It addresses the problem of the number of network scales with respect to the number of output frames. However, they still have to be trained anew each time the number of output frames changes. The method proposed by Zhang *et al.* [4] was one of the first attempts to restore continuous frames. Their method extracts a total of 42 sharp frames from a blurry image by cascading three GANs trained in three stages. However, this approach is limited to restoring a fixed number of frames. Instead of training the entire model in multiple stages, Argaw *et al.* [3] proposed a single framework that can be trained in an end-to-end manner. They proposed a feature transformer network consisting of a single encoder and multiple decoders, where each decoder was specified to output a specific frame. Thus, this method still requires changing the number of decoders when the number of output frames changes.

In short, existing studies are inherently limited in restoring only a fixed number of frames, owing to their rigorous architectural design or training strategies. In contrast, the proposed method differs in that 1) it restores continuous sharp frames beyond a fixed number, 2) a single deblurring network with a single training step is utilized, and 3) the proposed method can be trained in an end-to-end manner.

B. CONDITIONAL GENERATIVE ADVERSARIAL NETWORKS

Generative Adversarial Networks (GANs) [55] are among the most widely used frameworks in image generation and have been extensively studied over the past few years. Conditional GANs (cGANs) [27] are variants of GANs that synthesize realistic and diverse images using conditional information, such as class labels. Depending on how the framework incorporates the data and class labels, most cGANs can be categorized into classifier-based cGANs [28], [33]–[35] and projection-based cGANs [29], [30], [56], [57]. Classifier-based cGANs utilize conditional information (class labels) by training an additional classifier as well as a standard GAN discriminator. Meanwhile, projection-based cGANs propose a projection discriminator that takes an inner product between the embedded class labels and the feature vector extracted from the data.

The proposed method draws inspiration from all existing cGANs. To the best of our knowledge, this is the first attempt to apply continuous conditional information to deblurring task.

III. PRELIMINARIES

Generative Adversarial Networks (GANs) [55] are well-established method for mimicking the probability distribution of the real data by playing a min-max game between the generator G and discriminator D. Whereas G learns to fool D by generating realistic samples, D learns to classify whether the given samples are true data (real) or generated data (fake). Their objective, $V(G, D)$ is formulated as follows.

$$\begin{aligned}\min_G \max_D V(G, D) &= \mathbb{E}_{x\sim p(x)}[\log(D(x))] \\ &\quad + \mathbb{E}_{z\sim p(z)}[\log(1 - D(G(z)))], \qquad (1)\end{aligned}$$

where $p(x)$ denotes the real data distribution, and $p(z)$ denotes a pre-defined distribution, *e.g.* , Gaussian distribution. A key property of GANs is that a well-trained G successfully captures the data manifold even if there are missing data in the training set [58]–[60].

Conditional GANs (cGANs) [27]–[29] are an extended GAN framework developed for conditional image synthesis. Given a pair of images x and class labels c sampled from the joint distribution of the real dataset $(x, c) \sim p(x, c)$, the goal of G is to learn the class-conditional image synthesis by utilizing c as an additional input with z. Let $p_G(x|c)$ denote the generative distribution specified by $G(x, c)$ and $p_G(x, c) := p_G(x|c)p(c)$. The objective of generic cGANs [27], $V_{\text{cGAN}}(G, D)$, minimizes the Jensen-Shannon Divergence (JSD) between $p(x, c)$ and $p_G(x, c)$ as

$$\begin{aligned}&\min_G \max_D V_{\text{cGAN}}(G, D) \\ &\quad = \mathbb{E}_{(x,c)\sim p(x,c)}[\log(D(x, c))] \\ &\qquad + \mathbb{E}_{z\sim p(z),c\sim p(c)}[\log(1 - D(G(z, c), c))]. \qquad (2)\end{aligned}$$

As one of the most representative classifier-based cGANs, AC-GAN [28] introduces an auxiliary classifier Q to provide feedback on the class-conditional image synthesis of G. In AC-GAN, D and Q share all weights of the feature extractor, except for the final output layer. Let $p_Q(c|x)$ denote the conditional distribution induced by classifier Q. Then, their loss, $V_{\text{AC-GAN}}(G, Q, D)$ can be expressed as follows

$$\begin{aligned}\min_{G,Q} \max_D V_{\text{AC-GAN}}(G, Q, D) &= \mathbb{E}_{(x,c)\sim p(x,c)}[\log(D(x))] \\ &\quad + \mathbb{E}_{z\sim p(z),c\sim p(c)}[\log(1 - D(G(z, c)))] \\ &\quad - \lambda_c \underbrace{\mathbb{E}_{(x,c)\sim p(x,c)}[\log(p_Q(c|x)]}_{\text{(a)}} \\ &\quad - \lambda_c \underbrace{\mathbb{E}_{(x,c)\sim p_G(x,c)}[\log(p_Q(c|x))]}_{\text{(b)}}, \qquad (3)\end{aligned}$$

where λ_c is the balancing weight between the GAN and the auxiliary classification losses. In Eq. (3), the first two lines are loss functions similar to the original GANs (Eq. (1)), where D serves as a binary classifier that distinguishes between real and fake samples. Terms (a) and (b) represent the auxiliary classification losses that enable Q to determine the class labels of the input samples. Through this auxiliary classifier, AC-GAN can generate class-conditional image synthesis.

IV. PROPOSED METHOD

In this section, we first introduce the facial motion-based reordering (FMR) process, which is proposed to mitigate the temporal ambiguity problem by utilizing human face information (Sec. IV-A). Next, detailed explanation of the key components of the proposed CFMD-GAN is provided, which recovers the continuous moment latent in a blurry face image via a moment control factor (Sec. IV-B). Lastly, we introduce the training objectives of the proposed model (Sec. IV-C).

A. FACIAL MOTION-BASED REORDERING

One of the main challenges in restoring multiple images from a single blurred image is to resolve the *temporal (sequence)*

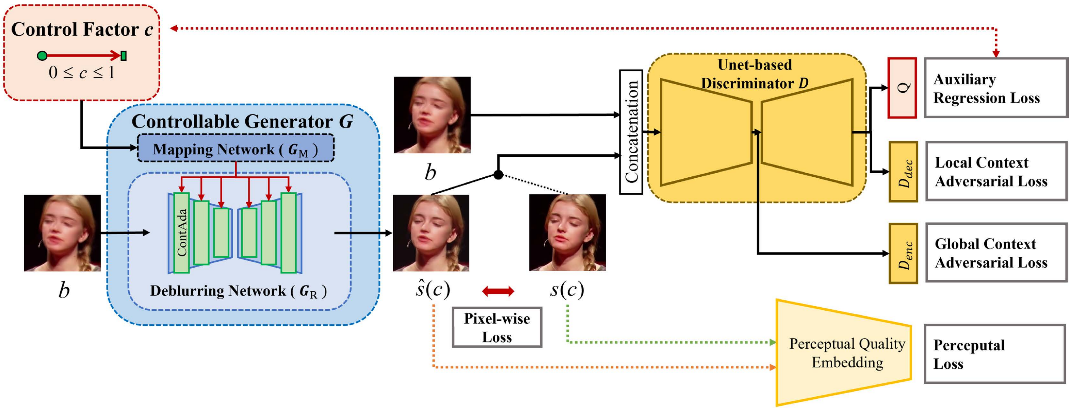

Control Factor c
$0 \leq c \leq 1$
Controllable Generator G
Mapping Network (G_M)
ContAda
Deblurring Network (G_R)
b
b
$\hat{s}(c)$
$s(c)$
Pixel-wise Loss
Concatenation
Unet-based Discriminator D
Q
D_{dec}
D_{enc}
Auxiliary Regression Loss
Local Context Adversarial Loss
Global Context Adversarial Loss
Perceptual Quality Embedding
Perceptual Loss

FIGURE 3. An overview of our CFMD-GAN framework. Given a single motion-blurred face image, the proposed generator restores the multiple sharp moments by varying a moment control factor. Subsequently, the proposed auxiliary regressor in the discriminator helps the generator learn to estimate more accurate result during training.

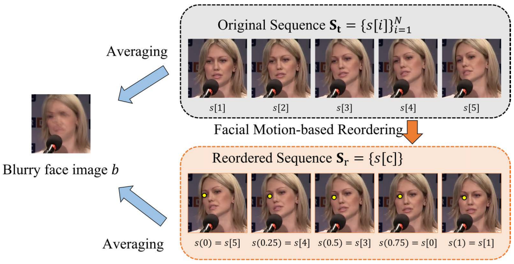


FIGURE 4. Facial motion-based reordering process (FMR). We rearrange the original sequence based on the position of the left eye *i.e.* from top-left to right-bottom.

ambiguity of sharp moments. A motion-blurred image is the averaged result of a continuous sharp sequence during the exposure time [16], [39]. As averaging destroys the information of the temporal order [1], [3], [4], reconstructing the original sequence of sharp moments is non trivial. For example, suppose a blurry facial image and its corresponding original sharp sequence are given, as shown in Fig. 4. The problem is that the same blurry image can be obtained even if the face moves in a reverse or shuffled order during the exposure time. Owing to this ill-posed nature of the temporal ambiguity, finding the underlying sequence of the blurry image is one of the unsolved issues [3]. In this regard, previous studies [1]–[3] have found that temporal ambiguity causes unstable training of the network because it is difficult to uniquely define the temporal sequence of object movements.

To alleviate this, we leverage the information of the human face to apply effective yet strong constraints. In a recent study on face landmark detection, Sun *et al.* [26] proposed defining the intensity of facial motion as the movement of the left eye during the time unit. Inspired by this, we devised a facial motion-based reordering (FMR) that enables the network to restore sharp face images in a generalized order based on the position of the left eye.

Specifically, as depicted in Fig. 4, FMR is a motion-based reordering process of the ground-truth (GT) sequence in a training dataset consisting of a single facial motion per single video clip. Let $\mathbf{S_t}$ be a time-ordered set of GT frames sampled from a high-frame-rate facial video, which is denoted by

$$\mathbf{S_t} = \{s[i] \in \mathbb{R}^{H\times W\times 3} \mid i \in [1, N]\}, \tag{4}$$

where i denotes the frame index within the total number of frames N. Then, a blurry image $b \in \mathbb{R}^{H\times W\times 3}$ can be approximated by averaging these GT frames as follows:

$$b \simeq g(\frac{1}{N}\sum\nolimits_{i=1}^{N} s[i]), \tag{5}$$

where $g(\cdot)$ denotes the camera response function [16]. We rearrange $s[i]$ according to the position of the left eye (x, y) [1] in each $s[i]$ so that the frame that includes the eye in the top-left position comes first, and the frame that includes the eye in the bottom-right position follows the last. Concretely, the proposed FMR process rearranges the sharp sequence according to the following criteria: **(c1)** The order is primarily determined by the ascending order of x values. It generalizes the erratic movement of the face as a *left-to-right movement*. **(c2)** If there are frames with the same x values, those frames are sorted in ascending of y values. It also regularizes the direction of facial motion to a *top-to-bottom movement*. **(c3)** When frames have the same (x, y), they are sorted in ascending order of temporal sequence.

Following the above procedure, we further transform the frame index i into a continuous motion index value $u \in [0, 1]$ by applying $u = \frac{i-1}{N}$. Then, we can denote this reordered set $\mathbf{S_r}$ as follows:

$$\mathbf{S_r} = \{s(u) \in \mathbb{R}^{H\times W\times 3} \mid u \in [0, 1]\}. \tag{6}$$

Note that the real number u becomes a *moment control factor* in the proposed framework.

In this study, the network learns to restore the facial motion-based order in $\mathbf{S_r}$. It should be noted that this reordered sequence does not match the temporal sequence. Instead, the proposed framework restores all possible sharp

[1] We utilize the public landmark detector provided by OpenCV [61] to obtain the position of the left eye in the face image.

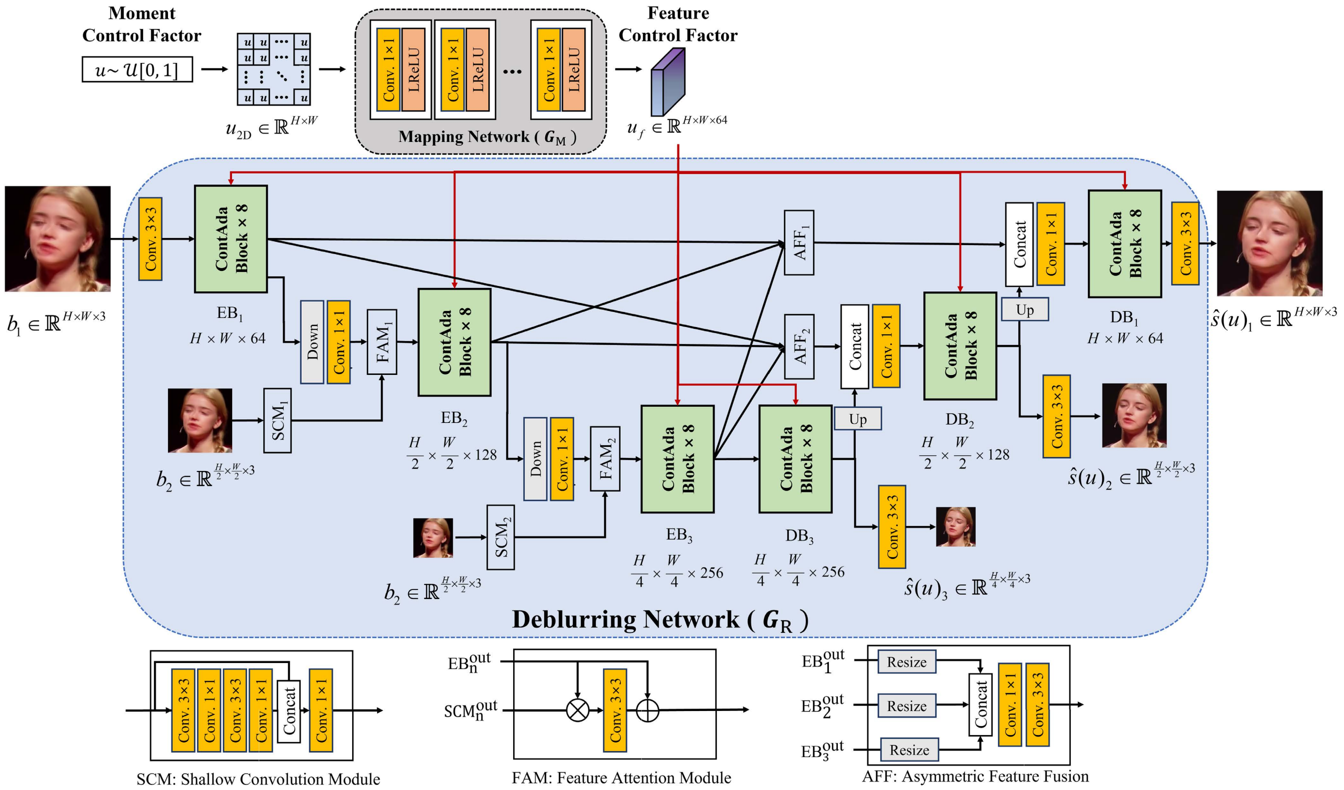

Moment Control Factor
$u \sim \mathcal{U}[0,1]$
$u_{2D} \in \mathbb{R}^{H\times W}$
Conv. 1×1
LReLU
Mapping Network (G_M)
Feature Control Factor
$u_f \in \mathbb{R}^{H\times W\times 64}$
$b_1 \in \mathbb{R}^{H\times W\times 3}$
Conv. 3×3
ContAda Block × 8
EB$_1$
$H \times W \times 64$
Down
Conv. 1×1
FAM$_1$
SCM$_1$
$b_2 \in \mathbb{R}^{\frac{H}{2}\times\frac{W}{2}\times 3}$
EB$_2$
$\frac{H}{2} \times \frac{W}{2} \times 128$
FAM$_2$
SCM$_2$
$b_3 \in \mathbb{R}^{\frac{H}{4}\times\frac{W}{4}\times 3}$
EB$_3$
$\frac{H}{4} \times \frac{W}{4} \times 256$
DB$_3$
$\frac{H}{4} \times \frac{W}{4} \times 256$
AFF$_1$
AFF$_2$
Concat
Up
DB$_2$
$\frac{H}{2} \times \frac{W}{2} \times 128$
DB$_1$
$H \times W \times 64$
$\hat{s}(u)_1 \in \mathbb{R}^{H\times W\times 3}$
$\hat{s}(u)_2 \in \mathbb{R}^{\frac{H}{2}\times\frac{W}{2}\times 3}$
$\hat{s}(u)_3 \in \mathbb{R}^{\frac{H}{4}\times\frac{W}{4}\times 3}$
Deblurring Network (G_R)
SCM: Shallow Convolution Module
EB$_n^{out}$
SCM$_n^{out}$
FAM: Feature Attention Module
EB$_1^{out}$
EB$_2^{out}$
EB$_3^{out}$
Resize
AFF: Asymmetric Feature Fusion

FIGURE 5. The architecture of our generator consisting of a mapping network and a deblurring network. In the deblurring network, the proposed control-adaptive block incorporates features of control factors and features of blurred image.

moments latent in a blurry facial image. The FMR process allows the frames in the sequence $\mathbf{S_r}$ to have regularity of face motion, which helps the network stabilize the training. The effects of the FMR are analyzed in Sec. V.

B. CONTINUOUS FACIAL MOTION DEBLURRING GAN

Inspired by the success of AC-GAN [28], the proposed continuous facial motion deblurring framework CFMD-GAN consists of a generator G and a discriminator D with an auxiliary regressor Q. An overview of the CFMD-GAN is depicted in Fig. 3. Given a blurry face image and a control factor, G performs the role of a deblurring network to perform conditional image restoration. Unlike most single image deblurring methods that only recover a single deblurred image from a single blurry image, the proposed G is a function that restores a deblurred image conditioned on a control factor. That is, G predicts continuous sharp moments latent in a blurry image by changing the value of the control factor. To achieve this, D learns to predict 1) decisions of images belonging to real or fake [62] and 2) regression for control factor at the additional output layer Q.

1) OVERALL PIPELINE OF GENERATOR

Given a blurry face image $b \in \mathbb{R}^{H\times W\times 3}$ and moment control factor $u \in [0, 1]$ as the *condition*, G generates a restored face image $\hat{s}(u) \in \mathbb{R}^{H\times W\times 3}$, which is defined as

$$\hat{s}(u) = G(b, u). \tag{7}$$

Specifically, the proposed G comprises two parts, a mapping network G_{M} and a deblurring network G_{R}. First, G_{M} translates the moment control factor $u \in [0, 1]$ into the *feature control factor* $u_f \in \mathbb{R}^{H\times W\times 64}$. Second, G_{R} incorporates u_f with features extracted from b and then outputs the final deblurred face image $\hat{s}(u)$. In the proposed deblurring network, we deign a ContAda block so that G can focus on important spatial locations and channels of features extracted from b according to c_f.

a: MAPPING NETWORK

In recent GANs studies [63]–[66], the additional mapping network has proven to provide more disentangled semantics for the generator than directly using input codes. Inspired by this, we set the mapping network G_{M} that outputs the *feature map control factor* $u_f \in \mathbb{R}^{H\times W\times 64}$ from the given moment control factor $u \in [0, 1]$ as

$$u_f = G_{\mathrm{M}}(u). \tag{8}$$

As shown in Fig. 5, G_{M} first expands u into a 2-dimensional matrix $u_{2\mathrm{D}} \in \mathbb{R}^{H\times W}$ where each position is filled with u. Then, G_{M} outputs u_f from $u_{2\mathrm{D}}$ through several convolutional layers. Similar to [63], we design G_{M} consisting of eight

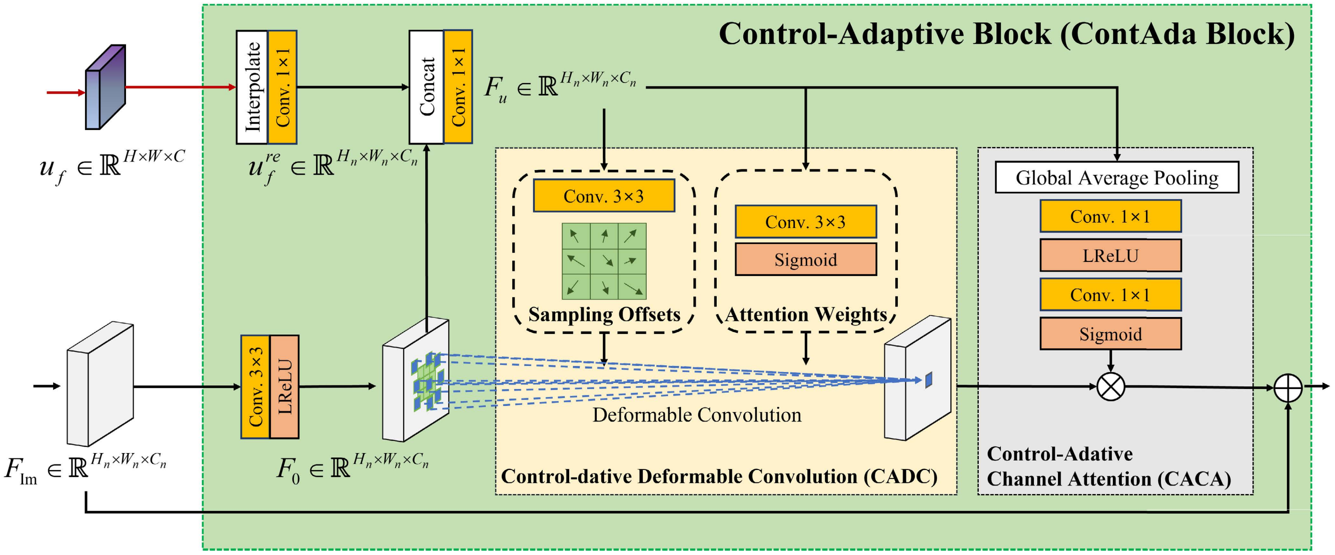

Control-Adaptive Block (ContAda Block)
Interpolate
Conv. 1×1
Concat
Conv. 1×1
$F_u \in \mathbb{R}^{H_n \times W_n \times C_n}$
$u_f \in \mathbb{R}^{H \times W \times C}$
$u_f^{re} \in \mathbb{R}^{H_n \times W_n \times C_n}$
Conv. 3×3
Sampling Offsets
Conv. 3×3
Sigmoid
Attention Weights
Global Average Pooling
Conv. 1×1
LReLU
Conv. 1×1
Sigmoid
Conv. 3×3
LReLU
Deformable Convolution
$F_{\text{Im}} \in \mathbb{R}^{H_n \times W_n \times C_n}$
$F_0 \in \mathbb{R}^{H_n \times W_n \times C_n}$
Control-dative Deformable Convolution (CADC)
Control-Adative Channel Attention (CACA)

FIGURE 6. **A structure of the proposed control-adaptive block.**

layers, each of which includes 1 × 1 convolutions and a leaky ReLU [67].

b: DEBLURRING NETWORK

As mentioned earlier, the deblurring network G_{R} generates a restored image $\hat{s}(u) \in \mathbb{R}^{H \times W \times 3}$ from the blurry face image $b \in \mathbb{R}^{H \times W \times 3}$ and the feature map control factor $u_f \in \mathbb{R}^{H \times W \times 64}$, as

$$\hat{s}(u) = G_{\text{R}}(b, u_f). \tag{9}$$

In this work, we employ the high-level structure of MIMO-UNet [49], which has exhibited impressive performance in a single image deblurring field. Specifically, as shown in Fig. 5, MIMO-Unet is based on the encoder-decoder architecture and comprises three encoder blocks (EB_1, EB_2 and EB_3) and three decoder blocks (DB_1, DB_2 and DB_3). Each of these encoder and decoder blocks contain eight modified residual blocks [46]. Unlike the original MIMO-UNet, the network developed in this study can focus on important spatial positions and channels of the feature map depending on the control factor by replacing the residual blocks with the proposed ContAda blocks. Note that SCM, FAM and AFF are modules used in the original MIMO-UNet that represent the shallow convolutional module, feature attention module and asymmetric feature fusion module, respectively. The details of each module, including the high-level architecture, can be found in [49]. In the following section, we discuss the proposed control-adaptive (ContAda) block.

2) CONTROL-ADAPTIVE BLOCK

There is a major challenge in applying existing building blocks (*e.g.* variants of residual blocks [68]) that are widely used in single image deblurring networks in the proposed continuous facial motion deblurring. First, standard convolution-based layers have an inherent drawback in modelling geometric transformations. This drawback stems from the fact that a convolutional unit samples the input feature map at fixed spatial locations [69]–[71]. To alleviate this, deformable convolution [69], [70] has exhibited promising results in object detection by learning the offsets of the convolution grid to adjust the receptive field dynamically. Inspired by this, several motion deblurring studies [72]–[74] applied a deformable convolution module to handle the complex and various latent movements in a given blurred image [72], [73]. However, these methods are still inadequate for our task because of the inability to focus on the adaptive positions of the feature maps depending on the control factor.

To this end, as shown in Fig. 6, we propose a Control-Adpative (ContAda) block that comprises a control-adaptive deformable convolution (CADC) module and a control-adaptive channel-attention (CACA) module. Let $F_{\text{Im}} \in \mathbb{R}^{H_n \times W_n \times C_n}$ denote an input feature map of the ContAda block extracted from the input blurred image $b \in \mathbb{R}^{H \times W \times 3}$. Here, H_n, W_n and C_n represent the height, width, and number of channels in the n^{th} encoder/decoder block, respectively. The ContAda block starts with a 3 × 3 convolutional layer and LeakyReLU to extract the initial feature map $F_o \in \mathbb{R}^{H_n \times W_n \times C_n}$. Meanwhile, the feature control factor $u_f \in \mathbb{R}^{H \times W \times C}$, which is the output of the mapping network G_{M}, is reshaped to $u_f^{(n)} \in \mathbb{R}^{H_n \times W_n \times C_n}$ using bilinear interpolation and 1 × 1 convolutional layer. Then, $u_f^{(n)}$ is concatenated with F_o along the channel dimension and then reshaped into $F_u \in \mathbb{R}^{H_n \times W_n \times C_n}$ by applying 1 × 1 convolution layer. F_u is utilized as an input feature for the CADC and CACA modules. In the following section, we introduce CADC and CACA distinctly.

a: CONTROL-ADAPTIVE DEFORMABLE CONVOLUTION

(CADC) module is based on deformable convolution [69], [70] that enhances the ability of network in modeling spatial variations. Unlike [69], [70], where deformable offsets and attention weights are solely determined by internal information regarding the features of the input image, the proposed CADC learns the offsets and attention weights from the combined features of the control factor and image features. Let K denote the sampling locations of a convolutional kernel. We denote the weight and pre-specified offset for the k^{th} location as w_k and p_k, respectively. For example, 3 × 3 convolutional kernel of dilation 1 has 9 sampling locations ($K = 9$) and $p_k \in \{(-1, -1), (-1, 0), \ldots, (1, 1)\}$. Let $F_u(p)$ and $F_{dc}(p)$ denote the features at location p of the input feature map F_u and output feature map F_{dc}, respectively. Accordingly, the proposed CADC can be formulated as

$$F_{dc}(p) = \sum_{k=1}^{K} w_k \cdot F_u(p + p_k + \Delta p_k) \cdot \Delta m_k, \tag{10}$$

where Δp_k and Δm_k denote the learned offset and attention weight scalar for the k^{th} location, respectively. As shown in Fig. 6, Δp_k and Δm_k are determined by separate convolutional layers. The output of the sampling offsets branch has $2K$ channels, corresponding to $\{\Delta p_k\}_{k=1}^{K}$. The output of the attention weights branch is of K channels, as $\{\Delta m_k\}_{k=1}^{K}$, and each Δm_k is in the range of [0, 1] by the sigmoid function. Following [70], the initial values of Δp_k and Δm_k are set to 0 and 0.5, respectively.

b: CONTROL-ADAPTIVE CHANNEL ATTENTION

(CACA) module is mainly motivated by [75]–[77], which benefits from applying the channel-wise attention mechanism for convolutional layers. In short, both CADC and CACA can be considered as attention functions of two variables: features extracted from blurry images and those extracted from the control factor. They are complementary in that CADC performs spatial attention to select important geometric properties of features, whereas CACA focuses on significant semantic and contextual attributes [75], [77]. Given F_u, as can be seen in Fig. 6, global average pooling is applied to transform channel-wise information into channel descriptors, following [77]. Subsequently, we obtain the channel-wise attention weights from two 1×1 convolutional layers and a sigmoid function. The learned attention weights are multiplied by F_{dc}, the output of the CADC, in an element-wise manner.

3) DISCRIMINATOR

As shown in Fig. 3, the proposed discriminator D is based on the U-net structure discriminator [62] with an auxiliary regressor. In our framework, G receives as inputs a blurred face image b and a control factor u, and outputs an image $\hat{s}(u) = G(b, u)$. Following [78], the discriminator D takes as inputs as a blurred face image and the corresponding sharp face image. Here, a face image is either a real sharp image $s(u)$ drawn from the training dataset or a restored image $\hat{s}(u)$ from G. Then, D provides three types of outputs from the encoder output layer D_{enc}, decoder output layer D_{dec}, and auxiliary regression layer Q.

Following [62], D_{enc} determines whether the global input context is real or fake. Similarly, the final outputs of D_{dec} are used to classify whether the local context of the input is sampled from the real or fake. On the other hand, the proposed Q provides a regression value for the estimated control factor. Instead of predicting a single scalar value of c, our Q outputs $\hat{u}_{2D} \in \mathbb{R}^{H\times W}$ and is trained to estimate the ground-truth control factor $u_{2D} \in \mathbb{R}^{H\times W}$.

C. MODEL OBJECTIVES

Following [55], D and G are optimized alternately using loss functions, which are described as follows.

1) DISCRIMINATOR LOSS

To estimate the global and per-pixel probability distributions, the encoder loss $\mathcal{L}_{D_{enc}}$ and decoder loss $\mathcal{L}_{D_{dec}}$ are formulated as follows:

$$\begin{aligned}\mathcal{L}_{D_{enc}} &= -\log D_{enc}(b, s(u)) + \log D_{enc}(b, G(b, u)),\\ \mathcal{L}_{D_{dec}} &= \frac{1}{WH}\sum_{i,j}^{W,H}\Big(-\log[D_{dec}(b, s(u))]_{(i,j)}\\ &\qquad + \log[D_{dec}(b, G(b, u))]_{(i,j)}\Big).\end{aligned} \tag{11}$$

Here, $[D_{dec}(\cdot)]_{(i,j)}$ represents the decision of the discriminator decoder at pixel coordinate (i, j).

To ensure that the restored image is an accurate moment of the blurry image, the auxiliary regression loss $\mathcal{L}_Q$ is defined by

$$\begin{aligned}\mathcal{L}_Q = \frac{1}{WH}\sum_{i,j}^{W,H}\Big(&\|u_{2\mathrm{D}} - Q(b, s(u))\|_2^2\\ &+ \|u_{2\mathrm{D}} - Q(b, G(b, u))\|_2^2\Big)\end{aligned} \tag{12}$$

The total loss of D is formulated as the sum of the above objectives:

$$\mathcal{L}_D = \mathcal{L}_{D_{enc}} + \mathcal{L}_{D_{dec}} + \lambda_Q \mathcal{L}_Q, \tag{13}$$

where λ_Q denotes a weight parameter, which is empirically set to 0.05.

2) GENERATOR LOSS

a: AUXILIARY REGRESSION LOSS

To accurately restore the output image conditioned by the control factor, an auxiliary regression loss $\mathcal{L}_{ar}$ is optimized as follows:

$$\mathcal{L}_{ar} = \frac{1}{WH}\sum_{i,j}^{W,H}\|u_{2\mathrm{D}} - Q(b, G(b, u))\|_2^2. \tag{14}$$

TABLE 1. Configuration of facial motion deblurring testset synthesized using 300VW dataset [5].

# of averaged frames	# of blurred images	# of sharp images
5	2753	13765
7	2677	18739
9	2605	23445
11	2530	27830
13	2493	32409
Total	13058	116188

TABLE 2. Quantitative comparison of single-to-single general deblurring methods. The best and the second best results are highlighted in bold and underline, respectively.

Methods	PSNR (↑)	SSIM (↑)	LPIPS (↓)	FID (↓)	ArcFace (↓)
Nah *et al.* [16]	31.4144	0.9232	0.0935	13.9722	1.1250
SRN [46]	32.1485	0.9249	0.0930	11.6292	1.1488
DMPHN [48]	33.1797	0.9284	0.0847	13.0407	1.0338
DMPHN*	33.8182	0.9345	0.0916	14.3071	1.0126
MIMO [49]	34.0372	0.9350	0.0795	7.8606	1.0205
MIMO*	**34.8496**	**0.9401**	<u>0.0794</u>	<u>7.2459</u>	<u>0.9918</u>
CFMD-GAN	<u>34.2684</u>	<u>0.9362</u>	**0.0697**	**5.1448**	**0.9338**

TABLE 3. Quantitative comparison of single-to-single face deblurring methods. The best and the second best results are highlighted in bold and underline, respectively.

Methods	PSNR (↑)	SSIM (↑)	LPIPS (↓)	FID (↓)	ArcFace (↓)
Shen *et al.* [18]	23.1795	0.6873	0.2310	78.3630	2.2112
UMSN [22]	27.0050	0.8276	0.1460	39.4150	1.4908
UMSN*	30.5884	0.9140	0.0833	16.8517	1.2716
MSPL-GAN [23]	28.2286	0.8936	0.1092	27.9441	1.2947
MSPL-GAN*	**34.2711**	<u>0.9359</u>	<u>0.0638</u>	<u>10.3597</u>	<u>1.0983</u>
CFMD-GAN$_{128}$	<u>33.8475</u>	**0.9379**	**0.04910**	**6.5449**	**1.0246**

b: ADVERSARIAL LOSS

We use the Unet-discriminator to ensure that the generated image is indistinguishable from the real data for both global and local contexts. The adversarial loss $\mathcal{L}_{adv}$ is formulated as follows:

$$\mathcal{L}_{adv} = -\Big(\log D_{enc}(b, G(b,u)) + \frac{1}{WH} \sum_{i,j}^{W,H} \log[D_{dec}(b, G(b,u))]_{(i,j)} \Big). \quad (15)$$

c: PIXEL-WISE LOSS

To restore accurate pixel intensities, following [79], we employ the Charbonnier loss [80] to minimize the pixel-wise distance between a ground-truth moment and a restored image as follows:

$$\mathcal{L}_{pix} = \sum_{n=1}^{3} \frac{1}{W_n H_n} \sum_{i,j}^{W_n,H_n} \sqrt{\| s(u)_n - G(b,u)_n \|^2 + \varepsilon^2}, \quad (16)$$

where n denotes the number of multi-scale levels. H_n and W_n represent the height and width at the corresponding n^{th} level of output image, respectively. Following [79], ε is set to 10^{-3}.

d: PERCEPTUAL LOSS

Furthermore, we use perceptual loss to obtain perceptually satisfactory images. Similar to [81], LPIPS [82] is employed for perceptual loss.

$$\mathcal{L}_{per} = \sum_{l}^{M} \omega^l \left\| \phi^l(s(u)) - \phi^l(G(b,u)) \right\|_2^2 \quad (17)$$

Here, $\phi(\cdot)$ is a feature extractor, ω denotes a learned vector to measure the LPIPS score, and the total score is averaged over M layers.

In overall, the total loss of G combines the aforementioned loss functions,

$$\mathcal{L}_G = \lambda_{ar}\mathcal{L}_{ar} + \lambda_{adv}\mathcal{L}_{adv} + \lambda_{pix}\mathcal{L}_{pix} + \lambda_{per}\mathcal{L}_{per}, \quad (18)$$

where λ_{ar}, λ_{adv}, λ_{pix} and λ_{per} denote the balancing weights empirically set to 0.05, 0.1, 1 and 0.01, respectively.

V. EXPERIMENTS

A. EXPERIMENTAL SETUP

1) DATASET

We use the 300VW dataset [5] which consists of a large number of high-quality facial videos recorded in the wild. Each video has a duration of about one minute at 25-30 fps. Following the face deblurring study by Ren *et al.* [54], the training and test datasets are extracted from 83 videos and 9 videos, respectively. Each blurry image is synthesized by averaging various numbers (5-13) of consecutive sharp frames, as in recent motion deblurring studies [16], [54]. Thus, the testset consists of total 13,058 blurred images and 116,188 sharp frames. The details of the number of test images are listed in Table 1.

2) IMPLEMENTATION DETAILS

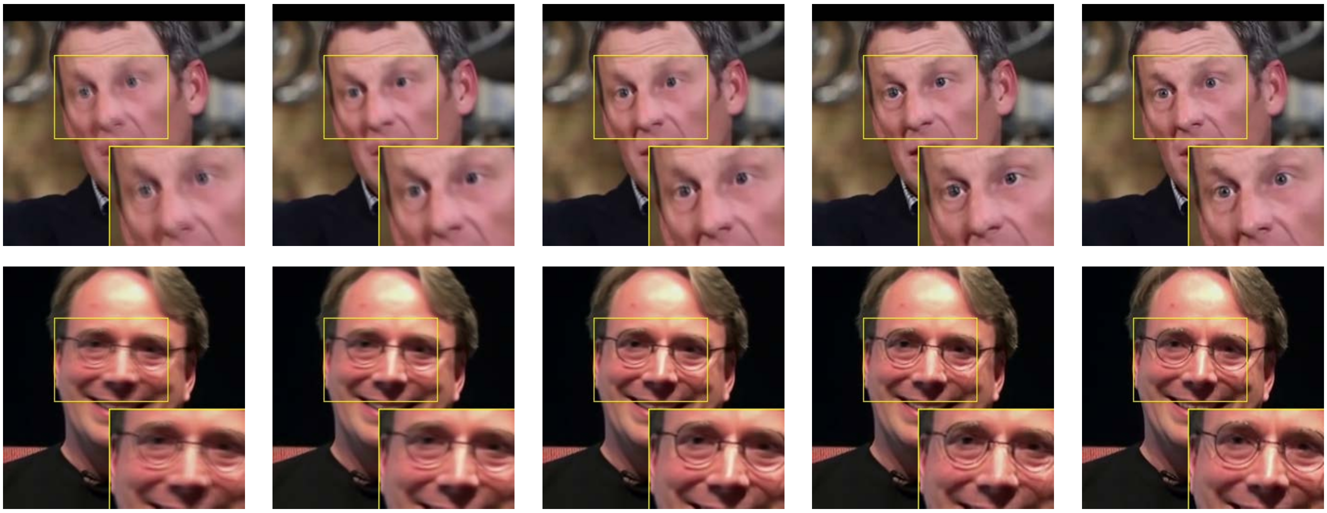
The proposed framework is implemented with Pytorch [83] and trained with NVIDIA TITAN-RTX GPUs. We train our networks using the Adam optimizer [84] with $\beta_1 = 0.9$, and $\beta_2 = 0.999$. The initial learning rate is set as 1×10^{-4} and it decayed exponentially by a factor of 0.99 for every epoch. For data augmentation, we randomly scale the image from 1.0 to 1.5 and then randomly crop the image with a spatial size of $256 \times 256 \times 3$. During training, we set the batch size as 8 and train our model for 200 epochs.

3) EVALUATION METRICS

For a quantitative evaluation, we measure the PSNR and SSIM [85], which are traditionally used for image quality assessment. We also report two metrics of learning-based perceptual quality, FID [86] and LPIPS [82]. Moreover, we employ the ArcFace [10] model to measure the distance of facial identity between the ground truth (GT) and the resulting image, as [87].

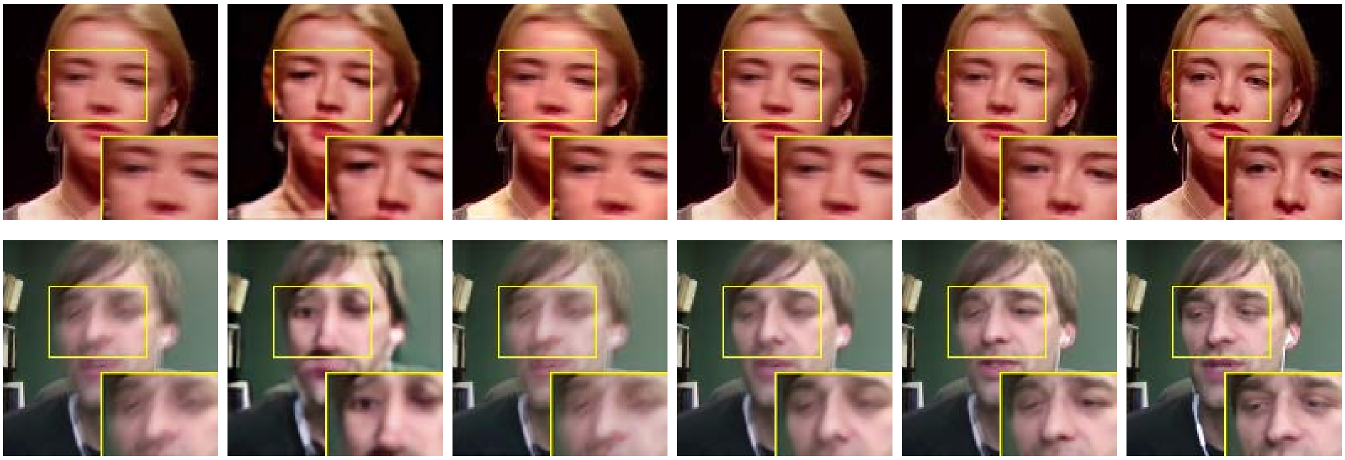
B. COMPARISONS WITH THE STATE-OF-THE-ARTS

To the best of our knowledge, the proposed method is the first attempt for single-to-video face deblurring. Hence, we conduct extensive and faithful comparisons with



Input DMPHN* [48] MIMO* [49] CFMD-GAN (ours) Ground truth

FIGURE 7. Qualitative comparisons of single-to-single general deblurring methods. Zoom in for the best view.



Input Shen *et al.* [18] UMSN* [22] MSPL-GAN* [23] CFMD-GAN$_{128}$ (ours) Ground truth

FIGURE 8. Qualitative comparisons of single-to-single face deblurring methods. Zoom in for the best view.

state-of-the-art methods in single image deblurring. Specifically, the proposed CFMD-GAN is compared with single-to-single (s2s) general deblurring (*i.e.* Nah *et al.* [16], SRN [46], DMPHN [48], MIMO [49]), s2s face deblurring (*i.e.* Shen *et al.* [18], UMSN [22], MSPL [23]), and single-to-video (s2v) general deblurring (*i.e.* Jin *et al.* [1]). To facilitate fair comparisons, we retrain the existing methods using the same training dataset used in this study. The retrained models are marked with asterisks (*). All experiments are performed using the official codes provided by the authors.

1) SINGLE-TO-SINGLE GENERAL DEBLURRING

In this comparison, we evaluate the performance of the center frame prediction, as most s2s general methods are proposed to restore the center frame. For the proposed method, the control factor is set to $c = 0.5$ to obtain the center frame results.

TABLE 4. Quantitative comparison of single-to-video deblurring methods. "# of GT" indicates the number of GT frames per a single blurry image, " # of pairs" is the total number of test GT frames, and "ALL" represents the entire results of 300VW testset. Note that all the results of CFMD-GAN are measured with the same model. The best results are highlighted in bold.

Methods	# of GT	# of pairs	PSNR (↑)	SSIM (↑)	LPIPS (↓)	FID (↓)	ArcFace (↓)
Jin *et al.* [1]	7	18739	29.2407	0.8754	0.1471	25.6946	1.1574
CFMD-GAN	7	18739	**33.1360**	**0.9336**	**0.0691**	**3.4238**	**0.9078**
CFMD-GAN	5	13765	34.6556	0.9498	0.0538	2.9080	0.7548
CFMD-GAN	9	23445	32.0300	0.9192	0.0823	4.0663	1.0384
CFMD-GAN	11	27830	31.1256	0.9060	0.0939	4.7120	1.1466
CFMD-GAN	13	32409	30.3970	0.8949	0.1041	5.3640	1.2367
CFMD-GAN	ALL	116188	31.8474	0.9153	0.0857	4.0948	1.0650

Table 2 reports the comparisons of s2s general deblurring methods. Despite the significant improvements in the performance of retrained DMPHN* and MIMO* compared to the original DMPHN and MIMO, our CFMD-GAN shows

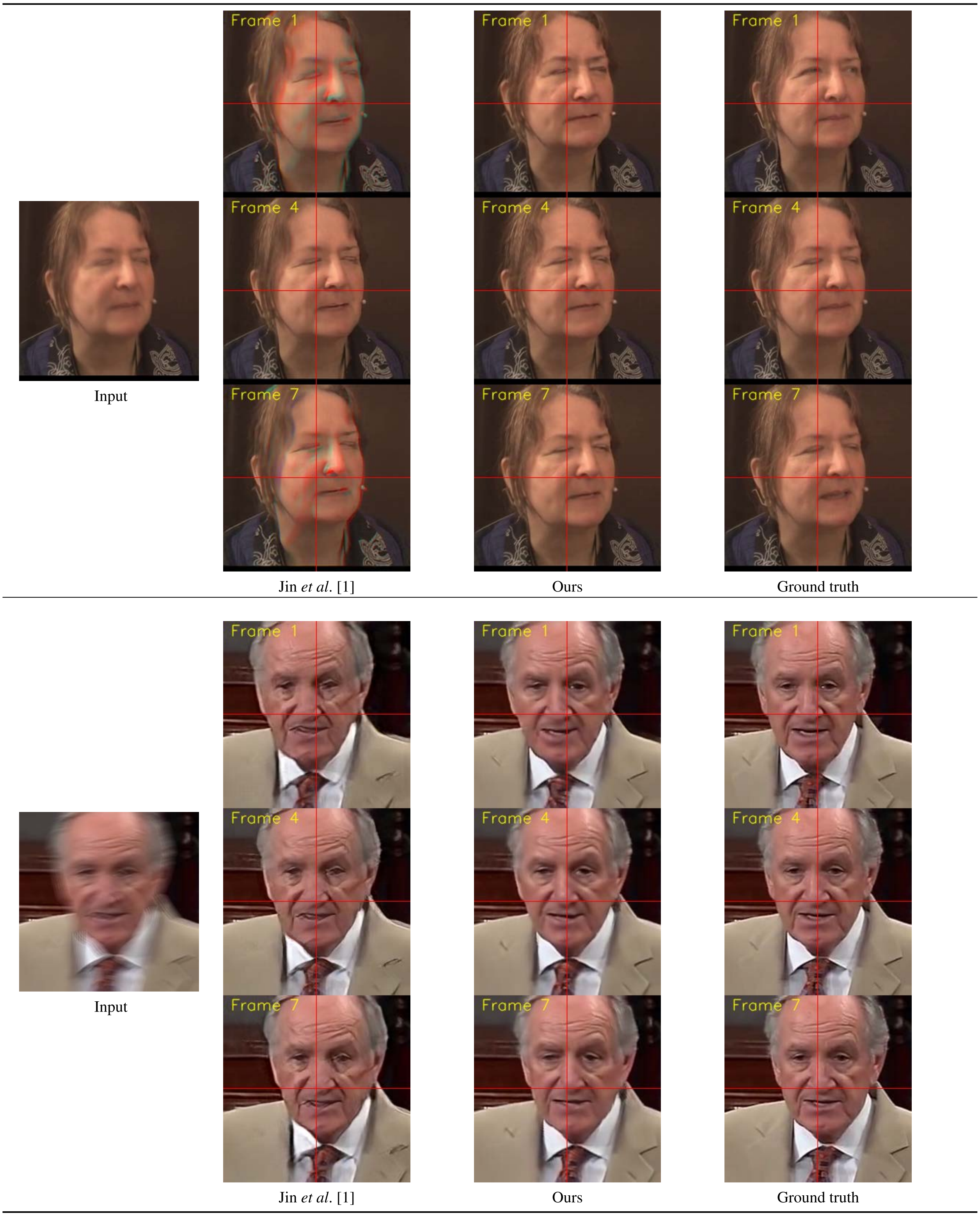


FIGURE 9. Qualitative comparisons of single-to-video deblurring methods. Due to space constraints, the initial frame (Frame 1), and the center frame (Frame 4) and the last frame (Frame 7) are displayed in this figure. To clearly observe the face movements between successive frames, horizontal and vertical lines are displayed in the center coordinates of each image. Please refer to the supplementary material for comparisons on restored video sequences.

the best results in LPIPS, FID and ArcFace distance, and the second best in PSNR and SSIM.

As investigated in recent GAN-based restoration studies [81], [87]–[92], PSNR and SSIM may be lower because

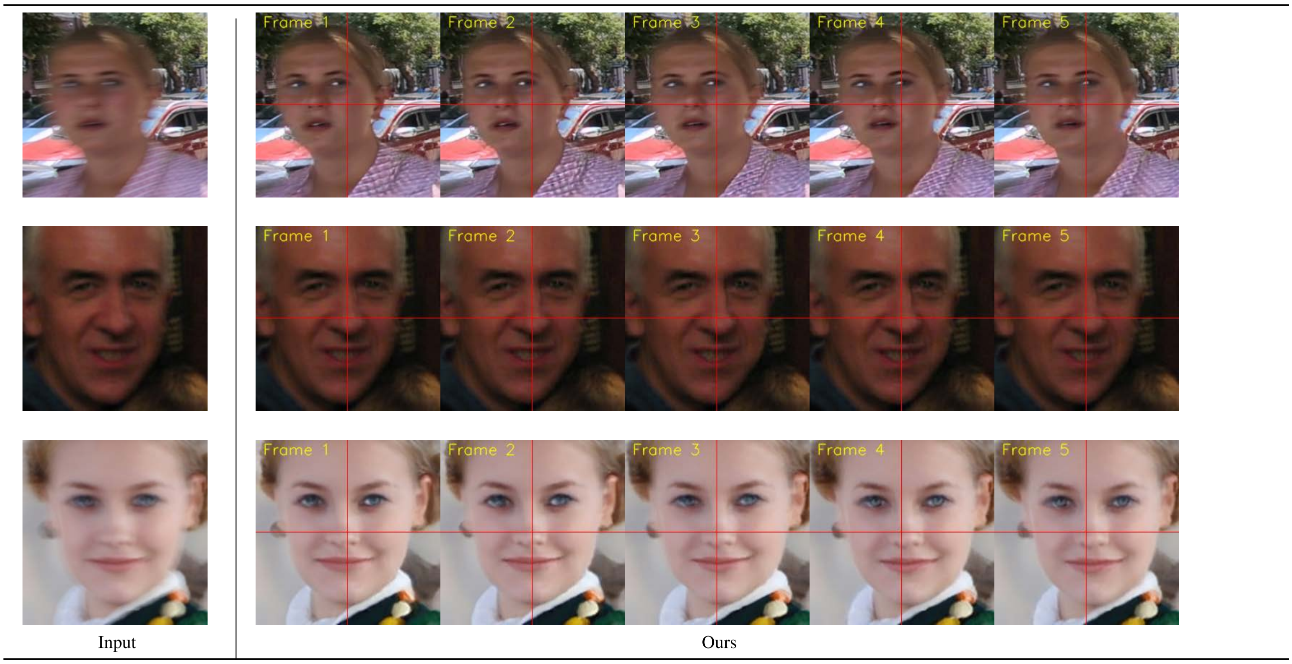


FIGURE 10. Qualitative results of the proposed CFMD-GAN on REDS dataset [93] (1^{st} row) and Lai dataset [94] (2^{nd} and 3^{rd} rows). Our resulting frames (Frame 1 to 5) are the outputs when the control factors are set to [0.0, 0.25, 0.5, 0.75, 1.0], respectively. To clearly observe the face movements between successive frames, horizontal and vertical lines are displayed in the center coordinates of each image. Please refer to the supplementary material for videos restored with various frame rates.

the GAN-based model tends to generate fake yet realistic details and textures [92]. This effect of GANs can be clearly observed in the visual comparisons in Fig. 7. Compared with other methods, the proposed CFMD-GAN restores more realistic textures and finer details of facial components, such as the eyes, nose, and eyelids. Based on these results, we can confirm that the proposed model can predict a more accurate center frame than the other methods.

2) SINGLE-TO-SINGLE FACE DEBLURRING

Most existing s2s face deblurring methods [18], [22], [23] are developed to remove spatially-uniform blurs. However, our training and test datasets contain spatially-variant blurs. Besides, their models only handle input images of 128 × 128 × 3. For these reasons, we downsample our dataset to 128 × 128 × 3 and use it to retrain UMSN [22], MSPL [23] and our model (termed as CFMD-GAN$_{128}$). The retrained models, UMSN* and MSPL*, are trained to predict the center frame, similar to the s2s general deblurring approaches. Note that we do not retrain Shen *et al.* [18] because they do not release the training code.

Table 3 and Fig. 8 provide the quantitative and qualitative comparisons of the s2s face deblurring methods, respectively. In this experiment, the proposed method achieved significantly better performance on SSIM, LPIPS, FID and ArcFace than the existing face deblurring methods. For PSNR, our method achieved the second best. Shen *et al.* [18] fails to restore plausible results because they are not trained to remove spatially-variant blurs, as shown in Fig. 8. Although the retrained models (UMSN* and MSPL*) show improved performance, they are still inferior to the CFMD-GAN.

3) SINGLE-TO-VIDEO GENERAL DEBLURRING

For s2v general deblurring methods, we compare our method with Jin *et al.* [1] which officially released their test model. Since this method is strictly fixed to extract seven sequential frames from a single blurry image, we compare the results only for blurry images averaged by seven sharp frames. None of the s2v deblurring methods [1]–[4] have released their training codes. [1] is the only work that provides the test code.

Table 4 reports quantitative comparisons with Jin *et al.* [1] and detailed results of our model according to the number of GT frames. The model of Jin *et al.* [1] is limited to predicting only a fixed number of frames when the model is trained once. However, it is worth to note that the proposed single model can predict various numbers of output frames without additional network changes or training processes.

Visual comparisons are shown in Fig. 9. Among the restored sequences, the initial, central and last frames are reported due to space constraints. The results for the entire restored frames are provided in the supplementary material. The results of the proposed method are visually more plausible than those of Jin *et al.* [1].

C. ANALYSIS ON CFMD-GAN

1) EVALUATION ON OTHER TEST DATASETS

Since our model is trained and evaluated with synthetically blurred images using the 300VW dataset [5], we verify how our model performs on other motion-blur benchmark datasets such as REDS [93] and Lai *et al.* [94]. The REDS dataset is generated using 120 fps videos, synthesizing blurry frames by merging consecutive frames. The Lai dataset contains real-blur images where the GT images do not exist.

TABLE 5. **Ablations on the proposed ContAda block. The best results are highlighted in bold.**

CADC	CACA	FMR	PSNR (↑)	SSIM (↑)	LPIPS (↓)	FID (↓)	ArcFace (↓)
✓		✓	33.5546	0.9263	0.0814	7.0030	1.0457
	✓	✓	34.0271	0.9279	0.0810	6.8322	1.0389
✓	✓		33.4478	0.9201	0.0880	9.6110	1.1547
✓	✓	✓	**34.2684**	**0.9362**	**0.0697**	**5.1448**	**0.9338**

We manually crop the facial regions of images in the REDS validation set and the Lai dataset.

Fig. 10 shows that our method restores satisfactory images for recent benchmark deblurring datasets. In 1^{st} row of Fig. 10, we can see that our method produces not only a sharp face, but also the background that was occluded by the face in the previous frame. For the real-blurred images in 2^{nd} row of Fig. 10, our model restores plausible results containing consecutive frames. Our framework can provide all sharp moments that user wants from a single motion-blurred face image.

2) ABLATION STUDY

In Table 5, we evaluate the impact of the proposed ContAda block consisting of ContAda deformable convolution (CADC) and ContAda channel attention (CACA). With the CADC module, the proposed method can focus on the spatially important sampling points of the feature maps by the feature map control factor. Notably, using only CACA module improves the average PSNR by about 0.5dB compared to using only CADC module. This demonstrates that the channel attention plays a more important role in the proposed model. More importantly, using both CADC and CACA achieves the best results. This indicates that both spatial and channel-wise modulations are required for the continuous facial motion deblurring. Furthermore, we conduct an ablation study to investigate the contribution of FMR to the network training. The 3^{rd} row in Table 5 indicates that without FMR, the performance of the model drops drastically when it learns the original temporal order.

VI. CONCLUSION

In this study, we introduce CFMD-GAN, a novel framework for continuous facial motion deblurring with a single network and a single training process. Subsequently, we apply facial motion-based reordering (FMR) to mitigate the difficulty of temporal ordering by utilizing domain-specific facial information. This ensures a stable learning process for the framework. We devise an auxiliary regressor to learn continuous motion deblurring by integrating the concept of conditional GANs into a single image deblurring framework. In addition, we propose a control-adaptive (ContAda) block that focuses on deformable locations and important channels according to the control factor. In our extensive experiments, we demonstrate that the proposed method outperforms state-of-the-art methods in facial image deblurring. The proposed framework can provide continuous sharp moments that users want to obtain from a single motion-blurred facial image. Since the proposed method restores facial motion in the order of FMR, there may be a limitation in predicting the accurate temporal order of the facial motion. However, we believe that the proposed method will be the basis for future studies on continuous facial motion deblurring. In addition, incorporating various facial priors can be a fundamental issue for future research to improve the quality of this study.

REFERENCES

[1] M. Jin, G. Meishvili, and P. Favaro, "Learning to extract a video sequence from a single motion-blurred image," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2018, pp. 6334–6342.

[2] K. Purohit, A. Shah, and A. N. Rajagopalan, "Bringing alive blurred moments," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 6830–6839.

[3] D. M. Argaw, J. Kim, F. Rameau, C. Zhang, and I. S. Kweon, "Restoration of video frames from a single blurred image with motion understanding," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Jun. 2021, pp. 701–710.

[4] K. Zhang, W. Luo, B. Stenger, W. Ren, L. Ma, and H. Li, "Every moment matters: Detail-aware networks to bring a blurry image alive," in *Proc. 28th ACM Int. Conf. Multimedia*, Oct. 2020, pp. 384–392.

[5] J. Shen, S. Zafeiriou, G. G. Chrysos, J. Kossaifi, G. Tzimiropoulos, and M. Pantic, "The first facial landmark tracking in-the-wild challenge: Benchmark and results," in *Proc. IEEE Int. Conf. Comput. Vis., Workshop (ICCVW)*, Dec. 2015, pp. 50–58.

[6] Y. Sun, X. Wang, and X. Tang, "Deep convolutional network cascade for facial point detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2013, pp. 3476–3483.

[7] 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.

[8] F. Saeed, M. J. Ahmed, M. J. Gul, K. J. Hong, A. Paul, and M. S. Kavitha, "A robust approach for industrial small-object detection using an improved faster regional convolutional neural network," *Sci. Rep.*, vol. 11, no. 1, pp. 1–13, Dec. 2021.

[9] H. Wang, Y. Wang, Z. Zhou, X. Ji, D. Gong, J. Zhou, Z. Li, and W. Liu, "CosFace: Large margin cosine loss for deep face recognition," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2018, pp. 5265–5274.

[10] 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.

[11] H. Yang, U. Ciftci, and L. Yin, "Facial expression recognition by de-expression residue learning," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2018, pp. 2168–2177.

[12] Z. Zhang, P. Luo, C. C. Loy, and X. Tang, "From facial expression recognition to interpersonal relation prediction," *Int. J. Comput. Vis.*, vol. 126, no. 5, pp. 550–569, May 2018.

[13] K. Sanjar, O. Bekhzod, J. Kim, J. Kim, A. Paul, and J. Kim, "Improved U-Net: Fully convolutional network model for skin-lesion segmentation," *Appl. Sci.*, vol. 10, no. 10, p. 3658, May 2020.

[14] J. Kim, J. K. Lee, and K. M. Lee, "Accurate image super-resolution using very deep convolutional networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2016, pp. 1646–1654.

[15] K. Zhang, W. Zuo, Y. Chen, D. Meng, and L. Zhang, "Beyond a Gaussian denoiser: Residual learning of deep CNN for image denoising," *IEEE Trans. Image Process.*, vol. 26, no. 7, pp. 3142–3155, Jul. 2017.

[16] 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.

[17] S. Din, A. Paul, and A. Ahmad, "Lightweight deep dense demosaicking and denoising using convolutional neural networks," *Multimedia Tools Appl.*, vol. 79, nos. 45–46, pp. 34385–34405, Dec. 2020.

[18] 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.

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

[20] 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.

[21] Z. Shen, T. Xu, J. Zhang, J. Guo, and S. Jiang, "A multi-task approach to face deblurring," *EURASIP J. Wireless Commun. Netw.*, vol. 2019, no. 1, pp. 1–11, Jan. 2019.

[22] 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.

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

[24] S. H. Jung, T. B. 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.

[25] M. Hirsch, C. J. Schuler, S. Harmeling, and B. Schölkopf, "Fast removal of non-uniform camera shake," in *Proc. IEEE Int. Conf. Comput. Vis.*, Nov. 2011, pp. 463–470.

[26] K. Sun, W. Wu, T. Liu, S. Yang, Q. Wang, Q. Zhou, Z. Ye, and C. Qian, "FAB: A robust facial landmark detection framework for motion-blurred videos," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 5462–5471.

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

[28] A. Odena, C. Olah, and J. Shlens, "Conditional image synthesis with auxiliary classifier GANs," in *Proc. 34th Int. Conf. Mach. Learn. (ICML)*, Aug. 2017, pp. 2642–2651.

[29] T. Miyato and M. Koyama, "cGANs with projection discriminator," in *Proc. Int. Conf. Learn. Represent.*, Jan. 2018, pp. 1–21.

[30] A. Brock, J. Donahue, and K. Simonyan, "Large scale GAN training for high fidelity natural image synthesis," in *Proc. Int. Conf. Learn. Represent.*, May 2019, pp. 1–35.

[31] H. Zhang, I. Goodfellow, D. Metaxas, and A. Odena, "Self-attention generative adversarial networks," in *Proc. Int. Conf. Mach. Learn.*, Jun. 2019, pp. 7354–7363.

[32] X. Ding, Y. Wang, Z. Xu, W. J. Welch, and Z. J. Wang, "CcGAN: Continuous conditional generative adversarial networks for image generation," in *Proc. Int. Conf. Learn. Represent.*, May 2021, pp. 1–30. [Online]. Available: https://openreview.net/forum?id=PrzjugOsDeE

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

[34] M. Kang and J. Park, "ContraGAN: Contrastive learning for conditional image generation," in *Proc. Adv. Neural Inf. Process. Syst. (NIPS)*, Dec. 2020, pp. 1–31.

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

[36] J. He, C. Dong, and Y. Qiao, "Interactive multi-dimension modulation with dynamic controllable residual learning for image restoration," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*. Cham, Switzerland: Springer, Nov. 2020, pp. 53–68.

[37] H. Kim, S. Baik, M. Choi, J. Choi, and K. M. Lee, "Searching for controllable image restoration networks," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 14234–14243.

[38] H. Cai, J. He, Y. Qiao, and C. Dong, "Toward interactive modulation for photo-realistic image restoration," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Jun. 2021, pp. 294–303.

[39] T. H. Kim, B. Ahn, and K. M. Lee, "Dynamic scene deblurring," in *Proc. IEEE Int. Conf. Comput. Vis.*, Dec. 2013, pp. 3160–3167.

[40] M. Noroozi, P. Chandramouli, and P. Favaro, "Motion deblurring in the wild," in *Proc. German Conf. Pattern Recognit. (GCPR)*. Cham, Switzerland: Springer, Aug. 2017, pp. 65–77.

[41] S. Su, M. Delbracio, J. Wang, G. Sapiro, W. Heidrich, and O. Wang, "Deep video deblurring for hand-held cameras," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jul. 2017, pp. 1279–1288.

[42] Z. Shen, W. Wang, X. Lu, J. Shen, H. Ling, T. Xu, and L. Shao, "Human-aware motion deblurring," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, Oct. 2019, pp. 5572–5581.

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

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

[45] 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.*, Jun. 2017, pp. 2319–2328.

[46] X. Tao, H. Gao, Y. Wang, X. Shen, J. Wang, and J. Jia, "Scale-recurrent network for deep image deblurring," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Feb. 2018, pp. 8174–8182.

[47] 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.

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

[49] 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*.

[50] 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.

[51] 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.

[52] 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.

[53] 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.

[54] 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.

[55] 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.*, Jun. 2014, pp. 2672–2680.

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

[57] 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.

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

[59] A. Dosovitskiy, J. T. Springenberg, M. Tatarchenko, and T. Brox, "Learning to generate chairs, tables and cars with convolutional networks," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 39, no. 4, pp. 692–705, Apr. 2017.

[60] A. Kumar, P. Sattigeri, and T. Fletcher, "Semi-supervised learning with GANs: Manifold invariance with improved inference," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 30, Dec. 2017, pp. 1–11.

[61] G. Bradski, "The OpenCV library," *Dr. Dobb's J., Softw. Tools Prof. Programmer*, vol. 25, no. 11, pp. 120–123, 2000.

[62] 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.

[63] T. Karras, S. Laine, and T. Aila, "A style-based generator architecture for generative adversarial networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2019, pp. 4401–4410.

[64] Y. Shen, J. Gu, X. Tang, and B. Zhou, "Interpreting the latent space of GANs for semantic face editing," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 9243–9252.

[65] E. Härkönen, A. Hertzmann, J. Lehtinen, and S. Paris, "GANSpace: Discovering interpretable GAN controls," 2020, *arXiv:2004.02546*.

[66] J. Zhu, Y. Shen, D. Zhao, and B. Zhou, "In-domain GAN inversion for real image editing," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*. Cham, Switzerland: Springer, Aug. 2020, pp. 592–608.

[67] A. L. Maas, A. Y. Hannun, and A. Y. Ng, "Rectifier nonlinearities improve neural network acoustic models," in *Proc. ICML*, 2013, vol. 30, no. 1, pp. 1–6.

[68] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Feb. 2016, pp. 770–778.

[69] J. Dai, H. Qi, Y. Xiong, Y. Li, G. Zhang, H. Hu, and Y. Wei, "Deformable convolutional networks," in *Proc. IEEE Int. Conf. Comput. Vis.*, Oct. 2017, pp. 764–773.
[70] X. Zhu, H. Hu, S. Lin, and J. Dai, "Deformable ConvNets V2: More deformable, better results," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 9308–9316.
[71] X. Zhu, W. Su, L. Lu, B. Li, X. Wang, and J. Dai, "Deformable DETR: Deformable transformers for end-to-end object detection," 2020, *arXiv:2010.04159*.
[72] X. Wang, K. C. K. Chan, K. Yu, C. Dong, and C. C. Loy, "EDVR: Video restoration with enhanced deformable convolutional networks," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Jun. 2019, pp. 1–10.
[73] K. Purohit and A. Rajagopalan, "Region-adaptive dense network for efficient motion deblurring," in *Proc. AAAI*, Feb. 2020, vol. 34, no. 7, pp. 11882–11889.
[74] Y. Yuan, W. Su, and D. Ma, "Efficient dynamic scene deblurring using spatially variant deconvolution network with optical flow guided training," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 3555–3564.
[75] L. Chen, H. Zhang, J. Xiao, L. Nie, J. Shao, W. Liu, and T.-S. Chua, "SCA-CNN: Spatial and channel-wise attention in convolutional networks for image captioning," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Aug. 2017, pp. 5659–5667.
[76] J. Hu, L. Shen, and G. Sun, "Squeeze-and-excitation networks," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 7132–7141.
[77] Y. Zhang, K. Li, K. Li, L. Wang, B. Zhong, and Y. Fu, "Image super-resolution using very deep residual channel attention networks," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, Sep. 2018, pp. 286–301.
[78] P. Isola, J.-Y. Zhu, T. Zhou, and A. A. Efros, "Image-to-image translation with conditional adversarial networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2017, pp. 1125–1134.
[79] 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*.
[80] P. Charbonnier, L. Blanc-Féraud, G. Aubert, and M. Barlaud, "Two deterministic half-quadratic regularization algorithms for computed imaging," in *Proc. IEEE Int. Conf. Image Process. (ICIP)*, vol. 2, Sep. 1994, pp. 168–172.
[81] 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.
[82] 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.
[83] A. Paszke *et al.*, "PyTorch: An imperative style, high-performance deep learning library," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 32, Dec. 2019, pp. 8026–8037.
[84] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," in *Proc. 3rd Int. Conf. Learn. Represent. (ICLR)*, Y. Bengio and Y. LeCun, Eds., May 2015, pp. 1–15.
[85] 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.
[86] M. Heusel, H. Ramsauer, T. Unterthiner, B. Nessler, and S. Hochreiter, "GANs trained by a two time-scale update rule converge to a local Nash equilibrium," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 30, Dec. 2017, pp. 1–12.
[87] 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.
[88] 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.
[89] Y. Blau, R. Mechrez, R. Timofte, T. Michaeli, and L. Zelnik-Manor, "The 2018 PIRM challenge on perceptual image super-resolution," in *Proc. Eur. Conf. Comput. Vis. Workshops (ECCVW)*, Sep. 2018, pp. 1–22.
[90] 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. Workshops (ECCVW)*, 2018, pp. 1–16.
[91] 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.
[92] J. Gu, H. Cai, C. Dong, J. S. Ren, Y. Qiao, S. Gu, and R. Timofte, "NTIRE 2021 challenge on perceptual image quality assessment," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Jun. 2021, pp. 677–690.
[93] S. Nah, S. Baik, S. Hong, G. Moon, S. Son, R. Timofte, and K. M. Lee, "NTIRE 2019 challenge on video deblurring and super-resolution: Dataset and study," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Jun. 2019, pp. 1–10.
[94] 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.

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.

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 with Samsung Electronics, Digital Media and Communications Research and Development Center. Currently, he is 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.