Deep Feature Prior Guided Face Deblurring

Soo Hyun Jung[1], Tae Bok Lee[2], and Yong Seok Heo[1,2]
[1]Department of Electrical and Computer Engineering, Ajou University, South Korea
[2]Department of Artificial Intelligence, Ajou University, South Korea
{tngusdldlt,dolphin0104,ysheo}@ajou.ac.kr

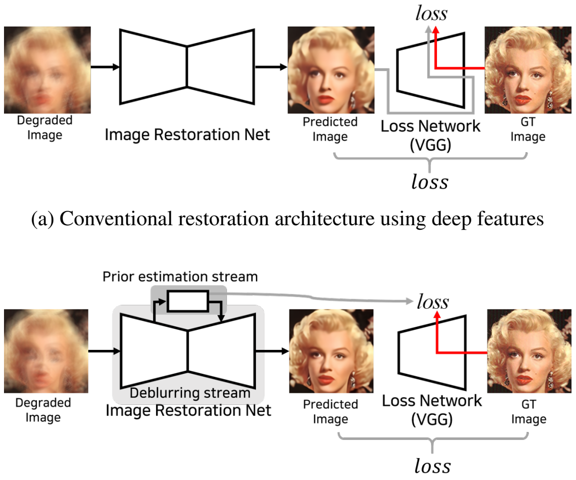
Abstract

Most recent face deblurring methods have focused on utilizing facial shape priors such as face landmarks and parsing maps. While these priors can provide facial geometric cues effectively, they are insufficient to contain local texture details that act as important clues to solve face deblurring problem. To deal with this, we focus on estimating the deep features of pre-trained face recognition networks (e.g., VGGFace network) that include rich information about sharp faces as a prior, and adopt a generative adversarial network (GAN) to learn it. To this end, we propose a deep feature prior guided network (DFPGnet) that restores facial details using the estimated the deep feature prior from a blurred image. In our DFPGnet, the generator is divided into two streams including prior estimation and deblurring streams. Since the estimated deep features of the prior estimation stream are learned from the VGGFace network which is trained for face recognition not for deblurring, we need to alleviate the discrepancy of feature distributions between the two streams. Therefore, we present feature transform modules at the connecting points of the two streams. In addition, we propose a channel-attention feature discriminator and prior loss, which encourages the generator to focus on more important channels for deblurring among the deep feature prior during training. Experimental results show that our method achieves state-of-the-art performance both qualitatively and quantitatively.

1. Introduction

Single facial image deblurring is to recover facial details from a blurred face image. The recovery of a sharp face image from a blurred face image is important for many computer vision tasks, such as face recognition [36, 6], face alignment [12, 2, 48], and face detection [34, 45, 19]. This is because performance degrades dramatically owing to the complex blur scenarios in which important shape and texture details of the face image often disappear.

As image deblurring is highly ill-posed, an appropriate prior is necessary to constrain the solution space. Fortunately, face deblurring can utilize strong prior knowledge of the face even for severe blur. Hence, many recent studies have exploited various facial priors, such as face landmark [5, 4], 2D face sketch [21], and face segmentation [32, 42, 18]. However, these priors are highly biased toward the global shape of the face rather than texture details [3]. They are insufficient for providing a high-dimensional texture, because they contain only location information [3].



(a) Conventional restoration architecture using deep features

(b) Proposed restoration architecture using deep features

Figure 1. **Illustration of the conventional and the proposed restoration architecture based on deep features from deep CNN model.** (a) The conventional architecture is designed to optimize both the perceptual loss for perceptual similarity and the pixel-wise loss for pixel-wise accuracy in a single restoration network. (b) The proposed architecture is designed to predict deep features in a separate stream and utilize them as a prior.

To restore the texture details, face deblurring approaches [32, 23, 42, 18] often rely on the perceptual loss function [11]. This loss function is based on the fact that features from deep convolutional neural networks (*e.g.*, VGGFace [27]) trained for high-level tasks (*e.g.*, face recognition), the so called *deep features* [46], contain rich representations of sharp face images, such as edges, textures, and semantic information [29]. They extract deep features from specific layers of this network to minimize the distance in the feature space between the predicted and ground truth images. Although it is helpful to generate realistic textures, it still may not match the ground truth images in pixel-wise accuracy [31]. Thus, it is conventional to optimize both the perceptual loss in feature space for perceptual similarity and the pixel-wise loss in image space for pixel-wise accuracy in a single restoration network, as shown in Fig. 1a.

However, minimizing these loss terms in a single restoration network may not be the optimal solution for image restoration. This is because image restoration and image recognition are different processes. The deep features from well-trained classifiers primarily contain useful information for recognition, not restoration. This means that some features of the pre-trained recognition network hinder the restoration network from learning to generate accurate appearances, because the features of the recognition network are designed to be robust to intra-class variations in appearance. This conflicts with the purpose of restoration by rendering the restoration network insensitive to appearance variations [24] such as blur. Due to this inherent characteristics of the deep features, it is not effective to reduce both the pixel-wise loss and the perceptual loss without considering the consistency of the purpose of deblurring in a single network. Thus, we argue that disentangling these representations by careful design, and focusing on information that is useful for deblurring among the deep features can lead to more accurate deblurring results.

In this work, we therefore propose a novel deep feature prior guided network (DFPGnet), that separates two streams, including a deep feature prior estimation stream and a deblurring stream in a end-to-end single restoration network, as shown in Fig. 1b. This enables the prediction of important deep features with the supervision of ground truth deep features from the VGGFace [27] in the prior estimation stream, and utilizing them as facial prior to restore facial details in the deblurring stream. By separating the two streams in a single network, they can focus on their own roles. For the information flow between two streams, we connect the input and output features of the prior estimation stream with the intermediate features of encoder and decoder in the deblurring stream, respectively. However, the appropriate feature transformations are required at the connecting points of the two streams due to the discrepancy of feature distributions between two streams. Thus, we place a self-spatial feature transform (SSFT) module which is a modified version of a spatial feature transform (SFT) module and a SFT module [46] at that points, respectively. In addition, we propose a channel-attention feature discriminator and prior loss based on generative adversarial networks (GANs) framework [7] to focus only on important information for deblurring among the deep features. The key idea of our discriminator is to learn the weight of important channels of the deep features for deblurring by using the channel attention (CA) module [9] in an unsupervised manner. Using the learned weight, we can define the prior loss that encourages our generator to focus on more important channels for deblurring among the deep features. We show that our method using our deep feature prior outperforms existing state-of-the-art methods that use the perceptual loss and the shape priors on various datasets.

Our contributions can be summarized as follows.

- We propose to incorporate the deep features as a prior in a restoration network.
- We propose a channel attention feature discriminator that enables the generator to focus on more important channels for deblurring among the deep feature prior.
- We achieve state-of-the-art performance for face deblurring against existing method and provide ablation studies to demonstrate the power of deep feature prior.

2. Related Works

Generic Image Deblurring. Recently, numerous deblurring methods based on deep learning (DL) [16] have been studied and achieved excellent performance. Nah *et al.* [25] proposed an end-to-end learning method that directly estimates the deblurred output using a multi-scale CNN to gradually restore sharp images. Since the seminal work of [25], various DL-based methods [35, 43] that predict images have directly contributed to improving the deblurring performance. While all of these deblurring methods are very accurate in generic image deblurring, they do not generalize well to domain-specific deblurring, such as text and face images.

Facial Image Deblurring. For face deblurring, most existing methods take advantage of facial geometric priors such as face landmarks [5, 4], facial parsing maps [32, 42, 18], and 3D facial model [30]. Chrysos *et al.* [5] introduced a deep architecture along with the pre-processing step to take advantage of the facial structure through landmark localization. Shen *et al.* [32] proposed a two-stage face deblurring network that generates facial parsing maps first and restores the facial images later. Yasarla *et al.* [42] suggested measuring the uncertainty of the estimated parsing map to address the side effects of inaccurate parsing maps. Lee *et al.* [18] proposed a multi-semantic progressive learning to exploit ground truth parsing maps for training purposes. While these methods are effective at providing geometric information of the facial components, they are insufficient for providing texture information, such as detailed edges or other low-level image contexts [30]. In contrast, we propose to utilize a deep feature prior that contains both the geometric and texture information of the face. Although deep

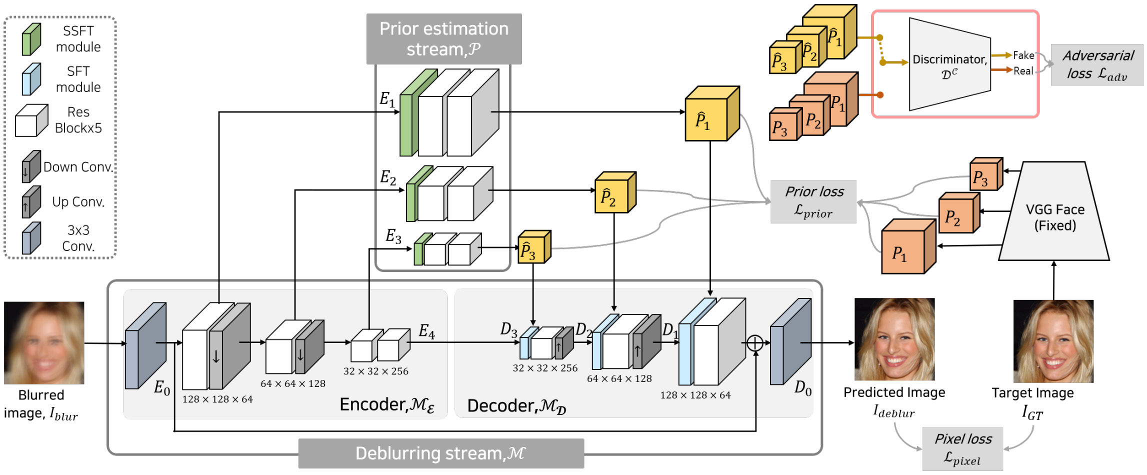

SSFT module
SFT module
Res Blockx5
Down Conv.
Up Conv.
3x3 Conv.
Prior estimation stream,$\mathcal{P}$
E_1
E_2
E_3
$\hat{P}_1$
$\hat{P}_2$
$\hat{P}_3$
P_1
P_2
P_3
Discriminator, $\mathcal{D}^{\mathcal{C}}$
Fake
Real
Adversarial loss $\mathcal{L}_{adv}$
Prior loss $\mathcal{L}_{prior}$
VGG Face (Fixed)
Blurred image, I_{blur}
E_0
$128 \times 128 \times 64$
$64 \times 64 \times 128$
$32 \times 32 \times 256$
E_4
Encoder,$\mathcal{M}_{\mathcal{E}}$
D_3
D_2
D_1
$32 \times 32 \times 256$
$64 \times 64 \times 128$
$128 \times 128 \times 64$
Decoder,$\mathcal{M}_{\mathcal{D}}$
D_0
Deblurring stream,$\mathcal{M}$
Predicted Image I_{deblur}
Target Image I_{GT}
Pixel loss $\mathcal{L}_{pixel}$

Figure 2. The overall scheme of our method.

dictionary feature prior [20] and generative prior [37] are recently utilized for texture generation in blind face restoration, they often fail to handle severely blurred face image.

Perceptual Loss for Restoration. Recently, the perceptual loss [11] has been widely used to reconstruct fine edges and textures in many image restoration studies [17, 39, 32, 29, 42, 18]. Subsequently, existing face deblurring methods [32, 23, 42, 18] also employed perceptual loss to capture facial specific features using face recognition model [27]. However, as image restoration and recognition task involve different processes, the perceptual loss is limited in providing optimal solutions for accurate restoration. Differently, we propose to estimate the rich information of deep features in separate processing streams and effectively integrate them into the restoration task.

3. Proposed Method

We propose a face deblurring framework, called **D**eep **F**eature **P**rior **G**uided network (DFPGnet) that restores a sharp face image I_{deblur} from a blurred face image I_{blur} with the help of the estimated deep feature prior. Our key idea is to estimate a strong prior containing the shape and texture information of the human face. For this, we adopt the VGGFace [27] model trained for face recognition with a sharp face dataset to extract the ground truth (GT) deep feature prior from the sharp face image.

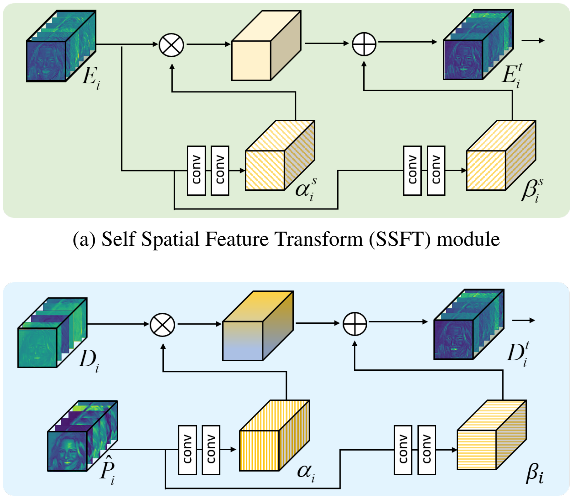
As shown in Fig. 2, our DFPGnet is based on generative adversarial networks (GANs) [7] that consists of a generator $\mathcal{G}$ and a discriminator $\mathcal{D}^{\mathcal{C}}$. The generator consists of three parts: an encoder of the deblurring stream $\mathcal{M}_{\mathcal{E}}$, a prior estimation stream $\mathcal{P}$, and a decoder of the deblurring stream $\mathcal{M}_{\mathcal{D}}$. First, $\mathcal{M}_{\mathcal{E}}$ extracts the features from I_{blur}. Then, $\mathcal{P}$ estimates the deep features prior from intermediate features of $\mathcal{M}_{\mathcal{E}}$. Finally, $\mathcal{M}_{\mathcal{D}}$ generates I_{deblur} by exploiting the output deep features of $\mathcal{P}$ as prior information. Although the GT deep features include rich information on human faces, they are designed for recognition not deblurring. Thus, there is a discrepancy of the feature distributions between the two different tasks of deblurring and recognition. For this reason, we present the self-spatial feature transform (SSFT) and spatial feature transform (SFT) [38] modules at the connecting point between the deblurring stream and the prior estimation stream. In addition, channel attention (CA) modules [9] are adopted when training the discriminator and the generator to emphasize important channels for image restoration task among the channels of the learned features using the VGGFace [27].

3.1. Encoder of Deblurring Stream

The encoder $\mathcal{M}_{\mathcal{E}}$ aims to extract features for deblurring. It takes a blurred face image I_{blur} as input and produces a set of multiple output features $\mathbb{E} = \{E_i | i = 0, 1, 2, 3, 4\}$. We group the intermediate outputs except for the first and last features into a subset $\mathbb{E}_P = \{E_i | i = 1, 2, 3\}$. Thus, we can formulate $[E_0, \mathbb{E}_P, E_4] = \mathcal{M}_{\mathcal{E}}(I_{blur})$. The first output E_0 from a blurred image is added to the last feature D_0 of $\mathcal{M}_{\mathcal{D}}$ for a sharp image using a global skip connection. The output $\mathbb{E}_P$ is passed into $\mathcal{P}$ to be transformed into deep features as a prior and the final output E_4 is directly fed into $\mathcal{M}_{\mathcal{D}}$ to help generate a sharp face image.

3.2. Prior Estimation Stream

This stream $\mathcal{P}$ aims to estimate deep feature prior through the supervision of GT deep features $\mathbb{P} = \{P_i | i = 1, 2, 3\}$. It takes $\mathbb{E}_P$ as input and returns a set of estimated deep features $\hat{\mathbb{P}} = \{\hat{P}_i | i = 1, 2, 3\}$. More concretely, $\mathcal{P}$ is divided into three sub-networks $\mathcal{P} = \{\mathcal{P}_i | i = 1, 2, 3\}$. Each sub-network $\mathcal{P}_i$ produce each element of output $\hat{P}_i$ from corresponding element of input E_i as $\hat{P}_i = \mathcal{P}_i(E_i)$ (Fig. 2). The $\mathcal{P}_i$ consist of the self-spatial feature transform



(a) Self Spatial Feature Transform (SSFT) module

(b) Spatial Feature Transform (SFT) module

Figure 3. The structure of two spatial feature transform modules: (a) SSFT and (b) SFT module

(SSFT) module and stacked Resblocks used in [47] in series. The purpose of the SSFT module is to transform the input feature E_i extracted from $\mathcal{M}$ into a feature for $\mathcal{P}$ to better estimate the deep feature prior. For this, we modify the spatial feature transform (SFT) module [38], which applies pixel-wise affine transformation on input features using additional prior. Unlike [38], this module can transform an input feature without extra data. As shown in Fig. 3a, the SSFT module learns a mapping function $\mathcal{F}_i^s$ based on its own input feature E_i to generate a pair of affine transformation parameters (α_i^s, β_i^s) as $(\alpha_i^s, \beta_i^s) = \mathcal{F}_i^s(E_i)$. Here α_i^s and β_i^s are the scale and shift parameters which have the same size as E_i, respectively. According to these internal parameters, we can adaptively modulate E_i to E_i^t as

$$E_i^t = SSFT_i(E_i|\alpha_i^s, \beta_i^s) = \alpha_i^s \otimes E_i + \beta_i^s, \qquad (1)$$

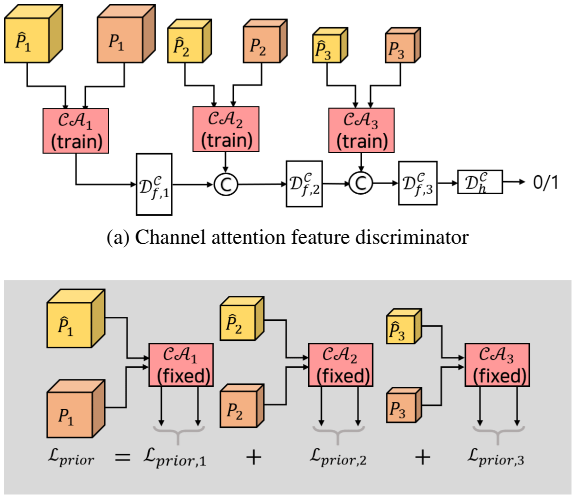
where $\otimes$ indicates element-wise multiplication. Thus, the output E_i^t is transformed to have an appropriate feature distribution to predict GT deep feature prior. Subsequently, the following Resblocks [47] are used to predict $\hat{P}_i$ using E_i^t.

3.3. Decoder of Deblurring Stream

Our key idea is that $\mathcal{M}_{\mathcal{D}}$ can restore a detailed face image with the help of the estimated feature prior, which contains a rich representation of sharp face images. It considers not only E_0 and E_4 from $\mathcal{M}_{\mathcal{E}}$ but also $\hat{\mathbb{P}}$ from $\mathcal{P}$ as facial priors and generates I_{deblur} as $I_{deblur} = \mathcal{M}_{\mathcal{D}}(E_0, E_4, \hat{\mathbb{P}})$. We construct $\mathcal{M}_{\mathcal{D}}$ using the SFT modules [38] to infuse prior features $\hat{\mathbb{P}}$ into deblurring features $\mathbb{D} = \{D_i | i = 1, 2, 3\}$ of $\mathcal{M}_{\mathcal{D}}$. The SFT module [38] is adopted for learning the transformation parameters to effectively incorporate prior conditions through affine transformation. Each SFT module progressively incorporates each prior feature $\hat{P}_i$ into the same size of deblurring feature D_i for better restoration. As shown in Fig. 3b, SFT module learns a mapping function $\mathcal{F}_i$ that produces a pair of transformation parameters (α_i, β_i) based on prior information $\hat{P}_i$ as $(\alpha_i, \beta_i) = \mathcal{F}_i(\hat{P}_i)$. These parameters adaptively transform D_i through a spatial-wise affine transformation as follows:

$$D_i^t = SFT_i(D_i|\alpha_i, \beta_i) = \alpha_i \otimes D_i + \beta_i, \qquad (2)$$

where α_i and β_i are the scale and shift parameters, respectively. They have the same spatial size as D_i.



(a) Channel attention feature discriminator

(b) Prior loss

Figure 4. (a) The structure of the proposed channel-attention feature discriminator. (b) Prior loss.

3.4. Deep Feature Guided GAN

According to the previous subsections, we need to estimate our prior features $\hat{\mathbb{P}}$ to be close to the GT deep features $\mathbb{P}$. Meanwhile, Zhang *et al.* [46] recently discovered that the channel-wise importance of deep features is not equal in perceptual judgment. Inspired by this, we speculate that there are more important channels among the deep features that act as guidance to restore the sharp image. However, it is not trivial to explicitly learn the important channels in the deblurring process due to the lack of direct supervision information. To resolve this problem, we notice that channels which facilitate to distinguish the sharp and blurred images are more important for restoration task in general. Based on this, we propose to implicitly learn important channels for deblurring via a novel *channel-attention feature discriminator* $\mathcal{D}^{\mathcal{C}}$. In contrast to conventional feature discriminators [26], we add the channel attention (CA) module [9] at each input layer, as depicted in Fig 4a. This allows to define the prior loss that encourage our generator to focus on training more important channels among $\mathbb{P}$ for deblurring and down-weight others.

Following the GAN-based framework [7], we train our generator $\mathcal{G}$ and discriminator $\mathcal{D}^{\mathcal{C}}$ jointly in a min-max function $V(\mathcal{G}, \mathcal{D}^{\mathcal{C}})$:

$$\min_{\mathcal{G}} \max_{\mathcal{D}^{\mathcal{C}}} V(\mathcal{G}, \mathcal{D}^{\mathcal{C}}) = \\ \mathbb{E}_{I_{GT}}[\log \mathcal{D}^{\mathcal{C}}(I_{GT})] + \mathbb{E}_{I_{blur}}[\log(1 - \mathcal{D}^{\mathcal{C}}(\mathcal{G}(I_{blur})))]. \tag{3}$$

Thus, it can be divided into two steps, *i.e.*, $\mathcal{D}^{\mathcal{C}}$ training step and $\mathcal{G}$ training step. We explain each step in detail below.

Discriminator Training. As shown in Fig. 4a, this step trains $\mathcal{D}^{\mathcal{C}}$ to alternately take a set of predicted deep features $\hat{\mathbb{P}}$ from $\mathcal{P}$ or a set of GT deep features $\mathbb{P}$ as input. It then tries to classify them into real (*i.e.*, sharp) or fake (*i.e.*, blurry) features. $\mathcal{D}^{\mathcal{C}}$ consists of three parts; a set of CA modules $\mathbb{CA} = \{\mathcal{CA}_i | i = 1, 2, 3\}$, a set of internal feature processing blocks $\mathbb{D}_f^{\mathbb{C}} = \{\mathcal{D}_{f,i}^{\mathcal{C}} | i = 1, 2, 3\}$, and the last classifier block $\mathcal{D}_h^{\mathcal{C}}$ (Fig. 4a). The key part is $\mathbb{CA}$ which receives the input prior features. They are trained to emphasize more helpful channels for classifying the input features. Thus the following $\mathbb{D}_f^{\mathbb{C}}$ can learn better representations from these weighted features instead of raw input features. Finally, $\mathcal{D}_h^{\mathcal{C}}$ classifies input features more easily based on this representation.

More formally, given $\mathbb{P}$ (or $\hat{\mathbb{P}}$) as input, each element P_i (or $\hat{P}_i$) is weighted by $\mathcal{CA}_i$ as

$$P_i{}' = \mathcal{CA}_i(P_i), \hat{P}_i{}' = \mathcal{CA}_i(\hat{P}_i). \tag{4}$$

These weighted features are classified as real or fake by

$$\mathcal{D}^{\mathcal{C}}(\mathbb{P}) = \mathcal{D}_h^{\mathcal{C}}(\mathcal{D}_{f,3}^{\mathcal{C}}(P'_3 \oplus \mathcal{D}_{f,2}^{\mathcal{C}}(P'_2 \oplus \mathcal{D}_{f,1}^{\mathcal{C}}(P'_1)))), \tag{5}$$

where $\oplus$ denotes the channel-wise concatenation. $\mathcal{D}^{\mathcal{C}}(\hat{\mathbb{P}})$ is defined similarly to $\mathcal{D}^{\mathcal{C}}(\mathbb{P})$ as Eq. (5).

The objective function for training $\mathcal{D}^{\mathcal{C}}$ is defined as follows:

$$\mathcal{L}_{\mathcal{D}^{\mathcal{C}}} = -\mathbb{E}[\log(\mathcal{D}^{\mathcal{C}}(\mathbb{P}))] - \mathbb{E}[\log(1 - \mathcal{D}^{\mathcal{C}}(\hat{\mathbb{P}}))]. \tag{6}$$

Generator Training. To train the generator, our objective function combines pixel loss $\mathcal{L}_{pixel}$, adversarial loss $\mathcal{L}_{G,adv}$ and prior loss $\mathcal{L}_{prior}$.

For pixel loss, we adopt pixel-wise L_1 distance to minimize the distance between the ground truth image I_{GT} and the deblurred image I_{deblur} as

$$\mathcal{L}_{pixel} = \|I_{deblur} - I_{GT}\|_1. \tag{7}$$

To encourage the generator to produce a more realistic deep feature prior, the adversarial loss for the generator is defined by

$$\mathcal{L}_{G,adv} = -\mathbb{E}[\log(\mathcal{D}^{\mathcal{C}}(\hat{\mathbb{P}}))]. \tag{8}$$

In particular, we propose a novel prior loss to transfer more useful knowledge of GT deep features for deblurring and suppressing the others. The main ingredient of our prior loss is weight of the important channels from the CA modules that are learned during the discriminator training step. As shown in Fig. 4b, the CA modules are fixed in generator training step and used to distill more useful knowledge by weighting both the GT deep feature P_i and predicted deep feature $\hat{P}_i$ as Eq. (4). Therefore, we can define our prior loss $\mathcal{L}_{prior}$ using L_2 distance as follows:

$$\mathcal{L}_{prior} = \sum_{i=1}^{3} \mathcal{L}_{prior,i} = \sum_{i=1}^{3} \sum_{w=1}^{W} \sum_{h=1}^{H} \left\| P'_{hw,i} - \hat{P}'_{hw,i} \right\|_2^2, \tag{9}$$

where $P'_{hw,i}$ and $\hat{P}'_{hw,i}$ are the normalized features of $P_i{}'$ and $\hat{P}'_i$ along the channel-axis for each spatial position.

As a result, the total loss for the training generator in our approach is defined by

$$\mathcal{L}_G = \mathcal{L}_{pixel} + \lambda_{G,adv}\mathcal{L}_{G,adv} + \lambda_{prior}\mathcal{L}_{prior}, \tag{10}$$

where $\lambda_{G,adv}$ and λ_{prior} are hyperparameters. They are empirically set as $\lambda_{G,adv} = 0.05$ and $\lambda_{prior} = 1$.

4. Experimental Results

4.1. Implementation Details

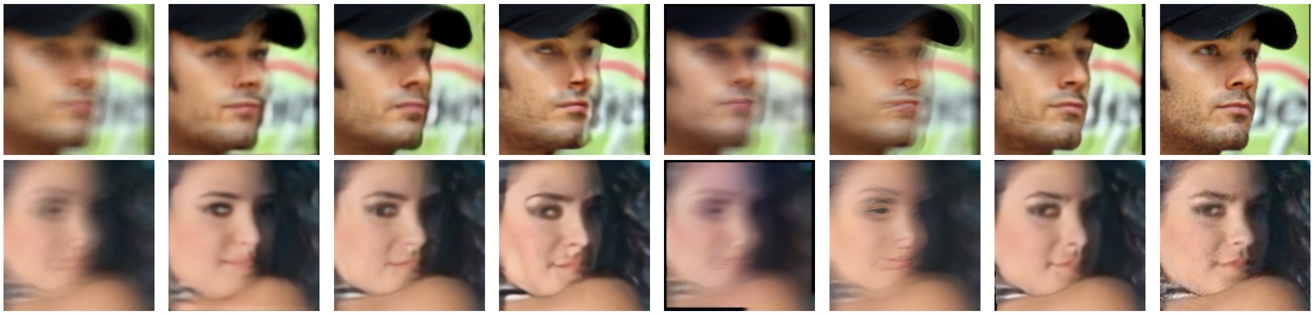
As DFPGnet consists of the generator and discriminator, We alternately trained them using the Adam optimizer [13] with β_1 = 0.9, β_2 = 0.999. The learning rates of the generator and discriminator were initialized as 1×10^{-4} and 1×10^{-5} respectively and decayed exponentially by a factor of 0.99 for every epoch. For ground truth prior features $\mathbb{P}$, we employed the VGGFace [27] model, which is trained on the VGG Face dataset [27] for the face recognition. Specifically, we selected the `relu1_2`, `relu2_2`, and `relu3_3` layers of VGGFace for P_1, P_2, and P_3, respectively. We trained the DFPGnet using a single NVIDIA TITAN-RTX GPU. Our method is implemented using Pytorch [28].

4.2. Datasets

We evaluated our method on two facial deblurring datasets: Shen test set [32] and MSPL test set [18]. Since the training sets for the Shen and MSPL test set are synthesized with different blur kernels and images, the evaluation of each test set was performed on the model trained on the corresponding training set for fair comparison, independently. The DFPGnet trained on Shen training set (termed as **DFPG-A**) was tested on the Shen test set which consists of 16,000 blurry face images synthesized using Helen [15] and CelebA [22]. And the DFPGnet trained on MSPL traning set (termed as **DFPG-B**) was evaluated on the MSPL test set which provides two subsets, MSPL-Center and MSPL-Random, each consisting of centered and randomly transformed face images.

	Helen				CelebA			
Method	PSNR(↑)	SSIM(↑)	d_{VGG}(↓)	LPIPS(↓)	PSNR(↑)	SSIM(↑)	d_{VGG}(↓)	LPIPS(↓)
Shen *et al.* [32]	25.58	0.861	91.06	0.1527	24.34	0.860	117.50	0.1832
Lu *et al.* [23]	20.25	0.705	241.93	0.1654	19.96	0.742	305.96	0.1688
Xia *et al.* [41]	26.13	0.886	55.97	0.1052	25.18	0.892	68.05	0.1199
Yasarla *et al.* [42]	**27.75**	<u>0.897</u>	86.87	0.1086	**26.62**	<u>0.908</u>	66.33	0.1401
Shen *et al.* [33]	25.91	0.869	–	–	24.89	0.875	–	–
Lee *et al.* [18]	25.91	0.881	<u>47.80</u>	**0.0828**	24.91	0.885	<u>57.54</u>	**0.0962**
Li *et al.* [20]	21.64	0.754	253.45	0.2275	20.96	0.761	327.00	0.2517
Wang *et al.* [37]	22.30	0.775	206.57	0.1592	21.62	0.792	261.50	0.1503
DFPG-A (ours)	<u>27.70</u>	**0.911**	**42.84**	<u>0.0928</u>	<u>26.56</u>	**0.915**	**53.38**	<u>0.1052</u>

Table 1. **Quantitative comparisons on Shen test set [32].** The best and second best results are **highlighted** and <u>underlined</u>, respectively.



Input Shen *et al.* [32] Yasarla *et al.* [42] Lee *et al.* [18] Li *et al.* [20] Wang *et al.* [37] DFPG-A (ours) Ground Truth

Figure 5. **Qualitative comparisons on Shen test set [32].**

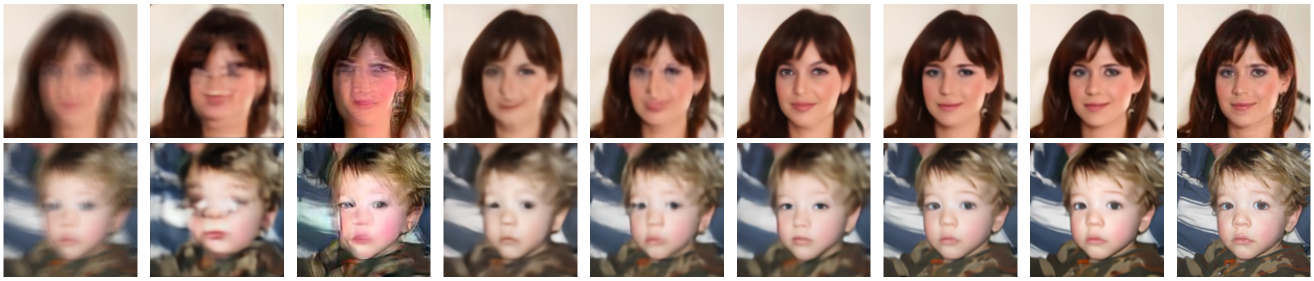
	MSPL-Center											
	CelebA				CelebA-HQ				FFHQ			
Method	PSNR	SSIM	d_{VGG}	LPIPS	PSNR	SSIM	d_{VGG}	LPIPS	PSNR	SSIM	d_{VGG}	LPIPS
Shen *et al.* [32]	19.75	0.740	113.66	0.301	19.95	0.755	267.41	0.287	19.57	0.723	220.87	0.342
Lu *et al.* [23]	17.93	0.617	123.35	0.228	18.63	0.649	243.06	0.190	18.26	0.630	177.00	0.226
Zhang *et al.* [44]	20.40	0.744	117.68	0.314	20.90	0.764	239.04	0.295	20.64	0.743	170.41	0.343
*Zhang *et al.*	23.98	0.824	45.13	0.241	24.84	0.844	83.36	0.212	23.52	0.813	71.51	0.287
Xia *et al.* [41]	25.03	0.873	39.58	0.179	25.79	0.886	83.46	0.161	24.66	0.859	57.66	0.208
Yasarla *et al.* [42]	22.73	0.817	55.01	0.213	23.02	0.827	102.97	0.196	22.19	0.795	86.43	0.251
*Yasarla *et al.*	24.71	0.857	37.80	0.183	26.11	0.882	50.67	0.148	24.31	0.843	58.46	0.218
Lee *et al.* [18]	<u>28.07</u>	<u>0.921</u>	<u>18.19</u>	<u>0.115</u>	<u>28.82</u>	<u>0.929</u>	<u>40.93</u>	<u>0.097</u>	<u>27.36</u>	<u>0.908</u>	<u>25.39</u>	<u>0.133</u>
DFPG-B (ours)	**29.06**	**0.933**	**14.76**	**0.102**	**29.86**	**0.940**	**20.95**	**0.085**	**28.76**	**0.921**	**20.28**	**0.118**

	MSPL-Random											
	CelebA				CelebA-HQ				FFHQ			
Method	PSNR	SSIM	d_{VGG}	LPIPS	PSNR	SSIM	d_{VGG}	LPIPS	PSNR	SSIM	d_{VGG}	LPIPS
Shen *et al.* [32]	18.89	0.711	90.37	0.331	19.18	0.729	157.49	0.319	19.03	0.713	127.71	0.336
Lu *et al.* [23]	17.41	0.631	46.05	0.269	18.04	0.664	72.56	0.230	17.94	0.654	65.06	0.259
Zhang *et al.* [44]	19.36	0.702	86.77	0.328	19.85	0.726	144.74	0.311	19.77	0.715	122.07	0.333
*Zhang *et al.*	23.35	0.794	30.46	0.254	24.09	0.817	54.06	0.227	23.54	0.804	46.03	0.255
Xia *et al.* [41]	23.66	0.849	30.94	0.204	24.48	0.861	60.95	0.194	23.95	0.855	44.62	0.202
Yasarla *et al.* [42]	21.24	0.777	45.05	0.245	21.46	0.789	72.56	0.230	21.28	0.778	65.06	0.241
*Yasarla *et al.*	22.92	0.789	33.83	0.234	23.56	0.812	48.17	0.214	23.16	0.793	49.94	0.227
Lee *et al.* [18]	<u>28.95</u>	<u>0.936</u>	<u>11.41</u>	<u>0.109</u>	<u>29.80</u>	<u>0.945</u>	<u>26.91</u>	<u>0.094</u>	<u>29.22</u>	<u>0.941</u>	<u>15.44</u>	<u>0.099</u>
DFPG-B (ours)	**29.96**	**0.945**	**8.37**	**0.100**	**30.76**	**0.953**	**23.05**	**0.084**	**30.29**	**0.951**	**10.93**	**0.089**

Table 2. **Quantitative comparisons on MSPL test set [18] .** The best and second best results are **highlighted** and <u>underlined</u>, respectively.

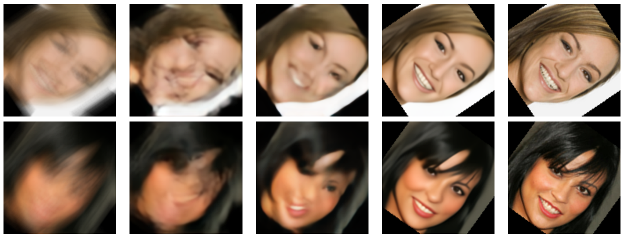
4.3. Comparisons with Existing Methods

We compare our DFPGnet with existing face deblurring methods [32, 23, 41, 42, 18, 33], blind face restoration methods [20, 37] and generic deblurring method [43]. For evaluation, we report two metrics, PSNR and SSIM [40], which are widely measured in image restoration studies. We also used the feature distance d_{VGG} of the pre-trained VGGFace [27] to measure the similarity of the facial identity between the deblurred and the ground truth images. In addition, we compared the LPIPS [46] distance, which measures perceptual image patch similarity. We also compared the performance of face detection and verification to determine how well the restored face image can be used for face-related applications.



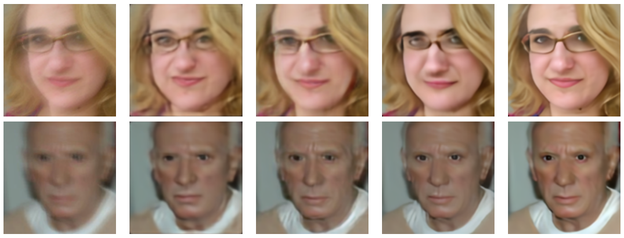
Input | Shen *et al.* | Lu *et al.* | *Zhang *et al.* | Xia *et al.* | *Yasarla *et al.* | Lee *et al.* | DFPG-B (ours) | Ground Truth

Figure 6. **Qualitative comparisons on MSPL-Center test set [18].**



Input | *[32] | *[42] | ours | GT

Figure 7. **Visual comparisons on MSPL-Random test set [18].**



Input | [32] | [42] | [18] | Ours

Figure 8. **Qualitative comparisons on Real-Blur test set [14].**

Method	Detection (%) (↑)	Acc (%) (↑)
GT of Shen test set [32]	96.00	93.47
Blurred images	77.40	77.05
Shen *et al.* [32]	94.80	87.03
Lu *et al.* [23]	89.03	80.56
Xia *et al.* [41]	95.95	89.12
Yasarla *et al.* [42]	94.49	87.84
Lee *et al.* [18]	**96.55**	89.59
DFPG-A (ours)	96.41	**89.87**

Table 3. **Face detection and verification comparisons on the CelebA from the Shen test set [32].** The best and second best results are **highlighted** and underlined, respectively.

Model	Per	DFP	FT	Dis	CA-Dis & Pr	PSNR	SSIM
B1	×	×	×	×	×	27.95	0.919
B2	✓	×	×	×	×	27.92	0.917
B3	×	✓	×	×	×	27.91	0.917
B4	×	✓	✓	×	×	28.67	0.927
B5	×	✓	✓	✓	×	28.87	0.929
B6	×	✓	✓	×	✓	**29.27**	**0.931**

Table 4. **Ablation Studies on MSPL-Center [18].** The best results are **highlighted**. The performance is the average result of all subsets of MSPL-Center. Our final DFPGnet is denoted as B6.

Comparisons Using Shen test set [32]. Table 1 shows the official results of the existing face deblurring methods reported by the authors for the Shen test set [32]. Note that since [33] only reports the PSNR and SSIM without any official code and visual results, we do not compare d_{VGG} and LPIPS results. We also compare the blind face restoration methods in Table 1. Since they only provide their test model, we did not retrain them with same degradation model. Our method is the best for SSIM and d_{VGG}, and the second best for PSNR and LPIPS. This indicates that our restored faces are the most structurally and perceptually to the ground truth faces. Fig. 5 shows the qualitative comparisons with existing methods. The 1^{st} row in Fig. 5 shows that the DFPGnet restores sharp edges and fine textures on notable face regions like eyes, lips and facial hair. When comparing the results of the 2^{nd} row in Fig. 5, our method can restore the most similar and sharp faces compared to the ground truth faces.

Comparisons Using MSPL test set [18]. The quantitative comparisons are shown in Table 2. Except for Lee *et al.* [18], the other methods are trained with different data from MSPL training set. For fair comparison, we retrained the methods [43, 42] that provide official training codes using MSPL training set [18] (termed as *). For the other methods that only provide the test model, we simply test with the provided model on MSPL test set. The results show that our method significantly surpasses the existing methods. Although retrained *Zhang *et al.* [43] has shown significant improvements, their performance is still worse than *Yasarla *et al.* [42], Lee *et al.* [18] and our DFPG-B. This is because the method of *Zhang *et al.* [43] is proposed for generic image deblurring, and they do not leverage any facial prior information. This demonstrates that the facial prior information plays an important role for face deblurring. Our method outperforms *Yasarla *et al.* [42] and Lee *et al.* [18], both of which are based on semantic priors and perceptual loss [11] using VGGFace [27]. In particular, it is remarkable that our method can restore faces more accurately in pixel values (PSNR), while being perceptually the best (SSIM, d_{VGG}, and LPIPS). These results can be attributed to our method of estimating the deep feature prior that is helpful for deblurring among the rich representation of VGGFace. A comparison of the visual results in Fig. 6 shows our method is best at restoring fine-grained facial texture components such as facial wrinkles, teeth, and hair.

The merit of our deep feature prior is that it is more robust to non-aligned face than geometric prior [32, 42]. Since we estimate the features of the early layers of VGGFace [27] as the prior, it can provide local low-level facial cues, which are not highly dependent to facial alignment. In contrast, existing methods [32, 42] that utilize the final outputs of segmentation model as prior are sensitive to non-aligned face because it contain global geometric facial cues. As shown in Fig. 7, DFPGnet can still restore more faithful face for randomly rotated face than other methods.

Face Detection and Verification. One of the major goals of face deblurring is to increase the accuracy of high-level tasks when the input image is blurry. For this reason, we compared performance of face detection and verification using deblurred images on the CelebA test set of Shen test set [32]. For the detection test, we measured the success rate of face detection using the OpenFace toolbox [1] similar to [32, 42]. As listed in Table 3, the success rates of face detection for GT images, blurry images, and deblurred images using our model are 96.00 %, 77.40 % and 96.10 %, respectively. To compare the verification performance, we measured the estimated mean accuracy (Acc) [10]. We employed MobileNet [8] trained with Arcface [6] loss. As demonstrated in Table 3, our model achieved the best performance in Acc. These results prove that our method is best suited for high-level tasks, such as detection and verification compared to other methods.

Visual comparison on real blurred images Since the existing face deblurring methods [32, 23, 41, 42, 18] are trained and evaluated with synthetically blurred datasets [32, 18], it is unclear how these methods would perform on real-world blurry images acquired in the wild [14]. To this end, we conduct the visual comparisons using facial images collected from real-blur dataset [14] where the ground-truth images do not exist. The comparison results are shown in Fig. 8. DFPGnet can produce more fine textures (*i.e.*, hair and wrinkles) and details of small facial components (*i.e.*, eyes, lips, and teeth) compared to other methods. This result shows that our DFPGnet can generalize well to real blurred images.

4.4. Ablation Study

To investigate the effects of our method, we gradually applied each component in the our method to the baseline model and compared the differences. The entire quantitative comparisons are presented in Table 4. Each row of the Table 4 represents a model trained with the configurations marked in the “Model”.

Effect of Deep Feature Prior. We set a baseline model (B1) by removing the prior estimation stream and all other modules except for $\mathcal{M}$ from the entire network. We trained it using only $\mathcal{L}_{pixel}$ (Eq. (7)). The model B2 is the same architecture with B1 but trained with the weighted sum of $\mathcal{L}_{pixel}$ (Eq. (7)) and the perceptual loss function [11] using VGGFace [27] (marked as “Per” in Table 4) similar to [17]. We then define model B3, which simply adds $\mathcal{P}$ (marked as “DFP” to B1 in Table 4) without feature transformation modules. In Table 4, the results show that the average PSNR and SSIM values of B3 are lower than those of B1 and B2. From this, we find that simply guiding the feature from the VGGFace without feature transform rather interferes with the restoration task owing to their different tasks.

Effect of Feature Transform Modules. To investigate the impacts of them, we added the feature transform modules (marked as ”FT” in Table 4) into the B3 model, as shown in the 4^{th} (B4) row in Table 4. The results of B4 show that the average PSNR and SSIM increase by 0.74 and 0.1 compared to those of B3, respectively. These results show that our feature transformation modules effectively reduce the difference of feature distributions between two different tasks of deblurring and recognition.

Effect of Channel-Attention Feature Discriminator and Prior loss. To study the effects of our channel-attention feature discriminator $\mathcal{D}^{\mathcal{C}}$ and prior loss, we compared the performances of models 1) without discriminator (B4), 2) with the feature discriminator without channel attention that is marked as “Dis” (B5), and 3) our channel-attention feature discriminator and prior loss marked as “CA-Dis & Pr” (B6). As shown in Table 4, the B6 model outperforms than the other methods. The average PSNR of the B6 model increases by 0.36, compared to the B4. This proves that it is effective to weight some important channels among features from VGGFace [27] higher, rather than to weight all channels equally.

5. Conclusions

We propose a deep feature prior guided face deblurring network (DFPGnet) which estimates deep features of the face recognition network as prior that includes rich information on sharp faces. Thanks to the feature transform modules and proposed channel attention mechanism, we can effectively utilize our prior in restoration task. In future work, we believe that our method can be generalized well to diverse restoration tasks using other pre-trained networks.

Acknowledgement. This work was supported in part by the National Research Foundation of Korea (NRF) grant funded by the Korea government (MSIT) (No. 2019R1C1C1007446), and in part by the BK21 FOUR program of the National Research Foundation of Korea funded by the Ministry of Education (NRF5199991014091).

References

[1] Brandon Amos, Bartosz Ludwiczuk, and Mahadev Satyanarayanan. *OpenFace: A general-purpose face recognition library with mobile applications*. Technical report, CMU-CS-16-118, CMU School of Computer Science, 2016.

[2] Adrian Bulat and Georgios Tzimiropoulos. *How far are we from solving the 2d & 3d face alignment problem?(and a dataset of 230,000 3d facial landmarks)*. In *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, pages 1021–1030, 2017.

[3] Yu Chen, Ying Tai, Xiaoming Liu, Chunhua Shen, and Jian Yang. *Fsrnet: End-to-end learning face super-resolution with facial priors*. In *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, pages 2492–2501, 2018.

[4] Grigorios G Chrysos, Paolo Favaro, and Stefanos Zafeiriou. *Motion deblurring of faces*. *Int. J. Comput. Vis.*, 127(6-7):801–823, Mar. 2019.

[5] Grigorios G Chrysos and Stefanos Zafeiriou. *Deep face deblurring*. In *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, pages 69–78, July 2017.

[6] Jiankang Deng, Jia Guo, Niannan Xue, and Stefanos Zafeiriou. *Arcface: Additive angular margin loss for deep face recognition*. In *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, pages 4690–4699, 2019.

[7] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. *Generative adversarial nets*. In *Adv. Neural Inf. Process. Syst. (NIPS)*, pages 2672–2680, June 2014.

[8] Andrew G Howard, Menglong Zhu, Bo Chen, Dmitry Kalenichenko, Weijun Wang, Tobias Weyand, Marco Andreetto, and Hartwig Adam. *Mobilenets: Efficient convolutional neural networks for mobile vision applications*. *arXiv preprint arXiv:1704.04861*, 2017.

[9] Jie Hu, Li Shen, and Gang Sun. *Squeeze-and-excitation networks*. In *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, pages 7132–7141, June 2018.

[10] Gary B. Huang, Manu Ramesh, Tamara Berg, and Erik Learned-Miller. Labeled faces in the wild: A database for studying face recognition in unconstrained environments. Technical Report 07-49, University of Massachusetts, Amherst, October 2007.

[11] Justin Johnson, Alexandre Alahi, and Li Fei-Fei. *Perceptual losses for real-time style transfer and super-resolution*. In *Proc. Eur. Conf. Comput. Vis. (ECCV)*, pages 694–711. Springer, Oct. 2016.

[12] Amin Jourabloo, Mao Ye, Xiaoming Liu, and Liu Ren. *Pose-invariant face alignment with a single cnn*. In *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, pages 3200–3209, 2017.

[13] Diederik P Kingma and Jimmy Ba. *Adam: A method for stochastic optimization*. *arXiv preprint arXiv:1412.6980*, Dec. 2014.

[14] Wei-Sheng Lai, Jia-Bin Huang, Zhe Hu, Narendra Ahuja, and Ming-Hsuan Yang. A comparative study for single image blind deblurring. In *IEEE Conferene on Computer Vision and Pattern Recognition*, 2016.

[15] Vuong Le, Jonathan Brandt, Zhe Lin, Lubomir Bourdev, and Thomas S Huang. Interactive facial feature localization. In *European conference on computer vision*, pages 679–692. Springer, 2012.

[16] Yann LeCun, Yoshua Bengio, and Geoffrey Hinton. *Deep learning*. *Nature*, 521(7553):436–444, May 2015.

[17] Christian Ledig, Lucas Theis, Ferenc Huszár, Jose Caballero, Andrew Cunningham, Alejandro Acosta, Andrew Aitken, Alykhan Tejani, Johannes Totz, Zehan Wang, et al. Photo-realistic single image super-resolution using a generative adversarial network. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4681–4690, 2017.

[18] Tae Bok Lee, Soo Hyun Jung, and Yong Seok Heo. Progressive semantic face deblurring. *IEEE Access*, 8:223548–223561, 2020.

[19] Jian Li, Yabiao Wang, Changan Wang, Ying Tai, Jianjun Qian, Jian Yang, Chengjie Wang, Jilin Li, and Feiyue Huang. Dsfd: dual shot face detector. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5060–5069, 2019.

[20] Xiaoming Li, Chaofeng Chen, Shangchen Zhou, Xianhui Lin, Wangmeng Zuo, and Lei Zhang. Blind face restoration via deep multi-scale component dictionaries. In *European Conference on Computer Vision*, pages 399–415. Springer, 2020.

[21] Songnan Lin, Jiawei Zhang, Jinshan Pan, Yicun Liu, Yongtian Wang, Jing Chen, and Jimmy Ren. *Learning to Deblur Face Images via Sketch Synthesis*. *Proc. AAAI Conf. Artif. Intell.*, 34:11523–11530, Apr. 2020.

[22] Ziwei Liu, Ping Luo, Xiaogang Wang, and Xiaoou Tang. *Deep Learning Face Attributes in the Wild*. In *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Dec. 2015.

[23] Boyu Lu, Jun-Cheng Chen, and Rama Chellappa. *Unsupervised domain-specific deblurring via disentangled representations*. In *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, pages 10225–10234, June. 2019.

[24] Aravindh Mahendran and Andrea Vedaldi. *Understanding deep image representations by inverting them*. In *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, pages 5188–5196, 2015.

[25] Seungjun Nah, Tae Hyun Kim, and Kyoung Mu Lee. Deep multi-scale convolutional neural network for dynamic scene deblurring. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, July 2017.

[26] Seong-Jin Park, Hyeongseok Son, Sunghyun Cho, Ki-Sang Hong, and Seungyong Lee. Srfeat: Single image super-resolution with feature discrimination. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 439–455, 2018.

[27] Omkar M Parkhi, Andrea Vedaldi, and Andrew Zisserman. *Deep face recognition*. 2015.

[28] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. Pytorch: An imperative style, high-performance deep learning library. In *Advances in neural information processing systems*, pages 8026–8037, 2019.

[29] Mohammad Saeed Rad, Behzad Bozorgtabar, Urs-Viktor Marti, Max Basler, Hazim Kemal Ekenel, and Jean-Philippe

Thiran. Srobb: Targeted perceptual loss for single image super-resolution. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2710–2719, 2019.

[30] Wenqi Ren, Jiaolong Yang, Senyou Deng, David Wipf, Xiaochun Cao, and Xin Tong. *Face Video Deblurring using 3D Facial Priors*. In *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, pages 9388–9397, Oct. 2019.

[31] Mehdi SM Sajjadi, Bernhard Scholkopf, and Michael Hirsch. Enhancenet: Single image super-resolution through automated texture synthesis. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 4491–4500, 2017.

[32] Ziyi Shen, Wei-Sheng Lai, Tingfa Xu, Jan Kautz, and Ming-Hsuan Yang. *Deep semantic face deblurring*. In *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, pages 8260–8269, June 2018.

[33] Ziyi Shen, Wei-Sheng Lai, Tingfa Xu, Jan Kautz, and Ming-Hsuan Yang. Exploiting semantics for face image deblurring. *International Journal of Computer Vision*, 128(7):1829–1846, 2020.

[34] Yi Sun, Xiaogang Wang, and Xiaoou Tang. Deep convolutional network cascade for facial point detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3476–3483, 2013.

[35] Xin Tao, Hongyun Gao, Xiaoyong Shen, Jue Wang, and Jiaya Jia. Scale-recurrent network for deep image deblurring. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 8174–8182, 2018.

[36] Hao Wang, Yitong Wang, Zheng Zhou, Xing Ji, Dihong Gong, Jingchao Zhou, Zhifeng Li, and Wei Liu. *Cosface: Large margin cosine loss for deep face recognition*. In *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, pages 5265–5274, June 2018.

[37] Xintao Wang, Yu Li, Honglun Zhang, and Ying Shan. Towards real-world blind face restoration with generative facial prior. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9168–9178, 2021.

[38] Xintao Wang, Ke Yu, Chao Dong, and Chen Change Loy. Recovering realistic texture in image super-resolution by deep spatial feature transform. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 606–615, 2018.

[39] Xintao Wang, Ke Yu, Shixiang Wu, Jinjin Gu, Yihao Liu, Chao Dong, Yu Qiao, and Chen Change Loy. Esrgan: Enhanced super-resolution generative adversarial networks. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 0–0, 2018.

[40] Zhou Wang, Alan C Bovik, Hamid R Sheikh, and Eero P Simoncelli. *Image quality assessment: from error visibility to structural similarity*. *IEEE Trans. Image Process.*, 13(4):600–612, Apr. 2004.

[41] Zhihao Xia and Ayan Chakrabarti. *Training Image Estimators without Image Ground Truth*. In *Adv. Neural Inf. Process. Syst. (NIPS)*, pages 2436–2446, June. 2019.

[42] Rajeev Yasarla, Federico Perazzi, and Vishal M Patel. *Deblurring face images using uncertainty guided multi-stream semantic networks*. *IEEE Trans. Image Process.*, Apr. 2020.

[43] Hongguang Zhang, Yuchao Dai, Hongdong Li, and Piotr Koniusz. Deep stacked hierarchical multi-patch network for image deblurring. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019.

[44] Hongguang Zhang, Yuchao Dai, Hongdong Li, and Piotr Koniusz. Deep stacked hierarchical multi-patch network for image deblurring. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5978–5986, 2019.

[45] Kaipeng Zhang, Zhanpeng Zhang, Zhifeng Li, and Yu Qiao. Joint face detection and alignment using multitask cascaded convolutional networks. *IEEE Signal Processing Letters*, 23(10):1499–1503, 2016.

[46] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. *The unreasonable effectiveness of deep features as a perceptual metric*. In *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, pages 586–595, 2018.

[47] Yulun Zhang, Kunpeng Li, Kai Li, Lichen Wang, Bineng Zhong, and Yun Fu. Image super-resolution using very deep residual channel attention networks. In *Proceedings of the European conference on computer vision (ECCV)*, pages 286–301, 2018.

[48] Xiangyu Zhu, Zhen Lei, Xiaoming Liu, Hailin Shi, and Stan Z Li. *Face alignment across large poses: A 3d solution*. In *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, pages 146–155, 2016.