Enhanced Blind Face Restoration with Multi-Exemplar Images and Adaptive Spatial Feature Fusion

Xiaoming Li[1], Wenyu Li[1], Dongwei Ren[2], Hongzhi Zhang[1], Meng Wang[3], Wangmeng Zuo[1](✉)
[1]School of Computer Science and Technology, Harbin Institute of Technology, China
[2]College of Intelligence and Computing, Tianjin University, China
[3]Hefei University of Technology, China
{csxmli, wmzuo}@hit.edu.cn

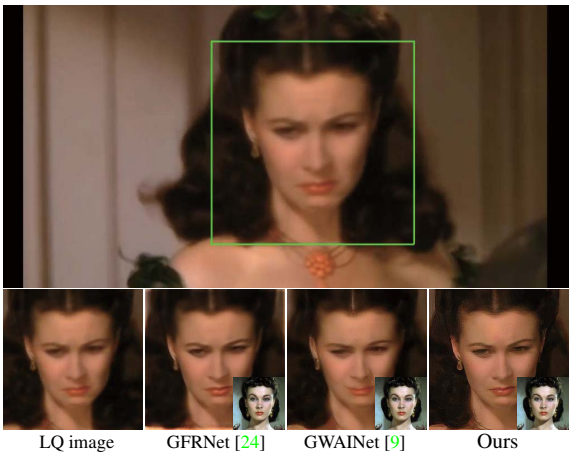
Abstract

In many real-world face restoration applications, e.g., *smartphone photo albums and old films, multiple high-quality (HQ) images of the same person usually are available for a given degraded low-quality (LQ) observation. However, most existing guided face restoration methods are based on single HQ exemplar image, and are limited in properly exploiting guidance for improving the generalization ability to unknown degradation process. To address these issues, this paper suggests to enhance blind face restoration performance by utilizing multi-exemplar images and adaptive fusion of features from guidance and degraded images. First, given a degraded observation, we select the optimal guidance based on the weighted affine distance on landmark sets, where the landmark weights are learned to make the guidance image optimized to HQ image reconstruction. Second, moving least-square and adaptive instance normalization are leveraged for spatial alignment and illumination translation of guidance image in the feature space. Finally, for better feature fusion, multiple adaptive spatial feature fusion (ASFF) layers are introduced to incorporate guidance features in an adaptive and progressive manner, resulting in our ASFFNet. Experiments show that our ASFFNet performs favorably in terms of quantitative and qualitative evaluation, and is effective in generating photo-realistic results on real-world LQ images. The source code and models are available at https://github.com/csxmli2016/ASFFNet.*

1. Introduction

Visual quality is always one of the high concerns of human perception and visual understanding, while recent years have witnessed the rapid progress in the acquisition and sharing of visual content. On the one hand, driven by the development of image capturing and display techniques, more and more high-quality (HQ) visual media are currently available. On the other hand, due to the diversity of acquisition devices, the effect of environment and the motion of objects, low quality (LQ) images and videos are still ubiquitous and inevitably in most real-world applications. Image restoration aims at estimating the clean HQ image from its degraded LQ observation [1, 5, 33], and remains a valuable research topic in computer vision.

(a) Restoration results of a real-world LQ image.



(b) Restoration result of a frame from an old film.

Figure 1: Comparison of exemplar-based face restoration methods. Close-up in the bottom right is the selected guidance for each method. (b) shows the restoration result of a frame from an old film. An animated figure is shown in our suppl.

In this work, we focus on the task of blind face restoration with multiple HQ exemplar images from the same person. The HQ face images play an important role in many applications such as entertainment, surveillance and human-computer interaction, making face restoration highly de-

sired for versatile vision tasks. Fortunately, benefited from the ubiquitous acquisition and sharing of face images, it is very likely that multiple HQ exemplar images of the same person are available for a given degraded LQ face image. Meanwhile, the unprecedented success of face recognition can be exploited to find the HQ exemplar images. For example, the face images in smartphone photo album usually are grouped according to the identities. As for old films, it is also practically feasible to find several HD exemplar images for main actors, which can then be utilized to guide the enhancement of LQ degraded face images in the video frames. The introduction of multi-exemplar guidance can greatly alleviate the difficulty of degradation estimation and image restoration, thereby offering a new perspective for improving blind face restoration.

Recently, several exemplar-based face restoration methods [9, 24] have been suggested. However, most existing methods, *e.g.*, GFRNet [24] and GWAINet [9], are based on single HQ exemplar image, failing to exploit multiple HQ exemplar images for improving face restoration. Consequently, performance degradation may occur when the guidance and degraded images are of very different poses. Moreover, GFRNet [24] and GWAINet [9] use direct concatenation to combine the degraded observation and warped guidance, which is limited in adapting to various degradation settings and exhibits poor generalization to real-world LQ images with unknown degradation process. Fig. 1(a) shows the restoration results of a real-world degraded image. GFRNet [24] and GWAINet [9] are still inadequate not only in reconstructing the details of eyelashes and teeth from the guidance image, but also in removing the noise and compression artifacts from the degraded observation.

In this paper, we present an ASFFNet to address the above issues by exploiting the multi-exemplar setting and properly combining the features from guidance and degraded images. First, we investigate the problem of selecting the optimal guidance image from multiple HQ exemplars. Intuitively, the exemplar with similar pose and expression is preferred given a degraded observation. Thus, we formulate the optimal guidance selection as a weighted least-square (WLS) model defined on landmark sets, where different weights are assigned to different face parts such as eyes and mouths. Moreover, the landmark weights are learned to make the selected guidance image be optimized for restoration performance.

Second, we further investigate the alignment and fusion issues for the guidance and degraded images. In [9, 24], a warping subnet usually is required for spatial alignment. As for our method, the pose difference can be largely alleviated via guidance selection, and we thus can leverage the moving least-square (MLS) [34] to align the guidance and degraded images in the feature space. Then, the adaptive instance normalization (AdaIN) [16] is utilized to translate the illumination of guidance image. Instead of direct concatenation, multiple adaptive spatial feature fusion (ASFF) blocks are adopted for combining the features from guidance and degraded images in an adaptive and progressive manner. In each ASFF block, both facial landmarks, guidance and restored features are considered to generate an attention mask for guiding adaptive feature fusion. When applying to real-world scenarios, the attention mask is still effective in finding where to incorporate guidance features, making our ASFF exhibit good generalization ability to unknown degradations.

Experiments are conducted to evaluate our ASFFNet on both synthetic and real-world degraded images. The quantitative and qualitative results show that our ASFFNet performs favorably against the state-of-the-art methods [9, 24]. As shown in Fig. 1, our ASFFNet exhibits good generalization ability to complex and unknown degradation process, and is effective in generating realistic results on real-world LQ images. The main contributions of this work include:

- For exploiting multiple exemplar images, we adopt a WLS model on landmark sets to select the optimal guidance image, and learn the landmark weights for optimizing the reconstruction performance.
- For compensating the pose and illumination difference between guidance and degraded images, MLS and AdaIN are leveraged to perform spatial alignment and illumination translation in the feature space.
- For combining the features from guidance and degraded images, multiple ASFF blocks are introduced for adaptive and progressive fusion, resulting in our ASFFNet.
- Experiments demonstrate the superiority of our ASFFNet in comparison to state-of-the-arts [9, 24], and also show its potential in handling real-world LQ images from several practical applications.

2. Related Work

2.1. Deep Single Face Image Restoration

Recent years have witnessed the unprecedented success of deep CNNs in several face image restoration tasks, *e.g.*, deblurring [8, 35, 39, 44] and super-resolution [6, 15, 47]. In terms of face hallucination, Huang *et al*. [15] proposed a wavelet-based CNN model that predicts the wavelet coefficients for reconstructing the high-resolution results from a very low resolution face image. Cao *et al*. [6] suggested a reinforcement learning based face hallucination method by specifying the next attended region via recurrent policy network and then recovering it via local enhancement network. As for blind face deblurring, Chrysos *et al*. [8] developed a domain-specific method by exploiting the well-documented face structure. Xu *et al*. [39] presented a generative adversarial network (GAN) for face and text deblurring. Shen *et al*. [35] incorporated the global semantic face priors for

better restoring the shape and details of face images. In general, existing single image restoration methods generalize poorly to real-world LQ face images due to the intrinsic ill-posedness and variety of unknown degradations.

2.2. Exemplar-based Deep Image Restoration

In contrast to single image restoration, the introduction of exemplar image can largely ameliorate the difficulty of image restoration and usually results in notable performance improvement. In guided depth image enhancement, the color guidance image is assumed to be spatially aligned with the degraded depth image. And several CNN methods [14, 17, 25] have been suggested to transfer structural details from intensity image to enhance depth images. However, as for blind face restoration, the guidance and degraded images are usually of different poses. Using a reference image with similar content, Zhang *et al*. [46] adopted a time- and memory-consuming searching scheme to align high-resolution guidance and low-resolution degraded patches in the feature space.

Exemplar-based methods [9, 24] have also been adopted for blind face restoration, where a warping subnet is usually adopted to spatially align the guidance and degraded images. Li *et al*. [24] presented a GFRNet by incorporating landmark loss and total variation regularization to train the warping subnet. Subsequently, Dogan *et al*. [9] suggested a GWANet which can be learned without requiring facial landmarks during training. Moreover, GWANet [9] adopts a feature fusion chain on multiple convolution layers to combine features from the warped guidance and degraded images. However, both GFRNet and GWANet are based on single exemplar image, while multiple HQ exemplar images are usually available in many real-world applications. Besides, when applied to real-world degraded images, the concatenation-based fusion in [9, 24] is still limited in translating the details from guidance to the reconstructed image.

2.3. Adaptive Feature Modulation

GFRNet [24] and GWANet [9] adopt the concatenation-based fusion, which does not consider the illumination difference and spatial variation between warped guidance and degraded images. In arbitrary style transfer, adaptive instance normalization (AdaIN) [16] has been suggested to translate the content image to the desired style. Perez *et al*. [31] suggested a FiLM method to learn feature-wise affine transformation from conditioning information for modulating the network's intermediate features. The feature modulation in AdaIN [16] and FiLM [31], however, is spatially agnostic and inadequate to translate and fuse warped guidance features for face restoration.

For spatially adaptive feature modulation, Wang *et al*. [36] presented a Spatial Feature Transform (SFT) method for super-resolution conditioned on segmentation maps. Besides super-resolution [13, 36], SFT has also been adopted in other vision tasks such as image manipulation [22] and makeup transfer [18]. In semantic image synthesis, Park *et al*. [29] suggested a spatially-adaptive denormalization (SPADE) method for modulating the activations via learning spatial-wise transform. For feature fusion, the gated module [32, 44] has been introduced to estimate a weighted map for combining features from different sources. In this work, both feature translation and fusion are considered for improving the restoration performance and generalization ability of our ASFFNet.

3. Proposed Method

To begin with, blind face restoration with multi-exemplar images is defined as the task of reconstructing the HQ image $\hat{I}^h$ from a degraded face image I^d conditioned on a set of exemplar images $\{I_k^g\}_{k=1}^K$. Without loss of generality, we assume that $\hat{I}^h$, I^d, and I_k^g are of the same image size 256×256. When an image is of different size, we simply resize it to 256×256 with bicubic sampling. Using the landmark detector [4], we further present the 68 landmarks for each image, including L^d, $L_k^g \in \mathbb{R}^{2\times 68}$ $(k = 1, ..., K)$. Then, the proposed blind face restoration model can be formulated as,

$$\hat{I}^h = \mathcal{F}\left(I^d | L^d, \{I_k^g, L_k^g\}_{k=1}^K; \Theta\right), \tag{1}$$

where I^d is the input, $L^d, \{I_k^g, L_k^g\}_{k=1}^K$ are the conditional variables, Θ denotes the model parameters. Benefited from the multi-exemplar guidance, the HQ image can be reconstructed by combining the information from the restoration of degraded input and the translation of the HQ guidance.

Fig. 2 illustrates the network architecture of the proposed ASFFNet consisting of guidance selection, feature extraction, MLS alignment, AdaIN, four ASFF blocks, and reconstruction modules. In particular, we focus on addressing three issues, *i.e*., guidance selection, spatial alignment and illumination translation, and adaptive feature fusion. First, a WLS model is presented to select the guidance image from the set of exemplar images. Second, considering that the pose difference can be largely alleviated after guidance selection, we can leverage the MLS and AdaIN respectively for spatial alignment and illumination translation of the guidance image in the feature space. Due to the MLS alignment is differentiable, the feature extraction subnet can also be end-to-end learnable during training. Finally, multiple ASFF blocks are incorporated to combine the warped guidance features with the restored features from degraded image. In the following, we first describe the methods for handling these three issues, and then give the learning objective for training the whole network. For more details of the network architecture, please refer to the suppl.

3.1. Guidance Selection

For most guided face restoration methods, the performance is diminished by the pose and expression difference between guidance and degraded images. Thus, it is natu-

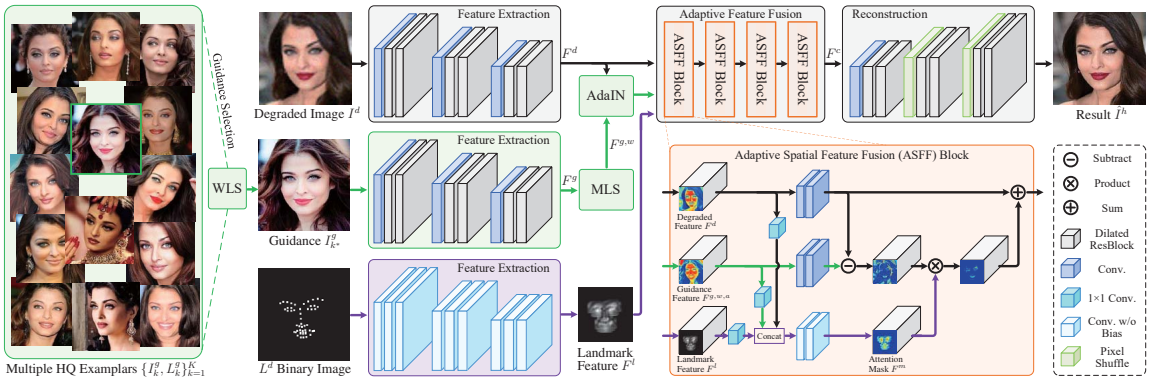


Figure 2: Overview of our ASFFNet.

ral to select the optimal guidance image as the one that has similar pose and expression with the degraded image. For similarity measuring, we adopt the the weighted affine distance between landmark sets by solving a weighted least-square (WLS) model. Taking both pose and expression into account, different weights are given to landmarks. Then, the optimal guidance image can be determined by finding the minimal weighted affine distance,

$$k^*=\arg\min_k \left\{ D_a^2(L^d, L_k^g)=\min_A \sum_{m=1}^{68} w_m \left\| A\tilde{L}_{k,m}^g - L_m^d \right\|^2 \right\}, \quad (2)$$

where $D_a(L^d, L_k^g)$ denotes the affine distance, and w_m denotes the weight for the m-th landmark. L_m^d and $L_{k,m}^g$ denote the m-th landmarks of the degraded image and k-th guidance image, respectively. In particular, $\tilde{L}_{k,m}^g$ is the homogeneous representation of $L_{k,m}^g$ (*e.g.*, a position $[x, y]^T$ in $L_{k,m}^g$ is defined as $[x, y, 1]^T$ in $\tilde{L}_{k,m}^g$ [11]). Given L^d and L_k^g, the closed-form solution of A can be written as,

$$A = L^d W \tilde{L}_k^{g\,T} (\tilde{L}_k^g W \tilde{L}_k^{g\,T})^{-1}, \quad (3)$$

where $W = Diag(w)$ is the diagonal matrix of the landmark weight vector w.

To determine the landmark weights, given a degraded image I^d, we enumerate all the exemplar images $\{I_k^g\}_{k=1}^K$, and find the one with the best performance, *i.e.*, $I_{k^*}^g$, in the forward propagation. Then, we introduce the following auxiliary loss for updating the landmark weights,

$$\ell_w = \sum_{k \neq k^*} \max\left\{0, 1 - \left(D_a^2(L^d, L_k^g) - D_a^2(L^d, L_{k^*}^g)\right)\right\}. \quad (4)$$

By substituting the closed-form solution of A into $D_a^2(L^d, L_{k^*}^g)$, we adopt the back-propagation algorithm to update w based on $\frac{\partial \ell_w}{\partial w}$. The loss constrains the selected guidance image to have a relatively smaller affine distance. For a given test image, the landmark weights are fixed, and we can simply use Eqn. (2) to select the guidance image.

3.2. MLS Alignment and Illumination Translation

Even though the selected guidance image has similar pose and expression with the degraded observation, misalignment remains unavoidable and may introduce visual artifacts to the reconstruction result. In GFRNet [24] and GWAINet [9], a warping subnet is adopted to spatially align the guidance and degraded images. However, the warping subnet is generally difficult to train and may exhibit poor generalization ability due to the lack of direct supervision information. Besides, the guidance and degraded images usually are of different illumination conditions, which should also be considered before feature fusion. In this work, we adopt the MLS method for spatial alignment and AdaIN for illumination translation, which will be described as follows.

MLS Alignment. Instead of learning warping subnet, we suggest to exploit a traditional image deformation method, *i.e.*, moving least-square (MLS), to align the guidance and degraded images in the feature space. Benefited from guidance selection, the pose and expression difference can be largely reduced. Furthermore, due to the differentiability of MLS, the feature extraction subnet can be learnable during training, making that feature extraction and MLS can work collaboratively for robust alignment. The experiments also empirically show that MLS works well in spatial alignment of the degraded image and selected guidance image.

Denote by F^g and L^g the features and landmarks of the optimal guidance image, and L^d the landmarks of the degraded image. For a given position $p = (x, y)$, we introduce a 68×68 position-specific diagonal matrix W_p with the m-th diagonal element $W_p(m, m) = \frac{1}{\|p - L_m^d\|^2}$. Then, the position-specific affine matrix can be obtained by,

$$M_p = L^g W_p \tilde{L}^{d\,T} (\tilde{L}^d W_p \tilde{L}^{d\,T})^{-1}, \quad (5)$$

where $\tilde{L}^d$ is the homogeneous representation of L^d. Let $\hat{p} = M_p \tilde{p}$, and $\mathcal{N}$ be the 4-nearest neighbors of $\hat{p} = (\hat{x}, \hat{y})$. The warped feature can be obtained by bilinear interpolation,

$$F^{g,w}(x, y) = \sum_{(x', y') \in \mathcal{N}} F^g(x', y') \max(0, 1 - |\hat{x} - x'|) \max(0, 1 - |\hat{y} - y'|), \quad (6)$$

where (x, y) is a position in the degraded input while $(\hat{x}, \hat{y})$ is the corresponding position in guidance image. We note that Eqns. (5) and (6) are differentiable. Thus, the feature extraction can also be end-to-end learnable during training.

AdaIN. For arbitrary style transfer, the AdaIN [16] is

introduced to translate the content features to the desired style. Analogously, we treat the illumination as a kind of style, and use AdaIN to adjust the warped guidance feature to have similar illumination with the restored feature of degraded image. Denote by F^d and $F^{g,w}$ the restored features from the degraded image and the warped guidance features from the guidance image. The AdaIN can be written as,

$$F^{g,w,a} = \sigma(F^d)\left(\frac{F^{g,w} - \mu(F^{g,w})}{\sigma(F^{g,w})}\right) + \mu(F^d), \quad (7)$$

where $\sigma(\cdot)$ and $\mu(\cdot)$ denote the mean and standard deviation. With MLS and AdaIN, $F^{g,w,a}$ can thus be aligned with F^d by space and illumination.

3.3. Adaptive Feature Fusion

After MLS alignment and AdaIN, the misalignment and illumination difference between degraded and guidance images can be largely reduced. Thus, we further combine the warped guidance features with the restored features to reconstruct the HQ image. In GFRNet [24], the guidance and degraded images are concatenated as the input to the reconstruction subnet. GWAINet [9] also adopts the concatenation-based fusion but is performed in multiple feature layers. However, the concatenation-based fusion is still limited in exploiting the complementarity between the guidance and degraded images. Thus, we present multiple adaptive spatial feature fusion (ASFF) blocks for progressively fusing the warped guidance and restored features.

On the one hand, the guidance image generally contains more high-quality facial details and is more reliable for most face components. On the other hand, considering that $F^{g,w,a}$ and F^d are spatially variant and convey complementary information, they can be combined for better reconstruction of the HQ image. Considering two examples: (i) The guidance image generally has a different background with the degraded image, and thus the restored features from degraded image are more reliable for background region. (ii) When the *mouth* of guidance is close while that of degraded image is open, we should reconstruct the *teeth* mainly based on restored features instead of warped guidance features. Therefore, we introduce an attention mask F^m to guide the fusion of $F^{g,w,a}$ and F^d. Naturally, the generation of F^m should consider $F^{g,w,a}$, F^d, and landmarks. And we adopt a landmark feature extraction subnet with the output F^l. Then, we take $F^{g,w,a}$, F^d and F^l as the input, and exploit a gating module to generate the attention mask F^m. For efficiency, 1×1 convolution is first applied to $F^{g,w,a}$, F^d and F^l to reduce the feature channels. Finally, the output of each ASFF block can be written as,

$$\begin{aligned}\mathcal{F}_{\text{ASFF}}(F^d, F^{g,w,a}) &= (1 - F^m) \circ \mathcal{F}_d(F^d) + F^m \circ \mathcal{F}_g(F^{g,w,a}) \\ &= \mathcal{F}_d(F^d) + F^m \circ (\mathcal{F}_g(F^{g,w,a}) - \mathcal{F}_d(F^d)),\end{aligned} \quad (8)$$

where $\circ$ denotes the element-wise product. Please refer to the suppl. for the detailed architectures for the updating of F^d, $F^{g,w,a}$, F^l, as well as the attention mask F^m.

As opposed to the concatenation-based fusion in GFRNet [24] and GWAINet [9], the ASFF is a more flexible fusion method and can be adapted to different degradation settings and image contents. Analogous to the multi-layer concatenation in [9], we deploy multiple ASSF blocks to facilitate progressive fusion. Benefited from the adaptive and progressive fusion, our ASFFNet can exhibit better generalization ability to real-world LQ face images with complex and unknown degradation process.

Given the combined feature F^c after the ASFF blocks, we further use a reconstruction subnet which consists of two pixel shuffle layers with each followed by two residual blocks. Therefore, the final result can be obtained by,

$$\hat{I}^h = \mathcal{F}_R\left(F^c; \Theta_R\right), \quad (9)$$

where Θ_R is the parameters of reconstruction subnet.

3.4. Learning Objective

Denote by $\hat{I}^h$ and I the reconstructed and ground-truth images. In general, the reconstructed image I is required to faithfully approximate the ground-truth image I and to be photo-realistic. Therefore, the objective involves two loss functions, *i.e.*, reconstruction loss and photo-realistic loss.

The reconstruction loss is introduced to constrain the reconstructed image to approximate the ground-truth, which involves two terms. First, the mean square error (MSE) is adopted to measure the difference between $\hat{I}^h$ and I,

$$\ell_{\text{MSE}} = \frac{1}{CHW}\|\hat{I}^h - I\|^2, \quad (10)$$

where C, H and W denote the channel, height and width of the image. Second, to improve the visual quality of the reconstructed image, we adopt the perceptual loss [19] defined on the VGGFace [30] feature space. In particular, the perceptual loss is adopted to constrain the reconstructed image $\hat{I}^h$ to approximate the ground-truth I in feature space,

$$\ell_{\text{perc}} = \sum_{u=1}^{4} \frac{1}{C_u H_u W_u} \left\|\Psi_u(\hat{I}^h) - \Psi_u(I)\right\|^2, \quad (11)$$

where Ψ_u represents the features from the u-th layer of pre-trained VGGFace model. In this work, we set $u \in [1, 2, 3, 4]$. The overall reconstruction loss is formulated as:

$$\mathcal{L}_{\text{rec}} = \lambda_{\text{MSE}}\ell_{\text{MSE}} + \lambda_{\text{perc}}\ell_{\text{perc}}, \quad (12)$$

where λ_{MSE} and λ_{perc} are the tradeoff parameters.

For photo-realistic reconstruction, we also consider two terms. First, we adopt the style loss [12], which can be used as an alternative of adversarial loss and is effective in generating visually plausible result with fine details [26]. In particular, style loss is defined on the Gram matrix of feature map for each layer from $u \in [1, 2, 3, 4]$,

$$\ell_{\text{style}} = \sum_{u=1}^{4} \frac{1}{C_u H_u W_u} \left\|\Psi_u(\hat{I}^h)^T\Psi_u(\hat{I}^h) - \Psi_u(I)^T\Psi_u(I)\right\|^2. \quad (13)$$

Second, adversarial loss has also been extensively used in many image generation and translation tasks as an effective method to improve visual quality. To stabilize the discriminator learning, we use SNGAN [28] by introducing spectral normalization on the weights of each convolution layer. Furthermore, we adopt the hinge version of adversarial loss to train the discriminator and generator [3, 42], which can be formulated as,

$$\begin{aligned}\ell_{\text{adv},D} = &\mathbb{E}_{I\sim P(I)}\left[\min(0,-1+D(I))\right] \\ &+\mathbb{E}_{\hat{I}^h\sim P(\hat{I}^h)}\left[\min(0,-1-D(\hat{I}^h))\right],\end{aligned} \tag{14}$$

$$\ell_{\text{adv},G} = -\mathbb{E}_{I^d\sim P(I^d)}\left[D\left(\mathcal{F}\left(I^d|L^d,\{I_k^g,L_k^g\}_{k=1}^K;\Theta\right)\right)\right]. \tag{15}$$

Here, $\ell_{\text{adv},D}$ is used to update the discriminator, while $\ell_{\text{adv},G}$ is adopted to update the ASFFNet for blind face restoration. Then, the overall photo-realistic loss can be written as,

$$\mathcal{L}_{\text{real}} = \lambda_{\text{style}}\ell_{\text{style}} + \lambda_{\text{adv}}\ell_{\text{adv},G}, \tag{16}$$

where λ_{style} and λ_{adv} are the tradeoff parameters.

To sum up, the overall objective function is defined as,

$$\mathcal{L} = \mathcal{L}_{\text{rec}} + \mathcal{L}_{\text{real}}. \tag{17}$$

4. Experiments

The proposed ASFFNet can be used to handle several usual degradation types, *e.g.*, noise, compression artifact, blurring and downsampling, and their combinations. For quantitative evaluation, we use the tasks of $\times 4$ and $\times 8$ super-resolution (SR) in conjunction with noise and blurring as the examples, and compare our ASFFNet with the state-of-the-art SR (*e.g.*, RCAN [45] and ESRGAN [37]), blind deblurring (*e.g.*, DeblurGANV2 [21]), face SR (*e.g.*, TDAE [41], WaveletSR [15], SCGAN [39], GWAINet [9], GFRNet [24]) methods. For a fair comparison, we retrain *ESRGAN, *RCAN, and *SCGAN, and finetune *GFRNet and *GWAINet using our training data. It is also noted that TDAE [41] and GWAINet [9] can only handle $\times 8$ SR while ESRGAN [37] and SCGAN [39] can only handle $\times 4$ SR. We also give the qualitative results on synthetic and real-world degraded face images. More visual results are provided in the suppl.

4.1. Dataset and Experimental Settings

Using the images from VGGFace2 [7], we build a dataset for face restoration with multi-exemplar guidance images. The Laplacian gradient is utilized to assess image quality and remove those with low scores. Then, we detect the 68 facial landmarks using [4], crop and resize the face region to 256×256 based on the convex hull of landmarks. By grouping the remained images based on the identity, we build our dataset containing 106,000 groups of face images, in which each group has 3$\sim$10 HQ exemplar images. Furthermore, we divide it into a training set of 100,000 groups, a validation set of 4,000 groups, and a testing set of 2,000 groups. We also note that the training, testing and validation sets are not overlapped in terms of either identity and image. For flexible quantitative evaluation, each test group is required to exactly have 10 exemplar images. We also build two other testing sets on CelebA [27] and CASIA-WebFace [40], where each set contains 2,000 groups and each group has 3$\sim$10 HQ exemplar images. PSNR, SSIM [38] and LPIPS [43] are adopted as the quantitative performance metrics.

In order to generate synthetic training and testing data, we adopt the degradation model adopted in [24],

$$I^d = ((I \otimes \mathbf{k}) \downarrow_s + \mathbf{n}_\sigma)_{JPEG_q}, \tag{18}$$

where $\otimes$ denotes the convolution operation, $\mathbf{k}$ denotes the blur kernel, $\downarrow_s$ denotes the $\times s$ bicubic downsampler, $\mathbf{n}_\sigma$ denotes Gaussian noise with the noise level σ, and $JPEG_q$ stands for the JPEG compression with quality factor q. In particular, we consider two types of blur kernels, *i.e.*, Gaussian blur with $\varrho \in \{1 : 0.1 : 3\}$ and 32 motion blur kernels from [2, 23]. We randomly sample the scale s from $\{1 : 0.1 : 8\}$, the noise level σ from $\{0 : 1 : 15\}$, and the compression quality factor q from $\{10 : 1 : 60\}$. According to [24], the degradation model can generate realistic LQ images for training guided face restoration model.

We adopt the ADAM optimizer [20] to train our ASFFNet with batch size of 8 and the momentum parameters $\beta_1 = 0.5$ and $\beta_2 = 0.999$. The initial learning rate is 2×10^{-4} and is decreased by 0.5 when the reconstruction loss on validation set is non-decreasing. Several common data augmentation methods, *e.g.*, randomly cropping and horizontal flipping, are also exploited during training. Chromatic transformations, *e.g.*, brightness and contrast [48], are also used to increase the image diversity. The tradeoff parameters of loss terms are set as: $\lambda_{\text{MSE}} = 300$, $\lambda_{\text{perc}} = 5$, $\lambda_{\text{style}} = 1$, and $\lambda_{\text{adv}} = 2$. All the experiments are conducted on a PC equipped with a RTX 2080 Ti GPU and it takes about 3 days to train an ASFFNet model.

4.2. Ablation Studies

Using the $\times 4$ and $\times 8$ SR tasks on the VGGFace2 testing set, three kinds of ablation studies are conducted to assess the effect of multi-exemplar images, ASFF based fusion, MLS and AdaIN modules.

(1) ***Multi-exempler images.*** We randomly select a fixed number of guidance images from each test group to implement four variants of our ASFFNet, *i.e.*, Ours (#10), Ours (#5), Ours (#3), Ours (#1). From Table 1, it can be seen that better quantitative results are obtained along with the use of more exemplar images. Moreover, the gain brought by increasing exemplar images is more obvious for $\times 8$ SR, indicating that the use of multi-exemplar images is more effective for difficult task. As shown in Fig. 3, when the number of exemplar images is inadequate, the selected guidance image is more likely to have different pose and expression with the degraded image, and visual artifacts can still be

Type	×4			×8		
	PSNR↑	SSIM↑	LPIPS↓	PSNR↑	SSIM↑	LPIPS↓
Ours (#1)	27.99	0.925	0.107	24.19	0.873	0.252
Ours (#3)	28.03	0.928	0.104	24.30	0.879	0.247
Ours (#5)	28.06	0.930	0.103	24.33	0.881	0.238
Ours (#10)	28.07	0.930	0.103	24.34	0.881	0.238

Table 1: Comparisons of ASFFNet on different exemplar numbers.

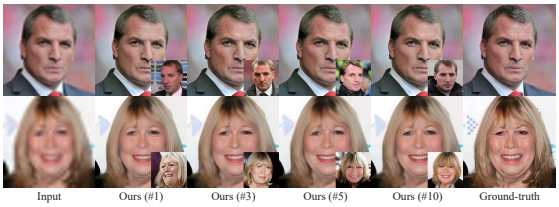


Figure 3: Visual comparisons of ASFFNet on different exemplar numbers. Close-up in the bottom right is the selected guidance.

observed from the reconstruction results. The comparisons with GFRNet and GWAINet on the same random guidance are reported in our suppl.

(2) ***ASFF-based fusion.*** For evaluating progressive fusion, we implement several ASFFNet models with different number of ASFF blocks, Ours (1-ASFF), Ours (2-ASFF), Ours (4-ASFF), Ours (8-ASFF). For evaluating adaptive feature fusion, we consider four variants of our ASFFNet, *i.e.*, Ours (1-Concat) by substituting ASFF with concatenation-based fusion in Ours (1-ASFF), Ours (4-Concat), Ours (w/o 1-Atten) by removing the attention mask in Ours (1-ASFF) and Ours (w/o 4-Atten). From Table 2, our ASFFNet outperforms Ours (Concat) and Ours (w/o Atten) in terms of the three quantitative metrics, clearly demonstrating the effectiveness of adaptive spatial feature fusion. Moreover, benefited from progressive fusion, better performance can be achieved by stacking more ASFF blocks, and the performance begin to be saturating when the number of ASFF blocks is higher than 4. Thus we adopt Ours (4-ASFF) as the default ASFFNet model. Fig. 4 shows the results by different fusion methods. One can see that Ours (4-ASFF) is effective in generating sharp result with fine details while suppressing visual artifacts.

(3) ***Spatial alignment and illumination translation.*** We consider three ASFFNet variants, *i.e.*, Ours (w/o AdaIN) by removing the AdaIN module, Ours (w/o MLS) by removing the MLS module, and Ours (Untrain F_g) by initializing the guidance feature extraction subnet F_g with VGGFace network and then keeping unchanged during training. It can be seen from Table 2 that both spatial alignment and illumination translation is beneficial to the reconstruction performance. The differentiability of MLS makes F_g be learnable and also benefits quantitative performance.

4.3. Experiments on Synthetic Datasets

Table 3 lists the quantitative results of ×4 and ×8 SR on three testing datasets, *i.e.*, VGGFace2, CelebA, and CASIA-WebFace. As for GFRNet [24] and GWAINet [9], we adopt three settings to report their results, (i) using the frontal guidance (*i.e.*, GFRNet and GWAINet), (ii) using the selected guidance by our method (*i.e.*, †GFRNet and †GWAINet), (iii) fine-tuning them using our training data and testing on our selected guidance (*i.e.*, *GFRNet and *GWAINet). For both the two tasks (*i.e.*, ×4 and ×8 SR) and the three datasets, our ASFFNet can achieve the best quantitative metrics. The results indicate that our ASFFNet is superior in guided face restoration and the model learned on VGGFace2 training can be well generalized to other datasets. Except GFRNet [24] and GWAINet [9], the other competing methods do not consider guidance image, which may explain their relatively poor performance. Our ASFFNet also outperforms than GFRNet [24] and GWAINet [9] on three settings, which may be ascribed to the effectiveness of adaptive and progressive ASFF blocks.

In terms of running time, our ASFFNet is comparable to GFRNet [24] (about 31 ms) for a 256×256 image, and can be 3× faster than GWAINet [9].

Figs. 5 and 6 present the visual comparison with the competing methods, including *RCAN, *ESRGAN, WaveletSR [15], SCGAN [39], DeblurGANV2 [21], *GFRNet and GWAINet [9]. More results are given in the suppl. *RCAN and *ESRGAN are suggested for SISR and cannot perform well even taking both degraded and guidance images as the input, thereby performing limited on blind face restoration. SCGAN [39] and WaveletSR [15] can be used in face deblurring but cannot faithfully recover the real face structures. By leveraging guidance image, *GFRNet and GWAINet [9] can well reconstruct HQ face images, but are limited in retaining small-scale details. In comparison, our ASFFNet is more effective in reconstructing HQ face images with more realistic details, especially in the regions with beard and eyelash.

Type	×4			×8		
	PSNR↑	SSIM↑	LPIPS↓	PSNR↑	SSIM↑	LPIPS↓
Ours (1-Concat)	27.56	0.913	0.243	23.32	0.863	0.301
Ours (4-Concat)	27.59	0.915	0.212	23.42	0.864	0.294
Ours (w/o 1-Atten)	27.57	0.915	0.167	23.46	0.865	0.279
Ours (w/o 4-Atten)	27.83	0.924	0.139	23.63	0.872	0.249
Ours (1-ASFF)	27.59	0.916	0.141	23.51	0.865	0.251
Ours (2-ASFF)	27.67	0.923	0.124	23.72	0.873	0.248
Ours (4-ASFF)	28.07	0.930	0.103	24.34	0.881	0.238
Ours (8-ASFF)	28.07	0.930	0.103	24.34	0.881	0.237
ASFF (w/o AdaIN)	28.01	0.929	0.112	24.15	0.880	0.240
ASFF (w/o MLS)	26.45	0.903	0.214	23.12	0.854	0.263
ASFF (untrain F_g)	27.58	0.917	0.155	23.47	0.865	0.281

Table 2: Comparisons of different ASFFNet variants.

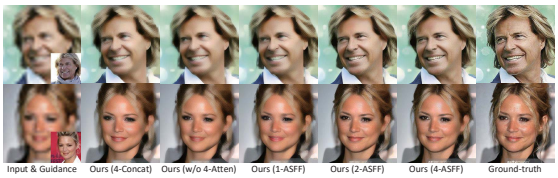


Figure 4: Visual comparison of different feature fusion methods.

4.4. Results on Real-world LQ Images

The AdaIN-based illumination translation and ASFF-based adaptive fusion are also helpful to the generalization ability of our ASFFNet in handling real-world LQ images. To illustrate this point, Fig. 7 provides the results of GFRNet [24] and our ASFFNet on four real-world LQ images with unknown degradation process, in which the face sizes

Methods	VGGFace2						CelebA						CASIA-WebFace					
	× 4			× 8			× 4			× 8			× 4			× 8		
	PSNR↑	SSIM↑	LPIPS↓	PSNR↑	SSIM↑	LPIPS↓	PSNR↑	SSIM↑	LPIPS↓	PSNR↑	SSIM↑	LPIPS↓	PSNR↑	SSIM↑	LPIPS↓	PSNR↑	SSIM↑	LPIPS↓
RCAN [45]	25.50	.896	.282	22.18	.833	.299	27.51	.913	.212	24.14	.876	.249	28.31	.931	.263	26.61	.907	.402
*RCAN	26.29	.902	.224	23.03	.845	.276	27.92	.924	.210	24.54	.880	.244	29.18	.934	.256	26.73	.907	.383
ESRGAN [37]	24.54	.880	.216	-	-	-	27.18	.910	.180	-	-	-	28.03	.929	.275	-	-	-
*ESRGAN	25.34	.892	.177	-	-	-	27.89	.922	.179	-	-	-	29.36	.937	.257	-	-	-
DeblurGANV2 [21]	24.73	.885	.219	21.87	.827	.310	27.39	.912	.206	23.83	.876	.259	29.13	.934	.234	26.58	.906	.393
TDAE [41]	-	-	-	18.38	.768	.392	-	-	-	18.98	.788	.388	-	-	-	19.79	.800	.381
WaveletSR [15]	24.33	.879	.234	21.49	.825	.278	26.52	.907	.220	24.02	.875	.230	29.11	.933	.283	25.11	.886	.379
SCGAN [39]	23.80	.877	.147	-	-	-	26.01	.901	.139	-	-	-	27.53	.914	.267	-	-	-
*SCGAN	23.86	.878	.142	-	-	-	26.12	.903	.135	-	-	-	27.68	.915	.260	-	-	-
GWAINet [9]	-	-	-	23.54	.871	.273	-	-	-	25.37	.897	.219	-	-	-	27.02	.909	.258
†GWAINet [9]	-	-	-	23.65	.876	.266	-	-	-	25.56	.900	.212	-	-	-	27.11	.909	.253
*GWAINet	-	-	-	23.87	.879	.261	-	-	-	25.77	.901	.210	-	-	-	27.18	.910	.250
GFRNet [24]	27.49	.910	.130	23.07	.857	.297	28.45	.929	.122	25.12	.893	.241	30.13	.936	.225	26.56	.906	.334
†GFRNet [24]	27.58	.914	.127	23.48	.864	.293	28.69	.932	.116	25.49	.898	.230	30.39	.939	.206	26.83	.908	.322
*GFRNet	27.66	.921	.122	23.85	.879	.263	29.01	.933	.113	25.93	.901	.227	30.80	.941	.181	27.19	.912	.307
Ours	**28.07**	**.930**	**.103**	**24.34**	**.881**	**.238**	**29.55**	**.937**	**.056**	**26.39**	**.905**	**.185**	**31.08**	**.948**	**.099**	**27.69**	**.921**	**.219**

Table 3: Quantitative results on image super-resolution (×4 and ×8). ↑ (↓) represents higher (lower) is better.

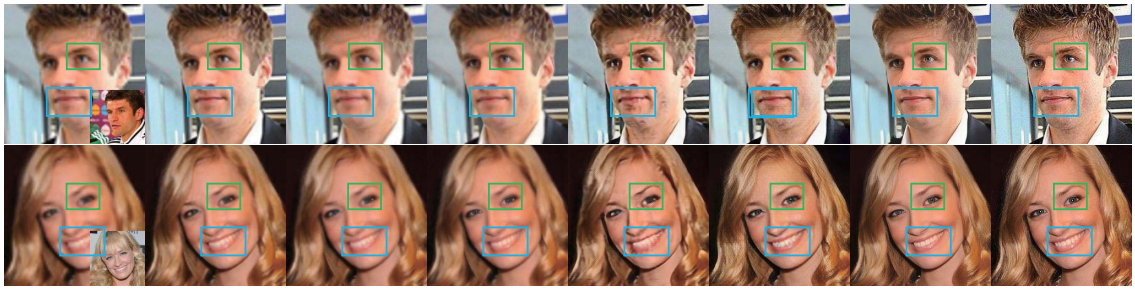


Figure 5: The ×4 SR results by the competing methods. Green and blue boxes are the improvement regions.

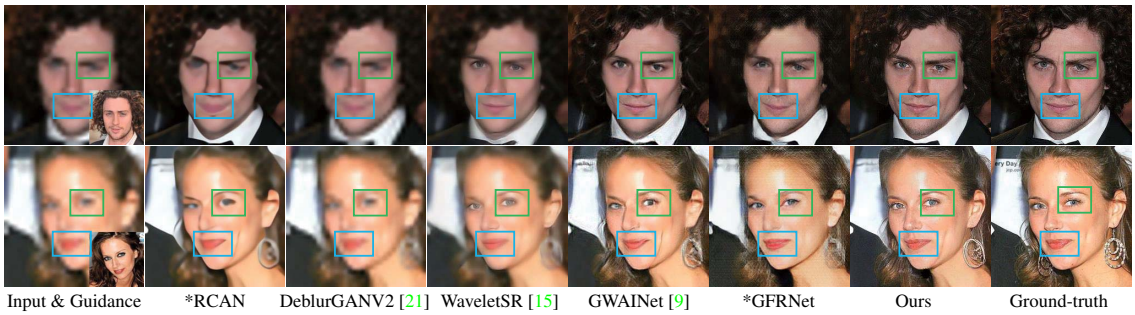


Figure 6: The ×8 SR results by the competing methods. Best view it by zooming in the screen.

are lower than 80×80. GFRNet [24] generally improves the visual quality in contrast to the input images. However, obvious artifacts are likely to be introduced in the reconstruction results, partially due to the poor adaptivity and generalization ability of concatenation-based fusion. In comparison, our ASFFNet can reconstruct more texture details with less artifacts, and exhibits better robustness to complex and unknown degradation process which is of the great value in real-world applications. More results on real-world LQ images can be found in the suppl.

5. Conclusion

This paper presented an enhanced blind face restoration model, *i.e.*, ASFFNet, by addressing three issues, *i.e.*, multi-exemplar images, spatial alignment and illumination translation, and adaptive feature fusion. For guidance selection from multiple exemplar images, we adopt a weighted least-square model on facial landmarks, and suggest a method to learn landmark weights. Moving least-square and adaptive instance normalization are then used for spatial alignment and illumination translation of guidance image in the feature space. And multiple ASFF blocks are finally deployed for adaptive and progressive fusion of restored features and guidance features for better HQ image reconstruction. Experiments show that our ASFFNet performs favorably against the competing methods, and exhibits better visual quality and generalization ability to real-world LQ images.

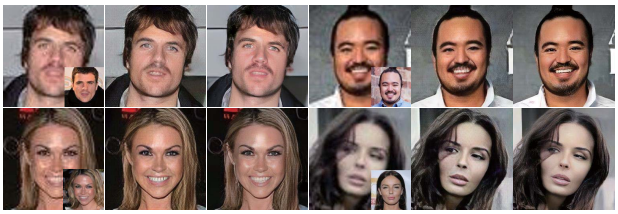


Figure 7: Visual comparison on real-world LQ images.

Acknowledgments. This work was supported in part by the National Natural Science Foundation of China under grant No. 61671182, 61872118, U19A2073.

References

[1] Harry C Andrews and Bobby Ray Hunt. Digital image restoration. *Prentice-Hall Signal Processing Series, Englewood Cliffs: Prentice-Hall, 1977*, 1977. 1

[2] Giacomo Boracchi and Alessandro Foi. Modeling the performance of image restoration from motion blur. *TIP*, 2012. 6

[3] Andrew Brock, Jeff Donahue, and Karen Simonyan. Large scale gan training for high fidelity natural image synthesis. *ICLR*, 2019. 6

[4] 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 *ICCV*, 2017. 3, 6

[5] Patrizio Campisi and Karen Egiazarian. *Blind image deconvolution: theory and applications*. CRC press, 2016. 1

[6] Qingxing Cao, Liang Lin, Yukai Shi, Xiaodan Liang, and Guanbin Li. Attention-aware face hallucination via deep reinforcement learning. In *CVPR*, 2017. 2

[7] Qiong Cao, Li Shen, Weidi Xie, Omkar M Parkhi, and Andrew Zisserman. Vggface2: A dataset for recognising faces across pose and age. In *FG*. IEEE, 2018. 6

[8] Grigorios G Chrysos and Stefanos Zafeiriou. Deep face deblurring. In *CVPRW*, 2017. 2

[9] Berk Dogan, Shuhang Gu, and Radu Timofte. Exemplar guided face image super-resolution without facial landmarks. In *CVPRW*, 2019. 1, 2, 3, 4, 5, 6, 7, 8

[10] Chao Dong, Chen Change Loy, Kaiming He, and Xiaoou Tang. Image super-resolution using deep convolutional networks. *TPAMI*, 2015.

[11] Wolfgang Förstner and Bernhard P Wrobel. *Photogrammetric computer vision*. Springer, 2016. 4

[12] Leon A Gatys, Alexander S Ecker, and Matthias Bethge. Image style transfer using convolutional neural networks. In *CVPR*, 2016. 5

[13] Jinjin Gu, Hannan Lu, Wangmeng Zuo, and Chao Dong. Blind super-resolution with iterative kernel correction. In *CVPR*, 2019. 3

[14] Shuhang Gu, Wangmeng Zuo, Shi Guo, Yunjin Chen, Chongyu Chen, and Lei Zhang. Learning dynamic guidance for depth image enhancement. In *CVPR*, 2017. 3

[15] Huaibo Huang, Ran He, Zhenan Sun, and Tieniu Tan. Wavelet-srnet: A wavelet-based cnn for multi-scale face super resolution. In *ICCV*, pages 1689–1697, 2017. 2, 6, 7, 8

[16] Xun Huang and Serge Belongie. Arbitrary style transfer in real-time with adaptive instance normalization. In *ICCV*, 2017. 2, 3, 4

[17] Tak-Wai Hui, Chen Change Loy, and Xiaoou Tang. Depth map super-resolution by deep multi-scale guidance. In *ECCV*. Springer, 2016. 3

[18] Wentao Jiang, Si Liu, Chen Gao, Jie Cao, Ran He, Jiashi Feng, and Shuicheng Yan. Psgan: Pose-robust spatial-aware gan for customizable makeup transfer. *arXiv preprint arXiv:1909.06956*, 2019. 3

[19] Justin Johnson, Alexandre Alahi, and Li Fei-Fei. Perceptual losses for real-time style transfer and super-resolution. In *ECCV*, 2016. 5

[20] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014. 6

[21] Orest Kupyn, Tetiana Martyniuk, Junru Wu, and Zhangyang Wang. Deblurgan-v2: Deblurring (orders-of-magnitude) faster and better. In *ICCV*, 2019. 6, 7, 8

[22] Cheng-Han Lee, Ziwei Liu, Lingyun Wu, and Ping Luo. Maskgan: towards diverse and interactive facial image manipulation. *arXiv preprint arXiv:1907.11922*, 2019. 3

[23] Anat Levin, Yair Weiss, Fredo Durand, and William T Freeman. Understanding and evaluating blind deconvolution algorithms. In *CVPR*, 2009. 6

[24] Xiaoming Li, Ming Liu, Yuting Ye, Wangmeng Zuo, Liang Lin, and Ruigang Yang. Learning warped guidance for blind face restoration. In *ECCV*, 2018. 1, 2, 3, 4, 5, 6, 7, 8

[25] Yijun Li, Jia-Bin Huang, Narendra Ahuja, and Ming-Hsuan Yang. Deep joint image filtering. In *ECCV*. Springer, 2016. 3

[26] Guilin Liu, Fitsum A Reda, Kevin J Shih, Ting-Chun Wang, Andrew Tao, and Bryan Catanzaro. Image inpainting for irregular holes using partial convolutions. In *ECCV*, 2018. 5

[27] Ziwei Liu, Ping Luo, Xiaogang Wang, and Xiaoou Tang. Deep learning face attributes in the wild. In *ICCV*, 2015. 6

[28] Takeru Miyato, Toshiki Kataoka, Masanori Koyama, and Yuichi Yoshida. Spectral normalization for generative adversarial networks. ICLR, 2018. 6

[29] Taesung Park, Ming-Yu Liu, Ting-Chun Wang, and Jun-Yan Zhu. Semantic image synthesis with spatially-adaptive normalization. In *CVPR*, 2019. 3

[30] Omkar M Parkhi, Andrea Vedaldi, and Andrew Zisserman. Deep face recognition. In *BMVC*, 2015. 5

[31] Ethan Perez, Florian Strub, Harm De Vries, Vincent Dumoulin, and Aaron Courville. Film: Visual reasoning with a general conditioning layer. In *AAAI*, 2018. 3

[32] Wenqi Ren, Lin Ma, Jiawei Zhang, Jinshan Pan, Xiaochun Cao, Wei Liu, and Ming-Hsuan Yang. Gated fusion network for single image dehazing. In *CVPR*, 2018. 3

[33] William Hadley Richardson. Bayesian-based iterative method of image restoration. *JoSA*, 1972. 1

[34] Scott Schaefer, Travis McPhail, and Joe Warren. Image deformation using moving least squares. In *ACM TOG*. 2

[35] Ziyi Shen, Wei-Sheng Lai, Tingfa Xu, Jan Kautz, and Ming-Hsuan Yang. Deep semantic face deblurring. In *CVPR*, 2018. 2

[36] Xintao Wang, Ke Yu, Chao Dong, and Chen Change Loy. Recovering realistic texture in image super-resolution by deep spatial feature transform. In *CVPR*, 2018. 3

[37] 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 *ECCVW*, 2018. 6, 8

[38] Zhou Wang, Alan C Bovik, Hamid R Sheikh, Eero P Simoncelli, et al. Image quality assessment: from error visibility to structural similarity. *IEEE transactions on image processing*, 2004. 6

[39] Xiangyu Xu, Deqing Sun, Jinshan Pan, Yujin Zhang, Hanspeter Pfister, and Ming-Hsuan Yang. Learning to super-resolve blurry face and text images. In *ICCV*, 2017. 2, 6, 7, 8

[40] Dong Yi, Zhen Lei, Shengcai Liao, and Stan Z Li. Learning face representation from scratch. *arXiv preprint arXiv:1411.7923*, 2014. 6

[41] Xin Yu and Fatih Porikli. Hallucinating very low-resolution unaligned and noisy face images by transformative discriminative autoencoders. In *CVPR*, 2017. 6, 8

[42] Han Zhang, Ian Goodfellow, Dimitris Metaxas, and Augustus Odena. Self-attention generative adversarial networks. *arXiv preprint arXiv:1805.08318*, 2018. 6

[43] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *CVPR*, 2018. 6

[44] Xinyi Zhang, Hang Dong, Zhe Hu, Wei-Sheng Lai, Fei Wang, and Ming-Hsuan Yang. Gated fusion network for joint image deblurring and super-resolution. 2018. 2, 3

[45] Yulun Zhang, Kunpeng Li, Kai Li, Lichen Wang, Bineng Zhong, and Yun Fu. Image super-resolution using very deep residual channel attention networks. In *ECCV*, 2018. 6, 8

[46] Zhifei Zhang, Zhaowen Wang, Zhe Lin, and Hairong Qi. Image super-resolution by neural texture transfer. In *CVPR*, 2019. 3

[47] Shizhan Zhu, Sifei Liu, Chen Change Loy, and Xiaoou Tang. Deep cascaded bi-network for face hallucination. In *ECCV*. Springer, 2016. 2

[48] Barret Zoph, Ekin D Cubuk, Golnaz Ghiasi, Tsung-Yi Lin, Jonathon Shlens, and Quoc V Le. Learning data augmentation strategies for object detection. *arXiv preprint arXiv:1906.11172*, 2019. 6