Learning Warped Guidance for Blind Face Restoration

Xiaoming Li[1], Ming Liu[1], Yuting Ye[1], Wangmeng Zuo[1(✉)], Liang Lin[2],
and Ruigang Yang[3]

[1]School of Computer Science and Technology, Harbin Institute of Technology, China
csxmli@hit.edu.cn, csmliu@outlook.com, yeyuting.jlu@gmail.com,
wmzuo@hit.edu.cn
[2]School of Data and Computer Science, Sun Yat-sen University, China
linliang@ieee.org
[3]Department of Computer Science, University of Kentucky, USA
ryang@cs.uky.edu

Abstract. This paper studies the problem of blind face restoration from an unconstrained blurry, noisy, low-resolution, or compressed image (i.e., degraded observation). For better recovery of fine facial details, we modify the problem setting by taking both the degraded observation and a high-quality guided image of the same identity as input to our guided face restoration network (GFRNet). However, the degraded observation and guided image generally are different in pose, illumination and expression, thereby making plain CNNs (e.g., U-Net [1]) fail to recover fine and identity-aware facial details. To tackle this issue, our GFRNet model includes both a warping subnetwork (WarpNet) and a reconstruction subnetwork (RecNet). The WarpNet is introduced to predict flow field for warping the guided image to correct pose and expression (i.e., warped guidance), while the RecNet takes the degraded observation and warped guidance as input to produce the restoration result. Due to that the ground-truth flow field is unavailable, landmark loss together with total variation regularization are incorporated to guide the learning of WarpNet. Furthermore, to make the model applicable to blind restoration, our GFRNet is trained on the synthetic data with versatile settings on blur kernel, noise level, downsampling scale factor, and JPEG quality factor. Experiments show that our GFRNet not only performs favorably against the state-of-the-art image and face restoration methods, but also generates visually photo-realistic results on real degraded facial images.

Keywords: Face hallucination · blind image restoration · flow field · convolutional neural networks

1 Introduction

Face restoration aims to reconstruct high quality face image from degraded observation for better display and further analysis [2–12]. In the ubiquitous imaging era, imaging sensors are embedded into many consumer products and surveillance devices, and more and more images are acquired under unconstrained

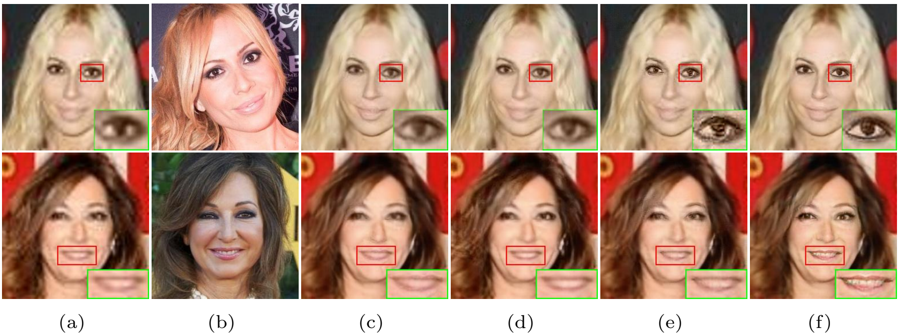


Fig. 1. Restoration results on real low quality images: (a) real low quality image, (b) guided image, and the results by (c) U-Net [1] by taking low quality image as input, (d) U-Net [1] by taking both guided image and low quality image as input, (e) our GFRNet without landmark loss, and (f) our full GFRNet model. Best viewed by zooming in the screen.

scenarios. Consequently, low quality face images cannot be completely avoided during acquisition and communication due to the introduction of low-resolution, defocus, noise and compression. On the other hand, high quality face images are sorely needed for human perception, face recognition [13] and other face analysis [14] tasks. All these make face restoration a very challenging yet active research topic in computer vision.

Many studies have been carried out to handle specific face restoration tasks, such as denoising [15,16], hallucination [2–11] and deblurring [12]. Most existing methods, however, are proposed for handling a single specific face restoration task in a non-blind manner. In practical scenario, it is more general that both the degradation types and degradation parameters are unknown in advance. Therefore, more attentions should be given to blind face restoration. Moreover, most previous works produce the restoration results purely relying on a single degraded observation. It is worth noting that the degradation process generally is highly ill-posed. By learning a direct mapping from degraded observation, the restoration result inclines to be over-smoothing and cannot faithfully retain fine and identity-aware facial details.

In this paper, we study the problem of guided blind face restoration by incorporating the degraded observation and a high-quality guided face image. Without loss of generality, the guided image is assumed to have the same identity with the degraded observation, and is frontal with eyes open. We note that such guided restoration setting is practically feasible in many real world applications. For example, most smartphones support to recognize and group the face images according to their identities[1]. In each group, the high quality face image can thus be exploited to guide the restoration of low quality images. In film restoration, it is also encouraging to use the high quality portrait of an actor to guide the

[1] https://support.apple.com/HT207103

restoration of low-resolution and corrupted face images of the same actor from an old film. For these tasks, further incorporation of guided image not only can ease the difficulty of blind restoration, but also is helpful in faithfully recovering fine and identity-aware facial details.

Guided blind face restoration, however, cannot be addressed well by simply taking the degraded observation and guided image as input to plain convolutional networks (CNNs), due to that the two images generally are of different poses, expressions and lighting conditions. Fig. 1(c) shows the results obtained using the U-Net [1] by only taking degraded observation as input, while Fig. 1(d) shows the results by taking both two images as input. It can be seen that direct incorporation of guided image brings very limited improvement on the restoration result. To tackle this issue, we develop a guided face restoration network (GFRNet) consisting of a warping subnetwork (WarpNet) and a reconstruction subnetwork (RecNet). Here, the WarpNet is firstly deployed to predict a flow field for warping the guided image to obtain the warped guidance, which is required to have the same pose and expression with degraded observation. Then, the RecNet takes both degraded observation and warped guidance as input to produce the final restoration result. To train GFRNet, we adopt the reconstruction learning to constrain the restoration result to be close to the target image (i.e., ground-truth), and further employ the adversarial learning for visually realistic restoration.

Nonetheless, even though the WarpNet can be end-to-end trained with reconstruction and adversarial learning, we empirically find that it cannot converge to the desired solution and fails to align the guided image to the correct pose and expression. Fig. 1(e) gives the results of our GFRNet trained by reconstruction and adversarial learning. One can see that its improvement over U-Net is still limited, especially when the degraded observation and guided images are distinctly different in pose. Moreover, the ground-truth flow field is unavailable, and the target and guided images may be of different lighting conditions, making it infeasible to directly use the target image to guide the WarpNet learning. Instead, we adopt the face alignment method [17] to detect the face landmarks of the target and guided images, and then introduce the landmark loss as well as the total variation (TV) regularizer to train the WarpNet. As in Fig. 1(f), our full GFRNet achieves the favorable visual quality, and is effective in recovering fine facial details. Furthermore, to make the learned GFRNet applicable to blind face restoration, our model is trained on the synthetic data generated by a general degradation model with versatile settings on blur kernel, noise level, downsampling scale factor, and JPEG quality factor.

Extensive experiments are conducted to evaluate the proposed GFRNet for guided blind face restoration. The results show that the WarpNet is effective in aligning the guided image to the desired pose and expression. The proposed GFRNet achieves significant performance gains over the state-of-the-art restoration methods, e.g., SRCNN [18], VDSR [19], SRGAN [20], DCP [21], DeepDeblur [22], DeblurGAN [23], DnCNN [24], MemNet [25], ARCNN [26], CBN [4], WaveletSRNet [9], TDAE [11], SCGAN [10] and MCGAN [10] in terms of both quantitative metrics (i.e., PSNR and SSIM) and visually perceptual quality. Moreover, our GFRNet also performs favorably on real degraded images as shown in Fig. 1(f). To sum up, the main contribution of this work includes:

- The GFRNet architecture for guided blind face restoration, which includes a warping subnetwork (WarpNet) and a reconstruction subnetwork (RecNet).
- The incorporation of landmark loss and TV regularization for training the WarpNet.
- The promising results of GFRNet on both synthetic and real face images.

2 Related Work

Recent years have witnessed the unprecedented success of deep learning in many image restoration tasks such as super-resolution [18–20], denoising [24,25], compression artifact removal [26, 27], compressed sensing [28–30], and deblurring [22, 23, 31, 32]. As to face images, several CNN architectures have been developed for face hallucination [4, 6, 7, 9], and the adversarial learning is also introduced to enhance the visual quality [5, 8]. Most of these methods, however, are suggested for non-blind restoration and are restricted by the specialized tasks. Benefitted from the powerful modeling capability of deep CNNs, recent studies have shown that it is feasible to train a single model for handling multiple instantiations of degradation (e.g., different noise levels) [24, 33]. As for face hallucination, Yu et al. [8, 11] suggest one kind of transformative discriminative networks to super-resolve different unaligned tiny face images. Nevertheless, blind restoration is a more challenging problem and requires to learn a single model for handling all instantiations of one or more degradation types.

Most studies on deep blind restoration are given to blind deblurring, which aims to recover the latent clean image from noisy and blurry observation with unknown degradation parameters. Early learning-based or CNN-based blind deblurring methods [31, 34, 35] usually follow traditional framework which includes a blur kernel estimation stage and a non-blind deblurring stage. With the rapid progress and powerful modeling capability of CNNs, recent studies incline to bypass blur kernel estimation by directly training a deep model to restore clean image from degraded observation [22, 23, 32, 36, 37]. As to blind face restoration, Chrysos and Zafeiriou [12] utilize a modified ResNet architecture to perform face deblurring, while Xu et al. [10] adopt the generative adversarial network (GAN) framework to super-resolve blurry face image. It is worth noting that the success of such kernel-free end-to-end approaches depends on both the modeling capability of CNN and the sufficient sampling on clean images and degradation parameters, making it difficult to design and train. Moreover, the highly ill-posed degradation further increases the difficulty of recovering the correct fine details only from degraded observation [38]. In this work, we elaborately tackle this issue by incorporating a high quality guided image and designing appropriate network architecture and learning objective.

Several learning-based and CNN-based approaches are also developed for color-guided depth image enhancement [39–41], where the structural interdependency between intensity and depth image is modeled and exploited to reconstruct high quality depth image. For guided depth image enhancement, Hui et al. [40] present a CNN model to learn multi-scale guidance, while Gu et al. [41] incorporate weighted analysis representation and truncated inference for dynamic guidance learning. For general guided filtering, Li et al. [39] construct CNN-based joint filters to transfer structural details from guided image to reconstructed image. However, these approaches assume that the guided image is spatially well aligned with the degraded observation. Due to that the guided image and degraded observation usually are different in pose and expression, such assumption generally does not hold true for guided face restoration. To address this issue, a WarpNet is introduced in our GFRNet to learn a flow field for warping the guided image to the desired pose and expression.

Recently, spatial transformer networks (STNs) are suggested to learn a spatial mapping for warping an image [42], and appearance flow networks (AFNs) are presented to predict a dense flow field to move pixels [43,44]. Deep dense flow networks have been applied to view synthesis [43, 45], gaze manipulation [44], expression editing [46], and video frame synthesis [47]. In these approaches, the target image is required to have the similar lighting condition with the input image to be warped, and the dense flow networks can thus be trained via reconstruction learning. However, in our guided face restoration task, the guided image and the target image usually are of different lighting conditions, making it less effective to train the flow network via reconstruction learning. Moreover, the ground-truth dense flow field is not available, further increasing the difficulty to train WarpNet. To tackle this issue, we use the face alignment method [17] to extract the face landmarks of guided and target images. Then, the landmark loss and TV regularization are incorporated to facilitate the WarpNet training.

3 Proposed Method

This section presents our GFRNet to recover high quality face image from degraded observation with unknown degradation. Given a degraded observation I^d and a guided image I^g, our GFRNet model produces the restoration result $\hat{I} = \mathcal{F}(I^d, I^g)$ to approximate the ground-truth target image I. Without loss of generality, I^g and I are of the same identity and image size 256×256. Moreover, to provide richer guidance information, I^g is assumed to be of high quality, frontal, non-occluded with eyes open. Nonetheless, we empirically find that our GFRNet is robust when the assumption is violated. For simplicity, we also assume I^d also has the same size with I^g. When such assumption does not hold, e.g., in face hallucination, we simply apply the bicubic scheme to upsample I^d to the size of I^g before inputting it to the GFRNet.

In the following, we first describe the GFRNet model as well as the network architecture. Then, a general degradation model is introduced to generate synthetic training data. Finally, we present the model objective of our GFRNet.

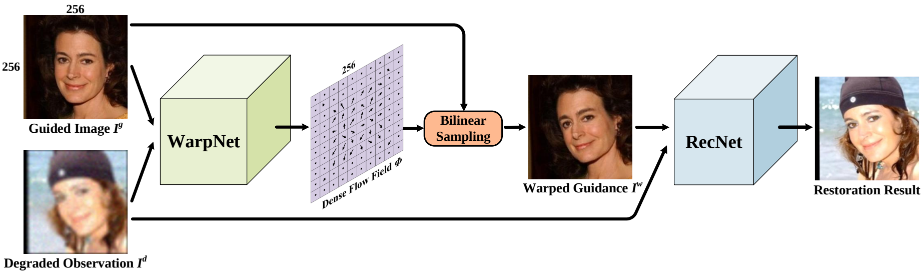


Fig. 2. Overview of our GFRNet. The WarpNet takes the degraded observation I^d and guided image I^g as input to predict the dense flow field Φ, which is adopted to deform I^g to the warped guidance I^w. I^w is expected to be spatially well aligned with ground-truth I. Thus the RecNet takes I^w and I^d as input to produce the restoration result $\hat{I}$.

3.1 Guided Face Restoration Network

The degraded observation I^d and guided image I^g usually vary in pose and expression. Directly taking I^d and I^g as input to plain CNNs generally cannot achieve much performance gains over taking only I^d as input (See Fig. 1(c)(d)). To address this issue, the proposed GFRNet consists of two subnetworks: (i) the warping subnetwork (WarpNet) and (ii) reconstruction subnetwork (RecNet).

Fig. 2 illustrates the overall architecture of our GFRNet. The WarpNet takes I^d and I^g as input to predict the flow field for warping guided image,

$$\Phi = \mathcal{F}_w(I^d, I^g; \Theta_w), \tag{1}$$

where Θ_w denotes the WarpNet model parameters. With Φ, the output pixel value of the warped guidance I^w at location (i, j) is given by

$$I^w_{i,j} = \sum_{(h,w)\in\mathcal{N}} I^g_{h,w} \max(0, 1 - |\Phi^y_{i,j} - h|) \max(0, 1 - |\Phi^x_{i,j} - w|), \tag{2}$$

where $\Phi^x_{i,j}$ and $\Phi^y_{i,j}$ denote the predicted x and y coordinates for the pixel $I^w_{i,j}$, respectively. $\mathcal{N}$ stands for the 4-pixel neighbors of $(\Phi^x_{i,j}, \Phi^y_{i,j})$. From Eqn. (2), we note that I^w is subdifferentiable to Φ [42]. Thus, the WarpNet can be end-to-end trained by minimizing the losses defined either on I^w or on Φ.

The predicted warping guidance I^w is expected to have the same pose and expression with the ground-truth I. Thus, the RecNet takes I^d and I^w as input to produce the final restoration result,

$$\hat{I} = \mathcal{F}_r(I^d, I^w; \Theta_r), \tag{3}$$

where Θ_r denotes the RecNet model parameters.

Warping Subnetwork (WarpNet). The WarpNet adopts the encoder-decoder structure and is comprised of two major components:

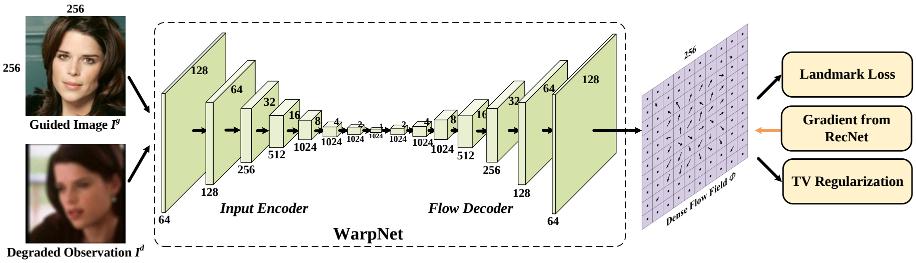


Fig. 3. Architecture of our WarpNet. It takes the degraded observation I^d and guided image I^g as input to predict the dense flow field Φ, which is adopted to deform I^g to the warped guidance I^w. I^w is expected to be spatially well aligned with ground-truth I. Landmark loss, TV regularization as well as gradient from RecNet are deployed to facilitate the learning of WarpNet.

- The input encoder extracts feature representation from I^d and I^g, consisting of eight convolution layers and each one with size 4×4 and stride 2.
- The flow decoder predicts the dense flow field for warping I^g to the desired pose and expression, consisting of eight deconvolution layers.

Except the first layer in encoder and the last layer in decoder, all the other layers adopt the convolution-BatchNorm-ReLU form. The detailed structure of WarpNet is shown in Fig 3.

Finally, some explanations are given to the design of WarpNet. (i) Instead of the U-Net architecture, we adopt the standard encoder-decoder structure by removing the skip connections. It is worth noting that the input to encoder is two color images I^d and I^g while the output of decoder is a dense flow field Φ. Due to the heterogeneity of the input and output, it is unappropriate to concatenate the encoder features to the corresponding decoder features as in U-Net. (ii) It is also improper to directly output the warped guidance instead of the flow field. I^w is of different pose and expression with I^g, making the U-Net architecture still suffer from the heterogeneity issue. Due to the effect of the bottleneck (i.e., the fully connected layer), the encoder-decoder structure inclines to produce over-smoothing I^w. Instead of directly predicting I^w, predicting the dense flow field Φ usually results in realistic facial image with fine details.

Reconstruction Subnetwork (RecNet). For the RecNet, the input (I^d and I^w) are of the same pose and expression with the output ($\hat{I}$), and thus the U-Net can be adopted to produce the final restoration result $\hat{I}$. The RecNet also includes two components, i.e., an encoder and a decoder. The encoder and decoder of RecNet are of the same structure with those adopted in WarpNet. To circumvent the information loss, the i-th layer is concatenated to the $(L-i)$-th layer via skip connections (L is the depth of the U-Net), which has been demonstrated to benefit the rich and fine details of the generated image [48]. The detailed structure of RecNet is shown in Fig 4.

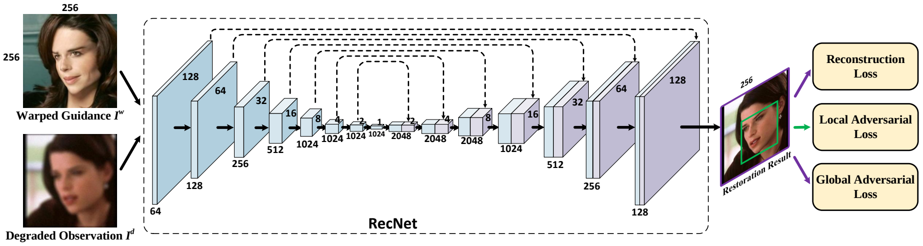


Fig. 4. Architecture of our RecNet. It takes I^w and I^d as input to produce the restoration result $\hat{I}$. Reconstruction loss and global adversarial loss are adopted across entire image (labeled in **purple**), while local adversarial loss is adopted across face region (labeled in **green**).

3.2 Degradation Model and Synthetic Training Data

To train our GFRNet, a degradation model is required to generate realistic degraded images. We note that real low quality images can be the results of either defocus, long-distance sensing, noise, compression, or their combinations. Thus, we adopt a general degradation model to generate degraded image $I^{d,s}$,

$$I^{d,s} = ((I \otimes \mathbf{k}_\varrho)\downarrow_s + \mathbf{n}_\sigma)_{JPEG_q}, \tag{4}$$

where $\otimes$ denotes the convolution operator. $\mathbf{k}_\varrho$ stands for the Gaussian blur kernel with the standard deviation ϱ. $\downarrow_s$ denotes the downsampling operator with scale factor s. $\mathbf{n}_\sigma$ denotes the additive white Gaussian noise (AWGN) with the noise level σ. $(\cdot)_{JPEG_q}$ denotes the JPEG compression operator with quality factor q.

In our general degradation model, $(I \otimes \mathbf{k}_\varrho)\downarrow_s + \mathbf{n}_\sigma$ characterizes the degradation caused by long-distance acquisition, while $(\cdot)_{JPEG_q}$ depicts the degradation caused by JPEG compression. We also note that Xu et al. [10] adopt the degradation model $(I \otimes \mathbf{k}_\varrho + \mathbf{n}_\sigma)\downarrow_s$. However, to better simulate the long-distance image acquisition, it is more appropriate to add the AWGN on the downsampled image. When $s \neq 1$, the degraded image $I^{d,s}$ is of different size with the ground-truth I. So we use bicubic interpolation to upsample $I^{d,s}$ with scale factor s, and then take $I^d = (I^{d,s})\uparrow_s$ and I^g as input to our GFRNet.

In the following, we explain the parameter settings for these operations:

- **Blur kernel.** In this work, only the isotropic Gaussian blur kernel $\mathbf{k}_\varrho$ is considered to model the defocus effect. We sample the standard deviation of Gaussian blur kernel from the set $\varrho \in \{0, 1:0.1:3\}$.
- **Downsampler.** We adopt the bicubic downsampler as [4, 6, 7, 9, 10]. The scale factor s is sampled from the set $s \in \{1:0.1:8\}$.
- **Noise.** As for the noise level σ, we adopt the set $\sigma \in \{0:1:7\}$ [10].
- **JPEG compression.** For economic storage and communication, JPEG compression with quality factor q is further operated on the degraded image, and we sample q from the set $q \in \{0, 10:1:40\}$. When $q = 0$, the image is only losslessly compressed.

By including $\varrho = 0$, $s = 1$, $\sigma = 0$ and $q = 0$ in the set of degradation parameters, the general degradation model can simulate the effect of either the defocus, long-distance acquisition, noising, compression or their versatile combinations.

Given a ground-truth image I_i together with the guided image I_i^g, we can first sample ϱ_i, s_i, σ_i and q_i from the parameter set, and then use the degradation model to generate a degraded observation I_i^d. Furthermore, the face alignment method [17] is adopted to extract the landmarks $\{(x_j^{I_i}, y_j^{I_i})|_{j=1}^{68}\}$ for I_i and $\{(x_j^{I_i^g}, y_j^{I_i^g})|_{j=1}^{68}\}$ for I_i^g. Therefore, we define the synthetic training set as $\mathcal{X} = \{(I_i, I_i^g, I_i^d, \{(x_j^{I_i}, y_j^{I_i})|_{j=1}^{68}\}, \{(x_j^{I_i^g}, y_j^{I_i^g})|_{j=1}^{68}\})|_{i=1}^{N}\}$, where N denotes the number of samples.

3.3 Model Objective

Losses on Restoration Result $\hat{I}$. To train our GFDNet, we define the reconstruction loss on the restoration result $\hat{I}$, and the adversarial loss is further incorporated on $\hat{I}$ to improve the visual perception quality.

Reconstruction loss. The reconstruct loss is used to constrain the restoration result $\hat{I}$ to be close to the ground-truth I, which includes two terms. The first term is the ℓ_2 loss defined as the squared Euclidean distance between $\hat{I}$ and I, i.e., $\ell_r^0(I, \hat{I}) = \|I - \hat{I}\|^2$. Due to the inherent irreversibility of image restoration, only the ℓ_2 loss inclines to cause over-smoothing result. Following [49], we define the second term as the perceptual loss on the pre-trained VGG-Face [50]. Denote by ψ the VGG-Face network, $\psi_l(I)$ the feature map of the l-th convolution layer. The perceptual loss on the l-th layer (i.e., Conv-4 in this work) is defined as

$$\ell_p^{\psi,l}(I, \hat{I}) = \frac{1}{C_l H_l W_l} \left\| \psi_l(\hat{I}) - \psi_l(I) \right\|_2^2 \tag{5}$$

where C_l, H_l and W_l denote the channel numbers, height and width of the feature map, respectively. Finally, we define the reconstruction loss as

$$\mathcal{L}_r(I, \hat{I}) = \lambda_{r,0} \ell_r^0(I, \hat{I}) + \lambda_{r,l} \ell_p^{\psi,l}(I, \hat{I}), \tag{6}$$

where $\lambda_{r,0}$ and $\lambda_{r,l}$ are the tradeoff parameters for the ℓ_2 and the perceptual losses, respectively.

Adversarial Loss. Following [51,52], both global and local adversarial losses are deployed to further improve the perceptual quality of the restoration result. Let $p_{data}(I)$ be the distribution of ground-truth image, $p_d(I^d)$ be the distribution of degraded observation. Using the global adversarial loss [53] as an example, the adversarial loss can be formulated as,

$$\ell_{a,g} = \min_{\Theta} \max_{D} \mathbb{E}_{I \sim p_{data}(I)}[\log D(I)] + \mathbb{E}_{I^d \sim p_d(I^d)}[\log(1 - D(\mathcal{F}(I^d, I^g; \Theta)))], \tag{7}$$

where $D(I)$ denotes the global discriminator which predicts the possibility that I is from the distribution $p_{data}(I)$. $\mathcal{F}(I^d, I^g; \Theta)$ denotes the restoration result by our GFRNet with the model parameters $\Theta = (\Theta_w, \Theta_r)$.

Following the conditional GAN [48], the discriminator has the same architecture with pix2pix [48], and takes the degraded observation, guided image and restoration result as the input. The network is trained in an adversarial manner, where our GFRNet is updated by minimizing the loss $\ell_{a,g}$ while the discriminator is updated by maximizing $\ell_{a,g}$. To improve the training stability, we adopt the improved GAN [54], and replace the labels 0/1 with the smoothed 0/0.9 to reduce the vulnerability to adversarial examples. The local adversarial loss $\ell_{a,l}$ adopts the same settings with the global one but its discriminator is defined only on the minimal bounding box enclosing all facial landmarks. To sum up, the overall adversarial loss is defined as

$$\mathcal{L}_a = \lambda_{a,g}\ell_{a,g} + \lambda_{a,l}\ell_{a,l}. \tag{8}$$

where $\lambda_{a,g}$ and $\lambda_{a,l}$ are the tradeoff parameters for the global and local adversarial losses, respectively.

Losses on Flow Field $\boldsymbol{\Phi}$. Although the WarpNet can be end-to-end trained based on the reconstruction and adversarial losses, it cannot be learned to correctly align I^w with I in terms of pose and expression (see Fig. 13). In [44,46], the appearance flow network is trained by minimizing the MSE loss between the output and the ground-truth of the warped image. But for guided face restoration, I generally has different illumination with I^g, and cannot serve as the ground-truth of the warped image. To circumvent this issue, we present the landmark loss as well as the TV regularization to facilitate the learning of WarpNet.

Landmark loss. Using the face alignment method TCDCN [17], we detect the 68 landmarks $\{(x_j^{I^g}, y_j^{I^g})|_{j=1}^{68}\}$ for I^g and $\{(x_j^I, y_j^I)|_{j=1}^{68}\}$ for I. In order to align I^w and I, it is natural to require that the landmarks of I^w are close to those of I, i.e., $\Phi^x(x_j^I, y_j^I) \approx x_j^{I^g}$ and $\Phi^y(x_j^I, y_j^I) \approx y_j^{I^g}$. Thus, the landmark loss is defined as

$$\ell_{lm} = \sum_i (\Phi_x(x_i^I, y_i^I) - x_i^{I^g})^2 + (\Phi_y(x_i^I, y_i^I) - y_i^{I^g})^2. \tag{9}$$

In our implementation, all the coordinates (including x, y, Φ_x and Φ_y) are normalized to the range $[-1, 1]$.

TV regularization. The landmark loss can only be imposed on the locations of the 68 landmarks. For better learning WarpNet, we further take the TV regularization into account to require that the flow field should be spatially smooth. Given the 2D dense flow field (f_x, f_y), the TV regularizer is defined as

$$\ell_{TV} = \|\nabla_x \Phi_x\|^2 + \|\nabla_y \Phi_x\|^2 + \|\nabla_x \Phi_y\|^2 + \|\nabla_y \Phi_y\|^2, \tag{10}$$

where ∇_x (∇_y) denotes the gradient operator along the x (y) coordinate.

Combining landmark loss with TV regularizer, we define the flow loss as

$$\mathcal{L}_{flow} = \lambda_{lm}\ell_{lm} + \lambda_{TV}\ell_{TV}, \tag{11}$$

where λ_{lm} and λ_{TV} denote the tradeoff parameters for landmark loss and TV regularizer, respectively.

Overall Objective. Finally, we combine the reconstruction loss, adversarial loss, and flow loss to give the overall objective,

$$\mathcal{L} = \mathcal{L}_r + \mathcal{L}_a + \mathcal{L}_{flow}. \tag{12}$$

4 Experimental Results

Extensive experiments are conducted to assess our GFRNet for guided blind face restoration. Peak Signal-to-Noise Ratio (PSNR) and structural similarity (SSIM) indices are adopted for quantitative evaluation with the related state-of-the-arts (including image super-resolution, deblurring, denoising, compression artifact removal and face hallucination). As for qualitative evaluation, we illustrate the results by our GFRNet and the competing methods. Results on real low quality images are also given to evaluate the generalization ability of our GFRNet. Code is available at: `https://github.com/csxmli2016/GFRNet`.

4.1 Dataset

We adopt the CASIA-WebFace [55] and VggFace2 [56] datasets to constitute our training and test sets. The WebFace contains 10,575 identities and each has about 46 images with the size 256 × 256. The VggFace2 contains 9,131 identities (8,631 for training and 500 for testing) and each has an average of 362 images with different sizes. The images in the two datasets are collected in the wild and cover a large range of pose, age, illumination and expression. For each identity, at most five high quality images are selected, in which a frontal image with eyes open is chosen as the guided image and the others are used as the ground-truth to generate degraded observations. By this way, we build our training set of 20,273 pairs of ground-truth and guided images from the VggFace2 training set. Our test set includes two subsets: (i) 1,005 pairs from the VggFace2 test set, and (ii) 1,455 pairs from WebFace. The images whose identities have appeared in our training set are excluded from the test set. Furthermore, low quality images are also excluded in training and testing, which include: (i) low-resolution images, (ii) images with large occlusion, (iii) cartoon images, and (iv) images with obvious artifacts. The face region of each image in VGGFace2 is cropped and resized to 256 × 256 based on the bounding box detected by MTCNN [57]. All training and test images are not aligned to keep their original pose and expression. Facial landmarks of the ground-truth and guided images are detected by TCDCN [17] and are only used in training.

4.2 Training Details and Parameter Setting

Our model is trained using the Adam algorithm [58] with the learning rate of 2×10^{-4}, 2×10^{-5}, 2×10^{-6} and $\beta_1 = 0.5$. In each learning rate, the model is trained until the reconstruction loss becomes non-decreasing. Then a smaller learning rate is adopted to further fine-tune the model. The tradeoff parameters

Table 1. Quantitative results on two test subsets. Numbers in the parentheses indicate SSIM and the remaining represents PSNR (dB). The best results are highlighted in **red** and second best ones except our GFRNet variants are highlighted in blue.

Methods		VggFace2 [56]				WebFace [55]			
		4×		8×		4×		8×	
SR	SRCNN [18]	24.57	(.842)	22.30	(.802)	26.11	(.872)	23.50	(.842)
	VDSR [19]	25.36	(.858)	22.50	(.807)	26.60	(.884)	23.65	(.847)
	SRGAN [20]	25.85	(.911)	23.01	(.874)	27.65	(.941)	24.49	(.913)
	MSRGAN	26.55	(.906)	23.45	(.862)	28.10	(.934)	24.92	(.908)
Deblurring	DCP [21]	24.42	(.894)	21.54	(.848)	24.97	(.895)	23.05	(.887)
	DeepDeblur [22]	26.31	(.917)	22.97	(.873)	28.13	(.934)	24.63	(.910)
	DeblurGAN [23]	24.65	(.889)	22.06	(.846)	24.63	(.910)	23.38	(.896)
	MDeblurGAN	25.32	(.918)	22.46	(.867)	29.41	(.952)	23.49	(.900)
Denoising	DnCNN [24]	26.73	(.920)	23.29	(.877)	28.35	(.933)	24.75	(.912)
	MemNet [25]	26.85	(.923)	23.31	(.877)	28.57	(.934)	24.77	(.909)
	MDnCNN	27.05	(.925)	23.33	(.879)	29.40	(.942)	24.84	(.912)
AR	ARCNN [26]	22.05	(.863)	20.84	(.827)	23.39	(.876)	20.47	(.858)
	MARCNN	25.43	(.923)	23.16	(.876)	28.40	(.938)	25.15	(.914)
Non-blind FH	CBN [4]	24.52	(.867)	21.84	(.817)	25.43	(.899)	23.10	(.852)
	WaveletSRNet [9]	25.66	(.909)	20.87	(.831)	27.10	(.937)	21.63	(.869)
	TDAE [11]	-	(-)	20.19	(.729)	-	(-)	20.24	(.741)
Blind FH	SCGAN [10]	25.16	(.905)	-	-	26.37	(.923)	-	-
	MCGAN [10]	25.26	(.912)	-	-	26.35	(.931)	-	-
Ours	Ours($-WG$)	25.97	(.915)	22.91	(.838)	28.73	(.928)	24.76	(.884)
	Ours($-WG2$)	27.20	(.932)	23.22	(.863)	29.45	(.945)	25.93	(.914)
	Ours($-W$)	26.03	(.923)	23.29	(.843)	29.66	(.934)	25.20	(.897)
	Ours($-W2$)	27.25	(.933)	23.24	(.864)	29.73	(.948)	25.95	(.917)
	Ours($-F$)	26.61	(.927)	23.17	(.863)	31.43	(.920)	26.00	(.922)
	Ours(R)	27.90	(.943)	24.05	(.890)	31.46	(.962)	26.88	(.922)
	Ours($Full$)	**28.55**	**(.947)**	**24.10**	**(.898)**	**32.31**	**(.973)**	**27.21**	**(.935)**

are set as $\lambda_{r,0} = 100$, $\lambda_{r,l} = 0.001$, $\lambda_{a,g} = 1$, $\lambda_{a,l} = 0.5$, $\lambda_{lm} = 10$, and $\lambda_{TV} = 1$. We first pre-train the WarpNet for 5 epochs by minimizing the flow loss $\mathcal{L}_{flow}$, and then both WarpNet and RecNet are end-to-end trained by using the objective $\mathcal{L}$. The batch size is 1 and the training is stopped after 100 epochs. Data augmentation such as flipping is also adopted during training.

4.3 Results on Synthetic Images

Table 1 lists the PSNR and SSIM results on the two test subsets, where our GFRNet achieves significant performance gains over all the competing methods. Using the 4× SR on WebFace as an example, in terms of PSNR, our GFRNet outperforms the SR and blind deblurring methods by more than 4 dB, the denoising methods by more than 3.5 dB, the compression artifact removal (AR) methods by more than 8 dB, the non-blind and blind FH methods by more than 5 dB. To the best of our knowledge, guided blind face restoration remains an uninvestigated issue in literature. Thus, we compare our GFRNet with several relevant state-of-the-arts, including three non-blind image super-resolution (SR) methods (SRCNN [18], VDSR [19], SRGAN [20]), three blind deblurring methods (DCP [21], DeepDeblur [22], DeblurGAN [23]), two denoising methods (DnCNN [24], MemNet [25]), one compression artifact removal method (ARCNN [26]), three non-blind face hallucination (FH) methods (CBN [4], Wavelet-

SRNet [9], TDAE [11]), and two blind FH methods (SCGAN [10], MCGAN [10]). To keep consistent with the SR and FH methods, only two scale factors, i.e., 4 and 8, are considered for the test images. As for non-SR methods, we take the bicubic upsampling result as the input to the model. SRCNN [18] and VDSR [19] are only trained to perform 2×, 3× and 4× SR. To handle 8× SR, we adopt the strategy in [59] by applying the 2× model to the result produced by the 4× model. For SCGAN [10] and MCGAN [10], only the 4× models are available. For TDAE [11], only the 8× model is available.

Quantitative evaluation. It is worth noting that the promising performance of our GFRNet cannot be solely attributed to the use of our training data and the simple incorporation of guided image. To illustrate this point, we retrain four competing methods (i.e., SRGAN, DeblurGAN, DnCNN, and ARCNN) by using our training data and taking both degraded observation and guided image as input. For the sake of distinction, the retrained models are represented as MSRGAN, MDeblurGAN, MDnCNN, MARCNN. From Table 1, the retrained models do achieve better PSNR and SSIM results than the original ones, but still perform inferior to our GFRNet with a large margin, especially on WebFace. Therefore, the performance gains over the retrained models should be explained by the network architecture and model objective of our GFRNet.

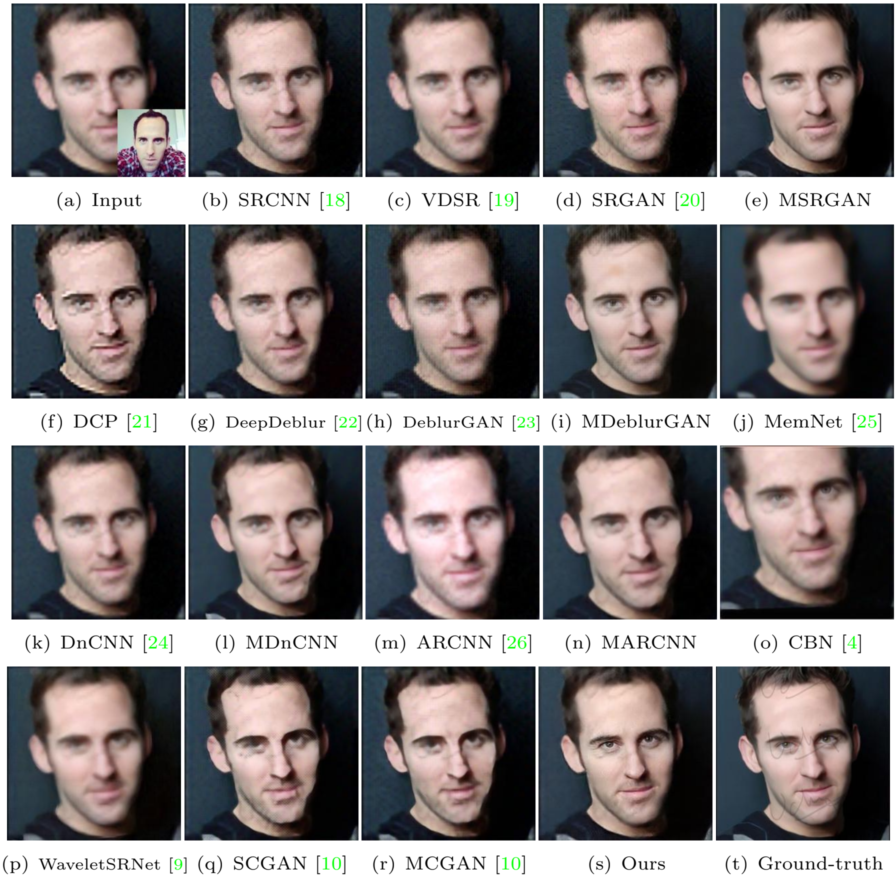
Qualitative evaluation. In Figs. 5 and 6, we compare results of all the competing methods in 4× SR and 8× SR. For better comparison, we select three competing methods with top quantitative performance, and compare their results with those by our GFRNet shown in Figs. 7 and 8. It is obvious that our GFRNet is more effective in restoring fine details while suppressing visual artifacts. In comparison with the competing methods, the results by GFRNet are visually photo-realistic and can correctly recover more fine and identity-aware details especially in eyes, nose, and mouth regions.

4.4 Results on Real Low Quality Images

Fig. 9 shows the results on real low quality images by all the competing methods. As for pose problem, Fig. 10 shows more restoration results of our GFRNet compared with the top-3 performance methods on real low quality images with different poses. One can see that our GFRNet can also show great robustness in restoring facial images with different poses. The real images are selected from VGGFace2 with the resolution lower than 60 × 60. Even the degradation is unknown, our method yields visually realistic and pleasing results in face region with more fine details, while the competing methods can only achieve moderate improvement on visual quality.

4.5 Ablation Studies

Three groups of ablative experiments are conducted to assess the components of our GFRNet. First, we consider five variants of our GFRNet: (i) Ours($Full$):



(a) Input (b) SRCNN [18] (c) VDSR [19] (d) SRGAN [20] (e) MSRGAN

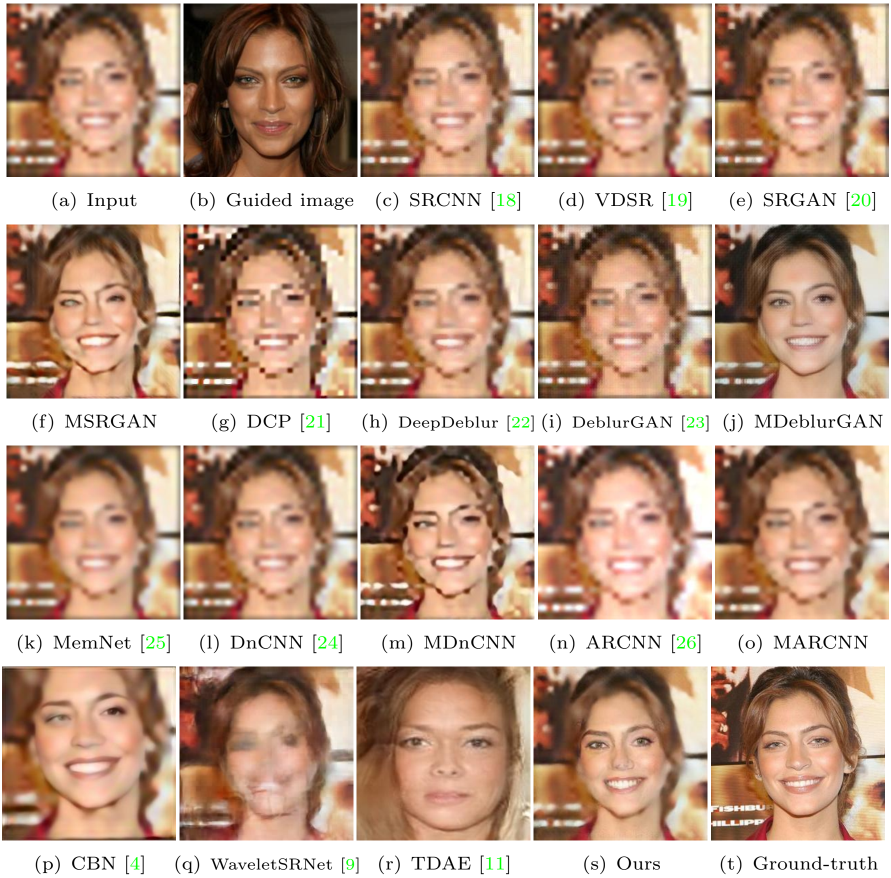
(f) DCP [21] (g) DeepDeblur [22] (h) DeblurGAN [23] (i) MDeblurGAN (j) MemNet [25]

(k) DnCNN [24] (l) MDnCNN (m) ARCNN [26] (n) MARCNN (o) CBN [4]

(p) WaveletSRNet [9] (q) SCGAN [10] (r) MCGAN [10] (s) Ours (t) Ground-truth

Fig. 5. The 4× SR results compared with all the competing methods.

the full GFRNet, (ii) Ours($-F$): GFRNet by removing the flow loss $\mathcal{L}_{flow}$, (iii) Ours($-W$): GFRNet by removing WarpNet (RecNet takes both I^d and I^g as input), (iv) Ours($-WG$): GFRNet by removing WarpNet and guided image (RecNet only takes I^d as input), and (v) Ours(R): GFRNet by using a random I^g with different identity to I^d. Table 1 also lists the PSNR and SSIM results of these variants, and we have the following observations. (i) All the three components, i.e., guided image, WarpNet and flow loss, contribute to the performance improvement. (ii) GFRNet cannot be well trained without the help of flow loss. As a result, although Ours($-F$) outperforms Ours($-W$) in most cases, sometimes Ours($-W$) can perform slightly better than Ours($-F$) by average PSNR, e.g., for 8× SR on VggFace2. (iii) It is worth noting that GFRNet with random guidance (i.e., Ours(R)) achieves the second best results, indicating that GFRNet is robust to the misuse of identity. Figs. 1, 11 and 12 give the restoration results by GFRNet variants. Ours($Full$) can generate much sharper and



(a) Input (b) Guided image (c) SRCNN [18] (d) VDSR [19] (e) SRGAN [20]

(f) MSRGAN (g) DCP [21] (h) DeepDeblur [22] (i) DeblurGAN [23] (j) MDeblurGAN

(k) MemNet [25] (l) DnCNN [24] (m) MDnCNN (n) ARCNN [26] (o) MARCNN

(p) CBN [4] (q) WaveletSRNet [9] (r) TDAE [11] (s) Ours (t) Ground-truth

Fig. 6. The 8× SR results compared with all the competing methods.

richer details, and achieves better perceptual quality than its variants. Moreover, Ours(R) also achieves the second best performance in qualitative results, but it may introduce the fine details of the other identity to the result (e.g., eye regions in Fig. 11(h)). Furthermore, to illustrate the effectiveness of flow loss, Fig. 13 shows the warped guidance by Ours($Full$) and Ours($-F$). Without the help of flow loss, Ours($-F$) cannot converge to stable solution and results in unreasonable warped guidance. In contrast, Ours($Full$) can correctly align guided image to the desired pose and expression, indicating the necessity and effectiveness of flow loss.

Second, it is noted that the parameters of Ours($Full$) are nearly two times of Ours($-W$) and Ours($-WG$). To show that the gain of Ours($Full$) does not come from the increase of parameter number, we include two other variants of GFRNet, i.e., Ours($-W2$) and Ours($-WG2$), by increasing the channels of Ours($-W$) and Ours($-WG$) to 2 times, respectively. From Table 1, in terms of

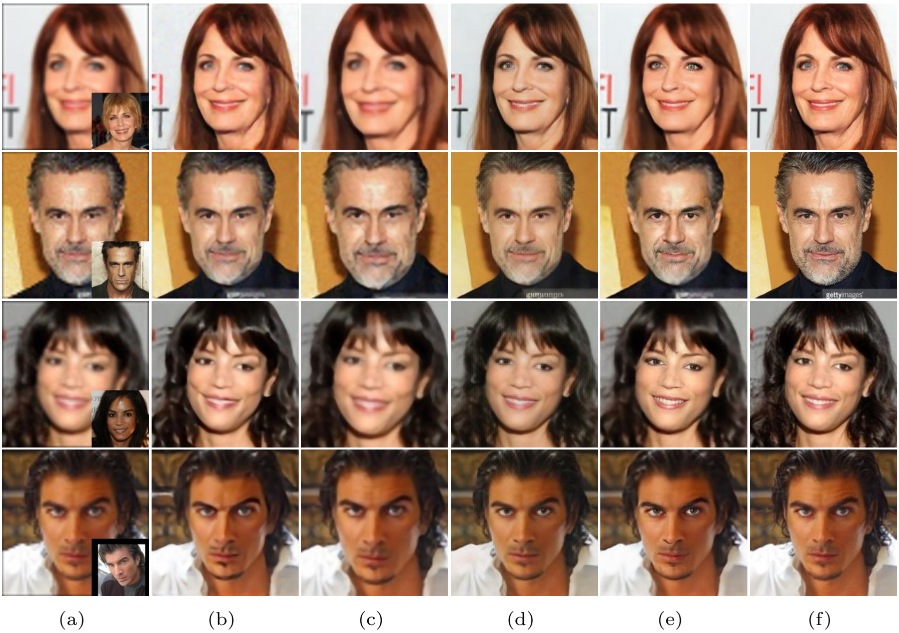


Fig. 7. The 4× SR results: (a) synthetic low quality image (Close-up in right bottom is the guided image), (b) MDnCNN [24], (c) MARCNN [26], (d) MDeblurGAN [23], (e) Ours, and (f) ground-truth.

PSNR, Ours($Full$) also outperforms Ours($-W2$) and Ours($-WG2$) with a large margin. Instead of the increase of model parameters, the performance improvement of Ours($Full$) should be mainly attributed to the incorporation of both WarpNet and flow loss.

Finally, four GFRNet models are trained based on four settings of our general degradation model: (i) only blurring, (ii) blurring+downsampling, (iii) blurring+downsampling+AWGN, (iv) our full degradation model in Eqn. (4). Due to that the four models are trained using different degradation settings, it is unfair to compared them using any synthetic test data. Thus, we test them on a real low quality image in Fig. 14. It can be seen that the results by the settings (i)∼(ii) are still blurry, while the results by the settings (i)∼(iii) contain visual artifacts. In comparison, the model by our full degradation model can produce sharp and clean result while suppressing most artifacts. The results indicate that our full degradation model is effective in simulating real low quality images which usually have unknown and complex degradation.

5 Conclusion

In this paper, we present a guided blind face restoration model, i.e., GFRNet, by taking both the degraded observation and a high-quality guided image from

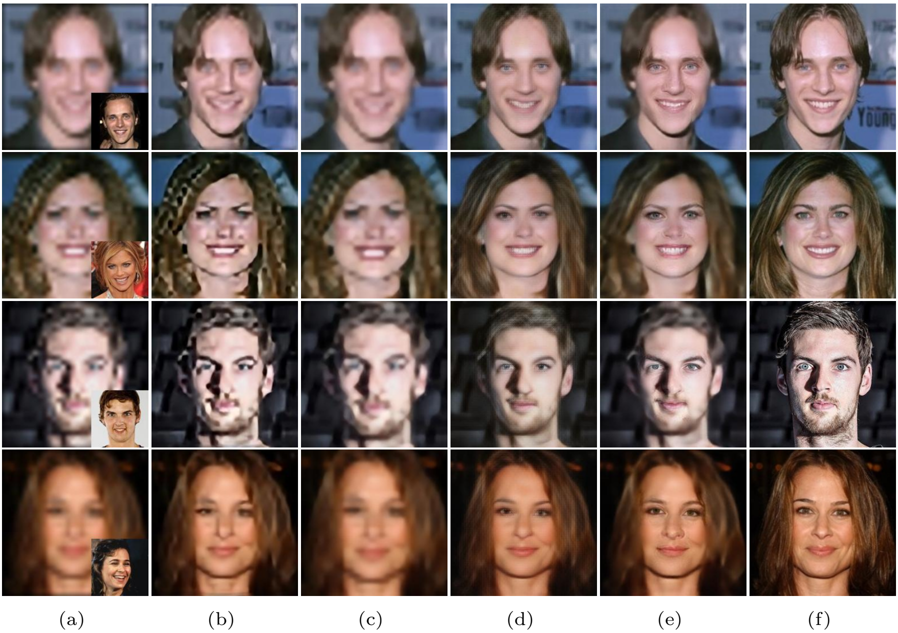
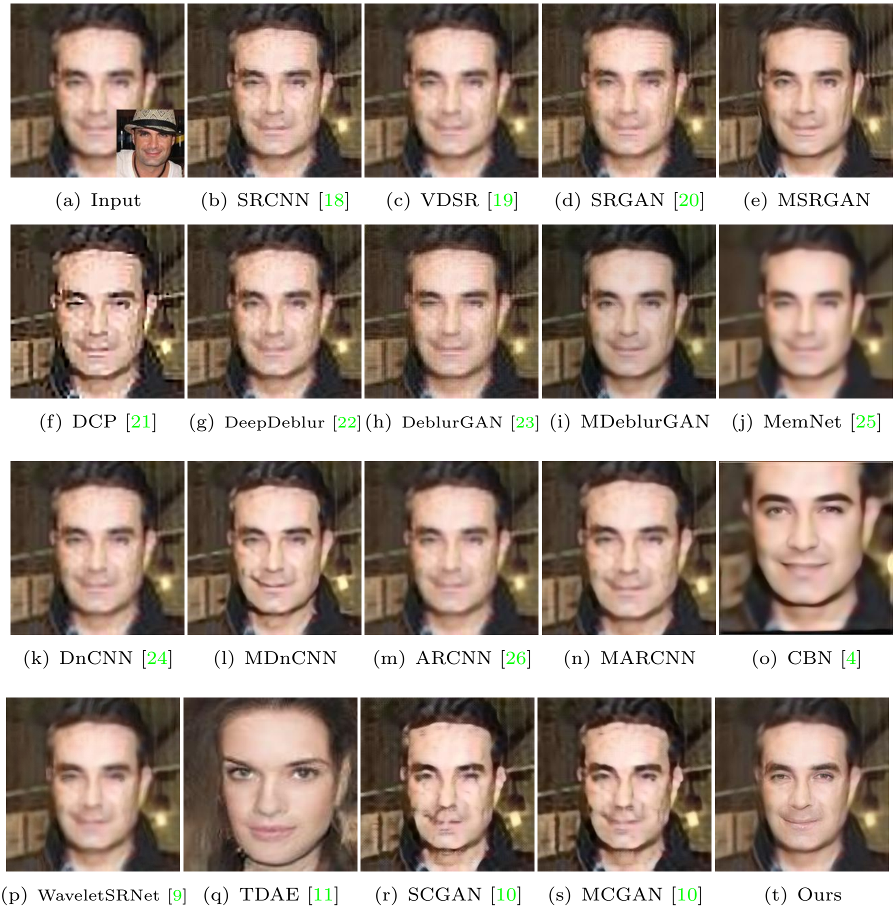


Fig. 8. The 8× SR results: (a) synthetic low quality image (Close-up in right bottom is the guided image), (b) MDnCNN [24], (c) MARCNN [26], (d) MDeblurGAN [23], (e) Ours, and (f) ground-truth.

the same identity as input. Besides the reconstruction subnetwork, our GFRNet also includes a warping subnetwork (WarpNet), and incorporates the landmark loss as well as TV regularizer to align the guided image to the desired pose and expression. To make our GFRNet be applicable to blind restoration, we further introduce a general image degradation model to synthesize realistic low quality face image. Quantitative and qualitative results show that our GFRNet not only performs favorably against the relevant state-of-the-arts but also generates visually pleasing results on real low quality face images.



(a) Input (b) SRCNN [18] (c) VDSR [19] (d) SRGAN [20] (e) MSRGAN

(f) DCP [21] (g) DeepDeblur [22] (h) DeblurGAN [23] (i) MDeblurGAN (j) MemNet [25]

(k) DnCNN [24] (l) MDnCNN (m) ARCNN [26] (n) MARCNN (o) CBN [4]

(p) WaveletSRNet [9] (q) TDAE [11] (r) SCGAN [10] (s) MCGAN [10] (t) Ours

Fig. 9. Restoration on real low quality images compared with all the competing methods.

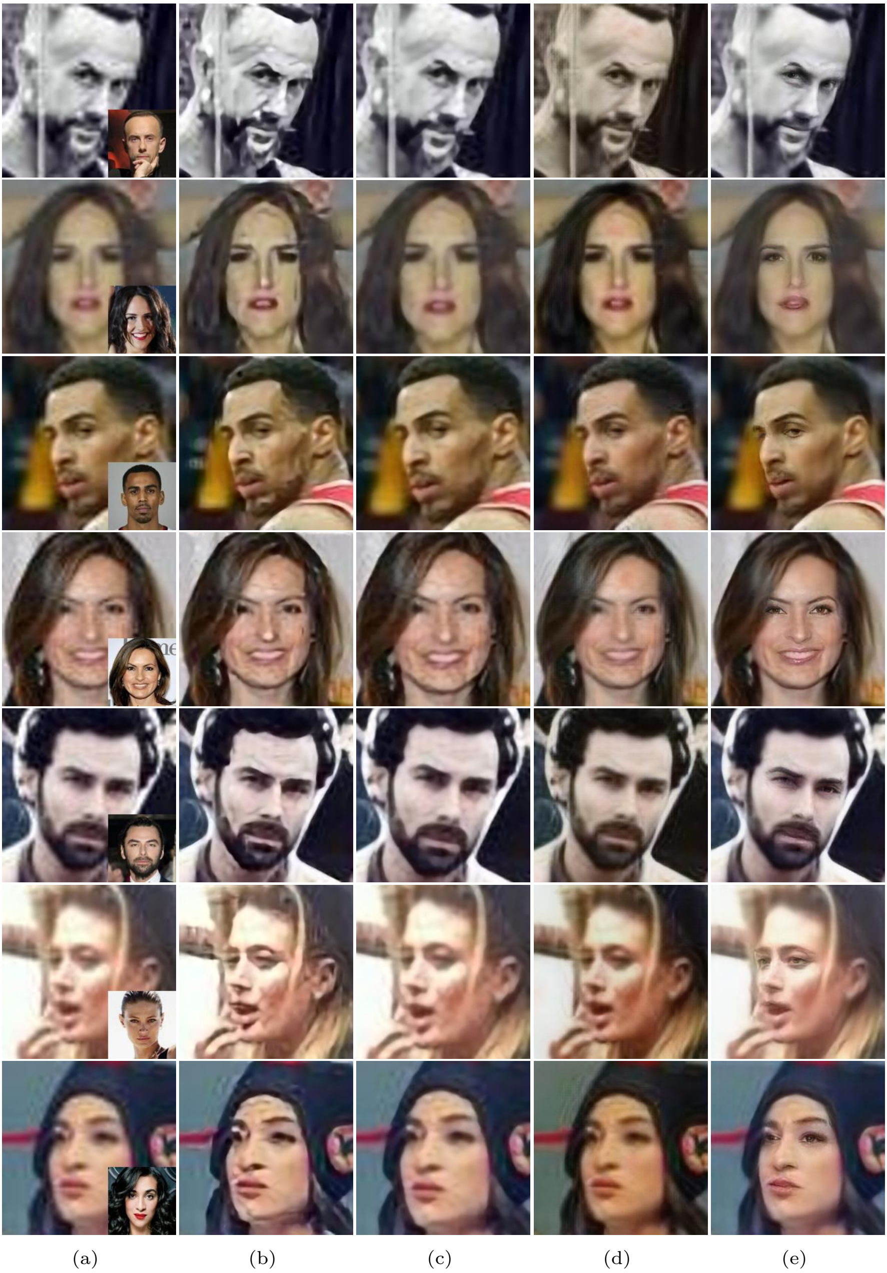


Fig. 10. Restoration results on real low quality images: (a) real low quality images (Close-up in right bottom is the guided image), (b) MDnCNN [24], (c) MARCNN [26], (d) MDeblurGAN [23], and (e) Ours.

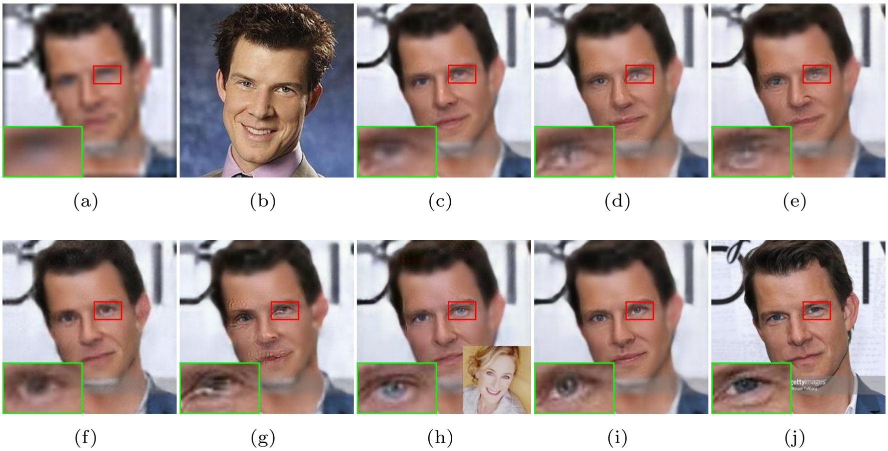


Fig. 11. Restoration results of our GFRNet variants: (a) input, (b) guided image. (c) Ours($-WG$), (d) Ours($-WG2$), (e) Ours($-W$), (f) Ours($-W2$), (g) Ours($-F$), (h) Ours(R) (Close-up in right bottom is the random guided image), (i) Ours($Full$), and (j) ground-truth. Best viewed by zooming in the screen.

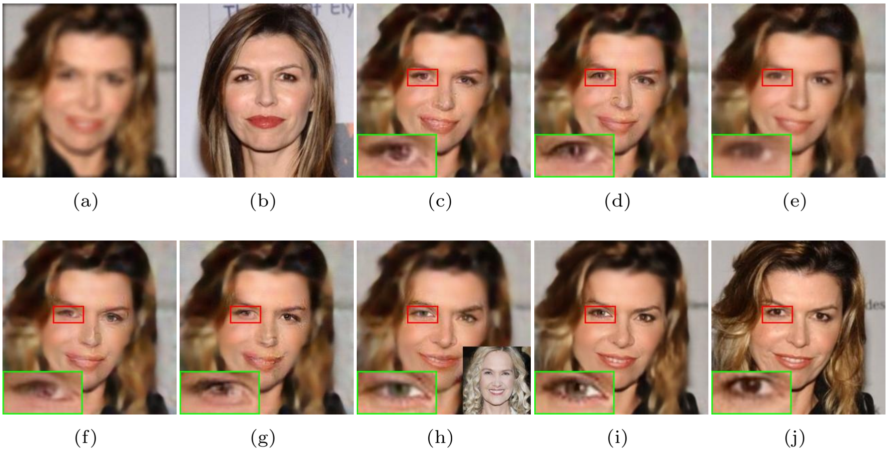


Fig. 12. Restoration results of our GFRNet variants: (a) input, (b) guided image. (c) Ours($-WG$), (d) Ours($-WG2$), (e) Ours($-W$), (f) Ours($-W2$), (g) Ours($-F$), (h) Ours(R) (Close-up in right bottom is the random guided image), (i) Ours($Full$), and (j) ground-truth. Best viewed by zooming in the screen.

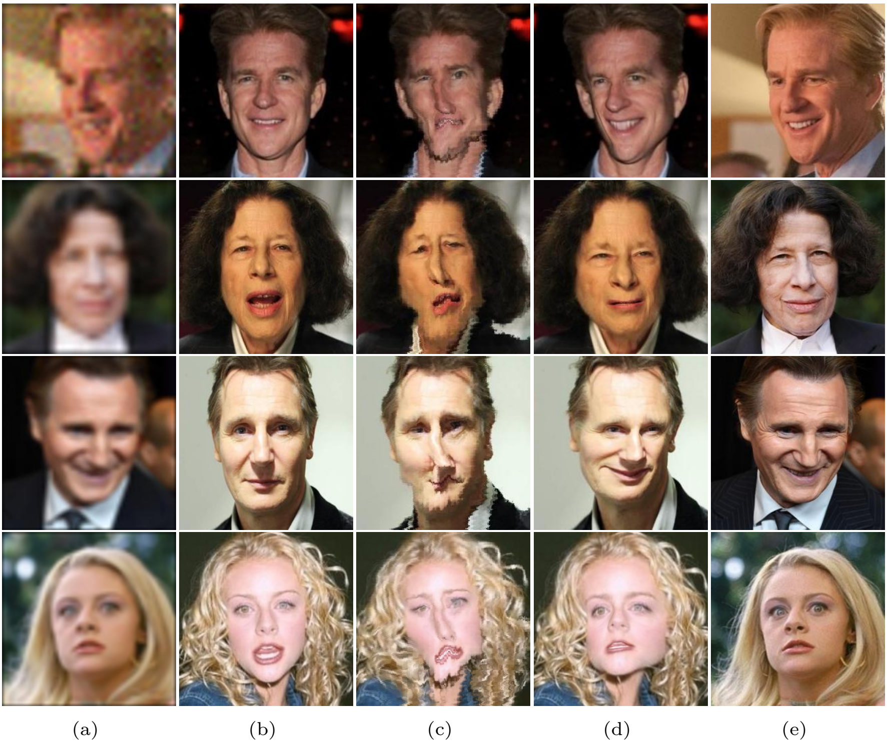


Fig. 13. Warped guidance by Ours($Full$) and Ours($-F$): (a) input, (b) guided image, (c) Ours($-F$), (d) Ours($Full$), and (e) ground-truth.

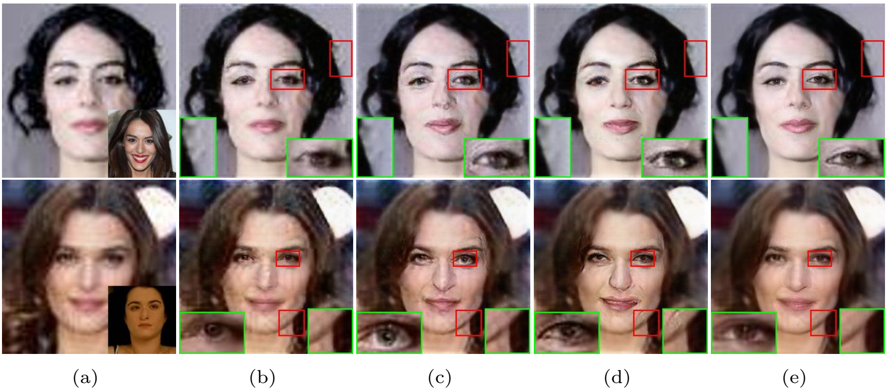


Fig. 14. Results on real low quality images by our GFRNet trained with different degradation settings: (a) real low quality image (Close-up in right bottom is the guided image), (b) only blurring, (c) blurring+downsampling, (d) blurring+downsampling+AWGAN, and (e) full degradation model. Best viewed by zooming in the screen.

References

1. Ronneberger, O., Fischer, P., Brox, T.: U-Net: convolutional networks for biomedical image segmentation. In: International Conference on Medical Image Computing and Computer-Assisted Intervention, Springer (2015) 234–241
2. Baker, S., Kanade, T.: Hallucinating faces. In: IEEE International Conference on Automatic Face and Gesture Recognition, IEEE (2000) 83–88
3. Liu, C., Shum, H.Y., Freeman, W.T.: Face hallucination: Theory and practice. International Journal of Computer Vision **75**(1) (2007) 115–134
4. Zhu, S., Liu, S., Loy, C.C., Tang, X.: Deep cascaded bi-network for face hallucination. In: European Conference on Computer Vision, Springer (2016) 614–630
5. Yu, X., Porikli, F.: Ultra-resolving face images by discriminative generative networks. In: European Conference on Computer Vision, Springer (2016) 318–333
6. Cao, Q., Lin, L., Shi, Y., Liang, X., Li, G.: Attention-aware face hallucination via deep reinforcement learning. In: IEEE Conference on Computer Vision and Pattern Recognition, IEEE (2017) 690–698
7. Chen, Y., Tai, Y., Liu, X., Shen, C., Yang, J.: FSRNet: End-to-end learning face super-resolution with facial priors. arXiv preprint arXiv:1711.10703 (2017)
8. Yu, X., Porikli, F.: Face hallucination with tiny unaligned images by transformative discriminative neural networks. In: AAAI Conference on Artificial Intelligence. (2017) 4327–4333
9. Huang, H., He, R., Sun, Z., Tan, T.: Wavelet-SRNet: A wavelet-based cnn for multi-scale face super resolution. In: IEEE International Conference on Computer Vision, IEEE (2017) 1689–1697
10. Xu, X., Sun, D., Pan, J., Zhang, Y., Pfister, H., Yang, M.H.: Learning to super-resolve blurry face and text images. In: IEEE International Conference on Computer Vision, IEEE (2017) 251–260
11. Yu, X., Porikli, F.: Hallucinating very low-resolution unaligned and noisy face images by transformative discriminative autoencoders. In: IEEE Conference on Computer Vision and Pattern Recognition, IEEE (2017) 3760–3768
12. Chrysos, G.G., Zafeiriou, S.: Deep face deblurring. In: IEEE Conference on Computer Vision and Pattern Recognition Workshops, IEEE (2017) 2015–2024
13. Phillips, P.J., Flynn, P.J., Scruggs, T., Bowyer, K.W., Chang, J., Hoffman, K., Marques, J., Min, J., Worek, W.: Overview of the face recognition grand challenge. In: IEEE Conference on Computer Vision and Pattern Recognition, IEEE (2005) 947–954
14. Andreu, Y., López-Centelles, J., Mollineda, R.A., García-Sevilla, P.: Analysis of the effect of image resolution on automatic face gender classification. In: IEEE International Conference Pattern Recognition, IEEE (2014) 273–278
15. Anwar, S., Porikli, F., Huynh, C.P.: Category-specific object image denoising. IEEE Transactions on Image Processing **26**(11) (2017) 5506–5518
16. Anwar, S., Huynh, C., Porikli, F.: Combined internal and external category-specific image denoising. In: British Machine Vision Conference. (2017)
17. Zhang, Z., Luo, P., Loy, C.C., Tang, X.: Learning deep representation for face alignment with auxiliary attributes. IEEE Transactions on Pattern Analysis and Machine Intelligence **38**(5) (2016) 918–930
18. Dong, C., Loy, C.C., He, K., Tang, X.: Learning a deep convolutional network for image super-resolution. In: European Conference on Computer Vision, Springer (2014) 184–199

19. Kim, J., Lee, J.K., Lee, K.M.: Accurate image super-resolution using very deep convolutional networks. In: IEEE Conference on Computer Vision and Pattern Recognition, IEEE (2016) 1646–1654
20. Ledig, C., Theis, L., Huszar, F., Caballero, J., Cunningham, A., Acosta, A., Aitken, A., Tejani, A., Totz, J., Wang, Z., Shi, W.: Photo-realistic single image super-resolution using a generative adversarial network. In: IEEE Conference on Computer Vision and Pattern Recognition, IEEE (2017) 4681–4690
21. Pan, J., Sun, D., Pfister, H., Yang, M.H.: Blind image deblurring using dark channel prior. In: IEEE Conference on Computer Vision and Pattern Recognition, IEEE (2016) 1628–1636
22. Nah, S., Kim, T.H., Lee, K.M.: Deep multi-scale convolutional neural network for dynamic scene deblurring. In: IEEE Conference on Computer Vision and Pattern Recognition, IEEE (2017) 3883–3891
23. Kupyn, O., Budzan, V., Mykhailych, M., Mishkin, D., Matas, J.: DeblurGAN: Blind motion deblurring using conditional adversarial networks. arXiv preprint arXiv:1711.07064 (2017)
24. Zhang, K., Zuo, W., Chen, Y., Meng, D., Zhang, L.: Beyond a gaussian denoiser: Residual learning of deep cnn for image denoising. IEEE Transactions on Image Processing **26**(7) (2017) 3142–3155
25. Tai, Y., Yang, J., Liu, X., Xu, C.: MemNet: A persistent memory network for image restoration. In: International Conference on Computer Vision, IEEE (2017) 4549–4557
26. Dong, C., Deng, Y., Change Loy, C., Tang, X.: Compression artifacts reduction by a deep convolutional network. In: IEEE International Conference on Computer Vision, IEEE (2015) 576–584
27. Galteri, L., Seidenari, L., Marco, B., Alberto, B.D.: Deep generative adversarial compression artifact removal. In: IEEE International Conference on Computer Vision, IEEE (2017) 4826–4835
28. Kulkarni, K., Lohit, S., Turaga, P., Kerviche, R., Ashok, A.: Reconnet: Non-iterative reconstruction of images from compressively sensed measurements. In: IEEE Conference on Computer Vision and Pattern Recognition, IEEE (2016) 449–458
29. Jin, K.H., McCann, M.T., Froustey, E., Unser, M.: Deep convolutional neural network for inverse problems in imaging. IEEE Transactions on Image Processing **26**(9) (2017) 4509–4522
30. Lucas, A., Iliadis, M., Molina, R., Katsaggelos, A.K.: Using deep neural networks for inverse problems in imaging: Beyond analytical methods. IEEE Signal Processing Magazine **35**(1) (2018) 20–36
31. Chakrabarti, A.: A neural approach to blind motion deblurring. In: European Conference on Computer Vision, Springer (2016) 221–235
32. Nimisha, T., Singh, A.K., Rajagopalan, A.: Blur-invariant deep learning for blind-deblurring. In: IEEE Conference on Computer Vision and Pattern Recognition, IEEE (2017) 4752–4760
33. Mao, X.J., Shen, C., Yang, Y.B.: Image denoising using very deep fully convolutional encoder-decoder networks with symmetric skip connections. arXiv preprint arXiv:1603.09056 (2016)
34. Schuler, C.J., Hirsch, M., Harmeling, S., Schölkopf, B.: Learning to deblur. IEEE Transactions on Pattern Analysis and Machine Intelligence **38**(7) (2016) 1439–1451
35. Xiao, L., Wang, J., Heidrich, W., Hirsch, M.: Learning high-order filters for efficient blind deconvolution of document photographs. In: European Conference on Computer Vision, Springer (2016) 734–749

36. Noroozi, M., Chandramouli, P., Favaro, P.: Motion deblurring in the wild. In: German Conference on Pattern Recognition, Springer (2017) 65–77
37. Hradiš, M., Kotera, J., Zemcík, P., Šroubek, F.: Convolutional neural networks for direct text deblurring. In: British Machine Vision Conference. (2015)
38. Lin, Z., He, J., Tang, X., Tang, C.K.: Limits of learning-based superresolution algorithms. International Journal of Computer Vision **80**(3) (2008) 406–420
39. Li, Y., Huang, J.B., Ahuja, N., Yang, M.H.: Deep joint image filtering. In: European Conference on Computer Vision, Springer (2016) 154–169
40. Hui, T.W., Loy, C.C., Tang, X.: Depth map super-resolution by deep multi-scale guidance. In: European Conference on Computer Vision, Springer (2016) 353–369
41. Gu, S., Zuo, W., Guo, S., Chen, Y., Chen, C., Zhang, L.: Learning dynamic guidance for depth image enhancement. In: IEEE Conference on Computer Vision and Pattern Recognition, IEEE (2017) 3769–3778
42. Jaderberg, M., Simonyan, K., Zisserman, A., Kavukcuoglu, K.: Spatial transformer networks. In: Advances in Neural Information Processing Systems. (2015) 2017–2025
43. Zhou, T., Tulsiani, S., Sun, W., Malik, J., Efros, A.A.: View synthesis by appearance flow. In: European Conference on Computer Vision, Springer (2016) 286–301
44. Ganin, Y., Kononenko, D., Sungatullina, D., Lempitsky, V.: Deepwarp: Photorealistic image resynthesis for gaze manipulation. In: European Conference on Computer Vision, Springer (2016) 311–326
45. Park, E., Yang, J., Yumer, E., Ceylan, D., Berg, A.C.: Transformation-grounded image generation network for novel 3d view synthesis. In: IEEE Conference on Computer Vision and Pattern Recognition, IEEE (2017) 3500–3509
46. Yeh, R., Liu, Z., Goldman, D.B., Agarwala, A.: Semantic facial expression editing using autoencoded flow. arXiv preprint arXiv:1611.09961 (2016)
47. Liu, Z., Yeh, R.A., Tang, X., Liu, Y., Agarwala, A.: Video frame synthesis using deep voxel flow. In: IEEE International Conference on Computer Vision, IEEE (2017) 4463–4471
48. Isola, P., Zhu, J.Y., Zhou, T., Efros, A.A.: Image-to-image translation with conditional adversarial networks. In: IEEE Conference on Computer Vision and Pattern Recognition, IEEE (2016) 1125–1134
49. Johnson, J., Alahi, A., Li, F.: Perceptual losses for real-time style transfer and super-resolution. In: European Conference on Computer Vision, Springer (2016) 694–711
50. Parkhi, O.M., Vedaldi, A., Zisserman, A.: Deep face recognition. In: British Machine Vision Conference. (2015) 41.1–41.12
51. Li, Y., Liu, S., Yang, J., Yang, M.H.: Generative face completion. In: IEEE Conference on Computer Vision and Pattern Recognition, IEEE (2017) 3911–3919
52. Iizuka, S., Simo-Serra, E., Ishikawa, H.: Globally and locally consistent image completion. ACM Transactions on Graphics **36**(4) (2017) 107:1–107:14
53. Goodfellow, I.J., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., Bengio, Y.: Generative adversarial networks. In: Advances in Neural Information Processing Systems. Volume 3. (2014) 2672–2680
54. Salimans, T., Goodfellow, I., Zaremba, W., Cheung, V., Radford, A., Chen, X.: Improved techniques for training gans. In: Advances in Neural Information Processing Systems. (2016) 2234–2242
55. Yi, D., Lei, Z., Liao, S., Li, S.Z.: Learning face representation from scratch. arXiv preprint arXiv:1411.7923 (2014)
56. Cao, Q., Shen, L., Xie, W., Parkhi, O.M., Zisserman, A.: Vggface2: A dataset for recognising faces across pose and age. arXiv preprint arXiv:1710.08092 (2017)

57. Zhang, K., Zhang, Z., Li, Z., Qiao, Y.: Joint face detection and alignment using multitask cascaded convolutional networks. IEEE Signal Processing Letters **23**(10) (2016) 1499–1503
58. Kingma, D., Ba, J.: Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980 (2014)
59. Tuzel, O., Taguchi, Y., Hershey, J.R.: Global-local face upsampling network. arXiv preprint arXiv:1603.07235 (2016)