Can We See More? Joint Frontalization and Hallucination of Unaligned Tiny Faces

Xin Yu, Fatemeh Shiri, Bernard Ghanem, and Fatih Porikli, *Fellow, IEEE*

Abstract—In popular TV programs (such as CSI), a very low-resolution face image of a person, who is not even looking at the camera in many cases, is digitally super-resolved to a degree that suddenly the person's identity is made visible and recognizable. Of course, we suspect that this is merely a cinematographic special effect and such a magical transformation of a single image is not technically possible. Or, is it? In this paper, we push the boundaries of super-resolving (hallucinating to be more accurate) a tiny, non-frontal face image to understand how much of this is possible by leveraging the availability of large datasets and deep networks. To this end, we introduce a novel Transformative Adversarial Neural Network (TANN) to jointly frontalize very-low resolution (i.e., 16×16 pixels) out-of-plane rotated face images (including profile views) and aggressively super-resolve them (8×), regardless of their original poses and without using any 3D information. TANN is composed of two components: a transformative upsampling network which embodies encoding, spatial transformation and deconvolutional layers, and a discriminative network that enforces the generated high-resolution frontal faces to lie on the same manifold as real frontal face images. We evaluate our method on a large set of synthesized non-frontal face images to assess its reconstruction performance. Extensive experiments demonstrate that TANN generates both qualitatively and quantitatively superior results achieving over 4 dB improvement over the state-of-the-art.

Index Terms—Face, super-resolution, hallucination, face frontalization

1 Introduction

Recovering high-resolution (HR) face images from their low-resolution (LR) counterparts, known as face hallucination, has received significant attention in recent years. Existing face hallucination methods mainly focus on super-resolving nearly frontal faces, which provide critical perceptual information for the human visual system [1]. However, in most cases, LR faces may not necessarily be frontal. Super-resolving such non-frontal LR faces requires either frontalizing them first and then applying existing face hallucination techniques, or super-solving first (which highly depends on an available pose-specific exemplar dataset) and then frontalizing. Nevertheless, both of these options are naturally very challenging.

Conventional and emerging face frontalization methods [1], [5], [6], [7], [8], [9], [10] often rely on facial landmarks for warping 2D face images onto 3D models, and thus require the input images to have a sufficient resolution where such landmarks are detectable. This renders them ineffective for tiny face images. Without a proper frontalization, directly employing face hallucination methods [11], [12], [13], [14], [15], [16], [17], [18], [19], [20], [21], [22], [23] may cause severe artifacts due to large pose variations and misalignments. As shown in Figs. 1 and 3, for very low-resolution non-frontal face images, applying either face frontalization followed by hallucination, or hallucination followed by frontalization produces degraded results.

In this paper, we aim to *jointly* frontalize and hallucinate a given input face image so as to avoid the artifacts produced by either of these tasks individually. To do so, we present a new Transformative Adversarial Neural Network (TANN) that automatically frontalizes the LR faces while hallucinating the frontalized LR feature maps by an upscaling factor of 8× in an end-to-end fashion. Considering that an LR input face may undergo large pose variations and misalignments as seen in Fig. 1, our motivation is to force a non-frontal LR face to share the same latent representation of its corresponding frontal LR face and then super-resolve the latent representation. Thus, we first design a transformative subnetwork to encode a non-frontal LR face into a latent representation, where the representation of the input non-frontal LR face is forced to be similar to the latent representation of its frontal counterpart in the latent subspace. Then, we pass the latent representations, i.e., the frontalized LR feature maps, through a subnetwork that is composed of deconvolutional and spatial transformer layers [3], whose goal is to generate HR outputs. Inspired by previous works [22], [24], [25], [26], [27], we choose to employ an adversarial network to make these HR outputs more closely resemble real human faces.

In order to train our network, we not only employ the traditional pixel-wise image appearance similarity and class-wise similarity constraints used in our previous works [2], [22], but also develop a triplet loss to constrain the similarity

- *X. Yu, F. Shiri, and F. Porikli are with the Research School of Engineering, Australian National University, Canberra, ACT 0200, Australia. E-mail: {xin.yu, fatemeh.shiri, fatih.porikli}@anu.edu.au.*
- *B. Ghanem is with the Department of Electronic Engineering, King Abdullah University of Science and Technology, Thuwal 23955, Saudi Arabia. E-mail: bernard.ghanem@kaust.edu.sa.*

Manuscript received 2 May 2018; revised 17 Mar. 2019; accepted 20 Apr. 2019. Date of publication 2 May 2019; date of current version 4 Aug. 2020. (Corresponding author: Xin Yu.)
Recommended for acceptance by E. Shechtman.
Digital Object Identifier no. 10.1109/TPAMI.2019.2914039

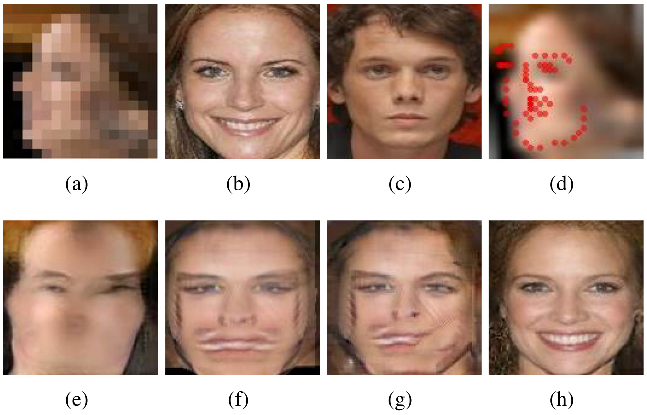


Fig. 1. Comparison with the combination of face hallucination [2] and frontalization [1] methods. (a) 16×16 LR non-frontal input image. (b) 128×128 HR original frontal image (not available in training). (c) The best possible match to the given LR image in the dataset after compensating for in-plane rotations by STN$_0$ [3]. (d) Detected landmarks by [4] after bicubic upsampling. (e) Result obtained by applying [1] first and then [2]. In [2], the first decoder and encoder are used to reduce image noise. Hereby, we only use the second decoder of [2] for super-resolving LR faces. (f) Result obtained by applying [2] first and then [1]. (g) Image generated by [2], which is retrained with LR non-frontal and HR frontal face images. (h) Our result.

of the latent representations between the input non-frontal faces and their ground-truth frontal LR ones. With the help of the proposed triplet loss, we are able to enforce that the representation of a side-view face to be close to its corresponding frontal LR face and far from other LR frontal faces in the latent subspace. In this manner, the faces upsampled from the encoded representations are not only similar to their HR frontal counterparts but also distinguishable from other hallucinated faces since the same upsampling subnetwork is used for super-resolution. In particular, the upsampled frontalized faces can share similar facial characteristics with their corresponding ground-truth ones after super-resolution. Thus, our triplet loss preserves the identity information implicitly. Note that, different from the traditional triplet loss, where both negative and positive examples are used to calculate the gradients of neural networks and updated simultaneously, we only update the latent representations of LR side-view faces by forcing them to be close to the representations of their ground-truth frontal faces without affecting positive and negative LR frontal faces. Furthermore, we exploit a feature-wise similarity constraint, known as perceptual loss [28], to make the hallucinated facial characteristics similar to the ground-truths, thus improving the visual quality.

Although deep neural networks have given rise to major advances in many computer vision tasks, they require very large datasets to train millions of parameters in their models. In our case, the existing large-scale face datasets [29], [30] do not provide a sufficient number of frontal and non-frontal face image pairs for training our TANN. To obtain a large corpus of frontal and non-frontal face image pairs for the goal of training our deep neural network, we construct a set of out-of-plane rotated images from available frontal faces mapped onto a 3D face model. We first map randomly chosen frontal images to a 3D model, and then render different views of the 3D face, similar to the work in [31]. This allows us to have high-quality HR frontal faces as our ground-truth images. It is important to note that this step is only to construct the *training dataset*, as we do not use any 3D models in our network (neither in training, nor in testing). In our experiments, we use non-frontal faces whose 3D models are unknown to demonstrate that TANN can hallucinate and frontalize different views of any unaligned LR face beyond the poses it is exposed to in training.

Overall, our contributions can be summarized as follows:

- We introduce a new transformative adversarial neural network to simultaneously hallucinate (by an upscaling factor of 8×) and frontalize tiny (16×16 pixels) unaligned face images with pose variations up to $\pm 75^\circ$.
- We propose a new triplet loss to encode non-frontal LR faces into a latent subspace without distorting the encoding of frontal LR ones. With the help of the proposed triplet loss, we can force non-frontal LR faces to be close to their ground-truth frontal ones while keeping away from other faces in the latent subspace. To the best of our knowledge, our method is the first attempt to employ the triplet training strategy in the face hallucination task.
- We perform the training of our network in an end-to-end fashion by incorporating the reconstruction, perceptual, discriminative and triplet loss terms. In order to train our network, we also provide a dataset of corresponding frontal and non-frontal view face image pairs, which will be made available on-line to the vision community at large.
- We achieve superior hallucination results and outperforms the state-of-the-art by a large margin of *4.0* dB PSNR. Our method eliminates the need for facial landmarks or 3D face models as it is agnostic to the underlying in-plane and out-of-plane pose variations and spatial deformations. In the testing phase, our method can successfully process faces that are imaged at views not seen during training.

This paper is an extension of our previous works [2], [22], [26]. Unlike our previous works [2], [22], [26], which only focus on super-resolving LR face images, this paper aims at upsampling LR faces while frontalizing them. However, our previous upsampling networks [2], [26] cannot upsample and frontalize profile faces even after retraining, as shown in Fig. 1g. Therefore, we first project LR faces in different poses into latent representations by an encoder network, and then upsample from the latent representations by a decoder network. To the best of our knowledge, our method is the first attempt to provide a unified framework for super-resolution and frontalization of unaligned very low-resolution face images, reducing significantly the artifacts introduced by either strategy, when considered individually.

2 Related Work

Our work mainly focuses on two aspects: face frontalization and hallucination. We briefly review noteworthy face frontalization and hallucination works below.

Face Frontalization. Generating a frontal face from a single non-frontal face image is very challenging due to self-occlusions and various pose variations, and has received significant attention in computer vision. Seminal works date back to the 3D Morphable Model (3DMM) [5], where a face is represented by the shape and texture bases in PCA subspace. After obtaining the the shape and texture coefficients of an input face image, Blanz and Vetter [5] render novel views of an input face. Driven by 3DMM, Yang et al. [6] estimate 3D surface from face appearance and then synthesize new expressions of the given face. However, these methods require the input face images to be nearly frontal in order to estimate the shape and appearance coefficients of input faces in PCA subspace. Dovgard and Basri [32] exploit the facial symmetry to estimate 3D geometry of the given faces and render frontal faces. Similarly, Hassner et al. [1] use facial symmetry to render out-of-view facial regions. Some methods, e.g., [7], [8], [31], [33], [34], attempt to reconstruct frontal views by mapping a 2D face image onto a 3D reference surface mesh after registering and normalizing the face image. Since they need to detect facial landmarks in the input images and establish correspondences of landmark points to 3D or 2D reference models, they require images in sufficiently high resolutions. Based on the fact that frontal faces have the minimum rank of all different poses, Sagonas et al. [9] propose a statistical face frontalization method, but the appearance of their frontalized faces may not be consistent with the input faces.

Deep learning based face frontalization methods have been proposed recently as well [34], [35], [36], [37], [38], [39], [40]. Zhu et al. [35] present a deep neural network to frontalize HR faces by exploiting the symmetry and similarity of facial components. Their method does not require estimation of a 3D model, but it cannot maintain appearance similarity between the frontalized and input faces either. Yim et al. [36] develop a multi-task deep neural network to rotate faces, but their method outputs blurry frontal faces due to the aggressive downsampling operations in the encoder. Similarly, Cole et al. [38] learn to generate facial landmarks and textures from features extracted by a face recognition network. Since Cole et al. warp input faces to the mean face geometry by using facial landmarks, the resolutions of their inputs need to be sufficiently large. Very recently, Huang et al. [39] employ two deep neural networks, i.e., global and local networks, to frontalize faces. However, their local network needs to extract HR facial components for identity preservation and to align HR facial components to pre-defined positions, and thus their method is not suitable for very LR unaligned non-frontal face images. Xi et al. [40] combine 3DMM and a generative adversarial network to frontalize faces with arbitrary poses. They also need to localize facial landmarks when mapping the input faces to the 3DMM. Thus their method requires sufficient resolutions for input images. Tran et al. [41] present a convolutional neural network (CNN) to regress 3DMM shape and texture parameters to speed up the optimization of 3DMM, but their method does not render frontalized faces which are similar to the input faces in terms of image intensity. Instead of localizing facial landmarks explicitly in the face images, Chang et al. [42] employ a simple CNN to regress 6 degrees of freedom (6DoF) 3D head poses from image intensities. Then the estimated 6DoF parameters can be used to align face images without localizing facial landmarks explicitly. By transforming input image intensities with the estimated parameters, [42] is able to preserve the appearance similarity between the input faces and their counterparts in the generated views.

Face Hallucination. Face super-resolution (FSR), also known as face hallucination, aims at magnifying an LR image to its HR version and can be roughly grouped into three categories: holistic-based, part-based, and deep network based solutions.

Holistic-based methods attempt to super-resolve an entire HR face by using global face models, often learned by PCA. Wang and Tang [14] establish a linear mapping between LR and HR face subspaces to super-resolve HR faces, while Liu et al. [15] learn a global appearance model for upsampling LR inputs and employ a local nonparametric model to enhance the facial details. Kolouri and Rohde [20] propose to morph an HR output from the aligned exemplar faces similar to LR inputs by the optimal transport and subspace learning techniques. Because holistic-based methods require LR inputs to be accurately aligned and to share the same pose and expression as HR references when learning global face models, they are very sensitive to misalignments and pose variations.

Instead of super-resolving entire faces, part-based methods upsample facial regions and thus can address various poses. They either use reference position patches, or employ facial components to restore the HR counterparts of LR inputs. For instance, Baker and Kanade [12] reconstruct high-frequency details of aligned frontal face images by finding the best mapping between LR and HR patches. Similarly, Ma et al. [17] employ position patches extracted from multiple aligned HR images to upsample aligned LR face images. Rather than reconstructing patches in the image domain, Yang et al. [18] and Li et al. [19] super-resolve HR image patches by employing sparse coding techniques to achieve better performance. Tappen and Liu [43] apply SIFT flow [44] to align the facial parts of LR images and reconstruct HR facial details by warping the reference HR images, while Yang et al. [45], [46] localize facial components in the LR images by a facial landmark detector and then reconstruct details from the similar HR reference components. Since these methods need to extract facial components in LR face images accurately, their performance degrades dramatically when the LR faces are tiny. We refer the readers to the work [21] for a more comprehensive survey on face hallucination using traditional approaches.

As large-scale datasets become available, Zhou et al. [47] propose a convolutional neural network to extract facial features and recover facial details from the extracted features. Yu and Porikli [23] consolidate deconvolutional and convolutional layers for super-resolving LR face images, but they improve the visual quality by a post-processing technique, i.e., an unsharp filter. The work presented in [22] develops a discriminative generative network to super-resolve aligned LR face images in an end-to-end fashion while Huang et al. [48] exploit wavelet coefficients learned by CNN to restore HR faces. In order to relax the requirement of face alignment, Yu and Porikli [26] embed multiple spatial transformer networks [3] into the generative network of [22]. Their follow-up work [2] employs a decoder-encoder-decoder structure to super-resolve noisy LR faces

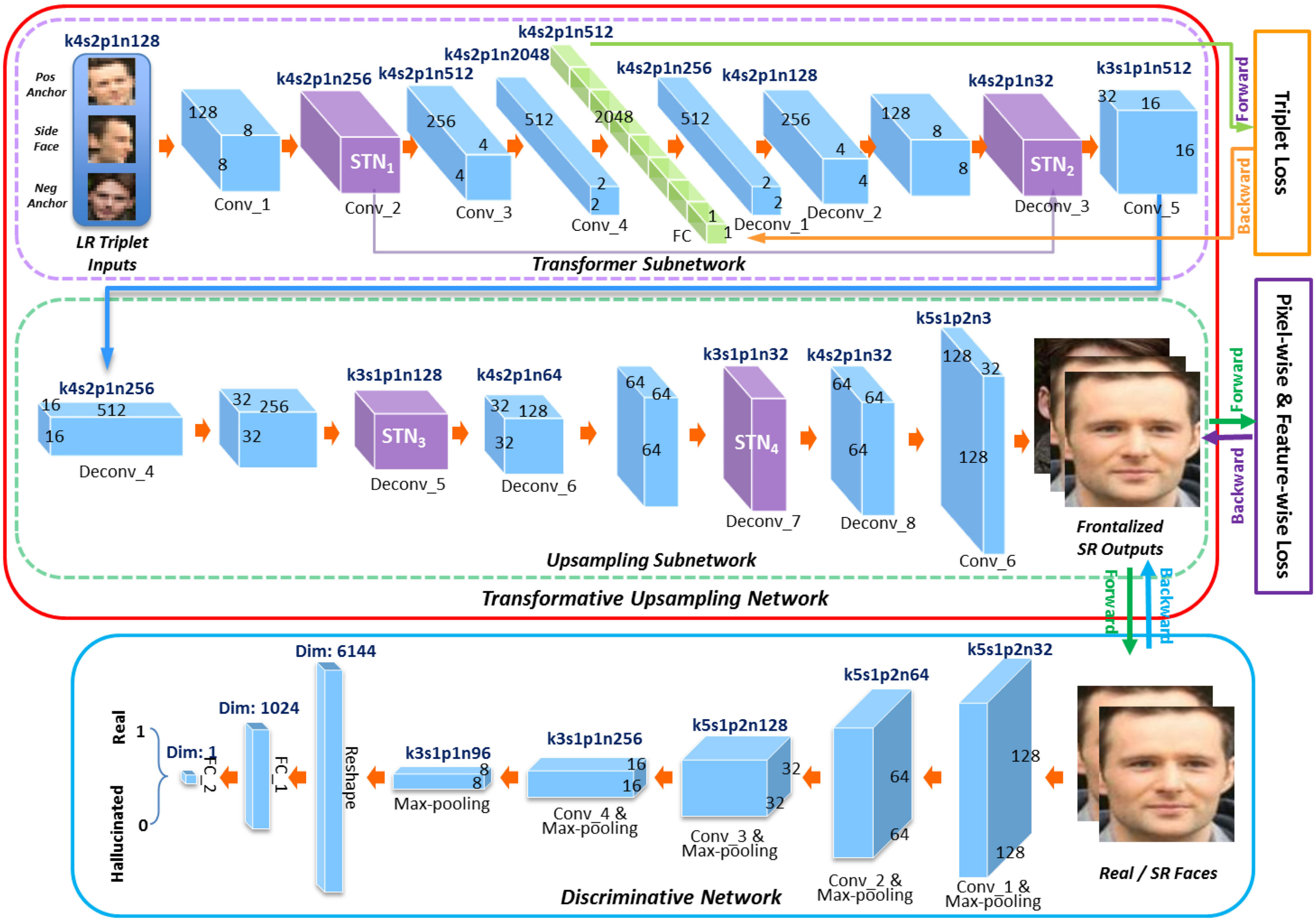


Fig. 2. TANN consists of two parts: A transformative upsampling network (red box) and a discriminative network (blue box). In our transformer subnetwork, we also employ skip connections between our encoding layers and decoding layers, indicated by the purple line. For simplicity, we only draw the first skip connections.

while suppressing image noise. Xu et al. [27] employ the generative adversarial framework [24] as well as a multi-class adversarial loss to upsample blurry and LR face and text images. Dahl et al. [49] exploit the framework of PixelCNN [50], known as an autoregressive generative model, to hallucinate very low-resolution face images. Towards the same goal, Zhu et al. [51] use a cascade bi-network to upsample very low-resolution and unaligned faces, of which one is used to super-resolve low-frequency components of face images and the other is employed to hallucinate high-frequency facial details. Since these deep learning based methods do not take out-of-plane rotations of faces into account and are restricted to small pose variations, (i.e., within $\pm 30^{\circ}$), they may fail to super-resolve LR faces with large pose variations.

Recently, some face hallucination methods have been proposed to handle large pose variations in LR face images by exploiting facial structure information [52], [53]. Chen et al. [52] first super-resolve low-frequency components of input faces and then enhance the facial details based on the facial landmarks estimated from the upsampled faces. Bulat et al. [53] upsample face images in different poses by imposing a loss to enforce the detected landmarks in the super-resolved faces to be close to the ground-truth ones. However, these methods only super-resolve profile faces rather than frontalizing them for better observation and analysis. Even though profile faces can be super-resolved with authentic details, localizing facial landmarks from those profile faces for frontalization is still challenging.

Due to the above limitations, simply cascading face hallucination and frontalization methods is not an acceptable solution for our problem.

3 Proposed Method: TANN

Our network has two components: (i) a transformative upsampling network, which transforms different poses to the frontal one and also super-resolves the frontalized LR feature maps; and (ii) a discriminative network, which forces the generated HR frontal faces to lie on the manifold of authentic HR face images. Fig. 2 illustrates the overall architecture of TANN.

In the training phase, the entire network is trained in an end-to-end fashion to compensate for possible artifacts induced by any of the frontalization and hallucination tasks. As shown in Fig. 3k, when we train the upsampling network separately, i.e., generating frontalized LR faces as intermediate results, the transformer subnetwork may suffer from the loss of information contained in its feature maps because it is enforced to output 3 channel LR faces as its objective function rather than 32 channel feature maps.

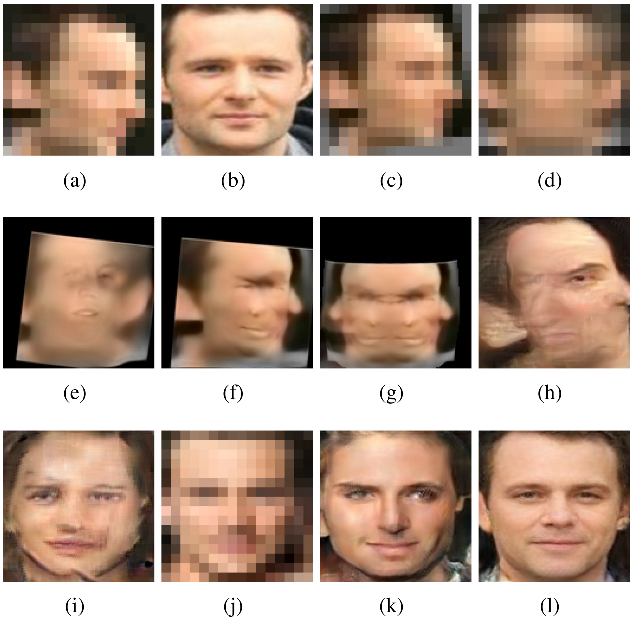


Fig. 3. Artifacts caused by the state-of-the-art face frontalization and hallucination methods. (a) The input 16×16 LR image. (b) The original 128×128 HR frontal image. (c) The aligned upright version of (a) by STN_0. (d) Frontalized result of (c) using [1]. Note that, we first upsample (c) by bicubic interpolation, then apply [1], and downsample the frontalized result. (e) HR image after applying [51] to (d). (f) HR image after applying [51] to (c) directly. (g) The frontalized version of (f) by [1]. (h) The result of applying [2] to (a). (i) The result of TANN without the transformer subnetwork, which is similar to the upsampling network [2], retrained with LR non-frontal and HR frontal faces. (j) The aligned and frontalized LR face by our transformer subnetwork. Note that, in our end-to-end trained TANN, the output of the transformer network is a set of feature maps not an image. (k) The hallucinated result of (j) by our upsampling subnetwork (here, we retrained the upsampling network). (l) Our final result.

This may lead to accumulated errors and obvious deviations in the output of the upsampling subnetwork due to the incorrect input images for upsampling. Thus, feeding 32 feature maps directly to the upsampling network is a better choice.

3.1 Transformative Upsampling Network (TUN)

In Fig. 2, our transformative upsampling network is shown (red box). TUN is composed of two parts: a transformer subnetwork and an upsampling subnetwork. The transformer part (purple box) aims at encoding non-frontal LR faces into latent representations which are close to the latent representations of their corresponding frontal LR ones. By doing so, we can achieve the latent codes of frontalized LR faces. Our transformer subnetwork is constructed by convolutional layers, a fully-connected layer, deconvolutional layers and spatial transformer layers. Since the input LR faces undergo in-plane rotations, translations and scale changes, multiple spatial transformer networks (STN) [3] are embedded as intermediate layers to compensate for such affine transformations. Moreover, because STNs learn 2D affine warps rather than out-of-plane rotations, they cannot recover self-occluded parts of faces. To solve this problem, our intuition is that we can project different views of a face into a subspace, where their encoded representations are enforced to lie close to the representations of their corresponding frontal one. Therefore, we incorporate a fully-connected layer to encode the feature maps of LR profile faces as well as design a triplet loss to force the similarity between the representations of LR profile and frontal ones.

To illustrate the effectiveness of the transformer subnetwork, we change the channel number of its output layer to 3, and use LR frontal faces as ground-truth images to train this subnetwork. As shown in Figs. 3j and 4d, it can successfully generate an LR frontal face image. Note that, when training our TANN, we do not employ LR frontal faces as supervision to prevent the aforementioned drift issue.

After obtaining the feature maps of LR frontal faces generated by the transformer subnetwork, we apply an upsampling subnetwork (green box in Fig. 2) to hallucinate the high-frequency facial details of frontal faces. Because the resolution of LR input images is very low, STNs in our transformer subnetwork may not align LR faces accurately. The LR feature maps generated by the transformer network may still contain misalignments. We employ the upsampling structure used in our previous works [2], [26] for further alignment and super-resolution.

As shown in Fig. 3h, simply applying the method of [2] to LR profile faces cannot provide high-quality HR frontal face images. This manifests that upsampling LR non-frontal faces with large pose variations is more difficult compared to LR frontal faces and also indicates the necessity of our transformer subnetwork. Since the mapping between common LR patterns and HR facial details can be easily learned from frontal faces, we frontalize LR inputs first and then hallucinate them.

3.2 Discriminative Network

As demonstrated in our previous works [2], [22], [26], only using euclidean distance (pixel-wise ℓ_2 loss) between the upsampled faces and the ground-truth HR faces tends to generate over-smoothed results. Therefore, a class-specific discriminative objective is also incorporated into our TUN, aiming to force the hallucinated HR face images to lie on the same manifold of real frontal face images.

As shown in Fig. 2 (blue box), the discriminative network consists of convolutional layers, max-pooling layers, dropout layers, and fully-connected layers. It is designed to determine whether an image is sampled from real face images or the hallucinated ones. The discriminative loss, also known as adversarial loss, will be back-propagated to update the parameters of TUN as well. With the help of the adversarial loss, we can generate more realistic HR frontal faces. Fig. 4 illustrates the impact of the adversarial loss on the final results.

3.3 Training Details of TANN

We construct LR profile and HR frontal ground-truth face image pairs $\{l_i, h_i\}$ for our training purpose, where h_i represents the aligned frontal HR face images (only eyes are aligned), and l_i is the synthesized LR side-view face images from h_i. For each HR frontal face h_i, we generate five different views, i.e., $\{0^\circ, \pm 40^\circ, \pm 75^\circ\}$, to construct LR/HR training pairs. Using these five distinct poses is a trade-off between a sufficient coverage of pose variations and the reasonable size of the training dataset and also suggested in [31]. More details are provided in Section 4.

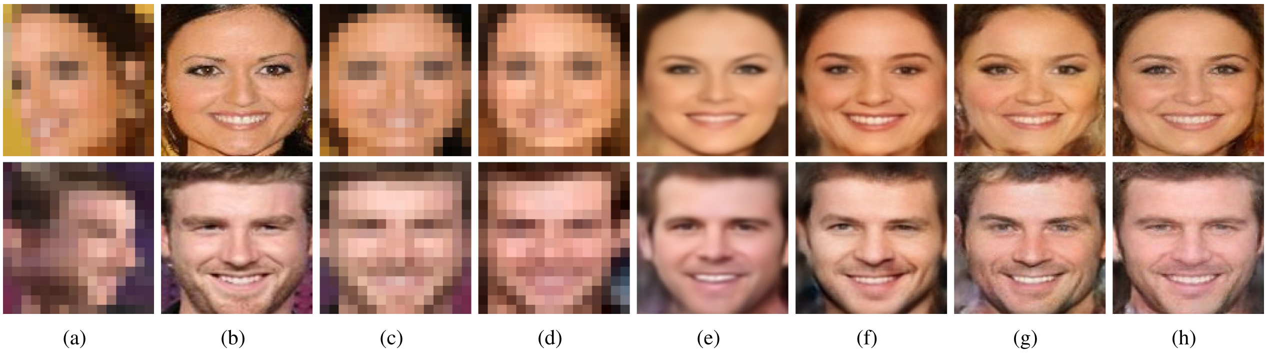


Fig. 4. Illustrations of influence of different losses. (a) The input 16×16 LR images. (b) The original 128×128 HR frontal images. (c) The downsampled version of (b). (d) The frontalized LR faces by our transformer subnetwork. (e) The upsampling results only using pixel-wise loss. (f) The upsampling results using the pixel-wise and perceptual losses. (g) The upsampling results without using the triplet loss. (h) Our final results.

In training our TANN, we not only enforce the conventional pixel-wise intensity similarity, known as pixel-wise ℓ_2 loss, but also the feature-wise similarity, known as perceptual loss [28], to obtain high-quality results. Similar to the works [22], [26], the adversarial loss is also employed to attain visually appealing frontalized HR face images. As mentioned in Section 3.1, we also develop a triplet loss to force the representations of LR profile faces to be similar to the representations of their frontal faces. In this manner, we can frontalize LR profile faces without degrading super-resolution of frontal ones.

Pixel-Wise Intensity Similarity Loss. We constrain the generated HR frontalized face $\hat{h}_i$ to be similar to its ground-truth frontal counterpart h_i in terms of image intensities. Thus we employ a pixel-wise ℓ_2 regression loss $\mathcal{L}_{pix}$ to impose the appearance similarity constraint, expressed as

$$\begin{aligned}\mathcal{L}_{pix} &= \mathbb{E}_{(\hat{h}_i,h_i)\sim p(\hat{h},h)}\|\hat{h}_i - h_i\|_F^2 \\ &= \mathbb{E}_{(l_i,h_i)\sim p(l,h)}\|\mathcal{T}_t(l_i) - h_i\|_F^2,\end{aligned} \tag{1}$$

where t and $\mathcal{T}$ are the parameters and the output of TUN, $p(\hat{h},h)$ represents the joint distribution of the frontalized HR faces and their corresponding frontal HR ground-truths, and $p(l,h)$ indicates the joint distribution of the LR and HR face images in the training dataset.

Feature-Wise Similarity Loss. As mention in [22], pixel-wise ℓ_2 loss leads to over-smoothed super-resolved results. Here, we employ a feature-wise similarity loss, known as perceptual loss [28], to constrain the super-resolved HR faces to share the same facial details as their ground-truth counterparts, thus attaining high-quality results with rich facial details. The perceptual loss $\mathcal{L}_{feat}$ measures euclidean distance between the feature maps of HR frontalized and ground-truth faces extracted by a deep neural network, written as

$$\begin{aligned}\mathcal{L}_{feat} &= \mathbb{E}_{(\hat{h}_i,h_i)\sim p(\hat{h},h)}\|\Phi(\hat{h}_i) - \Phi(h_i)\|_F^2 \\ &= \mathbb{E}_{(l_i,h_i)\sim p(l,h)}\|\Phi(\mathcal{T}_t(l_i)) - \Phi(h_i)\|_F^2,\end{aligned} \tag{2}$$

where $\Phi(\cdot)$ denotes feature maps extracted by the ReLU32 layer in VGG-19 [54], which gives good empirical performance in our experiments.

Adversarial Loss. In order to achieve visually appealing results, we infuse class-specific discriminative information into TUN by exploiting a discriminative network, similar to our previous works [2], [22], [26]. Our goal is to make the discriminative network fail to distinguish generated faces from real ones. In this manner, we enforce the super-resolved HR frontal faces to lie on the manifold of real frontal HR face images. Therefore, the discriminative network is used to categorize real HR frontal faces and generated ones, and thus its objective function is expressed as

$$\begin{aligned}\mathcal{L}_{\mathcal{D}} &= -\mathbb{E}_{(\hat{h}_i,h_i)\sim p(\hat{h},h)}\Big[\log \mathcal{D}_d(h_i) + \log\left(1 - \mathcal{D}_d(\hat{h}_i)\right)\Big] \\ &= -\mathbb{E}_{h_i\sim p(h)}\log \mathcal{D}_d(h_i) - \mathbb{E}_{\hat{h}_i\sim p(\hat{h})}\log\left(1 - \mathcal{D}_d(\hat{h}_i)\right) \\ &= -\mathbb{E}_{h_i\sim p(h)}\log \mathcal{D}_d(h_i) - \mathbb{E}_{l_i\sim p(l)}\log\left(1 - \mathcal{D}_d(\mathcal{T}(l_i))\right),\end{aligned} \tag{3}$$

where d represents the parameters of the discriminative network, $p(l)$, $p(h)$ and $p(\hat{h})$ indicate the distributions of the LR, HR ground-truth frontal and the generated faces respectively, and $\mathcal{D}_d(h_i)$ and $\mathcal{D}_d(\hat{h}_i)$ are the outputs of the discriminative network. To make the discriminative network distinguish hallucinated faces from real ones, we minimize the loss $\mathcal{L}_{\mathcal{D}}(d)$ and update the parameters d.

Meanwhile, our TUN aims to fool the discriminative network. Therefore, the adversarial loss for our TUN is

$$\begin{aligned}\mathcal{L}_{\mathcal{T}} &= -\mathbb{E}_{\hat{h}_i\sim p(\hat{h})}\log\left(\mathcal{D}(\hat{h}_i)\right) \\ &= -\mathbb{E}_{l_i\sim p(l)}\log\left(\mathcal{D}(\mathcal{T}_t(l_i))\right).\end{aligned} \tag{4}$$

Here, we minimize the loss $\mathcal{L}_{\mathcal{T}}(t)$ to update the parameters t. These two adversarial losses in Eqns. (3) and (4) are employed to update our TUN and discriminative network respectively in an alternating fashion.

Triplet Loss. In order to frontalize side-view LR faces, we present a triplet loss to constrain the encoded LR faces to be close to the latent representations of their corresponding frontal ones and far away from other frontal faces in the latent subspace. Therefore, our proposed triplet loss is expressed as

$$\mathcal{L}_{tri} = \mathbb{E}_{(l_i^+,l_i^-,l_i)\sim p(\mathcal{S})}\frac{\Big[\|\mathcal{F}(l_i) - \mathcal{F}(l_i^+)\|_F^2 - \|\mathcal{F}(l_i) - \mathcal{F}(l_i^-)\|_F^2\Big]_+}{\|\mathcal{F}(l_i)\|_F^2}, \tag{5}$$

where $\mathcal{F}(\cdot)$ indicates the encoded latent representation by the fully-connected layer in our transformer subnetwork, (l_i^+, l_i^-, l_i) represents a triplet sample from the set of all possible triplets $\mathcal{S}$ in the training set. l_i is an LR profile face, l_i^+, dubbed positive anchor, is the corresponding frontal LR face of l_i, and l_i^-, dubbed negative anchor, is any other frontal LR face. One example of the triplets is shown in Fig. 2. In addition, $[x]_+$ denotes the operator $\max\{x, 0\}$.

Since our network aims at super-resolving LR faces rather than clustering faces, it should not distort the mapping between LR and HR frontal faces. Considering that positive and negative anchors are LR frontal faces, updating the gradients with respect to the representations of the positive and negative anchors will distort the mapping between LR and HR frontal faces. In other words, clustering triplets by adjusting the latent representations of positive and negative anchors would damage the end-to-end mapping between LR and HR frontal faces and thus leads to inferior super-resolution performance. Different from the triplet loss presented in [55], we take positive and negative anchors as constant and thus only back-propagate gradients with respect to the latent representations of LR side-view faces. In this manner, we are able to upsample frontal faces without introducing distortions while forcing the LR profile faces to be close to their frontal counterparts in the latent space.

In our TANN, all the layers are differentiable and RMSprop [56] is used to update the parameters t and d. We update the parameters d by minimizing the adversarial loss $\mathcal{L}_{\mathcal{D}}$ as follows:

$$\begin{aligned} \Delta^{i+1} &= \gamma\Delta^{i} + (1-\gamma)(\frac{\partial \mathcal{L}_{\mathcal{D}}}{\partial d})^2, \\ d^{i+1} &= d^{i} - r\frac{\partial \mathcal{L}_{\mathcal{D}}}{\partial d}\frac{1}{\sqrt{\Delta^{i+1}+\epsilon}}, \end{aligned} \tag{6}$$

where r and γ represent the learning rate and the decay rate respectively, i indicates the index of the iterations, Δ is an auxiliary variable, and ϵ is set to 10^{-8} to avoid division by zero. We employ multiple losses, i.e., $\mathcal{L}_{pix}$, $\mathcal{L}_{feat}$, $\mathcal{L}_{\mathcal{T}}$ and $\mathcal{L}_{tri}$, to update our TUN and the object function is expressed as

$$\mathcal{L}_{TUN} = \mathcal{L}_{pix} + \eta\mathcal{L}_{feat} + \lambda\mathcal{L}_{\mathcal{T}} + \mu\mathcal{L}_{tri}, \tag{7}$$

where η, λ and μ are the trade-off weights. Since we aim at super-resolving frontal HR faces rather than generating random faces, we put lower weights on the feature-wise, adversarial and triplet losses and set λ, η and μ to $10e^{-2}$, $10e^{-2}$ and $10e^{-4}$ respectively. Then, the parameters of TUN t are updated by the gradient descent as follows:

$$\begin{aligned} \Delta^{i+1} &= \gamma\Delta^{i} + (1-\gamma)(\frac{\partial \mathcal{L}_{TUN}}{\partial t})^2, \\ t^{i+1} &= t^{i} - r\frac{\partial \mathcal{L}_{TUN}}{\partial t}\frac{1}{\sqrt{\Delta^{i+1}+\epsilon}}. \end{aligned} \tag{8}$$

As the iteration progresses, the output faces will be more similar to real faces. Therefore, we gradually reduce the impact of the discriminative network by decreasing λ

$$\lambda^{j} = \max\{\lambda \cdot 0.995^{j}, \lambda/2\}, \tag{9}$$

where j is the index of the epochs. Eqn. (9) not only increases the impact of the appearance similarity term but also preserves the class-specific discriminative information in the training phase.

3.4 Hallucinating Frontal HR from Non-Frontal LR

The discriminative network is only employed in the training phase. In the testing phase, we feed an unaligned LR profile face image into the transformative upsampling network to obtain its upright and frontal HR version. Note that, only in the training stage, we need to feed the network with triplet samples due to employing the triplet loss. In the testing stage, our network is able to super-resolve and frontalize a single image. Since aligned HR frontal face images are employed as ground-truths, TUN will output aligned and frontalized HR faces directly. As a result, our method does not need to estimate the face orientations or align very low-resolution images beforehand, and provides an end-to-end and highly nonlinear mapping from an unaligned LR profile face image to its frontal HR version.

3.5 Implementation Details

The STN layers, as shown in Fig. 2, are built by convolutional and ReLU layers (Conv+ReLU), max-pooling layers with a stride 2 (MP2) and fully connected layers (FC). Since STN is mainly used for calibrating in-plane transformations, we employ the similarity transformation for alignment. Specifically, STN$_1$ and STN$_2$ share the same architecture and consist of Conv+ReLU (filter size: $20 \times 128 \times 3 \times 3$ with 1 pixel padding), MP2, Conv+ReLU ($20 \times 20 \times 3 \times 3$), FC+ReLU (from 400 to 20 dimensions), and FC (from 20 to 4 dimensions). STN$_3$ is composed of MP2, Conv+ReLU ($20 \times 256 \times 5 \times 5$), MP2, Conv+ReLU ($20 \times 20 \times 5 \times 5$), FC+ReLU (from 80 to 20 dimensions) and FC (from 20 to 4 dimensions). STN$_4$ is composed of MP2, Conv+ReLU ($128 \times 64 \times 5 \times 5$), MP2, Conv+ReLU ($20 \times 128 \times 5 \times 5$), MP2, Conv+ReLU ($20 \times 20 \times 3 \times 3$), FC+ReLU (from 120 to 20 dimensions) and FC (from 20 to 4 dimensions).

Similar to the works [24], [57], batch normalization [58] is employed after each convolution except the final output layer of TUN and dropout is applied to the feature maps in the discriminative network. In the experimental part, some algorithms may require alignment of LR inputs, i.e. [17]. Hence, we employ another network STN$_0$ to align the LR face images to the upright position, and STN$_0$ consists of Conv+ReLU ($128 \times 3 \times 3 \times 3$ with 1 pixel padding), MP2, Conv+ReLU ($20 \times 20 \times 3 \times 3$), MP2, FC+ReLU (from 180 to 20 dimensions), and FC (from 20 to 4 dimensions).

We also use a triplet pair $\{(l_i^+, l_i, l_i^-), (h_i, h_i, h_i^-)\}$ as a unit to construct our mini-batch in training, where h_i is the HR frontal face image corresponding to the LR profile face l_i and the LR frontal face l_i^+, and h_i^- is the HR frontal version of the LR frontal face l_i^-. The triplet pairs are not only designed to calculate the triplet loss but also compatible with the other losses. Therefore, our network can be trained in an end-to-end fashion.

The learning rate r is set to 0.001 and multiplied by 0.99 after each epoch, η is set to 0.01, and the decay rate is set to 0.01. The training codes and details can be downloaded from https://github.com/XinYuANU/JFFH.

4 Synthesized Dataset

Training of a deep neural network requires a large number of samples to prevent models from overfitting to the training dataset. However, the publicly available large-scale face datasets [29], [30] only provide faces in the wild but not frontal/non-frontal pairs. For the training purpose, we opt to generate a large set of synthesized LR non-frontal faces from HR frontal face images.

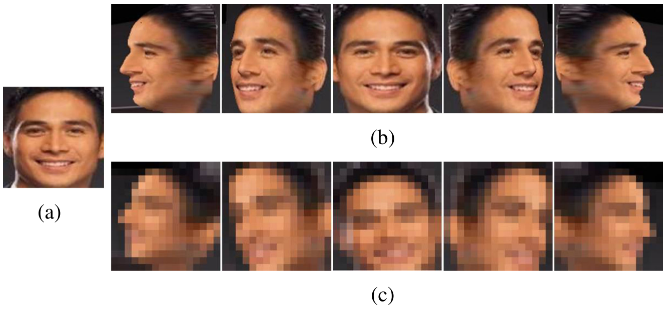


Fig. 5. Illustration of the synthesized dataset. (a) Original frontal HR face image. (b) The generated views of (a). (c) Spatially transformed and downsampled version of (b).

There are a number of alternative approaches available. For instance, Hassner et al. [1] render 2D frontal faces from different side-view faces using a single 3D reference mesh. However, when the out-of-view face regions are large, these methods are prone to artifacts. Similarly, landmark detection algorithms may fail to localize facial landmarks accurately in large poses.

We adopt the idea of [31] to generate different views from HR frontal ones. We use a single 3D face model to render HR out-of-plane rotated faces while taking advantage of the mirror-symmetry for the positive and negative angles to produce five different views of faces, i.e., $\{0^\circ, \pm 40^\circ, \pm 75^\circ\}$. Specifically, we first randomly select 10K cropped frontal faces (within $\pm 5^\circ$) from the CelebA [30], and resize them to 128×128 pixels. We use these images as our HR ground-truth faces h_i. Then we generate the non-frontal LR faces l_i by transforming and downsampling the reconstructed HR images down to 16×16 pixels. Here, we employ the bicubic interpolation method (*imresize* function in Matlab) to generate LR face images. Therefore, we obtain 50K LR/HR face pairs for training and testing of our network. Fig. 5 illustrates sample pairs $\{l_i, h_i\}$ generated from a single frontal face.

Since our side-view face images are generated from frontal faces by a generic 3D face model, some distorted areas or artifacts may appear in the synthesized side-view faces, such as ear and hair regions as visible in Fig. 5. When we downsample the generated profile faces, those artifacts can be largely reduced. Furthermore, those regions may not be visible in the final frontalized and upsampled HR face images, and thus deep neural networks might learn to ignore those artifacts. However, if localizations of facial landmarks are erroneous, the generated views may undergo obvious distortions. For instance, when noses or chins suffer severe misalignments to the 3D model, the synthesized profile faces can be severely distorted or even blended with backgrounds. Those artifacts cannot be alleviated in LR faces, thus bringing extra ambiguity to the super-resolution and frontalization process. Therefore, we manually choose frontal faces where landmarks are well localized on the facial components to avoid generating side-view faces in sub-quality.

5 Experimental Evaluation

We compare our method with ten state-of-the-art methods qualitatively and quantitatively. As mentioned in Section 4, we assemble 50K LR/HR face pairs, and randomly choose 9K frontal face images for training (45K LR/HR pairs), and 1K faces for testing (5K LR/HR pairs). In training TANN, we randomly choose a side-view LR face, its corresponding frontal LR face and any other frontal LR face to construct an input triplet (l_i^+, l_i, l_i^-) as well as employ their corresponding HR ground-truth triplet (h_i, h_i, h_i^-) as supervision. In all cases, the training data and test data do *not* overlap. We use different ground-truth HR frontal faces in the training and testing phases.

5.1 Qualitative Comparisons with the SoA

Since Ma et al.'s method [17] requires the input LR faces to be aligned uprightly, we train STN_0 to align the LR inputs to the upright position for a fair comparison. Note that, our method does not need any alignment or pose estimation in advance.

As illustrated in Figs. 6c and 7c, different combinations of bicubic interpolation and the frontalization method [1] cannot produce authentic frontal face details. Because of the low resolution of inputs, Hassner et al.'s method [1] fails to detect facial landmarks and outputs erroneous frontalized faces while bicubic interpolation is handicapped to generate necessary high-frequency facial details.

Kim et al. [59] propose a very deep CNN based general purpose super-resolution (SR) method, known as VDSR. Since VDSR is trained on natural image patches and does not provide an upscaling factor of 8×, we retrain VDSR with face patches extracted from CelebA dataset by an upscaling factor of 8×. As shown in Figs. 6d and 7d, VDSR fails to produce facial details and thus contaminates the outputs of [1] with ghosting artifacts.

Leigh et al. [60] present a generic super-resolution method, dubbed SRGAN. SRGAN employs the framework of generative adversarial networks [24], [57] to enhance the visual quality and is trained by using not only a pixel-wise ℓ_2 loss but also an adversarial loss. SRGAN provides an upscaling factor of 8×, but it is only trained on general patches. Thus, we retrain SRGAN on face images as well. As shown in Figs. 6e and 7e, the generated facial details by SRGAN are still blurry, and [1] fails to localize the landmarks accurately in the upsampled faces. Thus, the final results suffer from severe artifacts.

Ma et al. [17] super-resolve LR inputs by exploiting position patches, and require the LR inputs to be precisely aligned with the exemplar training dataset. Here, aligned HR face images from CelebA dataset are employed as the exemplar dataset. It spawns severe artifacts in the upsampled faces because of large pose variations that exist in the input LR images as visible in Fig. 7f. Due to the faulty frontalization by [1], this method also produces distorted facial details, as shown in Fig. 6f.

Zhu et al. [51] present a deep cascaded bi-network for face hallucination, called CBN, which first localizes facial landmarks and then aligns LR faces based on the localized landmarks. However, when the inputs undergo large pose variations, CBN cannot localize facial landmarks accurately, and thus causes severe artifacts as seen in Fig. 7g. Fig. 6g shows that CBN cannot hallucinate authentic HR faces from the incorrect frontalized LR faces either. Furthermore, CBN super-resolves high-frequency facial details by combing

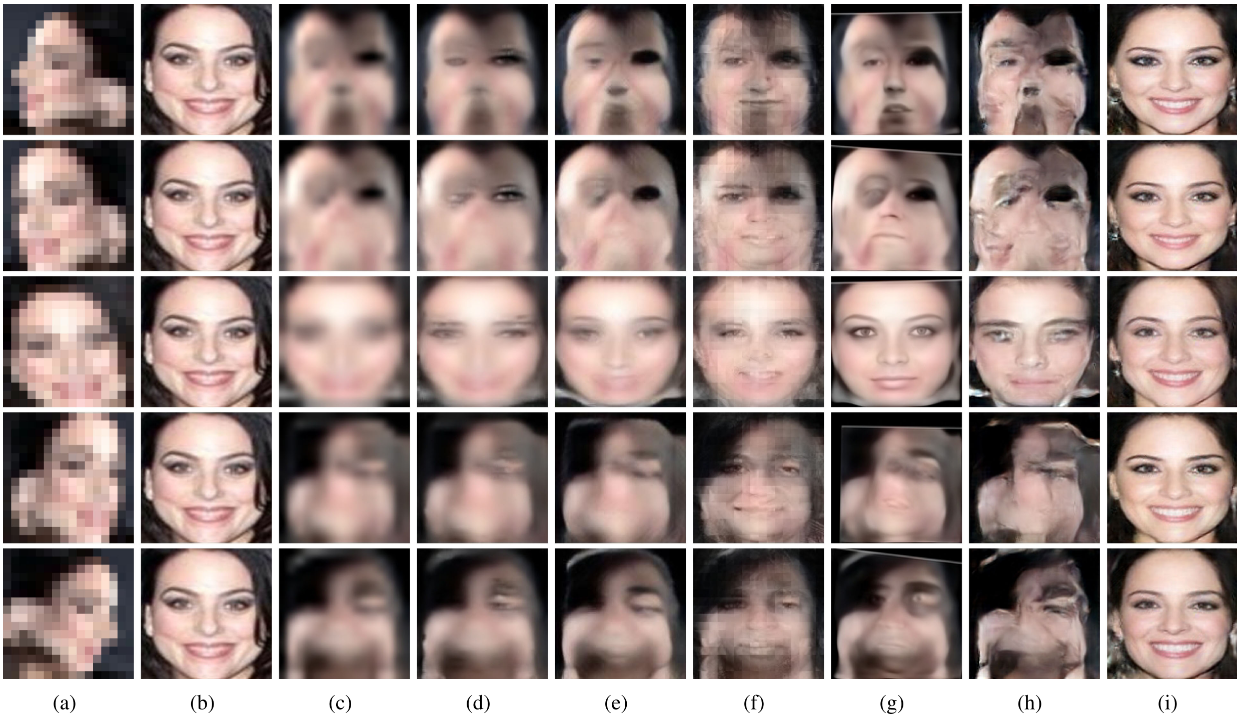


Fig. 6. Results of the state-of-the-art methods for *frontalization followed by hallucination*. The input faces are first frontalized by [1] and then hallucinated by different algorithms. Rows: +75°, +40°, 0°, −40°, and −75°. Columns: (a) Unaligned non-frontal LR inputs. (b) Original frontal HR images. (c) [1] + bicubic interpolation. (d) [1] + [59]. (e) [1] + [60]. (f) [1] + [17]. (g) [1] + [51]. (h) [1] + [2]. (i) Our method. Notice that, TANN does not need or use [1].

facial deformation bases. The bases are pre-defined and shared by all samples in CBN. When CBN fails to localize facial components accurately, it tends to output a mean face template composed by the bases as the high-frequency components of the upsampled faces. Therefore, the results seem very close to each other, as visible in Figs. 6g and 7g.

Yu and Porikli [2] propose a transformative discriminative autoencoder (TDAE) as an extension to [22] to upsample unaligned and noisy LR face images. TDAE interweaves deconvolutional and STN layers to align and super-resolve LR faces while employing a discriminative network that forces the generative network to produce sharper results.

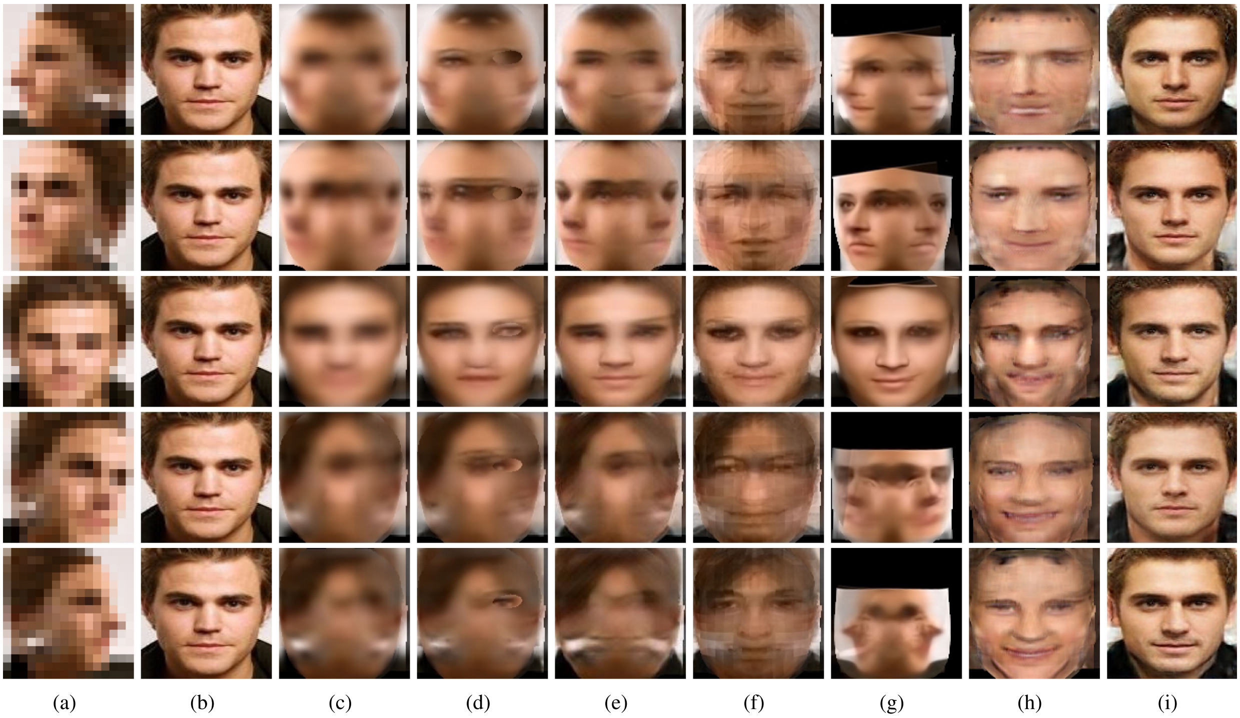


Fig. 7. Results of the state-of-the-art methods for *hallucination followed by frontalization* by [1]. Columns: (a) Unaligned non-frontal LR inputs. (b) Original frontal HR images. (c) Bicubic interpolation + [1]. (d) [59] + [1]. (e) [60] + [1]. (f) [17] + [1]. (g) [51] + [1]. (h) [2] + [1]. (i) Our method.

TABLE 1
Quantitative Evaluations on the Entire Test Dataset

H Method	F [1]+H			H+F [1]		
	PSNR	SSIM	PE	PSNR	SSIM	PE
Bicubic	20.99	0.80	2.36	20.41	0.79	2.48
VDSR [59]	21.04	0.80	2.27	20.47	0.79	2.39
SRGAN [60]	20.94	0.80	2.22	20.34	0.79	2.37
Ma et al. [17]	21.60	0.82	1.87	21.15	0.80	2.11
CBN [51]	20.61	0.79	2.33	19.40	0.77	2.76
TDAE [2]	20.68	0.79	2.28	19.89	0.77	2.52
Ours	**25.69**	**0.87**	**1.10**	**25.69**	**0.87**	**1.10**

However, TDAE can only hallucinate unaligned frontal faces rather than profile faces as demonstrated in Fig. 7h since it does not take out-of-plane rotations into account and the first decoder and encoder in TDAE are used for noise reduction rather than frontalization. Fig. 6h shows that TDAE cannot produce realistic HR faces due to the deteriorated LR facial patterns caused by the incorrect frontalization.

Our method reconstructs authentic facial details as shown in Figs. 6i and 7i and Figs. 8i and 9i. In the experiments, the face poses vary from -75° to $+75^{\circ}$. Since our transformer subnetwork can frontalize and align LR input faces more accurately, our upsampling subnetwork achieves superior reconstruction performance from the frontalized and aligned LR features.

5.2 Quantitative Comparisons to the SoA

We measure the reconstruction performance of all methods on the entire test dataset by the average PSNR and the structural similarity (SSIM) scores. Note that, when we hallucinate non-frontal faces, the hair and background regions may not be symmetric or the same compared to the original HR face images. Thus, for a fair comparison for all methods, we compute the PSNR and SSIM on the face regions.

We report results for two possible scenarios. In the first case, we first apply [1] to frontalize LR face images, and then super-resolve the frontalized LR images by the state-of-the-art SR/FSR methods (denoted as F+H). In the second case, we super-resolve LR face images first by the state-of-the-art SR/FSR methods and then frontalize the upsampled results by [1] (denoted as H+F). We apply STN_0 to align LR inputs uprightly in both cases. Table 1 shows that our method achieves the superior performance in comparison to the other methods, and outperforms the second best method over *4.0* dB in PSNR.

Table 2 indicates the PSNR and SSIM scores for different out-of-plane rotation degrees in the F+H and the H+F cases. In Table 2, the first and second numbers denote PSNR and SSIM scores respectively. As indicated in Table 2, first frontalizing and then upsampling faces can achieve slightly better results than first upsampling followed by frontalization. This also implies that it is easier to super-resolve frontal LR facial patterns than non-frontal ones. Because of the mirror symmetry operation in [1], the PSNR and SSIM scores of the other methods in the positive degrees are lower than those in the negative degrees, as seen in Table 2. However, our method does not have this effect and produces consistent PSNR scores in both negative and positive degrees. Furthermore, as the rotation degree increases, our method does not degrade like the other methods. From 0° to $\pm 75^{\circ}$, our performance only decreases 1.95 dB while the performance of the second best method decreases 3.75 dB.

In addition, as reported in our previous work [26], we also observe that blurry upsampled results may have higher PSNRs. Therefore, we introduce a perceptual error metric to measure the hallucination performance and the perceptual errors are more consistent with human perception. In particular, the perceptual errors are measured by the differences of feature maps between the hallucinated faces and their ground-truth ones, as indicated in Eqn. (2). As demonstrated in Tables 1, 2 and 3, our method achieves the lowest perceptual errors in comparison to other stat-of-the-art methods. This also implies that our method can attain more authentic frontalized upsampled HR face images.

We also conduct a user study. In the experiment, a cohort of twenty students are asked to rank the upsampled faces with respect to the ground-truth images. Since there are two strategies to obtain HR frontalized face images, we conduct the user study on these two scenarios separately.

TABLE 2
Quantitative Evaluations on Different Out-of-Plane Rotation Degrees

	H Methods	-75°	-40°	0°	$+40^{\circ}$	$+75^{\circ}$
F [1]+H	Bicubic	20.63 / 0.80 / 2.35	21.43 / 0.81 / 2.32	24.52 / 0.83 / 2.07	19.51 / 0.78 / 2.52	18.87 / 0.77 / 2.54
	VDSR [59]	20.69 / 0.80 / 2.23	21.47 / 0.81 / 2.23	24.59 / 0.84 / 1.90	19.54 / 0.78 / 2.49	18.90 / 0.77 / 2.50
	SRGAN [60]	20.58 / 0.80 / 2.21	21.34 / 0.80 / 2.21	24.53 / 0.83 / 1.76	19.41 / 0.78 / 2.46	18.81 / 0.77 / 2.46
	Ma et al. [17]	21.15 / 0.81 / 1.88	22.05 / 0.82 / 1.83	24.90 / 0.85 / 1.52	20.38 / 0.80 / 2.04	19.53 / 0.80 / 2.06
	CBN [51]	20.34 / 0.79 / 2.27	21.14 / 0.80 / 2.23	24.14 / 0.83 / 1.87	19.08 / 0.77 / 2.59	18.36 / 0.76 / 2.68
	TDAE [2]	20.44 / 0.79 / 2.26	20.69 / 0.79 / 2.30	23.13 / 0.82 / 1.80	19.74 / 0.78 / 2.50	19.43 / 0.78 / 2.52
H+F [1]	Bicubic	20.25 / 0.79 / 2.48	20.68 / 0.80 / 2.47	23.46 / 0.83 / 2.17	19.05 / 0.77 / 2.63	18.62 / 0.77 / 2.62
	VDSR [59]	20.41 / 0.80 / 2.38	20.83 / 0.80 / 2.39	23.43 / 0.83 / 1.98	19.04 / 0.77 / 2.64	18.66 / 0.77 / 2.59
	SRGAN [60]	20.36 / 0.79 / 2.34	20.69 / 0.79 / 2.40	23.12 / 0.82 / 1.92	18.98 / 0.77 / 2.61	18.53 / 0.77 / 2.55
	Ma et al. [17]	21.23 / 0.80 / 2.12	21.90 / 0.81 / 2.11	23.37 / 0.83 / 1.85	19.97 / 0.79 / 2.23	19.26 / 0.78 / 2.24
	CBN [51]	18.64 / 0.75 / 2.83	19.23 / 0.76 / 2.84	22.13 / 0.81 / 2.18	18.84 / 0.76 / 2.93	18.16 / 0.75 / 2.99
	TDAE [2]	19.35 / 0.77 / 2.59	19.97 / 0.77 / 2.56	22.62 / 0.80 / 2.18	19.36 / 0.77 / 2.58	18.13 / 0.76 / 2.69
	Ours $^{-}$	24.86 / 0.87 / 1.21	25.24 / 0.87 / 1.18	26.58 / 0.88 / 1.08	25.22 / 0.87 / 1.17	24.78 / 0.87 / 1.21
	Ours	**25.02 / 0.87 / 1.17**	**25.72 / 0.87 / 1.10**	**26.97 / 0.89 / 0.98**	**25.70 / 0.87 / 1.10**	**25.03 / 0.87 / 1.17**

TABLE 3
Quantitative Evaluations on the Frontal View

Method	Bicubic	VDSR [59]	SRGAN [60]	Ma et al. [17]	CBN [51]	TDAE [2]	Ours
PSNR	25.64	25.78	25.58	26.45	25.37	26.39	**26.97**
SSIM	0.86	0.86	0.85	0.88	0.86	0.87	**0.89**
PE	1.84	1.51	1.17	1.16	1.44	1.18	**0.98**

Specifically, in the first scenario all the face images are recovered by applying frontalization first and then hallucination, and in the second scenario all the images are reconstructed by employing hallucination first and then frontalization. In each scenario, we show twenty female faces as well as twenty male ones recovered by different methods to each user. Each user gives a score from 1 to 7 to different results. We average the scores for all the users. The average scores are illustrated in Fig. 15. As seen in Figs. 15a and 15b, all the users favor our method on the test images.

6 Discussions

6.1 Super-Resolving Different Levels of Downsampled Images

In order to super-resolve different levels of downsampled images, i.e., 2×, 4× and 8×, we need to modify our network slightly to accept images in larger resolutions. Due to the fixed size of the bottleneck layers of our network, merely increasing the number of layers of the encoder network does not necessarily improve the performance as the resolutions increase. Since increasing the size of the bottleneck of the network will increase the parameters of the network dramatically, and the network cannot be fed into GPU memory. Therefore, we employ skip connections between our encoder and decoder parts. Note that, we concatenate the feature maps of our encoder layers and their corresponding decoder layers rather than adding them. In this way, we can preserve more high-frequency details from inputs. The visual results for different resolutions are also shown in Fig. 10.

6.2 Comparisons with SoA on Face Recognition and Retrieval

It is important to notice that we do not claim our method is designed for face recognition for two reasons: (i) we do not explicitly incorporate an identification objective in our formulation, and (ii) it might seem fruitless to attempt recognizing people in such tiny images even for humans. However, we demonstrate that our hallucination method effectively facilitates the face recognition task in two difference scenarios: (1) we first use the hallucinated faces to train a face recognition network and then test its face recognition performance; (2) We use an off-the-shelf face recognition network which is trained on original HR face images, and then test its performance on our hallucinated face images.

For the first scenario, we use the standard faceNet [55] as the face recognition network and the same training protocols as indicated in [55]. We follow the standard divisions of the training and test datasets in the LFW benchmark to generate LR/HR pairs, as reported in [29]. The face recognition network is both trained and tested on the hallucinated faces by our network. Following the standard LFW face verification test protocol, we report the accuracy scores in Table 4. We also included another two baseline methods for more detailed comparisons. The first baseline network is trained and tested on the original HR faces, marked as HR, and the second baseline network is trained and tested on LR face images that are upsampled by bicubic interpolation to fit the resolution requirement of the network, marked as LR.

As indicated in Table 4, our method improves the face recognition performance by a large margin of 19.38 percent compared to the network that is only trained on LR face images. However, as seen in Table 4, the gap of the face recognition performance between the LR and the original HR faces is reduced by our method. We also test face recognition performance on different levels of downsampling, i.e., 2×, 4× and 8×. As indicated in Table 4, as the input resolutions increase, the face recognition performance improves.

For the second scenario, we employ a state-of-the-art pretrained face recognition network (SphereFaceNet [61]) to conduct standard face recognition tests on original HR faces, aggressive downsampled LR faces and hallucinated HR faces from LR ones by different methods. The face recognition performance is also evaluated on standard LFW face verification benchmark [29]. We demonstrate the face recognition performance in Table 5, where the performance on original HR faces is marked as HR, the performance on

TABLE 4
Results of Different Face Recognition Networks Trained on Different Source Images

Sources	HR	LR	8×	4×	2×
Accuracy	85.32%	62.15%	81.53%	83.33%	84.51%

TABLE 5
Face Recognition Results for Different Methods

H Method	Accuracy	
	F [1]+H	H+F [1]
Bicubic	66.57%	65.43%
VDSR [59]	64.15%	69.77%
SRGAN [60]	68.88%	70.92%
Ma et al. [17]	65.55%	68.83%
CBN [51]	65.05%	64.72%
TDAE [2]	64.02%	65.96%
HR	96.02%	
LR	77.27%	
Ours 8×	82.32%	
Ours 4×	86.18%	
Ours 2×	92.25%	

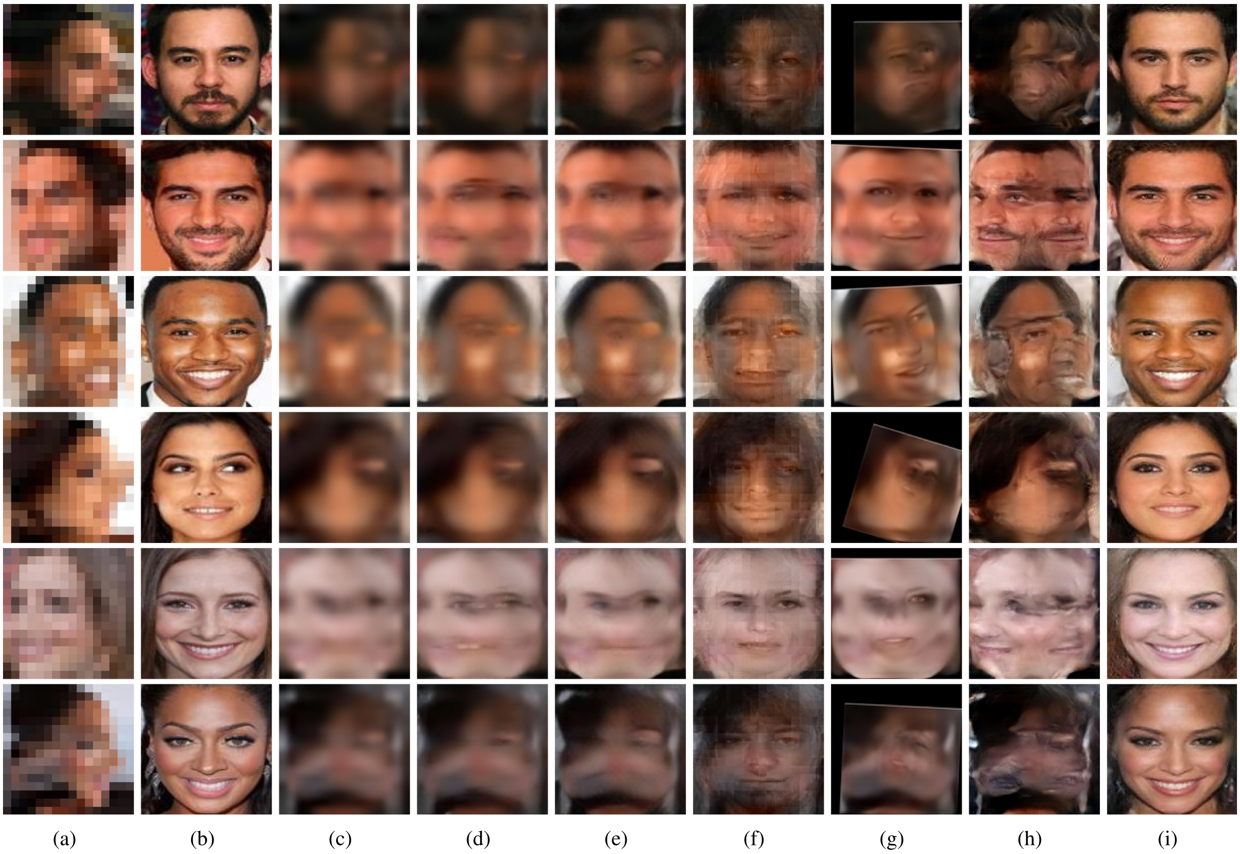


Fig. 8. Results of the state-of-the-art methods for *frontalization followed by hallucination*. Columns: (a) Unaligned non-frontal LR inputs. (b) Original frontal HR images. (c) [1] + bicubic interpolation. (d) [1] + [59]. (e) [1] + [60]. (f) [1] + [17]. (g) [1] + [51]. (h) [1] + [2]. (i) Our method.

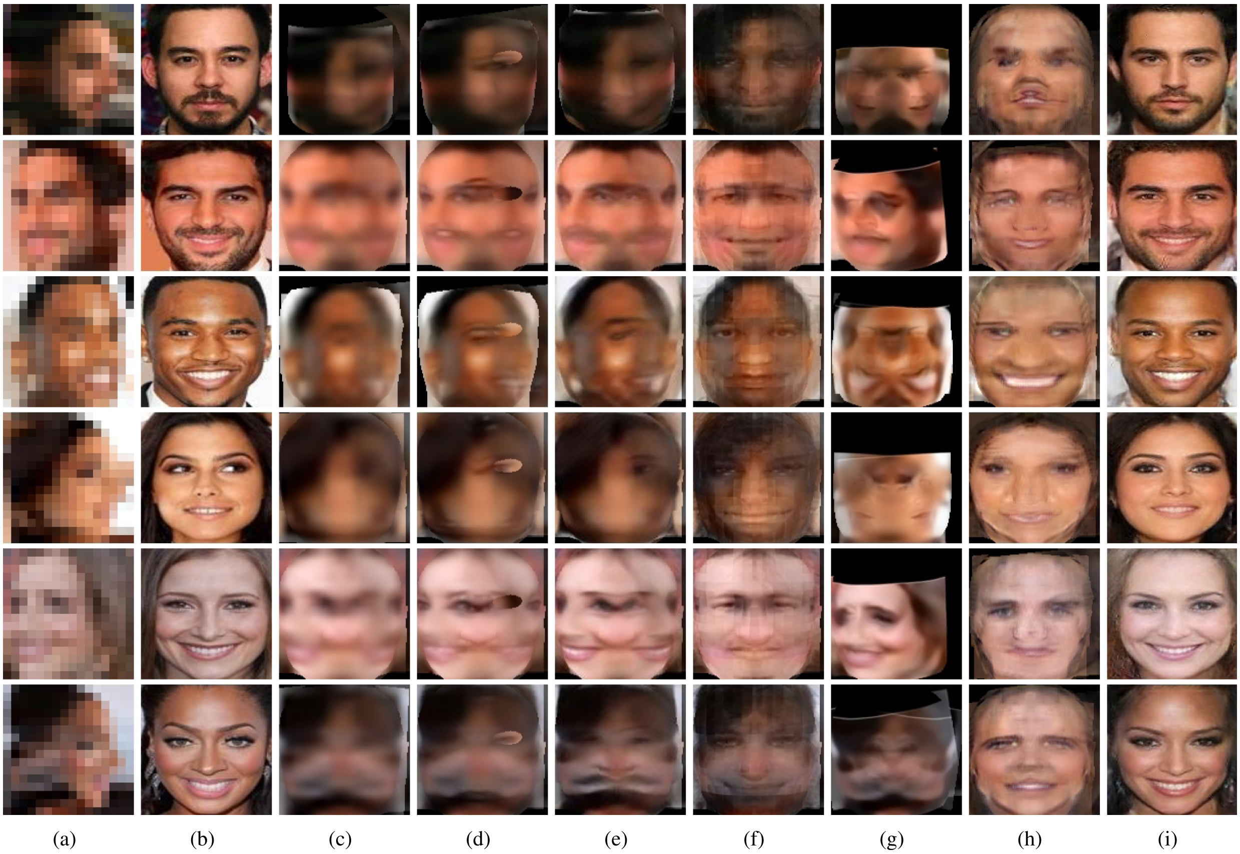


Fig. 9. Results of the state-of-the-art methods for *hallucination followed by frontalization* by [1]. Columns: (a) Unaligned non-frontal LR inputs. (b) Original frontal HR images. (c) Bicubic interpolation + [1]. (d) [59] + [1]. (e) [60] + [1]. (f) [17] + [1]. (g) [51] + [1]. (h) [2] + [1]. (i) Our method.

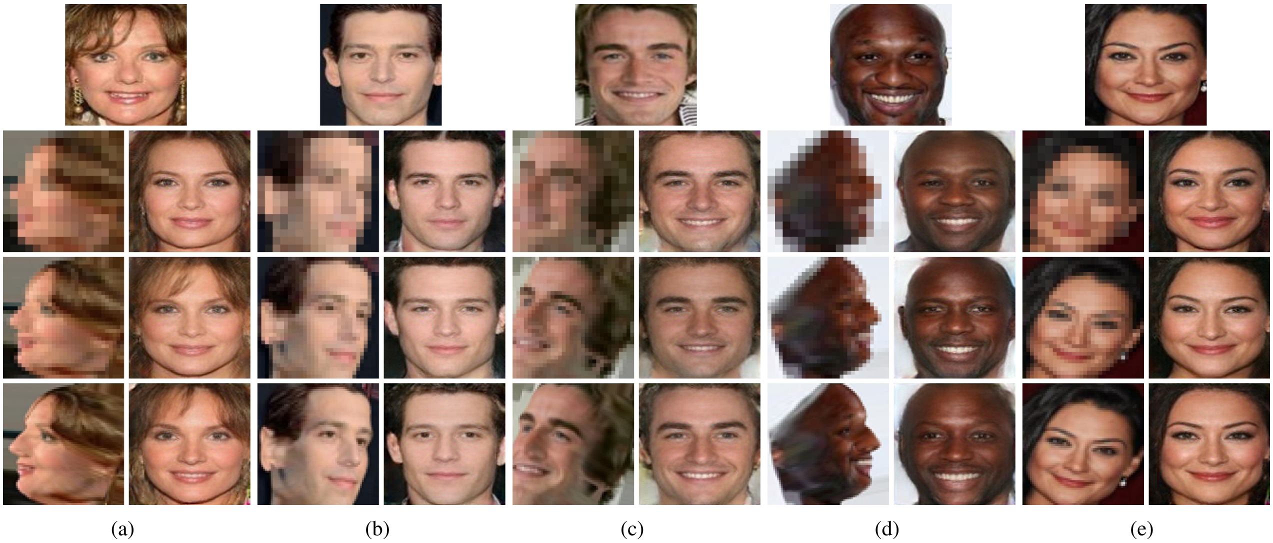


Fig. 10. Illustrations of super-resolving and frontalizing face images in different resolutions by our method. First row: Ground-truth frontal HR face images. Second row: Input LR faces (left) and our results with a magnification factor 8× (right). Third row: Input LR faces and our results with a magnification factor 4×. Fourth row: input LR faces and our results with a magnification factor 2×.

LR faces is marked as LR, and the combinations of the frontalization method and different upsampling methods are also listed. As shown in Table 5, our method improves the face recognition performance significantly compared to the other methods in both scenarios. Note that, since our cropped original HR faces might not be aligned to the positions of HR faces used for training the network of [61], the face recognition rate for original HR faces decreases slightly. Furthermore, we also demonstrate that when the resolutions of input images increase, the face recognition performance of our method improves as seen in Table 5.

Furthermore, to our advantage, our method achieves significant improvement in face retrieval performance as shown in Table 6. We use an off-the-shelf deep face recognition model [62] to evaluate the performance of all the methods. First, we randomly choose 100 frontal faces from the test data as our gallery. We generate their corresponding four LR non-frontal images, and employ six algorithms listed above to hallucinate the frontal HR faces on both F+H and H+F scenarios. Following the standard protocol in [62], we compute the accuracy score based on whether the correct person is included within the top-5 candidates (thus, the probability of random selection is 5 percent). Here, we notice that directly using off-the-shelf face recognition is inappropriate to measure the similarity between generated HR faces and real HR faces because there is still a domain gap between them. For instance, the features of real faces may be different from those of generated HR faces. In order to mitigate the domain gap, we train an autoencoder by using the same protocol of training TANN to transfer HR real faces to the domain where generated HR images lie in. In this way, we can significantly reduce the domain gap.

TABLE 6
Face Retrieval Results for Different Methods

H Method	Accuracy	
	F [1]+H	H+F [1]
Bicubic	5.8%	6.6%
VDSR [59]	7.0%	8.0%
SRGAN [60]	6.0%	9.0%
Ma et al. [17]	6.0%	9.0%
CBN [51]	6.2%	8.2%
TDAE [2]	7.2%	5.6%
Ours	**86.7%**	

As seen in Table 6, we improve the face retrieval accuracy with a large margin of 77.7 percent. This also implies that our method is able to preserve the appearance similarity rather than generating averaged HR faces when frontalizing and hallucinating LR faces.

6.3 Comparisons with SoA on Frontal Faces

Because we do not distinguish the views of LR faces deliberately before frontalization, the frontalization method [1] is applied to all the views of LR faces. As shown in Fig. 6, using the face frontalization method [1] distorts LR input faces due to the erroneous localization of facial components and its symmetrizing operations. Therefore, the super-resolution performance of frontal LR faces degrades dramatically.

For a fair comparison, we also include an evaluation for the frontal view case in Fig. 11 where the frontalization is not employed. As shown in Table 3, our method still outperforms all others in the frontal view case. Note that, our previous method TDAE [2] intends to increase the depth of its decoder to achieve better super-resolution performance but is limited by the GPU memory. In contrast, our network employs an autoencoder, i.e., our transformer subnetwork, before upsampling, and thus it does not require as much memory as TDAE yet achieves better performance. This also demonstrates that our transformer subnetwork can not only frontalize LR profile faces but also improve super-resolution performance.

6.4 Influence of Different Losses

Table 7 indicates the influences of different losses on the performance quantitatively. As indicated in Fig. 4f and Table 7, the feature-wise loss not only improves the visual

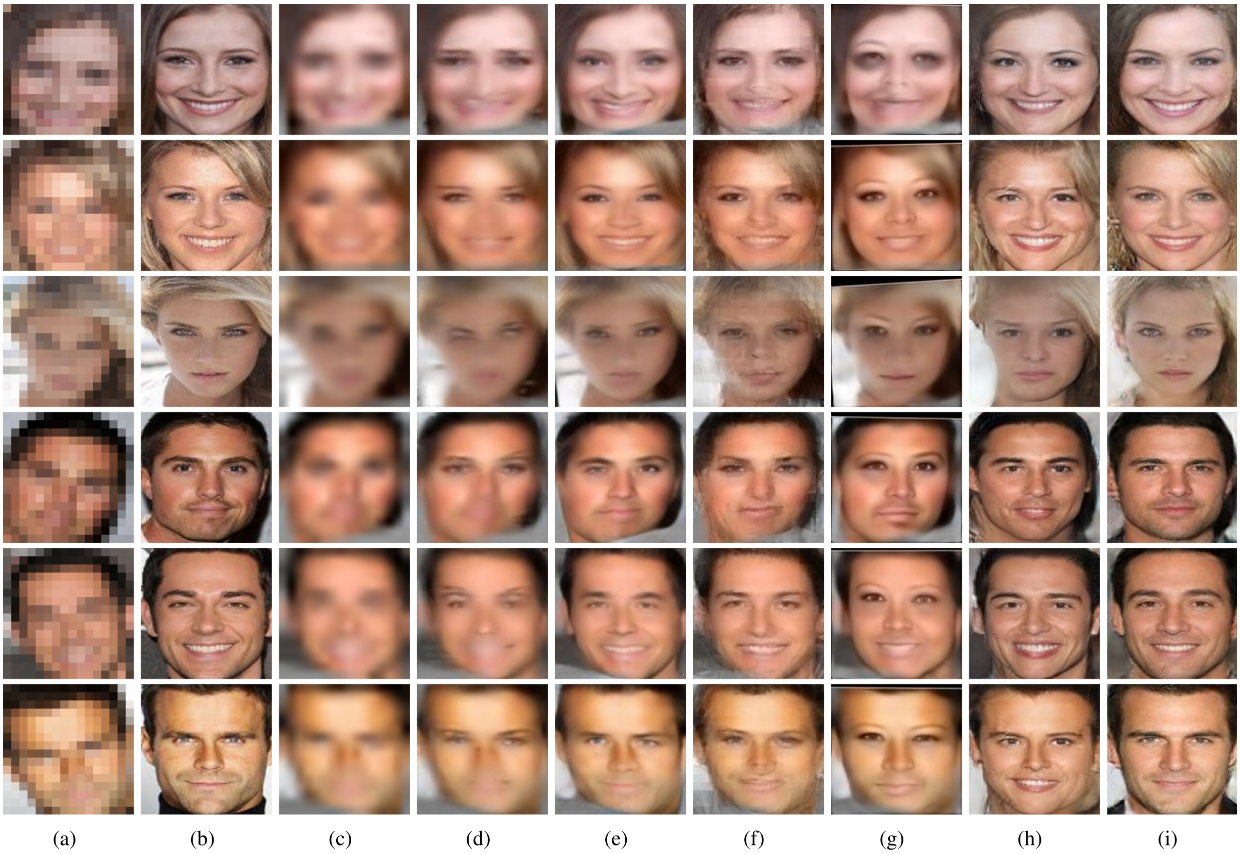


Fig. 11. Results of the state-of-the-art face hallucination methods for frontal LR faces. Columns: (a) Unaligned non-frontal LR inputs. (b) Original frontal HR images. (c) Bicubic interpolation. (d) [59]. (e) [60]. (f) [17]. (g) [51]. (h) [2]. (i) Our method.

quality but also increases the quantitative results. The adversarial loss makes the hallucinated faces sharper and more realistic, as shown in Fig. 4f. As illustrated in Table 7, using adversarial loss is also able to force the super-resolved face images to be frontal and thus improves the super-resolution performance.

As demonstrated in Table 7, using our triplet loss improves the final results. Because our triplet loss forces the LR profile faces to be close to their frontal ones in the latent subspace, the upsampled HR frontalized faces are more similar to their frontal ground-truths. Furthermore, we also illustrate the quantitative results without using our triplet loss for different out-of-plane rotation degrees in Table 2, marked as Ours $^{-}$. This experiment confirms that the triplet loss does not degrade the performance of upsampling frontal faces but improves the SR performance of LR profile faces. In addition, our triplet loss is able to reduce the reconstruction loss of LR profile faces earlier in the transformer subnetwork rather than spreading the loss through the entire upsampling network. Thus, the upsampling subnetwork can focus on learning mappings between LR and HR facial patterns as suggested in [22]. With the help of the triplet loss, we can even achieve better super-resolution performance on LR frontal faces, as indicated in Table 2.

TABLE 7
Quantitative Evaluations on the Influence of Different Losses

	w/o $\mathcal{L}_{tri}$			w/ $\mathcal{L}_{tri}$		
	$\mathcal{L}_{pix}$	$\mathcal{L}_{pix+feat}$	$\mathcal{L}_{pix+feat+\mathcal{T}}$	$\mathcal{L}_{pix}$	$\mathcal{L}_{pix+feat}$	$\mathcal{L}_{pix+feat+\mathcal{T}}$
PSNR	25.01	25.17	25.33	25.19	25.33	25.69
SSIM	0.87	0.87	0.87	0.88	0.87	0.87
PE	1.32	1.17	1.17	1.31	1.05	1.10

6.5 Performance on Faces Beyond 3D Models

Although our method is trained on a dataset of LR non-frontal and HR frontal image pairs synthesized by using a single 3D face model, our method can be effectively generalized to faces beyond the 3D model and the poses used in the training stage. To demonstrate this, we randomly choose face images from CelebA excluding the frontal faces used for generating our training dataset. Then we spatially deform, i.e., 2D transformation including rotations, translations and scale changes, and downsample these images to obtain LR face samples. The synthesized LR faces do not share 3D shapes or poses with the examples used in the training dataset, and thus these samples are much more challenging. As shown in Fig. 12, our network can hallucinate and frontalize such randomly chosen images, demonstrating it is not restricted to these five poses and certain models. Three reasons may account for this phenomenon: (1) When generating our dataset, the selected faces used for generating profile faces are not strictly frontal ones, which increases the variety of the training poses. (2) The differences between different HR faces become less obvious in LR faces, and faces in different poses can be approximated by one of the five poses in very low resolutions. (3) In the

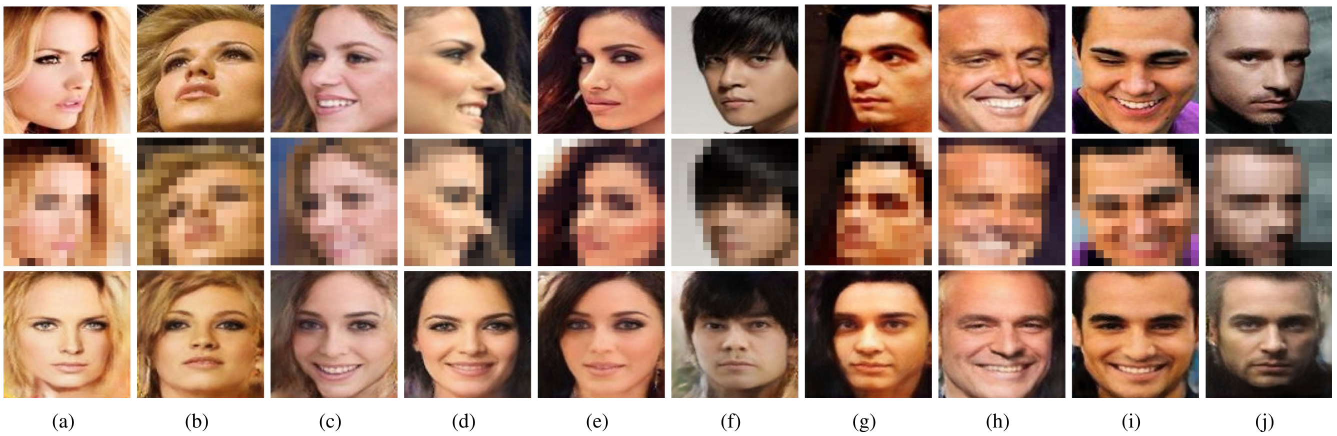


Fig. 12. Results on LR face images beyond 3D model and training poses. Top row: Real HR images. Middle row: Unaligned LR images. Bottom row: Our frontalized and hallucinated results.

process of encoding LR faces to latent representations, the max pooling layers also reduce the differences of 3D models and poses between the training and test LR faces.

We also apply our network to real LR face images chosen from the WiderFace dataset [63], where LR faces are captured in the wild. Notice that the real LR faces are even blurrier than our training samples. Our super-resolved results are shown in Fig. 13. Since our network does not need to select one specific model for a particular angle, our method does not require estimation of the face pose angles explicitly. Instead, our method frontalizes and hallucinates LR profile faces in different angles by a single network.

6.6 Super-Resolving LR Faces Without Frontalization

Since our method is an extension of our previous face super-resolution methods [2], [22], [26], our network can be also applied to super-resolve LR face images without frontalizing them. To this end, we use the ground-truth HR faces that have the same poses as the input LR ones as supervision and remove the triplet loss in training. As seen in Fig. 14, after retraining our network, our method can effectively super-resolve LR faces in different poses, similar to our previous methods [2], [26].

6.7 Limitations

Since our method uses a generic face model to generate faces in different poses, we do not contain different expressions in the training dataset. Therefore, our network does not account for different facial expressions. Furthermore, limited by the generic 3D face model, we do not model eye-glasses or sun-glasses in the training dataset either. When frontalizing occluded regions, general occluded regions and sun-glasses should yield different frontalization results due to the symmetry of sun-glasses and asymmetry of general occlusions. This may introduce further ambiguity in the frontalization process without exploiting any high-level semantic information. Our training dataset is generated from face images captured in normal illumination conditions where facial landmarks can be detected for generating different poses. Since facial landmark detectors may fail to localize landmarks accurately under extreme illumination conditions and the

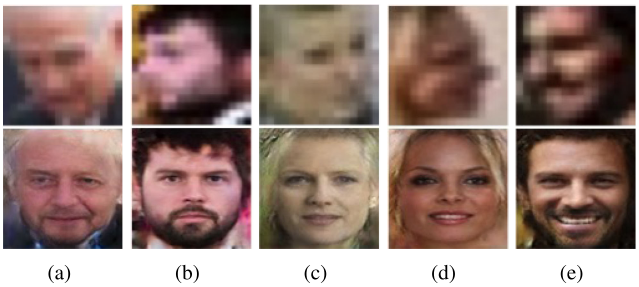


Fig. 13. Results on real LR face images. Top row: Real LR images. Bottom row: Our frontalized and hallucinated results.

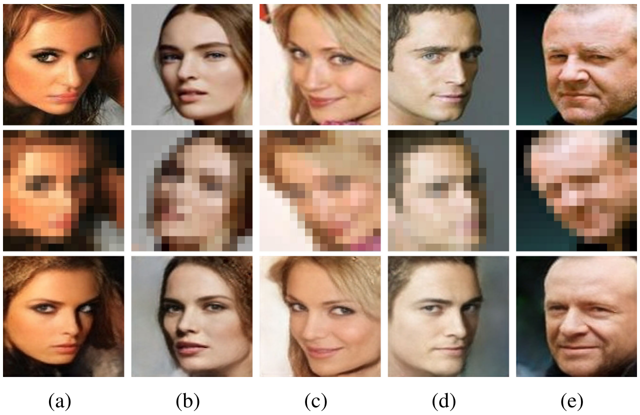


Fig. 14. Super-resolving LR faces without frontalization by our network. Top row: Ground-truth HR images. Middle row: LR face images. Bottom row: Our upsampled results.

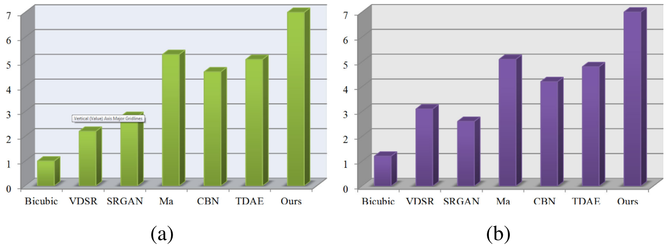


Fig. 15. Evaluation of user study on the test images. (a) Average scores of different methods for frontalization followed by hallucination. (b) Average scores of different methods for hallucination followed by frontalization.

generated faces by the 3D model may suffer severe artifacts, we do not contain those faces for training. Besides, since the illuminations on the faces are not symmetric, it is very challenging to frontalize realistic illumination conditions. Thus, our method does not tackle such face images acquired under extreme illumination conditions.

7 Conclusion

We introduced a transformative adversarial network to upsample and frontalize very low-resolution unaligned face images simultaneously in an end-to-end fashion. Our network is able to learn how to frontalize and align LR faces while upsampling 8×. Benefiting from our proposed triplet loss, we are able to enforce LR profile faces to be close to their frontal counterparts in the latent subspace and thus achieve better frontalization performance. With the help of the intra-class discriminative information and the feature constraints, our network generates realistic facial details.

Acknowledgments

This work was supported by the Australian Research Council's Discovery Projects funding scheme (project DP150104645), Australian Research Council Centre of Excellence for Robotic Vision (project number CE140100016) and King Abdullah University of Science and Technology (KAUST) Office of Sponsored Research.

References

[1] T. Hassner, S. Harel, E. Paz, and R. Enbar, "Effective face frontalization in unconstrained images," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2015, pp. 4295–4304.

[2] X. Yu and F. Porikli, "Hallucinating very low-resolution unaligned and noisy face images by transformative discriminative autoencoders," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 3760–3768.

[3] M. Jaderberg, K. Simonyan, A. Zisserman, et al., "Spatial transformer networks," in *Proc. Int. Conf. Neural Inf. Process. Syst.*, 2015, pp. 2017–2025.

[4] X. Zhu and D. Ramanan, "Face detection, pose estimation, and landmark localization in the wild," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2012, pp. 2879–2886.

[5] V. Blanz and T. Vetter, "A morphable model for the synthesis of 3D faces," in *Proc. 26th Annu. Conf. Comput. Graph. Interactive Techn.*, 1999, pp. 187–194.

[6] F. Yang, J. Wang, E. Shechtman, L. Bourdev, and D. Metaxas, "Expression flow for 3D-aware face component transfer," *ACM Trans. Graph.*, vol. 30, no. 4, 2011, Art. no. 60.

[7] T. Hassner, "Viewing real-world faces in 3D," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2013, pp. 3607–3614.

[8] Y. Taigman, M. Yang, M. Ranzato, and L. Wolf, "DeepFace: Closing the gap to human-level performance in face verification," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2014, pp. 1701–1708.

[9] C. Sagonas, Y. Panagakis, S. Zafeiriou, and M. Pantic, "Robust statistical face frontalization," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2015, pp. 3871–3879.

[10] J. Thies, M. Zollhöfer, M. Stamminger, C. Theobalt, and M. Nießner, "Face2Face: Real-time face capture and reenactment of RGB videos," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2016, pp. 2387–2395.

[11] S. Baker and T. Kanade, "Hallucinating faces," in *Proc. 4th IEEE Int. Conf. Autom. Face Gesture Recognit.*, 2000, pp. 83–88.

[12] S. Baker and T. Kanade, "Limits on super-resolution and how to break them," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 24, no. 9, pp. 1167–1183, Sep. 2002.

[13] C. Liu, H. Shum, and C. Zhang, "A two-step approach to hallucinating faces: Global parametric model and local nonparametric model," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, 2001, pp. 192–198.

[14] X. Wang and X. Tang, "Hallucinating face by eigentransformation," *IEEE Trans. Syst. Man Cybern. Part C: Appl. Rev.*, vol. 35, no. 3, pp. 425–434, Aug. 2005.

[15] C. Liu, H. Y. Shum, and W. T. Freeman, "Face hallucination: Theory and practice," *Int. J. Comput. Vis.*, vol. 75, no. 1, pp. 115–134, 2007.

[16] P. H. Hennings-Yeomans, S. Baker, and B. V. Kumar, "Simultaneous super-resolution and feature extraction for recognition of low-resolution faces," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, 2008, pp. 1–8.

[17] X. Ma, J. Zhang, and C. Qi, "Hallucinating face by position-patch," *Pattern Recognit.*, vol. 43, no. 6, pp. 2224–2236, 2010.

[18] J. Yang, J. Wright, T. S. Huang, and Y. Ma, "Image super-resolution via sparse representation," *IEEE Trans. Image Process.*, vol. 19, no. 11, pp. 2861–2873, Nov. 2010.

[19] Y. Li, C. Cai, G. Qiu, and K. M. Lam, "Face hallucination based on sparse local-pixel structure," *Pattern Recognit.*, vol. 47, no. 3, pp. 1261–1270, 2014.

[20] S. Kolouri and G. K. Rohde, "Transport-based single frame super resolution of very low resolution face images," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, 2015, pp. 4876–4884.

[21] N. Wang, D. Tao, X. Gao, X. Li, and J. Li, "A comprehensive survey to face hallucination," *Int. J. Comput. Vis.*, vol. 106, no. 1, pp. 9–30, 2014.

[22] X. Yu and F. Porikli, "Ultra-resolving face images by discriminative generative networks," in *Proc. Eur. Conf. Comput. Vis.*, 2016, pp. 318–333.

[23] X. Yu and F. Porikli, "Imagining the unimaginable faces by deconvolutional networks," *IEEE Trans. Image Process.*, vol. 27, no. 6, pp. 2747–2761, Jun. 2018.

[24] I. Goodfellow, J. Pouget-Abadie, and M. Mirza, "Generative adversarial networks," in *Proc. Int. Conf. Neural Inf. Process. Syst.*, 2014, pp. 2672–2680.

[25] E. Denton, S. Chintala, A. Szlam, and R. Fergus, "Deep generative image models using a Laplacian pyramid of adversarial networks," in *Proc. Int. Conf. Neural Inf. Process. Syst.*, 2015, pp. 1486–1494.

[26] X. Yu and F. Porikli, "Face hallucination with tiny unaligned images by transformative discriminative neural networks," in *Proc. 31st AAAI Conf. Artif. Intell.*, 2017, pp. 4327–4333.

[27] X. Xu, D. Sun, J. Pan, Y. Zhang, H. Pfister, and M.-H. Yang, "Learning to super-resolve blurry face and text images," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 251–260.

[28] J. Johnson, A. Alahi, and L. Fei-Fei, "Perceptual losses for real-time style transfer and super-resolution," in *Proc. Eur. Conf. Comput. Vis.*, 2016, pp. 694–711.

[29] G. B. Huang, M. Ramesh, T. Berg, and E. Learned-Miller, "Labeled faces in the wild: A database for studying face recognition in unconstrained environments," Univ. Massachusetts, Amherst, MA, USA, Tech. Rep. 07–49, 2007.

[30] Z. Liu, P. Luo, X. Wang, and X. Tang, "Deep learning face attributes in the wild," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2015, pp. 3730–3738.

[31] I. Masi, A. T. Tran, T. Hassner, J. T. Leksut, and G. Medioni, "Do we really need to collect millions of faces for effective face recognition?" in *Proc. Eur. Conf. Comput. Vis.*, 2016, pp. 579–596.

[32] R. Dovgard and R. Basri, "Statistical symmetric shape from shading for 3D structure recovery of faces," in *Proc. Eur. Conf. Comput. Vis.*, 2004, pp. 99–113.

[33] A. Asthana, T. K. Marks, M. J. Jones, K. H. Tieu, and M. Rohith, "Fully automatic pose-invariant face recognition via 3D pose normalization," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2011, pp. 937–944.

[34] X. Zhu, Z. Lei, J. Yan, D. Yi, and S. Z. Li, "High-fidelity pose and expression normalization for face recognition in the wild," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2015, pp. 787–796.

[35] Z. Zhu, P. Luo, X. Wang, and X. Tang, "Recover canonical-view faces in the wild with deep neural networks," arXiv preprint arXiv:1404.3543, 2014.

[36] J. Yim, H. Jung, B. Yoo, C. Choi, D. Park, and J. Kim, "Rotating your face using multi-task deep neural network," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2015, pp. 787–796.

[37] L. Tran, X. Yin, and X. Liu, "Disentangled representation learning GAN for pose-invariant face recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2017, Art. no. 7.

[38] F. Cole, D. Belanger, D. Krishnan, A. Sarna, I. Mosseri, and W. T. Freeman, "Face synthesis from facial identity features," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 3703–3712.

[39] R. Huang, S. Zhang, T. Li, and R. He, "Beyond face rotation: Global and local perception GAN for photorealistic and identity preserving frontal view synthesis," arXiv:1704.04086, 2017.

[40] X. Yin, X. Yu, K. Sohn, X. Liu, and M. Chandraker, "Towards large-pose face frontalization in the wild," in *Proc. IEEE Int. Conf. Comput. Vis.*, Oct. 2017, pp. 4010–4019.

[41] A. T. Tran, T. Hassner, I. Masi, and G. Medioni, "Regressing robust and discriminative 3D Morphable models with a very deep neural network," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 1493–1502.

[42] F.-J. Chang, A. T. Tran, T. Hassner, I. Masi, R. Nevatia, and G. Medioni, "FacePoseNet: Making a case for landmark-free face alignment," in *Proc. IEEE Int. Conf. Comput. Vis. Workshop*, 2017, pp. 1599–1608.

[43] M. F. Tappen and C. Liu, "A Bayesian approach to alignment-based image hallucination," in *Proc. Eur. Conf. Comput. Vis.*, 2012, pp. 236–249.

[44] C. Liu, J. Yuen, and A. Torralba, "SIFT flow: Dense correspondence across scenes and its applications," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 33, no. 5, pp. 978–994, May 2011.

[45] C. Y. Yang, S. Liu, and M. H. Yang, "Structured face hallucination," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, 2013, pp. 1099–1106.

[46] C.-Y. Yang, S. Liu, and M.-H. Yang, "Hallucinating compressed face images," *Int. J. Comput. Vis.*, vol. 126, pp. 597–614, 2018.

[47] E. Zhou and H. Fan, "Learning face hallucination in the wild," in *Proc. 29th AAAI Conf. Artif. Intell.*, 2015, pp. 3871–3877.

[48] H. Huang, R. He, Z. Sun, and T. Tan, "Wavelet-SRNet: A wavelet-based CNN for multi-scale face super resolution," in *Proc. IEEE Int. Conf. Comput. Vis.*, Oct. 2017, pp. 1698–1706.

[49] R. Dahl, M. Norouzi, and J. Shlens, "Pixel recursive super resolution," in *Proc. Int. Conf. Comput. Vis.*, 2017, pp. 5439–5448.

[50] A. Van Den Oord, N. Kalchbrenner, and K. Kavukcuoglu, "Pixel recurrent neural networks," in *Proc. IEEE Int. Conf. Int. Conf. Mach. Learn.*, 2016, pp. 1747–1756.

[51] S. Zhu, S. Liu, C. C. Loy, and X. Tang, "Deep cascaded Bi-network for face hallucination," in *Proc. Eur. Conf. Comput. Vis.*, 2016, pp. 614–630.

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

[53] A. Bulat and G. Tzimiropoulos, "Super-FAN: Integrated facial landmark localization and super-resolution of real-world low resolution faces in arbitrary poses with GANs," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 109–117.

[54] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," arXiv:1409.1556, 2014.

[55] F. Schroff, D. Kalenichenko, and J. Philbin, "FaceNet: A unified embedding for face recognition and clustering," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2015, pp. 815–823, http://www.cs.toronto.edu/~hinton/coursera/lecture6/lec6.pdf.

[56] G. Hinton, "Neural networks for machine learning lecture 6a: Overview of mini-batch gradient descent reminder: The error surface for a linear neuron,", http://www.cs.toronto.edu/~hinton/coursera/lecture6/lec6.pdf.

[57] A. Radford, L. Metz, and S. Chintala, "Unsupervised representation learning with deep convolutional generative adversarial networks," arXiv:1511.06434, pp. 1–15, 2015.

[58] S. Ioffe and C. Szegedy, "Batch normalization: Accelerating deep network training by reducing internal covariate shift," in *Proc. 32nd Int. Conf. Mach. Learn.*, 2015, pp. 448–456.

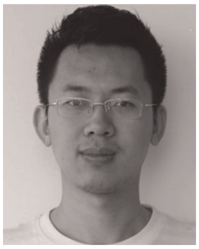
[59] J. Kim, J. Kwon Lee, and K. Mu Lee, "Accurate image super-resolution using very deep convolutional networks," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, 2016, pp. 1646–1654.

[60] C. Ledig, L. Theis, F. Huszár, J. Caballero, A. Cunningham, A. Acosta, A. Aitken, A. Tejani, J. Totz, Z. Wang, et al., "Photo-realistic single image super-resolution using a generative adversarial network," arXiv:1609.04802, 2016.

[61] W. Liu, Y. Wen, Z. Yu, M. Li, B. Raj, and L. Song, "SphereFace: Deep hypersphere embedding for face recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2017, Art. no. 1.

[62] O. M. Parkhi, A. Vedaldi, and A. Zisserman, "Deep face recognition," in *Proc. Brit. Mach. Vis. Conf.*, 2015, pp. 41.1–41.12.

[63] S. Yang, P. Luo, C.-C. Loy, and X. Tang, "WIDER FACE: A face detection benchmark," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2016, pp. 5525–5533.



Xin Yu received the BS degree in electronic engineering from the University of Electronic Science and Technology of China, Chengdu, China, in 2009, the PhD degree from the Department of Electronic Engineering, Tsinghua University, Beijing, China, in 2015, and the PhD degree from the College of Engineering and Computer Science, Australian National University, Canberra, Australia, in 2018. He is currently a research fellow with Australian National University. His interests include computer vision and image processing.

Fatemeh Shiri received the BS degree in electronic engineering from Razi University, Iran, and the MS degree from Tarbiat Modares University, Iran. She is currently working toward the PhD degree in the College of Engineering and Computer science, Australian National University, Australia. Her research interests include deep learning, computer vision, and image processing.

Bernard Ghanem received the PhD degree in electrical and computer engineering from the University of Illinois at Urbana-Champaign (UIUC), in 2010. He is currently an associate professor with the King Abdullah University of Science and Technology (KAUST), where he leads the Image and Video Understanding Lab (IVUL). His research interests focus on designing, implementing, and analyzing approaches to address computer vision problems (e.g., image understanding, object tracking and action recognition/detection in video), especially at large-scale.

Fatih Porikli received the PhD degree from New York University (NYU). He is a profe ssor with the Research School of Engineering, Australian National University, Canberra, Australia. Previously, he served as a distinguished research scientist with Mitsubishi Electric Research Laboratories, Cambridge. He has contributed broadly to object detection, motion estimation, tracking, image-based representations, and video analytics. He is the coeditor of two books on *Video Analytics for Business Intelligence* and Handbook on *Background Modeling and Foreground Detection for Video Surveillance*. He is an associate editor of five journals. His publications won four Best Paper Awards and he has received the RD100 Award in the Scientist of the Year category in 2006. He served as the general and program chair of numerous IEEE conferences in the past. He has 66 granted patents. He is a fellow of the IEEE.