Semantic Face Hallucination: Super-Resolving Very Low-Resolution Face Images with Supplementary Attributes

Xin Yu, Basura Fernando, Richard Hartley, *Fellow, IEEE*, and Fatih Porikli, *Fellow, IEEE*

Abstract—Given a tiny face image, existing face hallucination methods aim at super-resolving its high-resolution (HR) counterpart by learning a mapping from an exemplary dataset. Since a low-resolution (LR) input patch may correspond to many HR candidate patches, this ambiguity may lead to distorted HR facial details and wrong attributes such as gender reversal and rejuvenation. An LR input contains low-frequency facial components of its HR version while its residual face image, defined as the difference between the HR ground-truth and interpolated LR images, contains the missing high-frequency facial details. We demonstrate that supplementing residual images or feature maps with additional facial attribute information can significantly reduce the ambiguity in face super-resolution. To explore this idea, we develop an attribute-embedded upsampling network, which consists of an upsampling network and a discriminative network. The upsampling network is composed of an autoencoder with skip-connections, which incorporates facial attribute vectors into the residual features of LR inputs at the bottleneck of the autoencoder, and deconvolutional layers used for upsampling. The discriminative network is designed to examine whether super-resolved faces contain the desired attributes or not and then its loss is used for updating the upsampling network. In this manner, we can super-resolve tiny (16×16 pixels) unaligned face images with a large upscaling factor of $8\times$ while reducing the uncertainty of one-to-many mappings remarkably. By conducting extensive evaluations on a large-scale dataset, we demonstrate that our method achieves superior face hallucination results and outperforms the state-of-the-art.

Index Terms—Face, super-resolution, hallucination, attribute

1 Introduction

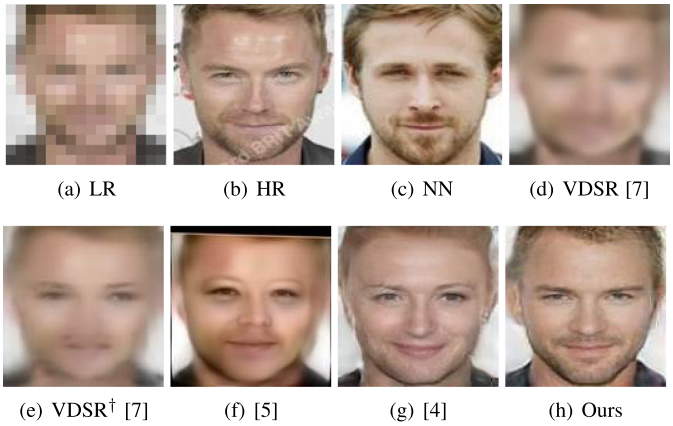
Face images provide important information for human visual perception as well as computer analysis [1], [2]. Depending on the imaging conditions, the resolution of a face area may be unfavorably low, thus raising a critical issue that would directly impede our understanding. Motivated by this challenge, recovering high-resolution (HR) face images from their low-resolution (LR) counterparts, also known as face hallucination, has received increasing attention recently [3], [4], [5], [6]. State-of-the-art face hallucination methods try to explore and utilize image domain priors for super-resolution. Even though they are trained on large-scale datasets benefiting from the development of deep learning techniques, ill-posed nature of the problem, which induces inherent ambiguities such as one-to-many correspondence between a given LR face and its possible HR counterparts, would still lead to drastically flawed outputs especially when the magnification factor is very large.

For instance, as shown in Fig. 1, hallucinated details generated by the state-of-the-art face super-resolution methods [4], [5] are semantically and perceptually inconsistent with the ground-truth HR image, and inaccuracies range from unnatural blur to attribute mismatches including the wrong facial hair and mixed gender features just to count a few. Note that Zhu et al.'s method [5], dubbed CBN, exploits facial structure information to super-resolve facial components while Yu and Porikli's method [4], known as TDAE, employ a class-specific discriminative prior. These methods explore either the low-level class-specific feature similarity or mid-level structure information as a spatial constraint in face super-resolution. However, they cannot capture high-level facial characteristic information and thus generate semantically inaccurate upsampled facial details in the outputs.

Unlike previous work, we utilize high-level semantic information, i.e., facial attributes, to reduce the ambiguity when super-resolving very low-resolution faces. However, a direct embedding of the binary facial attribute vector as an additional input channel to the network would still yield degraded results (see Fig. 3c). A simple combination of low-level visual information (an LR image) with high-level semantic information (attributes) in the input layer does not prevent ambiguity or provide consistent LR-HR mappings. We also note that the low-frequency facial components are visible in the LR input while the missing high-frequency details are often contained in the corresponding residual

- *X. Yu, R. Hartley, and F. Porikli are with the Research School of Engineering, Australian National University, Canberra, ACT 0200, Australia. E-mail: {xin.yu, richard.hartley, fatih.porikli}@anu.edu.au.*
- *B. Fernando is with the Artificial Intelligence Initiative, A*STAR, Singapore 138632. E-mail: basura.fernando@anu.edu.au.*

Manuscript received 12 Aug. 2018; revised 7 May 2019; accepted 10 May 2019. Date of publication 14 May 2019; date of current version 1 Oct. 2020. (Corresponding author: Xin Yu.)
Recommended for acceptance by J. Jia.
Digital Object Identifier no. 10.1109/TPAMI.2019.2916881



(a) LR (b) HR (c) NN (d) VDSR [7]

(e) VDSR† [7] (f) [5] (g) [4] (h) Ours

Fig. 1. Comparison with the state-of-the-art CNN based face hallucination methods. (a) 16×16 LR input image. (b) 128×128 HR original image (not used in training). (c) The corresponding HR image of the nearest neighbor of the given LR image in the dataset after compensating for misalignments. (d) Result of VDSR [7], which is a CNN based generic super-resolution method. (e) Result of VDSR† [7] retrained with LR and HR face image pairs. (f) Result of CBN [5]. (g) Result of TDAE [4]. (h) Our result.

between the HR face image and the upsampled LR image (e.g.interpolated by Bicubic interpolation). Thus, our intuition is to incorporate facial attribute information into the residual features that are extracted from LR inputs (as seen in the yellow block of Fig. 2) for super-resolution of high-frequency facial details.

Driven by our observations above, we present a novel LR face image upsampling network that is able to embed facial attributes into face super-resolution. In contrast to previous face super-resolution networks [3], [4], [5], [8], [9], [10], [11], our network employs an autoencoder with skip connections to amalgamate visual features obtained from LR face images and semantic cues provided from facial attributes. It progressively upsamples the concatenated feature maps through its deconvolutional layers. Inspired by the architecture of StackGAN [12], [13], we also employ a discriminative network that is used to examine whether a super-resolved face image is similar to authentic face images as well as the attributes extracted from the upsampled faces are faithful to the input attributes. As a result, our discriminative network can guide the upsampling network to incorporate the semantic information in the overall process. In this manner, the ambiguity in hallucination can be significantly reduced. Furthermore, since we apply the attribute information into the LR residual feature maps rather than concatenating it to the low-resolution input images, we can learn more consistent mappings between LR and HR facial patterns. This allows us to generate realistic high-resolution face images as shown in Fig. 1h.

Above all, the contributions of our work can be summarized as:

- We present a new semantics-embedded face hallucination framework to super-resolve LR face images. Instead of directly upsampling LR face images, we first encode LR images with facial attributes and then super-resolve the encoded feature maps.
- We propose an autoencoder with skip connections to extract residual feature maps from LR inputs and concatenate the residual feature maps with attribute information. This allows us to fuse visual and semantic information to achieve better visual results.
- Even though our network is trained to super-resolve very low-resolution face images, the upsampled HR faces can be further modified by tuning the face attributes in order to add or remove particular

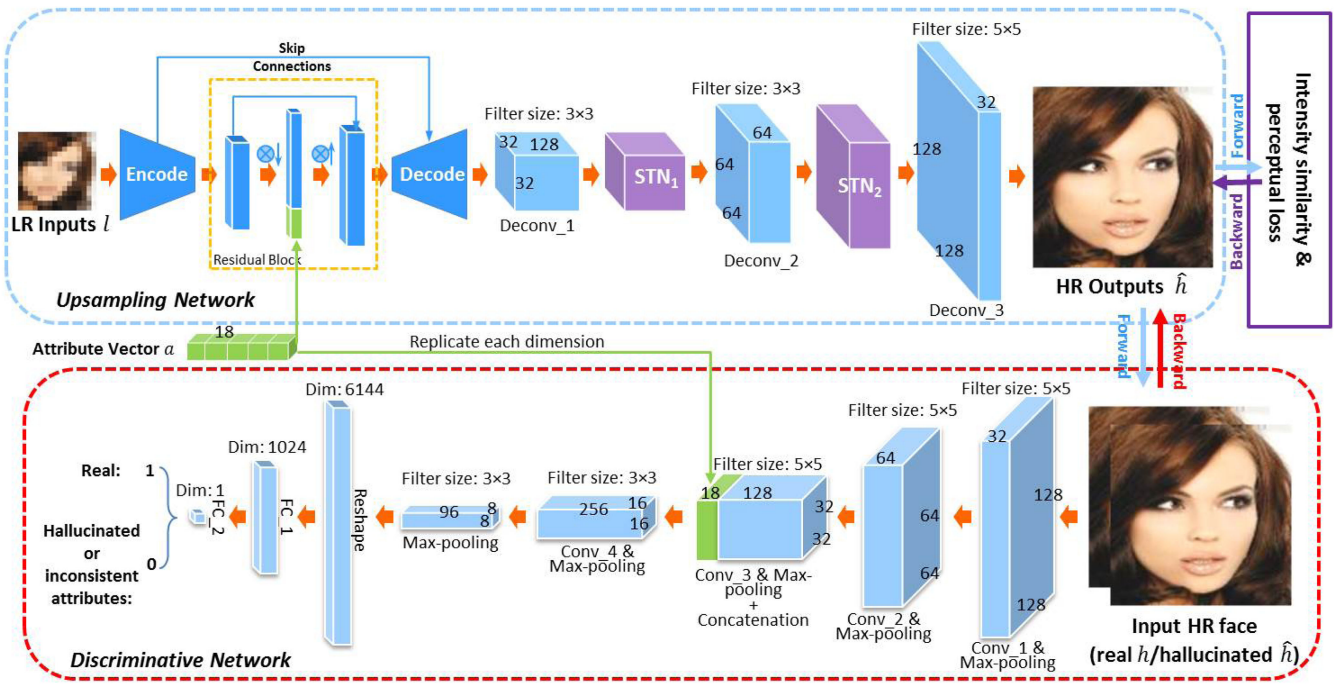


Fig. 2. The architecture of our attribute embedded upsampling network. The network consists of two parts: an upsampling network and a discriminative network. The upsampling network takes LR faces and attribute vectors as inputs while the discriminative network takes real/super-resolved HR face images and attribute vectors as inputs.

attributes. This property significantly increases the flexible of our face super-resolution method rather than only outputting a deterministic upsampled face.

- To the best of our knowledge, our method is the first attempt to utilize high-level semantic information, i.e., facial attribute, into face super-resolution, effectively reducing the ambiguity caused by the inherent nature of this task, especially when the upscaling factor is very challenging, i.e.8×.

2 Related Work

Since our work not only relates to traditional face hallucination methods but also has a close relationship with generative adversarial networks (GAN) [14], we briefly review the related literatures in these two fields.

Face hallucination methods can be roughly grouped into three categories: global model based, part based, and deep learning based. Global model based methods upsample a whole LR input image, often by a learned mapping between LR and HR face images such as Principal Component Analysis (PCA). The seminal works [8], [15] progressively transfer the pixels of HR faces to the given LR face in a Gaussian Pyramid by maximizing a posteriori estimate of the ground-truth HR face. Wang and Tang [16] learn a linear mapping between LR and HR face subspaces, and then reconstruct an HR output with the coefficients estimated from the LR input. Liu et al. [17] not only establish a global model for upsampling LR inputs by PCA but also exploit a local nonparametric model, i.e., Markov Random Field (MRF), to enhance the facial details as well as mitigate blocky and ghosting artifacts in the upsampled faces. Kolouri and Rohde [18] morph an HR output from the exemplar HR faces whose downsampled versions are similar to the LR input by optimal transport and subspace learning techniques. Global model based methods require LR inputs to be precisely aligned and share similar poses to exemplar HR images. However, aligning LR faces is difficult when the resolutions of LR faces are very low (e.g., 16×16 pixels). Therefore, global model based algorithms produce severe artifacts when there are misalignments and pose variations in LR inputs.

Aimed at addressing pose variations, part based methods super-resolve individual facial regions separately. They either exploit reference patches or facial components to reconstruct the HR counterparts of LR inputs. Ma et al. [9] blend position patches extracted from multiple aligned HR images to super-resolve aligned LR face images. In order to suppress image noise and achieve better performance, several follow-up methods [19], [20], [21] reconstruct the position patches in LR faces by sparse coding while Shi et al. [22] design a patch-based reconstruction model in the high-dimensional kernel space. Liu et al. [23] develop a bi-layer model to hallucinate face images and remove noise and outliers in LR inputs simultaneously. In [23], a weight vector is employed to identify whether a pixel is corrupted by noise or not and thus used to tune the contribution of each pixel for hallucination. Jiang et al. [24] exploit the neighboring information of position patches, known as context-patches, to reconstruct HR face images. Tappen and Liu [25] use SIFT flow [26] to align the facial components of LR images and reconstruct HR facial details by warping the reference HR images. Yang et al. [27] employ a facial landmark detector to localize facial components in the LR images and then reconstruct details from the similar HR reference components. Because part based methods need to extract and align facial parts in LR images accurately, their performance degrades dramatically when LR faces are tiny. More comprehensive survey of traditional face super-resolution methods can be referred to the literature review [28].

Recently, deep learning based models achieve significant progress in several image processing tasks and are now pushing forward the state-of-the-art in super-resolution. For instance, Yu and Porikli [11] employ deconvolutional layers to super-resolve aligned LR faces and convolutional layers to remove potential blocky artifacts. Their method also resorts an unsharp filter to enhance the edges of hallucinated faces. In order to train an end-to-end upsampling network, Yu and Porikli [3] introduce a discriminative generative network to super-resolve aligned tiny LR face images. Instead of restoring image intensities of HR faces, Huang et al. [29] estimate wavelet coefficients of an upsampled HR face in the framework of generative adversarial networks. Then the upsampled HR face is reconstructed from the estimated wavelet coefficients. Zhu and Fan [30] first extract feature maps from a blurry LR face image by a convolutional neural network (CNN) and then reconstruct a sharp HR version from the extracted feature maps. Cao et al. [6] employ an attention-aware mechanism to select facial regions from pre-aligned LR faces and then apply a local enhancement network to super-resolve the selected LR patches. Xu et al. [31] design a multi-class adversarial loss to super-resolve aligned LR blurry faces and text images in the framework of generative adversarial networks. Dahl et al. [32] exploit an autoregressive generative model, also known as Pixel-RNN [33], to upscale pre-aligned LR face images.

To relax the requirement of face alignments, Yu and Porikli [10] interweave multiple spatial transformer networks [34] with the deconvolutional layers. In this manner, their method can align LR faces while super-resolving them simultaneously. Based on the observation that mild distortions and artifacts in upsampled HR faces can be mitigated in their downsampled versions, their follow-up work [4] develops a decoder-encoder-decoder structure to super-resolve noisy and unaligned LR faces. Zhu et al. [5] develop a cascade bi-network to localize facial components first and then super-resolve the unaligned LR faces. Chen et al. [35] propose two-stage networks, where low-frequency components of LR faces are first super-resolved and then face priors (i.e., facial component locations) are used to enrich facial details. Later, Yu et al. [36] develop a facial component heatmap guided upsampling network, in which feature maps are first aligned and then the facial components are estimated from the upsampled aligned feature maps. In this way, Yu et al.ease the difficulty of estimating facial components from LR faces. Bulat et al. [37] employ a constraint that the landmarks of the upsampled faces should be close to the landmarks detected in their ground-truth images to handle various poses. However, due to the

inherent under-determined nature of super-resolution, they may still produce results unfaithful to the ground-truths, such as gender reversal and face rejuvenation. Grm et al. [38] and Hsu et al. [39] embed identity information into face hallucination in order to boost face recognition performance on the upsampled HR faces. Since the identity information is only employed in the loss function but not enforced in the testing phase, those methods may still suffer the inherent ambiguity of super-resolution in the testing phase.

Lee et al.'s method [40], concurrent with our work, also employs attributes in face super-resolution, where a feature extractor network is used to extract and combine the features of attributes and LR faces. However, their discriminative network is only designed to distinguish whether the upsampled faces are realistic or not and there is no mechanism to exam whether the attributes are successfully embedded or not. Lu et al. [41] present an attribute-guided face generation network based on conditional CycleGAN [42]. Similar to our method [43], their method also takes one LR input image and an attribute vector to generate an HR face which satisfies the given attributes. However, [41] only addresses aligned LR faces and requires to train four networks, i.e., two generators and two discriminators.

Image generation also has a close relationship to face hallucination when generated images are faces. Goodfellow et al. [14] propose a generative adversarial network (GAN) to construct images from noise, but the resolution of constructed images is limited (i.e.48×48 pixels) due to difficulty in training. Later, variants of GANs have been proposed to increase the resolutions and quality of generated images [44], [45], [46], [47], [48]. Rather than generating images from noise, conditional GANs [49] have been proposed to generate images from both noise as well as certain conditional inputs, and the conditional information is fed to both the generator and the discriminator. [50] and [12] generate images based on textual inputs. Yan et al. [13] use a conditional CNN to generate faces based on attribute vectors. Perarnau et al. [51] develop an invertible conditional GAN to generate new faces by manipulating facial attributes of the input images, while Shen and Liu [52] change attributes of an input image on its residual image by training two generative networks in a complementary fashion. Since their methods aim at generating new face images rather than super-resolving faces, they may change the identity information. In contrast, our work focuses on obtaining HR faces faithful to LR inputs. We employ the attribute information to reduce the uncertainty in face hallucination rather than producing new face images.

GANs based image-to-image translation networks have also been proposed, such as domain transfer [42], [53], super-resolution [3], [54] as well as photo editing [55], [56]. In particular, several face editing works, such as face aging [57], face completion [58], [59], and face attribute transferring [55], [56], also share many similarities with face hallucination. For instance, the inputs are face images and the outputs are another versions of the input images. However, face editing works mainly focus on changing certain attributes of the given faces or completing the missing parts of faces. On the contrary, our method aims at super-resolving LR input face images to their HR counterparts by exploiting the provided facial attributes instead of editing the input LR faces.

3 Super-Resolution with Attribute Embedding

Each low-resolution face image may correspond to many high-resolution face candidates during the process of increasing their resolutions. To reduce the ambiguity encountered in the super-resolution process, we present an upsampling network that takes LR faces and semantic information (i.e., facial attributes) as inputs and outputs super-resolved HR faces. The entire network consists of two parts: an upsampling network and a discriminative network. The upsampling network is used for embedding facial attributes into LR input images as well as upsampling the fused feature maps. The discriminative network is used to constrain the input attributes to be encoded and the hallucinated face images to be similar to real ones. The entire architecture of our network is illustrated in Fig. 2.

3.1 Attribute Embedded Upsampling Network

The upsampling network is composed of a facial attribute embedding autoencoder and upsampling layers (as shown in the blue frame). Previous works [3], [4], [10], [11] only take LR images as inputs and then super-resolve them by deconvolutional layers. They do not make use of any valuable semantic information into account during super-resolution. Indeed, obtaining semantic information such as facial attributes for face images is not difficult, yet it is logical to make use of semantic information, especially for face images. For instance, we can deduce gender information from the outfits. Unlike previous works, we incorporate low-level visual and high-level semantic information in face super-resolution to reduce the ambiguity of the mappings between LR and HR images.

Rather than concatenating LR input images with attribute vectors directly, in our proposed attribute embedding network we employ a convolutional-deconvolutional autoencoder with skip connections [60] to fuse visual features and attribute vectors. Due to the skip connections, we can utilize residual features obtained from LR input images to incorporate the attribute vectors. Specifically, at the bottleneck of the autoencoder, we concatenate the attribute vector with the residual feature vector as illustrated in the green and blue vectors of Fig. 2. As shown in Fig. 3d, when we encode attributes with the feature maps of LR faces at the bottleneck of the autoencoder without using the skip connections instead of residual feature maps, artifacts appear in the smooth regions of the super-resolved result. After combining the residual feature vectors of LR inputs with the attribute vectors, we employ deconvolutional layers to upsample the concatenated feature maps. Since LR input images may undergo misalignments, such as in-plane rotations, translations and scale changes, we use spatial transformer networks (STNs) [34] to compensate for misalignments similar to [4], [10], as shown in the purple blocks in Fig. 2. Since STNs employ bilinear interpolation to re-sample images, they will blur LR input images, as reported in [10]. Therefore, we only employ STNs in the upsampling layers.

To constrain the appearance similarity between the super-resolved faces and their HR ground-truth counterparts, we exploit a pixel-wise euclidean distance loss, also known as pixel-wise ℓ_2 loss, and a feature-wise ℓ_2 loss,

Fig. 3. Ablation study of our network. (a) 16×16 LR input image. (b) 128×128 HR ground-truth image, its ground-truth attributes are male and old. (c) Result without using an autoencoder. Here, the attribute vectors are replicated and then concatenated with the LR input directly. (d) Result without using skip connections in the autoencoder. (e) Result by only using an ℓ_2 loss. (f) Result without using the attribute embedding but with a standard discriminative network. In this case, the network is similar to the decoder in [4]. (g) Result without using the perceptual loss. (h) Our final result.

dubbed perceptual loss [61]. The pixel-wise ℓ_2 loss is employed to enforce image intensity similarity between the upsampled HR faces and their ground-truth images. As reported in [3], deconvolutional layers supervised by an ℓ_2 loss tend to output over-smoothed results as shown in Fig. 3e. Since the perceptual loss measures euclidean distance between features of two images, we use it to constrain feature similarity between the upsampled faces and their ground-truth ones. We use VGG-19 [62] to extract features from images (please refer to Section 3.3 for more details). Without the help of the perceptual loss, the network tends to produce ringing artifacts to mimic facial details, such as wrinkles, as seen in Fig. 3g.

3.2 Discriminative Network

In order to force the upsampling work to encode facial attribute information, we employ a conditional discriminative network. Specifically, the discriminative network is designed to distinguish whether the attributes of super-resolved face images are faithful to the desired attributes embedded in the upsampling network or not and is used to constrain the upsampled images to be similar to HR real face images too.

Even though our autoencoder concatenates attribute vectors with residual feature maps of the LR inputs, the upsampling network may simply learn to ignore them, e.g., the weights corresponding to the semantic information are zeros. Therefore, we need to design a discriminator network to enforce semantic attribute information into the generative process. As shown in Fig. 3f, by employing a standard discriminative network [3], [45], the output HR face still looks like a female face even if the expected figure should be an old male. It implies that the attribute information is not well embedded. Therefore, simply embedding a semantic vector into LR inputs may increase the ambiguity or deviate the learned mapping between the LR and correct HR face images.

We present a discriminative network to enforce the input attribute information to be embedded in LR inputs, thus generating the desired attributes in the hallucinated face images. As shown in the red frame of Fig. 2, our discriminative network is constructed by convolutional layers and fully connected layers. HR face images (real and upsampled faces) are fed into the network while attribute information is also fed into the middle layer of the network as conditional information. Here, an attribute vector is replicated and then concatenated with the feature maps of images. Because CNN filters in the first layers mainly extract low-level features while filters in higher layers extract image patterns or semantic information [63], we concatenate the attribute information with the extracted feature maps on the third layer, which yields good empirical results in our experiments. If the extracted features do not comply with the input attribute information, the discriminative network ought to pass that information to the upsampling network. Our discriminative network is a binary classifier which is trained with a binary cross-entropy loss. With the help of the discriminative network, the attribute information can be embedded into the upsampling network. As shown in Fig. 3h, our final result is faithful to the age and gender of the ground-truth image.

3.3 Training Procedure

Our face super-resolution network is trained in an end-to-end fashion. We use an LR face image denoted by l_i and its ground-truth attribute label vector a_i as the inputs and the corresponding HR ground-truth face image h_i as the target. Note that, since our network aims at super-resolving very low-resolution face images rather than manipulating facial attributes of HR face images, we only feed the correct attributes of LR face images into the upsampling network in the training phase.

In training the entire network, we employ a binary cross-entropy loss to update our discriminative network and then train the upsampling network using a pixel-wise ℓ_2 loss, a perceptual loss and the discriminative loss obtained from our discriminative network. Therefore, we first update the parameters of the discriminative network and then the parameters of the upsampling network because the upsampling network relies on the loss back-propagated from the discriminative network to update its weights.

3.3.1 *Training Discriminative Network*

Our discriminative network is designed to embed attribute information into the upsampling network as well as to force the super-resolved HR face images to be authentic. Similar to [12], [13], our goal is to make the discriminative network be able to tell whether super-resolved faces contains the desired attributes or not but fail to distinguish hallucinated faces from real ones. Hence, in order to train the discriminative network, we take real HR face images h_i and their corresponding ground-truth attributes a_i as positive sample pairs $\{h_i, a_i\}$. Negative data is constructed from super-resolved HR faces $\hat{h}_i$ by our upsampling network and their ground-truth attributes a_i as well as real HR faces and mismatched attributes $\tilde{a}_i$. Therefore, the negative sample pairs consist of both $\{\hat{h}_i, a_i\}$ and $\{h_i, \tilde{a}_i\}$. The objective function

for the discriminative network $\mathcal{L}_D$ is expressed as:

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

where d represents the parameters of the discriminative network $\mathcal{D}$, $\mathcal{D}_d(h_i,a_i)$, $\mathcal{D}_d(\hat{h}_i,a_i)$ and $\mathcal{D}_d(h_i,\tilde{a}_i)$ are the outputs of $\mathcal{D}$, $\mathcal{U}_t(l_i)$ is the output of our upsampling network and t represents the parameters of our upsampling network. In addition, $p(h,a)$ represents the joint distribution of positive sample pairs, $p(\hat{h},a)$ as well as $p(h,\tilde{a})$ represent the joint distributions of negative sample pairs, and $p(l,a)$ represents the joint distribution of the LR input faces and their ground-truth attributes.

Since all the layers in our discriminative network are differentiable, back-propagation is used to calculate the gradients with respect to the parameters of the discriminative network d. Thus, we minimize $\mathcal{L}_{\mathcal{D}}$ by RMSprop [64] as follows:

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

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.

3.3.2 Training Upsampling Network

Since our upsampling network aims at super-resolving LR input images, we only feed our upsampling network with LR face images l_i and their corresponding attributes a_i as inputs. To constrain the upsampled faces to be similar to the HR ground-truth face images, we employ a pixel-wise ℓ_2 loss on image intensities, expressed as:

$$\begin{aligned}\mathcal{L}_{pix} &= \mathbb{E}_{(\hat{h}_i,h_i)\sim p(\hat{h},h)}\|\hat{h}_i-h_i\|_F^2 \\ &= \mathbb{E}_{(l_i,a_i,h_i)\sim p(l,a,h)}\|\mathcal{U}_t(l_i,a_i)-h_i\|_F^2,\end{aligned} \tag{3}$$

where $p(\hat{h},h)$ is the joint distribution of the upsampled faces and their ground-truth counterparts and $p(l,h,a)$ represents the joint distribution of the LR and HR face images and their corresponding attributes in the training dataset.

As mentioned in Section 3.1, we also employ a perceptual loss $\mathcal{L}_{feat}$ to enforce the feature similarity between the super-resolved faces and their corresponding ground-truths, written as:

$$\begin{aligned}\mathcal{L}_{feat} &= \mathbb{E}_{(\hat{h}_i,h_i)\sim p(\hat{h},h)}\|\Phi(\hat{h}_i)-\Phi(h_i)\|_F^2 \\ &= \mathbb{E}_{(l_i,a_i,h_i)\sim p(l,a,h)}\|\Phi(\mathcal{U}_t(l_i,a_i))-\Phi(h_i)\|_F^2,\end{aligned} \tag{4}$$

where $\Phi(\cdot)$ denotes feature maps extracted by the ReLU32 layer in VGG-19 [62], which gives good empirical performance in our experiments.

To enforce the upsampling network to encode the attribution information, a discriminative loss $\mathcal{L}_{dis}$ is also exploited as follows:

$$\begin{aligned}\mathcal{L}_{dis} &= -\mathbb{E}_{(\hat{h}_i,a_i)\sim p(\hat{h},a)}\log\left(\mathcal{D}_d(\hat{h}_i,a_i)\right) \\ &= -\mathbb{E}_{(l_i,a_i)\sim p(l,a)}\log\left(\mathcal{D}_d(\mathcal{U}_t(l_i,a_i),a_i)\right),\end{aligned} \tag{5}$$

where $p(\hat{h},a)$ indicates the joint distribution of the upsampled faces and their corresponding attributes.

All the above three losses are used to update the parameters of our upsampling network, and the total loss $\mathcal{L}_{\mathcal{U}}$ is expressed as:

$$\mathcal{L}_{\mathcal{U}} = \mathcal{L}_{pix}+\alpha\mathcal{L}_{feat}+\beta\mathcal{L}_{dis}, \tag{6}$$

where α is a weight term which trades off between the image intensity similarity and the feature similarity, and β is a weight which trades off between the appearance similarity and the attribute similarity. Here, we also employ RMSprop to update the parameters of our upsampling network:

$$\begin{aligned}\Delta^{i+1} &= \gamma\Delta^{i}+(1-\gamma)\left(\tfrac{\partial \mathcal{L}_{\mathcal{U}}}{\partial t}\right)^2, \\ t^{i+1} &= t^{i}-r\tfrac{\partial \mathcal{L}_{\mathcal{U}}}{\partial t}\tfrac{1}{\sqrt{\Delta^{i+1}+\epsilon}}.\end{aligned} \tag{7}$$

After updating the upsampling network, we can obtained upsampled face images in better quality. Hence, we use HR faces hallucinated by the newly updated upsampling network to train the discriminative network again. By updating these two network alternatingly, we can achieve realistic super-resolved face images including correct attributes. The entire training procedure is illustrated in Algorithm 1.

Algorithm 1. Training Procedure of our Entire Network

Input: minibatch size N, LR and HR face image pairs $\{l_i,h_i\}$ and their corresponding attributes a_i, maximum number of iterations K.

1: **while** iter $< K$ **do**
2: Choose one minibatch of LR and HR image pairs $\{l_i,h_i\}$ and their corresponding attributes, $i=1,\ldots,N$.
3: Generate one minibatch of HR face images $\hat{h}_i$ from $\{l_i,a_i\}, i=1,\ldots,N$, where $\hat{h}_i=\mathcal{U}_t(l_i,a_i)$.
4: Generate mismatched attributes $\tilde{a}_i$ from a_i by randomly permuting one dimension in an attribute vector.
5: Generate positive sample pairs $\{h_i,a_i\}$ and negative sample pairs $\{\hat{h}_i,a_i\}$ and $\{h_i,\tilde{a}_i\}$.
6: Update the parameters of the discriminative network $\mathcal{D}_d$ by using Eqns. (1) and 2.
7: Update the parameters of the upsampling network $\mathcal{U}_t$ by using Eqns. (6) and 7.
8: **end while**

Output: Our attribute embedded upsampling network.

3.4 Super-Resolving LR Inputs with Attributes

The discriminative network $\mathcal{D}$ is only required in the training phase. In the super-resolving (testing) phase, we take

LR face images and their corresponding attributes as the inputs of the upsampling network $\mathcal{U}$, and the outputs of $\mathcal{U}$ are the hallucinated HR face images. In addition, although the attributes are binary values, i.e., either 0 or 1, in training, the attributes can be further scaled, such as negative values or values exceeding 1, to manipulate the final super-resolved results according to the users' descriptions in the testing phase.

3.5 Implementation Details

The detailed architectures of the upsampling and discriminative networks are illustrated in Fig. 2. We employ convolutional layers with kernels of size 4×4 in a stride 2 in the encoder and deconvolutional layers with kernels of size 4×4 in a stride 2 in the decoder. The feature maps in our encoder will be passed to the decoder by skip connections. We also use the same architectures of the STN layers in [4] to align feature maps. Specifically, the STN layers are constructed by convolutional and ReLU layers (Conv+ReLU), max-pooling layers with a stride 2 (MP2) and fully connected layers (FC). STN_1 layer is cascaded by: MP2, Conv +ReLU (with the filter size: $128 \times 20 \times 5 \times 5$), MP2, Conv +ReLU (with the filter size: $20 \times 20 \times 5 \times 5$), FC+ReLU (from 80 to 20 dimensions) and FC (from 20 to 4 dimensions). STN_2 is cascaded by: MP2, Conv+ReLU (with the filter size: $64 \times 128 \times 5 \times 5$), MP2, Conv+ReLU (with the filter size: $128 \times 20 \times 5 \times 5$), MP2, Conv+ReLU (with the filter size: $20 \times 20 \times 3 \times 3$), FC+ReLU (from 180 to 20 dimensions) and FC (from 20 to 4 dimensions). We do not use zero-padding in the convolution operations.

We set the learning rate to 0.001 and multiplied by 0.95 after each epoch, and α is set to 0.01. As suggested by [4], we also set β to 0.01 and gradually decrease it by a factor 0.995, thus emphasizing the importance of the appearance similarity. On the other hand, in order to guarantee the attributes to be embedded in the training phase, we stop decreasing β when it is lower than 0.005.

4 Experiments

We evaluate our network qualitatively and quantitatively, and compare with the state-of-the-art methods [4], [5], [6], [7], [9], [22], [24], [35], [36], [40], [54]. Ma et al.'s method [9] exploits position-patches in the exemplary dataset to reconstruct HR images while Jiang et al.'s method uses context-patches to upsample LR images. Shi et al. [22] establish the correspondences between LR and HR position patches in a high-dimensional kernel space. Kim et al.'s method [7], dubbed VDSR, is a generic CNN based super-resolution method. Ledig et al.'s method [54], also known as SRGAN, is also a generic CNN based super-resolution method, which employs an adversarial loss to enhance the super-resolved details. Since VDSR and SRGAN are trained on natural images, they may not capture LR facial patterns well for face super-resolution. We retrain VDSR and SRGAN on entire face images for fair comparisons. Zhu et al. [5] employ a cascaded deep convolutional neural network to hallucinate facial components of LR face images. Cao et al. [6] exploit a recurrent attention mechanism to localize and upsample facial regions. Yu and Porikli [4] use a decoder-encoder-decoder structure to super-resolve unaligned LR faces. Chen et al. [35] and Yu et al. [36] exploit the facial structure as a spatial constraint to hallucinate faces. Lee et al. [40] fuse attribute vectors and LR inputs in the feature space and then upsample the concatenated features to obtain HR faces.

4.1 Dataset

We use the Celebrity Face Attributes (CelebA) dataset [65] to train our network because CelebA dataset contains over 220K face images and also provides 40 binary-value attributes for each face image. Unlike previous face generation methods [13], [51], [52], our network focuses on super-resolving LR faces by exploiting facial attributes. Hence, we only choose the attributes related to facial details, such as gender, age and beard information, rather than the attributes which can be directly extracted from LR faces, such as hair and skin colors, and are not related to facial details, such as wearing hats, glasses and earrings. In particular, we select the 18 attributes from the 40 attributes, including 5 o'clock shadow, arched eyebrow, bags under eyes, big lips, big nose, bushy eyebrows, double chin, goatee, heavy makeup, high cheekybone, male, mouth open, mustache, narrow eyes, no beard, pointy nose, sideburns and young. In this way, we reduce the potential inconsistency between visual and semantic information imposed by the supplementary attributes.

When generating the LR and HR face pairs, we select 170K cropped face images from the CelebA dataset, and then resize them to 128×128 pixels as HR images. We manually transform the HR images, including rotations, translations and scale changes, and then downsample HR images to 16×16 pixels to attain their corresponding LR images. We use 160K LR and HR face pairs and their corresponding attributes for training, 2K LR and HR image pairs and their attributes for validation, and 2K LR face images and their ground-truth attributes for testing.

4.2 Qualitative Comparison with the SoA

Some algorithms [6], [7], [9], [22], [24], [54] need the alignments of LR inputs before face super-resolution while Yu and Porikli's method [4] and Yu et al.'s method [36] automatically generate upright HR face images. For a fair comparison and better illustration, we employ a spatial transformer network STN_0 to align LR faces. The aligned upright HR ground-truth images are shown for comparison. As reported in [4], [10], LR faces aligned by STN_0 may still suffer misalignments. Therefore, we employ multiple STNs in the upsampling network to reduce misalignments similar to [4], [10]. The only difference between STN_0 and STN_1 is that the first MP2 operation in STN_1 is removed in STN_0 and the input channel is 3.

Bicubic upsampling only interpolates new pixels from neighboring pixels rather than hallucinating new contents for new pixels. Furthermore, the resolution of our input face images is very small, and little information is contained in the input images. As shown in Figs. 4c, 5c, 6c, and 7c, conventional bicubic interpolation fails to generate facial details. The upsampled faces also suffer from obvious skew artifacts. This indicates that it is difficult to align very low-resolution faces accurately by a single STN_0.

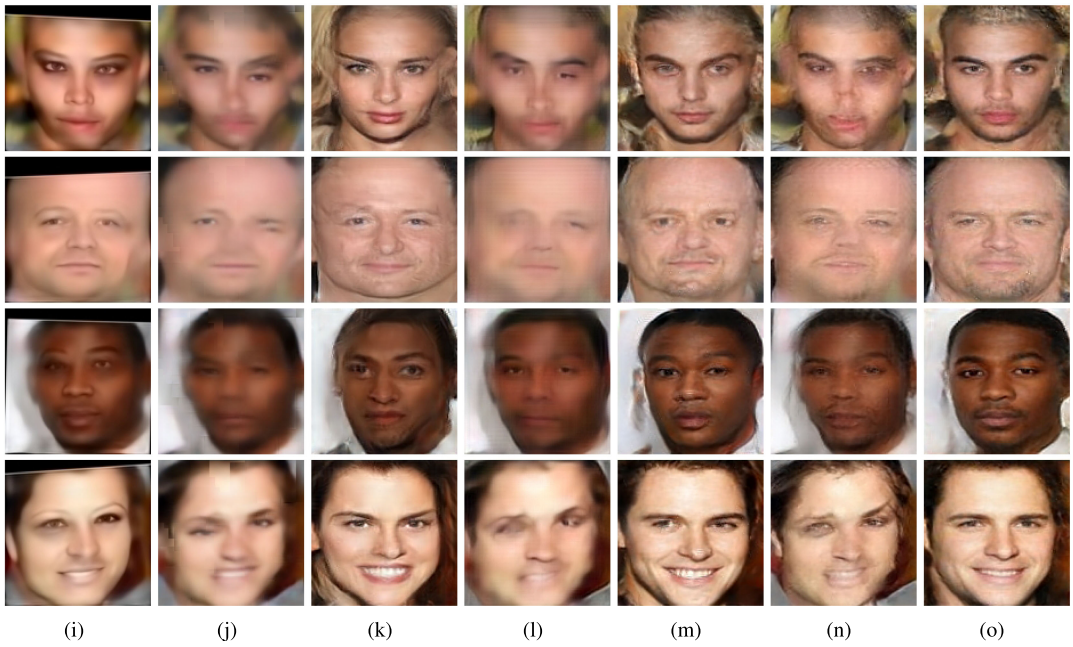


Fig. 4. Comparison with the state-of-the-arts methods on male images. (a) Unaligned LR inputs. (b) Original HR images. (c) Bicubic interpolation. (d) Results of Ma et al.'s method [9]. (e) Results of Shi et al.'s method [22]. (f) Results of Jiang et al.'s method (TLcR-RL) [24]. (g) Results of Kim et al.'s method (VDSR) [7]. (h) Results of Ledig et al.'s method (SRGAN) [54]. (i) Results of Zhu et al.'s method (CBN) [5]. (j) Results of Cao et al.'s method [6]. (k) Results of Yu and Porikli's method (TDAE) [4]. (l) Results of Chen et al.'s method (FSRNet) [35]. (m) Results of Yu et al.'s method [36]. (n) Results of Lee et al.'s method (AACNN) [40]. (o) Our results.

Ma et al. [9] super-resolve HR faces by position patches from HR exemplar face images. Thus, their method is sensitive to misalignments in LR inputs. As seen in Figs. 4d, 5d, 6d, and 7d, there are obvious blur artifacts along the profiles of hallucinated faces. In addition, the correspondences between LR and HR patches become inconsistent as the upscaling factor increases. Hence, severe blocky artifacts appear on the boundaries of different patches.

Shi et al. [22] project position-patches into a high-dimensional kernel space to better represent the non-linear relationship between exemplary HR patches. Thus, Shi et al.'s method avoids the assumption of local geometry consistency between LR and HR patches. However, this method is also sensitive to misalignments in LR inputs because it is still based on position-patches. As seen in Figs. 4e, 5e, 6e, and 7e, the upsampled facial details suffer from blur artifacts due to the misalignments of LR faces.

Jiang et al. [24] exploit context-patches from HR exemplar face images to upsample LR faces, where context-patches consist of the same position-patch and its neighboring patches. Moreover, Jiang et al.'s method also employs a thresholding strategy to select patches and reproduce learning to enhance the final results, known as TLcR-RL. Although context-patch based methods can tolerate slight

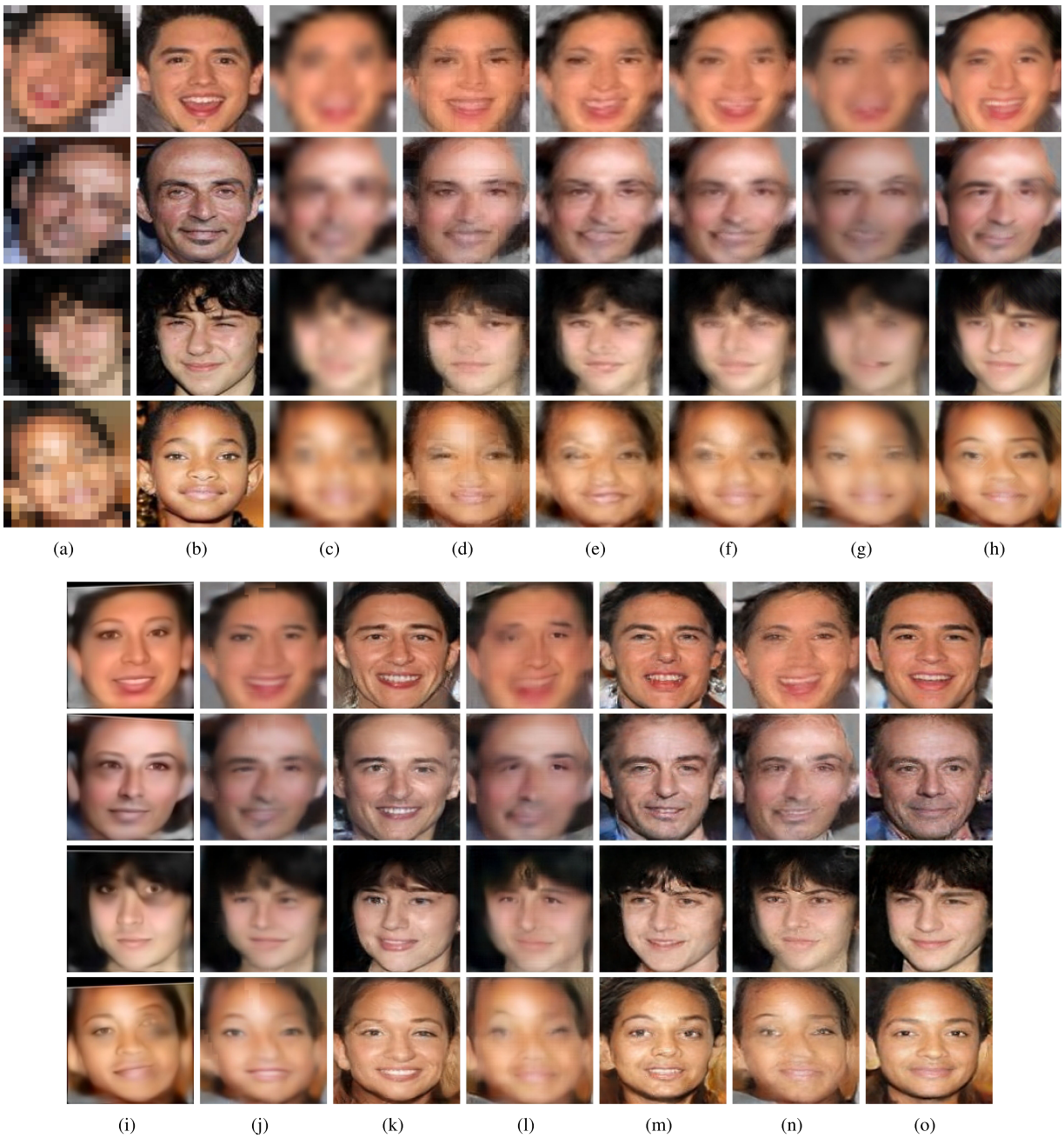


Fig. 5. Comparison with the state-of-the-arts methods on male images. (a) Unaligned LR inputs. (b) Original HR images. (c) Bicubic interpolation. (d) Results of Ma et al.'s method [9]. (e) Results of Shi et al.'s method [22]. (f) Results of Jiang et al.'s method (TLcR-RL) [24]. (g) Results of Kim et al.'s method (VDSR) [7]. (h) Results of Ledig et al.'s method (SRGAN) [54]. (i) Results of Zhu et al.'s method (CBN) [5]. (j) Results of Cao et al.'s method [6]. (k) Results of Yu and Porikli's method (TDAE) [4]. (l) Results of Chen et al.'s method (FSRNet) [35]. (m) Results of Yu et al.'s method [36]. (n) Results of Lee et al.'s method (AACNN) [40]. (o) Our results.

misalignments of LR faces, aliasing artifacts appear in the upsampled HR face images due to the variations of facial expressions and poses in the HR exemplar dataset, as visible in Figs. 4f, 5f, 6f, and 7f.

Kim et al. [7] present a deep CNN for generic purpose super-resolution, known as VDSR. Because VDSR is trained on natural image patches and does not provide an upscaling factor of 8×, it cannot capture the global face structure, as shown in Fig. 1d. We re-train the model with an upscaling factor of 8× on face images, marked as VDSR†. As shown in Figs. 4g, 5g, 6g, and 7g, this method also suffers from the distortion artifacts in the results due to misalignments. Furthermore, since VDSR† is only trained by a pixel-wise ℓ_2 loss, it outputs overly smoothed results as seen in Figs. 4g, 5g, 6g, and 7g.

Ledig et al. [54] develop a CNN based generic super-resolution method, dubbed SRGAN. In order to avoid producing overly smoothed super-resolved results, SRGAN employs an adversarial loss [14], [45]. Since original SRGAN is also trained on generic image patches, we also fine-tune SRGAN with entire face images for a fair comparison, named as SRGAN†. As seen in Figs. 4h, 5h, 6h, and 7h, SRGAN is able to capture LR facial patterns and achieves sharper upsampled results compared to VDSR. However, misalignments in LR faces result in severe distortions in the final results.

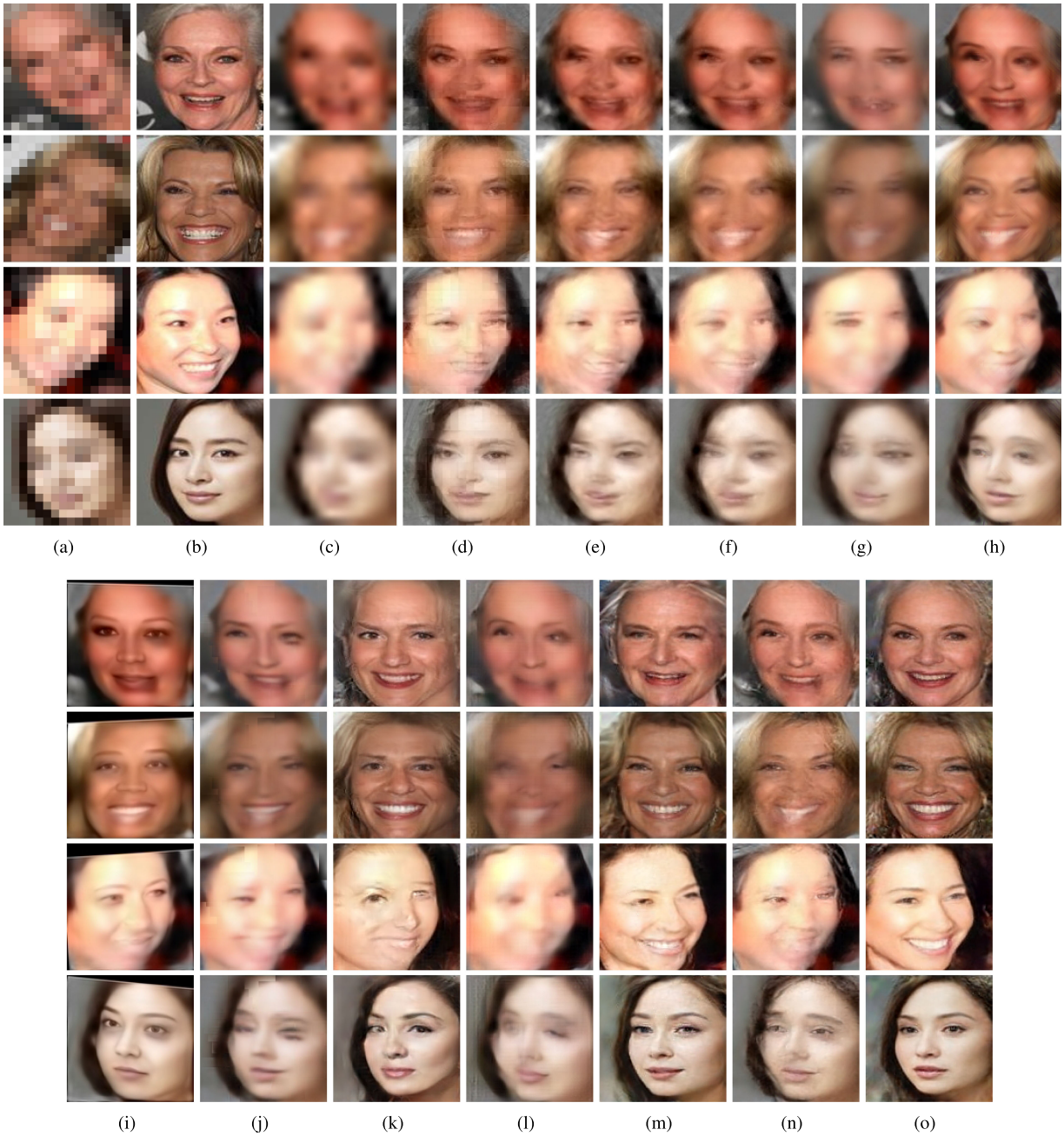


Fig. 6. Comparison with the state-of-the-arts methods on female images. (a) Unaligned LR inputs. (b) Original HR images. (c) Bicubic interpolation. (d) Results of Ma et al.'s method [9]. (e) Results of Shi et al.'s method [22]. (f) Results of Jiang et al.'s method (TLcR-RL) [24]. (g) Results of Kim et al.'s method (VDSR) [7]. (h) Results of Ledig et al.'s method (SRGAN) [54]. (i) Results of Zhu et al.'s method (CBN) [5]. (j) Results of Cao et al.'s method [6]. (k) Results of Yu and Porikli's method (TDAE) [4]. (l) Results of Chen et al.'s method (FSRNet) [35]. (m) Results of Yu et al.'s method [36]. (n) Results of Lee et al.'s method (AACNN) [40]. (o) Our results.

Zhu et al. [5] develop a cascaded bi-network (CBN) to super-resolve very low-resolution face images. CBN first localizes facial components in LR faces and then super-resolves facial details by a local network and entire face images by a global network. As shown in the first and fourth rows of Fig. 4i, CBN is able to generate HR facial components, but it also hallucinates feminine facial details in male face images. For instance, eye lines appear in male faces as seen in the fourth row of Fig. 4i. Furthermore, CBN fails to super-resolve faces of senior people, as shown in the first row of Fig. 6i. As the upscaling factor increases, the facial details in LR faces become more ambiguous. Therefore, it is difficult to recover the facial details of senior people, such as wrinkles and age spots which are even hard to observe in LR faces.

Cao et al. [6] propose an attention-aware face hallucination network. Their network jointly learns an attention mechanism to focus on local face regions and a local enhancement network to super-resolve the selected regions. Because the attention mechanism is learned on aligned faces, misalignments of LR faces lead to inferior super-resolution performance, as illustrated in Figs. 4j, 5j, 6j, and 7j. In addition, their method also suffers from obvious blocky artifacts since some facial regions are not chosen by the attention mechanism for super-resolution.

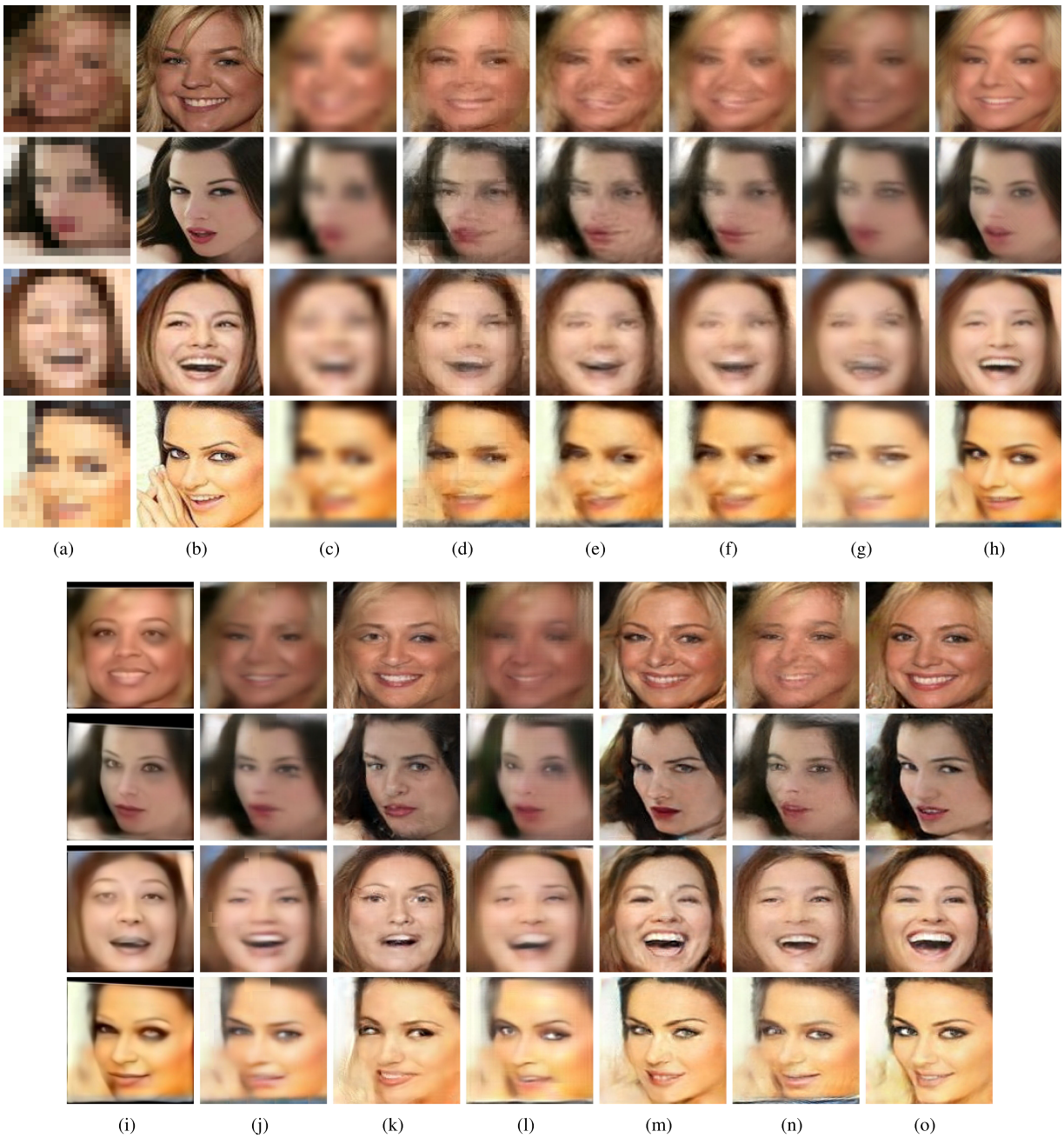
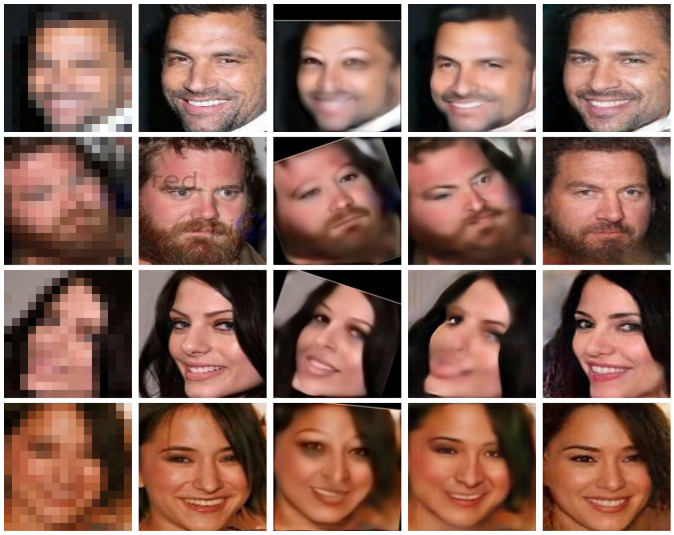


Fig. 7. Comparison with the state-of-the-arts methods on female images. (a) Unaligned LR inputs. (b) Original HR images. (c) Bicubic interpolation. (d) Results of Ma et al.'s method [9]. (e) Results of Shi et al.'s method [22]. (f) Results of Jiang et al.'s method (TLcR-RL) [24]. (g) Results of Kim et al.'s method (VDSR) [7]. (h) Results of Ledig et al.'s method (SRGAN) [54]. (i) Results of Zhu et al.'s method (CBN) [5]. (j) Results of Cao et al.'s method [6]. (k) Results of Yu and Porikli's method (TDAE) [4]. (l) Results of Chen et al.'s method (FSRNet) [35]. (m) Results of Yu et al.'s method [36]. (n) Results of Lee et al.'s method (AACNN) [40]. (o) Our results.

Yu and Porikli [4] exploit a transformative discriminative autoencoder (TDAE) to upsample very low-resolution face images. They also employ deconvolutional layers to upsample LR faces as well as STN layers to align LR faces, but their discriminative network is only used to force the upsampling network to produce sharper results without imposing any high-level semantic information, e.g., facial attributes, in super-resolution. As visible in Figs. 4k, 5k. 6k, and 7k, their method also reverses the genders of the upsampled faces as well as suffers from facial rejuvenation.

Chen et al. [35] develop two stage networks to super-resolve HR faces by exploiting face priors, named FSRNet. FSRNet first upsamples low-frequency components of LR faces by its first-stage network and then explores the face structure of upsampled faces as face priors to enhance facial details by its second-stage network. Although their method does not require alignment of LR faces, we apply their method to the LR faces aligned by STN_0 for comparisons. Since aligning LR faces may introduce extra blurriness and skew artifacts, FSRNet may fail to localize facial components from upsampled overly-smooth HR faces. Thus, FSRNet produces blurry HR faces, as shown in Figs. 4l, 5l, 6l, and 7l.

Yu et al. [36] present a facial component heatmap guided upsampling network. This method aligns feature maps by



(a) (b) (c) (d) (e)

Fig. 8. Results of the state-of-the-arts methods on unaligned LR face images. (a) Unaligned LR inputs. (b) Original HR images. (c) Results of Zhu et al.'s method (CBN) [5]. (d) Results of Chen et al.'s method (FSRNet) [35]. (e) Our results.

STN layers and then estimates the facial component heatmaps from the aligned feature maps rather than coarsely upsampled HR face images in [35]. Since the attributes are not embedded in this method, their super-resolved results may exhibit facial attributes different from the HR ground-truths. For instance, the upsampled mouths are open while the ground-truth ones are closed, as seen in the second, third and last rows of Fig. 5m. As visible in the second row of Fig. 7m, the upsampled mouth is different from the HR ground-truth one. The hallucinated eyes are almost closed while the ground-truth ones are open, as visible in the first row of Fig. 6m.

Lee et al. [40] introduce an attribute augmented convolutional neural network (AACNN) to super-resolve LR faces. Since AACNN is trained on aligned LR and HR face pairs, AACNN is sensitive to misalignments of LR faces. In addition, there is no mechanism to ensure the embedding of attribute information. Thus, AACNN may not fully exploit the attribute information to reduce the ambiguity of face super-resolution, as shown in Figs. 4n, 5n, 6n, and 7n.

In contrast, our method is able to reconstruct authentic facial details as shown in Figs. 4o, 5o, 6o, and 7o. Even though there are different poses, facial expressions and ages in the input faces, our method still produces visually pleasant HR faces which are similar to the ground-truth faces without suffering gender reversal and facial rejuvenation. For instance, we can super-resolve faces of senior persons as illustrated in the second row of Fig. 4o and the first rows of Fig. 6o as well as the child face in the last row of Fig. 5o.

Note that, FSRNet and CBN do not require LR faces to be aligned beforehand. The inferior super-resolution performance may be caused by the imperfect alignment of LR faces. Therefore, in Fig. 8, we also demonstrate the super-resolution results when FSRNet and CBN are directly applied to upsample unaligned LR faces. As seen in Fig. 8, FSRNet and CBN may still fail to localize facial components and then output artifacts in the final results. Those artifacts may handicap the process of aligning the upsampled HR face images. On the contrary, our method not only generates authentic HR faces but also aligns them to the upright position.

TABLE 1
Quantitative Evaluations on the Test Dataset

Method	PSNR	SSIM
Bicubic	19.23	0.56
Ma et al. [9]	19.11	0.54
Shi et al. [22]	19.12	0.55
TLcR-RL [24]	19.18	0.56
VDSR [7]	19.58	0.57
VDSR† [7]	20.12	0.57
SRGAN† [54]	19.06	0.57
CBN [5]	18.77	0.54
Cao et al. [6]	20.09	0.58
TDAE [4]	20.40	0.57
FSRNet [35]	19.25	0.54
Yu et al. [36]	21.25	0.60
AACNN [40]	19.33	0.54
Ours	**21.82**	**0.62**

4.3 Quantitative Comparison with the SoA

We quantitatively measure the performance of all the methods on the entire test dataset by the average Peak Single-to-Noise Ratio (PSNR) and Structural SIMilarity (SSIM) scores. Table 1 demonstrates that our method also achieves superior performance in comparison to other methods.

As indicated in Table 1, after retraining VDSR and SRGAN with face images, they achieve higher PSNRs but still output inferior quantitative results compared with our results. Yu et al.'s method [36] and TDAE [4] also employ multiple STNs to align LR face images and achieve the second and third best results respectively. Note that [36] exploits a multi-task network to super-resolve face images while TDAE [4] employs three networks, which are much larger than our network. This phenomenon also indicates that the ambiguity is significantly reduced by imposing attribute information into the super-resolution procedure rather than by increasing the capacity of a neural network. Therefore, our method is able to achieve better quantitative results.

5 Discussions

5.1 Attribute Manipulation in Super-Resolution

Given an LR face image, previous deep neural network based face hallucination methods [3], [4], [5] only produce a certain HR face image. There is no freedom for those methods to fine-tune the final results. In contrast, our method can output different super-resolved results by adjusting the attribute vectors. As shown in Fig. 9, by changing the gender attribute we can hallucinate face images either from male to female or from female to male. Our method can manipulate the age of the upsampled faces, i.e., more wrinkles and age spots, by changing the age attribute, as seen in Fig. 9b. Because gender and age information may become ambiguous in LR face images, combining that semantic information in super-resolution can produce more accurate results. In addition, after obtaining super-resolved faces,

Fig. 9. Our method can fine-tune the super-resolved results by adjusting the attributes. From top to bottom: the LR input faces, the HR ground-truth faces, our results with ground-truth attributes, our results by adjusting attributes. (a) Reversing genders of super-resolved faces. (b) Aging upsampled faces. (c) Removing makeups. (d) Changing noses. (The first two columns: making noses pointy, and the last two columns: making noses bigger.) (e) Adding and removing beard. (f) Narrowing and opening eyes. (g) Making and removing bushy Eyebrows. (h) Making lips bigger. (i) Opening and closing mouths.

our method is still able to post-edit the upsampled facial details in accordance with the desired attributes. For instance, our method removes the eye lines and shadows in Fig. 9c, makes noses bigger in Fig. 9d, removes and adds beard in Fig. 9e, opens and closes eyes in Fig. 9f, makes eyebrows bushy in Fig. 9g, makes lips bigger in Fig. 9h as well as opens and closes mouths in Fig. 9i by manipulating the corresponding attribute vectors. Therefore, infusing semantic information into LR face images significantly increases the flexibility of our method.

To demonstrate our upsampling network is able to embed attributes into the upsampled HR faces successfully, we choose 9 different attributes, i.e., gender, age, makeup, big nose, beard, open eyes, bushy eyebrows, big lips and open mouth, and train a attribute classifier for each attribute. Note that, some of our selected 18 attributes are

TABLE 2
Classification Results Impacted by Tuning Attributes

Attributes	Male	Young	Makeup	Big nose	Beard	Narrow eyes	Bushy eyebrows	Big lips	Mouth open
GT Attr. Acc.	100%	100%	91%	42%	100%	67%	88%	56%	100%
Increased Attr. Acc.	100%	100%	100%	100%	100%	100%	100%	94%	100%
Decreased Attr. Acc.	0%	0%	2.9%	8.3%	0%	0%	0%	0%	0%

coupled together, such as goatee and beard information, and some attributes may not be always consistent with human observation and are even hard to distinguish in upsampled faces in our experiments, such as eye bags. Therefore, we conduct the quantitative evaluations on the above 9 attributes as visible in Fig. 9 rather than all the selected attributes. By increasing and decreasing the corresponding attribute values, the true positive accuracies are changed accordingly, as illustrated in Table 2. This indicates that the attribute information has been successfully embedded in super-resolution.

5.2 Learn to Encode Attribute Vectors in Hallucination

Since our network directly accepts binary-value attributes, an option to improve the embedding might be using a shared CNN branch EN_s to encode attribute vectors. In the training stage, the encoding branch EN_s will be updated as well in order to embed attributes into the upsampling network. Because the output of EN_s, i.e., the embedded attribute vector, is the input of both the upsampling network and the discriminative network, the ℓ_2 and perceptual losses from the upsampling network $\mathcal{U}$ and the discriminative loss from the discriminative network $\mathcal{D}$ are used to update EN_s. Therefore, although the upsampling network and the discriminative network are updated alternatingly, EN_s is updated in every iteration.

In training our discriminative network, the discriminative labels for the faces upsampled by $\mathcal{U}$ are set to 0 regardless of the attribute information, the labels for real faces with matched attributes are set to 1, and the labels for real faces with mismatched attributes are set to 0. Different from the previous training protocol [3], [4], the discriminative loss is not only used to update the discriminative network but also employed to update the embedding branch EN_s. We only use one binary cross-entropy loss to update the discriminative network $\mathcal{D}$, but the training errors of $\mathcal{D}$ may come from either the face images or the mismatched attributes. Since the binary cross-entropy loss is not able to distinguish whether the faces are hallucinated or the attributes does not match the faces, it may cause ambiguity in the procedure of backpropagation.

On the other hand, in training our upsampling network, only the upsampled faces with their corresponding ground-truth attributes are fed into the discriminative network and the discriminative labels are set to 1. Note that, in training $\mathcal{D}$, the discriminative labels for super-resolved faces with their attributes should be 1 while in training $\mathcal{U}$, the labels are set to 0. Similar to previous works [3], [4], [10], the discriminative loss should be only used to update the upsampling network to make the super-resolved faces realistic, but here it is also used to update the encoding network EN_s. Thus, it is difficult for EN_s to learn a consistent encoder due to the contradicted discriminative labels in training $\mathcal{D}$ and $\mathcal{U}$. Therefore, the super-resolution performance using EN_s decreases 1.79 dB as indicated in Table 3 and the hallucinated faces suffer from obvious artifacts, as seen in Fig. 10c. Therefore, we directly feed a binary-value attribute vector into our upsampling and discriminative network.

5.3 Performance with/without Autoencoder

As shown in Fig. 3c, we demonstrate that it is not suitable to concatenate high-level semantic information with low-level image pixels directly. Specifically, we remove the autoencoder, replicate the attribute vector to the image size, and then concatenate the replicated attributes with the input LR image. In this way, all semantic labels will be applied to the whole images by the low-level convolutional filters. However, low-level filters are mainly responsible to extract image edges or corners [63]. It is unsuitable to employ low-level filters to fuse high-level semantic information and low-level visual information. This is also verified by the quantitative result, donated as woAE, in Table 3.

On the contrary, we first encode the LR input faces by an encoder and then fuse the high-level semantic information, i.e., attribute vectors, with the high-level feature maps extracted by the encoder. In this manner, the attribute labels are better associated with the feature maps qualitatively and quantitatively, as shown in Fig 3h and Table 3.

5.4 Performance with/without Skip-Connections

As shown in Fig. 2, we also employ skip-connections to pass low-frequency components of LR inputs to the decoder. In this fashion, we only focus on embedding the supplementary attributes into high-frequency facial details as well as preserve spatial information of LR input faces. Here, the low-frequency components are not strict low-frequency components of LR faces but relatively low-frequency compared to the components in the residual branch, i.e., high-frequency components. Without using skip-connections, the network will fuse the facial attributes with all the frequency components of LR faces. As seen in Fig. 3d, the hallucinated faces suffer from obvious artifacts at the smooth regions after removing the skip-connections. Therefore, the attribute information should be fused into high-frequency components of LR faces rather than low-frequency ones. We also demonstrate the quantitative result without using the skip-

TABLE 3
Ablation Study on Our Proposed Network

	EN_s	wrAttr	inAttr	woAttr	woAE	noSkip	stdDiscr	Ours
PSNR	20.03	20.42	21.43	21.64	21.03	21.21	21.65	**21.82**
SSIM	0.55	0.53	0.60	0.60	0.58	0.58	0.60	**0.62**

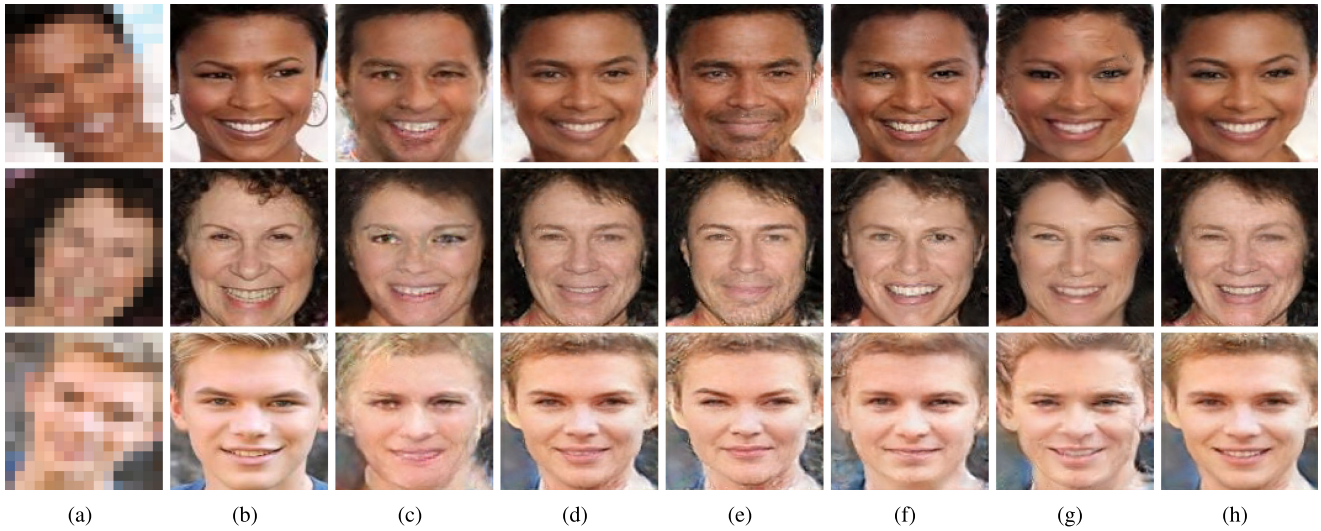


Fig. 10. Discussions on the variants of our network. (a) 16×16 LR input images. (b) 128×128 HR ground-truth images. (c) Results of using a shared CNN branch EN_s to encode attributes in super-resolution. (d) Results of using all neutral attributes. (e) Results of using completely wrong attributes. (f) Results without embedding attribute information. (g) Results of using a standard discriminative network. (h) Our results.

connections, denoted as noSkip, in Table 3. As indicated in Table 3, with the help of the skip-connections, our super-resolution performance increases 0.60 dB in PSNR.

5.5 Performance with Inaccurate Attributes

When super-resolving very low-resolution face images, we may not always obtain all the 18 ground-truth attributes. Therefore, we may use inaccurate attribute information in face hallucination. In this case, we set undetermined attributes to 0.5 as neutral attributes in super-resolution because an attribute is set either 1 or 0 in training. In an extreme case, we do not know any information about attributes. Hence, we use the neutral value for all the attributes in super-resolution, marked as inAttr, and the quantitative result is shown in Table 3. Fig. 10d also illustrates that our network can still generate high-quality results with inaccurate attributes.

Another case is that completely wrong attributes may be assigned to a given input, marked as wrAttr. Here, we reverse all the ground-truth attributes to their opposite values as completely wrong attributes. For instance, we change an attribute value 1 to 0 or vice verse. Notice that, some attributes are coupled together, such as gender and beard attributes. Thus, the wrong attributes are not only inconsistent with LR faces but also may contradict each other. Fig. 10e shows that the super-resolved face images with completely wrong attributes, and Table 3 demonstrates the quantitative results of using completely wrong attributes. As demonstrated in Table 3, using completely wrong attributes degrades the face super-resolution performance. Furthermore, the upsampled HR face images are different from their corresponding HR ground-truth ones, as visible in Fig. 10e. Therefore, when an attribute is uncertain, using neural attributes is more preferable to achieve better face hallucination performance.

5.6 Performance with/without Attribute Embedding

To demonstrate the influence of embedding attributes in face hallucination, we remove the branches of feeding attributes into $\mathcal{U}$ and $\mathcal{D}$ for comparisons, and denote this variant as woAttr. As shown in Fig. 10f, the final results upsampled by woAttr suffer from gender reversal and expression changes. The average PSNR without embedding attributes decreases 0.18 dB, as indicated in Table 3. Furthermore, we also employ two pretrained attribute classifiers, i.e., gender and age, to recognize the attributes recovered by our network and woAttr. For the age classification results, the error rate of our proposed network is 0 while the error rate of woAttr is 23.4 percent. For the gender classification results, the error rate of our proposed network is 0 while the error rate of woAttr is 6 percent. These experiments demonstrate that our method effectively reduces ambiguity in face hallucination by embedding supplementary attributes.

5.7 Impact of Embedding Layers in $\mathcal{D}$

As mentioned in Section 3.2, we embed attribute vectors into the third layer of the discriminative network. Here, we also demonstrate the quantitative results of embedding attributes into different layers of the discriminative network, (i.e., 1st, 2nd, 3rd and 4th convolutional layers). As reported in our previous work [10], overly smoothed upsampled results tend to achieve higher PSNR but their visual quality is inferior. Therefore, we compare the quantitative results when these variants generate similar visual quality results. As shown in Table 4, we achieve the best performance when embedding attribute vectors into the third layer of D.

Lee et al. [40] employ a vanilla discriminative network to distinguish whether the input faces are real or generated. In this manner, the discriminative network is not used to

TABLE 4
Embedding Attributes into Different Layers of $\mathcal{D}$

Layers	$\mathcal{D}_1$	$\mathcal{D}_2$	$\mathcal{D}_3$	$\mathcal{D}_4$
PSNR	21.59	21.76	**21.82**	21.63
SSIM	0.62	0.62	**0.62**	0.61

TABLE 5
Quantitative Evaluations of Impact of Different Losses

Losses	$\mathcal{L}_{pix}$	$\mathcal{L}_{pix}$+$\mathcal{L}_{feat}$	$\mathcal{L}_{pix}$+$\mathcal{L}_{dis}$	Ours
PSNR	22.45	22.31	20.96	21.82
SSIM	0.66	0.65	0.57	0.62

guarantee that the attributes are correctly embedded. Similar to the work [40], we replace our conditional discriminative network with a standard discriminative network, and then retrain our upsampling, marked as stdDiscr. As seen in the second row of Fig. 10g, the upsampled face has been rejuvenated. Since there is no mechanism to exam whether attribute information is fully embedded in the upsampled face images, the upsampled results may still suffer from ambiguity. Thus, artifacts appear in the super-resolved faces, as visible in Fig. 10g. As indicated in Table 3, using our conditional discriminative network can obtain higher PSNRs compared to employing a standard discriminative network. This also implies that a standard discriminative network cannot force our upsampling network to embed the facial attribute information.

5.8 Impact of Different Losses

As seen in Fig. 3, we only show the impact of different losses on the visual results. In Table 5, we also show the quantitative results of our network trained by using different losses. When only employing the pixel-wise ℓ_2 loss, the average PSNR is higher but the visual results suffer from severe blurriness, as shown in Fig. 3e. To avoid generating overly smoothed results, the feature-wise ℓ_2 loss is used in training the network. Due to the lack of the guidance of high-level semantic information in super-resolution, the network trained by using the pixel-wise and feature-wise losses still suffers from notorious ambiguity, such as gender reversal or facial rejuvenation. Using the discriminative loss $\mathcal{L}_{dis}$ and the pixel-wise ℓ_2 loss is able to embed the attribute information in the upsampled face images, but the facial characteristics may not be fully captured. Thus, the upsampling network generates ringing artifacts to mimic facial details, as shown in Fig. 3g. By employing these three losses altogether, our network is able to achieve the best visual quality. Similar to the phenomenon mentioned in our previous work [3], using the discriminative loss is a trade-off between the quantitative performance and the visual quality. Therefore, we set the weight for the discriminative loss to 0.001.

6 Conclusions

We introduced an attribute embedded discriminative network to super-resolve very low-resolution (16×16 pixels) unaligned face images by a large magnification factor $8\times$ in an end-to-end fashion. With the help of the conditional discriminative network, our network successfully embeds facial attribute information into the upsampling network to reduce the inherit ambiguity in super-resolution. After training, our network is not only able to super-resolve LR faces but also fine-tune the upsampled results by adjusting the attribute information. In this manner, our network can generate HR face images much closer to their corresponding ground-truth ones, thus achieving superior face hallucination performance.

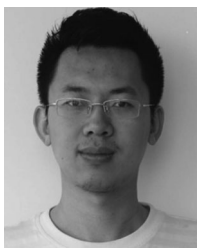
Acknowledgments

This work was supported under the Australian Research Council's Discovery Projects funding scheme (project DP150104645) and Australian Research Council Centre of Excellence for Robotic Vision (project number CE140100016).

References

[1] B. Fasel and J. Luettin, "Automatic facial expression analysis: A survey," *Pattern Recognit.*, vol. 36, no. 1, pp. 259–275, 2003.
[2] W. Zhao, R. Chellappa, P. J. Phillips, and A. Rosenfeld, "Face recognition: A literature survey," *ACM Comput. Surveys*, vol. 35, no. 4, pp. 399–458, 2003.
[3] X. Yu and F. Porikli, "Ultra-resolving face images by discriminative generative networks," in *Proc. Eur. Conf. Comput. Vis.*, 2016, pp. 318–333.
[4] 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.
[5] 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.
[6] Q. Cao, L. Lin, Y. Shi, X. Liang, and G. Li, "Attention-aware face hallucination via deep reinforcement learning," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 690–698.
[7] 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.
[8] S. Baker and T. Kanade, "Hallucinating faces," in *Proc. 4th IEEE Int. Conf. Autom. Face Gesture Recognit.*, 2000, pp. 83–88.
[9] C. Q. Xiang Ma, J. Zhang, "Hallucinating face by position-patch," *Pattern Recognit.*, vol. 43, no. 6, pp. 2224–2236, 2010.
[10] 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.
[11] 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.
[12] H. Zhang, T. Xu, H. Li, S. Zhang, X. Huang, X. Wang, and D. Metaxas, "Stackgan: Text to photo-realistic image synthesis with stacked generative adversarial networks," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2017, pp. 5907–5915.
[13] X. Yan, J. Yang, K. Sohn, and H. Lee, "Attribute2Image: Conditional image generation from visual attributes," *European Conf. Comput. Vision*. Springer: Cham, 2016, pp. 776–791.
[14] I. Goodfellow, J. Pouget-Abadie, and M. Mirza, "Generative Adversarial Networks," in *Proc. Advances Neural Inf. Process. Syst.*, 2014, pp. 2672–2680.
[15] 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.
[16] X. Wang and X. Tang, "Hallucinating face by eigen transformation," *IEEE Trans. Syst. Man Cybern. Part C: Appl. Rev.*, vol. 35, no. 3, pp. 425–434, Aug. 2005.
[17] 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.
[18] 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.
[19] 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. 2861–73, Nov. 2010.
[20] 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.
[21] R. A. Farrugia and C. Guillemot, "Face hallucination using linear models of coupled sparse support," *IEEE Trans. Image Process.*, vol. 26, no. 9, pp. 4562–4577, Sep. 2017.

[22] J. Shi, X. Liu, Y. Zong, C. Qi, and G. Zhao, "Hallucinating face image by regularization models in high-resolution feature space," *IEEE Trans. Image Process.*, vol. 27, no. 6, pp. 2980–2995, Jun. 2018.
[23] L. Liu, C. P. Chen, S. Li, Y. Y. Tang, and L. Chen, "Robust face hallucination via locality-constrained bi-layer representation," *IEEE Trans. Cybern.*, vol. 48, no. 4, pp. 1189–1201, Apr. 2018.
[24] J. Jiang, Y. Yu, S. Tang, J. Ma, G.-J. Qi, and A. Aizawa, "Context-patch based face hallucination via thresholding locality-constrained representation and reproducing learning," *IEEE Trans. Cybern.*, pp. 1–14, 2018.
[25] M. F. Tappen and C. Liu, "A bayesian approach to alignment-based image hallucination," in *Proc. Eur. Conf. Comput. Vis.*, 2012, vol. 7578, pp. 236–249.
[26] 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.
[27] 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.
[28] 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.
[29] 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.
[30] E. Zhou and H. Fan, "Learning Face Hallucination in the Wild," in *Proc. 29th AAAI Conf. Artif. Intell.*, 2015, pp. 3871–3877.
[31] 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.
[32] R. Dahl, M. Norouzi, and J. Shlens, "Pixel recursive super resolution," in *Proc. Int. Conf. Comput. Vis.*, 2017, pp. 5439–5448.
[33] A. Van Den Oord, N. Kalchbrenner, and K. Kavukcuoglu, "Pixel recurrent neural networks," in *Proc. Int. Conf. Int. Conf. Mach. Learn.*, 2016, pp. 1747–1756.
[34] M. Jaderberg, K. Simonyan, A. Zisserman, et al., "Spatial transformer networks," in *Proc. Advances Neural Inf. Process. Syst.*, 2015, pp. 2017–2025.
[35] 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.
[36] X. Yu, B. Fernando, B. Ghanem, F. Porikli, and R. Hartley, "Face super-resolution guided by facial component heatmaps," in *Proc. Eur. Conf. Comput. Vis.*, 2018, pp. 219–235.
[37] 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.
[38] K. Grm, S. Dobrisek, W. J. Scheirer, and V. Struc, "Face hallucination using cascaded super-resolution and identity priors," arXiv:1805.10938, 2018.
[39] C.-C. Hsu, C.-W. Lin, W.-T. Su, and G. Cheung, "Sigan: Siamese generative adversarial network for identity-preserving face hallucination," arXiv:1807.08370, 2018.
[40] C.-H. Lee, K. Zhang, H.-C. Lee, C.-W. Cheng, and W. Hsu, "Attribute augmented convolutional neural network for face hallucination," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. Workshops*, 2018, pp. 721–729.
[41] Y. Lu, Y.-W. Tai, and C.-K. Tang, "Attribute-guided face generation using conditional cyclegan," in *Proc. Eur. Conf. Comput. Vis.*, 2018, pp. 293–308.
[42] J.-Y. Zhu, T. Park, P. Isola, and A. A. Efros, "Unpaired image-to-image translation using cycle-consistent adversarial networks," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2017, pp. 2242–2251.
[43] X. Yu, B. Fernando, R. Hartley, and F. Porikli, "Super-resolving very low-resolution face images with supplementary attributes," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 908–917.
[44] E. Denton, S. Chintala, A. Szlam, and R. Fergus, "Deep generative image models using a laplacian pyramid of adversarial networks," in *Proc. Advances Neural Inf. Process. Syst.*, 2015, pp. 1486–1494.
[45] A. Radford, L. Metz, and S. Chintala, "Unsupervised representation learning with deep convolutional generative adversarial networks," arXiv:1511.06434, pp. 1–15, 2015.
[46] J. Zhao, M. Mathieu, and Y. LeCun, "Energy-based generative adversarial network," arXiv:1609.03126, 2016.
[47] M. Arjovsky, S. Chintala, and L. Bottou, "Wasserstein generative adversarial networks.," in *Proc. Int. Conf. Mach. Learning.*, 2017, pp. 214–223.
[48] D. Berthelot, T. Schumm, and L. Metz, "Began: Boundary equilibrium generative adversarial networks," arXiv:1703.10717, 2017.
[49] M. Mirza and S. Osindero, "Conditional generative adversarial nets," arXiv:1411.1784, pp. 1–7, 2014.
[50] S. Reed, Z. Akata, X. Yan, L. Logeswaran, B. Schiele, and H. Lee, "Generative adversarial text to image synthesis," arXiv:1605.05396, 2016.
[51] G. Perarnau, J. van de Weijer, B. Raducanu, and J. M. Álvarez, "Invertible conditional GANs for image editing," in *Proc. NIPS Workshop Adversarial Training*, 2016, pp. 1–9.
[52] W. Shen and R. Liu, "Learning residual images for face attribute manipulation," in *in Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 4030–4038.
[53] P. Isola, J.-Y. Zhu, T. Zhou, and A. A. Efros, "Image-to-image translation with conditional adversarial networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 5967–5976.
[54] C. Ledig, L. Theis, F. Huszár, et al., "Photo-realistic single image super-resolution using a generative adversarial network," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 4681–4690.
[55] Y. Choi, M. Choi, M. Kim, J.-W. Ha, S. Kim, and J. Choo, "StarGAN: Unified generative adversarial networks for multi-domain image-to-image translation," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 8789–8797.
[56] Z. Shu, E. Yumer, S. Hadap, K. Sunkavalli, E. Shechtman, and D. Samaras, "Neural face editing with intrinsic image disentangling," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 5444–5453.
[57] Z. Wang, X. Tang, W. Luo, and S. Gao, "Face aging with identity-preserved conditional generative adversarial networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 7939–7947.
[58] Y. Li, S. Liu, J. Yang, and M.-H. Yang, "Generative face completion," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2017, vol. 1, Art. no. 3.
[59] X. Yu, F. Shiri, B. Ghanem, and F. Porikli, "Can we see more? joint frontalization and hallucination of unaligned tiny faces," *IEEE Trans. Pattern Anal. Mach. Intell.*, pp. 1–1, 2019.
[60] J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2015, pp. 3431–3440.
[61] 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.
[62] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," arXiv:1409.1556, 2014.
[63] M. D. Zeiler and R. Fergus, "Visualizing and understanding convolutional networks," in *Proc. Eur. Conf. Comput. Vis.*, 2014, pp. 818–833.
[64] 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
[65] X. W. Ziwei Liu, P. Luo, and X. Tang, "Deep learning face attributes in the wild," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2015, pp. 3730–3738.



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 also the PhD degree from the College of Engineering and Computer Science, Australian National University, Canberra, Australia, in 2019. He is currently a research fellow in Australian National University. His interests include computer vision and image processing.

Bsura Fernando received the PhD degree from VISICS group of KU Leuven, Belgium, in Mar. 2015 under the supervision of Professor Tinne Tuytelaars. He is a research scientist at the Artificial Intelligence Initiative (A*AI) of Agency for Science, Technology and Research (A*STAR) Singapore. He is also a honorary lecturer at the Australian National University (ANU). Prior to that he was a research fellow at the Australian Centre for Robotic Vision (ACRV), the Australian National University. He was the project leader of "SR1: Understanding Human and Robot Actions and Interactions" of ACRV. He is interested in Computer Vision and Machine Learning research. Has has contributed to statistical visual domain adaptation and human action recognition.

Richard Hartley is a professor and has been a member of the Computer Vision Group, Research School of Engineering, Australian National University, Acton, Australia, since January 2001. He is a member of the Computer Vision Research Group, NICTA, a Government Funded Research Laboratory. He was with the General Electric Research and Development Center from 1985 to 2001, where he was involved in VLSI design and later in computer vision. He became involved with Image Understanding and Scene Reconstruction working with GEs Simulation and Control Systems Division. He is the author of the book Multiple View Geometry in Computer Vision with A. Zisserman. He is a fellow of the IEEE.

Fatih Porikli received the PhD degree from NYU. He is a professor with the Research School of Engineering, Australian National University, Canberra, Australia. Previously he served as a distinguished research Scientist at Mitsubishi Electric Research Laboratories, Cambridge. He has contributed broadly to object detection, motion estimation, tracking, image-based representations, and video analytics. 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 is a fellow of the IEEE.