Super-Resolving Very Low-Resolution Face Images with Supplementary Attributes*

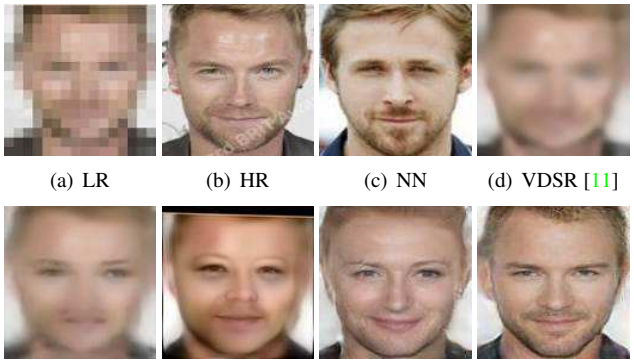
Xin Yu Basura Fernando Richard Hartley Fatih Porikli
Australian National University
{xin.yu,basura.fernando,Richard.Hartley,fatih.porikli}@anu.edu.au

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 exemplar 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. 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× 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.

1. Introduction

Face images provide important information for human visual perception as well as computer analysis [6, 33]. Depending on the imaging conditions, the resolution of a face area may be unfavorably low, raising a critical issue that would directly impede our understanding. Motivated by this challenge, recovering frontalized high-resolution (HR) face images from their low-resolution (LR) counterparts, also known as face hallucination, has received increasing attention recently [27, 29, 35, 4]. Existing face hallucination methods only utilize image domain priors for super-resolution. Even though they are trained on large-scale datasets, 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. For instance, as shown in Fig. 1, the hallucinated details generated by the state-of-the-art face super-resolution methods [35, 29] 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.



(a) LR (b) HR (c) NN (d) VDSR [11]
(e) VDSR† [11] (f) [35] (g) [29] (h) Ours

Figure 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 [11], which is a CNN based generic super-resolution method. (e) Result of VDSR† [11] retrained with LR and HR face image pairs. (f) Result of [35]. (g) Result of [29]. (h) Our result.

*This work was supported under the Australian Research Council's Discovery Projects funding scheme (project DP150104645)

Unlike previous work, we aim to utilize 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. 3(c)). A simple combination of low-level visual information (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 contained in the corresponding residual between the HR face image and the upsampled LR image (*e.g.*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 new LR face image upsampling network that is able to embed facial attributes into face super-resolution. In contrast to face super-resolution networks [27, 28, 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. Based on the StackGAN [31, 24] architecture, we also employ a discriminative network that examines whether a super-resolved face image is similar to authentic face images and 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. 1(h).

Contributions of our work can be summarized as:

- We present a new framework to hallucinate 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 attributes.
- To the best of our knowledge, our method is the first attempt to utilize facial attribute information 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

Face hallucination methods can be roughly grouped into three categories: global model based, part based, and deep learning based.

Global model based methods upsample the whole LR input image, often by a learned mapping between LR and HR face images such as PCA. Wang and Tang [22] 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.* [14] not only establish a global model for upsampling LR inputs but also exploit a local nonparametric model to enhance the facial details. Kolouri and Rohde [12] 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. Since global model based methods require LR inputs to be precisely aligned and share similar poses to exemplar HR images, they produce severe artifacts when there are 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. Baker and Kanade [2] reconstruct high-frequency details of aligned LR face images by searching the best mapping between LR and HR patches. Following this idea, [23, 26, 13] blend position patches extracted from multiple aligned HR images to super-resolve aligned LR face images. Tappen and Liu [21] use SIFT flow [15] to align the facial components of LR images and reconstruct HR facial details by warping the reference HR images. Yang *et al.* [25] 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.

Recently, deep learning based models achieve significant progress in several image processing tasks and is now pushing forward the state-of-the-art in super-resolution. For instance, Yu and Porikli [27] introduce a discriminative generative network to super-resolve aligned tiny LR face images. Follow-up works [28, 29] interweave multiple spatial transformer networks with the deconvolutional layers to relax the requirement of face alignment. Zhu *et al.* [35] use a cascade bi-network to upsample very low-resolution and unaligned faces. Zhu and Fan [34] exploit feature maps extracted from a blurry LR face image by a convolutional

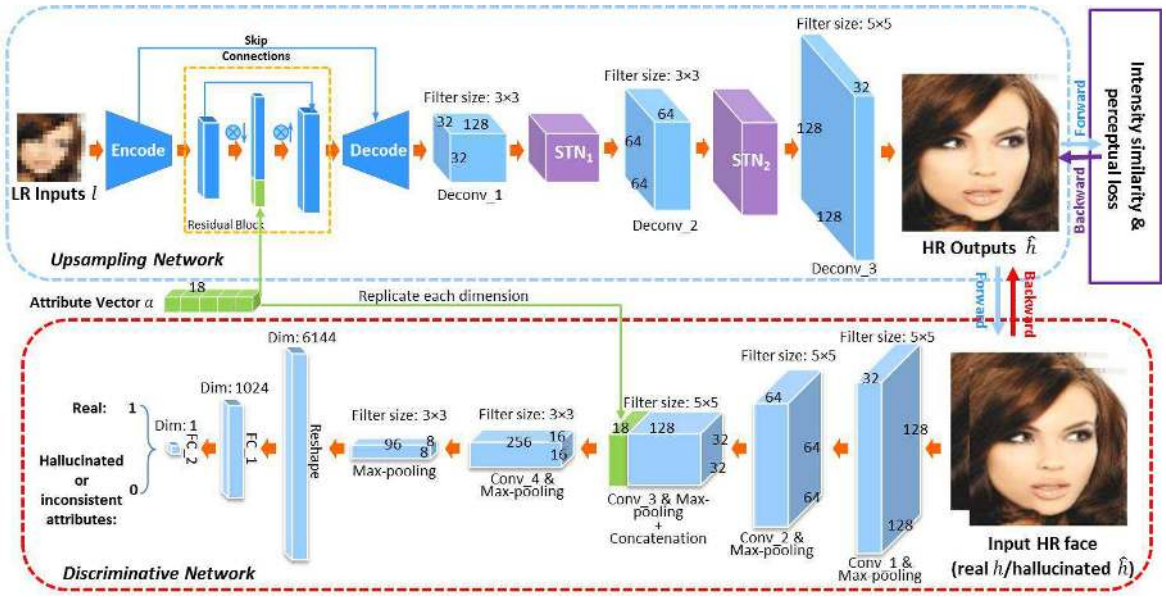


Figure 2. The architecture of our attribute embedded upsampling network. The network consists of two parts: an upsampling network and a discriminative network. The upsampling network takes LR faces and attribute vectors as inputs while the discriminative network takes real/super-resolved HR face images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

neural network (CNN) to reconstruct its sharp HR face image. However, due to the inherent under-determined nature of super-resolution, they may still produce results unfaithful to the ground truth, such as gender reversal and face rejuvenation.

Image generation also has a close relationship to face hallucination when generated images are faces. Goodfellow *et al.* [7] 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 [5, 32, 1, 3]. Rather than generating images from noise, [18, 31] generate images based on textual inputs. Yan *et al.* [24] use a conditional CNN to generate faces based on attributes. Perarnau *et al.* [17] develop an invertible conditional GAN to generate new faces by manipulating facial attributes of the input images, while Shen and Liu [19] change attributes of an input image on its residual image. 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 editing input face images.

3. Super-resolution with Attribute Embedding

Each low resolution face image might map to many high resolution face candidates during the process of making them high resolution. 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 work [27, 29] only take LR images as inputs and then super-resolve them by deconvolutional layers. They do not make use of valuable semantic information into account during super-resolution. Indeed, obtaining semantic information such as facial attributes for face images is not hard, yet it is logical to make use of it, especially for face images. Unlike previous work, we incorporate low-level visual and high-level semantic information in face super-resolution to reduce the ambiguity of mappings between LR and HR image patches.

Rather than concatenating LR input images with attribute vectors directly, in our proposed facial attribute embedding subnetwork, we employ an convolutional autoencoder with skip connections [16]. Due to the skip connections, we can utilize residual features obtained from LR

(a) (b) (c) (d) (e) (f) (g) (h)

Figure 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 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 [29]. (g) Result without using the perceptual loss. (h) Our final result.

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. 3(d), if we encode LR faces with attributes instead of residual feature maps, artifacts may appear in the smooth regions of the super-resolved result. After combining the residual feature vector of LR inputs with the attribute vector, we employ deconvolutional layers to upsample it. Since LR input images may undergo misalignments, such as in-plane rotations, translations and scale changes, we use spatial transformer networks (STNs) [9] to compensate for misalignments similar to [29], as shown in the purple blocks in Fig. 2. Since STNs employ bilinear interpolation to resample images, they will blur LR input images or feature maps. 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 perceptual loss [10]. 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 [27], deconvolutional layers supervised by ℓ_2 loss tend to output over-smoothed results as shown in Fig. 3(e). Since the perceptual loss measures Euclidean distance between features of two images, we use it to constrain feature similarity between the upsample faces and their ground-truth ones. We use VGG-19 [20] 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. 3(g).

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 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. 3(f), the output HR face 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 LR and the 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 [30], we concatenate the attribute information with the extracted feature maps on the third layer. 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 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. 3(h), our final result is faithful to the age and gender of the ground-truth image.

3.3. Training Procedure

Our facial 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 transferring facial attributes of HR face images, we only feed the correct attributes of LR face images into the upsampling network in the training phase.

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

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 [24, 31], 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 $\hat{h}_i$ 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 L_D is expressed as:

$$\begin{aligned} L_D = & -\mathbb{E}\left[\log D_d(h,a)\right] \\ & -\mathbb{E}\left[\log(1-D_d(\hat{h},a))+\log(1-D_d(h,\tilde{a}))\right] \\ = & -\mathbb{E}_{(h_i,a_i)\sim p(h,a))}\left[\log D_d(h_i,a_i)\right] \\ & -\mathbb{E}_{(\hat{h}_i,a_i)\sim p(\hat{h}_i,a))}\left[\log(1-D_d(\hat{h}_i,a_i))\right] \\ & -\mathbb{E}_{(h_i,\tilde{a}_i)\sim p(h_i,\tilde{a}))}\left[\log(1-D_d(h_i,\tilde{a}_i))\right], \end{aligned} \tag{1}$$

where d represents the parameters of the discriminative network D, and $D_d(h_i, a_i)$, $D_d(\hat{h}_i, a_i)$ and $D_d(h_i, \tilde{a}_i)$ are the outputs of D. We update the discriminative network by minimizing L_D.

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 ℓ_2 losses on both image intensity differences and differences of feature maps. In addition, the discriminative loss is also exploited to force the attribute information to be embedded. Hence, we minimize the objective function L_U of the upsampling network as follows:

$$\begin{aligned} L_U = & \mathbb{E}\left[\|\hat{h}-h\|_F^2+\alpha\|\phi(\hat{h})-\phi(h)\|_F^2-\beta\log D_d(\hat{h},a)\right] \\ = & \mathbb{E}_{(l_i,h_i,a_i)\sim p(l,h,a)}\left[\|U_t(l_i,a_i)-h_i\|_F^2\right. \\ & \left.+\alpha\|\phi(U_t(l_i,a_i))-\phi(h_i)\|_F^2-\beta\log D_d(U_t(l_i,a_i),a_i)\right], \end{aligned} \tag{2}$$

where t indicates the parameters of our upsampling network U, $p(l, h, a)$ represents the joint distribution of the LR and HR face images and the corresponding attributes in the training dataset, α is a weight term which trades off between the image intensity similarity and the feature similarity, β is a weight which trades off between the appearance similarity and the attribute similarity, and $\phi(\cdot)$ denotes extracted feature maps from the layer "ReLU32" in VGG-19. Since every layer in our network is differentiable, we employ RMSprop [8] to update t and d.

3.4. Super-Resolving LR Inputs with Attributes

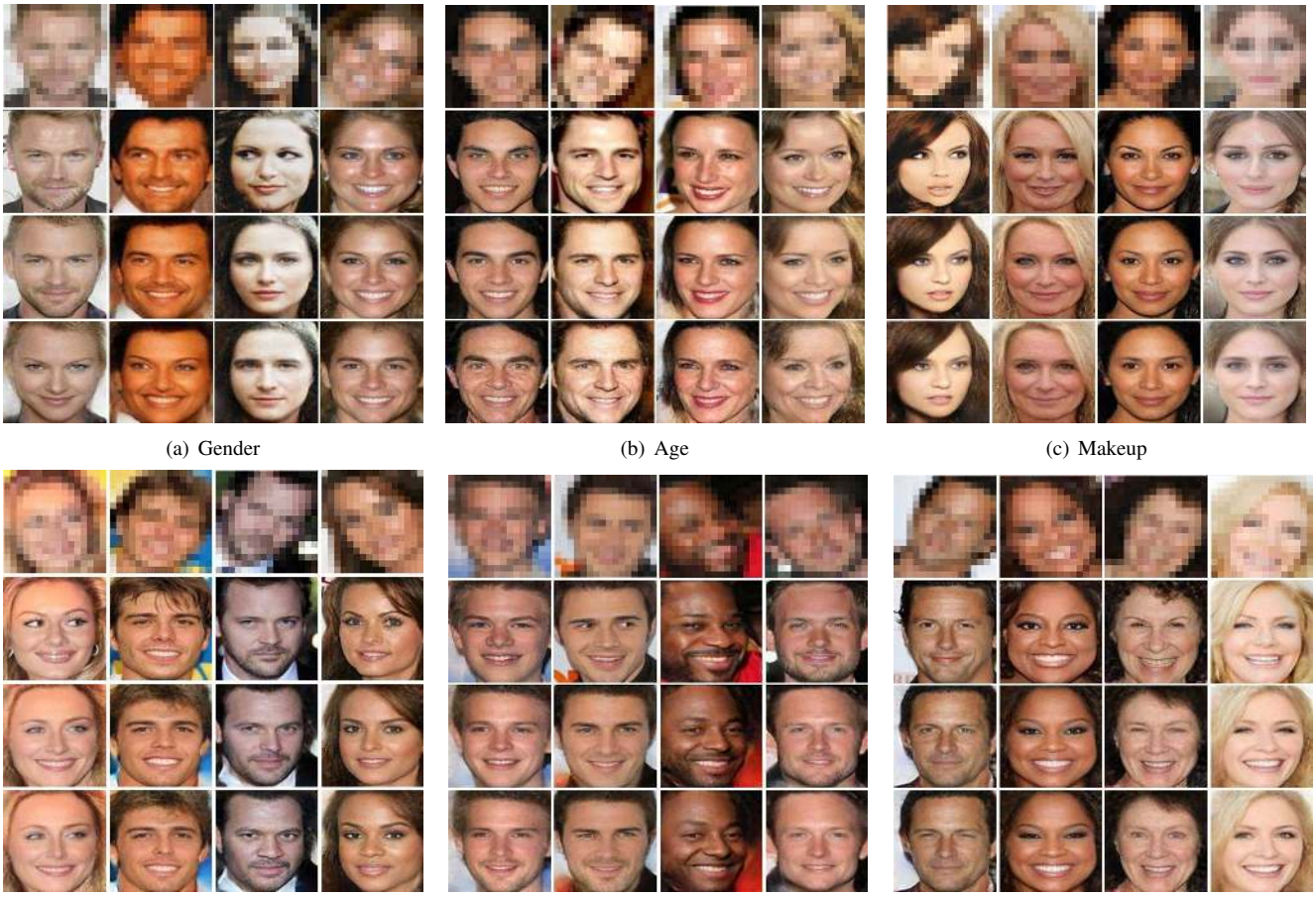
The discriminative network D is only required in the training phase. In the super-resolving (testing) phase, we take LR face images and their corresponding attributes as the inputs of the upsampling network U, and the outputs of U are the hallucinated HR face images. In addition, although the attributes are normalized between 0 and 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 spatial transformer networks in [29] to align feature maps. 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 [29], 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. (All the codes and pretrained model will be released.)

4. Experiments

We evaluate our network qualitatively and quantitatively, and compare with the state-of-the-art methods [11, 23, 35,



(a) Gender (b) Age (c) Makeup

(d) Nose (e) Beard (f) Eyes

Figure 4. 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. (d) Narrowing and opening eyes.

29]. Kim *et al.*'s method [11] is a generic CNN based super-resolution method. Ma *et al.*'s method [23] exploits position-patches in the exemplar dataset to reconstruct HR images. Zhu *et al.* [35] employ a cascaded deep convolutional neural network to hallucinate facial components of LR face images. Yu and Porikli [29] use a decoder-encoder-decoder structure to super-resolve unaligned LR faces[1].

4.1. Dataset

Similar to [27, 29, 17], we use the Celebrity Face Attributes (CelebA) dataset [36] to train our network. 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.

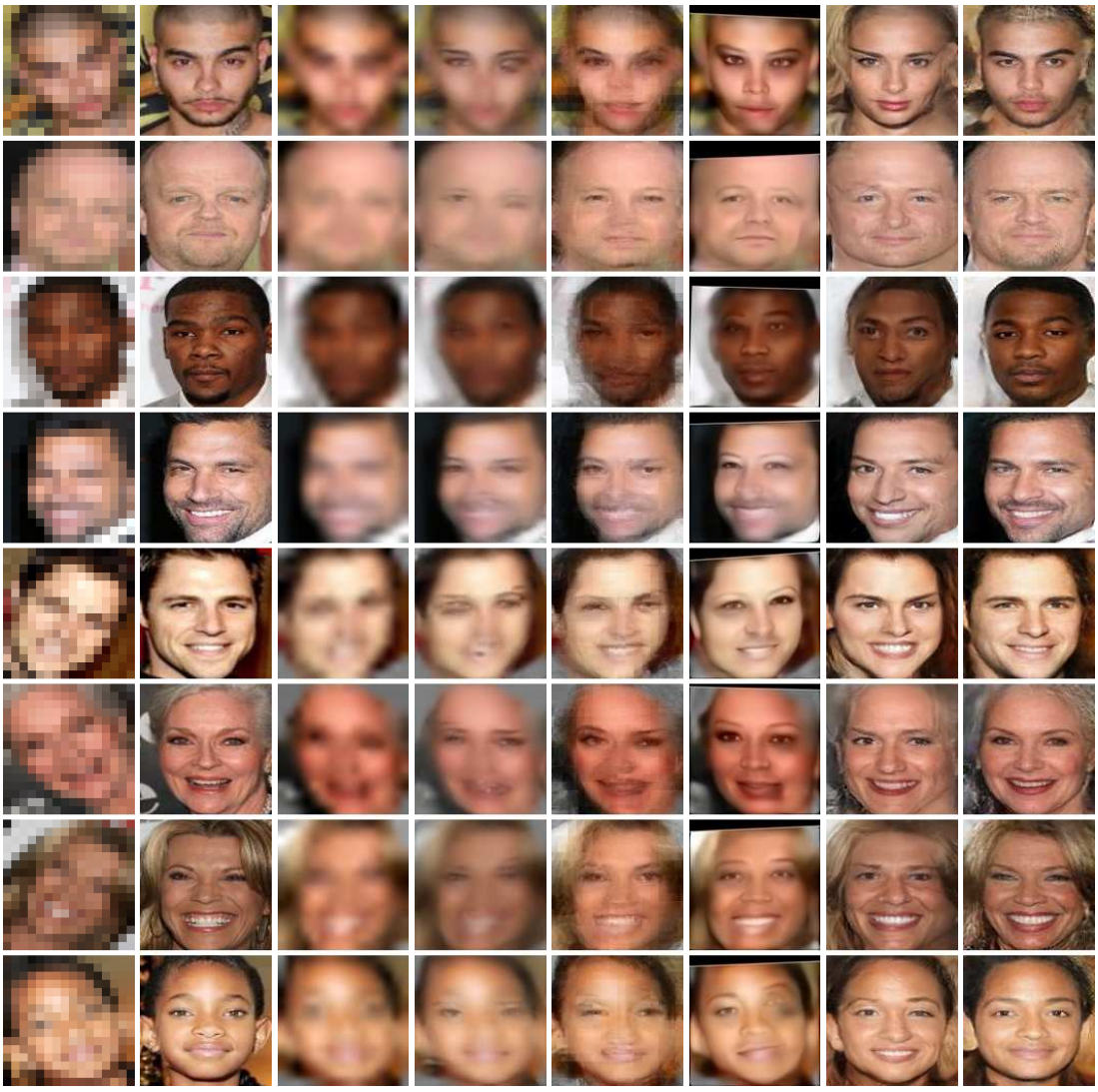
[1]The codes and models are provided from authors' websites.

Table 1. Classification results impacted by tuning attributes.

Attributes	GT Attr. Acc.	Increased Attr. Acc.	Decreased Attr. Acc.
Young	100%	100%	0%
Male	100%	100%	0%
Big nose	42%	100%	8.33%

We use 160K LR and HR face pairs and their corresponding attributes for training, and randomly choose 2K LR face images for testing.

Furthermore, since color information can be directly extracted from LR input faces, such as hair colors and skin colors, we do not include those attributes in super-resolution. Hence, we choose 18 attributes, such as gender, age, and beard information, from 40 attributes in CelebA. In this way, we reduce the potential inconsistency between information extracted from visual information and informa-



(a) (b) (c) (d) (e) (f) (g) (h)

Figure 5. Comparison with the state-of-the-arts methods. (a) Unaligned LR inputs. (b) Original HR images. (c) Bicubic interpolation. (d) Results of Kim *et al.*'s method (VDSR) [11]. (e) Results of Ma *et al.*'s method [23]. (f) Results of Zhu *et al.*'s method (CBN) [35]. (g) Results of Yu and Porikli's method (TDAE) [29]. (h) Our results.

Table 2. Quantitative evaluations on the test dataset.

Method	Bicubic	VDSR [11]	VDSR† [11]	Ma [23]	CBN [35]	TDAE [29]	Ours
PSNR	19.23	19.58	20.12	19.11	18.77	20.40	**21.82**
SSIM	0.56	0.57	0.57	0.54	0.54	0.57	**0.62**

tion imposed by semantic information.

4.2. Attribute Manipulation in Super-Resolution

Given an LR face image, previous deep neural network based face hallucination methods [27, 29, 35] 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. 4, 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. 4(b). Because gender and age infor-

mation may become ambiguous in LR face images, combining that semantic information in super-resolution can produce more accurate results. In addition, we also post-edit our super-resolved results. For instance, our method removes the eye lines and shadows in Fig. 4(c), makes noses bigger in Fig. 4(d), removes and adds beard in Fig. 4(e), as well as makes eyes open in Fig. 4(f) by manipulating the attribute vectors. Furthermore, we choose 3 different attributes, *i.e.*young, male and big nose, and train a attribute classifier for each attribute. By increasing and decreasing the corresponding attribute values, the true positive accuracies are changed accordingly, as illustrated in Tab. 1. This indicates that the attribute information has been successfully embedded in super-resolution. Therefore, infusing semantic information into LR face images significantly increases the flexibility of our method.

4.3. Qualitative Comparison with the SoA

We provide sample results in Fig 5. Note that, Ma *et al.* [23] require input LR faces to be aligned before hallucination while Yu and Porikli's method [29] automatically generates upright HR face images. For a fair comparison and better illustration, we employ a spatial transformer network STN_0 to align LR faces similar to [29]. The aligned upright HR ground-truth images are shown for comparison. As in [28], LR faces aligned by STN_0 may still suffer misalignments, which indicates that alignment of LR faces is difficult. Therefore, we employ multiple STNs in the upsampling network to reduce misalignments similar to [28, 29].

Bicubic upsampling only interpolates new pixels from neighboring pixels rather than hallucinating new contents for new pixels. Furthermore, because the resolution of input face images is very small, little information is contained in the input images. As shown in Fig. 5(c), conventional bicubic interpolation fails to generate facial details.

Kim *et al.* [11] present a deep CNN for generic purpose super-resolution known as VDSR. Because VDSR is trained on natural image patches, it cannot capture the global face structure, as shown in Fig. 1(d). We re-train the model with entire face images. As shown in Fig. 5(d), the artifacts appear in their results due to misalignments, and their method also suffers the gender reversal problem.

Ma *et al.* [23] super-resolve HR faces by position-patches from HR exemplar face images. Hence, their method is sensitive to misalignments in LR inputs. As seen in Fig. 5(e), 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 block artifacts appear on the boundaries of different patches.

Zhu *et al.* [35] develop a cascaded bi-network (CBN) to super-resolve very low-resolution face images. CBN firstly 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 fifth rows of Fig. 5(f), CBN is able to generate HR facial components, but it also hallucinates feminine facial details in male face images, *e.g.*, eye lines appear in male faces as seen in the fifth row of Fig. 5(f). Furthermore, CBN fails to super-resolve faces of senior people, as shown in the sixth row of Fig. 5(f).

Yu and Porikli [29] exploit a transformative discriminative autoencoder (TDAE) to super-resolve very low resolution face images. They also employ deconvolutional layers to upsample LR faces, but their discriminative network is only used to force the upsampling network to produce sharper results without imposing attribute information in super-resolution. As visible in Fig. 5(g), their method also reverses the genders.

In contrast, our method is able to reconstruct authentic facial details as shown in Fig. 1(h). Even though there are different poses, facial expressions and ages in the input faces, our method still produces visually pleasing 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 and sixth rows of Fig. 1(h) as well as the child face in the last row of Fig. 5(h).

4.4. Quantitative Comparison with the SoA

We quantitatively measure the performance of all methods on the entire test dataset by the average PSNR and the structural similarity (SSIM) scores. Table 2 presents that our method achieves superior performance in comparison to other methods, outperforming the second best with a large margin of **1.42** dB in PSNR.

TDAE [29] also employs multiple STNs to align LR face images and achieves second best results. Note that TDAE employs three networks to super-resolve face images, which is much larger than our network. This 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. Thus, our method is able to achieve better quantitative results.

4.5. Conclusions

We introduced an attribute embedded discriminative network to super-resolve very low-resolution (16×16 pixels) unaligned face images $8\times$ 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. After training our network, it is not only able to super-resolve LR faces but also manipulate the upsampled results by adjusting the attribute information.

References

[1] M. Arjovsky, S. Chintala, and L. Bottou. Wasserstein gan. *arXiv preprint arXiv:1701.07875*, 2017. 3

[2] S. Baker and T. Kanade. Limits on super-resolution and how to break them. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 24(9):1167–1183, 2002. 2

[3] D. Berthelot, T. Schumm, and L. Metz. Began: Boundary equilibrium generative adversarial networks. *arXiv preprint arXiv:1703.10717*, 2017. 3

[4] Q. Cao, L. Lin, Y. Shi, X. Liang, and G. Li. Attention-aware face hallucination via deep reinforcement learning. In *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 690–698, 2017. 1

[5] E. Denton, S. Chintala, A. Szlam, and R. Fergus. Deep Generative Image Models using a Laplacian Pyramid of Adversarial Networks. In *Advances In Neural Information Processing Systems (NIPS)*, pages 1486–1494, 2015. 3

[6] B. Fasel and J. Luettin. Automatic facial expression analysis: a survey. *Pattern recognition*, 36(1):259–275, 2003. 1

[7] I. Goodfellow, J. Pouget-Abadie, and M. Mirza. Generative Adversarial Networks. In *Advances in Neural Information Processing Systems (NIPS)*, pages 2672–2680, 2014. 3

[8] G. Hinton. Neural Networks for Machine Learning Lecture 6a: Overview of mini-batch gradient descent Reminder: The error surface for a linear neuron. 5

[9] M. Jaderberg, K. Simonyan, A. Zisserman, et al. Spatial transformer networks. In *Advances in Neural Information Processing Systems (NIPS)*, pages 2017–2025, 2015. 4

[10] J. Johnson, A. Alahi, and L. Fei-Fei. Perceptual losses for real-time style transfer and super-resolution. In *European Conference on Computer Vision (ECCV)*, 2016. 4

[11] J. Kim, J. Kwon Lee, and K. Mu Lee. Accurate image super-resolution using very deep convolutional networks. In *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1646–1654, 2016. 1, 2, 6, 7, 8

[12] S. Kolouri and G. K. Rohde. Transport-based single frame super resolution of very low resolution face images. In *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4876–4884, 2015. 2

[13] Y. Li, C. Cai, G. Qiu, and K. M. Lam. Face hallucination based on sparse local-pixel structure. *Pattern Recognition*, 47(3):1261–1270, 2014. 2

[14] C. Liu, H. Y. Shum, and W. T. Freeman. Face hallucination: Theory and practice. *International Journal of Computer Vision*, 75(1):115–134, 2007. 2

[15] C. Liu, J. Yuen, and A. Torralba. Sift flow: Dense correspondence across scenes and its applications. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 33(5):978–994, 2011. 2

[16] J. Long, E. Shelhamer, and T. Darrell. Fully convolutional networks for semantic segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3431–3440, 2015. 3

[17] G. Perarnau, J. van de Weijer, B. Raducanu, and J. M. Álvarez. Invertible Conditional GANs for image editing. In *NIPS Workshop on Adversarial Training*, 2016. 3, 6

[18] S. Reed, Z. Akata, X. Yan, L. Logeswaran, B. Schiele, and H. Lee. Generative adversarial text to image synthesis. *arXiv preprint arXiv:1605.05396*, 2016. 3

[19] W. Shen and R. Liu. Learning residual images for face attribute manipulation. *arXiv preprint arXiv:1612.05363*, 2016. 3

[20] K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014. 4

[21] M. F. Tappen and C. Liu. A Bayesian Approach to Alignment-Based Image Hallucination. In *Proceedings of European Conference on Computer Vision (ECCV)*, volume 7578, pages 236–249, 2012. 2

[22] X. Wang and X. Tang. Hallucinating face by eigen transformation. *IEEE Transactions on Systems, Man and Cybernetics Part C: Applications and Reviews*, 35(3):425–434, 2005. 2

[23] C. Q. Xiang Ma, Junping Zhang. Hallucinating face by position-patch. *Pattern Recognition*, 43(6):2224–2236, 2010. 2, 6, 7, 8

[24] X. Yan, J. Yang, K. Sohn, and H. Lee. Attribute2Image: Conditional Image Generation from Visual Attributes. *arXiv:1512.00570*, page 10, 2015. 2, 3, 5

[25] C. Y. Yang, S. Liu, and M. H. Yang. Structured face hallucination. In *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1099–1106, 2013. 2

[26] J. Yang, J. Wright, T. S. Huang, and Y. Ma. Image super-resolution via sparse representation. *IEEE transactions on image processing*, 19(11):2861–73, 2010. 2

[27] X. Yu and F. Porikli. Ultra-resolving face images by discriminative generative networks. In *Proceedings of European Conference on Computer Vision (ECCV)*, pages 318–333, 2016. 1, 2, 3, 4, 6, 7

[28] X. Yu and F. Porikli. Face hallucination with tiny unaligned images by transformative discriminative neural networks. In *Thirty-First AAAI Conference on Artificial Intelligence*, 2017. 2, 8

[29] X. Yu and F. Porikli. Hallucinating very low-resolution unaligned and noisy face images by transformative discriminative autoencoders. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3760–3768, 2017. 1, 2, 3, 4, 5, 6, 7, 8

[30] M. D. Zeiler and R. Fergus. Visualizing and understanding convolutional networks. In *European Conference on Computer Vision (ECCV)*, pages 818–833, 2014. 4

[31] 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. *arXiv preprint arXiv:1612.03242*, 2016. 2, 3, 5

[32] J. Zhao, M. Mathieu, and Y. LeCun. Energy-based generative adversarial network. *arXiv preprint arXiv:1609.03126*, 2016. 3

[33] W. Zhao, R. Chellappa, P. J. Phillips, and A. Rosenfeld. Face recognition: A literature survey. *ACM computing surveys (CSUR)*, 35(4):399–458, 2003. 1

[34] E. Zhou and H. Fan. Learning Face Hallucination in the Wild. In *Twenty-Ninth AAAI Conference on Artificial Intelligence*, pages 3871–3877, 2015. 2

[35] S. Zhu, S. Liu, C. C. Loy, and X. Tang. Deep cascaded bi-network for face hallucination. In *Proceedings of European Conference on Computer Vision (ECCV)*, pages 614–630, 2016. 1, 2, 6, 7, 8

[36] X. W. Ziwei Liu, Ping Luo and X. Tang. Deep learning face attributes in the wild. In *Proceedings of International Conference on Computer Vision (ICCV)*, pages 3730–3738, 2015. 6