Face Super-resolution Guided by Facial Component Heatmaps

Xin Yu[1], Basura Fernando[1], Bernard Ghanem[2], Fatih Porikli[1], Richard Hartley[1]

[1]Australian National University,[2]King Abdullah University of Science and Technology
{xin.yu,fatih.porikli,richard.hartley}@anu.edu.au, Basuraf@gmail.com,
Bernard.Ghanem@kaust.edu.sa

Abstract. State-of-the-art face super-resolution methods leverage deep convolutional neural networks to learn a mapping between low-resolution (LR) facial patterns and their corresponding high-resolution (HR) counterparts by exploring local appearance information. However, most of these methods do not account for facial structure and suffer from degradations due to large pose variations and misalignments. In this paper, we propose a method that explicitly incorporates structural information of faces into the face super-resolution process by using a multi-task convolutional neural network (CNN). Our CNN has two branches: one for super-resolving face images and the other branch for predicting salient regions of a face coined *facial component heatmaps*. These heatmaps encourage the upsampling stream to generate super-resolved faces with higher-quality details. Our method not only uses low-level information (*i.e.*, intensity similarity), but also middle-level information (*i.e.*, face structure) to further explore spatial constraints of facial components from LR inputs images. Therefore, we are able to super-resolve very small unaligned face images (16×16 pixels) with a large upscaling factor of 8×, while preserving face structure. Extensive experiments demonstrate that our network achieves superior face hallucination results and outperforms the state-of-the-art.

Keywords: Face, super-resolution, hallucination, facial component localization, multi-task neural networks.

1 Introduction

Face images provide crucial clues for human observation as well as computer analysis [1,2]. However, the performance of most existing facial analysis techniques, such as face alignment [3,4] and identification [5], degrades dramatically when the resolution of a face is adversely low. Face super-resolution (FSR) [8], also known as face hallucination, provides a viable way to recover a high-resolution (HR) face image from its low-resolution (LR) counterpart and has attracted increasing interest in recent years. Modern face hallucination methods employ deep learning [9,10,7,11,6,12,13,14,15,16] and achieve state-of-the-art performance. These methods explore image intensity correspondences between LR and HR faces from large-scale face datasets. Since near-frontal faces prevail in popular

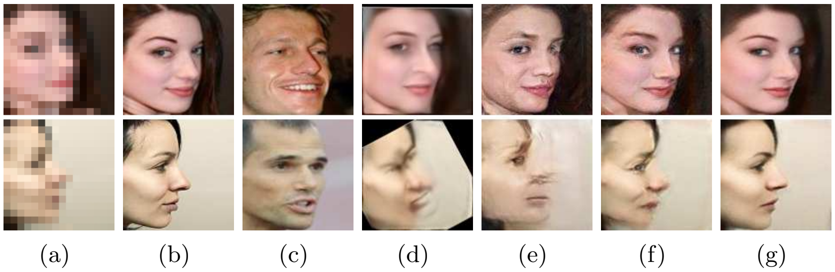

Fig. 1. Comparison of state-of-the-art face super-resolution methods on very low-resolution (LR) face images. Columns: (a) Unaligned LR inputs. (b) Original HR images. (c) Nearest Neighbors (NN) of aligned LR faces. Note that image intensities are used to find NN. (d) CBN [6]. (e) TDAE [7]. (f) TDAE†. We retrain the original TDAE with our training dataset. (g) Our results.

large-scale face datasets [17,18], deep learning based FSR methods may fail to super-resolve LR faces under large pose variations, as seen in the examples of Fig. 1. In fact, in these examples, the face structure has been distorted and facial details are not fully recovered by state-of-the-art super-resolution methods.

A naive idea to remedy this issue is to augment training data with large pose variations (*i.e.*, [19]) and then retrain the neural networks. As shown in Fig. 1(f), this strategy still leads to suboptimal results where facial details are missing or distorted due to erroneous localization of LR facial patterns. This limitation is common in intensity-based FSR methods that only exploit local intensity information in super-resolution and do not take face structure or poses into account. We postulate that methods that explicitly exploit information about the locations of facial components in LR faces have the capacity to improve super-resolution performance.

Another approach to super-resolve LR face images is to localize facial components in advance and then upsample them [20,6] progressively. However, localizing these facial components with high accuracy is generally a difficult task in very LR images, especially under large pose variations. As shown in Fig. 1(e), the method of Zhu *et al.* [6] fails to localize facial components accurately and produces an HR face with severe distortions. Therefore, directly detecting facial components or landmarks in LR faces is suboptimal and may lead to ghosting artifacts in the final result.

In contrast to previous methods, we propose a method that super-resolves LR face images while predicting face structure in a collaborative manner. Our intuition is that, although it is difficult to accurately detect facial landmarks in LR face images, it is possible to localize facial components (not landmarks) and identify the visibility of the components on the super-resolved faces or the intermediate upsampled feature maps because they can provide enough resolution for localization. Obtaining the locations of facial components can in turn facilitate face super-resolution.

Driven by this idea, we propose a multi-task deep neural network to upsample LR images. In contrast to the state-of-the-art FSR methods [7,6,12,13], our network not only super-resolves LR images but also estimates the spatial positions of their facial components. Then the estimated locations of the facial components are regarded as a guidance map which provides the face structure in super-resolution. Here, face structure refers to the locations and visibility of facial components as well as the relationship between them and we use heatmaps to represent the probability of the appearance of each component. Since the resolution of the input faces is small, (*i.e.*, 16×16 pixels), localizing facial components is also very challenging. Instead of detecting facial components in LR images, we opt to localize facial components on super-resolved feature maps. Specifically, we first super-resolve features of input LR images, and then employ a spatial transformer network [21] to align the feature maps. The upsampled feature maps are used to estimate the heatmaps of facial components. Since the feature maps are aligned, the same facial components may appear at the corresponding positions closely. This also provides an initial estimation for the component localization. Furthermore, we can also largely reduce the training examples for localizing facial components when input faces or feature maps are pre-aligned. For instance, we only use $30K$ LR/HR face image pairs for training our network, while a state-of-the-art face alignment method [4] requires about $230K$ images to train a landmark localization network.

After obtaining the estimated heatmaps of facial components, we concatenate them with the upsampled feature maps to infuse the spatial and visibility information of facial components into the super-resolution procedure. In this fashion, higher-level information beyond pixel-wise intensity similarity is explored and used as an additional prior in FSR. As shown in Fig. 1(g), our presented network is able to upsample LR faces in large poses while preserving the spatial structure of upsampled face images.

Overall, the contributions of our work can be summarized as:

- We present a novel multi-task framework to super-resolve LR face images of size 16×16 pixels by an upscaling factor of $8\times$, which not only exploits image intensity similarity but also explores the face structure prior in face super-resolution.
- We not only upsample LR faces but also estimate the face structure in the framework. Our estimated facial component heatmaps provide not only spatial information of facial components but also their visibility information, which cannot be deduced from pixel-level information.
- We demonstrate that the proposed two branches, *i.e.*, upsampling and facial component estimation branches, collaborate with each other in super-resolution, thus achieving better face hallucination performance.
- Due to the design of our network architecture, we are able to estimate facial component heatmaps from the upsampled feature maps, which provides enough resolutions and details for estimation. Furthermore, since the fea-

ture maps are aligned before heatmap estimation, we can largely reduce the number of training images to train the heatmap estimation branch.

To the best of our knowledge, our method is the first attempt to use a multi-task framework to super-resolve very LR face images. We not only focus on learning the intensity similarity mappings between LR and HR facial patterns, similar to [7,13,22], but also explore the face structure information from images themselves and employ it as an additional prior for super-resolution.

2 Related Work

Exploiting facial priors, such as spatial configuration of facial components, in face hallucination is the key factor different from generic super-resolution tasks. Based on the usage of the priors, face hallucination methods can be roughly grouped into global model based and part based approaches.

Global model based approaches aim at super-resolving an LR input image by learning a holistic appearance mapping such as PCA. Wang and Tang [23] learn subspaces from LR and HR face images respectively, and then reconstruct an HR output from the PCA coefficients of the LR input. Liu *et al.* [24] employ a global model for the super-resolution of LR face images but also develop a markov random field (MRF) to reduce ghosting artifacts caused by the misalignments in LR images. Kolouri and Rohde [25] employ optimal transport techniques to morph an HR output by interpolating exemplar HR faces. In order to learn a good global model, LR inputs are required to be precisely aligned and to share similar poses to the exemplar HR images. When large pose variations and misalignments exit in LR inputs, these methods are prone to produce severe artifacts.

Part based methods are proposed to super-resolve individual facial regions separately. They reconstruct the HR counterparts of LR inputs based on either reference patches or facial components in the training dataset. Baker and Kanade [26] search the best mapping between LR and HR patches and then use the matched HR patches to recover high-frequency details of aligned LR face images. Motivated by this idea, [22,27,28,29] average weighted position patches extracted from multiple aligned HR images to upsample aligned LR face images in either the image intensity domain or sparse coding domain. However, patch based methods also require LR inputs to be aligned in advance and may produce blocky artifacts when the upscaling factor is too large. Instead of using position patches, Tappen and Liu [30] super-resolve HR facial components by warping the reference HR images. Yang *et al.* [20] localize facial components in the LR images by a facial landmark detector and then reconstruct missing high-frequency details from similar HR reference components. Because facial component based methods need to extract facial parts in LR images and then align them to exemplar images accurately, their performance degrades dramatically when the resolutions of input faces become unfavorably small.

Recently, deep learning techniques have been applied to the face hallucination field and achieved significant progress. Yu and Porikli [10] present a discriminative generative network to hallucinate aligned LR face images. Their follow-up works [31,7] interweave multiple spatial transformer networks [21] with the deconvolutional layers to handle unaligned LR faces. Xu *et al.* [32] employ the framework of generative adversarial networks [33,34] to recover blurry LR face images by a multi-class discriminative loss. Dahl *et al.* [13] leverage the framework of PixelCNN [35] to super-resolve very low-resolution faces. Since the above deep convolutional networks only consider local information in super-resolution without taking the holistic face structure into account, they may distort face structure when super-resolving non-frontal LR faces. Zhu *et al.* [6] present a cascade bi-network, dubbed CBN, to localize LR facial components first and then upsample the facial components, but CBN may produce ghosting faces when localization errors occur. Concurrent to our work, the algorithms [15,14] also employ facial structure in face hallucination. In contrast to their works, we propose a multi-task network which can be trained in an end-to-end manner. In particular, our network not only estimates the facial heatmaps but also employs them for achieving high-quality super-resolved results.

3 Our Proposed Method

Our network mainly consists of two parts: a multi-task upsampling network and a discriminative network. Our multi-task upsampling network (MTUN) is composed of two branches: an upsampling branch and a facial component heatmap estimation branch (HEB). Figure 2 illustrates the overall architecture of our proposed network. The entire network is trained in an end-to-end fashion.

3.1 Facial Component Heatmap Estimation

When the resolution of input images is too small, facial components will be even smaller. Thus, it is very difficult for state-of-the-art facial landmark detectors to localize facial landmarks in very low-resolution images accurately. However, we propose to predict facial component heatmaps from super-resolved feature maps rather than localizing landmarks in LR input images, because the upsampled feature maps contain more details and their resolutions are large enough for estimating facial component heatmaps. Moreover, since 2D faces may exhibit a wide range of poses, such as in-plane rotations, out-of-plane rotations and scale changes, we may need a large number of images for training HEB. For example, Bulat and Tzimiropoulos [4] require over $200K$ training images to train a landmark detector, and there is still a gap between the accuracy of [4] and human labeling. To mitigate this problem, our intuition is that when the faces are roughly aligned, the same facial components lie in the corresponding positions closely. Thus, we employ a spatial transformer network (STN) to align the upsampled features before estimating heatmaps. In this way, we not only ease the heatmap estimation but also significantly reduce the number of training images used for learning HEB.

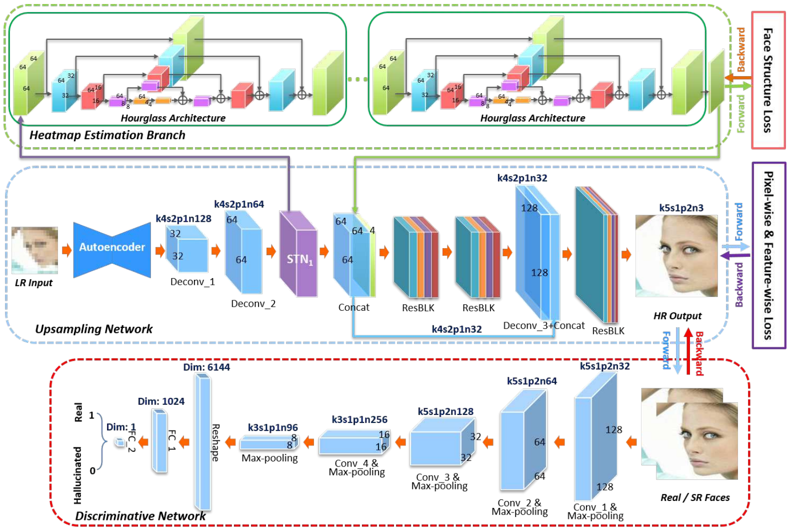

Fig. 2. The pipeline of our multi-task upsampling network. In the testing phase, the upsampling branch (blue block) and the heatmap estimation branch (green block) are used.

We use heatmaps instead of landmarks based on three reasons: (i) localizing each facial landmark individually is difficult in LR faces even for humans and erroneous landmarks would lead to distortions in the final results. On the contrary, it is much easier to localize each facial components as a whole. (ii) Even state-of-the-art landmark detectors may fail to output accurate positions in high-resolution images, such as in large pose cases. However, it is not difficult to estimate a region represented by a heatmap in those cases. (iii) Furthermore, our goal is to provide clues of the spatial positions and visibility of each component rather than the exact shape of each component. Using heatmaps as a probability map is more suitable for our purpose.

In this paper, we use four heatmaps to represent four components of a face, *i.e.*, eyes, nose, mouth and chain, respectively. We exploit 68 point facial landmarks to generate the ground-truth heatmaps. Specifically, each landmark is represented by a Gaussian kernel and the center of the kernel is the location of the landmark. By adjusting the standard variance of Gaussian kernels in accordance with the resolutions of feature maps or images, we can generate a heatmap for each component. The generated ground-truth heatmaps are shown in Fig. 3(c). Note that, when self-occlusions appear, some components are not visible and they will not appear in the heatmaps. In this way, heatmaps not only provides the locations of components but also their visibility in the original LR input images.

In order to estimate facial component heatmaps, we employ the stacked hourglass network architecture [36]. It exploits a repeated bottom-up and top-down

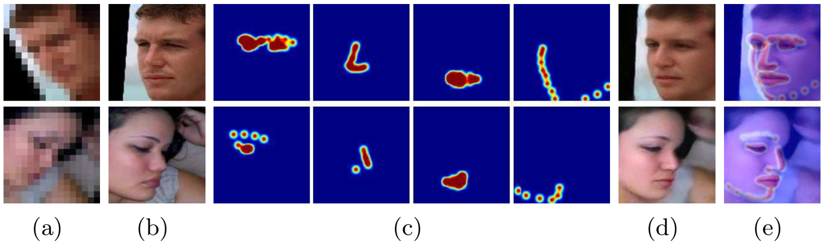

Fig. 3. Visualization of estimated facial component heatmaps. Columns: (a) Unaligned LR inputs. (b) HR images. (c) Ground-truth heatmaps generated from the landmarks of HR face images. (d) Our results. (e) The estimated heatmaps overlying over our super-resolved results. Note that, we overlap four estimated heatmaps together and upsample the heatmaps to fit our upsampled results.

fashion to process features across multiple scales and is able to capture various spatial relationships among different parts. As suggested in [36], we also use the intermediate supervision to improve the performance. The green block in Fig. 2 illustrates our facial component heatmap estimation branch. We feed the aligned feature maps to HEB and then concatenate the estimated heatmaps with the upsampled feature maps for super-resolving facial details. In order to illustrate the effectiveness of HEB, we resize and then overlay the estimated heatmaps over the output images as visible in Fig. 3(e). The ground-truth heatmaps are shown in Fig. 3(c) for comparison.

3.2 Network Architecture

Multi-task Upsampling Network: Figure 2 illustrates the architecture of our proposed multi-task upsampling network (MTUN) in the blue and green blocks. MTUN consists of two branches: an upsampling branch (blue block) and a facial component heatmap estimation branch (green block). The upsampling branch firstly super-resolves features of LR input images and then aligns the feature maps. When the resolution of the feature maps is large enough, the upsampled feature maps are fed into HEB to estimate the locations and visibility of facial components. Thus we obtain the heatmaps of the facial components of LR inputs. The estimated heatmaps are then concatenated with the upsampled feature maps to provide the spatial positions and visibility information of facial components for super-resolution.

In the upsampling branch, the network is composed of a convolutional autoencoder, deconvolutional layers and an STN. The convolutional autoencoder is designed to extract high-frequency details from input images while removing image noise before upsampling and alignment, thus increasing the super-resolution performance. The deconvolutional layers are employed to super-resolve the feature maps. Since input LR faces undergo in-plane rotations, translations and

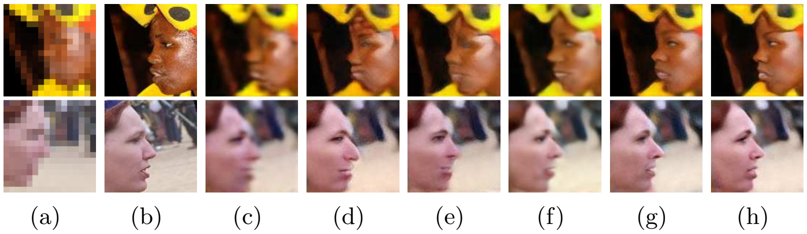

Fig. 4. Comparisons of different losses for the super-resolution. Columns: (a) Unaligned LR inputs. (b) Original HR images. (c) $\mathcal{L}_p$. (d) $\mathcal{L}_p + \mathcal{L}_f$. (e) $\mathcal{L}_p + \mathcal{L}_f + \mathcal{L}_{\mathcal{U}}$. (f) $\mathcal{L}_p + \mathcal{L}_h$. (g) $\mathcal{L}_p + \mathcal{L}_f + \mathcal{L}_h$. (h) $\mathcal{L}_p + \mathcal{L}_f + \mathcal{L}_{\mathcal{U}} + \mathcal{L}_h$. For simplicity, we omit the trade-off weights.

scale changes, STN is employed to compensate for those affine transformations, thus facilitating facial component heatmap estimation.

After obtaining aligned upsampled feature maps, those feature maps are used to estimate facial component heatmaps by an HEB. We construct our HEB by a stacked hourglass architecture [36], which consists of residual blocks and upsampling layers, as shown in the green block of Fig. 2.

Our multi-task network aims at super-resolving input face images as well as predicting heatmaps of facial components in the images. As seen in Fig. 4(c), when we only use the upsampling branch to super-resolve faces without using HEB, the facial details are blurred and some facial components, *e.g.*, mouth and nose, are distorted in large poses. Furthermore, the heatmap supervision also forces STN to align the upsampled features more accurately, thus improving super-resolution performance. Therefore, these two tasks collaborate with each other and benefit from each other as well. As shown in Fig. 4(f), our multi-task network achieves better super-resolved results.

Discriminative Network: Recent works [10,7,32,37] demonstrate that only using Euclidean distance (ℓ_2 loss) between the upsampled faces and the ground-truth HR faces tends to output over-smoothed results. Therefore, we incorporate a discriminative objective into our network to force super-resolved HR face images to lie on the manifold of real face images.

As shown in the red block of Fig. 2, the discriminative network is constructed by convolutional layers and fully connected layers similar to [34]. It is employed to determine whether an image is sampled from real face images or hallucinated ones. The discriminative loss, also known as adversarial loss, is back-propagated to update our upsampling network. In this manner, we can super-resolve more authentic HR faces, as shown in Fig. 4(h).

3.3 Loss Function

Pixel-wise Loss: Since the upsampled HR faces should be similar to the input LR faces in terms of image intensities, we employ the Euclidean distance, also known as pixel-wise ℓ_2 loss, to enforce this similarity as follows:

$$\mathcal{L}_p(w) = \mathbb{E}_{(\hat{h}_i, h_i)\sim p(\hat{h}, h)} \|\hat{h}_i - h_i\|_F^2 = \mathbb{E}_{(l_i, h_i)\sim p(l, h)} \|\mathcal{U}_w(l_i) - h_i\|_F^2, \quad (1)$$

where $\hat{h}_i$ and $\mathcal{U}_w(l_i)$ both represent the upsampled faces by our MTUN, w is the parameters of MTUN, l_i and h_i denote the LR input image and its HR ground-truth counterpart respectively, $p(l, h)$ represents the joint distribution of the LR and HR face images in the training dataset, and $p(\hat{h}, h)$ indicates the joint distribution of the upsampled HR faces and their corresponding HR ground-truths.

Feature-wise Loss: As mentioned in [10,37,32], only using pixel-wise ℓ_2 loss will produce over-smoothed super-resolved results. In order to achieve high-quality visual results, we also constrain the upsampled faces to share the same features as their HR counterparts. The objective function is expressed as:

$$\mathcal{L}_f(w) = \mathbb{E}_{(\hat{h}_i, h_i)\sim p(\hat{h}, h)} \|\psi(\hat{h}_i) - \psi(h_i)\|_F^2 = \mathbb{E}_{(l_i, h_i)\sim p(l, h)} \|\psi(\mathcal{U}_w(l_i)) - \psi(h_i)\|_F^2, \quad (2)$$

where $\psi(\cdot)$ denotes feature maps of a layer in VGG-19 [38]. We use the layer ReLU32, which gives good empirical results in our experiments.

Discriminative Loss: Since super-resolution is inherently an under-determined problem, there would be many possible mappings between LR and HR images. Even imposing intensity and feature similarities may not guarantee that the upsampling network can output realistic HR face images. We employ a discriminative network to force the hallucinated faces to lie on the same manifold of real face images, and our goal is to make the discriminative network fail to distinguish the upsampled faces from real ones. Therefore, the objective function for the discriminative network $\mathcal{D}$ is formulated as:

$$\mathcal{L}_{\mathcal{D}}(d) = \mathbb{E}_{(\hat{h}_i, h_i)\sim p(\hat{h}, h)} \left[\log \mathcal{D}_d(h_i) + \log(1 - \mathcal{D}_d(\hat{h}_i))\right] \quad (3)$$

where d represents the parameters of the discriminative network $\mathcal{D}$, $p(h)$, $p(l)$ and $p(\hat{h})$ indicate the distributions of the real HR, LR and super-resolved faces respectively, and $\mathcal{D}_d(h_i)$ and $\mathcal{D}_d(\hat{h}_i)$ are the outputs of $\mathcal{D}$. To make our discriminative network distinguish the real faces from the upsampled ones, we maximize the loss $\mathcal{L}_{\mathcal{D}}(d)$ and the loss is back-propagated to update the parameters d.

In order to fool the discriminative network, our upsampling network should produce faces as much similar as real faces. Thus, the objective function of the upsampling network is written as:

$$\mathcal{L}_{\mathcal{U}}(w) = \mathbb{E}_{(\hat{h}_i)\sim p(\hat{h})} \left[\log \mathcal{D}_d(\hat{h}_i)\right] = \mathbb{E}_{l_i \sim p(l)} \left[\log \mathcal{D}_d(\mathcal{U}_w(l_i))\right]. \quad (4)$$

We minimize Eqn. 4 to make our upsampling network generate realistic HR face images. The loss $\mathcal{L}_{\mathcal{U}}(w)$ is back-propagated to update the parameters w.

Face Structure Loss: Unlike previous works [7,32,10], we not only employ image pixel information (*i.e.*, pixel-wise and feature-wise losses) but also explore the face structure information during super-resolution. In order to achieve spatial relationships between facial components and their visibility, we estimate the heatmaps of facial components from the upsampled features as follows:

$$\mathcal{L}_h(w) = \mathbb{E}_{(l_i,h_i)\sim p(l,h)} \frac{1}{M} \sum_{k=1}^{M} \frac{1}{N} \sum_{j=1}^{N} \|\mathcal{H}_j^k(h_i) - \mathcal{H}_j^k(\tilde{\mathcal{U}}_w(l_i))\|_2^2, \tag{5}$$

where M is the number of the facial components, N indicates the number of Gaussian kernels in each component, $\tilde{\mathcal{U}}_w(l_i)$ is the intermediate upsampled feature maps by $\mathcal{U}$, $\mathcal{H}_j^k$ represents the j-th kernel in the k-th heatmap, and $\mathcal{H}_j^k(h_i)$ and $\mathcal{H}_j^k(\tilde{\mathcal{U}}_w(l_i))$ denote the ground-truth and estimated kernel positions in the heatmaps. Due to self-occlusions, some parts of facial components are invisible and thus N varies according to the visibility of those kernels in the heatmaps. Note that, the parameters w not only refer to the parameters in the upsampling branch but also those in the heatmap estimation branch.

Training Details: In training our discriminative network $\mathcal{D}$, we only use the loss $\mathcal{L}_{\mathcal{D}}(d)$ in Eqn. 3 to update the parameters d. Since the discriminative network aims at distinguishing upsampled faces from real ones, we maximize $\mathcal{L}_{\mathcal{D}}(d)$ by stochastic gradient ascent.

In training our multi-task upsampling network $\mathcal{U}$, multiple losses, *i.e.*, $\mathcal{L}_p$, $\mathcal{L}_f$, $\mathcal{L}_{\mathcal{U}}$ and $\mathcal{L}_h$, are involved to update the parameters w. Therefore, in order to achieve authentic super-resolved HR face images, the objective function $\mathcal{L}_{\mathcal{T}}$ for training the upsampling network $\mathcal{U}$ is expressed as:

$$\mathcal{L}_{\mathcal{T}} = \mathcal{L}_p + \alpha\mathcal{L}_f + \beta\mathcal{L}_{\mathcal{U}} + \mathcal{L}_h, \tag{6}$$

where α, β are the trade-off weights. Since our goal is to recover HR faces in terms of appearance similarity, we set α and β to 0.01. We minimize $\mathcal{L}_{\mathcal{T}}$ by stochastic gradient descent. Specifically, we use RMSprop optimization algorithm [39] to update the parameters w and d. The discriminative network and upsampling network are trained in an alternating fashion. The learning rate r is set to 0.001 and multiplied by 0.99 after each epoch. We use the decay rate 0.01 in RMSprop.

3.4 Implementation Details

In our multi-task upsampling network, we employ similarity transformation estimated by STN to compensate for in-plane misalignments. In Fig. 2, STN is built by convolutional and ReLU layers (Conv+ReLU), max-pooling layers with a stride 2 (MP2) and fully connected layers (FC). Specifically, our STN is composed of MP2, Conv+ReLU (k5s1p0n20), MP2, Conv+ReLU (k5s1p0n20), MP2, FC+ReLU (from 80 to 20 dimensions) and FC (from 20 to 4 dimensions), where k, s and p indicate the sizes of filters, strides and paddings respectively, and n represents the channel number of the output feature maps. Our HEB is constructed by stacking four hourglass networks and we also apply intermediate supervision to the output of each hourglass network. The residual block is constructed by BN, ReLU, Conv (k3s1p1nN$_i$), BN, ReLU and Conv (k1s1p0nN$_o$), where N$_i$ and N$_o$ indicate the channel numbers of input and output feature maps.

In the experimental part, some algorithms require alignment of LR inputs, *e.g.*, [22]. Hence, we employ an STN$_0$ to align the LR face images to the upright position. STN$_0$ is composed of Conv+ReLU (k5s1p0n64), MP2, Conv+ReLU (k5s1p0n20), FC+ReLU (from 80 to 20 dimensions), and FC (from 20 to 4 dimensions).

4 Experimental Results

In order to evaluate the performance of our proposed network, we compare with the state-of-the-art methods [40,37,22,6,7] qualitatively and quantitatively. Kim *et al.* [40] employ a very deep convolutional network to super-resolve generic images, known as VDSR. Ledig *et al.*'s method [37], dubbed SRGAN, is a generic super-resolution method, which employs the framework of generative adversarial networks and is trained with pixel-wise and adversarial losses. Ma *et al.*'s method [22] exploits position patches in the dataset to reconstruct HR images. Zhu *et al.*'s method [6], known as CBN, first localizes facial components in LR input images and then super-resolves the localized facial parts. Yu and Porikli [7] upsample very low-resolution unaligned face images by a transformative discriminative autoencoder (TDAE).

4.1 Dataset

Although there are large-scale face datasets [17,18], they do not provide structural information, *i.e.*, facial landmarks, for generating ground-truth heatmaps. In addition, we found that most of faces in the celebrity face attributes (CelebA) dataset [17], as one of the largest face datasets, are near-frontal. Hence, we use images from the Menpo facial landmark localization challenges (Menpo) [19] as well as images from CelebA to generate our training dataset. Menpo [19] provides face images in different poses and their corresponding 68 point landmarks or 39 point landmarks when some facial parts are invisible. Because Menpo only contains about $8K$ images, we also collect another $22K$ images from CelebA. We crop the aligned faces and then resize them to 128×128 pixels as our HR ground-truth images h_i. Our LR face images l_i are generated by transforming and downsampling the HR faces to 16×16 pixels. We choose 80 percent of image pairs for training and 20 percent of image pairs for testing.

4.2 Qualitative Comparisons with SoA

Since [22] needs to align input LR faces before super-resolution and [7] automatically outputs upright HR face images, we align LR faces by a spatial transformer network STN$_0$ for a fair comparison and better illustration. The upright HR ground-truth images are also shown for comparison.

Bicubic interpolation only upsamples image intensities from neighboring pixels instead of generating new contents for new pixels. As shown in Fig. 5(c), bicubic interpolation fails to generate facial details.

Table 1. Quantitative comparisons on the entire test dataset

Methods	Bicubic	VDSR[40]	SRGAN[37]	Ma[22]	CBN[6]	TDAE[7]	TDAE†	Ours†	Ours‡	Ours
PSNR	18.83	18.65	18.57	18.66	18.49	18.87	21.39	22.69	22.83	**23.14**
SSIM	0.57	0.57	0.55	0.53	0.55	0.52	0.62	0.66	0.65	**0.68**

VDSR only employs a pixel-wise ℓ_2 loss in training and does not provide an upscaling factor 8×. We apply VDSR to an LR face three times by an upscaling factor 2×. As shown in Fig. 5(d), VDSR fails to generate authentic facial details and the super-resolved faces are still blurry.

SRGAN is able to super-resolve an image by an upscaling factor of 8× directly and employs an adversarial loss to enhance details. However, SRGAN does not take the entire face structure into consideration and thus outputs ringing artifacts around facial components, such as eyes and mouth, as shown in Fig. 5(e).

Ma *et al.*'s method is sensitive to misalignments in LR inputs because it hallucinates HR faces by position-patches. As seen in Fig. 5(f), obvious blur artifacts and ghosting facial components appear in the hallucinated faces. As the upscaling factor increases, the correspondences between LR and HR patches become inconsistent. Thus, the super-resolved face images suffer severe blocky artifacts.

CBN first localizes facial components in LR faces and then super-resolves facial details and entire face images by two branches. As shown in Fig. 5(g), CBN generates facial components inconsistent with the HR ground-truth images in near-frontal faces and fails to generate realistic facial details in large poses. This indicates that it is difficult to localize facial components in LR faces accurately.

TDAE employs ℓ_2 and adversarial losses and is trained with near-frontal faces. Due to various poses in our testing dataset, TDAE fails to align faces in large poses. For a fair comparison, we retrain the decoder of TDAE with our training dataset. As visible in Fig. 5(h), TDAE still fails to realistic facial details due to various poses and misalignments.

Our method reconstructs authentic facial details as shown in Fig. 5(i). Our facial component heatmaps not only facilitate alignment but also provide spatial configuration of facial components. Therefore, our method is able to produce visually pleasing HR facial details similar to the ground-truth faces while preserving face structure. (More results are shown in the supplementary materials.)

4.3 Quantitative Comparisons with SoA

We also evaluate the performance of all methods quantitatively on the entire test dataset by the average PSNR and the structural similarity (SSIM) scores. Table 1 indicates that our method achieves superior performance compared to other methods, *i.e.*, outperforming the second best with a large margin of **1.75** dB in PSNR. Note that, the average PSNR of TDAE for its released model is

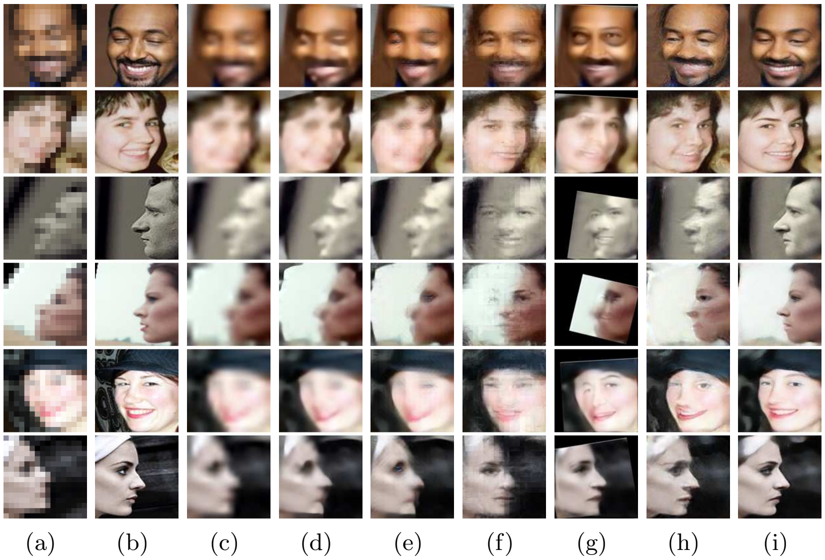

Fig. 5. Comparisons with the state-of-the-art methods. (a) Unaligned LR inputs. (b) Original HR images. (c) Bicubic interpolation. (d) Kim *et al.*'s method [40] (VDSR). (e) Ledig *et al.*'s method [37] (SRGAN). (f) Ma *et al.*'s method [22]. (g) Zhu *et al.*'s method [6] (CBN). (h) Yu and Porikli's method [7] (TDAE). Since TDAE is not trained with near-frontal face images, we retrain it with our training dataset. (i) Our method.

only 18.87 dB because it is trained with near-frontal faces. Even after retaining TDAE, indicated by TDAE†, its performance is still inferior to our results. It also implies that our method localizes facial components and aligns LR faces more accurately with the help of our estimated heatmaps.

5 Analysis and Discussion

Effectiveness of HEB: As shown in Fig. 4(c), Fig. 4(d) and Fig. 4(e), we demonstrate that the visual results without HEB suffer from distortion and blur artifacts. By employing HEB, we can localize the facial components as seen in Fig. 3, and then recover realistic facial details. Furthermore, HEB provides the spatial locations of facial components and an additional constraint for face alignments. Thus we achieve higher reconstruction performance as shown in Tab. 3.

Feature Sizes for HEB: In our network, there are several layers which can be used to estimate facial component heatmaps, *i.e.*, feature maps of sizes 16, 32, 64 and 128, respectively. We employ HEB at different layers and demonstrate the influence of the sizes of feature maps. Due to GPU memory limitations, we only compare the super-resolution performance of using features of sizes

Table 2. Ablation study of HEB

	Position		Depth			
	$\mathcal{R}16$	$\mathcal{R}32$	$\mathcal{S}1$	$\mathcal{S}2$	$\mathcal{S}3$	$\mathcal{S}4$
PSNR	21.97	21.98	22.32	22.91	22.93	**23.14**
SSIM	0.63	0.64	0.64	0.67	0.67	**0.68**

Table 3. Ablation study on the loss

	w/o $\mathcal{L}_h$			w/ $\mathcal{L}_h$		
	$\mathcal{L}_p$	$\mathcal{L}_{p+f}$	$\mathcal{L}_{p+f+\mathcal{U}}$	$\mathcal{L}_p$	$\mathcal{L}_{p+f}$	$\mathcal{L}_{p+f+\mathcal{U}}$
PSNR	21.43	21.57	21.55	23.23	23.35	23.14
SSIM	0.66	0.66	0.65	0.69	0.69	0.68

16 ($\mathcal{R}16$), 32 ($\mathcal{R}32$) and 64 ($\mathcal{S}4$) to estimate heatmaps. As shown in Tab. 2, as the resolution of feature maps increases, we obtain better super-resolution performance. Therefore, we employ the upsampled feature maps of size 64×64 to estimate heatmaps.

Depths of HEB: Table 2 demonstrates the performance influenced by the stack number of hourglass networks. Due to the limitation of GPU memory, we only conduct our experiments on the stack number ranging from 1 to 4. As indicated in Tab. 2, the final performance improves as the stack number increases. Hence, we set the stack number to 4 for our HEB.

Loss Functions: Table 3 also indicates the influences of different losses on the super-resolution performance. As indicated in Tab. 3 and Fig. 4, using the face structure loss improves the super-resolved results qualitatively and quantitatively. The feature-wise loss improves the visual quality and the discriminative loss makes the hallucinated faces sharper and more realistic, as shown in Fig. 4(h).

Skip Connection and Autoencoder: Considering there are estimation errors in the heatmaps, fusing feature maps with erroneous heatmaps may lead to distortions in the final outputs. Hence, we employ a skip connection to correct the errors in Fig. 2. As indicated in Tab. 1, using the skip connection, we can improve the final quantitative result by 0.45 dB in PSNR. The result without using skip connection is indicated by Ours†. We also remove our autoencoder and upsample LR inputs directly and the result is denoted as Ours‡. As shown in Tab. 1, we achieve 0.31 dB improvement with the help of the autoencoder.

6 Conclusion

We present a novel multi-task upsampling network to super-resolve very small LR face images. We not only employ the image appearance similarity but also exploit the face structure information estimated from LR input images themselves in the super-resolution. With the help of our facial component heatmap estimation branch, our method super-resolves faces in different poses without distortions caused by erroneous facial component localization in LR inputs.

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

References

1. Fasel, B., Luettin, J.: Automatic facial expression analysis: a survey. Pattern recognition **36**(1) (2003) 259–275 1
2. Zhao, W., Chellappa, R., Phillips, P.J., Rosenfeld, A.: Face recognition: A literature survey. ACM computing surveys (CSUR) **35**(4) (2003) 399–458 1
3. Xiong, X., De la Torre, F.: Supervised descent method and its applications to face alignment. In: Proceedings of the IEEE conference on computer vision and pattern recognition (CVPR). (2013) 532–539 1
4. Bulat, A., Tzimiropoulos, G.: How far are we from solving the 2d & 3d face alignment problem? (and a dataset of 230,000 3d facial landmarks). In: International Conference on Computer Vision (ICCV). (2017) 1, 3, 5
5. Taigman, Y., Yang, M., Ranzato, M., Wolf, L.: Deepface: Closing the gap to human-level performance in face verification. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). (2014) 1701–1708 1
6. Zhu, S., Liu, S., Loy, C.C., Tang, X.: Deep cascaded bi-network for face hallucination. In: Proceedings of European Conference on Computer Vision (ECCV). (2016) 614–630 1, 2, 3, 5, 11, 12, 13
7. Yu, X., Porikli, F.: 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). (2017) 3760–3768 1, 2, 3, 4, 5, 8, 9, 11, 12, 13
8. Baker, S., Kanade, T.: Hallucinating faces. In: Proceedings of 4th IEEE International Conference on Automatic Face and Gesture Recognition, FG 2000. (2000) 83–88 1
9. Zhou, E., Fan, H.: Learning Face Hallucination in the Wild. In: Twenty-Ninth AAAI Conference on Artificial Intelligence. (2015) 3871–3877 1
10. Yu, X., Porikli, F.: Ultra-resolving face images by discriminative generative networks. In: Proceedings of European Conference on Computer Vision (ECCV). (2016) 318–333 1, 4, 8, 9
11. Yu, X., Porikli, F.: Imagining the unimaginable faces by deconvolutional networks. IEEE Transactions on Image Processing (2018) 1
12. Cao, Q., Lin, L., Shi, Y., Liang, X., Li, G.: Attention-aware face hallucination via deep reinforcement learning. In: Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR). (2017) 690–698 1, 3
13. Dahl, R., Norouzi, M., Shlens, J.: Pixel recursive super resolution. In: International Conference on Computer Vision (ICCV). (2017) 5439–5448 1, 3, 4, 5
14. Bulat, A., Tzimiropoulos, G.: Super-fan: Integrated facial landmark localization and super-resolution of real-world low resolution faces in arbitrary poses with gans. arXiv preprint arXiv:1712.02765 (2017) 1, 5
15. Chen, Y., Tai, Y., Liu, X., Shen, C., Yang, J.: Fsrnet: End-to-end learning face super-resolution with facial priors. arXiv preprint arXiv:1711.10703 (2017) 1, 5
16. Yu, X., Fernando, B., Hartley, R., Porikli, F.: Super-resolving very low-resolution face images with supplementary attributes. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). (2018) 908–917 1
17. Ziwei Liu, Ping Luo, X.W., Tang, X.: Deep learning face attributes in the wild. In: Proceedings of International Conference on Computer Vision (ICCV). (2015) 3730–3738 2, 11

18. Huang, G.B., Ramesh, M., Berg, T., Learned-Miller, E.: Labeled faces in the wild: A database for studying face recognition in unconstrained environments. Technical Report 07-49, University of Massachusetts, Amherst (2007) 2, 11
19. Zafeiriou, S., Trigeorgis, G., Chrysos, G., Deng, J., Shen, J.: The menpo facial landmark localisation challenge: A step towards the solution. In: Proceedings of the IEEE conference on computer vision and pattern recognition workshops (CVPRW). (2017) 2116–2125 2, 11
20. Yang, C.Y., Liu, S., Yang, M.H.: Structured face hallucination. In: Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR). (2013) 1099–1106 2, 4
21. Jaderberg, M., Simonyan, K., Zisserman, A., et al.: Spatial transformer networks. In: Advances in Neural Information Processing Systems (NIPS). (2015) 2017–2025 3, 5
22. Xiang Ma, Junping Zhang, C.Q.: Hallucinating face by position-patch. Pattern Recognition **43**(6) (2010) 2224–2236 4, 11, 12, 13
23. Wang, X., Tang, X.: Hallucinating face by eigen transformation. IEEE Transactions on Systems, Man and Cybernetics Part C: Applications and Reviews **35**(3) (2005) 425–434 4
24. Liu, C., Shum, H.Y., Freeman, W.T.: Face hallucination: Theory and practice. International Journal of Computer Vision **75**(1) (2007) 115–134 4
25. Kolouri, S., Rohde, G.K.: 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). (2015) 4876–4884 4
26. Baker, S., Kanade, T.: Limits on super-resolution and how to break them. IEEE Transactions on Pattern Analysis and Machine Intelligence **24**(9) (2002) 1167–1183 4
27. Yang, J., Wright, J., Huang, T.S., Ma, Y.: Image super-resolution via sparse representation. IEEE transactions on image processing **19**(11) (2010) 2861–73 4
28. Li, Y., Cai, C., Qiu, G., Lam, K.M.: Face hallucination based on sparse local-pixel structure. Pattern Recognition **47**(3) (2014) 1261–1270 4
29. Jin, Y., Bouganis, C.S.: Robust multi-image based blind face hallucination. In: Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR). (2015) 5252–5260 4
30. Tappen, M.F., Liu, C.: A Bayesian Approach to Alignment-Based Image Hallucination. In: Proceedings of European Conference on Computer Vision (ECCV). Volume 7578. (2012) 236–249 4
31. Yu, X., Porikli, F.: Face hallucination with tiny unaligned images by transformative discriminative neural networks. In: Thirty-First AAAI Conference on Artificial Intelligence. (2017) 5
32. Xu, X., Sun, D., Pan, J., Zhang, Y., Pfister, H., Yang, M.H.: Learning to super-resolve blurry face and text images. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (ICCV). (2017) 251–260 5, 8, 9
33. Goodfellow, I., Pouget-Abadie, J., Mirza, M.: Generative Adversarial Networks. In: Advances in Neural Information Processing Systems (NIPS). (2014) 2672–2680 5
34. Radford, A., Metz, L., Chintala, S.: Unsupervised Representation Learning with Deep Convolutional Generative Adversarial Networks. arXiv:1511.06434 (2015) 1–15 5, 8
35. Van Den Oord, A., Kalchbrenner, N., Kavukcuoglu, K.: Pixel recurrent neural networks. In: Proceedings of International Conference on International Conference on Machine Learning (ICML). (2016) 1747–1756 5

36. Newell, A., Yang, K., Deng, J.: Stacked hourglass networks for human pose estimation. In: European Conference on Computer Vision (ECCV). (2016) 483–499 6, 7, 8
37. Ledig, C., Theis, L., Huszár, F., Caballero, J., Cunningham, A., Acosta, A., Aitken, A., Tejani, A., Totz, J., Wang, Z., et al.: Photo-realistic single image super-resolution using a generative adversarial network. In: Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR). (2017) 4681–4690 8, 9, 11, 12, 13
38. Simonyan, K., Zisserman, A.: Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:1409.1556 (2014) 9
39. Hinton, G.: Neural Networks for Machine Learning Lecture 6a: Overview of mini-batch gradient descent Reminder: The error surface for a linear neuron 10
40. Kim, J., Kwon Lee, J., Mu Lee, K.: Accurate image super-resolution using very deep convolutional networks. In: Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR). (2016) 1646–1654 11, 12, 13