International Journal of Computer Vision manuscript No.
(will be inserted by the editor)

Hallucinating Unaligned Face Images by Multiscale Transformative Discriminative Networks

Xin Yu · Fatih Porikli · Basura Fernando · Richard Hartley

Received: date / Accepted: date

Abstract Conventional face hallucination methods heavily rely on accurate alignment of low-resolution (LR) faces before upsampling them. Misalignment often leads to deficient results and unnatural artifacts for large upscaling factors. However, due to the diverse range of poses and different facial expressions, aligning an LR input image, in particular when it is tiny, is severely difficult. In addition, when the resolutions of LR input images vary, previous deep neural network based face hallucination methods require the interocular distances of input face images to be similar to the ones in the training datasets. Downsampling LR input faces to a required resolution will lose high-frequency information of the original input images. This may lead to suboptimal super-resolution performance for the state-of-the-art face hallucination networks. To overcome these challenges, we present an end-to-end multiscale transformative discriminative neural network (MTDN) devised for super-resolving unaligned and very small face images of different resolutions ranging from 16×16 to 32×32 pixels in a unified framework. Our proposed network embeds spatial transformation layers to allow local receptive fields to line-up with similar spatial supports, thus obtaining a better mapping between LR and HR facial patterns. Furthermore, we incorporate a class-specific loss designed to classify upright realistic faces in our objective through a successive discriminative network to improve the alignment and upsampling performance with semantic information. Extensive experiments on a large face dataset show that the proposed method significantly outperforms the state-of-the-art.

Keywords Face hallucination · Super-resolution · Multiscale · Transformative discriminative network

1 Introduction

Face images provide vital information for visual perception and identity analysis. Nonetheless, when the resolution of the face image is very small (*e.g.* in typical surveillance videos), there is little information that can be inferred from it. Very low-resolution (LR) face images not only degrade the performance of the recognition systems but also impede human interpretation. This challenge motivates the reconstruction of high-resolution (HR) images from given LR counterparts, known as face hallucination, and has attracted increasing interest in recent years.

Previous face hallucination methods based on holistic appearance models (Liu et al, 2001; Baker and Kanade, 2002; Wang and Tang, 2005; Liu et al, 2007; Hennings-Yeomans et al, 2008; Ma et al, 2010; Yang et al, 2010; Li et al, 2014; Arandjelović, 2014; Kolouri and Rohde, 2015) demand LR faces to be precisely aligned beforehand. However, aligning LR faces to appearance models is not a straightforward task itself, and more often, it requires expert feedback when the input image is small. Regarding pose and expression variations naturally exist in LR face images, aligning LR faces by state-of-the-art automatic alignment techniques (Zhu and Ramanan, 2012; Bulat and Tzimiropoulos, 2017) which usually assume facial landmarks are visible and detectable would be even more difficult. As a result, the performance of face hallucination degrades severely.

Xin Yu, Fatih Porikli, and Richard Hartley are with the Research School of Engineering, Australian National University, Canberra, Australia.
Basura Fernando is with Human Centric AI Programme, A*STAR Artificial Intelligence Initiative (A*AI), Singapore.
E-mail: xin.yu@anu.edu.au, fatih.porikli@anu.edu.au, fernando_basura@scei.a-star.edu.sg, richard.hartley@anu.edu.au

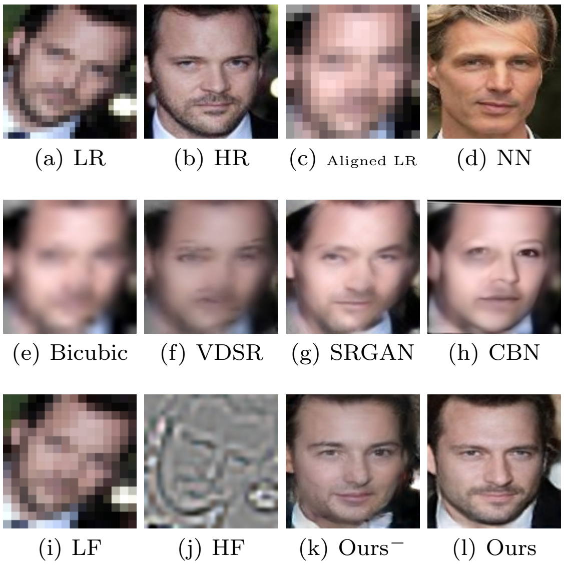


Fig. 1 Comparison of our method with the CNN based super-resolution. (a) The input 24 × 24 LR image. (b) The original 128 × 128 HR image. (c) Aligned LR image of (a). The resolution of the aligned LR image is 16 × 16 pixels since STN_0 only outputs a fixed resolution for all images. (d) The corresponding HR version of the nearest neighbor (NN) of (c) in the training set. (e) Bicubic interpolation of (c). (f) The image generated by a CNN based generic super-resolution, *i.e.*, VDSR (Kim et al, 2016a). We retrain VDSR with face images to better capture LR facial patterns in super-resolution. (g) The image upsampled by a GAN based generic super-resolution method, *i.e.*, SRGAN (Ledig et al, 2016). Here, SRGAN is also fine-tuned on face images. (h) The image super-resolved by a state-of-the-art face hallucination method, *i.e.*, CBN (Zhu et al, 2016a). (i) The low-frequency component of (a). (j) The high-frequency component of (a). (k) The upsampled face by our previous method (Yu and Porikli, 2017b), which only uses the image (i) as input. (l) The result of our MTDN.

Such a broad spectrum of pose and expression variations also makes learning a comprehensive appearance model even harder. For instance, Principal Component Analysis (PCA) based schemes become critically ineffective to learn a reliable face model while aiming to capture different in- and out-of-plane rotations, scale changes, translational shifts, and facial expressions. As a result, these methods lead to unavoidable artifacts when LR faces are misaligned or depict different poses and facial expressions from the base appearance model. Moreover, once appearance models are learned, input LR faces at different resolutions need to be downscaled to fit the input size of the learned models. By doing so, some high-frequency information of LR faces will be lost and different LR faces tend to be indistinguishable at a lower resolution. Thus, the downscaling operation may result in suboptimal super-resolution performance.

Rather than learning holistic appearance models, many methods upsample *facial components* by transferring references from an HR training dataset and then blending them into an HR version (Tappen and Liu, 2012; Yang et al, 2013, 2017). Although these methods do not need LR face images to be aligned in advance or to resize input images to a fixed resolution, they expect the resolution of input faces to be sufficient enough for detecting the facial landmarks and parts. When the resolution is very low, they fail to localize the components accurately, thus producing non-realistic faces. In other words, the facial component based methods are unsuitable to upsample very low-resolution faces.

By better exploring the information available in the natural structure of face images, appearance similarities between individuals and emerging large-scale face datasets (Huang et al, 2007; Liu et al, 2015), it becomes possible to derive competent models to reconstruct authentic 4×∼8× magnified HR face images. Deep neural networks, in particular convolutional neural networks (CNN), are inherently suitable for learning from large-scale datasets. Very recently, CNN based generic patch super-resolution methods have been proposed (Dong et al, 2016; Kim et al, 2016a; Ledig et al, 2016) without focusing on any image class. A straightforward retraining (fine-tuning) of the networks, *i.e.*, VDSR (Kim et al, 2016a) and SRGAN (Ledig et al, 2016) with face images cannot produce realistic and visually pleasant results, as shown in Fig. 1(f) and Fig. 1(g), because these networks cannot address misalignments of LR inputs inherently. Misalignments of LR faces lead to the degradation of the super-resolution performance.

Recently, deep neural network based face hallucination methods have been proposed, and achieve state-of-the-art performance (Yu and Porikli, 2016, 2017a,b, 2018; Zhu et al, 2016a; Huang et al, 2017). However, those networks are only designed to super-resolve fixed-sized LR face images. When the input images are larger than the desired input size of the networks, images are required to be downsampled to fit the input size of the networks. After downsampling, some high-frequency components are lost. Thus, those deep learning based methods cannot fully exploit all the information of input images and output suboptimal results.

In this paper, we present a new multiscale transformative discriminative neural network (MTDN) to overcome the above issues. Our proposed network is able to super-resolve a range of small and unaligned face images (*i.e.*, from 16×16 to 32×32 pixels) to HR images of 128×128 pixels. In particular, when the resolution of input images is 16×16 pixels, we upsample LR faces by a remarkable upscaling factor 8×, where we reconstruct 64 pixels for each single pixel of an input LR image. Unlike previous works (Yu and Porikli, 2016, 2017a,b), when the resolutions of input images are larger than the input size of the networks, *i.e.*, 16×16 pixels, our network can preserve all the information of input face images. Specifically, our MTDN develops two branches to receive an downscaled LR input image as well as its residuals. In this fashion, our MTDN is able to exploit the residuals from the downscaled images for super-resolution. As seen in Fig. 1(j), the high-frequency residual component can be regarded as an external attention mechanism and lets the network focus on learning the missing high-frequency parts in the upsampled HR face images. In order to retain the global structure of faces while being able to reconstruct instance specific details, we use whole face images to train our networks.

Our network consists of two components: an upsampling network that comprises deconvolutional and spatial transformation network (Jaderberg et al, 2015) layers, and a discriminative network. The upsampling network is designed to progressively improve the resolution of the latent feature maps at each deconvolutional layer. We do not assume the LR face is aligned in advance. Instead, we compensate for any misalignment and changes through the spatial transformation network layers that are embedded into the upsampling network. In order to avoid the loss of information caused by downsampling LR face images, we separate LR images into two branches, *i.e.*, a low-frequency branch and a high-frequency branch. For instance, we downsample an LR image of 24×24 pixels to 16×16 pixels to obtain the low-frequency image as well as upsample its residual image (*i.e.*, an image is subtracted from the original LR image by the resized low-frequency image) to 32×32 pixels to achieve the high-frequency image. Then, we extract features from these two branches and then combine the feature maps for further super-resolution without losing information of inputs. One can use the pixel-wise intensity similarity between the estimated and the ground-truth HR face images as the objective function in the training stage. However, when the upscaling factor becomes larger, employing only the pixel-wise intensity similarity causes over-smoothed outputs. In order to force the upsampled faces to share facial features similar to their ground-truth counterparts, we employ the perceptual loss (Johnson et al, 2016). Since face hallucination is an under-determined problem, there would be one-to-many mappings between image intensities and features. Thus, the upsampled HR faces may not be sharp and realistic-looking enough. To make the upsampled HR faces realistic, we incorporate class similarity information that is provided by a discriminative network. We back-propagate the discriminative errors to the upsampling network. Our end-to-end solution allows fusing the pixel-wise, feature-wise and class-wise information in a manner robust to spatial transformations and obtaining a super-resolved output with much richer details.

Overall, our main contributions have four aspects:

- We present a novel end-to-end multiscale transformative discriminative network (MTDN) to super-resolve very low-resolution face images to HR face images of 128×128 pixels, where the upscaling factor ranges from 4× to 8×.
- We propose a unified framework which super-resolves LR faces at different resolutions, *i.e.*, from 16×16 to 32×32 pixels, and outputs aligned upscaled HR faces by a single deep neural network.
- In order to accept different sizes of LR input face images, we firstly divide an input image into a low-frequency component and a high-frequency residual one, and then design a two-branch network to receive these two components for upsampling. In this manner, we do not need to discard the residuals of the downsample LR faces so as to fit the input size of deep neural networks, thus avoiding losing information of inputs.
- For tiny input images where landmark based methods inherently fail, our method is able to align and hallucinate an unaligned LR face image without requiring precise alignment in advance, which makes our method practical.

This paper is an extension of our previous conference papers (Yu and Porikli, 2016, 2017a,b). In this paper, we propose a new unified framework to super-resolve LR faces at different resolutions. Since our previous methods need to downsample LR faces at different resolutions to a fixed resolution, this downsampling operation lose some high-frequency details of the LR inputs, *i.e.*, residual images. Thus, they may lead to suboptimal super-resolution results, as shown in our experimental part. Different from our previous works, the proposed network can preserve all the information of LR faces by our newly proposed multiscale network, thus achieving better super-resolution performance. In addition, we also conduct more comprehensive qualitative and quantitative experiments and discussions on each component of our proposed network.

2 Related Work

Super-resolution can be classified into two categories: generic super-resolution methods and class-specific super-resolution methods. When upsampling LR images, generic methods employ priors that ubiquitously exist in natural images without considering any image class information. Class-specific methods aim to exploit statistical information of objects in a certain class and they usually attain better results than generic methods, *e.g.*, the task of super-resolving LR face images.

Generic single image super-resolution methods generally have three types: interpolation based methods, image statistics based methods and learning-based methods. Interpolation based methods such as linear and non-linear upsampling are simple and computationally efficient, but they may produce overly smooth edges and fail to generate HR details as the upscaling factor increases. Image statistics based methods employ natural image priors to enhance the details of upsampled HR images, such as image gradients are sparse and follow heavy-tailed distributions (Tappen et al, 2003), but these methods are also limited to smaller magnification factors (Lin and Shum, 2006).

Learning-based methods demonstrate their potentials to exceed this limitation of the maximum upscaling factor by learning a mapping from a large number of LR/HR pairs (Lin et al, 2008). Glasner et al (2009); Freedman and Fattal (2010); Singh et al (2014) and Huang et al (2015) exploit self-similarity of patches in an input image to generate HR patches. Freeman et al (2002) and Hong Chang et al (2004) construct LR and HR patch pairs from a training dataset, and then infer high-frequency details by searching the corresponding HR patch of the nearest neighbor of an input LR patch. Yang et al (2010) employ sparse representation to construct the corresponding LR and HR dictionaries and then reconstruct HR output images by the sparse coding coefficients inferred from LR images. Gu et al (2015) apply convolutional sparse coding instead of patch-based sparse coding to reconstruct HR images.

Deep learning based super-resolution methods have been also proposed. Dong et al (2016) incorporate convolutional neural networks to learn a mapping function between LR and HR patches from a large-scale dataset. Motivated by this idea, the follow-up works (Kim et al, 2016a; Ledig et al, 2016; Kim et al, 2016b; Shi et al, 2016; Lai et al, 2017; Tai et al, 2017) try to explore deeper network architectures to improve super-resolution performance. Since many different HR patches may correspond to one LR patch, output images may suffer from artifacts at the intensity edges. In order to reduce the ambiguity between the LR and HR patches, Bruna et al (2016) explore the statistical information learned from a deep convolutional network to reduce ambiguity between LR and HR patches. Johnson et al (2016) propose a perceptual loss to constrain the feature similarity by a pre-trained deep neural network. Ledig et al (2016) employ the framework of generative adversarial networks (GAN) (Goodfellow et al, 2014) to enhance image details by combining an image intensity loss and an adversarial loss. Since those generic super-resolution methods do not take class-specific information into account, they still suffer over-smoothed results when input sizes are tiny and magnification factors are large.

Class-specific super-resolution methods further exploit the statistical information in the image categories, thus leading to better performance. When the class is faces, they are also called face hallucination methods (Baker and Kanade, 2000; Liu et al, 2001; Baker and Kanade, 2002).

The seminal work (Baker and Kanade, 2000, 2002) builds the relationship between facial HR and LR patches using Bayesian formulation such that high-frequency details can be transferred from the dataset for face hallucination. It can generate face images with richer details. However, artifacts also appear due to the possible inconsistency of the transferred HR patches. Wang and Tang (2005) apply PCA to LR face images, and then hallucinate HR face images by an Eigen-transformation of LR images. Although their method is able to magnify LR images by a large scaling factor, the output HR images suffer from ghosting artifacts when the HR images in the exemplar dataset are not precisely aligned. Liu et al (2007) enforce linear constraints for HR face images using a subspace learned from the training set via PCA, and a patch-based Markov Random Field is proposed to reconstruct the high-frequency details in the HR face images. To mitigate artifacts caused by misalignments, a bilateral filtering is used as a post-processing step. Kolouri and Rohde (2015) employ optimal transport in combination with subspace learning to morph an HR image from the LR input. Their method still requires that face images in the dataset are precisely aligned and the test LR images have the same poses and facial expressions as the exemplar HR face images. Instead of imposing global constraints, Ma et al (2010) super-resolve local HR patches by a weighted average of exemplar HR patches and the weights are learned from the corresponding LR patches. Rather than hallucinating HR patches in terms of image intensities, Li et al (2014) resort to sparse representation on the local regions of faces. However, blocky artifacts may appear as magnification factors become large.

To handle various poses and expressions, Tappen and Liu (2012) integrate SIFT flow (Liu et al, 2011) to align facial components in LR images. Their method performs competently when the training face images are highly similar to the test face image in terms of identity, pose, and expression. Yang et al (2013) and

Yang et al (2017) first localize facial components, and then upsample each component by matching gradients with respect to the similar HR facial components in the exemplar dataset. However, these methods rely on accurate facial landmark points that are usually unavailable when the image size is very small. More comprehensive literature review of early face hallucination works can be referred to the work of Wang et al (2014).

Deep learning based face hallucination methods are proposed to fully exploit the face structure and priors from emerging large-scale face datasets (Liu et al, 2015; Huang et al, 2007; Yang et al, 2016). Zhou and Fan (2015) propose a convolutional neural network (CNN) to extract facial features and recover facial details from the extracted features. Yu and Porikli (2018) combine deconvolutional and convolutional layers to upsample LR face images, but they resort a post-processing step (Yu et al, 2014) to improve the visual quality of the super-resolved faces. Later, Yu and Porikli (2016) explore a discriminative generative network to super-resolve aligned LR face images in an end-to-end manner while Huang et al (2017) estimate wavelet coefficients for a face upsampled by a generative adversarial network and then reconstruct the HR image from the estimated coefficients. Xu et al (2017) employ a multi-class adversarial loss in the framework of generative adversarial networks to super-resolve LR blurry face and text images. Dahl et al (2017) exploit an autoregressive generative model (Van Den Oord et al, 2016) to hallucinate prealigned LR face images. In order to mitigate the ambiguity of the mappings between LR and HR faces, Yu et al (2018b, 2019a) embed high-level semantic information, *i.e.*, face attributes, into the procedure of face hallucination. To relax the requirement of face alignment, Bulat and Tzimiropoulos (2018) present a constraint that the landmarks of the upsampled faces should be close to the landmarks detected in their ground-truth images. Since ground-truth landmarks are not provided in the training stage and erroneous localization of landmarks may lead to distorted upsampled face images, their results are only restricted to 64×64 pixels and facial details are not sharp enough. Zhu et al (2016a) develop a cascade bi-network to super-resolve unaligned LR faces, where facial components are localized first and then upsampled. Chen et al (2018) present a two-stage network, where low-frequency components of LR face are first super-resolved and then face priors (*i.e.*, facial component locations) are also employed to enrich facial details. Their methods may produce ghosting artifacts when the facial component localization is erroneous. To reduce the difficulty of estimating facial landmarks from unaligned LR images, Yu et al (2018a) predict facial landmarks from intermediate aligned feature maps by upsampling and aligning LR input images. Considering LR face images may be affected by different degeneration factors in real world cases, Bulat et al (2018) present a network to learn not only the mappings between LR and HR face images but also the real-world degeneration process. Towards the same goal, our previous works (Yu and Porikli, 2017a,b; Yu et al, 2018b, 2019b) embed multiple spatial transformer networks (Jaderberg et al, 2015) into the upsampling networks. However, their networks are trained on a fixed input resolution, and thus LR faces at different resolutions have to be resized (*i.e.*, downsampling) to meet the input resolution of the networks. Therefore, these methods may lose information of input images and introduce extra ambiguity due to the downscaling operation.

3 Proposed Method: MTDN

3.1 Background

Our face hallucination method is motivated by the generative adversarial networks (Goodfellow et al, 2014) since they can generate an face image from random noise represented by a fairly low-dimensional vector. Specifically, the generative model $\mathcal{G}$ takes a noise vector z from a distribution $P_{noise}(z)$ as an input and then outputs an image $\hat{x}$. The discriminative model $\mathcal{D}$ takes an image stochastically chosen from either the generated image $\hat{x}$ or the real image x drawn from the training dataset with a distribution $P_{data}(x)$ as an input. $\mathcal{D}$ is trained to output a scalar probability, which is large for real images and small for generated images from $\mathcal{G}$. The generative model $\mathcal{G}$ is learned to maximize the probability of $\mathcal{D}$ making a mistake. Thus a minmax objective is used to train these two models simultaneously,

$$\min_{\mathcal{G}}\max_{\mathcal{D}}\mathbb{E}_{x\sim P_{data}(x)}\log\mathcal{D}(x)+\mathbb{E}_{z\sim P_{noise}(z)}\log(1-\mathcal{D}(\mathcal{G}(z))).$$

This equation encourages $\mathcal{G}$ to fit $P_{data}(x)$ so as to fool $\mathcal{D}$ with its generated samples $\hat{x}$. However, we cannot directly employ the above equation for the face hallucination task since GAN takes a fixed size noise vector as input to learn the distribution on the training dataset. In contrast, the input for our face super-resolution task is an LR face image, and its resolution is not fixed either. LR faces also undergo rotations, translations and scale changes.

In this paper, we propose a transformative discriminative neural network (MTDN) which achieves the image alignment and super-resolution simultaneously. Furthermore, our MTDN accepts LR input images in various sizes without losing image information. The entire pipeline is shown in Fig. 2.

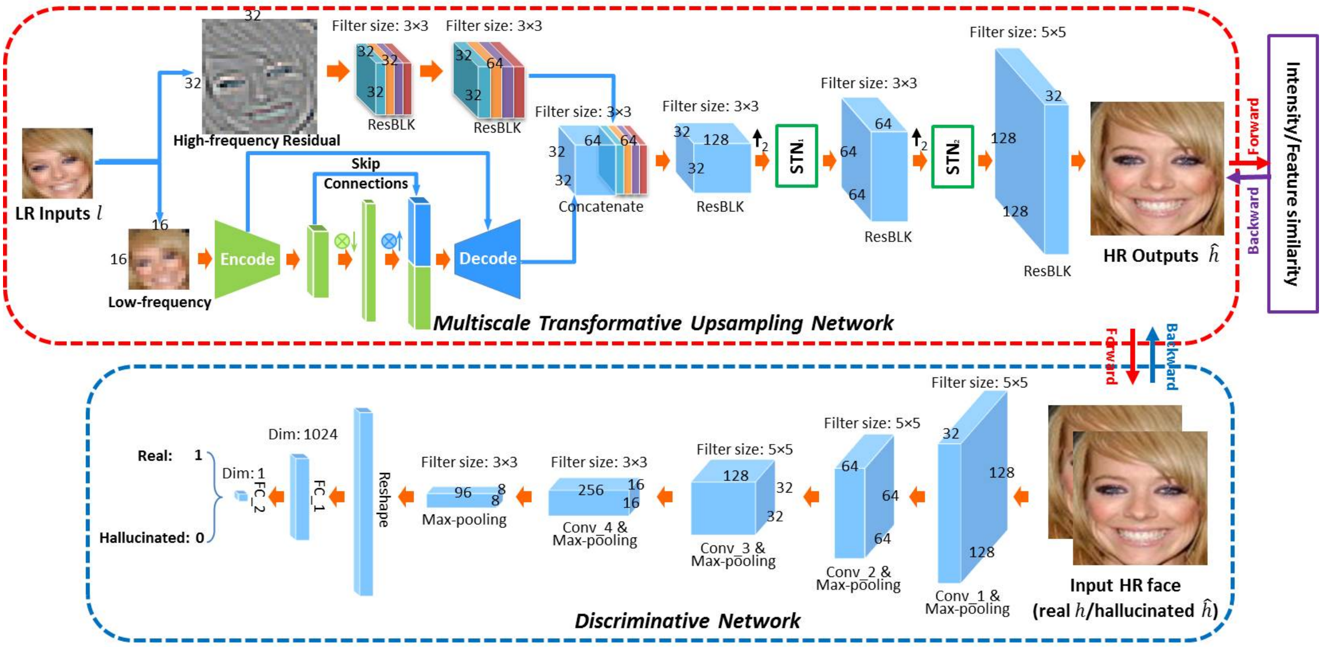


Fig. 2 Our MTDN consists of two parts: an upsampling network (in the red frame) and a discriminative network (in the blue frame).

3.2 Network Architecture

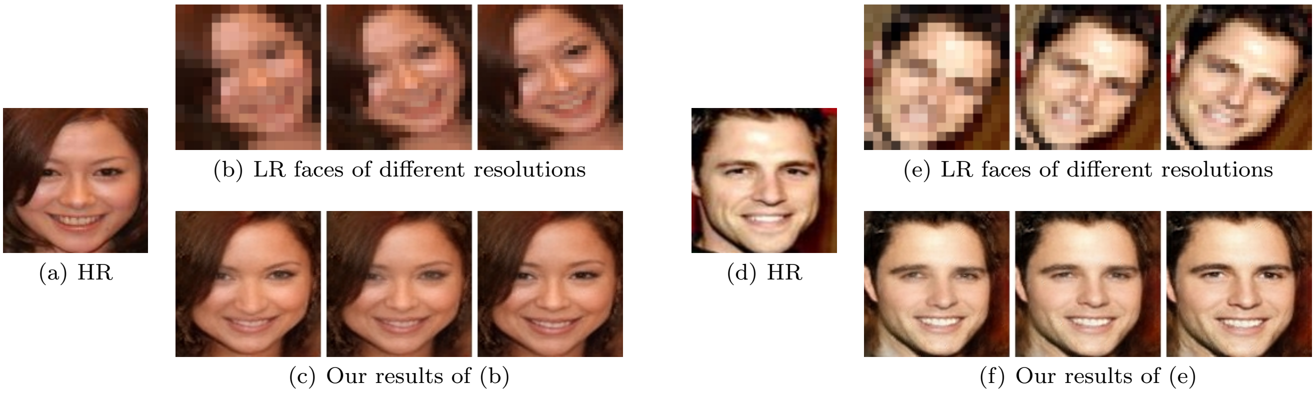
Our MTDN consists of two parts: a multiscale transformative upsampling network that combines autoencoder, spatial transformation network layers, upsampling layers and residual block layers, and a discriminative network that is composed of convolutional layers, max-pooling layers, and fully-connected layers. The multiscale transformative upsampling network is designed for receiving and super-resolving LR images at different resolutions while the discriminative network is developed to force the super-resolved faces to be realistic.

3.2.1 Multiscale Transformative Upsampling Network

Reception for LR Images in a Multiscale Manner: State-of-the-art CNN based super-resolution networks (Yu and Porikli, 2016, 2017a,b; Zhu et al, 2016a; Bulat and Tzimiropoulos, 2018; Chen et al, 2018) only accept LR inputs in a fixed resolution, *i.e.*, 16×16 pixels. When the resolutions of LR images are larger than the desired resolution, those methods need to downsample input images. However, downsampling input images may result in the loss of high-frequency details of LR inputs as well as more ambiguous mappings between LR and HR face images in super-resolution. In addition, we assume that the resolutions of LR images are smaller than 32×32 pixels. Otherwise, LR images can provide enough resolution for human observation and computer analysis. Hence, we only focus on LR images whose resolutions are smaller than 32×32 pixels in this paper.

Inspired by the Laplacian pyramid, we decompose an image into two components: a low-frequency part and a high-frequency part. We downsample an input image to 16×16 pixels as our low-frequency part, as illustrated in Fig. 1(i). The high-frequency part is obtained by subtracting the input image by the interpolated low-frequency components. Then, we upsample the high-frequency component to 32×32 pixels, as visible in Fig. 1(j). In this way, our transformative upsampling network can receive LR face images at different resolutions while preserving high-frequency residual details of the inputs for super-resolution.

In order to combine the information of the high-frequency and low-frequency branches together, we extract feature maps from the images of those two branches and then concatenate the feature maps for further super-resolution. Specifically, we firstly employ an autoencoder with skip connections to extract features from the low-frequency component and then upsample the feature maps by a deconvolutional layer. After the deconvolutional layer, the resolution of the low-frequency branch has been increased as the same as the resolution of the high-frequency branch. Rather than directly combining the high-frequency residual component with the feature maps of the low-frequency component, we employ two cascaded residual blocks to extract features from the high-frequency component as well. Then, we



(a) HR

(b) LR faces of different resolutions

(c) Our results of (b)

(d) HR

(e) LR faces of different resolutions

(f) Our results of (e)

Fig. 3 Illustrations of our results with respect to the different resolutions of LR input images. (a)(d) Ground-truth HR face images. (b)(e) unaligned LR face images. From left to right, the resolutions of the images are 16×16, 24×24 and 32×32. (c) Our results of (b). From left to right, the corresponding PSNRs are 22.79 dB, 23.59 dB and 24.63 dB. (f) Our results of (e). From left to right, the corresponding PSNRs are 17.80 dB, 19.96 dB and 21.94 dB.

concatenate the feature maps extracted from the high-frequency residual component with the upsampled feature maps of the low-frequency component and then employ a residual block to fuse the concatenated feature maps.

As shown in Fig. 3, our network is able to super-resolve LR face images at different resolutions. Note that, we do not need to fine-tune our network on images of different sizes. As expected, the PSNRs of our upsampled results become higher as the resolutions of LR faces increase. This indicates our network exploits all the information in LR input images for super-resolution.

Upsampling Layers: After obtaining the concatenated features maps of input images, we further super-resolve the feature maps by the deconvolutional layers and residual blocks. The deconvolutional layer, also known back-convolutional layer, can be made of a cascade of an upsampling layer and a convolutional layer, or a convolutional layer with a fractional stride (Zeiler et al, 2010; Zeiler and Fergus, 2014). Therefore, the resolution of the output of the deconvolutional layers is larger than the resolution of its input. To reduce potential blocky artifacts caused by deconvolutional layers (Yu and Porikli, 2018) as well as increase the capacity of the network, we cascade a residual block after each deconvolutional layer as our upsampling layer.

Spatial Transformation Layers: The spatial transformation network (STN) is proposed by Jaderberg et al (2015). It can estimate the motion parameters of images, and warp images to the canonical view. In our architecture, the spatial transformation network layers are represented as the green boxes in Fig. 2. These layers contain three modules: a localization module, a grid generator module, and a sampler. The localization module consists of a number of hidden layers and outputs the transformation parameters of an input relative to the canonical view. The grid generator module creates a sampling grid according to the estimated parameters. Finally, the sampler module maps the input onto the generated grid by bilinear interpolation.

Since we focus on in-plane rotations, translations, and scale changes without requiring a 3D face model, we employ the similarity transformation for face alignment. Although STNs can warp images, it is not straightforward to use them directly to align very LR face images. As shown in Fig. 1(c), directly applying an STN to LR images causes distortion artifacts due to the difficulty of spatial transformation estimation on very LR faces. There are several factors needed to be considered: (i) After the alignment of LR images, facial patterns are blurred due to the resampling of the aligned faces by bilinear interpolation. (ii) Since the resolution is very low and a wide range of poses exists, estimating spatial transformations on such small face images may lead to alignment errors. (iii) Due to the blur and alignment errors, the upsampling network may fail to generate realistic HR faces. (iv) If STNs are employed to the two branches separately, the estimated transformation parameters of these two branches may be different. This will result in misalignments between the low-frequency component and the high-frequency residual components. As a result of the misalignments, distortion artifacts or ghosting artifacts may appear in the final results. Therefore, we employ STNs to align the concatenated feature maps. In this way, we can align the low-frequency and high-frequency parts simultaneously.

Instead of using a single STN to align LR face images, we employ multiple STN layers to line up the feature maps. Using multiple layers significantly reduces the load on each spatial transformation network and further reduces the errors of misalignments. In addition, resampling feature maps by multiple STN layers prevents from damaging or blurring input LR facial patterns. Since STN layers and the upsampling layers are interwoven together (rather than being two individual networks), the upsampling network can learn to eliminate the undesired effects of misalignment in the training stage.

3.2.2 Discriminative Network

In generic super-resolution (Kim et al, 2016a,b), only the ℓ_2 regression loss, also known as Euclidean distance loss, is employed to constrain the similarity between the upsampled HR images and their original HR ground-truth versions. However, as reported in our previous work (Yu and Porikli, 2016), deconvolutional layers supervised by a ℓ_2 loss tend to produce over-smoothed results. As seen in Fig. 4(f), the hallucinated faces are not sharp enough because the common parts learned by the upsampling network are averaged from similar components shared by different individuals. Thus, there is a quality gap between the real face images and the hallucinated faces. To bridge this gap, we inject class information. We integrate a discriminative network to distinguish whether the generated image is classified as an upright real face image or not. A similar idea is employed in the generative adversarial networks (Goodfellow et al, 2014; Denton et al, 2015; Radford et al, 2015), which are designed to generate a new face. The architecture of the discriminative network is shown in the blue frame of Fig. 2. It consists of convolutional, max-pooling, fully-connected and non-linear transformation layers. We employ a binary cross-entropy as the loss function to distinguish whether the input HR faces are sampled from super-resolved or real images. We back-propagate the discriminative error to revise the coefficients of the multiscale transformative upsampling network (for simplicity, we also refer to it as the upsampling network), which enforces the facial parts learned by the deconvolutional layers to be as sharp and authentic as real face images. Furthermore, the use of class information also facilitates the performance of the STN layers for face alignment since only upright faces are classified as valid faces. Therefore, our discriminative network also determines whether the faces are upright or not. As shown in Fig. 4(h), with the help of the discriminative information, the hallucinated face embodies more authentic, much sharper and better aligned details.

3.3 Training Details of MTDN

We construct LR and HR face image pairs $\{l_i, h_i\}$ as our training dataset, where h_i represents the aligned HR face images (only eyes are aligned), and l_i is the synthesized LR face images downsampled from h_i. Notice that, different from our previous works, the resolutions of input LR images l_i are different. As mentioned in Sec. 3.2.1, the input LR faces l_i are further decomposed into two components: the low-frequency component l_i^L of size 16×16 pixels and the high-frequency residual component l_i^H of size 32×32 pixels.

In training our MTDN, we not only employ the conventional pixel-wise intensity similarity, known as pixel-wise ℓ_2 loss, but also the feature-wise similarity, known as perceptual loss (Johnson et al, 2016). The perceptual loss is able to enforce the upsampled facial characteristics to resemble their ground-truth counterparts. Even though pixel-wise and feature-wise similarity are applied in training our network, learning a mapping between LR and HR face images is still an ill-posed problem. Our network will tend to output blurry results to lower the training losses. Thus, in the testing stage, the upsampling network generates blurry faces. Similar to our previous works (Yu and Porikli, 2016, 2017a), the adversarial loss is also employed to attain visually appealing HR face images.

3.3.1 Pixel-wise Intensity Similarity Loss

We enforce the generated HR face $\hat{h}_i$ to be similar to its corresponding ground-truth h_i in terms of image intensities. Thus we employ a pixel-wise ℓ_2 regression loss $\mathcal{L}_{pix}$ to impose the appearance similarity constraint, expressed as:

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

where t and $\mathcal{U}$ are the parameters and the output of the upsampling network, $p(\hat{h}, h)$ represents the joint distribution of the frontalized HR faces and their corresponding frontal HR ground-truths, $p(l, h)$ indicates the joint distribution of the LR and HR face images in the training dataset, and the LR input l_i is decomposed into l_i^L and l_i^H before fed into the upsampling network. Here, we do not distinguish the parameters of the upsampling layers and the STN layers because all the parameters are learned simultaneously. We employ t to represent all the parameters in our multiscale transformative upsampling network.

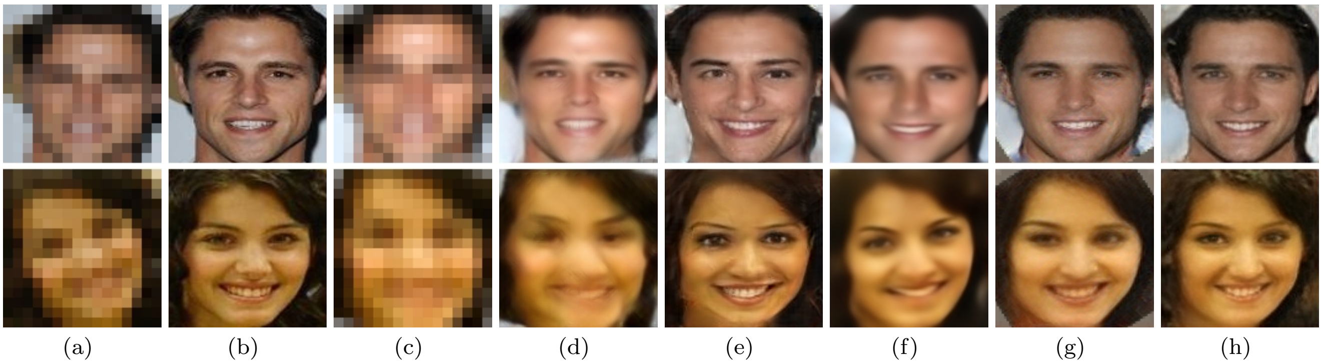


Fig. 4 Illustrations of different losses for super-resolution. (a) The input 16 × 16 LR images. (b) The original 128 × 128 HR images. (c) The aligned LR images. (d) The upsampled faces by SRGAN (Ledig et al, 2016). Here, SRGAN is applied to the aligned LR faces. Since SRGAN is trained on generic images patches, we re-train SRGAN on whole face images. (e) The face images super-resolved by our previous method (Yu and Porikli, 2017b). (f) The super-resolved faces by $\mathcal{L}_{pix}$. (g) The super-resolved faces by $\mathcal{L}_{pix} + \mathcal{L}_{feat}$. (h) The super-resolved faces by $\mathcal{L}_{pix} + \mathcal{L}_{feat} + \mathcal{L}_{\mathcal{U}}$. Here, we omit the trade-off weights for simplicity.

3.3.2 Feature-wise Similarity Loss

As illustrated in Fig. 4(f), the pixel-wise ℓ_2 loss leads to over-smoothed super-resolved results. Therefore, we employ a feature-wise similarity loss to force the super-resolved HR faces to share the same facial features as their ground-truth counterparts. The feature-wise loss $\mathcal{L}_{feat}$ measures Euclidean distance between the feature maps of super-resolved and ground-truth HR faces which are extracted by a deep neural network, written as:

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

where $\Phi(\cdot)$ denotes feature maps extracted by the ReLU32 layer in VGG-19 (Simonyan and Zisserman, 2014), which gives good empirical performance in our experiments.

3.3.3 Class-wise Discriminative Loss

In order to achieve visually appealing results, we infuse class-specific discriminative information into our upsampling network by exploiting a discriminative network, similar to our previous works (Yu and Porikli, 2016, 2017a,b). Since our goal is to output realistic HR faces, the upsampled face images should be able to fool the discriminative network. In other words, the upsampling network makes the discriminative network fail to distinguish generated faces from real ones. To do so, we enforce the super-resolved HR frontal faces to lie on the manifold of real HR face images. The discriminative network is used to classify real and super-resolved faces, and thus its objective function is written as:

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

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

Meanwhile, our upsampling network aims to fool the discriminative network. It needs to generate realistic HR face images and make the discriminative network classify the super-resolved faces as real faces. Therefore, the objective function of our upsampling network is written as:

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

By minimizing the loss $\mathcal{L}_{\mathcal{U}}$, we update the parameters t and thus the discriminative network will be prone to categorize the upsampled faces as real ones. These two discriminative losses in Eqn. 3 and Eqn. 4 are used to update our upsampling and discriminative networks respectively in an alternating fashion.

All the layers in our MTDN are differentiable and thus RMSprop (Hinton, 2012) is employed to update the parameters t and d. We update the parameters d by minimizing the loss $\mathcal{L}_{\mathcal{D}}$ as follows:

$$\begin{aligned} \varDelta^{i+1} &= \gamma\varDelta^{i} + (1-\gamma)(\frac{\partial \mathcal{L}_{\mathcal{D}}}{\partial d})^2, \\ d^{i+1} &= d^{i} - r\frac{\partial \mathcal{L}_{\mathcal{D}}}{\partial d}\frac{1}{\sqrt{\varDelta^{i+1}+\epsilon}}, \end{aligned} \tag{5}$$

where r and γ represent the learning rate and the decay rate respectively, i indicates the index of the iterations, $\varDelta$ is an auxiliary variable, and ϵ is set to 10^{-8} to avoid division by zero.

Multiple losses, *i.e.*, $\mathcal{L}_{pix}$, $\mathcal{L}_{feat}$, and $\mathcal{L}_{\mathcal{U}}$, are used for learning the parameters of our upsampling network and the object function is expressed as:

$$\mathcal{L}_{\mathcal{T}} = \mathcal{L}_{pix} + \eta\mathcal{L}_{feat} + \lambda\mathcal{L}_{\mathcal{U}}, \tag{6}$$

where η and λ are the trade-off weights. We employ lower weights on the feature-wise and discriminative losses because we aim at super-resolving HR faces rather than generating random faces. Thus, λ and η are both set to 0.01. Then, the parameters of our upsampling network t are updated by the gradient descent as follows:

$$\begin{aligned} \varDelta^{i+1} &= \gamma\varDelta^{i} + (1-\gamma)(\frac{\partial \mathcal{L}_{\mathcal{T}}}{\partial t})^2, \\ t^{i+1} &= t^{i} - r\frac{\partial \mathcal{L}_{\mathcal{T}}}{\partial t}\frac{1}{\sqrt{\varDelta^{i+1}+\epsilon}}. \end{aligned} \tag{7}$$

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

$$\lambda^{j} = \max\{\lambda \cdot 0.995^{j}, \lambda/2\}, \tag{8}$$

where j is the index of the epochs. Equation 8 not only increases the impact of the appearance similarity term but also preserves the class-specific discriminative information in the training phase. The training procedure of our MTDN is illustrated in Algorithm 1.

3.4 Hallucinating a Very LR Face Image

The discriminative network is only used for training of the upsampling network. In the testing phase, we first decompose an LR image into a low-frequency component image and its high-frequency residual image and then feed them into the upsampling network to obtain a super-resolved HR face. Because the ground-truth HR face images are upright in the training stage of the entire network, the output of the upsampling network will be an upright face image. As a result, our method does not require alignment of the very low-resolution images in advance. Our network provides an end-to-end mapping from an unaligned LR face image to an upright HR version, which mitigates potential artifacts caused by misalignments and facilitates achieving high-quality super-resolved HR face images.

Algorithm 1 Minibatch stochastic gradient descent training of MTDN

```
Input: minibatch size $N$, LR and HR face image pairs $\{l_i, h_i\}$, maximum number of iterations $K$.
1: while Iter < K do
2:     Choose one minibatch of LR and HR image pairs $\{l_i, h_i\}, i = 1, \dots, N$.
3:     Decompose LR images into the low-frequency and high-frequency components $\{l_i^L, l_i^H\}$.
4:     Generate one minibatch of HR face images $\hat{h}_i$ from $\{l_i^L, l_i^H\}, i = 1, \dots, N$, where $\hat{h}_i = \mathcal{U}_t(l_i^L, l_i^H)$.
5:     Update the parameters of the discriminative network $\mathcal{D}_d$ by using Eqn. 3 and Eqn. 5.
6:     Update the parameters of the multiscale transformative upsampling network $\mathcal{U}_t$ by using Eqn. 6 and Eqn. 7.
7:     Update the trade-off weight $\lambda$ by using Eqn. 8.
8: end while
Output: MTDN.
```

Moreover, our two-branch architecture network is able to upsample LR input images to HR images of size 128×128 pixels by a upscaling factor ranging from 4× to 8×. For example, our MTDN super-resolves an LR image by 8× when zeros are fed into the high-frequency branch. In other words, our low-frequency branch focuses on super-resolving images by a upscaling factor of 8× while our high-frequency residual branch extracts complementary information from input images for super-resolution. In this fashion, we can upsample LR faces by different magnification factors in a uniform framework.

3.5 Flexiblity of Halluciatining Faces with Different Interocular Distances

Since previous methods, such as FSRNet Chen et al (2018) and CBN Zhu et al (2016b), require the interocular distances of input LR faces to be similar, they may fail to super-resolve faces when the interocular distances of input faces are different from their training ones, as seen in Fig. 5. On the contrary, we employ STN layers to align LR faces in different sizes to the pre-defined upright position, and thus the interocular distances of LR faces are also rectified to similar distances by the STN layers. Therefore, we do not suffer from artifacts when input LR faces undergo different interocular distances. As indicated in Fig. 5, Fig. 6 and Fig. 7, our method is able to super-resolve LR faces from 16×16 to 32×32 pixels.

3.6 Implementation Details

In Fig. 2, 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). In particular, STN_1 layer is cascaded by: MP2, Conv+ReLU (with the filter size: 128×20×5×5), MP2, Conv+ReLU (with the filter size: 20×20×5×5), 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: 64×128×5×5), MP2, Conv+ReLU (with the filter size: 128×20×5×5), MP2, Conv+ReLU (with the filter size: 20×20×3×3), FC+ReLU (from 180 to 20 dimensions) and FC (from 20 to 4 dimensions). We do not use zero-padding in the convolution operations.

In order to merge the low-frequency images with the information extracted from the high-frequency branch, we employ an autoencoder with skip connections. The encoder is composed of convolutional layers with a stride of 2 and zero-paddings. The decoder consists of deconvolutional layers with a stride of 2 and zero-paddings as well. The feature maps from the encoder and decoder are concatenated by skip connections. The residual block is composed of a convolutional layer with a kernel size 3×3, batch normalization, ReLU, a convolutional layer with a kernel size 1×1 and a high-pass connection.

In the following experimental part, some algorithms require the alignments of LR inputs (Ma et al, 2010; Ledig et al, 2016; Kim et al, 2016b). Thus, we use STN_0 to align the LR inputs images (*i.e.*, 16×16 pixels) for those methods. 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. Since some algorithms can super-resolve unaligned LR faces, we use an STN network to align the upsampled HR face images, marked as STN_{HR}. STN_{HR} is constructed by: Conv+ReLU (with the filter size: 3×16×5×5), MP2, Conv+ReLU (with the filter size: 16×32×3×3), MP2, Conv+ReLU (with the filter size: 32×64×3×3), MP2, Conv+ReLU (with the filter size: 64×128×3×3), MP2, Conv+ReLU (with the filter size: 128×20×5×5), FC+ReLU (from 80 to 20 dimensions) and FC (from 20 to 4 dimensions). All the training details, codes and pre-trained models will be released.

4 Experiments

In this section, we compare our method with the state-of-the-art methods (Ma et al, 2010; Kim et al, 2016a; Ledig et al, 2016; Zhu et al, 2016b; Yu and Porikli, 2016, 2017a,b; Chen et al, 2018; Yu et al, 2018b,a) qualitatively and quantitatively. Kim et al (2016a) employ very deep CNN to upsample images. Ledig et al (2016) use the generative adversarial framework to enhance super-resolved details. Ma et al (2010) exploit position-patches in the dataset to reconstruct HR images. Zhu et al (2016b) develop a deep CNN to localize facial components and then super-resolve them in a cascaded manner. Yu and Porikli (2016) not only employ a pixel-wise similarity loss to train a deconvolutional network but also present a discriminative loss to enforce the upsampled HR faces to be realistic. Yu and Porikli (2017a) propose a single-scale face hallucination method, which also employs STN layers and deconvolutional layers for super-resolution. Yu and Porikli (2017b) develop a decoder-encoder-decoder network architecture to suppress noise in LR face images while upsampling. Chen et al (2018) exploit facial priors (*i.e.*, facial landmarks) to enrich upsampled face details. Yu et al (2018b) embed facial attributes into the procedure of face hallucination to reduce the inherent ambiguity of super-resolution with a large upscaling factor. Yu et al (2018a) first estimate facial component structure from the intermediate aligned feature maps and then hallucinated facial details based on the estimated locations of facial components.

4.1 Dataset

Our network is trained on the Celebrity Face Attributes (CelebA) dataset (Liu et al, 2015). There are more than 200K face images in this dataset, and the images cover different pose variations and facial expressions. In training our network, we disregard these variations without grouping the face images into different pose and facial expression subcategories.

When generating the LR and HR face pairs, we crop the aligned HR face images from the CelebA dataset, and then resize them to 128×128 pixels as HR images. We manually transform the HR images including 2D translations, rotations and scale changes while constraining the faces in the image region, and then downsample the HR images to generate their corresponding LR images, where the resolutions of LR images are also randomly set between 16 to 32 pixels. We use 70%, 10% and 20% of LR and HR image pairs for training, validation and testing, respectively.

4.2 Experimental Setup

Since our method is able to super-resolve an image with a substantial upscaling factor of 8×, for the methods that do not provide 8× (Kim et al, 2016a; Ledig et al,

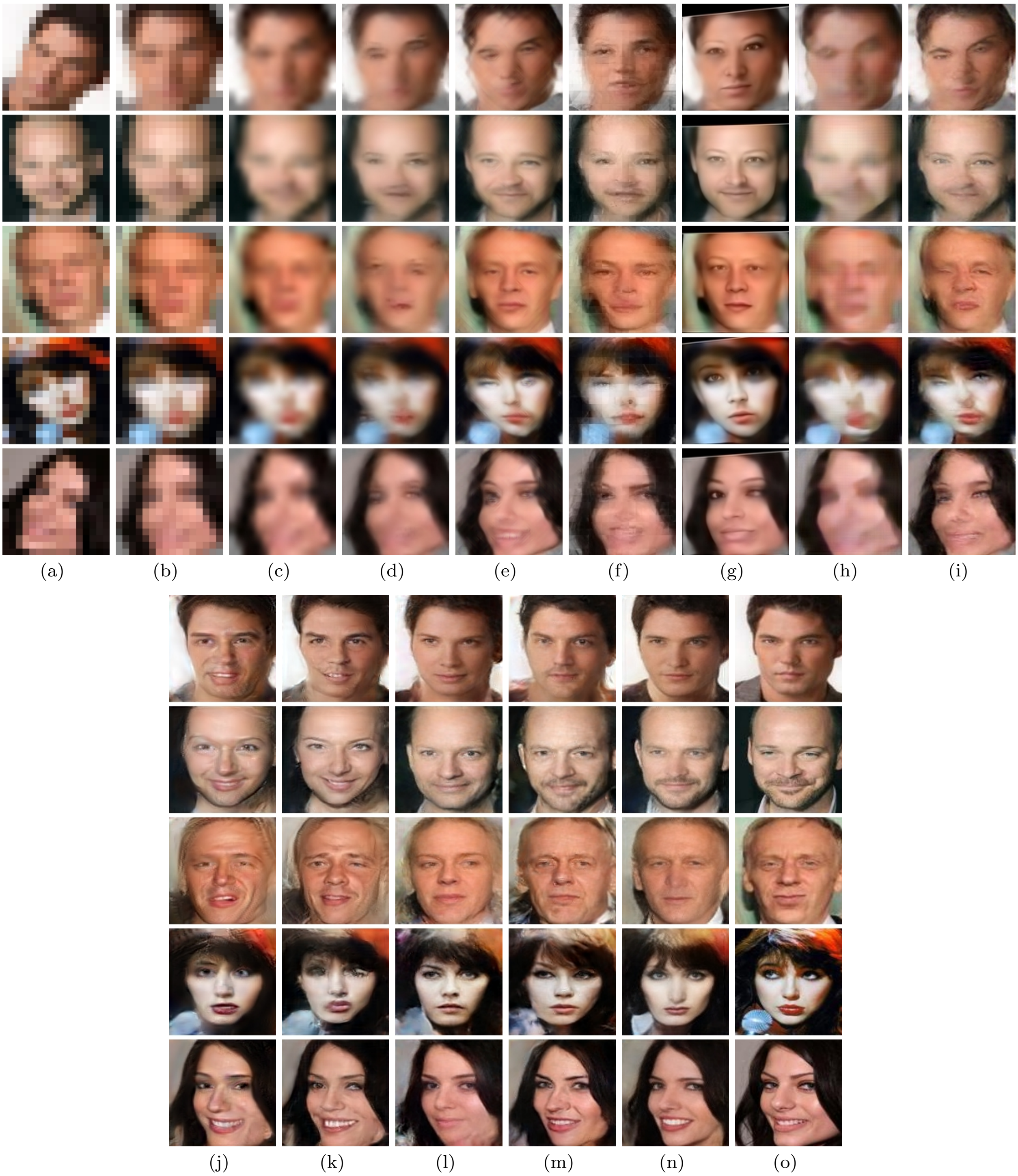


Fig. 5 Comparisons with the state-of-the-art methods on the input images of size 16×16 pixels. The results are obtained in the scenario of first aligning LR faces and then super-resolving them. (a) Unaligned LR inputs. (b) Aligned LR faces. (c) Bicubic interpolation. (d) Kim *et al.*'s method (Kim et al, 2016a) (VDSR). (e) Ledig *et al.*'s method (Ledig et al, 2016) (SRGAN). (f) Ma *et al.*'s method (Ma et al, 2010). (g) Zhu *et al.*'s method (Zhu et al, 2016a) (CBN). (h) Chen *et al.*'s method (Chen et al, 2018) (FSRNet). (i) Yu and Porikli's method (Yu and Porikli, 2016) (URDGN). (j) Yu and Porikli's method (Yu and Porikli, 2017a) (TDN). (k) Yu and Porikli's method (Yu and Porikli, 2017b) (TDAE). (l) Yu *et al.*'s method (Yu et al, 2018b). (m) Yu *et al.*'s method (Yu et al, 2018a). (n) Our method (MTDN). (o) Original HR images.

2016), we retrain their network on face images with a

magnification factor 8×. Furthermore, the resolutions

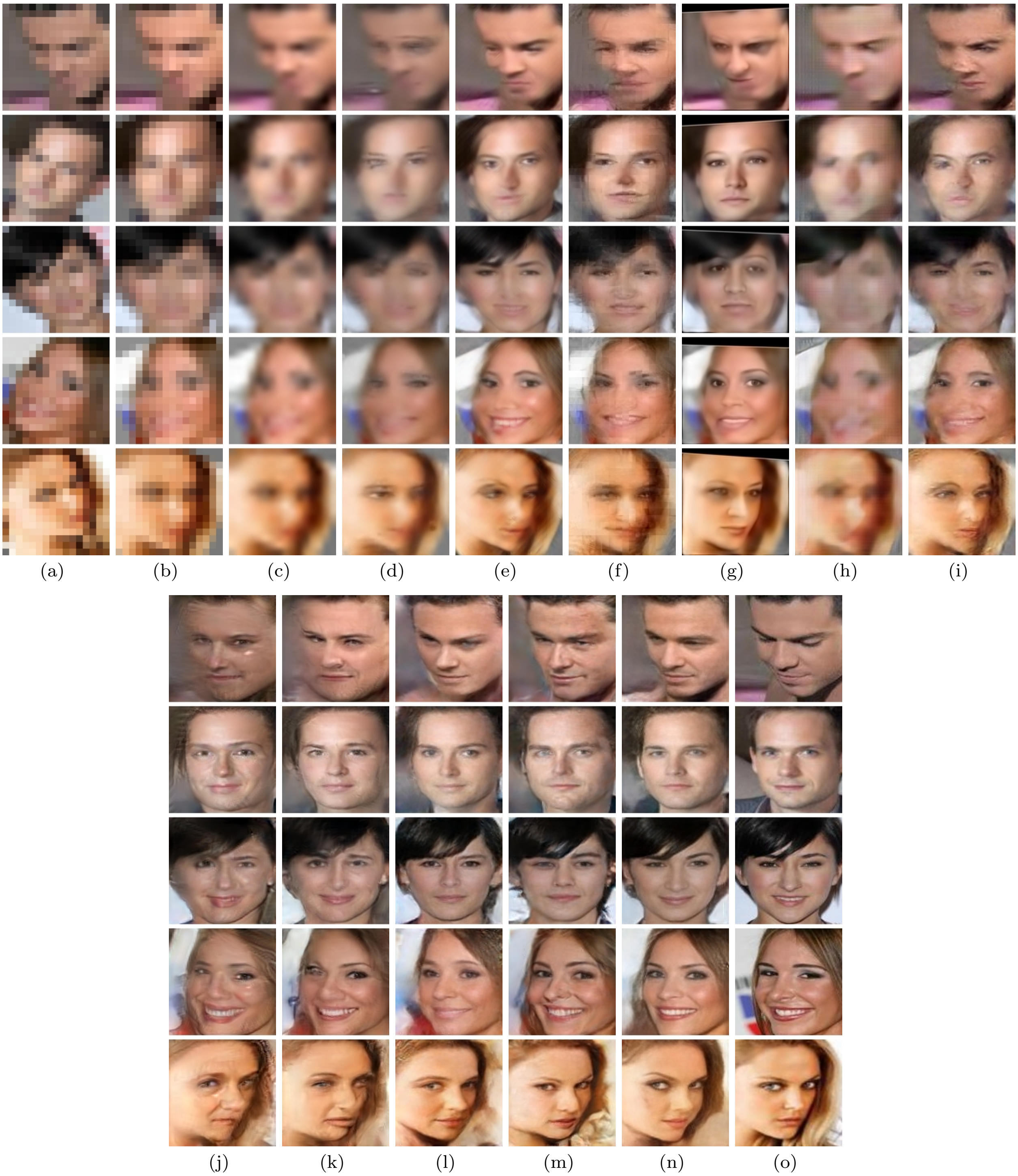


Fig. 6 Comparisons with the state-of-the-art methods on the input images of size 24×24 pixels. The results are obtained in the scenario of first aligning LR faces and then super-resolving them. (a) Unaligned LR inputs. (b) Aligned LR faces. (c) Bicubic interpolation. (d) Kim *et al.*'s method (Kim et al, 2016a) (VDSR). (e) Ledig *et al.*'s method (Ledig et al, 2016) (SRGAN). (f) Ma *et al.*'s method (Ma et al, 2010). (g) Zhu *et al.*'s method (Zhu et al, 2016a) (CBN). (h) Chen *et al.*'s method (Chen et al, 2018) (FSRNet). (i) Yu and Porikli's method (Yu and Porikli, 2016) (URDGN). (j) Yu and Porikli's method (Yu and Porikli, 2017a) (TDN). (k) Yu and Porikli's method (Yu and Porikli, 2017b) (TDAE). (l) Yu *et al.*'s method (Yu et al, 2018b). (m) Yu *et al.*'s method (Yu et al, 2018a). (n) Our method (MTDN). (o) Original HR images.

of LR inputs are various, *i.e.*, 16×16∼32×32 pixels, but STNs can only accept an image in a fixed resolution

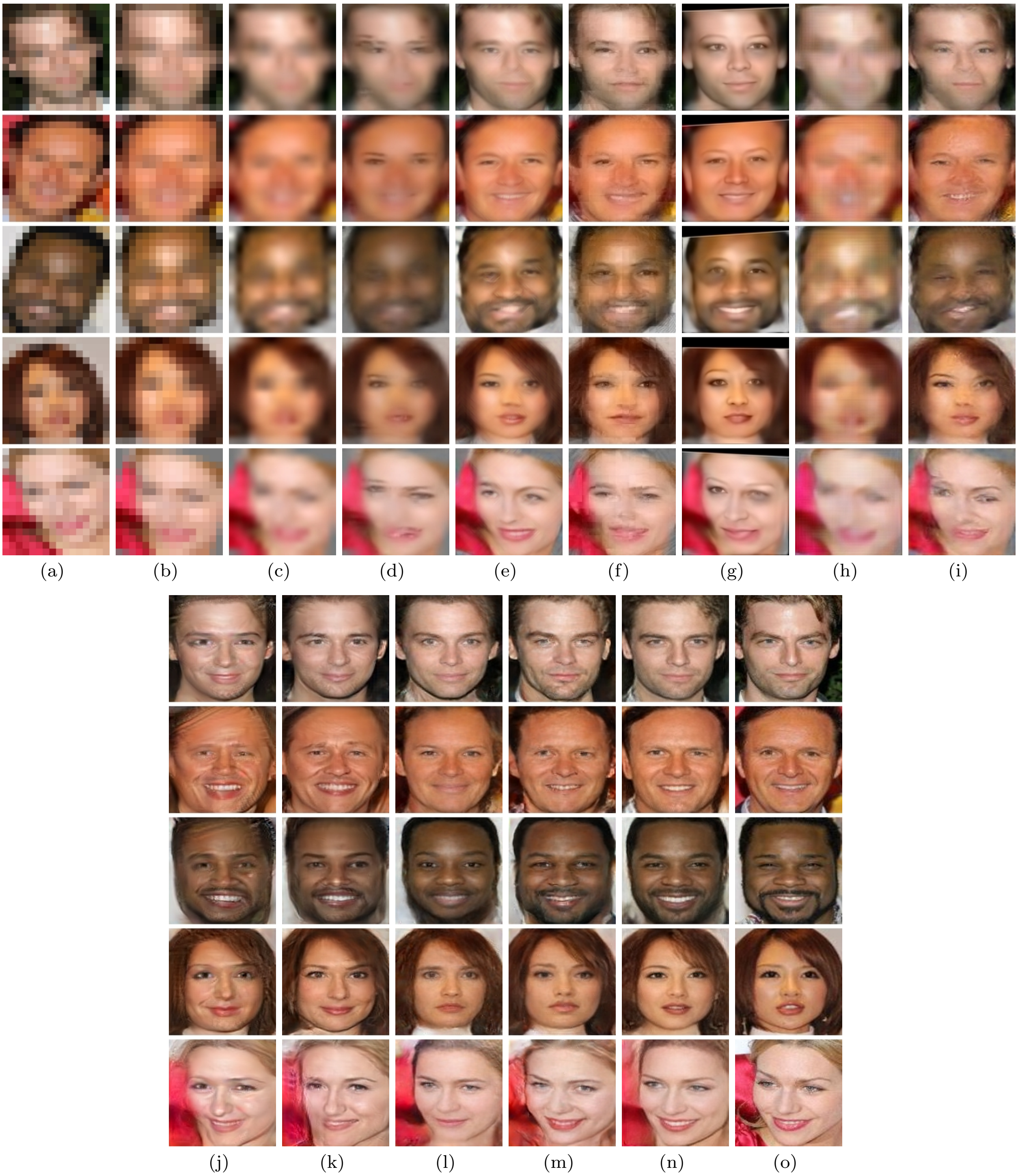


Fig. 7 Comparisons with the state-of-the-art methods on the input images of size 32×32 pixels. The results are obtained in the scenario of first aligning LR faces and then super-resolving them. (a) Unaligned LR inputs. (b) Aligned LR faces. (c) Bicubic interpolation. (d) Kim *et al.*'s method (Kim et al, 2016a) (VDSR). (e) Ledig *et al.*'s method (Ledig et al, 2016) (SRGAN). (f) Ma *et al.*'s method (Ma et al, 2010). (g) Zhu *et al.*'s method (Zhu et al, 2016a) (CBN). (h) Chen *et al.*'s method (Chen et al, 2018) (FSRNet). (i) Yu and Porikli's method (Yu and Porikli, 2016) (URDGN). (j) Yu and Porikli's method (Yu and Porikli, 2017a) (TDN). (k) Yu and Porikli's method (Yu and Porikli, 2017b) (TDAE). (l) Yu *et al.*'s method (Yu et al, 2018b). (m) Yu *et al.*'s method (Yu et al, 2018a). (n) Our method (MTDN). (o) Original HR images.

due to the network architecture of its localization module. Considering approaches (Zhu et al, 2016b; Yu and

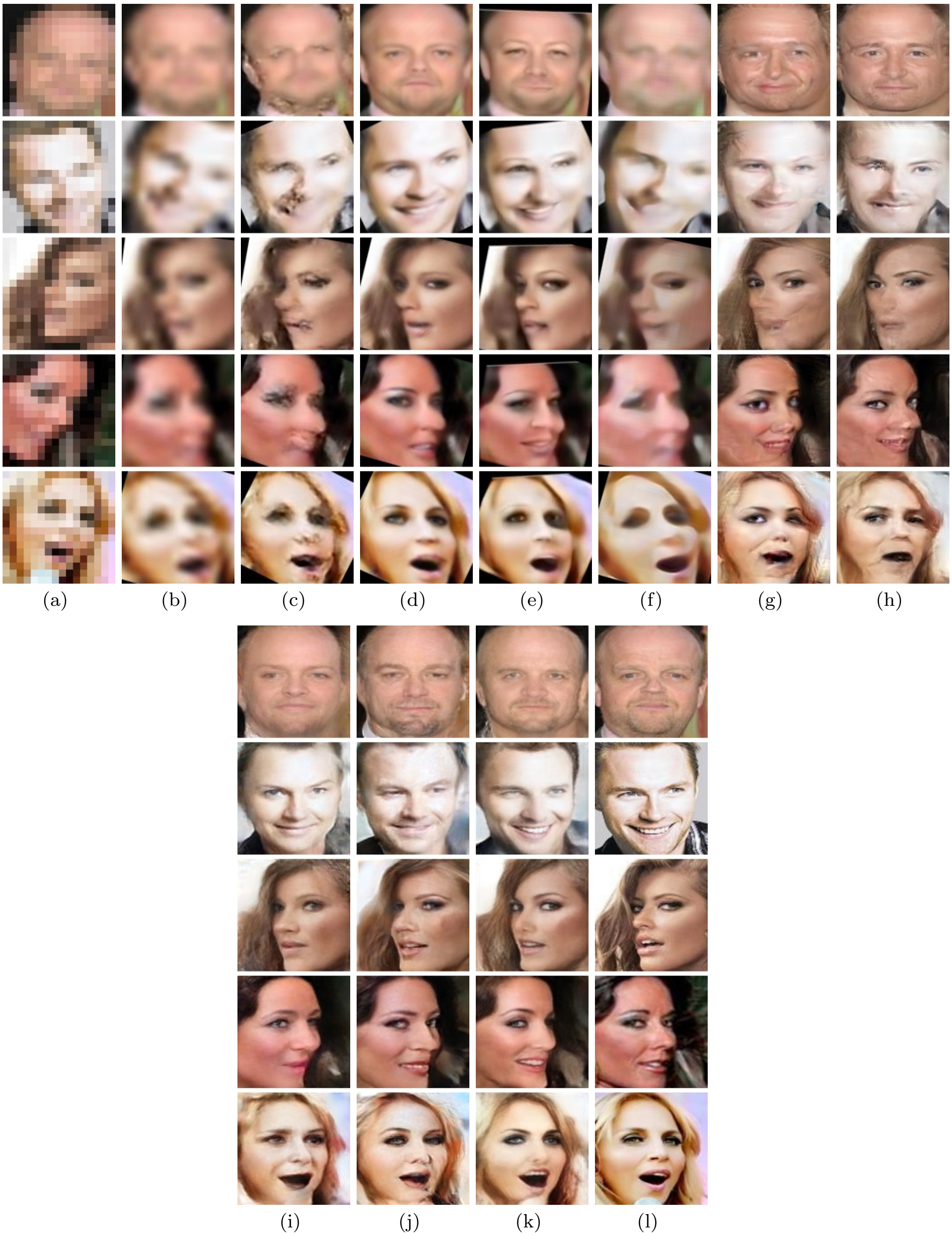


Fig. 8 Comparisons with the state-of-the-art methods on the input images of size 16×16 pixels. The results are obtained in the scenario of first upsampling LR faces and then aligning the super-resolved faces by STN$_{HR}$. (a) Unaligned LR inputs. (b) Bicubic interpolation. (c) Kim *et al.*'s method (Kim et al, 2016a) (VDSR). (d) Ledig *et al.*'s method (Ledig et al, 2016) (SRGAN). (e) Zhu *et al.*'s method (Zhu et al, 2016a) (CBN). (f) Chen *et al.*'s method (Chen et al, 2018) (FSRNet). (g) Yu and Porikli's method (Yu and Porikli, 2017a) (TDN). (h) Yu and Porikli's method (Yu and Porikli, 2017b) (TDAE). (i) Yu *et al.*'s method (Yu et al, 2018b). (j) Yu *et al.*'s method (Yu et al, 2018a). (k) Our method (MTDN). (l) Original HR images.

Porikli, 2017b; Chen et al, 2018; Yu et al, 2018b,a) only accept the input resolution of 16×16 pixels, the input images are resized to 16×16 pixels to meet the requirements. Since the methods of Ma et al (2010) and Yu

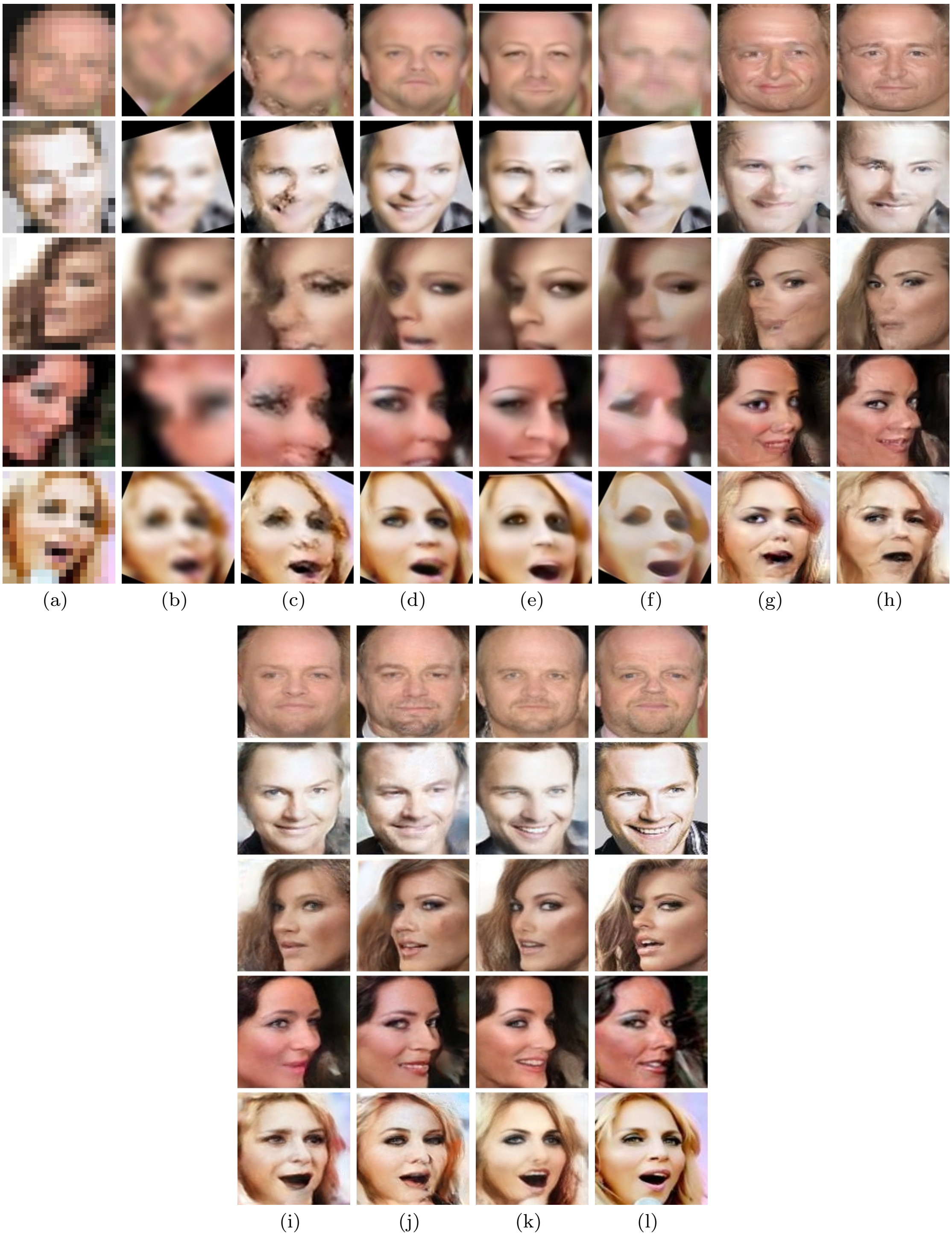


Fig. 9 Comparisons with the state-of-the-art methods on the input images of size 16×16 pixels. The results are obtained in the scenario of first upsampling LR faces and then aligning the super-resolved faces by Bulat *et al.*'s method (Bulat and Tzimiropoulos, 2017). (a) Unaligned LR inputs. (b) Bicubic interpolation. (c) Kim *et al.*'s method (Kim et al, 2016a) (VDSR). (d) Ledig *et al.*'s method (Ledig et al, 2016) (SRGAN). (e) Zhu *et al.*'s method (Zhu et al, 2016a) (CBN). (f) Chen *et al.*'s method (Chen et al, 2018) (FSRNet). (g) Yu and Porikli's method (Yu and Porikli, 2017a) (TDN). (h) Yu and Porikli's method (Yu and Porikli, 2017b) (TDAE). (i) Yu *et al.*'s method (Yu et al, 2018b). (j) Yu *et al.*'s method (Yu et al, 2018a). (k) Our method (MTDN). (l) Original HR images.

and Porikli (2016) require LR faces to be aligned before super-resolution while some networks (Kim et al, 2016a; Ledig et al, 2016; Chen et al, 2018) can super-resolve unaligned LR images, we propose two scenarios to compare with the state-of-the-art methods. In the first scenario, we first employ an STN network (*i.e.*, STN_0) to align LR input images and then super-resolve aligned LR faces by the state-of-the-art methods. In the second scenario, we first upsample unaligned LR input faces by the state-of-the-art approaches and then use an alignment method to transform the upsampled HR faces to the upright position. Since our method exploits STN layers to align feature maps, we employ an STN network (*i.e.*, STN_{HR}) to align the upsampled HR face images. Moreover, we also employ a state-of-the-art facial landmark based alignment method (Bulat and Tzimiropoulos, 2017) to align upsampled HR faces. Due to the misalignments of LR input faces, the aligned and upsampled HR faces cannot contain all the regions in the ground-truth images, and the unmatching regions may lead to inaccurate quantitative evaluation. Thus, we employ a mask to remove unmatching regions for quantitative comparisons.

4.3 Qualitative Comparisons with the State-of-the-Art

As shown in Fig. 5(c) and Fig. 8(b), traditional upsampling methods, *i.e.*, bicubic interpolation, cannot hallucinate authentic facial details. Since the resolution of inputs is very small, little information is contained in the input images. Simply interpolating input LR images cannot recover extra high-frequency details. As seen in Fig. 5(c), Fig. 6(c) and Fig. 7(c), the images upsampled by bicubic interpolation have some skew effects rather than laying in the upright view. This also indicates that aligning input images by STN_0 suffers from misalignments because it is difficult to estimate transformation parameters accurately from images in such a small size. As shown in Fig. 8(b) and Fig. 9(b), STN_{HR} and the landmark based face alignment algorithm (Bulat and Tzimiropoulos, 2017) fail to align upsampled HR faces to the upright position since the hallucinated faces are too blurry. On the contrary, we apply multiple STNs on the upsampled feature maps, which improves the alignment of the LR inputs. Therefore, our method outputs well-aligned faces. Moreover, with the help of our discriminative network, our method can achieve much sharper results.

Kim et al (2016a) propose a very deep convolutional neural network based general purpose super-resolution method, dubbed VDSR. Since VDSR is trained on natural image patches, it may be not suitable to super-resolve face images. Furthermore, VDSR does not provide a magnification factor of 8×. Thus, we fine-tune VDSR on both aligned and unaligned LR/HR face image pairs with an upscaling factor of 8×. However, VDSR is only composed of convolutional layers, and cannot generate aligned HR face images. Hence, STN_0 is employed to align LR faces before super-resolution in the first scenario, while STN_{HR} and Bulat *et al.*'s method (Bulat and Tzimiropoulos, 2017) are used to align upsampled HR ones in the second scenario. As shown in Fig. 5(d), Fig. 8(c) and Fig. 9(c), VDSR fails to produce realistic facial details in both scenarios. This indicates that only using a pixel-wise loss as supervision leads to overly smoothed super-resolved results.

Ledig et al (2016) develop a generic super-resolution method, known as SRGAN. SRGAN employs the framework of generative adversarial networks (Goodfellow et al, 2014; Radford et al, 2015) to enhance the visual quality. It is trained by using not only a pixel-wise ℓ_2 loss but also an adversarial loss. Similar to VDSR, original SRGAN is also trained on image patches, and thus it is hard to capture the global structure of face images. Therefore, we also retrain SRGAN on both aligned and unaligned LR/HR face image pairs. As seen in Fig. 5(e), Fig. 6(e) and Fig. 7(e), SRGAN captures LR facial patterns and achieves sharper upsampled results compared to VDSR, but misalignments in LR faces cause severe distortions and artifacts in the final hallucinated faces. As visible in Fig. 8(d) and Fig. 9(d), although SRGAN is able to generate unaligned LR facial details, the aligned HR faces still undergo obvious skew effects. This implies that directly combining a face super-resolution method and a face alignment method cannot yield satisfying results.

Ma et al (2010) exploit position patches to hallucinate HR faces. Thus their method requires the LR inputs to be precisely aligned with the reference images in the training dataset. As seen in Fig. 5(f), Fig. 6(f) and Fig. 7(f), as the upscaling factor increases, the correspondences between LR and HR patches become more inconsistent. As a result, this method suffers from obvious blocky artifacts around the boundaries of different patches. In addition, when there are obvious alignment errors in the aligned LR faces or large poses exist, their method will output mixed and blurry facial components in their results. Since Ma et al (2010) require the input faces to be aligned in advance, we do not compare with their method in the second scenario.

Zhu et al (2016a) present a deep cascaded bi-branch network for face hallucination, named CBN, where one branch first localizes facial components, then aligns and upsamples LR facial components while the other branch is used to upsample global face profiles. However, due to the inaccurate localization of facial components in both scenarios, CBN may produce severe artifacts as seen in

Table 1 Quantitative comparisons on the entire test dataset

SR Methods	A+SR			SR+A$_{STN}$			SR+A$_{LM}$		
	PSNR	SSIM	FLLE	PSNR	SSIM	FLLE	PSNR	SSIM	FLLE
Bicubic	22.14	0.75	56.32	19.21	0.68	51.25	17.30	0.64	58.96
VDSR	23.28	0.76	8.61	21.64	0.73	10.34	20.81	0.70	12.88
SRGAN	22.24	0.76	3.66	22.95	0.79	3.44	22.30	0.75	5.42
CBN	20.98	0.72	5.48	20.94	0.72	4.73	19.57	0.67	6.89
FSRNet	20.53	0.72	28.48	19.44	0.70	16.68	19.30	0.69	19.81
Ma et al (2010)	22.06	0.74	4.48	–	–	–	–	–	–
URDGN	23.24	0.75	4.54	–	–	–	–	–	–
				PSNR	SSIM	FLLE			
TDN				23.32	0.75	4.03			
TDAE				23.53	0.76	3.94			
TDAE$_{32}$				23.96	0.78	3.00			
Yu et al (2018b)				23.89	0.79	2.56			
Yu et al (2018a)				24.28	0.79	2.35			
IBSR				24.26	0.78	3.53			
Ours				**25.02**	**0.80**	**2.32**			

Table 2 Quantitative comparisons on the frontal faces

SR Methods	PSNR	SSIM	FLLE
Bicubic	24.49	0.67	36.07
VDSR	24.68	0.69	12.88
SRGAN	26.90	0.76	1.96
Ma et al (2010)	26.07	0.79	3.42
CBN	22.62	0.65	6.89
FSRNet	22.54	0.78	8.59
URDGN	26.04	0.72	2.11
TDN	23.68	0.76	3.22
TDAE	24.04	0.78	3.12
Yu et al (2018b)	25.02	0.81	2.32
Yu et al (2018a)	24.71	0.80	2.30
Ours	25.56	0.81	2.19
Ours†	**28.19**	**0.83**	**1.48**

Fig. 5(g), Fig. 6(g), Fig. 7(g), Fig. 8(e) and Fig. 9(e). In contrast, our method estimates the 2D deformations of LR faces and aligns them by multiple STNs in the procedure of super-resolution, where misalignments from the previous STN layer can be eliminated by the latter STN layer. Therefore, our results do not suffer the ghosting artifacts as shown in Fig. 5(n), Fig. 6(n) and Fig. 7(n). Note that, the facial component localization branch in CBN requires the input resolution to be fixed, *i.e.*16×16 pixels. Thus, if the resolutions of input images are larger than 16×16 pixels, CBN needs to downsample input images first. In that case, CBN may lose high-frequency information of inputs and achieves suboptimal hallucination results.

Chen et al (2018) present a network to super-resolve HR faces in two stages by exploiting face priors, named FSRNet. FSRNet firstly super-resolves the low-frequency part of LR input faces by its first-stage network and then exploits the face structure of the upsampled faces as face priors to enrich facial details by its second-stage network. Since FSRNet does not align upsampled HR face images, we compare with FSRNet in the two scenarios. We use the pre-trained model released by Chen et al (2018) to super-resolve LR faces. Because aligning LR faces by STN$_0$ may introduce extra blurriness and skew artifacts, FSRNet fail to localize facial components from upsampled overly-smoothed HR faces, as seen in Fig. 5(h), Fig. 6(h) and Fig. 7(h). In addition, since unaligned LR faces undergo different 2D transformations, the interocular distances of the testing face images are different from the ones for training FSRNet. FSRNet may fail to localize facial landmarks. As shown in Fig. 8(f) and Fig. 9(f), FSRNet generates overly-smoothed faces due to the erroneous localization of facial components in LR faces. This implies that FSRNet might be overfitted to super-resolve face images with a certain interocular distance.

Yu and Porikli (2016) develop a discriminative generative network, known as URDGN, to upsample very low-resolution face images. Their method uses deconvolutional layers to upsample LR faces and a discriminative network to force the generated faces to be realistic. URDGN is trained on aligned LR/HR face pairs at a fixed scale. Thus, we resize input images to the required resolution for URDGN and compare with URDGN in the first scenario. As visible in Fig. 5(i), Fig. 6(i) and Fig. 7(i), URDGN suffers severe artifacts when LR facial patterns are distorted by the face alignment network STN_0. Moreover, URDGN is only composed of three deconvolutional layers and two convolutional layers for upsampling and thus this shallow network architecture may hinder the face hallucination performance. By increasing the network capacity as well as embedding STN layers, our MTDN not only obtains aligned HR face images but also achieves superior super-resolution performance to URDGN, as seen in Fig. 5, Fig. 6 and Fig. 7.

To super-resolve unaligned LR face images, Yu and Porikli (2017a) embed spatial transformer networks (Jaderberg et al, 2015) as intermediate layers into an upsampling network, as well as exploit a discriminative network to enforce the upsampling network to produce sharper results. Since STNs are used to align feature maps of LR input images, their transformative discriminative network (TDN) is able to generate upright HR face images. Thus, we do not need to align the super-resolved HR images or the input LR images. Similar to URDGN, TDN also employs a relative shallow network architecture. Thus, the super-resolution performance of TDN tends to saturate. As seen in Fig. 6(j), Fig. 7(j) and Fig. 8(g), ringing artifacts and distortions appear in the upsampled results of TDN. In contrast, due to the larger network capacity, our MTDN attains much sharper and clearer HR face images, as shown in Fig. 5(n), Fig. 6(n) and Fig. 8(k).

Yu and Porikli (2017b) design a transformative discriminative autoencoder, called TDAE, to upsample noisy and unaligned LR face images. However, TDAE only upsamples LR images in a fixed resolution, and it has to downsample LR images to a lower-resolution when the resolutions of input images are larger than the required resolution, *i.e.*, 16×16 pixels. Therefore, TDAE will lose details of input images and may generate inaccurate facial characteristics, such as gender reversal as visible in the second row of Fig. 5(k) and the third row of Fig. 6(k). Note that, we apply TDAE to the unaligned LR face images directly for super-resolution. Furthermore, benefiting from the feature-wise loss, our MTDN is able to hallucinate facial characteristics akin to the ground-truth HR faces. Furthermore, TDAE is trained mainly on near-frontal face images. It does not super-resolves LR faces in large poses well. In contrast, we enlarge the training dataset with more examples and more challenging poses to train our MTDN. Therefore, our network attains better super-resolution performance.

Yu et al (2018b) embed high-level semantic information into face hallucination by designing a conditional generative discriminative network. Their method can significantly reduce the ambiguity of super-resolution when the facial attribute information of an input face image is provided. Since all the other methods do not use any high-level information for face hallucination, we feed "neutral attributes" (*i.e.*, 0.5) to their network for super-resolving LR face images. Due to the employment of STN layers in their network, the network accepts LR images at a fixed resolution and aligns LR inputs while hallucinating them. As seen in the first and third rows of Fig. 5(l), the results of Yu et al (2018b) exhibit different facial attributes from their corresponding ground-truth ones when the input attributes are inaccurate. This also implies that their performance highly relies on the accuracy of the input facial attributes, which may not always be available in practice.

Yu et al (2018a) develop a facial component heatmap guided upsampling network. Unlike the methods (Chen et al, 2018; Bulat and Tzimiropoulos, 2018), this method aligns feature maps by STN layers and then estimates the facial components from the intermediate aligned feature maps instead of the coarsely upsampled HR face images. Similar to TDN and TDAE, their method does not need to apply face alignment to the LR input images. Due to the usage of STN layers in the upsampling network, Yu et al (2018b) also need to resize input images before super-resolution, thus losing some high-frequency details of input images. Although their method is able to super-resolve LR faces in very large poses with the help of the facial component localization module, it may fail to super-resolve HR faces authentically when it estimates facial landmarks incorrectly. The inaccurate facial component estimation degrades the final super-resolution performance, as shown in the first row of Fig. 5(m) and the fourth row of Fig. 8(j).

As shown in Fig. 5(n), Fig. 6(n), Fig. 7(n), Fig. 8(k) and Fig. 9(k), our method reconstructs authentic facial details and the reconstructed faces have different poses and facial expressions. Since our method applies multiple STNs on feature maps to align face images, we can achieve better alignment results without damaging input LR facial patterns. Furthermore, our method does not warp input images directly, so there are no blank regions in our results. Since our network is able to receive LR images at different resolutions without discarding residual images, our method can exploit information better than the other methods. Notice that, we only use a single network to super-resolve all the LR face images in various resolutions.

4.4 Quantitative Comparisons with the State-of-the-Art

We report the quantitative comparison results using the average Peak Single-to-Noise Ratio (PSNR), Structural SIMilarity scores (SSIM) as well as average Facial Landmark Localization Error (FLLE) on the entire test dataset in Tab. 1. FLLE measures the Euclidean distance between the estimated facial landmarks and the ground-truth ones. We employ a state-of-the-art face alignment method (Bulat and Tzimiropoulos, 2017) to detect 68 point facial landmarks. Note that, in the test dataset the resolutions of LR input face images ranges from 16×16 to 32×32. We use all the methods to upsample LR face images to the HR images of size 128×128 pixels and then compare the upsampled HR faces with their corresponding ground-truths. As mentioned in Sec. 4.2, the state-of-the-art methods need to downsample input images to 16×16 pixels for super-resolution.

In Tab. 1, we compare with the state-of-the-art methods in the two possible scenarios quantitatively. The first scenario, *i.e.*, LR face alignment followed by super-resolution, is marked as A+SR. In the second scenario, we first upsample LR input images by the state-of-the-art methods and then align the upsampled HR faces by a face alignment method. For the second scenario, we employ two alignment methods (*i.e.*, STN-based and landmark-based alignment methods) to align upsampled HR faces and report these two possible combinations. We employ STN_{HR} to align super-resolved HR faces and this combination is named as SR+A_{STN}. The other combination using a facial landmark-based aligning method (Bulat and Tzimiropoulos, 2017) is marked as SR+A_{LM}.

As indicated in Tab. 1, our MTDN attains the best PSNR and SSIM results and outperforms the second best with a large margin of 0.74 dB in PSNR. Although (Yu et al, 2018a) interweave STN layers and deconvolutional layers as well as exploit face structure to upsample unaligned face images, their method only accepts input images in a fixed resolution, *i.e.*16×16 pixels, and thus achieves the second best performance. This indicates that our previous works lose important high-frequency information of LR images in the downsampling operation. As indicated in Tab. 1, by using the multi-scale strategy and the two-branch architecture network, we can preserve all the information of the LR inputs in super-resolution and thus obtain superior performance to our previous work Yu and Porikli (2017b). Futhermore, Tab. 1 demonstrates that it is difficult to obtain well-aligned hallucinated faces by either aligning LR faces before super-resolution or aligning upsampled HR face images. In contrast, we can achieve lower face alignment errors by embedding multiple STN layers into our upsampling networks. This also implies that the necessity of using STN layers.

4.5 Comparisons with the State-of-the-Art on Aligned Faces

To further demonstrate the effectiveness of our methods, we compare with the state-of-the-art algorithms when the LR input faces are aligned in different resolutions in Fig. 10 and Fig. 11. Similar to the procedure of generating unaligned face dataset, we construct 10K aligned LR/HR face pairs in different resolutions (*i.e.*, from 16×16 to 32×32 pixels) for testing.

As illustrated in Fig. 10 and Fig. 11, our method achieves visually appealing results and authentic facial details which are akin to the ground-truth ones. Furthermore, we also demonstrate the quantitative results in comparisons to the state-of-the-art methods in Tab. 2. Since several works (Yu and Porikli, 2017a,b; Yu et al, 2018b,a) and our MTDN employ STN layers, STN layers will try to align input LR images even though the input faces are aligned. Thus, STN layers may deform (such as, rescale, translate and rotate) input images and lead to inferior quantitative results as seen in Tab. 2. Thus, we remove the STN layers from our MTDN to super-resolve aligned LR faces. As shown in Tab. 2, the results of our MTDN, marked as Ours†, achieves the best quantitative performance. Our method outperforms the second best method SRGAN by a large margin of 1.29 dB in PSNR.

5 Abalation Study

5.1 Impacts of Residual Branch

As indicated by the quantitative result of TDAE in Tab. 1, the downsampling operation leads to suboptimal super-resolution performance. Since our MTDN also employs extra residual blocks, the improvement of the performance may be caused by the increased capacity of the network. In order to evaluate the impacts of the high-frequency residual branch, similar to our previous methods (Yu and Porikli, 2016, 2017b), we only employ one branch, *i.e.*, the low-frequency branch, to upsample LR input face images. Note that, we do not need to re-train our MTDN network. As shown in Tab. 6, the performance of only using low-frequency branch is marked by noHF, and its performance degrades 1.30 dB in PSNR. It indicates that the high-frequency residual information extracted from the input images contains useful clues for super-resolution.

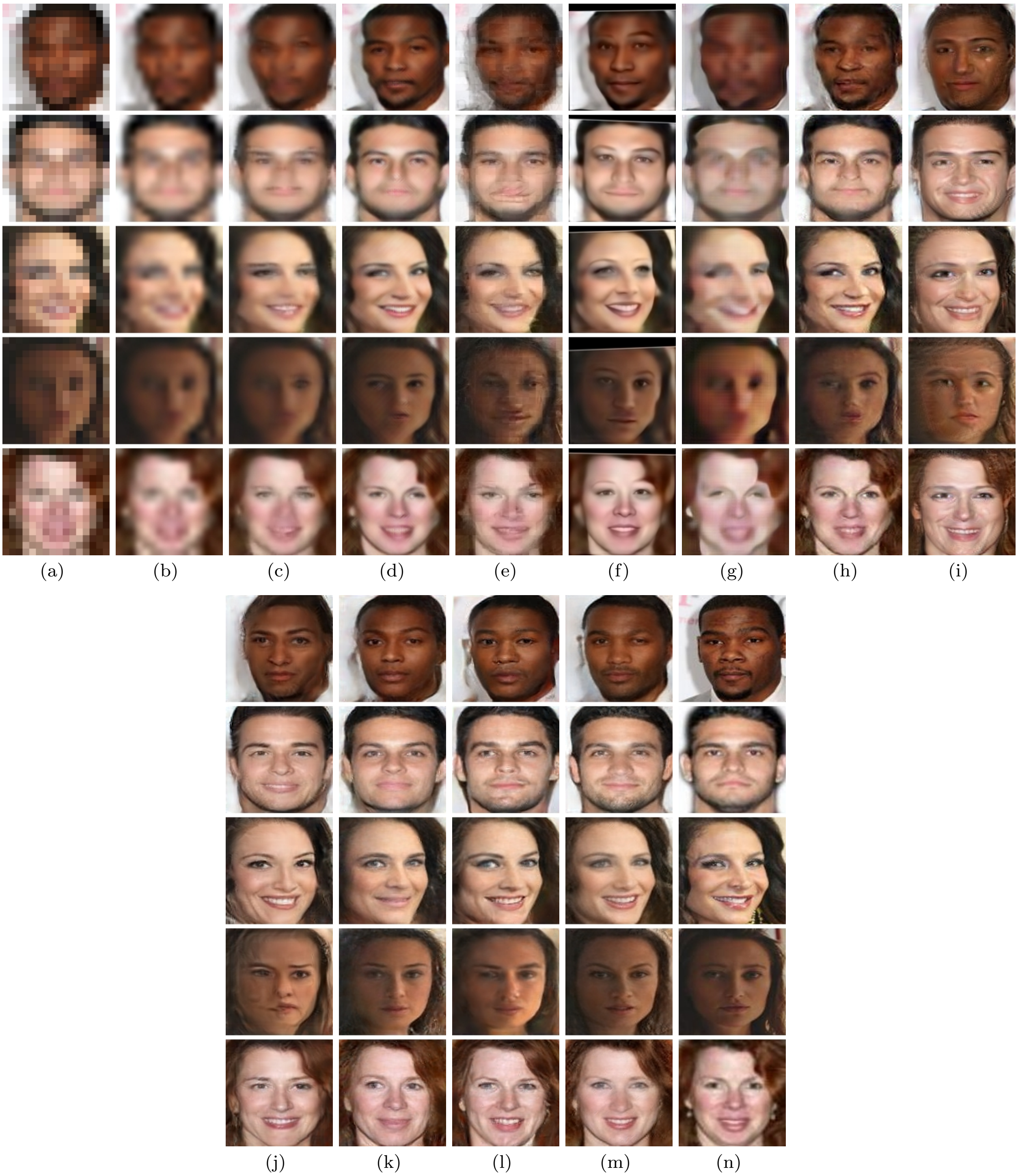


Fig. 10 Comparisons with the state-of-the-art methods on the *aligned* input images of size 16×16 pixels. (a) Unaligned LR inputs. (b) Bicubic interpolation. (c) Kim *et al.*'s method (Kim et al, 2016a) (VDSR). (d) Ledig *et al.*'s method (Ledig et al, 2016) (SRGAN). (e) Ma *et al.*'s method (Ma et al, 2010). (f) Zhu *et al.*'s method (Zhu et al, 2016a) (CBN). (g) Chen *et al.*'s method (Chen et al, 2018) (FSRNet). (h) Yu and Porikli's method (Yu and Porikli, 2016) (URDGN). (i) Yu and Porikli's method (Yu and Porikli, 2017a) (TDN). (j) Yu and Porikli's method (Yu and Porikli, 2017b) (TDAE). (k) Yu *et al.*'s method (Yu et al, 2018b). (l) Yu *et al.*'s method (Yu et al, 2018a). (m) Our method (MTDN). (n) Original HR images.

Thus, providing more high-frequency details improves face super-resolution performance.

5.2 Effects of Different Losses

As mentioned in Sec. 3.3, there are three different losses employed to train our network, *i.e.*, pixel-wise and feature-

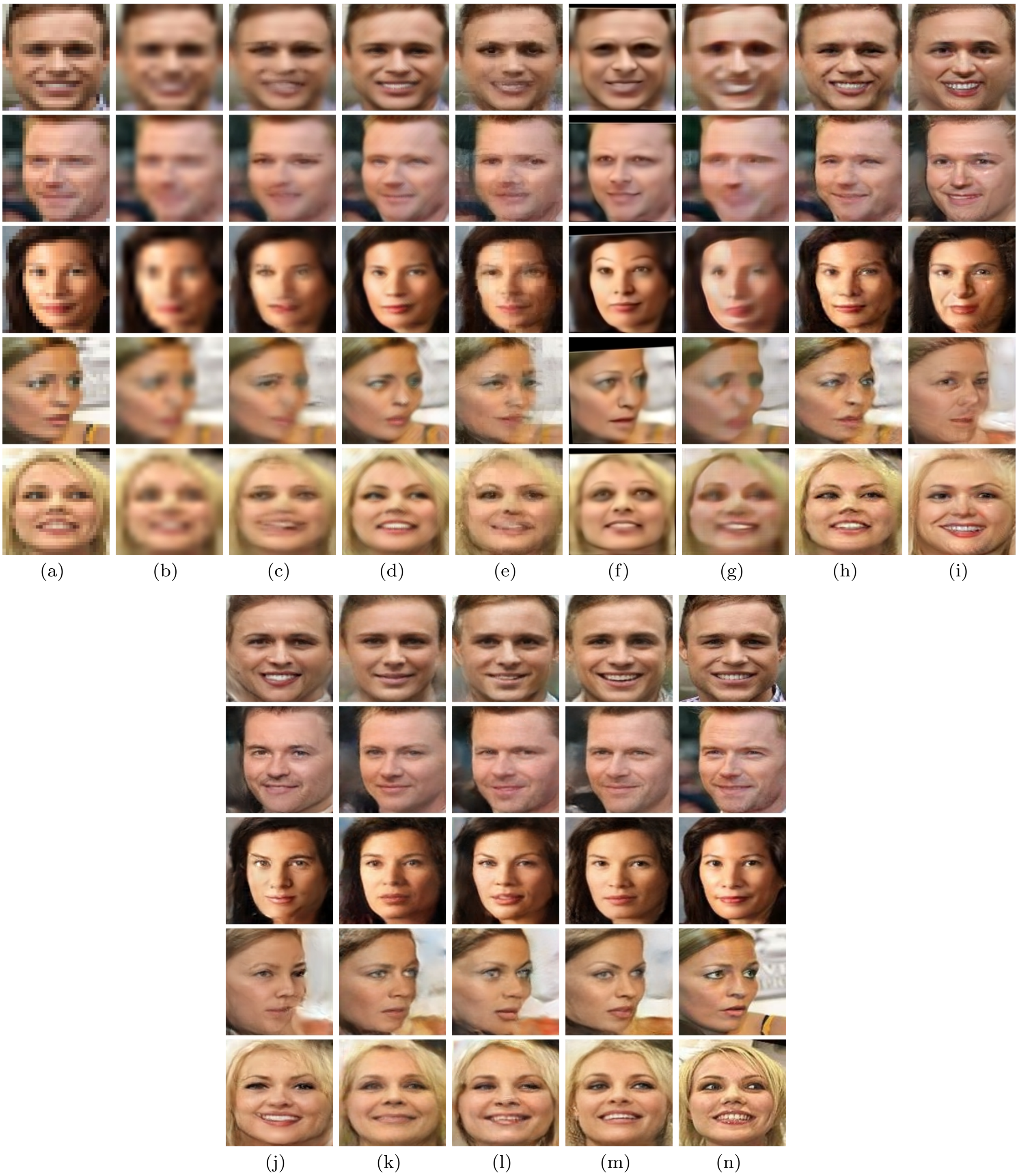


Fig. 11 Comparisons with the state-of-the-art methods on the *aligned* input images of size 32×32 pixels. (a) Unaligned LR inputs. (b) Bicubic interpolation. (c) Kim *et al.*'s method (Kim et al, 2016a) (VDSR). (d) Ledig *et al.*'s method (Ledig et al, 2016) (SRGAN). (e) Ma *et al.*'s method (Ma et al, 2010). (f) Zhu *et al.*'s method (Zhu et al, 2016a) (CBN). (g) Chen *et al.*'s method (Chen et al, 2018) (FSRNet). (h) Yu and Porikli's method (Yu and Porikli, 2016) (URDGN). (i) Yu and Porikli's method (Yu and Porikli, 2017a) (TDN). (j) Yu and Porikli's method (Yu and Porikli, 2017b) (TDAE). (k) Yu *et al.*'s method (Yu et al, 2018b). (l) Yu *et al.*'s method (Yu et al, 2018a). (m) Our method (MTDN). (n) Original HR images.

wise ℓ_2 losses and a class-wise discriminative loss. Pixel-wise ℓ_2 loss is used to constrain the appearance similarity. As reported in our previous work (Yu and Porikli, 2016) and as indicated in Tab. 4, the upsampling net-

Table 3 Quantitative evaluations on different STN layers

STNs	STN$_1$	STN$_2$	Ours
PSNR	24.29	24.69	**25.02**
SSIM	0.79	**0.80**	**0.80**

Table 4 Quantitative evaluations on different losses

Losses	$\mathcal{L}_{pix}$	$\mathcal{L}_{pix+feat}$	$\mathcal{L}_{pix+\mathcal{U}}$	$\mathcal{L}_{\mathcal{T}}$
PSNR	25.35	24.93	24.69	25.02
SSIM	0.81	0.81	0.79	0.80

Table 5 Quantitative evaluations on different input resolutions

Resolutions	16×16	24×24	32×32
PSNR	23.93	25.30	25.40
SSIM	0.78	0.81	0.81

Table 6 Quantitative evaluations on different components in our MTDN

Modules	NoAE	NoSkip	NoHF	Ours
PSNR	23.87	24.64	23.72	**25.02**
SSIM	0.79	0.80	0.78	**0.80**

Table 7 Face recognition performance on different input resolutions

Resolution	HR	LR	16×16	24×24	32×32
Accuracy	95.68%	77.27%	84.53%	89.68%	92.30%

work which is trained only by a pixel-wise ℓ_2 loss to super-resolve LR faces obtains the highest PSNR but produces over-smoothed results as shown in Fig. 4(f).

The feature-wise loss is able to make the super-resolved results sharper without suffering over smoothness because it forces the high-order moments of upsampled faces, *i.e.*, feature maps of faces, to be similar to their ground-truths. In addition, we also incorporate a class-wise discriminative loss to force the upsampling network to generate realistic faces. Since the class-specific loss is not used to measure the similarity between two images, too large discriminative loss will distort our super-resolution performance. Therefore, there is a trade-off between the upsampling and discriminative networks and we gradually decrease the influence of the discriminative network as iterations progress.

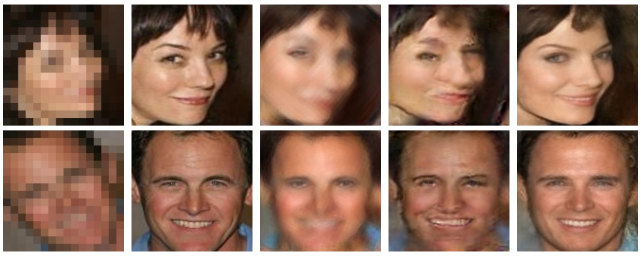


Fig. 12 Comparisons of different variants of our network. (a) The input 16 × 16 LR images. (b) The original 128 × 128 HR images. (c) Results of the network without using the autoencoder. (d) Results of IBSR. (e) Our results.

Because PSNR is designed to measure the similarity of appearance intensities but does not reflect visual quality of reconstructed images, using the feature-wise and class-wise losses decreases the PSNR, as seen in Tab. 4 but improves the visual quality significantly, as visible in Fig. 4.

5.3 Impacts of Multiple STN Layers

As illustrated in Fig. 2, we employ two STN layers to align feature maps in our network. Our previous works (Yu and Porikli, 2017a,b) only use one branch to upsample LR faces and they align feature maps at the resolution of 16×16 pixels. However, our MTDN has two branches and the resolutions of these two-branch inputs are different. Therefore, we apply STN layers after the concatenation layer, where the resolution of the feature maps is 32×32 pixels. In this manner, all feature maps can be aligned simultaneously. As mentioned in Jaderberg et al (2015), using multiple STNs can achieve more accurate alignment. Due to the GPU memory limitation, we cannot use an STN layer to align the feature maps of size 128×128 pixels. Hence, we only employ two STN layers to the feature maps of size 32×32 and 64×64 pixels in our network. As shown in Tab. 3, we demonstrate the contributions of different STN layers to the final performance. Note that, for the cases STN$_1$ and STN$_2$ in Tab. 3, a network is trained by only employing one STN layer. Table 3 also indicates that using multiple STN layers can improve face alignment, thus obtaining better face hallucination performance.

5.4 Effects of Autoencoder in Low-frequency Branch

Different from our previous works (Yu and Porikli, 2016, 2017a, 2018), our MTDN does not super-resolve LR faces directly by deconvolutional layers. Since our method needs to fuse two branch images together, we first extract feature maps from the two branch input images separately. In order to make the resolutions of the feature maps from the two branches compatible, we upsample the feature maps of the low-frequency branch to 32×32 pixels. In particular, we exploit an autoencoder with skip connections to extract features and then upsample features by a deconvolutional layer in the low-frequency branch while residual blocks are applied to extract features from the high-frequency branch. Since the resolution of the low-frequency branch is very small, the autoencoder does not require much GPU memory but increases the capacity of our network.

We replace the autoencoder with a convolutional layer and use a deconvolutional layer to upsample the LR feature maps in the low-frequency branch, and we represent this variant as noAE in Tab. 6. As demonstrated in Tab. 6, the performance of noAE degrades 1.15 dB compared to our MTDN. Therefore, by increasing the network capacity, *i.e.*, the employment of the autoencoder, our MTDN achieves better quantitative super-resolution performance. Furthermore, the upsampled faces also achieve better visual quality by using the autoencoder, as shown in Fig. 12. It also demonstrates that our autoencoder can extract feature maps better than a single convolutional layer. Since skip connections are employed in the autoencoder, we can also preserve the spatial information from the encoder to the decoder. Removal of the skip connections in our network, marked as NoSkip, causes 0.38 dB degradation in PSNR, as shown in Tab. 6. However, we do not observe significant deterioration in visual quality.

5.5 PSNR and SSIM at Different Input Resolutions

Since our test dataset consists of LR face images at different resolutions, it cannot reflect the performance of our network as the input resolutions increase. Hence, we generate another test dataset where each HR face image corresponds to three different LR image versions, *i.e.*, 16×16, 24×24 and 32×32 pixels. We group and super-resolve input LR images according to their resolutions and then measure the performance of our network in each group. As indicated in Tab. 5, our network generates better super-resolved results in terms of PSNR as the input resolution increases. It implies our proposed two-branch network can fully exploit input information when more information is provided in LR input images.

5.6 Interpolation before Super-resolution

There is another option for preserving all the information in the LR input images: we can first resize the different LR image sizes to 32×32 pixels by bicubic interpolation and then super-resolve the interpolated images by deconvolutional layers and residual blocks similar to the ones as illustrated in Fig. 2. We name this super-resolution approach as IBSR. As reported in previous generic super-resolution methods (Kim et al, 2016a; Ledig et al, 2016), using convolutional and deconvolutional layers can achieve better super-resolution performance than traditional interpolation methods, *e.g.*bicubic interpolation. Therefore, we use an autoencoder and a deconvolutional layer to upsample low-frequency part as well as residual blocks to extract features from high-frequency residuals. After obtaining the feature maps from the low-frequency and high-frequency branches, we fuse those feature maps by a residual block. In this fashion, we achieve 128 channel features maps of size 32×32 for further super-resolution rather than only 3 channel interpolated images in IBSR. Thus, our network architecture achieves better performance qualitatively and quantitatively, as demonstrated in Fig. 12(d) and Tab. 1.

Since IBSR increases the depth of the low-frequency branch of our network to receive LR face images of 32×32 pixels as inputs, the capacity of IBSR is larger than our low-frequency branch in terms of the number of network layers and parameters. As seen in Tab. 1, our network still outperforms IBSR by a margin of 0.76 dB. This implies that the main contribution to the final performance comes from the high-frequency residual branch in our network. Therefore, exploiting the high-frequency residuals explicitly improves the super-resolution performance significantly.

5.7 Two-Branch Architecture Versus TDAE

In order to investigate the necessity of our two-branch architecture modification, we retrain TDAE with input images of size 32×32 pixels, marked as TDAE$_{32}$. Then, we resize input images in different resolutions to 32×32 pixels by bicubic interpolation. Since TDAE is originally designed to receive LR face images of 16×16 pixels, we replace its first deconvolutional layer with a convolutional layer to accept input images of size 32×32 pixels. In this way, TDAE$_{32}$ does not lose any information of inputs. However, since TDAE$_{32}$ is composed of three separate sequential networks, the input of each network is a 3-channel image rather than feature maps from the previous network. Therefore, the bottlenecks of TDAE may restrict the super-resolution performance. On the contrary, our network employs an autoencoder with skip connections to extract more informative features from LR face images and the extracted 128-channel feature maps are upsampled by the following layers. As indicated in Tab. 1, our network outperforms TDAE_{32} by a margin of 1.06 dB. Therefore, our two-branch architecture facilitates the procedure of face hallucination.

5.8 Identity Preservation by MTDN

We employ a state-of-the-art pre-trained face recognition network (Liu et al, 2017), named SphereFaceNet, to conduct the standard face recognition tests on our hallucinated HR faces. Following the standard divisions of the training and test datasets in the LFW benchmark (Huang et al, 2007), we generate LR and HR image pairs and report the standard LFW face verification performance in Tab. 7. Furthermore, we also employ two baselines in the experiments, the face recognition performance on original HR faces as well as aggressively downsampled LR faces (*i.e.*16×16 pixels). The face recognition performance on original HR faces is marked as HR, and the performance on LR faces is marked as LR. For the second baseline of face recognition on LR faces, LR faces are firstly upsampled by bicubic interpolation and then fed into SphereFaceNet to meet the image resolution requirement of SphereFaceNet. As seen in Tab. 7, our method improves the face recognition performance on LR faces. Thus, our generated HR face images preserve the identity information with respect to their corresponding ground-truth ones. In addition, as the resolutions of LR inputs increase, our method achieves better face recognition performance. This indicates that our MTDN is able to exploit all the information in the LR face images.

6 Discussions

6.1 Robustness to LR faces in Large Poses

In Fig. 13, we illustrate that our method is able to super-resolve LR faces in large poses. Since the input LR faces in large poses are unaligned and also contain self-occlusions, it is challenging to align as well as super-resolve them, as seen in the third and fourth rows of Fig. 9. As shown in Fig. 13, our MTDN is able to align and upsample LR faces in different large poses effectively and the upsampled faces are similar to their corresponding HR ground-truths. This indicates that our method is robust to upsample LR faces in different views.

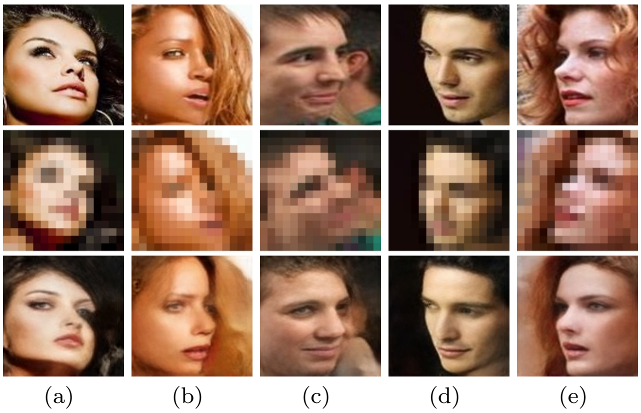


Fig. 13 Illustration of upsampling LR faces in large poses by our MTDN. From top to bottom: the original 128 × 128 HR images, the input 16 × 16 LR images and our results.

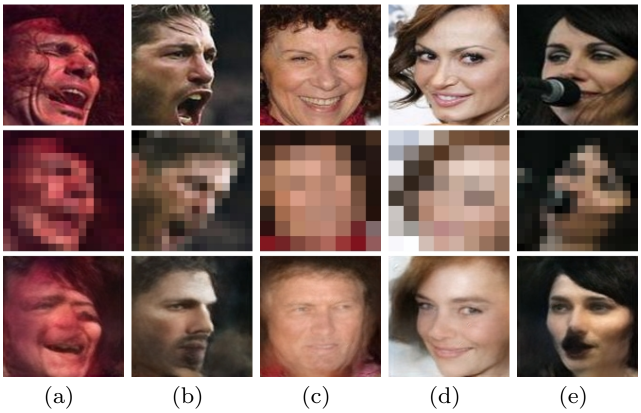


Fig. 14 Failure cases. From top to bottom: the original 128× 128 HR images, the input 16 × 16 LR images and our results.

6.2 Real World Cases

Since it is easy to obtain real-world LR face images but very difficult to attain their corresponding HR images, we use bicubic downsampling to mimic the degradation process. Although our network is trained on CelebA dataset, our model can also super-resolve real-world LR face images effectively, as seen in Fig. 15. In Fig. 15, we randomly choose LR face images from 16×16 pixels to 32×32 pixels in WiderFace dataset (Yang et al, 2016) where LR faces are captured in the wild. As visible in real-world LR faces, the mosaic artifacts and noise are obvious, which can degrade the super-resolution performance. We believe with proper data augmentation our network is able to super-resolve real-world LR faces even better.

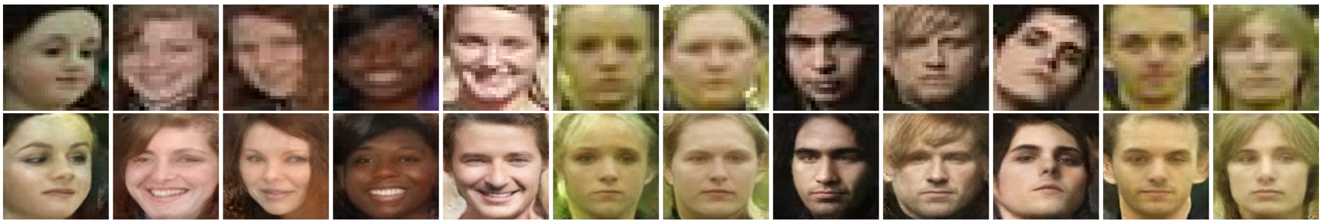


Fig. 15 Real-world cases. The top row: real-world LR faces captured in the wild. The bottom row: our super-resolved results.

6.3 Limitations

Although our MTDN is able to hallucinate LR faces in different views, it may suffer severe artifacts when LR facial patterns are indistinguishable as shown in Fig. 14(a). Furthermore, since some facial expressions in large views do not have enough samples in the training dataset, our network fails to learn the mappings as seen in Fig. 14(b). As visible in Fig. 14(c) and Fig. 14(d), when the input resolutions are even smaller, *i.e.*, 8×8 pixels, our network may not recognize the LR facial patterns accurately and thus generates blurry results. Note that, we do not retrain our MTDN on LR images of size 8×8 pixels. Since STN layers as an attention mechanism focus on aligning facial features and deconvolutional layers aim at super-resolving them instead of features of generic objects, our MTDN is not designed to upsample generic objects. Figure 14(e) illustrates that the occluded regions in the image are not well super-resolved.

7 Conclusion

We present a novel and capable multiscale transformative discriminative network to super-resolve very small LR face images. By designing a two-branch input neural network, our network can upsample LR images in various resolutions without discarding the residuals of resized input images. In this manner, our method is able to utilize all the information from inputs for face super-resolution. Furthermore, our algorithm can increase the input LR image size significantly, *e.g.* 8×, and reconstruct much richer facial details. Since our method does not require any alignments of LR faces and learns an end-to-end mapping between LR and HR face images, it preserves well the global structure of faces and is more practical.

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

References

Arandjelović O (2014) Hallucinating optimal high-dimensional subspaces. Pattern Recognition 47(8):2662–2672

Baker S, Kanade T (2000) Hallucinating faces. In: Proceedings of 4th IEEE International Conference on Automatic Face and Gesture Recognition, FG 2000, pp 83–88

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

Bruna J, Sprechmann P, LeCun Y (2016) Super-resolution with deep convolutional sufficient statistics. In: International Conference on Learning Representations (ICLR)

Bulat A, Tzimiropoulos G (2017) How far are we from solving the 2d and 3d face alignment problem? (and a dataset of 230,000 3d facial landmarks). In: Proceeding of International Conference on Computer Vision (ICCV)

Bulat A, Tzimiropoulos G (2018) Super-fan: Integrated facial landmark localization and super-resolution of real-world low resolution faces in arbitrary poses with gans. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)

Bulat A, Yang J, Tzimiropoulos G (2018) To learn image super-resolution, use a gan to learn how to do image degradation first. In: Proceedings of European Conference on Computer Vision (ECCV), pp 185–200

Chen Y, Tai Y, Liu X, Shen C, Yang J (2018) Fsrnet: End-to-end learning face super-resolution with facial priors. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)

Dahl R, Norouzi M, Shlens J (2017) Pixel recursive super resolution. In: Proceeding of International Conference on Computer Vision (ICCV), pp 5439–5448

Denton E, Chintala S, Szlam A, Fergus R (2015) Deep Generative Image Models using a Laplacian Pyramid of Adversarial Networks. In: Advances In Neural Information Processing Systems (NIPS), pp 1486–1494

Dong C, Loy CC, He K (2016) Image Super-Resolution Using Deep Convolutional Networks. IEEE Transactions on Pattern Analysis and Machine Intelligence 38(2):295–307

Freedman G, Fattal R (2010) Image and video upscaling from local self-examples. ACM Transactions on Graphics 28(3):1–10

Freeman WT, Jones TR, Pasztor EC (2002) Example-based super-resolution. IEEE Computer Graphics and Applications 22(2):56–65

Glasner D, Bagon S, Irani M (2009) Super-Resolution from a Single Image. In: Proceedings of IEEE International Conference on Computer Vision (ICCV), pp 349–356

Goodfellow I, Pouget-Abadie J, Mirza M (2014) Generative Adversarial Networks. In: Advances in Neural Information Processing Systems (NIPS), pp 2672—-2680

Gu S, Zuo W, Xie Q, Meng D, Feng X, Zhang L (2015) Convolutional Sparse Coding for Image Super-resolution. In: Proceedings of the IEEE International Conference on Computer Vision (ICCV)

Hennings-Yeomans PH, Baker S, Kumar BV (2008) Simultaneous super-resolution and feature extraction for recognition of low-resolution faces. In: Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR), pp 1–8

Hinton G (2012) Neural Networks for Machine Learning Lecture 6a: Overview of mini-batch gradient descent Reminder: The error surface for a linear neuron. Tech. rep.

Hong Chang, Dit-Yan Yeung, Yimin Xiong (2004) Super-resolution through neighbor embedding. In: Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR), vol 1, pp 275–282

Huang GB, Ramesh M, Berg T, Learned-Miller E (2007) Labeled faces in the wild: A database for studying face recognition in unconstrained environments. Tech. Rep. 07-49, University of Massachusetts, Amherst

Huang H, He R, Sun Z, Tan T (2017) Wavelet-srnet: A wavelet-based cnn for multi-scale face super resolution. In: Proceeding of International Conference on Computer Vision (ICCV)

Huang JB, Singh A, Ahuja N (2015) Single image super-resolution from transformed self-exemplars. In: Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR), pp 5197–5206

Jaderberg M, Simonyan K, Zisserman A, et al (2015) Spatial transformer networks. In: Advances in Neural Information Processing Systems (NIPS), pp 2017–2025

Johnson J, Alahi A, Fei-Fei L (2016) Perceptual losses for real-time style transfer and super-resolution. In: Proceedings of European Conference on Computer Vision (ECCV)

Kim J, Kwon Lee J, Mu Lee K (2016a) Accurate image super-resolution using very deep convolutional networks. In: Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR), pp 1646–1654

Kim J, Kwon Lee J, Mu Lee K (2016b) Deeply-recursive convolutional network for image super-resolution. In: Proceedings of the IEEE conference on computer vision and pattern recognition (CVPR), pp 1637–1645

Kolouri S, Rohde GK (2015) 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)

Lai WS, Huang JB, Ahuja N, Yang MH (2017) Deep laplacian pyramid networks for fast and accurate super-resolution. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp 624–632

Ledig C, Theis L, Huszár F, Caballero J, Cunningham A, Acosta A, Aitken A, Tejani A, Totz J, Wang Z, et al (2016) Photo-realistic single image super-resolution using a generative adversarial network. arXiv preprint arXiv:160904802

Li Y, Cai C, Qiu G, Lam KM (2014) Face hallucination based on sparse local-pixel structure. Pattern Recognition 47(3):1261–1270

Lin Z, Shum HY (2006) Response to the comments on ”Fundamental limits of reconstruction-based super-resolution algorithms under local translation”. IEEE Transactions on Pattern Analysis and Machine Intelligence 28(5):83–97

Lin Z, He J, Tang X, Tang CK (2008) Limits of learning-based superresolution algorithms. International journal of computer vision 80(3):406–420

Liu C, Shum H, Zhang C (2001) A two-step approach to hallucinating faces: global parametric model and local nonparametric model. In: Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR), vol 1, pp 192–198

Liu C, Shum HY, Freeman WT (2007) Face hallucination: Theory and practice. International Journal of Computer Vision 75(1):115–134

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

Liu W, Wen Y, Yu Z, Li M, Raj B, Song L (2017) Sphereface: Deep hypersphere embedding for face recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), vol 1, p 1

Liu Z, Luo P, Wang X, Tang X (2015) Deep learning face attributes in the wild. In: Proceedings of International Conference on Computer Vision (ICCV)

Ma X, Zhang J, Qi C (2010) Hallucinating face by position-patch. Pattern Recognition 43(6):2224–2236

Radford A, Metz L, Chintala S (2015) Unsupervised Representation Learning with Deep Convolutional Generative Adversarial Networks. arXiv:151106434 pp 1–15

Shi W, Caballero J, Huszár F, Totz J, Aitken AP, Bishop R, Rueckert D, Wang Z (2016) Real-time single image and video super-resolution using an efficient sub-pixel convolutional neural network. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp 1874–1883

Simonyan K, Zisserman A (2014) Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:14091556

Singh A, Porikli F, Ahuja N (2014) Super-resolving noisy images. In: Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR), pp 2846–2853

Tai Y, Yang J, Liu X (2017) Image super-resolution via deep recursive residual network. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), vol 1

Tappen MF, Liu C (2012) A Bayesian Approach to Alignment-Based Image Hallucination. In: Proceedings of European Conference on Computer Vision (ECCV), vol 7578, pp 236–249

Tappen MF, Russell BC, Freeman WT (2003) Exploiting the sparse derivative prior for super-resolution and image demosaicing. In: In IEEE Workshop on Statistical and Computational Theories of Vision

Van Den Oord A, Kalchbrenner N, Kavukcuoglu K (2016) Pixel recurrent neural networks. In: Proceedings of International Conference on International Conference on Machine Learning (ICML), pp 1747–1756

Wang N, Tao D, Gao X, Li X, Li J (2014) A comprehensive survey to face hallucination. International Journal of Computer Vision 106(1):9–30

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

Xu X, Sun D, Pan J, Zhang Y, Pfister H, Yang MH (2017) Learning to super-resolve blurry face and text images. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (ICCV), pp 251–260

Yang CY, Liu S, Yang MH (2013) Structured face hallucination. In: Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR), pp 1099–1106

Yang CY, Liu S, Yang MH (2017) Hallucinating compressed face images. International Journal of Computer Vision pp 1–18

Yang J, Wright J, Huang TS, Ma Y (2010) Image super-resolution via sparse representation. IEEE transactions on image processing 19(11):2861–73

Yang S, Luo P, Loy CC, Tang X (2016) Wider face: A face detection benchmark. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp 5525–5533

Yu X, Porikli F (2016) Ultra-resolving face images by discriminative generative networks. In: European Conference on Computer Vision (ECCV), pp 318–333

Yu X, Porikli F (2017a) Face hallucination with tiny unaligned images by transformative discriminative neural networks. In: Thirty-First AAAI Conference on Artificial Intelligence

Yu X, Porikli F (2017b) 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), pp 3760–3768

Yu X, Porikli F (2018) Imagining the unimaginable faces by deconvolutional networks. IEEE Transactions on Image Processing 27(6):2747–2761

Yu X, Xu F, Zhang S, Zhang L (2014) Efficient patch-wise non-uniform deblurring for a single image. IEEE Transactions on Multimedia 16(6):1510–1524

Yu X, Fernando B, Ghanem B, Porikli F, Hartley R (2018a) Face super-resolution guided by facial component heatmaps. In: Proceedings of European Conference on Computer Vision (ECCV), pp 217–233

Yu X, Fernando B, Hartley R, Porikli F (2018b) Super-resolving very low-resolution face images with supplementary attributes. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp 908–917

Yu X, Fernando B, Hartley R, Porikli F (2019a) Semantic face hallucination: Super-resolving very low-resolution face images with supplementary attributes. IEEE transactions on pattern analysis and machine intelligence

Yu X, Shiri F, Ghanem B, Porikli F (2019b) Can we see more? joint frontalization and hallucination of unaligned tiny faces. IEEE transactions on pattern analysis and machine intelligence

Zeiler MD, Fergus R (2014) Visualizing and understanding convolutional networks. In: European Conference on Computer Vision (ECCV), pp 818–833

Zeiler MD, Krishnan D, Taylor GW, Fergus R (2010) Deconvolutional networks. In: Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR), pp 2528–2535

Zhou E, Fan H (2015) Learning Face Hallucination in the Wild. In: Twenty-Ninth AAAI Conference on Artificial Intelligence, pp 3871–3877

Zhu S, Liu S, Loy CC, Tang X (2016a) Deep cascaded bi-network for face hallucination. In: Proceedings of European Conference on Computer Vision (ECCV), pp 614–630

Zhu S, Liu S, Loy CC, Tang X (2016b) Deep cascaded bi-network for face hallucination. In: European Conference on Computer Vision (ECCV), pp 614–630

Zhu X, Ramanan D (2012) Face detection, pose estimation, and landmark localization in the wild. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp 2879–2886