Hallucinating Very Low-Resolution Unaligned and Noisy Face Images by Transformative Discriminative Autoencoders*

Xin Yu, Fatih Porikli
Australian National University
{xin.yu, fatih.porikli}@anu.edu.au

Abstract

Most of the conventional face hallucination methods assume the input image is sufficiently large and aligned, and all require the input image to be noise-free. Their performance degrades drastically if the input image is tiny, unaligned, and contaminated by noise.

*In this paper, we introduce a novel transformative discriminative autoencoder to 8× super-resolve unaligned noisy and tiny (16×16) low-resolution face images. In contrast to encoder-decoder based autoencoders, our method uses decoder-encoder-decoder networks. We first employ a transformative discriminative decoder network to upsample and denoise simultaneously. Then we use a transformative encoder network to project the intermediate HR faces to aligned and noise-free LR faces. Finally, we use the second decoder to generate hallucinated HR images. Our extensive evaluations on a very large face dataset show that our method achieves superior hallucination results and outperforms the state-of-the-art by a large margin of **1.82** dB PSNR.*

1. Introduction

Face images provide critical information for visual perception and identity analysis. However, when they are noisy and their resolutions are inadequately small (*e.g.*as in some surveillance videos), there is little information available to be inferred reliably from them. Very low-resolution and noisy face images not only impede human perception but also impair computer analysis.

To tackle this challenge, face hallucination techniques aim at recovering high-resolution (HR) counterparts from low-resolution (LR) face images and have received significant attention in recent years. Previous state-of-the-art methods mainly focus on recovering HR faces from aligned and noise-free LR face images. More specifically, face hallucination methods based on holistic appearance models [1, 2, 14, 20, 15, 8, 22, 24, 13, 12, 19, 25] require LR faces to be precisely aligned beforehand. However, when the LR images are contaminated by noise, the accuracy of face alignment degrades dramatically. Besides, due to the wide range of pose and expression variations, it is difficult to learn a comprehensive, holistic appearance model for LR images not aligned appropriately. As a result, these methods often produce ghosting artifacts for noisy unaligned LR inputs.

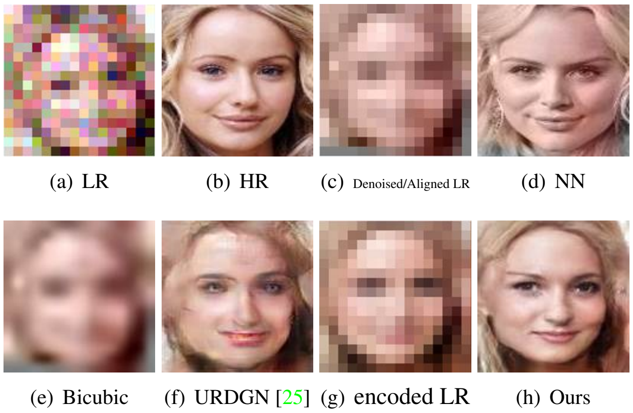

Figure 1. Comparison of our method with the CNN based face hallucination URDGN [25]. (a) 16 × 16 LR input image. (b) 128 × 128 HR original image. (c) Denoised and aligned LR image. We firstly apply BM3D [4] and then STN [10]. (d) The corresponding most similar face in the training dataset. (e) Bicubic interpolation of (c). (f) Image generated by URDGN. Note that, URDGN super-resolves the denoised and aligned LR image, not the original LR input (in favor of URDGN). (g) The denoised and aligned LR image by our decoder-encoder as an intermediate output. (h) The final hallucinated face by our TDAE method.

Rather than learning holistic appearance models, facial components based face hallucination methods have been proposed [18, 23, 28, 29]. They transfer HR facial components from the training dataset to the input LR images without requiring alignment of LR input images in advance. These methods heavily rely on the successful localization of

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

facial landmarks. Because facial landmarks are difficult to detect in very low resolution (16×16 pixels) images, they fail to localize the facial components accurately and thus produce artifacts in the upsampled face images. In other words, the facial component based methods are not suitable to upsample noisy unaligned LR faces either.

Considering the resolution of faces is too small and the presence of noise, face detectors may also fail to locate such tiny noisy faces. Thus, using pose specific face detectors as a preprocessing step to compensate for misalignments is also impractical.

In this paper, we propose a new transformative discriminative autoencoder (TDAE) to super-resolve a tiny (16×16 pixels) unaligned and noisy face image by a remarkable upscaling factor of 8×, where we estimate 64 pixels for each single pixel of the input LR image. Furthermore, each pixel has also been contaminated by noise, making the task even more challenging.

Our TDAE consists of three serial components: a decoder, an encoder, and a second decoder. Our decoder network comprises deconvolutional and spatial transformation layers [10]. It can progressively upsample the resolutions of the feature maps by its deconvolutional layers while aligning the feature maps by its spatial transformation layers. Similar to [25], we employ not only the pixel-wise intensity similarity between the hallucinated face images and the ground-truth HR face images but also the class similarity constraint that enforces the upsampled faces to lie on the manifold of real faces by a discriminative network. Hence, we achieve a transformative decoder that is also discriminative. Since the LR inputs are noisy, the hallucinated faces after the decoder may still contain artifacts. In order to obtain aligned and noise-free LR faces, we project the upsampled HR faces back onto the LR face domain by a transformative encoder. Finally, we train our second decoder on the projected LR faces to attain hallucinated HR face images. In this manner, the artifacts are greatly reduced and our TDAE produces authentic HR face images.

Overall, the contributions of this paper are mainly in four aspects:

- We propose a new transformative-discriminative architecture to hallucinate tiny (16×16 pixels) unaligned and noisy face images by an upscaling factor of 8×.
- In contrast to conventional autoencoders, we first device a decoder-encoder structure to generate noise-free and aligned LR faces, and then a second decoder trained on the encoded LR faces to hallucinate high-quality HR face images.
- Our method does not require to model or estimate noise parameters. It is agnostic to the underlying spatial deformations and contaminated noise.
- To the best of our knowledge, our method is the first attempt to address the super-resolution of tiny and noisy face images without requiring alignment of LR faces beforehand, which makes our method practical.

2. Related Work

Face hallucination has received significant attention in recent years [18, 23, 19, 12, 28, 29, 25]. Previous face hallucination methods mainly focus on recovering HR faces from aligned and noise-free LR face images, and in general, they can be grouped into two categories: holistic methods and part-based methods.

Holistic methods use global face models learned by P-CA to hallucinate entire HR faces. In [20], an eigen-transformation is proposed to generate HR face images by establishing a linear mapping between LR and HR face subspaces. Similarly, [15] employs a global appearance model learned by PCA to upsample aligned LR faces and a local non-parametric model to enhance the facial details. The work in [12] explores optimal transport and subspace learning to morph an HR output according to the given aligned LR faces. Since holistic methods require LR face images to be precisely aligned and share the same pose and expression as the HR references, they are very sensitive to the misalignments of LR images. Besides, image noise makes the alignment of LR faces even more difficult.

Part-based methods upsample facial parts rather than entire faces, and thus they can handle various poses and expressions. They either employ a training dataset of reference patches to reconstruct the HR counterparts of the input LR patches or exploit facial components. In [2], high-frequency details of aligned frontal face images are reconstructed by finding the best mapping between LR and HR patches. The work in [24] uses coupled LR/HR dictionaries to enhance the details. In [22], an LR face image is super-resolved with position patches sampled from multiple aligned HR images. [13] models the local face patches as a sparse coding problem rather than averaging the reference HR patches directly. In [18], SIFT flow [16] is exploited to align the facial parts of LR images, and then the details of LR images are reconstructed by warping the reference HR images. [23] first localizes facial components in the LR images and then transfers the most similar HR facial components in the dataset to the LR inputs. Since part-based methods often require extraction of facial components in L-R inputs, their performance degrades dramatically when the LR faces are tiny or noisy.

As large-scale data becomes available, convolutional neural network (CNN) based SR methods have been proposed and achieved the state-of-the-art performance [11, 21, 6, 3]. However, because these SR methods are designed to upsample generic patches and do not fully exploit class-specific information, they are not suitable to hallucinate tiny faces. The work in [28] employs a CNN to extract facial features and then generates high-frequency facial details based

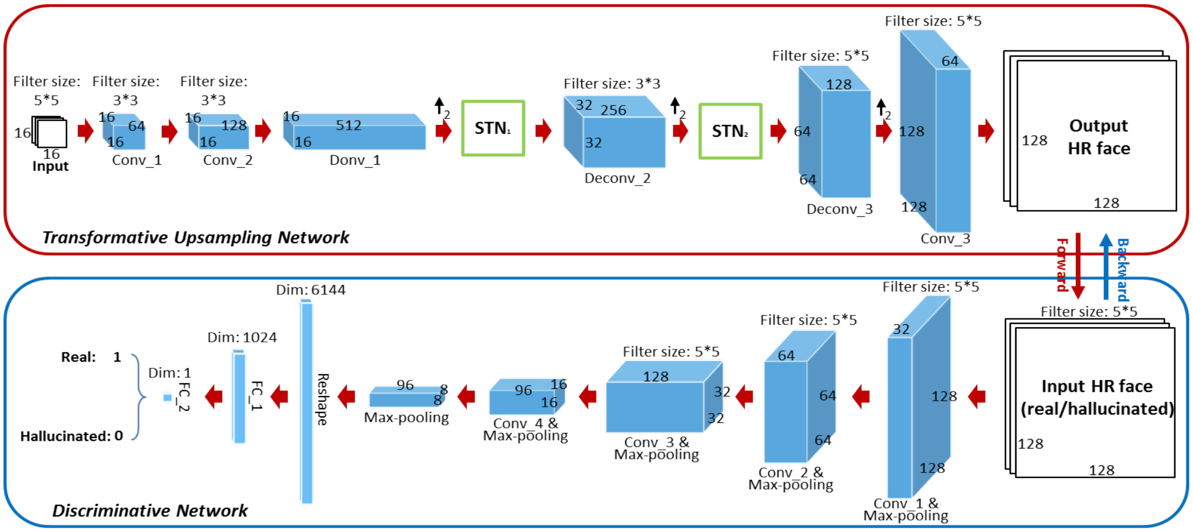

Figure 2. Our transformative discriminative decoder consists of two parts: a transformative upsampling network (in the red frame) and a discriminative network (in the blue frame).

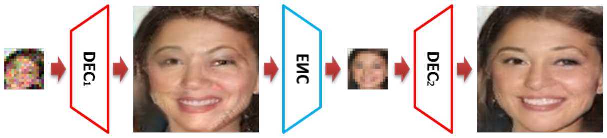

Figure 4. Workflow of our transformative discriminative autoencoder. Colors of the boxes refer to the networks in Fig.2 and Fig.3.

on the extracted features. Due to the requirement of the facial feature extraction, the resolution of the input cannot be low. Very recently, [25] presents a discriminative generative network to super-resolve LR face images. This method addresses different facial expressions and head poses without requiring facial landmarks, but it needs the eyes to be aligned in advance. [29] proposes a cascade bi-network to super-resolve very low-resolution and unaligned faces. However, when there is noise in the LR images, this method may fail to localize the face parts accurately, thus producing artifacts in the outputs.

3. Proposed Method: TDAE

Our transformative discriminative autoencoder has three complementary components: two transformative discriminative decoders (as shown in Fig. 2) and a transformative encoder (as shown in Fig. 3). In the training phase, our parameters of TDAE are learned in three steps (§.3.3). In the testing phase, we cascade the transformative upsampling network of the first decoder DEC_1, the encoder ENC, and the second decoder DEC_2 together to hallucinate the final HR faces in an end-to-end manner. The whole pipeline is illustrated in Fig. 4

3.1. Architecture of Decoder

Our decoder architecture is composed of two subnetworks, a transformative upsampling network (TUN) and a discriminative network. In the transformative upsampling network, we first apply two convolutional layers with larger receptive fields to partially reduce noise artifacts rather than feeding noisy images into the deconvolutional layers directly. The deconvolutional layer can be made of a cascade of an upsampling layer and a convolutional layer, or a convolutional layer with a fractional stride [27, 26]. Therefore, the resolution of the output image of the deconvolutional layer is larger than the resolution of its input image. We employ the ℓ_2 regression loss, also known as Euclidean distance loss, to constrain the similarity between the hallucinated HR faces and their HR ground-truth versions.

As reported in [25], deconvolutional layers supervised by ℓ_2 loss tend to produce over-smoothed results. To tackle this, we embed the class-specific discriminative information into the deconvolutional layers by a discriminative network (as shown in the blue frame in Fig. 2). The discriminative network is able to distinguish whether an image (its input) is sampled from authentic face images or hallucinated ones. The corresponding discriminative information is backpropagated to the deconvolutional layers. Hence, the deconvolutional layers can generate HR face images more similar to the real faces.

We notice that rotational and scale misalignments of LR face images will lead to apparent artifacts in the upsampled face images in [25]. By contrast, our decoder can align the LR faces automatically and hallucinate face images simultaneously. In order to align LR faces, we incorporate the spatial transformation network (STN) [10] into our network, as shown in the green box in Fig. 2. STN can estimate the

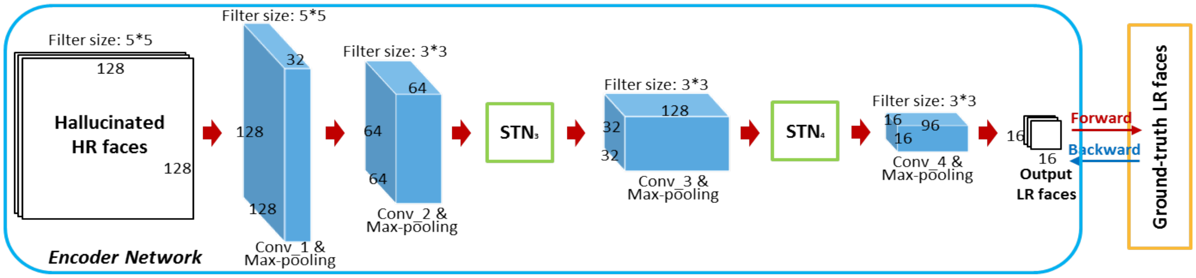

Figure 3. Architecture of our transformative encoder.

transformation parameters of images, and then warp images to a canonical view.

There are three modules in STN: 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 constructs a sampling grid according to the estimated parameters, and then the sampler module maps the input onto the generated grid by bilinear interpolation.

Here, we mainly focus on in-plane rotations, translations, and scale changes, and thus use the similarity transformation to align faces. Considering the resolution of our inputs is very small and input images are noisy, using state-of-the-art denoising algorithms to reduce noise and then employing an STN to align LR faces will introduce extra blurriness, as shown in Fig. 1(c) and Fig. 5(c). Therefore, aligning LR faces in the image domain may blur the original LR facial patterns and leads to artifacts as visible in the results of [25] in Fig. 1(f). To prevent from this, we apply STNs to *align feature maps*. As reported in [10], using multiple STNs can improve the accuracy of the alignment. As a trade-off between the accuracy and GPU memory usage, we employ two STNs following the first two deconvolutional layers.

Our decoder not only embeds discriminative information but also processes multiple tasks (denoising, alignment, and upsampling) simultaneously. As shown in Fig. 5(f), our transformative discriminative decoder can reconstruct more salient high-frequency details and aligned upsampled HR face images as well.

3.2. Architecture of Encoder

By feeding an unaligned and noisy LR input to our transformative discriminative decoder network DEC$_1$, we obtain an intermediate HR face image. As shown in Fig. 5(f), the intermediate HR face contains more high-frequency details and it is roughly aligned. The noise is comparatively reduced as well. However, the intermediate images may still contain artifacts, which are mainly caused by noise. We observe that noise not only distorts the LR facial patterns but also affects the face alignment. In order to achieve authentic HR face images, these artifacts should be removed while preserving the high-frequency facial details.

Our intuition is that projecting intermediate HR images to LR images, artifacts and noise can be suppressed further, which would allow us to apply our decoder to super-resolve these almost noise-free and approximately aligned LR faces. However, a decimation with anti-aliasing or simple downsampling may introduce additional artifacts into the LR face images. Therefore, we design another CNN, regarded as the encoder ENC, to project intermediate HR images to noise-free LR versions as illustrated in Fig. 3. Considering the upsampled HR faces may still have misalignments, we also incorporate STNs into our encoder to provide further alignment improvement.

When training the encoder, we constrain the projected LR faces to be similar to the aligned ground-truth LR faces. This helps us to generate aligned and noise-free LR faces, as shown in Fig. 1(g) and Fig. 5(g).

To obtain HR face images, we employ a second decoder DEC$_2$ to super-resolve the LR faces projected by the ENC. The decoder DEC$_2$ shares the same architecture as the one in Fig. 2. By employing the decoder-encoder structure, we can jointly align the input LR faces and handle noise as shown in Fig. 1(g) and Fig. 5(g). By exploiting the encoder-decoder structure, we are able to remove artifacts in the upsampled HR faces, thus achieving high-quality, more authentic, hallucinated HR face images as shown in Fig. 5(h).

3.3. Training Details of TDAE

We divide the training phase of our TDAE into three stages: i) Training the transformative discriminative decoder network DEC$_1$, as illustrated in Fig. 2. ii) Training the encoder ENC, as shown in Fig. 3. iii) Training the decoder DEC$_2$, which shares the same architecture as DEC$_1$.

3.3.1 Training Discriminative Decoder

We construct LR and HR face image pairs $\{l_i^n, h_i\}$ as our training dataset for the training of our transformative discriminative decoder DEC$_1$. Here, h_i represents aligned HR face images, and l_i^n is *not* directly downsampled from the HR face image h_i. We apply rotations, translations, and scale changes to h_i to obtain unaligned HR image h_i^u. Then,

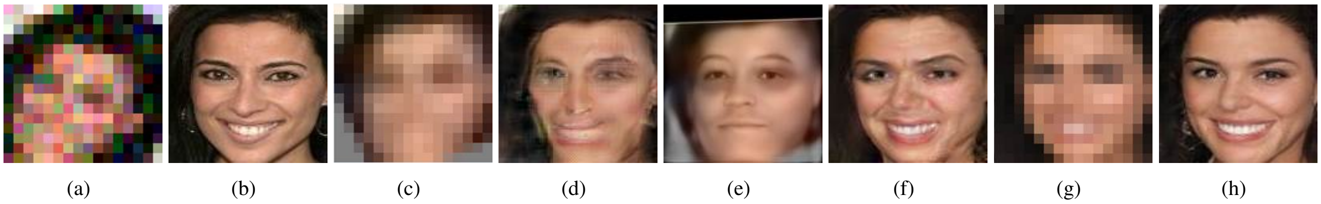

Figure 5. Comparison of our method with the CNN based face hallucination methods. (a) The input 16 × 16 LR image. (b) The original upright 128 × 128 HR image (for comparison purposes). (c) The denoised and aligned version of (a). (d) The result of URDGN [25]. (e) The result of CBN [29]. (f) The result of our DEC$_1$. (g) The aligned and noise-free LR face projected by our ENC. (h) Our final result.

we downsample h_i^u and then add Gaussian noise to obtain the noisy unaligned LR faces l_i^n.

Since we impose the upsampled image $\hat{h}_i$ by our decoder should be similar to its corresponding reference HR image h_i, we use pixel-wise Euclidean distance, as known as ℓ_2 regression loss, to enforce the intensity similarity. The loss function $U(s)$ of the TUN is modeled as,

$$\min_s U(s) = \mathbb{E}_{(l_i^n, h_i)\sim p(l^n, h)} \|\hat{h}_i - h_i\|_F^2, \tag{1}$$

where s indicates the parameters of the TUN. The convolutional layers, the STN layers, and the deconvolutional layers are updated jointly in the TUN. The STN layers align the feature maps while the deconvolutional layers upsample the resolution of the feature maps gradually. Here, $p(l^n, h)$ indicates the joint distribution of the LR and HR face images in the training dataset.

As mentioned in [25], only applying intensity similarity constraint will lead to over-smoothed results. Similar to [7, 5, 17, 25], we infuse class-specific discriminative information into the TUN by exploiting a discriminative network. The architecture of the discriminative network is illustrated in the blue frame in Fig. 2. It is designed to distinguish whether an image is realistic or hallucinated. If an HR face super-resolved by our decoder can convince the discriminative network that it is a real face image, our hallucinated faces will be similar to real face images. In other words, our goal is to make the discriminative network fail to distinguish hallucinated faces from real ones. Hence, we maximize the cross-entropy of the discriminative network L as follows:

$$\begin{aligned}\max_t L(t) &= \mathbb{E}\left[\log D(h_i) + \log(1 - D(\hat{h}_i))\right] \\ &= \mathbb{E}_{h_i\sim p(h)}[\log D(h_i)] + \mathbb{E}_{\hat{h}_i\sim p(\hat{h}))}[\log(1 - D(\hat{h}_i))],\end{aligned} \tag{2}$$

where t represents the parameters of the discriminative network, $p(h)$ and $p(\hat{h})$ indicate the distributions of the real faces and the hallucinated faces, and $D(h_i)$ and $D(\hat{h}_i)$ are the outputs of the discriminative network. The loss L is backpropagated to the TUN in order to update the parameters s. By injecting discriminative information to s, our decoder can hallucinate more authentic HR faces.

In our decoder network, every layer is differentiable, and thus we use backpropagation to learn its parameters. RMSprop [9] is employed to update s and t. To maximize the discriminative network objective L, we use the stochastic gradient ascent that updates the parameters t as follows:

$$\begin{aligned}\Delta^{i+1} &= \gamma\Delta^i + (1-\gamma)(\frac{\partial L}{\partial t})^2, \\ t^{i+1} &= t^i + r\frac{\partial L}{\partial t}\frac{1}{\sqrt{\Delta^{i+1}+\epsilon}},\end{aligned} \tag{3}$$

where r and γ are the learning rate and decay rate, respectively, i is the index of iteration, Δ is an auxiliary variable, and ϵ is set to 10^{-8} to avoid division by zero. For the TUN, both losses U and L are used to update the parameters s by the stochastic gradient descent,

$$\begin{aligned}\Delta^{i+1} &= \gamma\Delta^i + (1-\gamma)(\frac{\partial U}{\partial s} + \lambda\frac{\partial L}{\partial s})^2, \\ s^{i+1} &= s^i - r(\frac{\partial U}{\partial s} + \lambda\frac{\partial L}{\partial s})\frac{1}{\sqrt{\Delta^{i+1}+\epsilon}},\end{aligned} \tag{4}$$

where λ is a trade-off weight between the intensity similarity term and the class similarity term. Since our goal is to hallucinate an HR face, we put a higher weight on the intensity similarity term and set λ to 0.01. As the iteration progresses, the super-resolved faces will be more similar to real faces. Therefore, we gradually reduce the impact of the discriminative network by decreasing λ as,

$$\lambda^j = \max\{\lambda \cdot 0.99^j, \lambda/2\}, \tag{5}$$

where j indicates the index of the epochs. Eqn. 5 also guarantees that the class-specific discriminative information is preserved in the decoder network during the training phase.

3.3.2 Training Encoder

In training our transformative encoder, we use the outputs of DEC$_1$ $\hat{h}_i$ and the ground-truth aligned LR images l_i as our training dataset. Since there may be misalignment in $\hat{h}_i$, we also embed STNs into our encoder ENC to align the LR faces. During the training of the transformative encoder, the downsampled LR faces $\hat{l}_i$ is constrained to be similar to the ground-truth aligned LR faces l_i. Therefore, the objective function of the transformative encoder $E(e)$ is modeled as,

$$\begin{aligned}\min_{e} E(e) &= \mathbb{E}_{(l_i,\hat{h}_i)\sim p(l,\hat{h})}\|\Psi(\hat{h}_i) - l_i\|_F^2 \\ &= \mathbb{E}_{(l_i,\hat{h}_i)\sim p(l,\hat{h})}\|\hat{l}_i - l_i\|_F^2,\end{aligned} \tag{6}$$

where e is the parameters of the transformative encoder, and $\Psi(\hat{h}_i)$ represents the mapping from the intermediate upsampled HR faces $\hat{h}_i$ to the projected LR faces $\hat{l}_i$. Similar to Eqn. 1, we also use RMSprop to update e by the stochastic gradient descent.

To obtain the final HR faces, we integrate a second decoder DEC$_2$ to super-resolve the projected LR face images. DEC$_2$, as shown in Fig. 4, is trained on the encoded LR and aligned ground-truth HR image pairs $\{\hat{l}_i, h_i\}$.

After training the encoder network, we use the encoder ENC to generate the training dataset $\hat{l}_i$, and then train DEC$_2$ by using the image pairs $\{\hat{l}_i, h_i\}$. The training procedure of DEC$_2$ is as the same as §. 3.3.1.

3.4. Hallucinating HR from Unaligned & Noisy LR

The discriminative network is only employed in training our decoders. When hallucinating HR faces, the discriminative work is not used. In the testing phase, we first feed an unaligned and noisy LR face l_i^n into the decoder DEC$_1$ to obtain an upsampled intermediate HR image $\hat{h}_i$. Then, we use our encoder ENC to project the intermediate HR face $\hat{h}_i$ to an aligned LR face $\hat{l}_i$. Finally, we use the decoder DEC$_2$ to super-resolve the aligned LR face $\hat{l}_i$ and attain our final hallucinated face $\tilde{h}_i$.

Since in the training phase we use upright HR faces as targets, our TDAE not only super-resolves the LR faces but also aligns HR face images simultaneously. Although we need to train our network in three steps, it can hallucinate an unaligned and noisy LR face to an upright HR version in an end-to-end fashion.

3.5. Implementation Details

The STN layers, as shown in Fig. 2 and Fig. 3, are built by convolutional and ReLU layers (Conv+ReLU), max-pooling layers with a stride 2 (MP2) and fully connected layers (FC). Specifically, STN$_1$ layer is built by cascading the layers: MP2, Conv+ReLU (filter size: 512×20×5×5), MP2, Conv+ReLU (20×20×5×5), FC+ReLU (from 400 to 20 dimensions) and FC (from 20 to 4 dimensions). STN$_2$ is constructed by cascading the layers: MP2, Conv+ReLU (256×128×5×5), MP2, Conv+ReLU (128×20×5×5), MP2, Conv+ReLU (20×20×3×3), FC+ReLU (from 180 to 20 dimensions) and FC (from 20 to 4 dimensions). STN$_3$ is constructed by cascading the layers: MP2, Conv+ReLU (128×20×5×5), MP2, Conv+ReLU (filter size: 20×20×5×5), MP2, FC+ReLU (from 80 to 20 dimensions) and FC (from 20 to 4 dimensions). STN$_4$ layer is built by cascading the layers: Conv+ReLU (96×20×5×5), MP2, Conv+ReLU (20×20×5×5), FC+ReLU (from 80 to 20 dimensions) and FC (from 20 to 4 dimensions). In the convolution operations, padding is not used.

In the following experimental part, some algorithms require the alignment of LR inputs [22, 25]. Thus, we employ STN$_0$ to align the LR images for those methods. The only difference between STN$_0$ and STN$_1$ is that the first MP2 step in STN$_1$ is removed in STN$_0$.

In training our decoders and encoder, we use the same learning rate r and decay rate γ. We set the learning rate r to 0.001 and multiply 0.99 after each epoch, and the decay rate is set to 0.01.

4. Experiments

We compare our method with the state-of-the-art methods qualitatively and quantitatively. We employ BM3D [4] to reduce the image noise, and then align the LR inputs by STN$_0$. In the experiments, we only show the upright HR ground-truth faces h_i for comparison purposes.

4.1. Dataset

We use the Celebrity Face Attributes (CelebA) dataset [30] to train our TDAE. There are more than 200K face images in this dataset, and the images cover different pose variations and facial expressions. We use these images without grouping them into different pose and facial expression subcategories.

When generating the LR and HR face pairs, we randomly select 30K cropped aligned face images from the CelebA dataset, and then resize them to 128×128 pixels as HR images. We use 28K images for training and 2K for our tests. We manually transform the HR images while constraining the faces to be visible in the in the image, downsample the HR images to generate LR images, and add Gaussian noise. In the training of the decoder DEC$_1$, we apply zero mean Gaussian noise with the standard deviation 10% of the maximum image intensity to the LR images.

4.2. Qualitative Comparison with the SoA

Since some super-resolution baselines [22, 25] require the input LR faces to be aligned, for a fair comparison we align the LR faces by STN$_0$ for the compared methods. We present only the aligned upright HR ground-truth faces for easy comparisons.

As shown in Fig. 6(c), conventional bicubic interpolation cannot generate facial details. Since the resolution of inputs is very small, little information is contained in the input images. Furthermore, the upsampled images also have some deformations. This indicates that aligning very LR images is more difficult when there is noise in the images.

Dong *et al.* [6] present a CNN based general purpose super-resolution method, also known as SRCNN. Since SR-

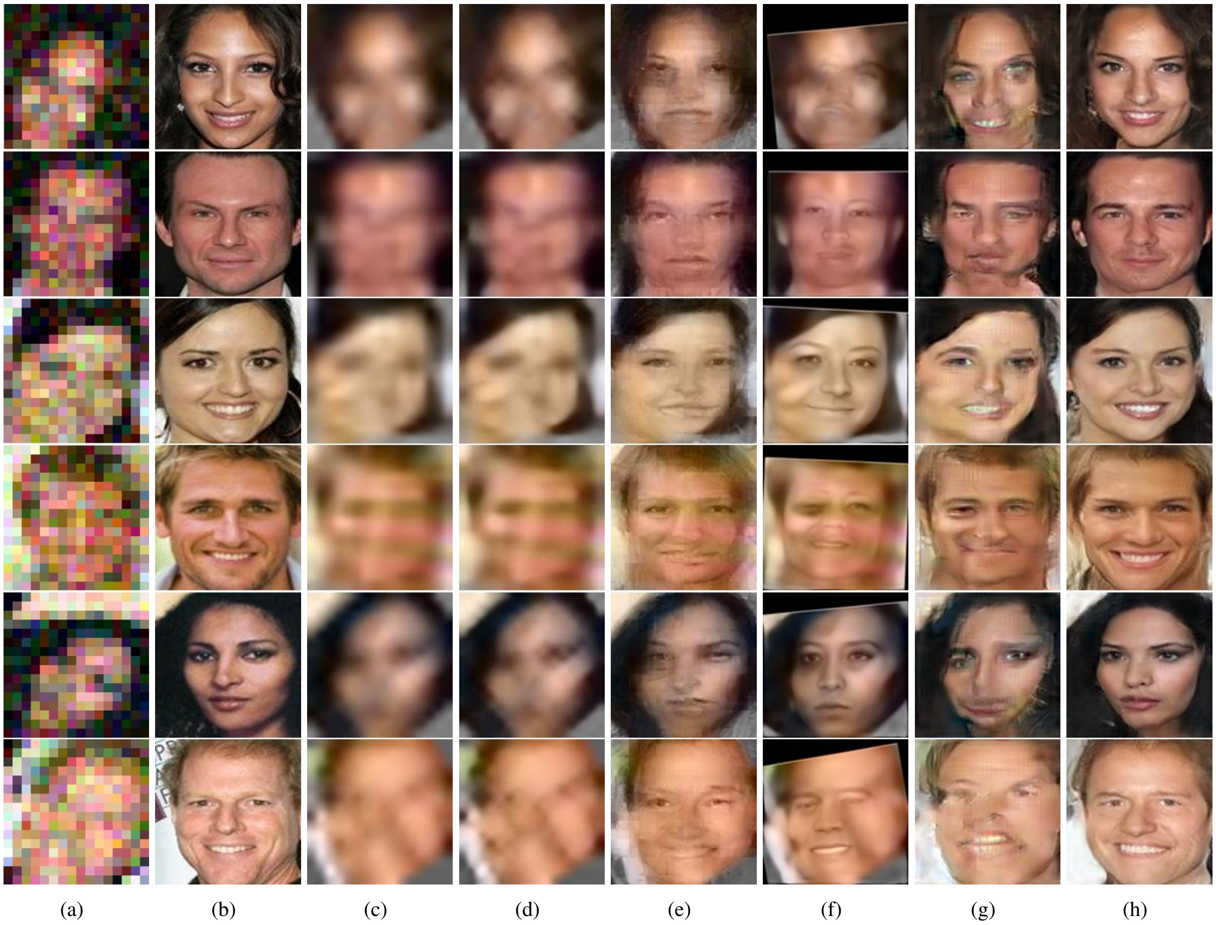

Figure 6. Comparison with the state-of-the-arts methods at the noise level 10%. (a) Unaligned and noisy LR inputs. (b) Original HR images. (c) Bicubic interpolation. (d) Results of [6]. (e) Results of [22]. (f) Results of [29]. (g) Results of [25]. (h) Our method.

CNN is patch based, it cannot capture the global face structure. Training SRCNN with the full face images introduces more ambiguity because the patch size (*i.e.*128×128) is too large to learn a valid non-linear mapping. Hence, we employ an upscaling factor of 8× to retrain it. As seen in Fig. 6(d), SRCNN cannot produce authentic facial details.

Ma *et al.* [22] exploit position patches to hallucinate HR faces. This method requires the LR inputs to be precisely aligned with the reference images in the training dataset. As visible in Fig. 6(e), when there are alignment errors, it produces deformed faces. Moreover, as the upscaling factor increases, the correspondences between LR and HR patches become inconsistent. Hence, it suffers from severe block artifacts around the boundaries of different patches.

Zhu *et al.* [29] propose a deep cascaded bi-network for face hallucination, known as CBN. This method has its own aligning process that localizes facial landmarks used to fit a global face model. When the noise level is low, it can align LR faces based on the landmarks. However, when the noise is not negligible, it fails to localize landmarks thus produces ghosting artifacts (see Fig. 6(f)). Since noise impedes the landmark detection, we apply BM3D as a remedy. However, LR faces becomes smooth, and detecting facial landmarks becomes even difficult. Our observation is that CBN is not designed for noisy images.

Yu and Porikli [25] developed a discriminative generative network to super-resolve very low resolution face images, known as URDGN. Their method also employs deconvolutional layers to upsample LR faces and a discriminative network is used to force the generate network to produce sharper results. However, this method requires aligned images and cannot super-resolve unaligned faces. In addition, noise may damage the LR facial patterns, which may degrade the performance as visible in Fig. 6(g).

In comparison, our method reconstructs authentic facial details as shown in Fig. 6(h). We note that the input faces have different poses and facial expressions. Since our method applies multiple STNs on feature maps to align face images and remove noise simultaneously, it achieves much better alignment. With the help of the encoder, it obtains aligned and noise-free LR images. With its second decoder, it produces visually pleasing results, which are similar to the

Table 1. Quantitative evaluations on the entire test dataset. Different configurations: (1) STN+SR+BM3D, (2) STN+BM3D+SR, (3) BM3D+STN+SR. Here, SR is the compared super-resolution method. Our method does not use BM3D or a separate STN.

		PSNR		SSIM	
	Noise	5%	10%	5%	10%
1	Bicubic	17.93	17.77	0.51	0.49
	SRCNN [6]	17.77	17.53	0.51	0.48
	Ma [22]	17.98	17.90	0.51	0.50
	CBN [29]	17.16	16.93	0.47	0.44
	URDGN [25]	16.58	16.45	0.38	0.36
2	Bicubic	18.59	18.30	0.52	0.51
	SRCNN [6]	18.59	18.32	0.53	0.51
	Ma [22]	18.63	18.37	0.50	0.49
	CBN [29]	18.34	18.26	0.52	0.52
	URDGN [25]	16.95	16.79	0.41	0.40
3	Bicubic	17.87	17.63	0.52	0.50
	SRCNN [6]	17.74	17.53	0.51	0.50
	Ma [22]	17.86	17.65	0.49	0.48
	CBN [29]	17.39	17.28	0.49	0.48
	URDGN [25]	18.95	18.65	0.49	0.47
	Ours	**21.02**	**20.47**	**0.58**	**0.56**

ground-truth faces as well. Our method does not need any landmark localization or any information about the noise. When the noise is low, it also attains superior performance.

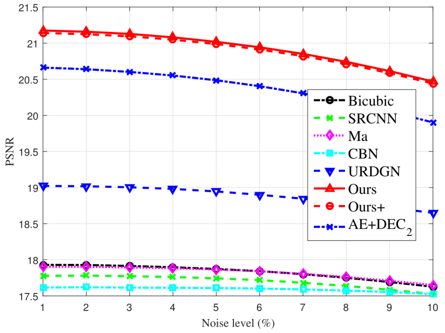

Figure 7. The PSNR curves of the state-of-the-art methods on synthetic test datasets with noise level from 1% to 10%.

4.3. Quantitative Comparison with the SoA

We quantitatively measure the performance of all methods on the entire test dataset in different noise levels by the average PSNR and the structural similarity (SSIM) scores. Table 1 presents that our method achieves superior performance in comparison to other methods, outperforming the second best with a large margin of **1.82** dB in PSNR.

For an objective comparison with the SoA methods, we report results for three possible scenarios. In the first case, we first apply STN_0 to align noisy LR faces, then super-resolve the aligned LR images by the SoA, and finally use BM3D to remove the noise in the upsampled HR images. In the second case, we apply STN_0 followed by BM3D and then super-resolution. In the third case, we first denoise by BM3D, then align by STN_0, and finally super-resolve. When aligning noisy LR images, we train STN_0 with noisy LR faces. Otherwise, if we first use BM3D to reduce noise, we train STN_0 with noise-reduced LR faces.

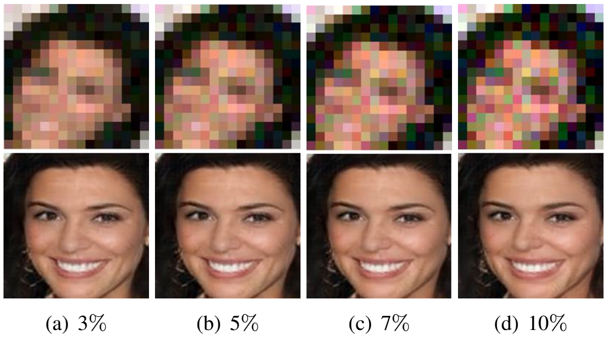
(a) 3% (b) 5% (c) 7% (d) 10%

Figure 8. Visualization of our results for different noise levels. Please refer to Fig. 5(b) for the ground-truth HR image.

Table 1 also indicates that simply denoising and then aligning, or aligning and then denoising LR faces cannot lead to good performance by the SoA methods.

Furthermore, we demonstrate that our method can successfully hallucinate faces in different noise levels in Fig. 8. When the noise level increases, our hallucinated faces remain consistent and retain their visual quality, which implies that our method is robust to noise variations.

Figure 7 shows the PSNR curves for different noise levels. We observe that our method achieves higher PSNRs over the other methods, and for lower noise levels it performs even better. Furthermore, we apply Gaussian blur with $\sigma = 2.4$ to the spatially transformed HR images, downsample HR faces, and add noise to the LR images. As shown in Fig. 7, our network still performs well without obvious degradation (dashed red line). Note that, we do not need to know the noise level or re-train our network with blurred LR inputs. We also combine DEC1 and ENC together as another baseline, denoted as AE.

5. Conclusion

We presented a transformative autoencoder network to super-resolve very low-resolution (16×16 pixels) unaligned and noisy face images with a challenging upsampling factor of 8×. We leverage on a new decoder-encoder-decoder architecture. Our networks jointly align, remove noise, and discriminatively hallucinate input images. Since our method is agnostic to image noise, face pose, and spatial deformations, it is very practical. At the same time, it can generate rich and authentic facial details.

References

[1] S. Baker and T. Kanade. Hallucinating faces. In *Proceedings of 4th IEEE International Conference on Automatic Face and Gesture Recognition, FG 2000*, pages 83–88, 2000. 1

[2] S. Baker and T. Kanade. Limits on super-resolution and how to break them. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 24(9):1167–1183, 2002. 1, 2

[3] J. Bruna, P. Sprechmann, and Y. LeCun. Super-resolution with deep convolutional sufficient statistics. In *ICLR*, 2016. 2

[4] K. Dabov, A. Foi, V. Katkovnik, and K. Egiazarian. Image denoising by sparse 3-d transform-domain collaborative filtering. *IEEE Transactions on image processing*, 16(8):2080–2095, 2007. 1, 6

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

[6] C. Dong, C. C. Loy, and K. He. Image Super-Resolution Using Deep Convolutional Networks. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 38(2):295–307, 2016. 2, 6, 7, 8

[7] I. Goodfellow, J. Pouget-Abadie, and M. Mirza. Generative Adversarial Networks. In *Advances in Neural Information Processing Systems (NIPS)*, pages 2672—-2680, 2014. 5

[8] P. H. Hennings-Yeomans, S. Baker, and B. V. Kumar. 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)*, pages 1–8. IEEE, 2008. 1

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

[10] M. Jaderberg, K. Simonyan, A. Zisserman, et al. Spatial transformer networks. In *Advances in Neural Information Processing Systems*, pages 2017–2025, 2015. 1, 2, 3, 4

[11] J. Kim, J. K. Lee, and K. M. Lee. Accurate Image Super-Resolution Using Very Deep Convolutional Networks. *arXiv:1511.04587*, 2015. 2

[12] S. Kolouri and G. K. Rohde. Transport-based single frame super resolution of very low resolution face images. In *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015. 1, 2

[13] Y. Li, C. Cai, G. Qiu, and K. M. Lam. Face hallucination based on sparse local-pixel structure. *Pattern Recognition*, 47(3):1261–1270, 2014. 1, 2

[14] C. Liu, H. Shum, and C. Zhang. 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)*, volume 1, pages 192–198, 2001. 1

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

[16] C. Liu, J. Yuen, and A. Torralba. Sift flow: Dense correspondence across scenes and its applications. *IEEE transactions on pattern analysis and machine intelligence*, 33(5):978–994, 2011. 2

[17] A. Radford, L. Metz, and S. Chintala. Unsupervised Representation Learning with Deep Convolutional Generative Adversarial Networks. *arXiv:1511.06434*, pages 1–15, 2015. 5

[18] M. F. Tappen and C. Liu. A Bayesian Approach to Alignment-Based Image Hallucination. In *Proceedings of European Conference on Computer Vision (ECCV)*, volume 7578, pages 236–249, 2012. 1, 2

[19] N. Wang, D. Tao, X. Gao, X. Li, and J. Li. A comprehensive survey to face hallucination. *International Journal of Computer Vision*, 106(1):9–30, 2014. 1, 2

[20] X. Wang and X. Tang. Hallucinating face by eigen transformation. *IEEE Transactions on Systems, Man and Cybernetics Part C: Applications and Reviews*, 35(3):425–434, 2005. 1, 2

[21] Z. Wang, Y. Yang, Z. Wang, S. Chang, W. Han, J. Yang, and T. Huang. Self-tuned deep super resolution. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 1–8, 2015. 2

[22] C. Q. Xiang Ma, Junping Zhang. Hallucinating face by position-patch. *Pattern Recognition*, 43(6):2224–2236, 2010. 1, 2, 6, 7, 8

[23] C. Y. Yang, S. Liu, and M. H. Yang. Structured face hallucination. In *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1099–1106, 2013. 1, 2

[24] J. Yang, J. Wright, T. S. Huang, and Y. Ma. Image super-resolution via sparse representation. *IEEE transactions on image processing*, 19(11):2861–73, 2010. 1, 2

[25] X. Yu and F. Porikli. Ultra-resolving face images by discriminative generative networks. In *Proceedings of European Conference on Computer Vision (ECCV)*, pages 318–333, 2016. 1, 2, 3, 4, 5, 6, 7, 8

[26] M. D. Zeiler and R. Fergus. Visualizing and understanding convolutional networks. In *European Conference on Computer Vision*, pages 818–833, 2014. 3

[27] M. D. Zeiler, D. Krishnan, G. W. Taylor, and R. Fergus. Deconvolutional networks. In *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2528–2535, 2010. 3

[28] E. Zhou and H. Fan. Learning Face Hallucination in the Wild. In *Twenty-Ninth AAAI Conference on Artificial Intelligence*, pages 3871–3877, 2015. 1, 2

[29] S. Zhu, S. Liu, C. C. Loy, and X. Tang. Deep cascaded bi-network for face hallucination. In *Proceedings of European Conference on Computer Vision (ECCV)*, pages 614–630, 2016. 1, 2, 3, 5, 7, 8

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