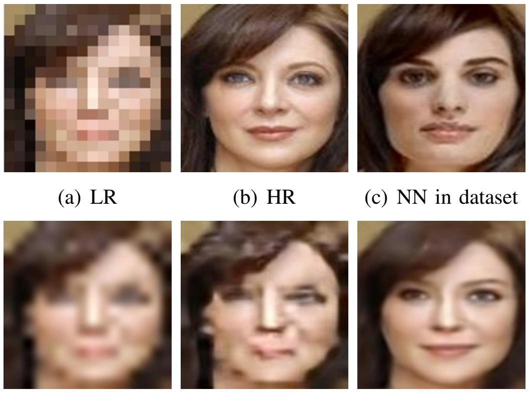
Imagining the Unimaginable Faces by Deconvolutional Networks

Xin Yu, Fatih Porikli, *Fellow, IEEE,*

***Abstract*—We tackle the challenge of constructing 64 pixels for each individual pixel of a thumbnail face image. We show that such an aggressive super-resolution objective can be attained by taking advantage of the global context and making the best use of the prior information portrayed by the image class. Our input image is so small (*e.g.*, 16×16 pixels) that it can be considered as a patch of itself. Thus, conventional patch-matching based super-resolution solutions are unsuitable. In order to enhance the resolution while enforcing the global context, we incorporate a pixel-wise appearance similarity objective into a deconvolutional neural network, which allows efficient learning of mappings between low-resolution input images and their high-resolution counterparts in the training dataset. Furthermore, the deconvolutional network blends the learned high-resolution constituent parts in an authentic manner where the face structure is naturally imposed and the global context is preserved. To account for the possible artifacts in upsampled feature maps, we employ a sub-network composed of additional convolutional layers. During training, we use roughly aligned images (only eye locations), yet demonstrate that our network has the capacity to super-resolve face images regardless of pose and facial expression variations. This significantly reduces the requirement of precisely face alignments in the dataset. Owing to the network topology we apply, our method is robust to translational misalignments. In addition, our method is able to upsample rotational unaligned faces with data augmentation. Our extensive experimental analysis manifests that our method achieves more appealing and superior results than the state-of-the-art.**

***Index Terms*—Face hallucination, deconvolutional neural network, super-resolution.**



(a) LR (b) HR (c) NN in dataset

(d) Bicubic (e) RED [10] (f) Ours

Fig. 1. Comparison of our method with the CNN based super-resolution. (a) The input 16 × 16 LR image. (b) The original 128 × 128 HR image. (c) The corresponding HR version of the nearest neighbor of (a) in the training set. (d) Bicubic interpolation of (a). (e) The image generated by the CNN based super-resolution [10]. Notice that, the CNN based approach is further *fine-tuned* with a large corpus of face images. (f) Our result.

I. Introduction

THE human face is perhaps the most powerful channel of nonverbal communication. It provides valuable clues to our own feelings and those of the people around us. Even in the most simple interaction, our attention naturally gravitates to the face, seeking to read some of the vital information is "written" there. Faces also play an important role in physical attractiveness.

Naturally, face perception is possible *if* the face is visible in *sufficient* detail and resolution. When the face image is imperceptibly small, its resolution has to be super-resolved with a large upscaling factor. However, conventional super-resolution (SR) methods are mostly limited up to $2 \sim 4\times$ upscaling factors. As reported in [1], when the upscaling factor increases to $8\times$, the performance of most SR techniques decreases rapidly, rendering them unsuitable for this challenge.

Existing state-of-the-art SR methods highly rely on a variety of assumptions about the quality of the given low-resolution (LR) image and the availability of an associated set of high-resolution (HR) images. They are applicable only when (i) accurate facial features and landmarks can be found in LR images [2], [3], (ii) similar appearances of the "same" person are included in the reference HR dataset [4], and (iii) the exemplar HR face images are "densely" aligned in order to derive a representative subspace [5], [6], [7], [8], [9]. When the input image resolution is inadequately small, the performance of the face SR methods that require detection of precise landmarks for a dense alignment degrades dramatically due to the problematic localization of such refined features and landmark points. This is a consequence of the fact that there is little margin for error or flexibility when the LR image is tiny. Typical pose, facial expression and illumination differences between the input LR image and exemplary HR images hinder the ability of subspace-based face SR methods in capturing local variations and lead to unavoidable ghosting artifacts in the reconstructed HR images.

Several super-resolution methods based on deep neural networks have been proposed [11], [12], [13], [14], [15], [16], [10] recently. However, these methods are all patch based and ignore image class information. As shown in Fig. 1(e), the Convolutional Neural Network (CNN) based network [10], even when it has been retrained with face images, fails to produce authentic facial details.

When super-resolving an LR image with an $8\times$ upscaling factor, 98.5% of the original information is missing. Hallucinating such a significant chunk of missing information is an

Xin Yu and Fatih Porikli are with the Research School of Engineering, Australian National University, Canberra, Australia, e-mail: (xin.yu@anu.edu.au, fatih.porikli@anu.edu.au).

ill-posed problem without a competent regularization term and efficient exploitation of strong priors.

As a solution, we exploit a variant of deconvolutional neural networks [17] to learn the mappings between the LR facial patterns and HR facial details across individual samples while maintaining the underlying global structure of face images by taking advantage of the collective representation power of large-scale face datasets [18], [19]. Deconvolutional layers, also known as backwards-convolutional layers, are convolutional layers where the forward and backward passes are reversed. In other words, for a stride larger than 1, the output of such a deconvolutional layer has larger resolution than its inputs. They are first utilized in [20], [17] to visualize the features a CNN has learned by back-projecting activations in the low-dimensional feature maps to the high-dimensional image domain. Rather than projecting feature activations to the image domain, Long *et al.* [21] use a deconvolutional network to upsample heat maps while Fischer *et al.* [22] upsample optical flow fields. However, the upsampling results of these methods tend to be over-smoothed without pronounced high-frequency details. To enhance image details, Shi *et al.* [23] present a variant of deconvolutional networks that rearranges multiple LR feature maps into an HR image as its output. These deconvolutional networks do not formulate the super-resolution task on class-specific settings; hence, they fail to model and generate valuable class-specific cues. Furthermore, since our deconvolutional layers are not used for back-projecting activations of feature maps, our method does not require unpooling layers for super-resolution.

Our intuition is that, deconvolutional networks can be trained to generate certain HR image patterns given specific LR activations by presenting the network with a set of well-structured LR-HR image pairs. Such well-structured data conveniently exists for the face class. Our analysis in section III-C demonstrates that deconvolutional networks can be trained to recognize particular facial patterns.

In the training stage of our deconvolutional neural network, we feed the *entire* images, *i.e.*, not patches but whole faces, into our network. This allows maintaining the global structure of faces while reconstructing instance specific details. As a result, our deconvolutional network produces realistic HR facial components that seamlessly blend into an HR face image. Since the filters in each layer of our deconvolutional neural network architecture are applied to the entire image, our method achieves robustness to spatial translations and deformations of input faces. For training, we use approximately frontal HR face images that are only aligned at eye locations, which is readily available for most face datasets. We do not make any assumption on facial landmarks and facial expressions.

Overall, our contributions are fourfold:

- We present a novel method to super-resolve with an $8\times$ upscaling factor a very small (16×16 pixels) face image.
- Our method consolidates a deconvolutional network for hallucinating face images. We demonstrate that without using an adversarial loss, our network is still able to super-resolve realistic HR face images and achieves an impressive 1.16 dB PSNR improvement over the state-of-the-art.
- Since only convolution operations are used in our network, our method is not sensitive to translational misalignments, which significantly reduces the accuracy requirement of the face localization in the LR image. This means, even when the face detector response may not be accurate since the face region is very small, our network can still super-resolve it.
- When training our network, we only require approximately frontal and roughly aligned images regardless of pose and facial expression variations, which makes the training datasets more attainable.

II. Related Work

Image super-resolution methods aim to magnify an LR image to its HR version that comprises authentic high-frequency details. In general, there are three categories of *generic* super-resolution approaches: interpolation based techniques, image statistics based schemes [24], [25] and example/patch based methods [26], [27], [28], [8], [29], [30]. Interpolation based techniques such as bilinear and bicubic upsampling are computationally efficient. However, they fail to establish high-frequency details since they generate overly smooth edges as the upscaling factor increases. Image statistics based schemes employ image priors to reconstruct HR images with sharper edges, but they are still limited to smaller scaling factors [31].

Example based methods have the potential to break this limitation. They can be further classified into two groups: internal and external example methods depending on how the reference samples are derived. The first group of methods [28], [32], [33], [30] exploit self-similarity of patches in the input image. Alternatively, several methods [26], [27], [8] aim to learn mappings between LR and HR patches from external reference datasets, and then utilize the learned correspondences to upsample LR images. Nevertheless, when the input image size is very small, it is difficult for internal example based methods to find similar patches across different scales. When the scaling factor is large, it is hard for external example based methods to determine the correct correspondences between LR and HR patches because many different HR patches can correspond to a single LR patch, which induces artifacts at intensity edges.

Recently, many generic super-resolution methods based on deep neural networks have been proposed [11], [12], [14], [15], [16], [10], [23], [34]. For instance, SRCNN [11] applies cascaded convolutional layers to obtain a mapping function between LR and HR patches from a large-scale dataset, while Kim *et al.* [15] learn to upsample the residuals between the HR and interpolated LR patches. To improve the performance of super-resolution without introducing extra parameters of the networks, Kim *et al.* [16] employ recursive convolutional layers to increase the depth of the convolutional layers. Mao *et al.* [10] apply symmetric-skip connections between convolutional layers and deconvolutional layers to pass information to the latter layers, thus mitigating the difficulty of training their very deep network. Shi *et al.* [23] employ convolutional layers to extract LR features and then rearrange the LR feature maps into HR images by a sub-pixel convolutional layer, which can be considered as a variant of deconvolutional layers. Dong *et al.* [12] use convolutional and deconvolutional layers with smaller filter sizes to speed up SRCNN [11]. Ledig *et al.* [34] exploit an adversarial loss and a perceptual loss [35] to obtain more realistic upsampled results. Bruna *et al.* [14] extracts statistical priors using CNN to regularize the super-resolution process. Since these generic SR methods based on neural networks do not consider class-specific priors, they cannot achieve high performance when they are employed for super-resolving faces. Retraining (fine-tuning) of these networks with face image patches cannot capture the global structure of faces either.

Related to face hallucination, the work in generative adversarial networks (GAN) [36], [37] and variational auto-encoders [38] exploit neural networks to generate an entirely new image that endows similar properties to the training data distribution, from a random noise input.

Unlike generic SR methods, *class-specific* super-resolution approaches, such as face hallucination [39], [40], [41], [5], [6], [7], [42], [4], [2], [43], [44], [9], [45], [3], [46], explore the underlying patterns of a certain class, thus leading to better performance. Baker and Kanade [39] transfer high-frequency details from a face dataset by building the relationships between LR and HR patches. Due to the possible inconsistency of the transferred HR patches, their method tends to produce artifacts. Eigen-transformation is employed to hallucinate face images by establishing a mapping between the LR and HR face subspaces in [5]. Similarly, Liu *et al.* [6] employ a subspace that is learned from the training set via Principle Component Analysis (PCA) as a linear constraint for HR face images and proposes a patch-based Markov Random Field (MRF) to reconstruct the missing high-frequency details. Kolouri *et al.* [9] use optimal transport in combination with subspace learning to morph an HR image from the LR input. Since the subspace based face hallucination methods require the HR images in the reference dataset to be precisely aligned and the LR test image to have the same pose and facial expression as the reference ones, they are overly sensitive to the misalignments of LR images. In particular, methods that depend on PCA based holistic appearance models suffer from ghosting artifacts.

Rather than imposing global constraints, Ma *et al.* [42] construct a super-resolved HR patch by multiple reference HR patches at the corresponding spatial position. Li *et al.* [47] model the local structures of faces as a sparse representation problem. Jin and Bouganis [45] process multiple LR face images to recover an HR image by exploiting a patch-wise mixture of probabilistic PCA prior instead of the holistic PCA prior in [6]. Hence, face hallucination methods that constrain the spatial positions of patches may avoid ghosting artifacts caused by PCA, but their performance degrades dramatically when LR image is not aligned precisely to the reference HR images. To handle various poses and expressions, [4] integrates the SIFT flow to align images. By exploiting local patterns, Yang *et al.* [2] present a structured face hallucination method. It first detects facial components in the given LR image and then transfers the corresponding HR facial components in the reference dataset to the LR input. Zhu *et al.* [3] present a deep bi-network to super-resolve LR faces. It uses a CNN to localize facial components and then recovers the high-frequency of the localized facial components by another CNN. Nevertheless, these facial component based methods may fail to produce authentic HR face images due to potentially inaccurate landmark localization. Zhou and Fan [43] propose a bi-channel CNN to hallucinate face images in wild scenes. Since they require extraction of local features from the input images, the smallest input image size is limited to 48×48 pixels. Yu and Porikli [46] extend the framework of GAN for very low-resolution face super-resolution. Their follow-up works [48], [49] employ an adversarial loss to distinguish whether super-resolved HR faces are realistic, and use spatial transformer networks (STN) [50] in their deconvolutional networks to compensate for misalignments. When LR face images are aligned and in low noise levels, [49] super-resolves face images similar to the results of [46] because they employ similar architectures for upsampling. Due to the sensitive training procedure of GAN, artifacts may appear in the HR outputs; as a result, their high-frequency details may be inconsistent with the ground-truth data.

III. Our Face Super-Resolution Network

As shown in Fig. 2, our complete network consists of two parts: an upsampling part (deconvolutional), and an image enhancement part (convolutional).

In the upsampling part, we employ deconvolutional layers, as our upsampling part, to super-resolve the LR face images as well as we exploit convolutional layers, as our enhancement part, to remove the blocking artifacts caused by the deconvolutional layers [51]. We utilize the ℓ_2 regression loss, also known as the Euclidean distance loss, as the objective of the entire network to attain appearance similarity between the reconstructed images and the original HR images in the training stage.

We first feed the input LR images into a convolutional layer to extract low-level patterns (features). Since the resolution of input images is very small, *i.e.*, 16×16, the filter size is set to 3×3. The reason for applying a convolutional layer to LR inputs is to mitigate the artifacts introduced by the following deconvolutional layers. As reported in [51], a direct application of deconvolutional layers to input images may lead to severe blocking artifacts due to the overlapping regions between the receptive fields. We exploit deconvolutional layers to upsample feature maps, in which most of the activations are close to zero, and thus the artifacts can be mitigated. After the feature extraction, three deconvolutional layers are employed to upsample the feature maps. Each layer upsamples the previous feature maps by an upscaling factor of 2. Since upsampling images is an under-determined problem, we intend to increase the capacity of the network as the neural network goes deeper, *i.e.*, the resolutions of feature maps become larger. Hereby, we double the channel numbers of feature maps of previous layers. The filter sizes of these three deconvolutional layers are $3\times3\times64$, $5\times5\times128$, and $5\times5\times256$, respectively. We apply batch normalization [52] after each deconvolutional layer to accelerate the convergence behavior of the network.

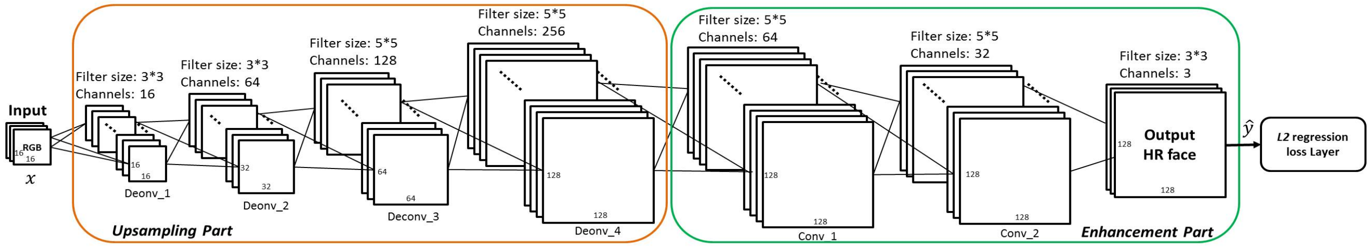


Fig. 2. Our deconvolutional network consists of two parts: an upsampling part (the orange block) and an enhancement part (the green block).

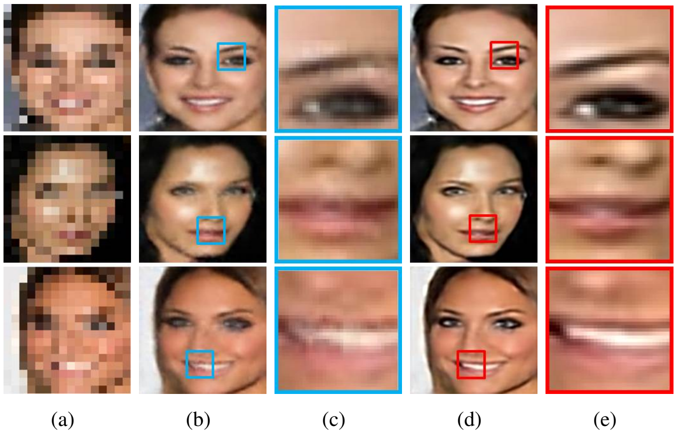


Fig. 3. Blocking artifacts caused by the deconvolutional layers are effectively removed by the enhancement part. (a) LR input images. (b) Results upsampled only by the deconvolutional layers (the upsampling part). (c) The close-ups of (b). (d) Results upsampled by the entire network. (e) The close-ups of (d).

Since deconvolutional layers introduce aliasing artifacts in the output images, we incorporate convolutional layers as a subsequent enhancement subnetwork to remove such artifacts. We use three convolutional layers with the filter sizes of 5×5×64, 5×5×32 and 3×3×3 in the enhancement part. We note that Dong *et al.* [53] indicate adding more convolutional layers does not suppress artifacts (in their case compression artifacts) but makes the training convergence of the network more difficult. This phenomenon also appears in SRCNN [11], where they show that using more than three layers does not provide a significant improvement in the super-resolution performance. Moreover, a larger network cannot be fed into the GPU memory, either. Hence, we employ a three convolutional layers network to remove aliasing artifacts rather than using a deeper enhancement network.

To illustrate the effectiveness of the two parts of our network, we present the outputs of each part separately in Fig. 3. For visualization of the images that are super-resolved only by the upsampling part, we switch the output channel of the last deconvolutional layer to 3 and remove the enhancement part from the entire network. To retrain the upsampling part, we employ the ℓ_2 regression loss between the upsampled images and the HR ground-truth as the object function. As shown in Fig. 3(b), the upsampling part generates HR facial details, but the results suffer from the blocking and aliasing artifacts. As shown in Fig. 3(d), the artifacts are significantly suppressed, and the facial details are sharpened by the image enhancement part when we train the entire network comprised of the upsampling and enhancement parts. Additionally, the output of the entire network obtains almost 1.3 dB PSNR improvement over the output of the upsampling part on the test dataset. Notice that, the upsampling part produces a total of 256 feature maps.

A. *Training of the Entire Network*

We use LR-HR face image pairs $\{x_i, y_i\}$ as our training data. Since the output of the entire network $\hat{y}_i$ is imposed to be similar to the corresponding HR image y_i, a pixel-wise ℓ_2 regularization term is integrated to induce similarity. The loss E of the complete network for a mini-batch of N face image pairs becomes

$$E = \frac{1}{ACN} \sum_{i=1}^{N} \|\Phi(x_i) - y_i\|_2^2, \tag{1}$$

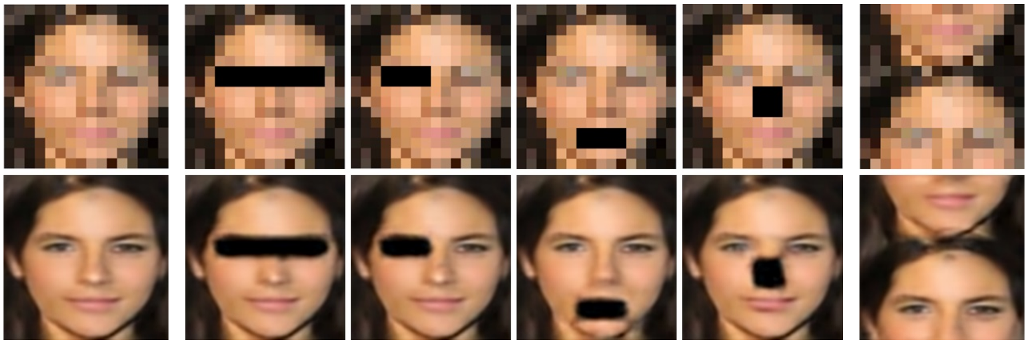
where $\Phi(x_i) = \hat{y}_i$ denotes the output of the entire network. Here, A and C represent the area and the number of the channels of the training HR images.

The loss E in Eqn. 1 is back-propagated to update the parameters of the complete network. Since each layer of our network is differentiable, RMSprop [54] is used for back-propagation. In RMSprop, we set the learning rate to 10^{-3} and the decay rate α to 0.9. In addition, the learning rate η is multiplied by 0.99 after each epoch.

B. *Super-Resolution of an LR Face Image*

We input the LR image x into our network to construct its upsampled HR image $\hat{y}$. In our previous work [46], we used a discriminative network to enforce the final results to be similar to typical face images, yet that discriminative network has potential to inject ringing artifacts in the final results. To improve the overall visual quality, we also apply an unsharp filtering [55] to the upsampled HR results, which is an image enhancement technique and widely used in low-level image processing tasks, such as super-resolution [56] and deblurring [57]. Specifically, unsharp filtering is used to generate a sharp image by adding an difference image, which is obtained from subtracting an image a blurred version of itself, to the original version. In this way, we preserve the visual fidelity while avoiding the artifacts introduced by the discriminative network.

Since only convolutional operations are used in the network, our end-to-end mapping can maintain the global structure of

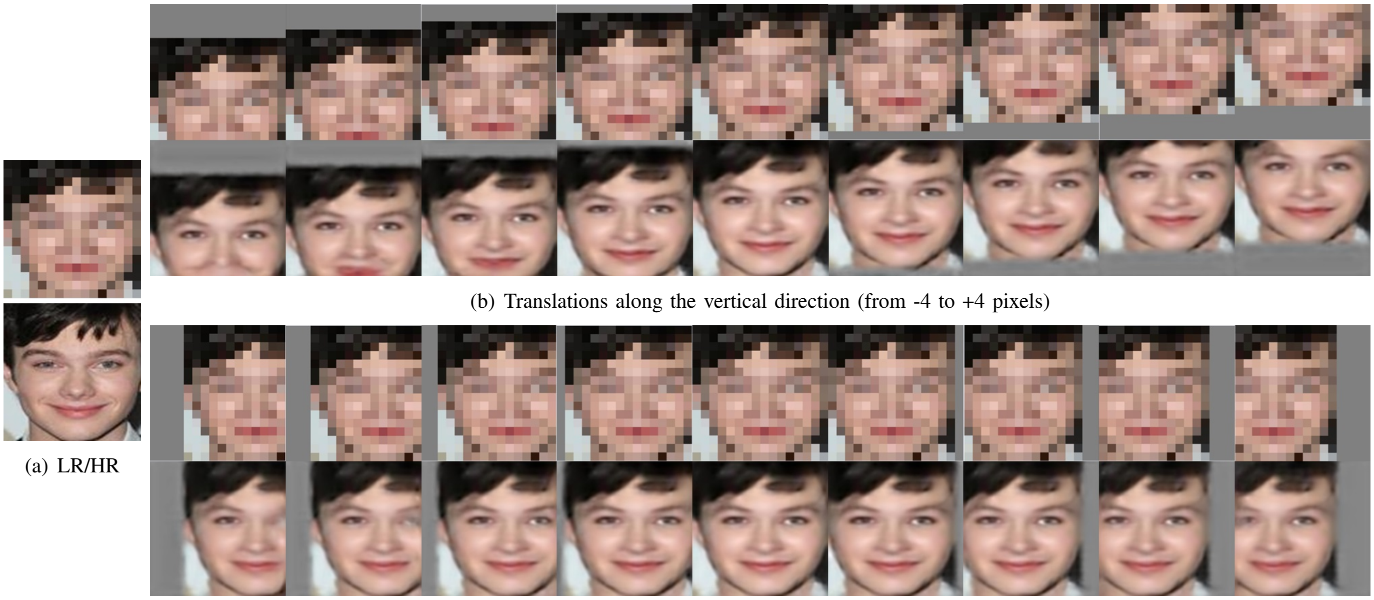


(a) LR,HR

(b) Our results with occlusions

(c) Alteration

Fig. 4. Illustrations of influence of occlusions. Top row: the LR images, bottom row: the results of our deconvolutional network. (a) Result without occlusions. (b) Results for partially occluded input images. (c) Result when the upper-lower parts are altered.



(a) LR/HR

(b) Translations along the vertical direction (from -4 to +4 pixels)

(c) Translations along the horizontal direction (from -4 to +4 pixels)

Fig. 5. Our method is robust against the translational misalignments of the LR image.

HR face images while infusing rich and localized details. It is also robust to translational misalignments of LR images. As illustrated in Fig. 5, our method can accurately reconstruct the corresponding HR face images even if the LR images are shifted in horizontal and vertical directions.

Thanks to its feed-forward architecture, our method runs in real-time on GPU when it super-resolves an LR image.

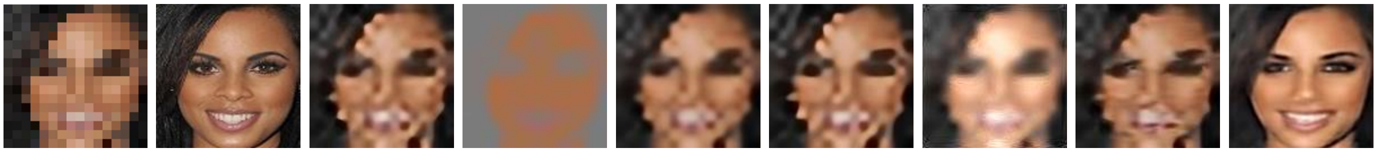
C. What does the Deconvolutional Network Learn?

In our deconvolutional network, the hallucination of the entire face and the formation of individual facial components are implemented seamlessly. To dissect what our deconvolutional network learns, we apply a set of masks to occlude different parts and facial components of the input image. Our assumption here is that a holistic face model based neural network can still generate a complete face without missing parts, even if the reconstructions of the originally occluded parts may be not realistic. Otherwise, it is more likely that the network learns face components.

Figure 4 suggests that our deconvolutional network learns facial components and their relative local arrangements. Figure 4(b) shows that the visible parts of the input images are super-resolved well while the masked parts are not recovered. Even when we switch the upper and lower parts of the face as shown in Fig. 4(c), which does not look like a face, the corresponding parts can be super-resolved by our network. As presented in Fig. 5, our network can reconstruct the translated versions of the HR face images consistent with the LR face images when the input face undergoes large translations. This also indicates that our network learns the facial components rather than a rigid holistic face model, and generates HR facial components given specific LR facial patterns.

D. Differences Between Our Network and CNN based Nets

One major difference between our network and CNN based super-resolution networks, such as SRCNN [11] and RED [10], lies on the network architecture. Our method employs deconvolutional layers for upsampling LR face images,



(a) LR (b) Original (c) SRCNN [11] (d) SRCNN$_{wf}$ (e) SRCNN$_{p}$ (f) RED [10] (g) RED$_{wf}$ (h) RED$_{p}$ (i) Ours

Fig. 6. Comparison with *fine-tuned* SRCNNs [11] and REDs [10]. (a) The LR image. (b) The original HR image. (c) Result of the original SRCNN applying an upscaling factor of 2× three times. (d) Result of the SRCNN fine-tuned and retrained with whole face images. (e) Result of the SRCNN retrained with patches with an upscaling factor of 8×. (f) Result of the original RED applying an upscaling factor of 2× three times. (g) Result of the RED fine-tuned and retrained with whole face images. (h) Result of the RED retrained with patches with an upscaling factor of 8×. (i) Our result.

while CNN based super-resolution networks apply convolutional layers. For instance, [11] and [10] firstly upsample the input LR patches by bicubic interpolation and then use convolutional layers to enhance the corresponding details of the interpolated LR patches. Since the corresponding receptive fields of the filters in the HR images are just the same as the filter sizes, only local information is incorporated in the generated high-frequency details. As shown in Fig. 6(c), when SRCNN is directly applied to the face hallucination task, the output HR face image is severely blurred due to the small size of the input image and the large upscaling factor. The same phenomenon for the RED can be seen in Fig. 6(f) as well.

Another difference is that generic super-resolution methods [11], [10] are patch based while our method uses the entire image. Since SRCNN released its training code, we can compare its variants more objectively. To achieve the most objective comparison, we not only assess the performance of the original SRCNN but also its possible adaptations for face hallucination. The original SRCNN does not provide a direct upscaling factor of 8× but requires 2× upsampling of the input image three times. When sequentially upsampling, facial components that appear in different scales cannot be learned by the original SRCNN. Hence, we first retrain SRCNN with face patches with an upscaling factor 8×. We use the same architecture and hyperparameters of SRCNN and retrain the network by using face patches with the scaling factor 8×. As shown in Fig. 6(e), SRCNN cannot produce an HR face image with authentic high-frequency details. Because the scaling factor is large, the interpolated LR images are too smooth for SRCNN to manage. In other words, local neighbors provide little information in enhancing the details. Moreover, when retraining SRCNN with entire face images, the large size of training patches, *i.e.*, 128×128 pixels, introduces more ambiguity in learning of the parameters, compared with the 33×33 pixels patch size the original SRCNN employs. During the training, the weights of SRCNN gets stuck into erroneous local minima and decrease to zero, thus produce a zero-valued image. As shown in Fig. 6(d), the SRCNN retrained with entire face images fails to provide high-quality HR face images.

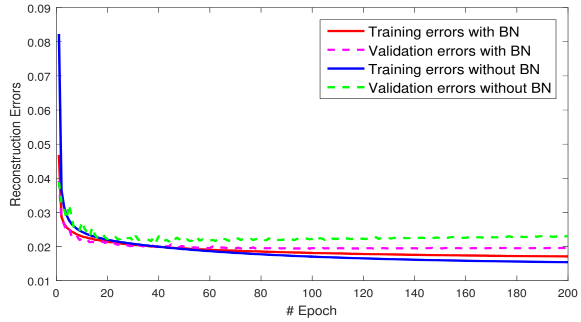


Fig. 7. Comparisons of the training and validation errors with and without using batch normalization.

One factor that affects the super-resolution performance is the depth of neural networks. Since SRCNN only has 3 convolutional layers, its performance may be limited. We also compare with another CNN based method, RED, which consists of 15 convolutional layers and 15 deconvolutional layers, much deeper than our network and trained on image patches of size 50×50 pixels. Note that, the deconvolutional layers employed in RED are different from our deconvolution layers; the deconvolutional layers in RED only implement backward convolutional operations without increasing the output resolutions. To tackle the vanishing gradient problem and obtain an efficient training scheme, RED passes information from the convolutional layers to their corresponding deconvolutional layers by exploiting skip connections. Similar to SRCNN, RED firstly upsamples inputs by bicubic interpolation and then enhances details. As shown in Fig. 6(f), directly applying RED to the LR face by an upscaling factor 2× three times cannot achieve realistic facial details, *e.g.*, the LR eye regions only consist of dark colors. It only enhances edges and textures rather than generating semantically new pixels, such as the white color in the eyeballs. As presented in Fig. 6(h), retraining RED with face patches by an upscaling factor 8× cannot obtain authentic facial details since the large upscaling factor introduces severe ambiguity between LR and HR patches. We also retrain RED with the whole face images as well as the same training protocol that we use. As seen in Fig. 6(g), RED fails to generate realistic facial details; instead, it outputs ringing artifacts. Hence, simply increasing the depth of convolutional networks cannot super-resolve LR faces either.

In contrast to SRCNN and RED, our deconvolutional network upsamples the LR face images gradually without any bicubic interpolation. This strategy can be regarded as leveraging the image pyramid to address the under-determined task of 8× super-resolution. In a hierarchical manner, we hallucinate facial details, thus mitigating the ambiguity between LR and HR face images. In contrast, bicubic interpolation employed in CNN based super-resolution methods cannot reduce the ambiguity between the interpolated LR and HR faces since it only relies on upsampling of pixels without any hallucination.

Furthermore, the receptive field of the filters of our first deconvolutional layer is 24×24 pixels in the HR images, which is much larger than the largest receptive field of the filters in SRCNN, *i.e.*, 9×9 pixels. As a result, our network can better capture LR facial patterns, and it can access expanded spatial neighborhood to generate HR faces. Our deconvolutional layers are able to project the low-dimensional feature maps to the high-dimensional image domain and the learned feature patterns are embedded in the weights of the network. Hence, our deconvolutional network is more suitable to construct a mapping from LR face images to their HR versions.

Generic CNN based super-resolution methods, such as SRCNN and RED, do not incorporate batch normalization either. Batch normalization is originally invented to reduce internal covariate shift by whitening feature maps and widely used for classification tasks. Since batch normalization will change the intensity distributions of feature maps in each layer, it may distort the mapping relationships between LR and HR patches in the super-resolution problem. Specifically, generic CNN based SR methods construct a nonlinear mapping between different LR and HR patches on image intensities. Considering the intensity distributions of different patches may vary dramatically, the distributions of their corresponding feature maps in each layer would be significantly different because image patches are not normalized when they are fed into super-resolution networks. Thus, the mean and variance for each layer vary in a mini-batch. Using a statistical mean and variance to normalize the feature maps in each layer will shift activations of input patches. This effect will increase the ambiguity in super-resolution, thus increasing the training loss. As a result, the intensity of the reconstructed HR patches would be distorted. This phenomenon that embedding batch normalization into the CNN based super-resolution, *e.g.*, SRCNN and VDSR [15], degrades the super-resolution performance is also observed in the very recent works [58], [59]. Therefore, it is not suitable to use batch normalization in generic patch based super-resolution convolutional networks.

Since our inputs are class-specific, the feature maps share similar distributions in each layer. Using batch normalization allows speeding up the training phase without shifting the reconstructed faces in our network. In Fig. 7, we compare the training errors with and without using batch normalization. As seen in the first 50 epochs, our network achieves lower training and validation errors by using batch normalization. It indicates that batch normalization speeds up the learning process of our network. Even though after 50 epochs the training errors of the network without using batch normalization become lower than the one using batch normalization, their validation errors stop decreasing and the validation errors of the network without using batch normalization are higher than the one using batch normalization. It implies that batch normalization facilitates the generalization ability of our network.

IV. Experimental Analysis

We compare our method with a large set of eleven state-of-the-art methods [6], [8], [2], [11], [42], [15], [16], [10], [45], [3], [46] both qualitatively and quantitatively. Liu *et al.* [6] employ a subspace based face hallucination method. Yang *et al.* [8] use sparse representations to super-resolve HR images by constructing LR and HR dictionaries. The method in [2] hallucinates face images by using facial components from an exemplar image dataset while CBN [3] super-resolves facial components by deep cascaded bi-networks. SRCNN [11], VDSR [15], DRCN [16], and RED [10] apply CNNs to upsample images. Ma *et al.* [42] use same position reference patches to reconstruct HR images. Jin and Bouganis [45] exploit multiple LR faces to recover an HR version by a patch-wise mixture of probabilistic PCA prior (MPPCA).

A. Datasets

Our network is trained on the Celebrity Face Attributes (CelebA) dataset [19]. There are more than 200K face images in this dataset where only the similarity transformation is employed to align the locations of eye centers [19]. The images cover different pose variations and facial expressions. We simply use all available data regardless of these variations and do not require grouping the face images into different pose and facial expression subcategories.

We randomly select 30K cropped face images from the CelebA dataset, and then resize them to 128×128 pixels as HR images. We downsample the HR face images to 16×16 pixels to obtain the LR counterparts. We use 29K images for the training, 1K images for validation and 1K images for testing.

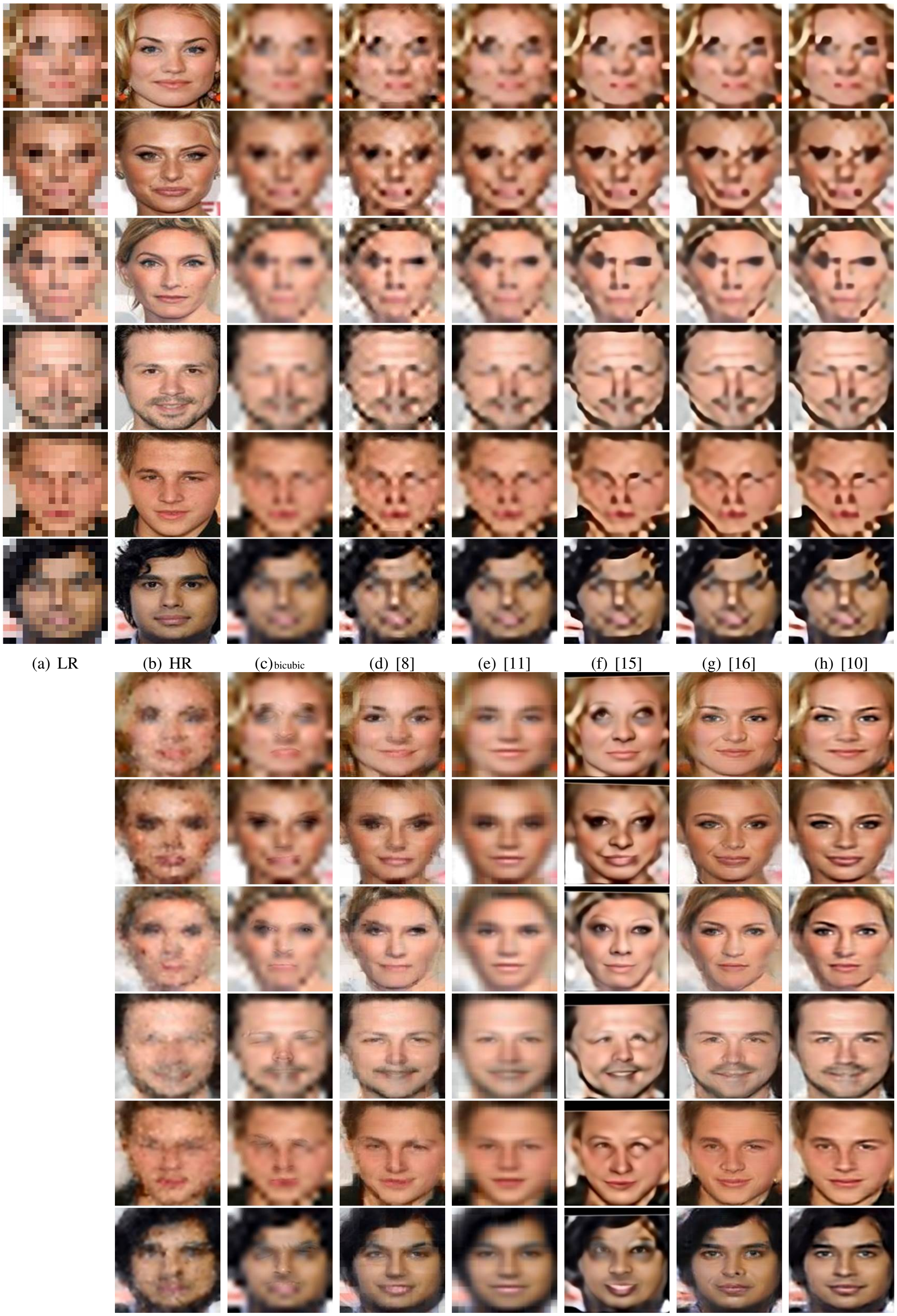
Our network never sees the test LR images in the training phase. The test and training images are substantially different. To illustrate this, we find the best matching LR image in the training data for a random input test LR image. As shown in Fig. 1, the corresponding HR version of the best match has significant differences from the original HR version of the LR test image. (All the protocol details, data, and code for this paper will be released.)

B. Qualitative Comparisons

We perform side-by-side comparisons with eleven state-of-the-art face hallucination methods. In case an approach does not allow an 8× scaling factor directly, *e.g.*, [8], [11], [15], [16], [10], we repeatedly apply a scaling factor of 2× three times. For fair comparisons, we use the same CelebA dataset for the training of all other algorithms. As another baseline, we present the bicubic interpolation results.

Comparison with Yang *et al.*'s method [8]: As depicted in Fig. 8(d), Fig. 9(d), Fig. 10(d) and Fig. 11(d), Yang *et al.*'s method does not recover high-frequency facial details. Besides, irregular over-emphasized edge artifacts appear in the results. As the scaling factor increases, the correspondence between LR and HR patches becomes ambiguous. Therefore, the results suffer from exaggerated pixelation patterns.

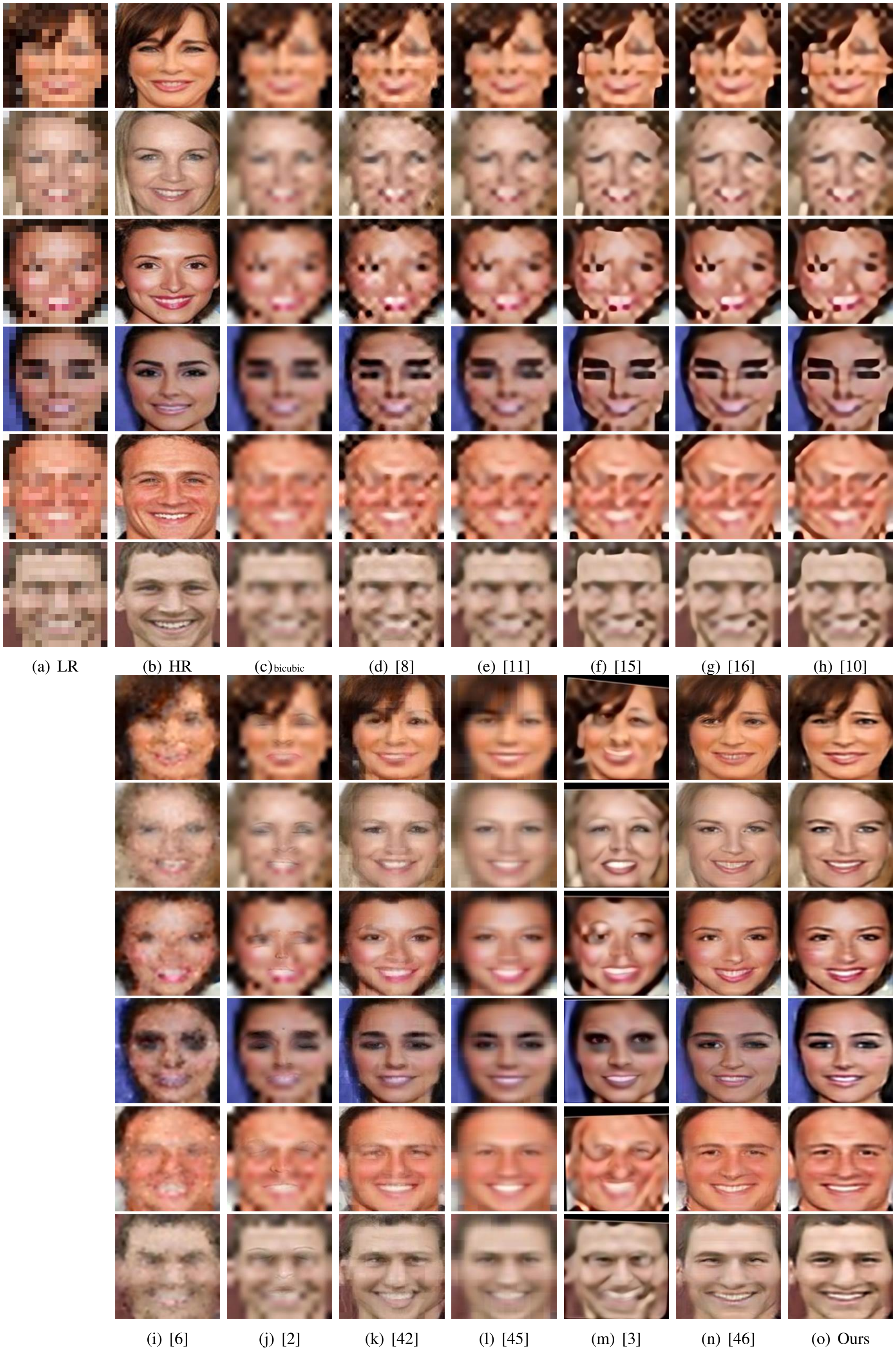
Comparison with Dong *et al.*'s method [11]: SRCNN applies convolutional layers to learn a generic patch-based mapping function. Even though we retrain their CNN on face images, SRCNN cannot generate high-frequency facial details in the HR images as shown in Fig. 8(e), Fig. 9(e), Fig. 10(e) and Fig. 11(e). This demonstrates that our deconvolutional



(a) LR (b) HR (c)bicubic (d) [8] (e) [11] (f) [15] (g) [16] (h) [10]

(i) [6] (j) [2] (k) [42] (l) [45] (m) [3] (n) [46] (o) Ours

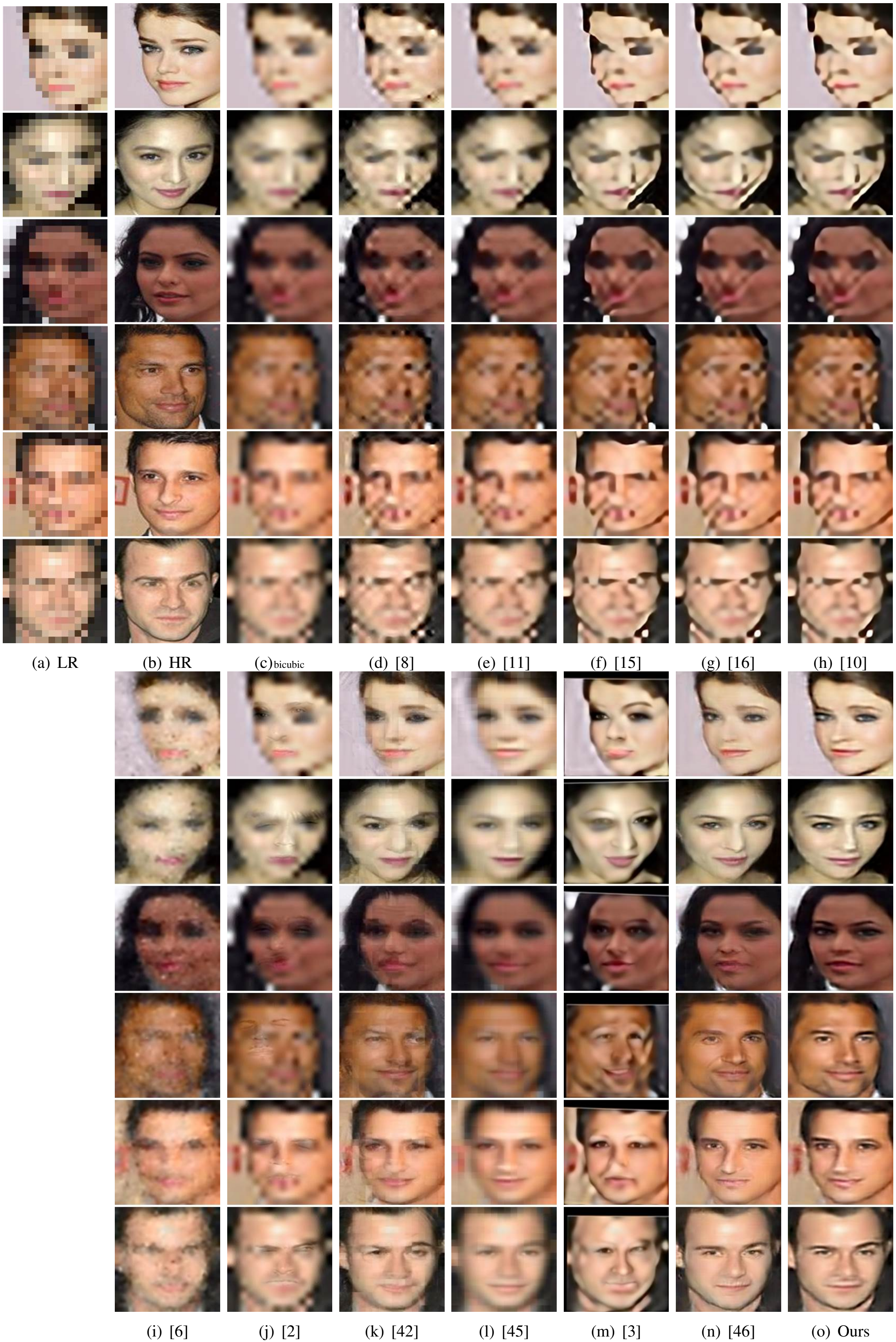
Fig. 8. Comparison with the state-of-the-art on **frontal** face images. (a) LR inputs. (b) Original HR images. (c) Bicubic interpolation. (d) Yang *et al.*'s method [8]. (e) Dong *et al.*ś method (SRCNN) [11]. (f) Kim *et al.*'s method (VDSR) [15]. (g) Kim *et al.*'s method (DRCN) [16]. (h) Mao *et al.*'s method (RED) [10]. (i) Liu *et al.*'s method [6]. (j) Yang *et al.*'s method [2]. (k) Ma *et al.*'s method [42]. (l) Jin and Bouganis's method (MPPCA) [45]. (m) Zhu *et al.*'s method (CBN) [3]. (n) Yu and Porikli's method (URDGN) [46]. (o) Our method. (Please see the electronic version for fine-grained details)



(a) LR (b) HR (c)bicubic (d) [8] (e) [11] (f) [15] (g) [16] (h) [10]

(i) [6] (j) [2] (k) [42] (l) [45] (m) [3] (n) [46] (o) Ours

Fig. 9. Comparison with the state-of-the-art on images with **facial expressions**. (a) LR inputs. (b) Original HR images. (c) Bicubic interpolation. (d) Yang *et al.*'s method [8]. (e) Dong *et al.*ś method (SRCNN) [11]. (f) Kim *et al.*'s method (VDSR) [15]. (g) Kim *et al.*'s method (DRCN) [16]. (h) Mao *et al.*'s method (RED) [10]. (i) Liu *et al.*'s method [6]. (j) Yang *et al.*'s method [2]. (k) Ma *et al.*'s method [42]. (l) Jin and Bouganis's method (MPPCA) [45]. (m) Zhu *et al.*'s method (CBN) [3]. (n) Yu and Porikli's method (URDGN) [46]. (o) Our method.



(a) LR (b) HR (c)bicubic (d) [8] (e) [11] (f) [15] (g) [16] (h) [10]

(i) [6] (j) [2] (k) [42] (l) [45] (m) [3] (n) [46] (o) Ours

Fig. 10. Comparison with the state-of-the-art on **different pose** face images. (a) LR inputs. (b) Original HR images. (c) Bicubic interpolation. (d) Yang *et al.*'s method [8]. (e) Dong *et al.*ś method (SRCNN) [11]. (f) Kim *et al.*'s method (VDSR) [15]. (g) Kim *et al.*'s method (DRCN) [16]. (h) Mao *et al.*'s method (RED) [10]. (i) Liu *et al.*'s method [6]. (j) Yang *et al.*'s method [2]. (k) Ma *et al.*'s method [42]. (l) Jin and Bouganis's method (MPPCA) [45]. (m) Zhu *et al.*'s method (CBN) [3]. (n) Yu and Porikli's method (URDGN) [46]. (o) Our method.

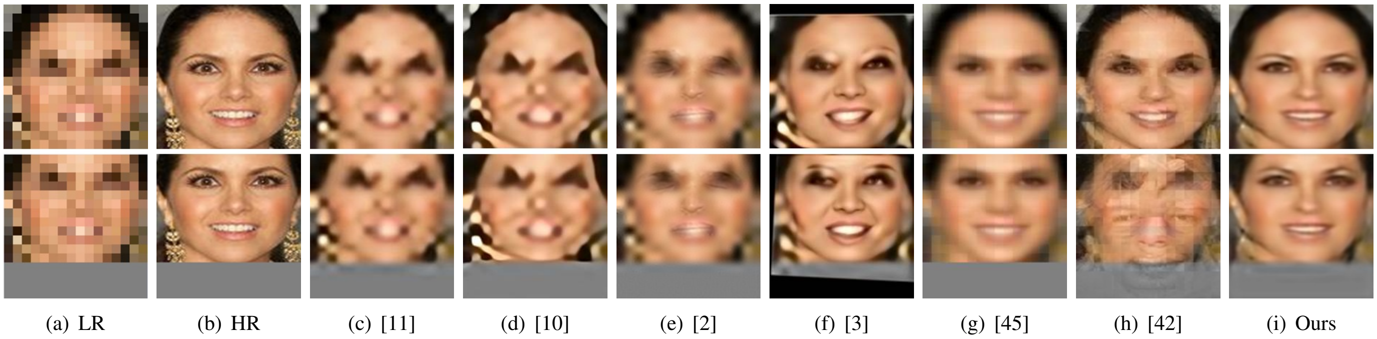


Fig. 11. Comparison with the state-of-the-art on translational **misaligned** face images. (a) LR inputs. (b) Original HR images. (c) Dong *et al.*'s method (SRCNN) [11]. (d) Mao *et al.*'s method (RED) [10]. (e) Yang *et al.*'s method [2]. (f) Zhu *et al.*'s method (CBN) [3]. (g) Jin and Bouganis's method (MPPCA) [45]. (h) Ma *et al.*'s method [42]. (i) Our method.

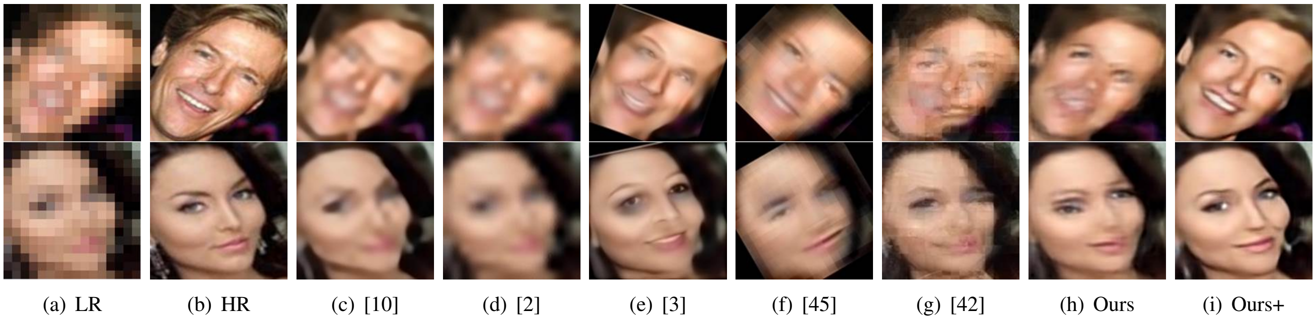


Fig. 12. Comparison with the state-of-the-art on rotational **misaligned** face images. (a) LR inputs. (b) Original HR images. (c) Mao *et al.*'s method (RED) [10]. (d) Yang *et al.*'s method [2]. (e) Zhu *et al.*'s method (CBN) [3]. (f) Jin and Bouganis's method (MPPCA) [45]. (g) Ma *et al.*'s method [42]. (h) Our method. (i) Our method with rotated face augmentation.

network is more suitable to address the face hallucination task. In contrast to SRCNN, our deconvolutional network incorporates class-specific information to induce fine-grained patterns authentic to faces, thus leads to better performance.

Comparison with Kim *et al.*'s method [15]: Kim *et al.* propose a very deep convolutional network for generic image super-resolution, known as VDSR, where they increase the number of the convolutional layers to 20 while SRCNN uses only 3. To accelerate the training of its network, VDSR learns the high-frequency residuals between the upsampled input patches and their HR ground truths instead of producing HR patches directly. Similar to SRCNN, VDSR also firstly upsamples LR input patches by bicubic interpolation and then reconstructs high-frequency details by a deep CNN. As shown in Fig. 8(f), Fig. 9(f) and Fig. 10(f), VDSR fails to output realistic facial details and over-enhances edges of the upsampled LR facial patterns. This also indicates that just increasing the depth of traditional convolutional networks may not necessarily generate authentic facial details.

Comparison with Kim *et al.*'s method [16]: Kim *et al.* develop a deeply recursive convolutional network (DRCN) to super-resolve generic images. DRCN employs 16 recursive convolutional layers followed by ReLU layers to increase the super-resolution performance without introducing extra parameters. Similar to VDSR, the high-frequency residuals are learned from the neural network. As shown in Fig. 8(g), Fig. 9(g) and Fig. 10(g), DRCN over-emphasizes edges and cannot hallucinate authentic high-frequency facial textures, *i.e.*, eyes and mouths. In contrast, our network can reconstruct realistic facial details.

Comparison with Mao *et al.*'s method [10]: Mao *et al.* employ a very deep residual encoder-decoder network to upsample images, named as RED, which has 15 convolutional and 15 deconvolutional layers to recover the missing high-frequency contents in LR patches. Different from our deconvolutional layers, the deconvolutional layers in RED do not increase the resolution of feature maps. RED is a patch-based generic super-resolution method, and it is trained with generic image patches. As shown in Fig. 8(h), Fig. 9(h), Fig. 10(h), Fig. 11(d) and Fig. 12(c), RED cannot produce authentic HR face images either. Hence, we conclude that directly upsampling LR inputs by bicubic interpolation and then generating image details from the interpolated images by CNNs is not suitable for the face hallucination task.

Comparison with Liu *et al.*'s method [6]: Since Liu *et al.*'s method requires the face images in the dataset to be precisely aligned, it is difficult for their method to learn a representative subspace from the CelebA dataset where large variations exist. Therefore, the global model of the input LR image cannot be represented by the learned subspace, and its local model leads to patchy artifacts in the results. As shown in Fig. 8(i), Fig. 9(i) and Fig. 10(i), this method cannot recover face details correctly, and noisy artifacts appear in the final results.

Comparison with Yang *et al.*'s method [2]: This method

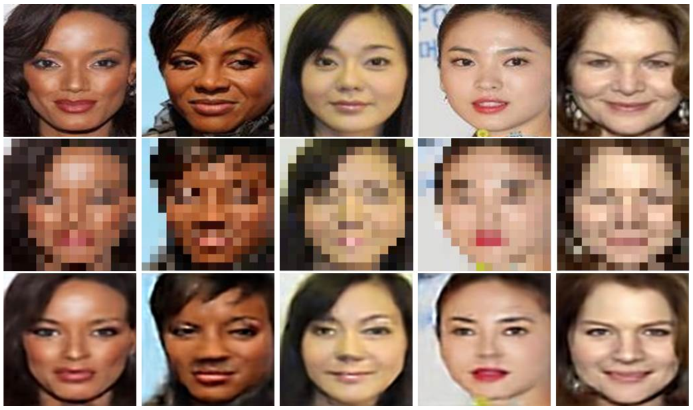


Fig. 13. Our method can hallucinate face images regardless of the racial profiles of the input images. Top row: the original HR face images. Middle row: the input LR face images. Bottom row: our results.

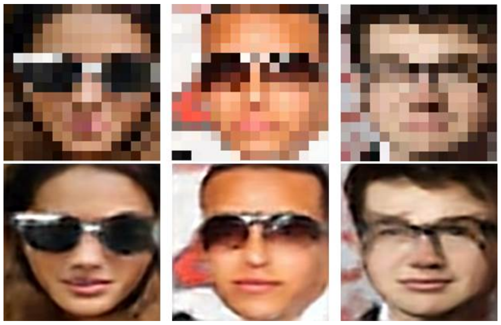
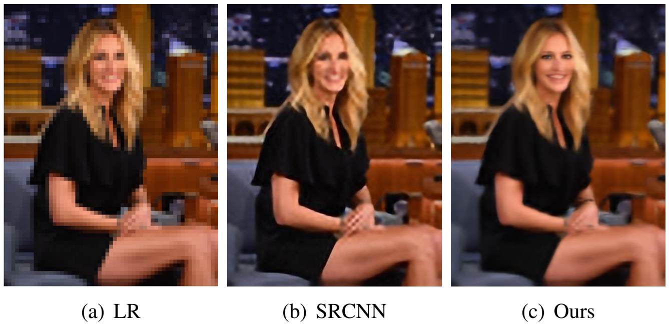


Fig. 14. Hallucinating face images with eyeglasses. Top row: the input LR face images. Bottom row: our results.



(a) LR (b) SRCNN (c) Ours

Fig. 15. Hallucinating face images without detecting and cropping faces. (a) The input LR image. (b) The result of SRCNN. (c) Our result. Note that the face region upsampled by our method contains much richer high-frequency details, such as the eyes and mouth. (please see electronic version for details)

requires landmarks of facial components. It reconstructs LR images by transferring high-resolution facial components. In a 16 × 16 input image, it is extremely difficult to localize landmarks. Hence, this method cannot correctly transfer facial components as shown in Fig. 8(j), Fig. 9(j), Fig. 10(j) and Fig. 12(d). Moreover, as seen in Fig. 11(e), facial details cannot be recovered either due to the very large upscaling factor. To our advantage, our method does not need landmark localization and still preserves the global structure of the faces.

Comparison with Ma *et al.*'s method [42]: This method requires the reference images to be precisely aligned. As shown in Fig. 8(k), Fig. 9(k) and Fig. 10(k), it suffers from obvious blocking artifacts and uneven over-smoothing as a result of unaligned reference patches in the training dataset and the large scaling factor. As illustrated in Fig. 11(h) and Fig. 12(g), this method mixes the magnified input face with a reference positioned ghost face due to translational and rotational misalignments. Our method, on the other hand, can still upsample the misaligned LR face images with rich high-frequency details.

Comparison with the method of Jin and Bouganis [45]: Instead of generating a holistic face model by PCA, this method, also known as MPPCA, super-resolves each patch of an LR face by exploiting a prior of the mixture probabilistic principal component analysis [60]. MPPCA uses multiple LR images to recover an HR face. As reported in their experimental part, MPPCA utilizes multiple LR images synthesized from a single HR image to evaluate its performance. Hence, following its experimental protocol, we also generate multiple LR faces from an HR ground-truth image and then apply MPPCA to reconstruct the HR face. Because MPPCA needs to estimate the motion transformations between LR images, any error in transformation parameter estimation causes reconstruction errors. To prevent from this, we use the ground-truth motion transformation parameters to align LR images in our experiments. Since each pixel of the LR inputs corresponds to an MPPCA model and the upscaling factor is large, *i.e.*, 8×, inconsistency may appear along the boundaries of generated HR patches. As seen in Fig. 8(l), Fig. 9(l), Fig. 10(l), Fig. 11(g) and Fig. 12(f), MPPCA suffers visible blocking artifacts and produces overly smooth HR faces due to the large upscaling factor.

Comparison with Zhu *et al.*'s method [3]: Zhu *et al.*'s method, called as CBN, first detects the facial components and then applies a deep neural network to super-resolve facial components. Since the resolution of the input faces is very small, it is difficult to detect and localize facial components accurately. Such errors directly lead to ghosting artifacts. As illustrated in Fig. 8(m), Fig. 9(m) and Fig. 10(m), CBN fails to output authentic HR faces when erroneous localization of the LR facial components occurs. As shown in Fig. 11(f) and Fig. 12(e), the upsampled facial details are inconsistent with the LR faces. CBN firstly aligns the LR inputs to its predefined coordinates and then generates high-frequency details. When we transform the hallucinated faces back onto the original coordinates, the black regions appear in the final results.

Comparison with the method of Yu and Porikli [46]: Yu and Porikli's method, also known as URDGN, exploits the framework of the generative adversarial network [36] to super-resolve HR faces. Its discriminator network enforces the generated HR face images to be similar to the real ones, but it may also introduce artifacts and thus distorts the hallucinated facial details. As shown in Fig. 8(n), Fig. 9(n) and Fig. 10(n), although the results of URDGN are sharp, the high-frequency details may not comply with the HR ground-truth as indicated in the quantitative evaluation. In contrast, our method can recover facial details more faithfully to the ground-truth faces. Note that the artifacts caused by deconvolutional layers as well as the adversarial loss are not suppressed by URDGN while they are significantly reduced by our convolutional layers. Furthermore, URDGN employs the procedure of generative adversarial networks (GAN) to train its entire network, and it is difficult to maintain the balance between the generative and discriminative networks. Thus, the convergence of URDGN is not as stable as our method.

C. Quantitative Comparisons

We also measure the performance by the average PSNR and the structural similarity (SSIM) scores on the entire test dataset. Table I shows that our method achieves the best performance with an impressive 1.16 dB PSNR improvement. In Tab. I, we also compare the PSNR and SSIM scores without using batch normalization, as indicated by Ours^{-}. Benefiting from batch normalization, our method is able to achieve higher PSNR and SSIM scores.

Notice that, the bicubic interpolation explicitly builds on pixel-wise intensities without any hallucination, and attains better performance than several state-of-the-art methods. This implies that either the high-frequency details reconstructed by the state-of-the-art methods are *not* authentic or the artifacts caused by those methods severely degrade their quantitative results.

Unlike the existing approaches, our method consistently provides visually appealing super-resolved HR face images that contain rich details, and at the same time, exhibit close similarity to the original ones (not used in the training). Since our method takes the input LR image as a whole and learns facial components in a data-driven manner, it reduces the ambiguity of the correspondence between LR and HR patches, leading to superior results both qualitatively and quantitatively.

D. Sensitivity to Translational Misalignments

Since the low-resolution of the input face images is very small, state-of-the-art face detectors may not localize the face precisely. In particular, when the translational alignments occur, the previous face hallucination methods may fail as seen in Fig. 11. By contrast, our method is able to upsample the LR face images without any degradation. In our method the translational alignment requirement is significantly relaxed. Even when the face detector fails to localize LR faces accurately, our method can still upsample the face images that have the similar sizes as the faces in the training dataset.

E. Sensitivity to Rotational Misalignments

As shown in Fig. 5 and Fig. 11, our method significantly reduces the requirement of face alignment, in particular, it can tolerate the translational misalignments of LR face images. Having said that, our network is trained with only upright face images; thus its performance would decrease when LR face images undergo large rotations as shown in Fig. 12(h). The rotated facial parts are not explicitly learned in the training stage. Therefore, our network may not recognize the corresponding low-dimensional features. As a result, we crop the HR faces from CelebA, randomly rotate HR faces and then downsample the HR faces to 16×16 pixels as LR faces. We augment our training and testing datasets with the rotated faces and then retrain our network on the augmented dataset. Notice that, the ground-truth images may not be upright due to the data augmentation. As shown in Fig. 12(i), our method can super-resolve LR faces with rotational misalignments as well.

F. Face Super-Resolution without a Face Detector

In Fig. 15, we present an example where the face region in the LR image is directly super-resolved without a face detector, *i.e.*, the face region is not detected and cropped before it is applied to our network. As visible, the face region is restored with sufficient and pleasant high-frequency facial details while the background regions are also upsampled without artifacts. Our method can efficiently remove the blocking artifacts along the edges in the background. In comparison, the CNN based super-resolution not only fails to generate authentic facial features such as mouth and eyes but also injects faulty checkerboard patterns (around fingers, hair, etc.) and overemphasized edges (around the black dress).

This example demonstrates that our deconvolutional network allows generating high-frequency details for faces without creating artifacts in the generic regions. Our method can recognize and super-resolve the LR facial features regardless of the locations of the features. We can upsample the LR faces without using a face detector when the LR faces approximately have the size of 16×16 pixels while the existing face hallucination methods rely on face detectors to crop faces in advance.

G. Different Racial Profiles

When training our deconvolutional network, we do not partition the training face images into different training sets based on their racial profiles. Instead, we use all available face images. We observe that our network can still conceive the shared characteristics of each race and upsample LR input images without requiring different models for different races. In other words, our method does not need a face attribute for the input image. As shown in Fig. 13, our method can super-resolve while maintaining the original racial profiles without mixing different racial characteristics.

H. Glasses

There are three cases around the super-resolution of faces with eyeglasses. The first one is the people wearing sunglasses, as shown in the first column of Fig. 14. In this case, eyes are occluded by the sunglasses. Obviously, the eyes cannot be super-resolved while the other facial parts including the sunglasses can be well reconstructed. The second case is that the frames of eyeglasses are thin and invisible in the

TABLE I
QUANTITATIVE EVALUATION ON THE ENTIRE TEST DATASET

Methods	Bicubic	[8]	[11]	[15]	[16]	[10]	[6]	[2]	[42]	[45]	[3]	[46]	Ours$^-$	Ours
PSNR	23.15	21.29	22.25	20.17	20.75	20.11	21.54	23.05	23.09	22.96	20.27	23.88	24.39	**25.04**
SSIM	0.67	0.60	0.65	0.58	0.60	0.58	0.55	0.66	0.64	0.64	0.58	0.71	0.72	**0.74**

small LR images. Since the eyeglasses are not visible, they cannot be reconstructed in the HR outputs, *i.e.*, the eyeglasses will not affect the face super-resolution. Lastly, the frames of eyeglasses might be thick enough to be hinted in the small LR images. Since the resolution of the LR image is only $16{\times}16$ pixels, the pixels corresponding to the eyeglasses frames are blended with the pixels of the eyes (the last column of Fig. 14). This may introduce some degradation around the upsampled eyes yet the rest of the face is well hallucinated. Since the styles and colors of eyeglass frames vary remarkably, using a proportionally larger dataset of annotated training images with eyeglasses can provide a remedy. However, this may not be practical.

I. Training Dataset Bias

In the CelebA dataset, the most common facial expression is the smile, which constitutes 48.2 percent of all faces in the dataset. This is the reason that in Fig. 9 most of the samples have smiling expressions. Although there are other expressions in the dataset, they do not exist in sufficient numbers to train a deep neural network. Given enough training samples, our deconvolutional network can be devised to hallucinate any facial expressions.

J. Limitations

Since our deconvolutional network hallucinates facial parts from a very low-resolution face image and then assembles them in an authentic manner into an HR face image, our method does not generate a complete face image when some regions are occluded in the LR input image. Nevertheless, as shown in Fig. 4(b), such occlusions do not degrade the super-resolution performance of the visible parts.

V. Conclusion

We presented an effective method to super-resolve very small LR face images by exploiting deconvolutional neural networks. Our method increases the input LR image size significantly, *i.e.*, 8×, and reconstructs rich facial details. Since it learns an end-to-end mapping between the LR and HR face images and uses only convolutional operations, it preserves the global structure of faces while mitigating the alignment requirements of LR inputs. Due to the simple feed-forward network architecture, our method runs in real-time.

Acknowledgment

The authors would like to thank the anonymous reviewers and the Associate Editor for their helpful suggestions and valuable comments. This work was supported under the Australian Research Council‘s Discovery Projects funding scheme (project DP150104645).

References

[1] C.-Y. Yang, C. Ma, and M.-H. Yang, “Single-image super-resolution: A benchmark,” in *Proceedings of European Conference on Computer Vision (ECCV)*, 2014.

[2] 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)*, 2013, pp. 1099–1106.

[3] S. Zhu, S. Liu, C. C. Loy, and X. Tang, “Deep cascaded bi-network for face hallucination,” in *European Conference on Computer Vision (ECCV)*, 2016, pp. 614–630.

[4] M. F. Tappen and C. Liu, “A Bayesian Approach to Alignment-Based Image Hallucination,” in *Proceedings of European Conference on Computer Vision (ECCV)*, vol. 7578, no. c, 2012, pp. 236–249.

[5] X. Wang and X. Tang, “Hallucinating face by eigen transformation,” *IEEE Transactions on Systems, Man and Cybernetics Part C: Applications and Reviews*, vol. 35, no. 3, pp. 425–434, 2005.

[6] C. Liu, H. Y. Shum, and W. T. Freeman, “Face hallucination: Theory and practice,” *International Journal of Computer Vision*, vol. 75, no. 1, pp. 115–134, 2007.

[7] K. Jia and S. Gong, “Generalized face super-resolution,” *IEEE Transactions on Image Processing*, vol. 17, no. 6, pp. 873–886, 2008.

[8] J. Yang, J. Wright, T. S. Huang, and Y. Ma, “Image super-resolution via sparse representation.” *IEEE transactions on image processing*, vol. 19, no. 11, pp. 2861–73, 2010.

[9] 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.

[10] X. Mao, C. Shen, and Y.-B. Yang, “Image restoration using very deep convolutional encoder-decoder networks with symmetric skip connections,” in *Advances in Neural Information Processing Systems (NIPS)*, 2016, pp. 2802–2810.

[11] C. Dong, C. C. Loy, and K. He, “Image Super-Resolution Using Deep Convolutional Networks,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 38, no. 2, pp. 295–307, 2016.

[12] C. Dong, C. C. Loy, and X. Tang, “Accelerating the super-resolution convolutional neural network,” in *European Conference on Computer Vision (ECCV)*, 2016, pp. 391–407.

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

[14] J. Bruna, P. Sprechmann, and Y. LeCun, “Super-resolution with deep convolutional sufficient statistics,” in *ICLR*, 2016.

[15] J. Kim, J. Kwon Lee, and K. Mu Lee, “Accurate image super-resolution using very deep convolutional networks,” in *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016, pp. 1646–1654.

[16] ——, “Deeply-recursive convolutional network for image super-resolution,” in *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016, pp. 1637–1645.

[17] M. D. Zeiler and R. Fergus, “Visualizing and understanding convolutional networks,” in *European Conference on Computer Vision (ECCV)*, 2014, pp. 818–833.

[18] G. B. Huang, M. Ramesh, T. Berg, and E. Learned-Miller, “Labeled faces in the wild: A database for studying face recognition in unconstrained environments,” University of Massachusetts, Amherst, Tech. Rep. 07-49, 2007.

[19] 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.

[20] 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)*, 2010, pp. 2528–2535.

[21] J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentation," in *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015.

[22] P. Fischer, A. Dosovitskiy, E. Ilg, P. Häusser, C. Hazırbaş, V. Golkov, P. van der Smagt, D. Cremers, and T. Brox, "Flownet: Learning optical flow with convolutional networks," in *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2015, pp. 2758–2766.

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

[24] T. Peleg and M. Elad, "A statistical prediction model based on sparse representations for single image super-resolution." *IEEE Transactions on Image Processing*, vol. 23, no. 6, pp. 2569–2582, 2014.

[25] C.-Y. Yang and M.-H. Yang, "Fast Direct Super-Resolution by Simple Functions," in *Proceedings of IEEE International Conference on Computer Vision (ICCV)*, 2013, pp. 561–568.

[26] W. T. Freeman, T. R. Jones, and E. C. Pasztor, "Example-based super-resolution," *IEEE Computer Graphics and Applications*, vol. 22, no. 2, pp. 56–65, 2002.

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

[28] D. Glasner, S. Bagon, and M. Irani, "Super-Resolution from a Single Image," in *Proceedings of IEEE International Conference on Computer Vision (ICCV)*, 2009, pp. 349–356.

[29] S. Schulter and C. Leistner, "Fast and Accurate Image Upscaling with Super-Resolution Forests," in *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015, pp. 3791–3799.

[30] J.-B. Huang, A. Singh, and N. Ahuja, "Single image super-resolution from transformed self-exemplars," in *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015, pp. 5197–5206.

[31] Z. Lin and H. Y. Shum, "Response to the comments on "Fundamental limits of reconstruction-based superresolution algorithms under local translation"," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 28, no. 5, pp. 83–97, 2006.

[32] G. Freedman and R. Fattal, "Image and video upscaling from local self-examples," *ACM Transactions on Graphics*, vol. 28, no. 3, pp. 1–10, 2010.

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

[34] C. Ledig, L. Theis, F. Huszár, J. Caballero, A. Cunningham, A. Acosta, A. Aitken, A. Tejani, J. Totz, Z. Wang *et al.*, "Photo-realistic single image super-resolution using a generative adversarial network," *arXiv preprint arXiv:1609.04802*, 2016.

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

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

[37] 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)*, 2015, pp. 1486–1494.

[38] D. P. Kingma and M. Welling, "Auto-Encoding Variational Bayes," *arXiv:1312.6114*, no. Ml, pp. 1–14, 2013.

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

[40] 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)*, vol. 1, 2001, pp. 192–198.

[41] S. Baker and T. Kanade, "Limits on super-resolution and how to break them," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 24, no. 9, pp. 1167–1183, 2002.

[42] C. Q. Xiang Ma, Junping Zhang, "Hallucinating face by position-patch," *Pattern Recognition*, vol. 43, no. 6, pp. 2224–2236, 2010.

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

[44] N. Wang, D. Tao, X. Gao, X. Li, and J. Li, "A comprehensive survey to face hallucination," *International Journal of Computer Vision*, vol. 106, no. 1, pp. 9–30, 2014.

[45] Y. Jin and C.-S. Bouganis, "Robust multi-image based blind face hallucination," in *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015, pp. 5252–5260.

[46] X. Yu and F. Porikli, "Ultra-resolving face images by discriminative generative networks," in *European Conference on Computer Vision (ECCV)*, 2016, pp. 318–333.

[47] Y. Li, C. Cai, G. Qiu, and K. M. Lam, "Face hallucination based on sparse local-pixel structure," *Pattern Recognition*, vol. 47, no. 3, pp. 1261–1270, 2014.

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

[49] ——, "Hallucinating very low-resolution unaligned and noisy face images by transformative discriminative autoencoders," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017, pp. 3760–3768.

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

[51] A. Odena, V. Dumoulin, and C. Olah, "Deconvolution and checkerboard artifacts," *Distill*, 2016. [Online]. Available: http://distill.pub/2016/deconv-checkerboard

[52] S. Ioffe and C. Szegedy, "Batch normalization: Accelerating deep network training by reducing internal covariate shift," in *Proceedings of The 32nd International Conference on Machine Learning (ICML)*, 2015, pp. 448–456.

[53] C. Dong, Y. Deng, C. Change Loy, and X. Tang, "Compression artifacts reduction by a deep convolutional network," in *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2015, pp. 576–584.

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

[55] R. Gonzalez and P. Wintz, "Digital image processing (second edition)," pp. 187–191, 1977.

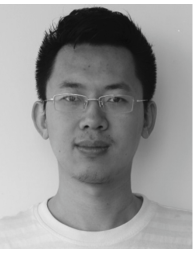
[56] S. Gu, W. Zuo, Q. Xie, D. Meng, X. Feng, and L. Zhang, "Convolutional sparse coding for image super-resolution," in *Proceedings of the IEEE International Conference on Computer Vision*, 2015, pp. 1823–1831.

[57] X. Yu, F. Xu, S. Zhang, and L. Zhang, "Efficient patch-wise non-uniform deblurring for a single image," *IEEE Transactions on Multimedia*, vol. 16, no. 6, pp. 1510–1524, 2014.

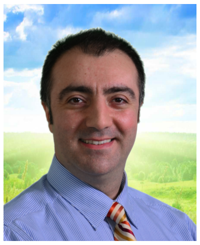
[58] M. Ren, R. Liao, R. Urtasun, F. H. Sinz, and R. S. Zemel, "Normalizing the normalizers: Comparing and extending network normalization schemes," in *5th International Conference on Learning Representations (ICLR)*, 2017.

[59] Z. Yang, K. Zhang, Y. Liang, and J. Wang, "Single image super-resolution with a parameter economic residual-like convolutional neural network," in *International Conference on Multimedia Modeling (MMM)*, 2017, pp. 353–364.

[60] M. E. Tipping and C. M. Bishop, "Mixtures of probabilistic principal component analyzers," *Neural computation*, vol. 11, no. 2, pp. 443–482, 1999.



Xin Yu received his B.S. degree in Electronic Engineering from University of Electronic Science and Technology of China, Chengdu, China, in 2009, and received his Ph.D. degree in the Department of Electronic Engineering, Tsinghua University, Beijing, China, in 2015. He is currently pursuing the Ph.D. degree in the College of Engineering and Computer Science, Australian National University, Canberra, Australia. His interests include computer vision and image processing.



Fatih Porikli is an IEEE Fellow and a Professor with the Research School of Engineering, Australian National University, Canberra, Australia. He received his Ph.D. degree from NYU. Previously he served as a Distinguished Research Scientist at Mitsubishi Electric Research Laboratories, Cambridge, USA. He has contributed broadly to object detection, motion estimation, tracking, image-based representations, and video analytics. He is the coeditor of two books on Video Analytics for Business Intelligence and Handbook on Background Modeling and Foreground Detection for Video Surveillance. He is an Associate Editor of five journals. His publications won four Best Paper Awards and he has received the RD100 Award in the Scientist of the Year category in 2006. He served as the General and Program Chair of numerous IEEE conferences in the past. He has 66 granted patents.