Ultra-Resolving Face Images by Discriminative Generative Networks

Xin Yu(✉) and Fatih Porikli

Australian National University, Canberra, Australia
{xin.yu,fatih.porikli}@anu.edu.au

Abstract. Conventional face super-resolution methods, also known as face hallucination, are limited up to $2 \sim 4\times$ scaling factors where $4 \sim 16$ additional pixels are estimated for each given pixel. Besides, they become very fragile when the input low-resolution image size is too small that only little information is available in the input image. To address these shortcomings, we present a discriminative generative network that can ultra-resolve a very low resolution face image of size 16×16 pixels to its $8\times$ larger version by reconstructing 64 pixels from a single pixel. We introduce a pixel-wise ℓ_2 regularization term to the generative model and exploit the feedback of the discriminative network to make the upsampled face images more similar to real ones. In our framework, the discriminative network learns the essential constituent parts of the faces and the generative network blends these parts in the most accurate fashion to the input image. Since only frontal and ordinary aligned images are used in training, our method can ultra-resolve a wide range of very low-resolution images directly regardless of pose and facial expression variations. Our extensive experimental evaluations demonstrate that the presented ultra-resolution by discriminative generative networks (UR-DGN) achieves more appealing results than the state-of-the-art.

Keywords: Super-resolution · Discriminative Generative Networks · Face

1 Motivation

Face images arguably carry the most interesting and valuable visual information and can be obtained in a non-intrusive manner. Still, for many applications from content enhancement to forensics, face images require significant magnification.

In order to generate high-resolution (HR) face images from low-resolution (LR) inputs, face hallucination [1–12] attracted great interest in the past. These state-of-the-art face hallucination methods can achieve exciting results up to 4× upscaling factors when accurate facial features and landmarks can be found in

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

B. Leibe et al. (Eds.): ECCV 2016, Part V, LNCS 9909, pp. 318–333, 2016.
DOI: 10.1007/978-3-319-46454-1_20

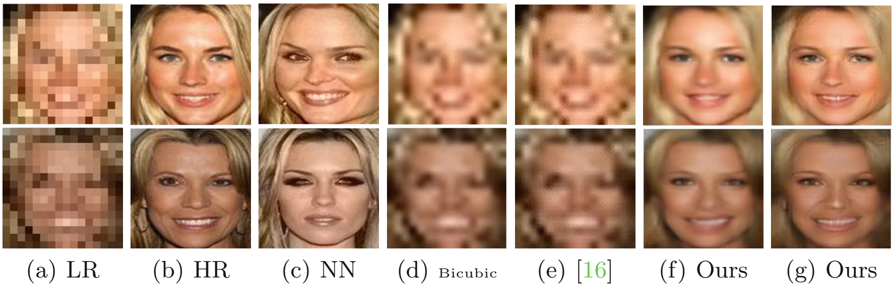


Fig. 1. Comparison of our UR-DGN over CNN based super-resolution. (a) 16×16 pixels LR face images [given]. (b) 128×128 original HR images [not given]. (c) The corresponding HR version of the nearest neighbors of (a) in the training set. (d) Upsampling by bicubic interpolation. (e) The results generated by the CNN based super-resolution [16]. This network is *retrained* with face images. (f) Our UR-DGN without the feedback of the discriminative model. (g) Our UR-DGN.

LR images [9,10], manual supervision is provided, suitably similar HR images of the same person are included in the support dataset, and the exemplar HR face images are densely aligned [4–7]. When the input image resolution becomes smaller, landmark based methods fail gravely because of erroneous landmark localization. In other words, their performances highly depend on the input image size. Furthermore, when the appearances of the input LR images are different from the HR images in the dataset due to pose, lighting and expression changes, subspace based methods degrade by producing ghosting artifacts in the outputs.

When ultra-resolving (8× scaling factor) a low-resolution image, almost 98.5 % of the information is missing. This is a severely ill-posed problem. As indicated in [13], when the scaling factor increases to 8×, the performances of existing approaches degrade acutely.

Our intuition is that by better exploring the information available in the natural structure of face images, appearance similarities between individuals, and emerging large-scale face datasets [14,15], it may be possible to derive competent models to reconstruct authentic 8× magnified HR face images. Deep neural networks, in particular convolutional neural networks (CNN), are inherently suitable for learning from large-scale datasets. Very recently, CNN based generic patch super-resolution methods have been proposed [16,17] without focusing on any image class. A straightforward retraining (fine-tuning) of these networks with face image patches cannot capture the global structure of faces. As shown in Fig. 1(e), these networks fail to produce realistic and visually pleasant results. In order to retain the global structure of faces while being able to reconstruct instance specific details, we use whole face images to train our networks.

We are inspired from the generative adversarial network (GAN) [18] that consists of two topologies: a generative network G that is designed to learn the distribution of the training data samples and generate a new sample similar to the training data, and a discriminative network D that estimates the probability that a sample comes from the training dataset rather than G. This work is empowered with a Laplacian pyramid by [19] to progressively generate images due to the higher dimensional nature of the training images. One advantage of GAN is that it generates face images yet sharp images from nothing but noise. However, it has two serious shortcomings: (i) The output faces are totally random. (ii) GAN has fixed output size limitation (32×32 [18] and 64×64 [19]). Therefore, GAN cannot be used for ultra-resolution directly.

Instead of noise, we apply the LR face image l as the input for our discriminative-generative network (DGN) and then generate a HR face image $\hat{h}$. In order to enforce the similarity between the generated HR face image $\hat{h}$ and the exemplar HR image h, we impose a pixel-wise ℓ_2 regularization on the differences between $\hat{h}$ and h in the generative network. This enables us to constrain the affinity between the exemplar HR images and the generated HR images. Hereby, a loss function layer is added to G. Finally, the generative network G produces a HR image consistent with the exemplar HR image. In training DGN, the discriminative network D provides feedback to G to distinguish whether the upsampled face image is considered (classified by the D) as real (sharp) or as generated (smooth). As shown in Fig. 1(f), by directly upsampling images by the generative network G, we are not able to obtain face images with sharp details. In contrast, with the help of the network D, we can generate much sharper HR face images, as shown in Fig. 1(g). Since the discriminative network is designed to distinguish between the real face images and generated ones, the generative network can produce HR face images more similar to real images.

Our method does not make any explicit assumption or require the location of the facial landmarks. Because the convolutional neural network topologies we use provide robustness to translations and deformations, our method does not need densely aligned HR face images or constrain the face images to controlled settings, such as the same pose, lighting and facial expression. Our approach only requires frontal and approximately nearby eye locations in the training images, which can be easily satisfied in most of face datasets. Hence, our UR-DGN method can ultra-resolve 8× a wide range of LR images without taking other information into account.

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

- We present a novel method to ultra-resolve, 8× scaling factor, low-resolution face images. The size of our input low-resolution images is tiny, 16×16 pixels, which makes the magnification task even more challenging as almost all facial details are missing. We reconstruct 64 pixels from only 1 pixel.
- To the best of our knowledge, our method is the first attempt to develop discriminative generative networks for generating authentic face images. We demonstrate that our UR-DGN achieves better visual results than the state-of-the-art.
- We show that by introducing a pixel-wise ℓ_2 regularization term into the network and backpropagating its residual, it is possible to ultra-resolve in any size while GANs can only generate images in fixed and small sizes.

– When training our network, we only require frontal and approximately aligned images, which makes the training datasets more attainable. Our UR-DGN can ultra-resolve regardless of pose, lighting and facial expressions variations.

2 Related Work

Super-resolution can be basically classified into two categories: generic super-resolution methods and class-specific super-resolution methods. When upsampling LR images, generic methods employ priors that ubiquitously exist in natural images without considering any image class information. Class-specific methods, also called face hallucination [1] if the class is face, aim to exploit statistical information of objects in a certain class. Thus, they usually attain better results than generic methods when super-resolving images of a known class.

Generic super-resolution: In general, generic single image super-resolution methods have three types: interpolation based methods, image statistics based methods [20,21] and example (patch)-based methods [7,22–26]. Interpolation based methods such as bicubic upsampling are simple and computationally efficient, but they generate overly smooth edges as the scaling factor increases. Image statistics based methods employ natural image priors to predict HR images, but they are limited to smaller scaling factors [27]. [24,26,28,29] exploit self-similarity of patches in an input image to generate high resolution patches. [22,23] constructs LR and HR patch pairs from a training dataset, and then the nearest neighbor of the input patch is searched in the LR space. The HR output is reconstructed from the corresponding HR patch. [7] proposes a sparse representation formulation by reconstructing corresponding LR and HR dictionaries, while [30] applies convolutional sparse coding instead of patch-based sparse coding. Recently, several deep learning based methods [16,17] have been proposed. Dong *et al.* [16] incorporates convolutional neural networks to learn a mapping function between LR and HR patches from a large-scale dataset. Since many different HR patches may correspond to one LR patch, the output images would suffer from artifacts at the intensity edges. In order to reduce the ambiguity between the LR and HR patches, [31] exploits the statistical information learned from deep convolutional network to reduce ambiguity between LR and HR patches.

Face hallucination: Unlike generic methods, class-specific super-resolution methods [1–6,8–12] further exploit the statistical information in the image categories, thus leading to better performances. In one of the earlier works, [1] builds the relationship between HR and LR patches using Bayesian formulation such that high-frequency details can be transferred from the dataset for face hallucination. It can generate face images with richer details. However, artifacts also appear due to the possible inconsistency of the transferred HR patches.

The work in [4] employs constraints on both LR and HR images, and then hallucinate HR face images by an eigen-transformation. Although it is able to magnify LR images by a large scaling factor, the output HR images suffer from ghosting artifacts as a result of using a subspace. Similarly, [5] enforces linear constraints for HR face images using a subspace learned from the training set via Principle Component Analysis (PCA), and a patch-based Markov Random Field is proposed to reconstruct the high-frequency details in the HR face images. This method works only when the images are precisely aligned at fixed poses and expressions. In other cases, the results usually contain ghosting artifacts due to PCA based holistic appearance model. To mitigate artifacts a blind bilateral filtering is used as a post-processing step. Instead of imposing global constraints, [8] uses multiple local constraints learned from exemplar patches, and [32] reserves to sparse representation on the local structures of faces. [33] uses optimal transport in combination with subspace learning to morph a HR image from the LR input. These subspace based methods require that face images in the dataset are precisely aligned and the test LR image has the same pose and facial expression as the HR face images.

In order to handle various poses and expressions, [9] integrates SIFT flow to align images. This method performs adequately when the training face images are highly similar to the test face image in terms of identity, pose, and expression. Since it uses local features to match image segments, the global structure is not preserved either.

By exploiting local structures of face images, [10] presents a structured face hallucination method. It divides a face image into facial components, and then maintains the structure by matching gradients in the reconstructed output. However, this method relies on accurate facial landmark points that are usually unavailable when the image size is very small. The recent work in [11] proposes a bichannel CNN to hallucinate face images in the wild. Since it needs to extract features from the input images, the smallest input image size is 48 × 48.

Some generative network [18,19,34,35] can generate random face images from nothing but random noise. Among those generative models, generative adversarial networks (GANs) [18,19] can generate face images with much sharper details due to the discriminative network. However, the generated images are only similar in the class domain but different in the appearance domain. In other words, GAN is capable of generating only random faces. Moreover, GAN only uses the cross entropy loss function of discriminative models to optimize the entire network. Hence, the generative models in GAN are difficult to generate images in high resolutions. For instance, [18] only produces images of size 32 × 32 pixels.

3 Proposed Ultra-Resolution Method

A processing pipeline of UR-DGN is shown in Fig. 2. Below, we present the pipeline of UR-DGN and describe the details of training the network. We also discuss the differences between UR-DGN and GAN.

3.1 Model Architecture

Let us first recap the generative model G that takes a noise vector z from a distribution $P_{noise}(z)$ as an input and then outputs an image $\hat{x}$ in [18]. The

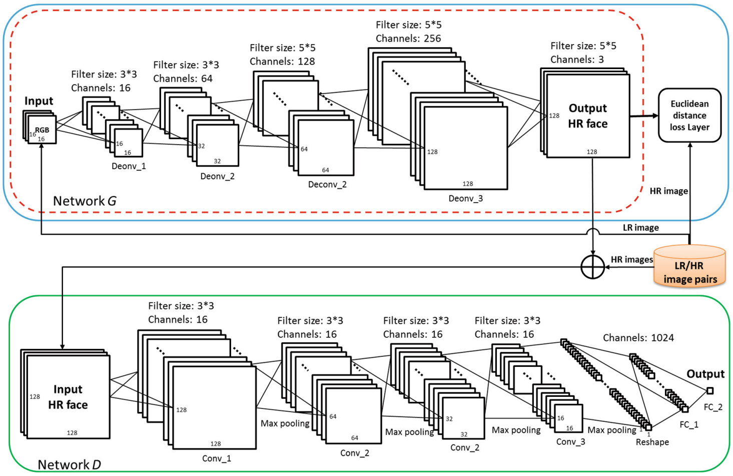


Fig. 2. The pipeline of UR-DGN. In the testing phase, only the generative network in the red dashed block is employed. (Color figure online)

discriminative model D takes an image stochastically chosen from either the generated image $\hat{x}$ or the real image x drawn from the training dataset with a distribution $P_{data}(x)$ as an input. D is trained to output a scalar probability, which is large for real images and small for generated images from G. The generative model G is learned to maximize the probability of D making a mistake. Thus a minmax objective is used to train these two models simultaneously

$$\min_G \max_D \mathbb{E}_{x \sim P_{data}(x)}[\log D(x)] + \mathbb{E}_{z \sim P_{noise}(z)}[\log(1 - D(G(z)))]. \tag{1}$$

This equation encourages G to fit $P_{data}(x)$ so as to fool D with its generated samples $\hat{x}$.

We cannot directly employ Eq. 1 for the ultra-resolution task since GAN takes noise as input to learn the distribution on the training dataset. In UR-DGN, we design a deconvolutional network [36] as the generative model G to ultra-resolve LR inputs, and a convolutional network as the discriminative model D. We construct LR and HR face image pairs $\{l_i, h_i\}$ as the training dataset. Because the generated HR face image $\hat{h}_i$ should be similar to its corresponding HR image h_i, a pixel-wise ℓ_2 regularization term induces the similarity. Thus, the objective function $F(G, D)$ is modeled as follows:

$$\begin{aligned}\min_G \max_D F(G, D) &= \mathbb{E}_{h_i \sim P_H(h)}[\log D(h_i)] + \mathbb{E}_{l_i \sim P_L(l)}[\log(1 - D(G(l_i)))] \\ &\quad + \lambda \mathbb{E}_{(h_i, l_i) \sim P_{HL}(h,l)}[\|\hat{h}_i - h_i\|_F^2] \\ &= \mathbb{E}_{h_i \sim P_H(h)}[\log D(h_i)] + \mathbb{E}_{l_i \sim P_L(l)}[\log(1 - D(G(l_i)))] \\ &\quad + \lambda \mathbb{E}_{(h_i, l_i) \sim P_{HL}(h,l)}[\|G(l_i) - h_i\|_F^2],\end{aligned} \tag{2}$$

where $P_L(l)$ and $P_H(h)$ represent the distributions of LR and HR face images respectively, $P_{HL}(h,l)$ represents the joint distribution of HR and LR face images, and λ is a trade-off weight to balance the cross entropy loss of D and the Euclidean distance loss of G.

3.2 Training of the Network

The parameters of the generative network G and the discriminative network D are updated by backpropagating the loss in Eq. 2 through their respective networks. Specifically, when training G, the loss of the last two terms in Eq. 2 is backpropagated through G to update its parameters. When training D, the loss of the first two terms in Eq. 2 is backpropagated through D to update its parameters.

Training D: Since D is a CNN with a negative cross-entropy loss function, backpropation is used to train the parameters of D. Thus, the derivative of the loss function $F(G,D)$ with respect to D is required when updating the parameters in D. It is formulated as follows:

$$\frac{\partial F(G,D)}{\partial D} = \nabla_{\theta_D}\left(\mathbb{E}_{h_i\sim P_H(h)}[\log D(h_i)] + \mathbb{E}_{l_i\sim P_L(l)}[\log(1-D(G(l_i)))]\right), \quad (3)$$

where θ_D is the parameters of D, and ∇ is the derivative operator. Specifically, given a batch of LR and HR image pairs $\{l_i, h_i\}, i = 1, \ldots, N$, the stochastic gradient of the discriminator D is written as

$$\frac{\partial F(G,D)}{\partial D} = \nabla_{\theta_D}\left(\frac{1}{N}\sum_{i=1}^{N}\log D(h_i) + \log(1-D(G(l_i)))\right), \quad (4)$$

where N is the number of LR and HR face image pairs in the batch. Since we need to maximize D, the parameters θ_D are updated by ascending their stochastic gradients. RMSprop [37] is employed to update the parameters θ_D as follows:

$$\begin{aligned}\delta^{j+1} &= \alpha\delta^j + (1-\alpha)(\frac{\partial F(G,D)}{\partial D})^2,\\ \theta_D^{j+1} &= \theta_D^j + \eta\frac{\partial F(G,D)}{\partial D}/\sqrt{\delta^{j+1}+\epsilon}.\end{aligned} \quad (5)$$

where η and α represent the learning rate and the decay rate respectively, j indicates the iteration index, ϵ is set to 10^{-8} as a regularizer to avoid division by zero, and δ is an auxiliary variable.

Training G: G is a deconvolutional neural network [36]. It is trained by backpropagation as well. Similar to training D, the derivative of the loss function $F(G,D)$ with respect to G is written as

$$\begin{aligned}\frac{\partial F(G,D)}{\partial G} = \nabla_{\theta_G}\big(&\mathbb{E}_{l_i\sim P_L(l)}[\log(1-D(G(l_i)))]\\ &+\lambda\mathbb{E}_{(h_i,l_i)\sim P_{HL}(h,l)}[\|G(l_i)-h_i\|_F^2]\big),\end{aligned} \quad (6)$$

Algorithm 1. Minibatch stochastic gradient descent training of UR-DGN

```
Input: minibatch size $N$, LR and HR face image pairs $\{l_i, h_i\}$, maximum number of
    iterations $K$.
while iter < K do
    Choose one minibatch of LR and HR image pairs $\{l_i, h_i\}, i = 1, \dots, N$.
    Generate one minibatch of HR face images $\hat{h}_i$ from $l_i, i = 1, \dots, N$, where $\hat{h}_i = G(l_i)$.
    Update the parameters of the discriminative network $D$ by using Eqs. 4 and 5.
    Update the parameters of the generative network $G$ by using Eqs. 7 and 8.
end while
Output: UR-DGN.
```

where θ_G denotes the parameters of G. Given a batch of LR and HR face image pairs $\{l_i, h_i\}, i = 1, \dots, N$, the stochastic gradient of the generator G is

$$\frac{\partial F(G, D)}{\partial G} = \nabla_{\theta_G} \left(\frac{1}{N} \sum_{i=1}^{N} \log(1 - D(G(l_i))) + \lambda \| G(l_i) - h_i \|_F^2 \right). \tag{7}$$

Since we will minimize the cost function for G, the parameters θ_G are updated by descending their stochastic gradients as follows:

$$\begin{aligned} \delta^{j+1} &= \alpha \delta^j + (1 - \alpha)(\frac{\partial F(G, D)}{\partial G})^2, \\ \theta_G^{j+1} &= \theta_G^j - \eta \frac{\partial F(G, D)}{\partial G} / \sqrt{\delta^{j+1} + \epsilon}. \end{aligned} \tag{8}$$

In our algorithm, we set the learning rate η to 0.001 and the decay rate to 0.01, and the learning rate is multiplied by 0.99 after each epoch. Since we super-resolve an image rather than generate a face image, we set λ to 100 to constrain the similarity between the generated face image $G(l_i)$ and the exemplar HR face image h_i. The training procedure of our UR-DGN is presented in Algorithm 1.

3.3 Ultra-Resolution of a Given LR Image

The discriminative network D and the pixel-wise ℓ_2 regularization are only required in the training phase. In the ultra-resolution (testing) phase, we take LR face images as the inputs of the generative network G, and the outputs of G are the ultra-resolved face images. This end-to-end mapping is able to keep the global structure of HR face images while reconstructing local details.

3.4 Differences Between GAN and UR-DGN

GAN of [18] consists of fully connected layers, while Denton *et al.* [19] use a fully connected layer and deconvolutional layers. In [19], the noise input is required to be fed into a fully connected layer first before fed into deconvolutional layers.

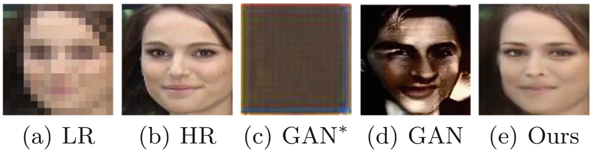


Fig. 3. Illustration of the differences between GAN and our UR-DGN. (a) Given LR image. (b) Original HR image (not used in training). (c) GAN*: GAN with no fully connected layer. Without a fully connected layer, GAN* cannot rearrange the convolutional layer features (activations) of the input noise to a face image. (d) GAN with fully connected layer. Given the test LR image (not noise!), GAN still outputs a random face image. (e) Result of our UR-DGN.

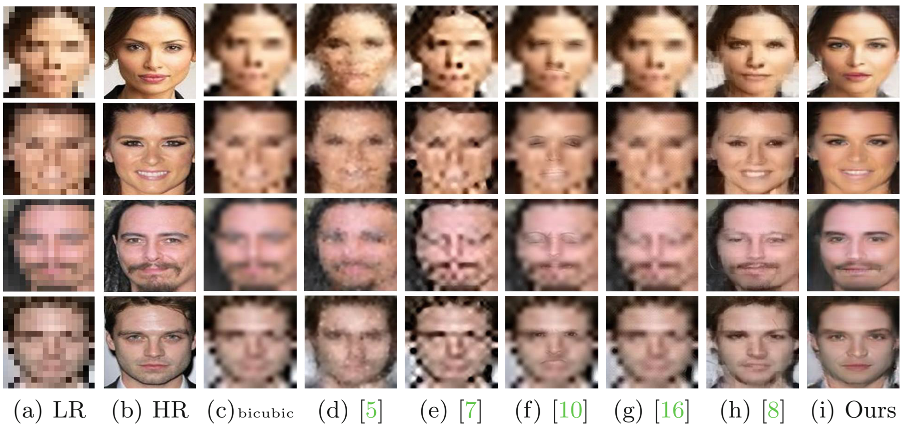
The fully connected layer can be considered as a nonlinear mapping from the noise to the activations of a feature map. If we remove the fully connected layer while leaving other layers unchanged, GAN will fail to produce face images, as shown in Fig. 3(c). Therefore, fully connected layers are necessary for GAN.

Since deconvolutional layers are able to project low-resolution feature maps back to high-resolution image space, we take a LR face image as a 3-channel feature map, and then project this LR feature map into the HR face image space. However, the fully connected layers are not necessary in our UR-DGN. Because LR face images are highly structured, they can be regarded as feature maps after normalization, which scales the range of intensities between -1.0 and 1.0. Feeding a LR face image into a fully connected layer may destroy the global structure of the feature map, *i.e.* the input LR face image. In other words, UR-DGN does not need a nonlinear mapping from an input LR image to a feature map via a fully connected layer.

Furthermore, since there is no pixel-wise regularization in GAN, it cannot produce HR results faithful to the input LR face images and generate high-quality face images as the output size increases as shown in Fig. 3(d). In conclusion, the original architecture of GAN cannot be employed in the ultra-resolution problem.

4 Experiments

In order to dissect the performance of UR-DGN, we evaluate it qualitatively and quantitatively, and compare with the state-of-the-art methods [5,7,8,10,16]. Liu *et al.*'s method [5] is a subspace based face hallucination method. The work in [7] uses sparse representations to super-resolve HR images by constructing LR and HR dictionaries. Yang *et al.*'s method [10] hallucinates face images by using facial components from exemplar images in the dataset. Dong *et al.* [16] employ CNN to upsample images. Ma *et al.* [8] use position-patches in the dataset to reconstruct HR images.



(a) LR (b) HR (c) bicubic (d) [5] (e) [7] (f) [10] (g) [16] (h) [8] (i) Ours

Fig. 4. Comparison with the state-of-the-art methods on frontal faces. (a) LR inputs. (b) Original HR images. (c) Bicubic interpolation. (d) Liu *et al.*'s method [5]. (e) Yang *et al.*'s method [7]. (f) Yang *et al.*'s method [10]. (g) Dong *et al.*'s method [16]. (h) Ma *et al.*'s method [8]. (i) UR-DGN. (please zoom-in to see the differences between (f) and (g). In (f), there are artificial facial edges while (g) has jitter artifacts.)

4.1 Datasets

We trained UR-DGN with the celebrity face attributes (CelebA) dataset [15]. There are more than 200K images in this dataset, where Liu *et al.* [15] use similarity transformation to align the locations of eye centers. We use the cropped face images for training. Notice that the images in this dataset cover remarkably large pose variations and facial expressions. We do not classify the face images into different subcategories according to their poses and facial expressions when training UR-DGN.

We randomly draw $16,000$ aligned and cropped face images from the CelebA dataset, and then resize them to 128×128. We use $15,000$ images for training, 500 images for validation, and 500 images for testing. Thus, our UR-DGN model never sees the test LR images in the training phase.

We downsample the HR face images to 16×16 pixels (without aliasing), and then construct the LR and HR image pairs $\{l_i, h_i\}$. The input of UR-DGN is an image of size 16×16 with 3 RGB channels, and the output is an image of size 128×128 with 3 RGB channels.

4.2 Comparisons with SoA

We do side-by-side comparisons with five state-of-the-art face hallucination methods. In case an approach does not allow $8\times$ scaling factor directly, *i.e.* [7, 16], we repeatedly (three times) apply a scaling factor $2\times$ when ultra-resolving a LR image. For a fair comparison, we use the same dataset CelebA for training of all other algorithms. Furthermore, we apply bicubic interpolation to all input LR images as another baseline.

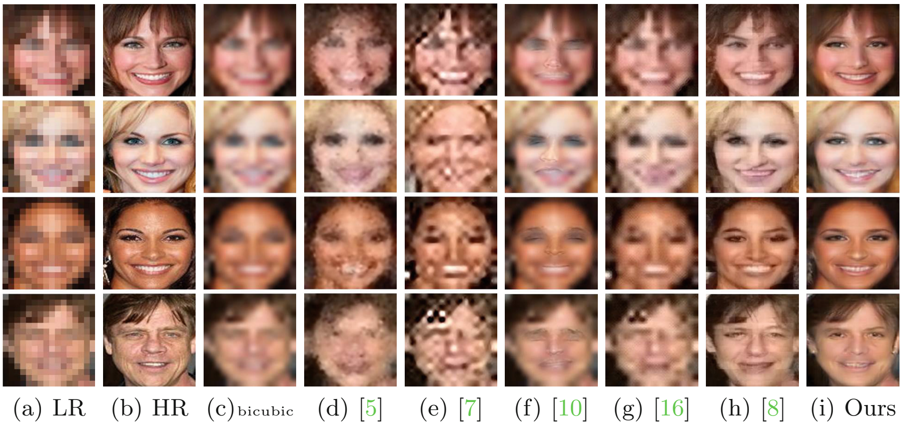


Fig. 5. Facial expression: Comparison with the state-of-the-art methods on images with facial expressions. (a) LR inputs. (b) Original HR images. (c) Bicubic interpolation. (d) Liu *et al.*'s method [5]. (e) Yang *et al.*'s method [7]. (f) Yang *et al.*'s method [10]. (g) Dong *et al.*'s method [16]. (h) Ma *et al.*'s method [8]. (i) UR-DGN. (please zoom-in to see the differences between (f) and (g))

Comparison with Liu *et al.*'s method [5]**:** Since this method requires the face images in the dataset to be precisely aligned, it is difficult for it to learn a representative subspace from the CelebA dataset where face images have large variations. Therefore, the global model of the input LR image cannot be represented by the learned subspace, and its local model impels patchy artifacts on the output. As shown in Figs. 4(d), 5(d) and 6(d), this method cannot recover face details accurately, and suffers from distorted edges and blob-like artifacts.

Comparison with Yang *et al.*'s method [7]**:** As illustrated in Figs. 4(e), 5(e) and 6(e), Yang *et al.*'s method does not recover high-frequency facial details. Besides, non-smooth over-emphasized edge artifacts appear in their results. As the scaling factor becomes larger, the correspondence between LR and HR patches becomes ambiguous. Therefore, their results suffer exaggerated pixellation pattern of the LR, similar to a contrast enhanced bicubic upsampled results.

Comparison with Yang *et al.*'s method [10]**:** This method requires landmarks of facial components and building on them, and reconstructs transferred high-resolution facial components over the low-resolution image. In 16 × 16 input images, it is extremely difficult to localize landmarks. Hence, this method cannot correctly transfer facial components as shown in Figs. 4(f), 5(f) and 6(f). In contrast, UR-DGN does not need landmark localization and still preserve the global structure.

Comparison with Dong *et al.*'s method [16]**:** It applies convolutional layers to learn a generic patch-based mapping function, and achieves state-of-the-art results on natural images. Even though we retrain their CNN on face images to

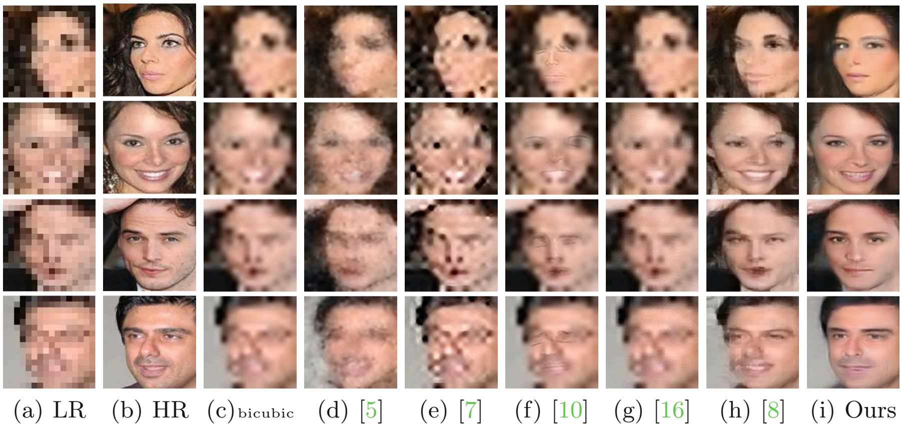


Fig. 6. Pose: Comparison with the state-of-the-art methods on face images with different poses. (a) LR inputs. (b) Original HR images. (c) Bicubic interpolation. (d) Liu *et al.*'s method [5]. (e) Yang *et al.*'s method [7]. (f) Yang *et al.*'s method [10]. (g) Dong *et al.*'s method [16]. (h) Ma *et al.*'s method [8]. (i) UR-DGN. (please zoom-in to see the differences between (f) and (g))

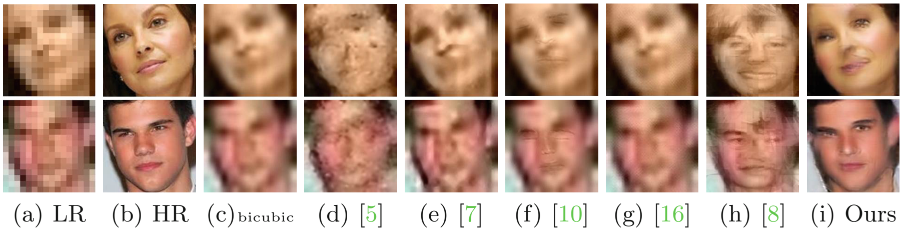


Fig. 7. Comparison with the state-of-the-art methods on unaligned faces. (a) LR inputs. (b) Original HR images. (c) Bicubic interpolation. (d) Liu *et al.*'s method [5]. (e) Yang *et al.*'s method [7]. (f) Yang *et al.*'s method [10]. (g) Dong *et al.*'s method [16]. (h) Ma *et al.*'s method [8]. (i) UR-DGN.

suit better for face hallucination, this method cannot generate high-frequency facial details except some noisy spots in the HR images as shown in Figs. 4(g), 5(g) and 6(g).

Comparison with Ma *et al.*'s method [8]**:** This method employs local constraints learned from positioned exemplar patches to avoid ghosting artifacts caused by a global model such as PCA. However, it requires the exemplar patches to be precisely aligned. As shown in Figs. 4(h), 5(h) and 6(h), this method suffers from obvious blocking artifacts and uneven oversmoothing as a result of the unaligned position patches in the dataset CelebA.

In contrast to the above approaches, our method provides more visually pleasant HR face images that not only contain richer details but also are similar to the original (not given to our method). UR-DGN takes the input LR image as

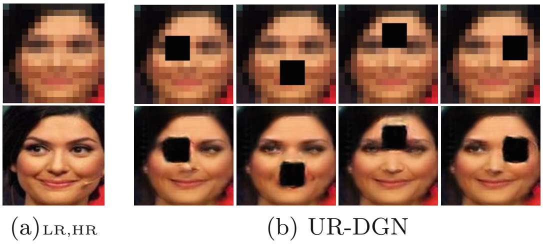


Fig. 8. Illustrations of influence of occlusions. Top row: the LR inputs, bottom row: the results of UR-DGN. (a) LR and HR images. (b) Results of UR-DGN with occlusions. As seen, occlusions of facial features and landmarks (eyes, mouth, etc.) do not cause any degradation of the unoccluded parts of the faces.

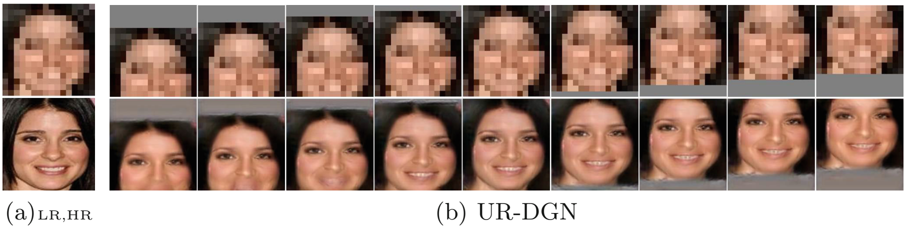


Fig. 9. Effects of misalignment. Top row: the LR images, bottom row: the results of UR-DGN. (a) LR and HR images. (b) Results with translations. From left to right, the y-axis translations are from -4 to +4 pixels. Notice that, the size of the LR image is 16×16 pixels. As visible, UR-DGN is robust against severe translational misalignments.

a whole and reduces the ambiguity of the correspondence between LR and HR patches. Our method attains much sharper results.

4.3 Quantitative Results

We also assess UR-DGN performance quantitatively by comparing the average PSNR and structural similarity (SSIM) on the entire test dataset. Table 1 shows that our method achieves the best performance. As expected, bicubic interpolation achieves better results than the other baselines since it explicitly builds on pixel-wise intensity values without any hallucination. Notice that bicubic interpolation achieves the second best results, which implies that the high-frequency details reconstructed by the state-of-the-art methods are not authentic. Our

Table 1. Quantitative comparisons on the entire test dataset

Methods	Bicubic	[5]	[7]	[10]	[16]	[8]	Ours
PSNR	23.22	21.60	21.35	23.07	23.11	23.12	**24.82**
SSIM	0.67	0.55	0.60	0.65	0.65	0.64	**0.70**

method on the other hand achieves facial details consistent with real faces as it attains the best PSNR and SSIM results while improving the PSNR an impressive 1.6 dB over the previous best.

5 Limitations

Since we use a generative model to ultra-resolve LR face images, if there are occlusions in the images, our method cannot resolve the occlusions. Still, occlusions of facial features do not adversely affect ultra-resolution of the unoccluded parts as shown in Fig. 8.

Our algorithm alleviates the requirements of exact face alignment. As shown in Figs. 7 and 9, it is robust against translations, but sensitive to rotations. As a future work, we plan to investigate incorporating an affine transformation estimator and adapting the generative network according to estimated transformation parameters.

6 Conclusion

We present a new and very capable discriminative generative network to ultra-resolve very small LR face images. Our algorithm can both increase the input LR image size significantly, *i.e.* 8×, and reconstruct much richer facial details. The larger scaling factors beyond 8× only require larger training datasets (e.g., larger than 128×128 training face images for 16×16 inputs), and it is straightforward to achieve even much extreme ultra resolution results.

By introducing a pixel-wise ℓ_2 regularization on the generated face images into the framework of UR-DGN, our method is able to generate authentic HR faces. Since our method learns an end-to-end mapping between LR and HR face images, it preserves well the global structure of faces. Furthermore, in training, we only assume the locations of eyes to be approximately aligned, which significantly makes the other face datasets more attainable.

References

1. Baker, S., Kanade, T.: Hallucinating faces. In: Proceedings - 4th IEEE International Conference on Automatic Face and Gesture Recognition, FG 2000, pp. 83–88 (2000)
2. Liu, C., Shum, H., Zhang, C.: A two-step approach to hallucinating faces: global parametric model and local nonparametric model. CVPR **1**, 192–198 (2001)
3. Baker, S., Kanade, T.: Limits on super-resolution and how to break them. IEEE Trans. Pattern Anal. Mach. Intell. **24**(9), 1167–1183 (2002)
4. Wang, X., Tang, X.: Hallucinating face by eigen transformation. IEEE Trans. Syst. Man Cybern. Part C Appl. Rev. **35**(3), 425–434 (2005)
5. Liu, C., Shum, H.Y., Freeman, W.T.: Face hallucination: theory and practice. Int. J. Comput. Vis. **75**(1), 115–134 (2007)

6. Jia, K., Gong, S.: Generalized face super-resolution. IEEE Trans. Image Process. **17**(6), 873–886 (2008)
7. Yang, J., Wright, J., Huang, T.S., Ma, Y.: Image super-resolution via sparse representation. IEEE Trans. Image Process. **19**(11), 2861–2873 (2010)
8. Ma, X., Zhang, J., Qi, C.: Hallucinating face by position-patch. Pattern Recogn. **43**(6), 2224–2236 (2010)
9. Tappen, M.F., Liu, C.: A Bayesian approach to alignment-based image hallucination. In: Fitzgibbon, A., Lazebnik, S., Perona, P., Sato, Y., Schmid, C. (eds.) ECCV 2012, Part VII. LNCS, vol. 7578, pp. 236–249. Springer, Heidelberg (2012)
10. Yang, C.Y., Liu, S., Yang, M.H.: Structured face hallucination. In: Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, pp. 1099–1106 (2013)
11. Zhou, E., Fan, H.: Learning face hallucination in the wild. In: Twenty-Ninth AAAI Conference on Artificial Intelligence, pp. 3871–3877 (2015)
12. Wang, N., Tao, D., Gao, X., Li, X., Li, J.: A comprehensive survey to face hallucination. Int. J. Comput. Vis. **106**(1), 9–30 (2014)
13. Yang, C.-Y., Ma, C., Yang, M.-H.: Single-image super-resolution: a benchmark. In: Fleet, D., Pajdla, T., Schiele, B., Tuytelaars, T. (eds.) ECCV 2014. LNCS, vol. 8692, pp. 372–386. Springer, Heidelberg (2014). doi:10.1007/978-3-319-10593-2_25
14. Huang, G.B., Ramesh, M., Berg, T., Learned-Miller, E.: Labeled faces in the wild: a database for studying face recognition in unconstrained environments. Technical report 07–49, University of Massachusetts, Amherst, October 2007
15. Liu, Z., Luo, P., Wang, X., Tang, X.: Deep learning face attributes in the wild. In: Proceedings of International Conference on Computer Vision (ICCV), December 2015
16. Dong, C., Loy, C.C., He, K.: Image super-resolution using deep convolutional networks. IEEE Trans. Pattern Anal. Mach. Intell. **38**(2), 295–307 (2016)
17. Kim, J., Lee, J.K., Lee, K.M.: Accurate image super-resolution using very deep convolutional networks. arXiv:1511.04587 (2015)
18. Goodfellow, I., Pouget-Abadie, J., Mirza, M.: Generative adversarial networks. In: Advances in Neural Information Processing Systems, pp. 2672–2680 (2014)
19. Denton, E., Chintala, S., Szlam, A., Fergus, R.: Deep generative image models using a Laplacian pyramid of adversarial networks. In: Advances In Neural Information Processing Systems, pp. 1486–1494 (2015)
20. Peleg, T., Elad, M.: A statistical prediction model based on sparse representations for single image super-resolution. IEEE Trans. Image Process. **23**(6), 2569–2582 (2014)
21. Yang, C.Y., Yang, M.H.: Fast direct super-resolution by simple functions. In: 2013 IEEE International Conference on Computer Vision (ICCV), pp. 561–568 (2013)
22. Freeman, W.T., Jones, T.R., Pasztor, E.C.: Example-based super-resolution. IEEE Comput. Graph. Appl. **22**(2), 56–65 (2002)
23. Chang, H., Yeung, D-Y., Xiong, Y.: Super-resolution through neighbor embedding. In: CVPR, vol. 1, pp. 275–282 (2004)
24. Glasner, D., Bagon, S., Irani, M.: Super-resolution from a single image. In: ICCV, pp. 349–356 (2009)
25. Schulter, S., Leistner, C.: Fast and accurate image upscaling with super-resolution forests. In: CVPR, pp. 3791–3799 (2015)
26. Huang, J.B., Singh, A., Ahuja, N.: Single image super-resolution from transformed self-exemplars. In: IEEE Conference on Computer Vision and Pattern Recognition, pp. 5197–5206 (2015)

27. Lin, Z., Shum, H.Y.: Response to the comments on fundamental limits of reconstruction-based superresolution algorithms under local translation. IEEE Trans. Pattern Anal. Mach. Intell. **28**(5), 847 (2006)
28. Freedman, G., Fattal, R.: Image and video upscaling from local self-examples. ACM Trans. Graph. **28**(3), 1–10 (2010)
29. Singh, A., Porikli, F., Ahuja, N.: Super-resolving noisy images. In: Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, pp. 2846–2853 (2014)
30. Gu, S., Zuo, W., Xie, Q., Meng, D., Feng, X., Zhang, L.: Convolutional sparse coding for image super-resolution. In: ICCV (2015)
31. Bruna, J., Sprechmann, P., LeCun, Y.: Super-resolution with deep convolutional sufficient statistics. In: ICLR (2016)
32. Li, Y., Cai, C., Qiu, G., Lam, K.M.: Face hallucination based on sparse local-pixel structure. Pattern Recogn. **47**(3), 1261–1270 (2014)
33. Kolouri, S., Rohde, G.K.: Transport-based single frame super resolution of very low resolution face images. In: Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR) (2015)
34. Kingma, D.P., Welling, M.: Auto-encoding variational Bayes. arXiv:1312.6114 (Ml), pp. 1–14 (2013)
35. Radford, A., Metz, L., Chintala, S.: Unsupervised representation learning with deep convolutional generative adversarial networks, pp. 1–15 (2015). arXiv:1511.06434
36. Zeiler, M.D., Taylor, G.W., Fergus, R.: Adaptive deconvolutional networks for mid and high level feature learning. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 2018–2025 (2011)
37. Hinton, G.: Neural Networks for Machine Learning Lecture 6a: Overview of mini-batch gradient descent Reminder: The error surface for a linear neuron