Deep CNN Denoiser and Multi-layer Neighbor Component Embedding for Face Hallucination

Junjun Jiang[1,2], **Yi Yu**[2], **Jinhui Hu**[3], **Suhua Tang**[4] and **Jiayi Ma**[5]
[1] Harbin Institute of Technology, Harbin, China
[2] National Institute of Informatics, Tokyo, Japan
[3] The Smart City Research Institute of CETC, Shenzhen, China
[4] The University of Electro-Communications, Tokyo, Japan
[5] Wuhan University, Wuhan, China
{jiangjunjun, yiyu}@nii.ac.jp, hujinhui@cetccity.com, shtang@uec.ac.jp, jyma2010@gmail.com

Abstract

Most of the current face hallucination methods, whether they are shallow learning-based or deep learning-based, all try to learn a relationship model between Low-Resolution (LR) and High-Resolution (HR) spaces with the help of a training set. They mainly focus on modeling image prior through either model-based optimization or discriminative inference learning. However, when the input LR face is tiny, the learned prior knowledge is no longer effective and their performance will drop sharply. To solve this problem, in this paper we propose a general face hallucination method that can integrate model-based optimization and discriminative inference. In particular, to exploit the model based prior, the Deep Convolutional Neural Networks (CNN) denoiser prior is plugged into the super-resolution optimization model with the aid of image-adaptive Laplacian regularization. Additionally, we further develop a high-frequency details compensation method by dividing the face image to facial components and performing face hallucination in a multi-layer neighbor embedding manner. Experiments demonstrate that the proposed method can achieve promising super-resolution results for tiny input LR faces.

1 Introduction

Face hallucination refers to the technique of reconstructing a High-Resolution (HR) face image with fine details from an observed Low-Resolution (LR) face image with the help of HR/LR training pairs [Baker and Kanade, 2000]. It is a domain specific image super-resolution method, which focuses on the human face, and can transcend the limitations of an imaging system, thus providing very important clues about objects for criminals recognition. Due to the highly underdetermined constraints and possible noise, image super-resolution is a seriously ill-posed problem and needs the prior information to regularize the solution space. Mathematically, let $\mathbf{y}$ denotes the observed LR face image, and the target HR face image $\mathbf{x}$ can be deduced by minimizing an energy function composed of a fidelity term and a regularization term balanced through a trade-off parameter λ,

$$\hat{\mathbf{x}} = \arg\min_{\mathbf{x}} \frac{1}{2}||\mathbf{y} - \mathbf{H}\mathbf{x}||^2 + \lambda\Omega(\mathbf{x}). \quad (1)$$

According to the source of the prior information of $\Omega(\mathbf{x})$, the super-resolution techniques can be divided into two categories, model-based optimization methods and discriminative inference learning methods. The former tries to solve the problem of Eq. (1) by some time-consuming iterative optimization algorithms, while the latter aims at learning the

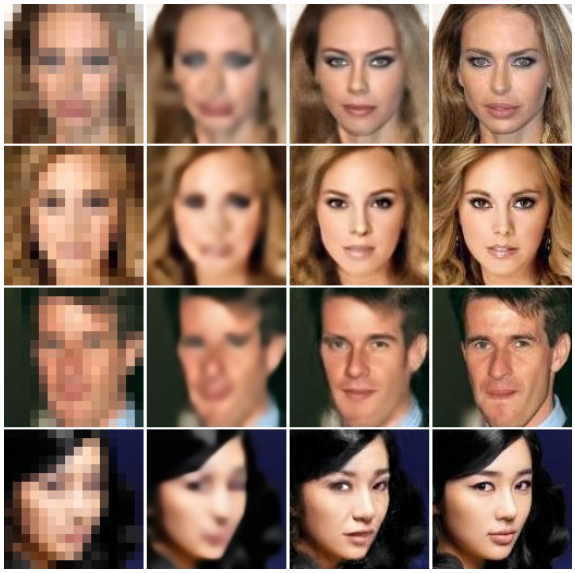


Figure 1: 8× face hallucination results of the proposed method. *Step1*: Global intermediate HR face generation via Deep CNN prior. *Step2*: High-frequency face details compensation. GT: Ground truth.

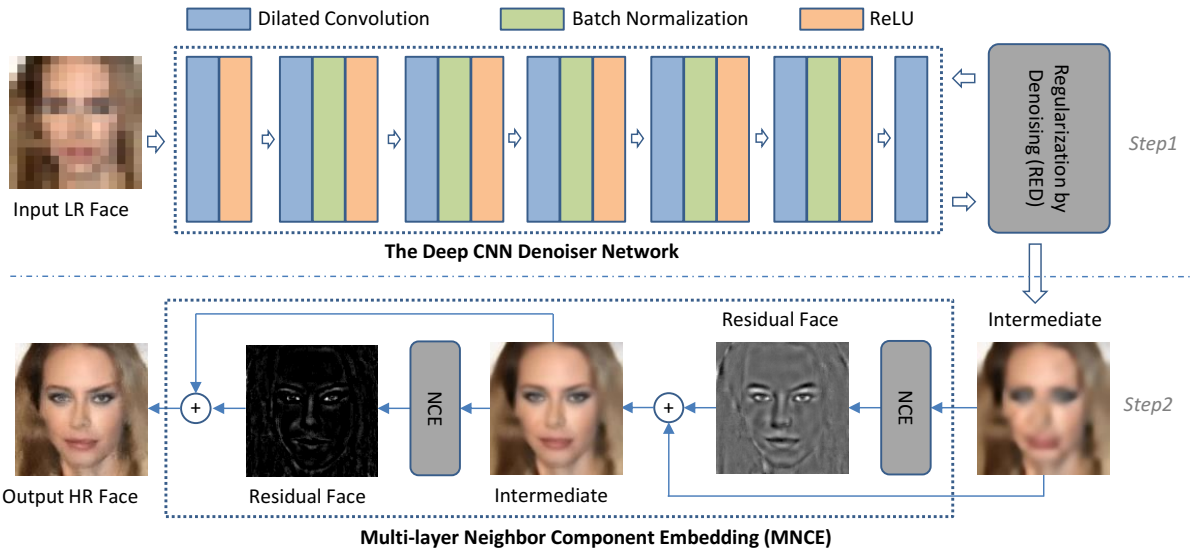


Figure 2: Main steps of the proposed face hallucination algorithm. *Step1*: Deep CNN denoiser based global face reconstruction. *Step2*: MNCE based residual compensation. For convenience, here we only show two layer NCE.

relationship between LR and HR images through a loss function on a training set containing LR and HR sample pairs. Therefore, the model-based optimization methods (such as LRTV [Shi *et al.*, 2015] and NCSR [Dong *et al.*, 2013]) are very general and can be used to handle various image degradation models by specifying the matrix $\mathbf{H}$. In contrast, these discriminative inference learning methods are restricted by specialized image degradation model $\mathbf{H}$. The representative discriminative learning methods include LLE [Chang *et al.*, 2004], ScSR [Yang *et al.*, 2010], ANR [Timofte *et al.*, 2013], SRCNN [Dong *et al.*, 2016], VDSR [Kim *et al.*, 2016], and some methods specifically for face images, TDN [Yu and Porikli, 2017], UR-DGN [Yu and Porikli, 2016], CBN [Zhu *et al.*, 2016], and LCGE [Song *et al.*, 2017]. Due to their end-to-end training strategy, given an LR input image, they can directly predict the target HR image in an efficient and effective way.

In order to overcome the shortcomings of model-based optimization methods and discriminative inference learning methods while leveraging their respective merits, recently, some approaches have been proposed to handle the fidelity term and the regularization term separately, with the aid of variable splitting techniques, such as ADMM optimization or Regularization by Denoising (RED) [Romano *et al.*, 2017]. A model-based super-resolution method tries to iteratively reconstruct an HR image, so that its degraded LR image matches the input LR image, while inference learning tries to train a denoiser by machine learning, using the pairs of LR and HR images. Therefore, the complex super-resolution reconstruction problem is decomposed into a sequence of image denoising tasks, coupled with quadratic norm regularized least-squares optimization problems that are much easier to deal with.

In many real surveillance scenarios, cameras are usually far from the interested object, and the bandwidth and storage resources of systems are limited, which generally result in very small face images, *i.e.*, tiny faces. Although the above-mentioned method is general and can be used to handle various image degradation processes, the performance of this method will become very poor when the sampling factor is very large, *i.e.*, the input LR face image is very small. The learned denoiser prior can not take full advantage of the structure of human face, thus the hallucinated HR faces still lack detailed features, as shown in the second column of Figure 1. In general, Deep Convolutional Neural Networks (CNN) denoiser prior based face hallucination method generates primary face structures fairly well, but fails to return much high-frequency content. To deal with the bottlenecks of very small input images, some deep neural networks based methods have been proposed [Yu and Porikli, 2016; 2017].

In this paper, we develop a novel face hallucination approach via Deep CNN Denoiser and Multi-layer Neighbor Component Embedding (MNCE). Inspired by the work of [Zhang *et al.*, 2017], we adopt CNN to learn the denoiser prior, which is then plugged into a model-based optimization to jointly benefit the merits of model-based optimization and discriminative inference. In this step, we can predict the intermediate results, which look smooth, by this Deep CNN denoiser. In order to enhance the detailed feature, we further propose a residual compensation method through MNCE. It extends NCE to a multi-layer framework to gradually mitigate the inconsistency between the LR and HR spaces (especially when the factor is very large), thus compensating for the missing details that have not been recovered in the first step. Figure 2 shows the pipeline of the proposed algorithm.

The contributions of this work are summarized as follows: (i) We proposed a novel two-step face hallucination method which combines the benefits of model-based optimization and discriminative inference Learning. The proposed framework makes it possible to learn priors from different sources (i.e., general and face images) to simultaneously regularize face

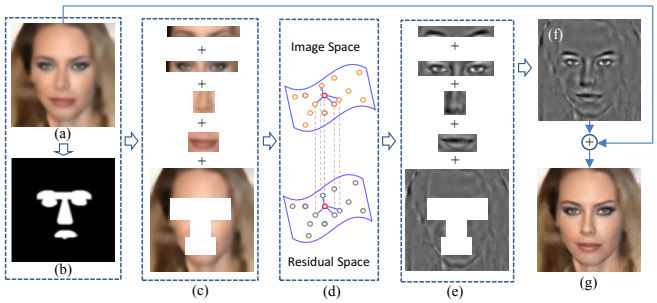


Figure 3: Illustration of neighbor component embedding based residual compensation. (a) Input image. (b) Face component masks. (c) Five facial components. (d) Neighbor embedding on the image and residual manifold spaces. (e) Constructed residual components. (f) Residual face image. (g) Hallucinated face image.

hallucination. (ii) To recover the missing detailed features, neighbor component embedding with multi-layer manner is proposed, and the hallucinated result can be gradually optimized and improved. It provides a scheme to mitigate the inconsistency between LR and HR spaces due to one-to-many mappings.

2 Related Work

There have been several attempts to incorporate advanced denoiser priors into general inverse problems. In [Danielyan *et al.*, 2012], BM3D denoising [Dabov *et al.*, 2007] is adapted to the inverse problem of image deblurring. It was later extended by [Zhang *et al.*, 2014] to other image restoration problems. Most recently, Zhang et al. [Zhang *et al.*, 2017] take advantage of Deep CNN discriminative learning and incorporated it to the model-based optimization methods to tackle with the inverse problems. It exhibits powerful prior modeling capacity. When the magnification is large, however, these denoiser prior based super-resolution methods cannot reconstruct the discriminant features. Therefore, residual face compensation is needed to improve the super-resolved results.

Two-step method was first proposed by Liu et al. [Liu *et al.*, 2001], in where the PCA based parametric model is used to generate the global face image and the MRF based local nonparametric model is adopted to compensate the lost face details in the first step. Manifold alignment based two-step methods [Huang *et al.*, 2010] have been proposed to predict the target HR face image in the aligned common space. In [Song *et al.*, 2017], a component generation and enhancement is proposed. They firstly divided the LR test image into five facial components and obtained the basic structure by several parallel CNNs, and then fine grained facial structures are predicted by a component enhancement method.

3 Proposed Algorithm

Our precise pipeline (as shown in Figure 2) works in the following two steps. Firstly, we construct a discriminative denoiser based on the Deep CNN model. Acquiring the denoiser, the super-resolution reconstruction problem can be iteratively solved by Deep CNN denoising and RED with an image-adaptive Laplacian regularizer [Milanfar, 2013]. The output of this step, one intermediate HR face image, suffers from lacking detailed face features (as shown in the second column of Figure 1). Secondly, we propose an MNCE based residual compensation to predict the missing detailed residual face image gradually.

3.1 Deep CNN Denoiser Prior for Global Face Reconstruction

Regularization by Denoising for the Inverse Problem

To solve the problem of (1), some methods have been proposed by transforming it to an image denoising task based on some variable splitting techniques, such as ADMM optimization [Boyd *et al.*, 2011; Afonso *et al.*, 2010] or RED based framework [Romano *et al.*, 2017]. Since the latter adopts a theoretically better founded method than the ADMM optimization, in this paper we apply the RED to handle the restoration task (1). In RED, the regularizer $\Omega(\mathbf{x})$ is defined by a denoiser,

$$\hat{\mathbf{x}} = \arg\min_{\mathbf{x}} \frac{1}{2}||\mathbf{y} - \mathbf{H}\mathbf{x}||^2 + \frac{\lambda}{2}\mathbf{x}(\mathbf{x} - h(\mathbf{x})), \tag{2}$$

where the function $h(\cdot)$ is an arbitrary denoiser. In Eq. (2), the second term is an image-adaptive Laplacian regularizer [Milanfar, 2013], which can lead to either a small inner product between $\mathbf{x}$ and the residual $(\mathbf{x} - h(\mathbf{x}))$, or a small residual image. Now, the problem is how to optimize the energy function:

$$\mathbf{E}(\mathbf{x}) = \frac{1}{2}||\mathbf{y} - \mathbf{H}\mathbf{x}||^2 + \frac{\lambda}{2}\mathbf{x}(\mathbf{x} - h(\mathbf{x})). \tag{3}$$

Following [Romano *et al.*, 2017], which states that the gradient of $\Omega(\mathbf{x})$ can be induced under the mild assumptions, *i.e.*, $\nabla_{\mathbf{x}}\Omega(\mathbf{x}) = \mathbf{x} - h(\mathbf{x})$. Thus, we can obtain the gradient of $\mathbf{E}(\mathbf{x})$ by

$$\nabla_{\mathbf{x}}\mathbf{E}(\mathbf{x}) = \mathbf{H}^T(\mathbf{H}\mathbf{x} - \mathbf{y}) + \lambda(\mathbf{x} - h(\mathbf{x})). \tag{4}$$

Therefore, we can easily get the update rule by setting $\nabla_{\mathbf{x}}\mathbf{E}(\mathbf{x}) = 0$,

$$\begin{aligned} 0 &= \mathbf{H}^T(\mathbf{H}\hat{\mathbf{x}}_{k+1} - \mathbf{y}) + \lambda(\hat{\mathbf{x}}_{k+1} - h(\hat{\mathbf{x}}_k)) \\ \Rightarrow \hat{\mathbf{x}}_{k+1} &= (\mathbf{H}^T\mathbf{H} + \lambda\mathbf{I})^{-1}(\mathbf{H}^T\mathbf{y} + \lambda h(\hat{\mathbf{x}}_k)). \end{aligned} \tag{5}$$

Through a sequence of image denoising problems and L_2 norm regularized least-squares optimization problems, we can take full advantage of model-based optimization methods and discriminative inference learning methods: various degradation process can be handled and advanced denoiser prior can be easily incorporated.

Learning the Deep CNN Denoiser Prior

Inspired by [Zhang *et al.*, 2017], we also introduce the Deep CNN denoiser to model the discriminative image prior for its efficiency sue to parallel computation ability of GPU and powerful prior modeling capacity with deep neural networks . The above part of Figure 2 shows the architecture of the Deep CNN denoiser network, which consists of seven hidden layers, "Dilated Convolution + ReLU" block in the first layer, five "Dilated Convolution + Batch Normalization + ReLU" blocks in the middle layers, and "Dilated Convolution" block in the last layer.

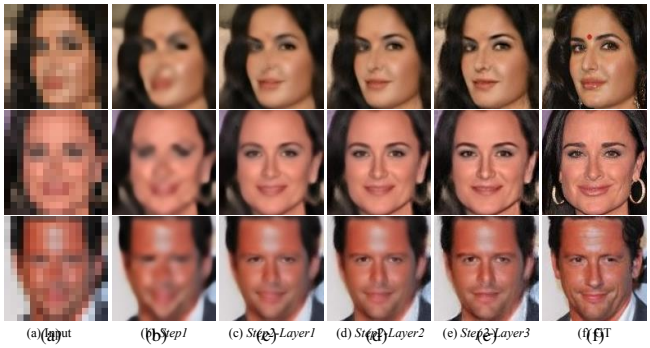


Figure 4: Face hallucination results of different steps of the proposed method. (a) Input. (b) *Step1*. (c) *Step2-Layer1*. (d) *Step2-Layer2*. (e) *Step2-Layer3*. (f) GT.

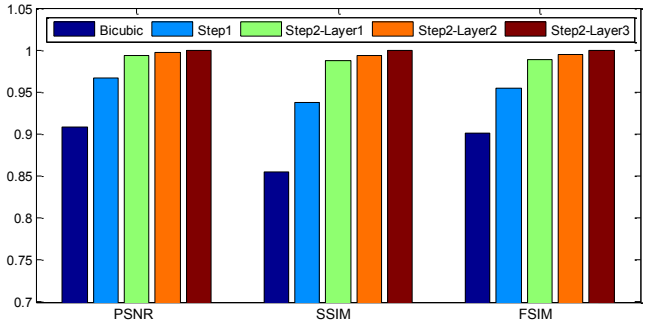


Figure 5: PSNR (dB), SSIM, and FSIM results of different steps of the proposed method. Note that we scale these three indices to [0, 1] by dividing their maximums, respectively.

Once the network is trained, we can predict the result by iterative Deep CNN based denoising and solving the L_2 norm regularized least-squares optimization problem. From previous discussion, we learn that this method will become very poor and fail to return much high-frequency content when the sampling factor is very large, due to ignoring the structure of human face, which is a highly structured object. In the following, we will introduce an improvement method to enhance the high-frequency content.

3.2 Multi-layer Neighbor Component Embedding (MNCE) based Residual Compensation

We take the assumption that similar LR contents will share similar potential HR contents. Let $f(\mathbf{y})$ denotes the prediction function, $\mathbf{x} - f(\mathbf{y})$ is the high-frequency residual face image. Therefore, we can construct the HR face $\mathbf{x}'$ with high-frequency residual information through the locality regularized neighbor embedding algorithm,

$$\mathbf{x}' = f(\mathbf{y}) + \sum_{k=1}^{K} w_k^*(\mathbf{x}_k - f(\mathbf{y}_k)) \quad where$$
$$\mathbf{w}^* = \arg\min_{\mathbf{w}} \left\| f(\mathbf{y}) - \sum_{k=1}^{K} w_k f(\mathbf{y}_k) \right\|_2^2 + \lambda \left\| \mathbf{d} \odot \mathbf{w} \right\|_2^2, \tag{6}$$

where $\odot$ denotes point-wise vector product, $f(\mathbf{y}_k)$ refers to K-nearest-neighbor (in the training set) to $f(\mathbf{y})$, $\mathbf{w} = [w_1, w_2, ..., w_K]$ is the embedding weight of $f(\mathbf{y})$ from the global face space to the residual face space, and $\mathbf{d}$ is a K-dimensional locality adaptor that gives different freedom for each training sample, $f(\mathbf{y}_1), f(\mathbf{y}_2), ..., f(\mathbf{y}_K)$, proportional to its similarity to the input $f(\mathbf{y})$. Specifically,

$$d_k = \left\| f(\mathbf{y}_k) - f(\mathbf{y}) \right\|_2. \tag{7}$$

In Eq. (6), the first term represents the reconstruction error with K-NN, the second term represents the local geometry constraint of manifold. Here, the regularization parameter λ represents the trade-off between the closeness to the data and the locality regularization term. Different from traditional LLE based reconstruction method [Roweis and Saul, 2010], which treats each K-NN equally, our method can give different weights to different K-NN, *i.e.*, the dissimilar samples will be penalized heavily and obtain very small reconstruction weights, while the similar samples will be given more freedom and obtain large reconstruction weights. Thus, our method can capture salient properties as well as yield minimized reconstruction error.

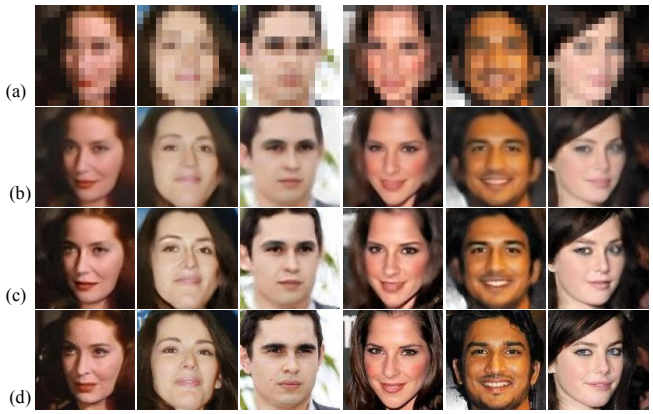


Figure 6: Visual comparisons when using different global face reconstruction methods. (a) Input. (b) Bicubic + MNCE. (c) Deep Denoiser + MNCE. (d) GT.

Neighbor Component Embedding

The above method is limited to reconstruct the entire high-frequency faces, but it is hard for us to find the entire faces that are very similar to the input one. Similar to [Song *et al.*, 2017; Yang *et al.*, 2013], we also divide a face image into five components, *e.g.*, eyes, eyebrows, noses, mouths, and the remaining region, as shown in Figure 3(c). By dividing a face image into different components, we can embed each component from the image space to the residual component face space separately,

$$\mathbf{x'}_j = f_j(\mathbf{y}) + \sum_{k=1}^{K} w_{jk}^*(\mathbf{x}_{jk} - f_j(\mathbf{y}_k)), \text{ where}$$
$$\mathbf{w}^*{}_j = \arg\min_{\mathbf{w}_j} \left\| f_j(\mathbf{y}) - \sum_{k=1}^{K} w_{jk} f_j(\mathbf{y}_k) \right\|_2^2 + \lambda \left\| \mathbf{d}_j \odot \mathbf{w}_j \right\|_2^2, \tag{8}$$

where $\mathbf{x}_{jk}$, $f_j(\mathbf{y}_k)$, and $f_j(\mathbf{y})$ are the j-th component of $\mathbf{x}_k$, $f(\mathbf{y}_k)$, and $f(\mathbf{y})$, respectively, and $\mathbf{w}_j$ is the corresponding embedding vector of $f_j(\mathbf{y})$. Illustration of neighbor component embedding is given in Figure 3. For each facial component, we transform it from the LR image space to the residual

Figure 7: 8× face hallucination comparisons with state-of-the-arts on near frontal input faces. Please zoom in to see the differences.

image space by neighbor embedding. In this way, the high-frequency residual face (Figure 3(f)) can capture tiny details.

Multi-layer Embedding Enhancement

From previous works, we learn that the similar local manifold structure assumption of LR and HR spaces is not always holden in practice. As reported in [Jiang *et al.*, 2014a], the neighborhood preservation rates decrease with the increase of downsampling factor or noise level. In order to reduce the gap between the LR and HR manifold spaces, we introduce a multi-layer embedding enhancement based on the observation that the reconstructed HR manifold of the LR training samples is much more consistent than that of the original LR manifold. With the reconstructed HR training samples and the corresponding HR training samples, we can perform super-resolution reconstruction in much more consistent coupled LR and HR spaces. Specially, in the training phase, we can leverage the "leave-one-out" strategy to obtain the global face based on Deep CNN denoiser, and then predict the residual face through neighbor component embedding for all the LR training face images. When all the LR training face image are updated (super-resoved), we generate a new "LR" training set and take it as the input of the next neighbor embedding layer. In the testing phase, the input LR face can be gradually super-resolved to a satisfactory result.

4 Experiments

The performance of the proposed algorithm has been evaluated on the large-scale Celebrity Face Attributes (CelebA) dataset [Liu *et al.*, 2015a], and we compared our method with the state-of-the-arts qualitatively and quantitatively on the dataset. We adopt the widely used Peak Signal-to-Noise Ratio (PSNR), structural similarity(SSIM) [Wang *et al.*, 2004] as well as feature similarity (FSIM) [Zhang *et al.*, 2011] as our evaluation measurements.

4.1 Dataset

We use the Celebrity Face Attributes (CelebA) dataset [Liu *et al.*, 2015b] as it consists of subjects with large diversities, large quantities, and rich annotations, including 10,177 identities and 202,599 face images. We select ten percent of the data, which includes 20K training images and 260 testing images. And then, these images are aligned and cropped to 128×128 pixels as HR images. The LR images are obtained by Bicubic 8× downsampling (default setting of Matlab function imresize), and thus the input LR faces are 16×16 pixels.

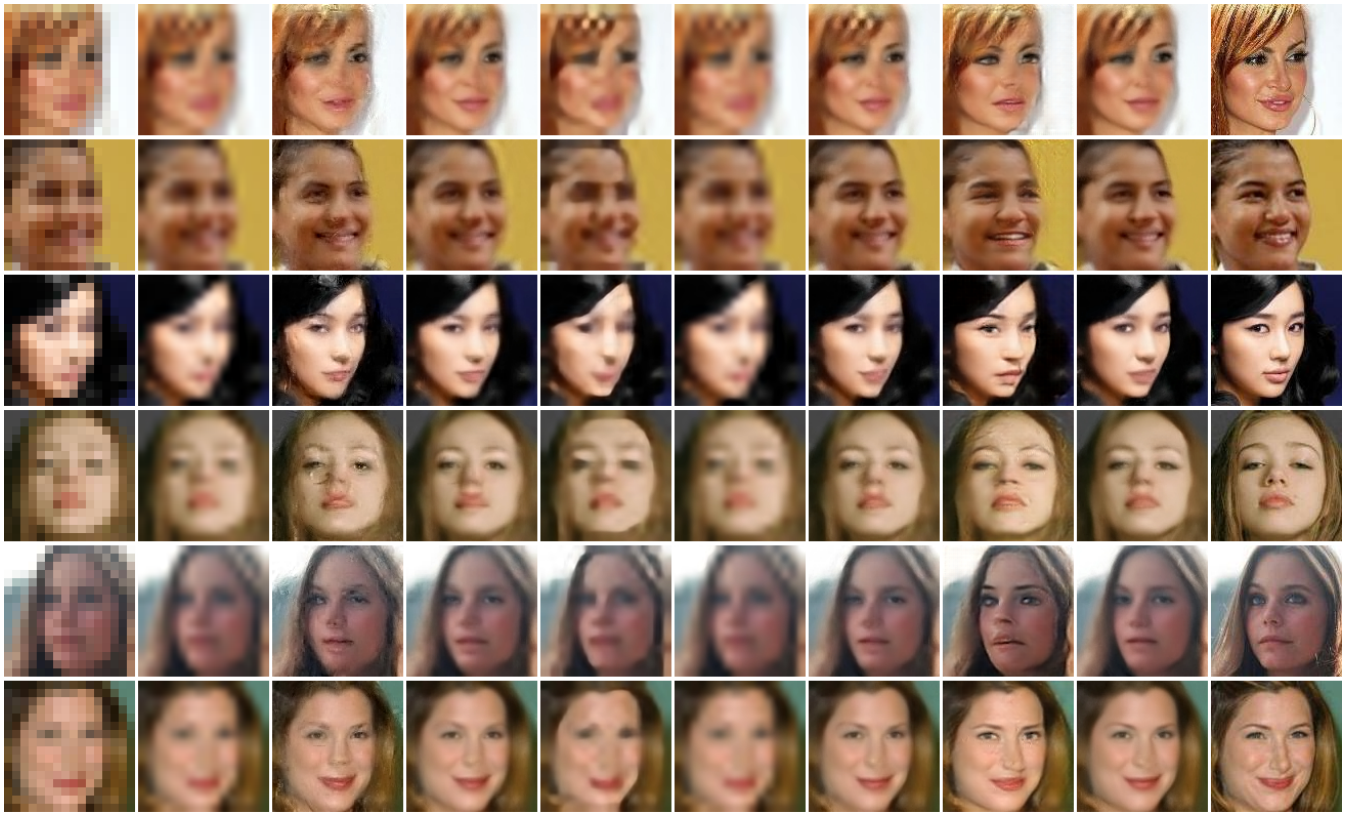


Figure 8: 8× face hallucination comparisons with state-of-the-arts on non-frontal input faces. Please zoom in to see the differences.

4.2 Effectiveness of the Proposed Two-step Methods

To demonstrate the effectiveness of the proposed two-step methods, we give the intermediate results of different steps. As shown in Figure 4, by performing the Deep CNN denoiser based global face reconstruction (*Step1*), it can well maintain the primary facial contours. Through layer-wise component embedding (*Step2*), we can expect to gradually enhance the characteristic details of the reconstructed results (please refer to the third to the fifth columns). As a learned general prior, the Deep CNN denoiser prior cannot be used to model the facial details. However, it can be used to mitigate the manifold inconsistence between the LR and HR image spaces, and this will benefit the following neighbor component embedding learning. At the second step, it is much easier to predict the relationship between the LR and HR spaces when the gap of manifold structure between them is small. Figure 5 quantitatively shows the effectiveness of multi-layer embedding. It demonstrates that by iteratively embedding, we can expect to gradually approach the ground truth.

To demonstrate the effectiveness of the Deep CNN denoiser based global face reconstruction model, we further show the hallucination results of replacing Deep CNN denoiser based global face reconstruction with Bicubic interpolation while keeping the second step (*i.e.*, MNCE) as the same. As shown in Figure 6, Deep CNN denoiser can produce clearer and shaper facial contours. In addition, we also noticed that Bicubic with MNEC can also infer reasonable results, which verifies the ability of MNCE when learning the relationship between the LR faces and residual images.

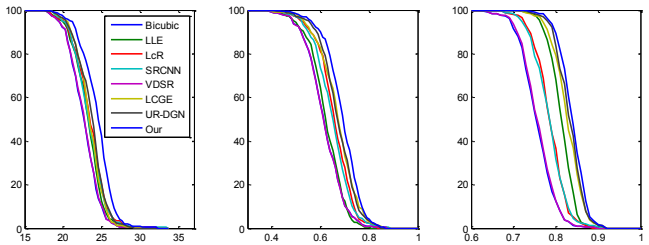
Index	Bicubic	LLE	LcR	SRCNN	VDSR	LCGE	UR-DGN	Our
PSNR	22.61	23.08	23.11	23.27	22.65	23.35	23.55	24.34
SSIM	0.6134	0.6208	0.6542	0.6463	0.6128	0.6673	0.6696	0.6883
FSIM	0.7541	0.8118	0.7843	0.7828	0.7558	0.8257	0.8309	0.8375

Table 1: Average scores in terms of PSNR (dB), SSIM, and FSIM of different face hallucination approaches.

4.3 Qualitative and Quantitative Comparisons

We compare our method with several representative methods, which include LLE [Chang *et al.*, 2004] and LcR [Jiang *et al.*, 2014b], two representative deep learning based methods, SRCNN [Dong *et al.*, 2016], VDSR [Kim *et al.*, 2016], and two most recently proposed face specific image super-resolution methods, *i.e.*, LCGE [Song *et al.*, 2017] and UR-DGN [Yu and Porikli, 2016]. Bicubic interpolation in also introduced as a baseline.

As shown in Figure 7, we also compare the visual results of different comparison methods. It shows that the basic Bicubic interpolation method cannot produce additional details, whereas LLE may introduce some high frequency that doesn't exist. LcR, which focuses on the well aligned frontal face reconstruction, will inevitably smooth the final result due to the misalignments between training samples. As for the



(a) PSNR (b) SSIM (c) FSIM

Figure 9: Image quality statistics using (a) PSNR (dB), (b) SSIM, and (c) FSIM. The horizontal axis labels the scores using PSNR, SSIM, or FSIM, while the vertical axis marks the percentage of hallucinated HR face images whose scores are larger than the score marked on the horizontal axis.

deep learning based technologies, such as SRCNN and VDSR, they can well maintain the face contours due to their global optimization scheme. However, they fail to capture high frequency details (please refer to the eyes, noses, and mouth). This is mainly because when the magnification factor is large, it is very difficult for them to learn the relationship between the LR and HR images with an end-to-end manner. As a gradual super-resolution approach, LCGE method and the proposed method can infer the original low-frequency global face structure as well as the high-frequency local face details simultaneously. When we look further at the results of LCGE and the proposed method, we learn that our method can produce clearer HR faces (please refer to the eyes, mouths, and facial contours). When compared with UR-DGN, which can be seen as the current most competitive face hallucination method for tiny input, our results are still very competitive and much more reasonable. UR-DGN achieves relatively sharper face contours, but the hallucinated faces are dirty.

In addition to the results on near frontal faces (Figure 7), in Figure 8 we also show some visual hallucination results with non-frontal faces, to further demonstrate the robustness of the proposed method. The advantages of the proposed method are still obvious, especially for the regions of eyes and mouth. For examples, the resultant faces of LcR, SRCNN, and VDSR lack detailed information, LLE introduces some unexpected high-frequency details, and UR-DGN may produce sharp but dirty faces. Although the same for component embedding based two-step method, the proposed method is much more robust to pose variety than the approach of LCGE.

Figure 9 shows the statistical curves of PSNR (dB), SSIM, and FSIM scores of different face hallucination approaches, and Table 1 tabulates their average scores. It shows a considerable quantitative advantage of our method compared to traditional shallow learning based methods and some recently proposed deep learning based methods. By comparing UR-DGN and our method, we learn that the proposed method can generate more reliable results, while UR-DGN can well maintain structure information but introduce dirty pixels.

4.4 Face Hallucination with Surveillance Faces

While existing methods can perform well on standard test databases, they often perform poorly when they encounter low-quality and LR face images obtained in real-world scenarios. Figure 10 shows some face hallucination results on the SCface dataset [Grgic *et al.*, 2011] in which images mimic the real world conditions. The first column is the input LR face image, while the last is the reference HR face image of the corresponding individual that can be seen as the ground truth. The middle four columns are the results of LcR, LCGE, UR-GDN, and the proposed method. We observe that these results are obviously worse than those under the CelebA dataset, which shows the shortcomings of learning based methods that require statistical consistency between the training and testing samples. For example, for the eye regions of the hallucinated results, there are more artifacts than the results in the CelebA dataset. This is mainly due to the self-occlusion problem caused by the pose (*e.g.*, looking down) of surveillance cameras, and it is hard to find such samples in a standard face dataset like CelebA.

LR Bicubic LcR LCGE UR-DGN Our Reference

Figure 10: Real-world face hallucination results of different approaches with low-quality surveillance face images.

5 Conclusions and Future Work

In this paper, we presented a novel two-step face hallucination framework for tiny face images. It jointly took into consideration the model-based optimization and discriminative inference, and presented a Deep CNN denoiser prior based global face reconstruction method. And then, the global intermediate HR face was gradually embedded into the HR manifold space with a multi-layer neighbor component embedding manner. Empirical studies on the large scale face dataset and real-world images demonstrated the effectiveness and robustness of the proposed face hallucination framework.

The input faces are aligned manually or by other algorithms. In future work, we need to consider the face alignment and parsing to hallucinate an LR face image with unknown and arbitrary poses [Zhu *et al.*, 2016; Chen *et al.*, 2018; Yu *et al.*, 2018].

Acknowledgments

The research was supported by the National Natural Science Foundation of China under Grants 61501413 and 61503288, and was also partially supported by JSPS KAKENHI Grant Number 16K16058.

References

[Afonso *et al.*, 2010] Manya V Afonso, José M Bioucas-Dias, and Mário AT Figueiredo. Fast image recovery using variable splitting and constrained optimization. *IEEE Trans. Image Process.*, 19(9):2345–2356, 2010.

[Baker and Kanade, 2000] Simon Baker and Takeo Kanade. Hallucinating faces. In *FG*, pages 83–88, 2000.

[Boyd *et al.*, 2011] Stephen Boyd, Neal Parikh, Eric Chu, Borja Peleato, and Jonathan Eckstein. Distributed optimization and statistical learning via the alternating direction method of multipliers. *Foundations and Trends® in Machine Learning*, 3(1):1–122, 2011.

[Chang *et al.*, 2004] Hong Chang, Dit-Yan Yeung, and Yimin Xiong. Super-resolution through neighbor embedding. In *CVPR*, volume 1, pages 275–282, 2004.

[Chen *et al.*, 2018] Yu Chen, Ying Tai, Xiaoming Liu, Chunhua Shen, and Jian Yang. FSRNet: End-to-end learning face super-resolution with facial priors. In *CVPR*, 2018.

[Dabov *et al.*, 2007] Kostadin Dabov, Alessandro Foi, Vladimir Katkovnik, and Karen Egiazarian. Image denoising by sparse 3-d transform-domain collaborative filtering. *IEEE Trans. Image Process.*, 16(8):2080–2095, 2007.

[Danielyan *et al.*, 2012] Aram Danielyan, Vladimir Katkovnik, and Karen Egiazarian. Bm3d frames and variational image deblurring. *IEEE Trans. Image Process.*, 21(4):1715–1728, 2012.

[Dong *et al.*, 2013] Weisheng Dong, Lei Zhang, Guangming Shi, and Xin Li. Nonlocally centralized sparse representation for image restoration. *IEEE Trans. Image Process.*, 22(4):1620–1630, 2013.

[Dong *et al.*, 2016] Chao Dong, Chen Change Loy, Kaiming He, and Xiaoou Tang. Image super-resolution using deep convolutional networks. *IEEE Trans. Pattern Anal. Mach. Intell.*, 38(2):295–307, 2016.

[Grgic *et al.*, 2011] Mislav Grgic, Kresimir Delac, and Sonja Grgic. Scface—surveillance cameras face database. *Multimedia Tools and Applications*, 51(3):863–879, 2011.

[Huang *et al.*, 2010] Hua Huang, Huiting He, Xin Fan, and Junping Zhang. Super-resolution of human face image using canonical correlation analysis. *Pattern Recogn.*, 43(7):2532–2543, 2010.

[Jiang *et al.*, 2014a] Junjun Jiang, Ruimin Hu, Zhongyuan Wang, and Zhen Han. Face super-resolution via multilayer locality-constrained iterative neighbor embedding and intermediate dictionary learning. *IEEE Trans. Image Process.*, 23(10):4220–4231, Oct 2014.

[Jiang *et al.*, 2014b] Junjun Jiang, Ruimin Hu, Zhongyuan Wang, and Zhen Han. Noise robust face hallucination via locality-constrained representation. *IEEE Trans. on Multimedia*, 16(5):1268–1281, Aug 2014.

[Kim *et al.*, 2016] Jiwon Kim, Jung Kwon Lee, and Kyoung Mu Lee. Accurate image super-resolution using very deep convolutional networks. In *CVPR*, pages 1646–1654, 2016.

[Liu *et al.*, 2001] Ce Liu, Heung-Yeung Shum, and Chang-Shui Zhang. A two-step approach to hallucinating faces: global parametric model and local nonparametric model. In *CVPR*, volume 1, pages 192–198, 2001.

[Liu *et al.*, 2015a] Ziwei Liu, Ping Luo, Xiaogang Wang, and Xiaoou Tang. Deep learning face attributes in the wild. In *ICCV*, pages 3730–3738, 2015.

[Liu *et al.*, 2015b] Ziwei Liu, Ping Luo, Xiaogang Wang, and Xiaoou Tang. Deep learning face attributes in the wild. In *ICCV*, pages 3730–3738, 2015.

[Milanfar, 2013] Peyman Milanfar. A tour of modern image filtering. *IEEE Singal Proc. Mag.*, 30(1):106–128, 2013.

[Romano *et al.*, 2017] Yaniv Romano, Michael Elad, and Peyman Milanfar. The little engine that could: Regularization by denoising (red). *Siam Journal on Imaging Sciences*, 10(4), 2017.

[Roweis and Saul, 2010] Sam T. Roweis and Lawrence K. Saul. Nonlinear dimensionality reduction by locally linear embedding. *Science*, 290(5500):2323–2326, 2010.

[Shi *et al.*, 2015] Feng Shi, Jian Cheng, Li Wang, Pew-Thian Yap, and Dinggang Shen. Lrtv: Mr image super-resolution with low-rank and total variation regularizations. *IEEE Trans. Med. Imag.*, 34(12):2459–2466, 2015.

[Song *et al.*, 2017] Yibing Song, Jiawei Zhang, Shengfeng He, Linchao Bao, and Qingxiong Yang. Learning to hallucinate face images via component generation and enhancement. In *IJCAI*, pages 4537–4543, 2017.

[Timofte *et al.*, 2013] Radu Timofte, Vivek De, and Luc Van Gool. Anchored neighborhood regression for fast example-based super-resolution. In *ICCV*, pages 1920–1927, 2013.

[Wang *et al.*, 2004] Zhou Wang, Alan C Bovik, Hamid R Sheikh, and Eero P Simoncelli. Image quality assessment: from error visibility to structural similarity. *IEEE Trans. Image Process.*, 13(4):600–612, 2004.

[Yang *et al.*, 2010] Jianchao Yang, John Wright, Thomas S Huang, and Yi Ma. Image super-resolution via sparse representation. *IEEE Trans. Image Process.*, 19(11):2861–2873, 2010.

[Yang *et al.*, 2013] Chih-Yuan Yang, Sifei Liu, and Ming-Hsuan Yang. Structured face hallucination. In *CVPR*, pages 1099–1106, 2013.

[Yu and Porikli, 2016] Xin Yu and Fatih Porikli. Ultra-resolving face images by discriminative generative networks. In *ECCV*, pages 318–333. Springer, 2016.

[Yu and Porikli, 2017] Xin Yu and Fatih Porikli. Face hallucination with tiny unaligned images by transformative discriminative neural networks. In *AAAI*, pages 4327–4333, 2017.

[Yu *et al.*, 2018] Xin Yu, Basura Fernando, Richard Hartley, and Fatih Porikli. Super-resolving very low-resolution face images with supplementary attributes. In *CVPR*, 2018.

[Zhang *et al.*, 2011] Lin Zhang, Lei Zhang, Xuanqin Mou, and David Zhang. Fsim: A feature similarity index for image quality assessment. *IEEE Trans. Image Process.*, 20(8):2378–2386, 2011.

[Zhang *et al.*, 2014] Jian Zhang, Debin Zhao, and Wen Gao. Group-based sparse representation for image restoration. *IEEE Trans. Image Process.*, 23(8):3336–3351, 2014.

[Zhang *et al.*, 2017] Kai Zhang, Wangmeng Zuo, Shuhang Gu, and Lei Zhang. Learning deep CNN denoiser prior for image restoration. In *CVPR*, pages 2808–2817, 2017.

[Zhu *et al.*, 2016] Shizhan Zhu, Sifei Liu, Chen Change Loy, and Xiaoou Tang. Deep cascaded bi-network for face hallucination. In *ECCV*, pages 614–630. Springer, 2016.