Learning to Hallucinate Face Images via Component Generation and Enhancement

Yibing Song[1], **Jiawei Zhang**[1], **Shengfeng He**[2], **Linchao Bao**[3], **and Qingxiong Yang**[4]
[1]City University of Hong Kong
[2]South China University of Technology
[3]Tencent AI Lab
[4]University of Science and Technology of China

Abstract

We propose a two-stage method for face hallucination. First, we generate facial components of the input image using CNNs. These components represent the basic facial structures. Second, we synthesize fine-grained facial structures from high resolution training images. The details of these structures are transferred into facial components for enhancement. Therefore, we generate facial components to approximate ground truth global appearance in the first stage and enhance them through recovering details in the second stage. The experiments demonstrate that our method performs favorably against state-of-the-art methods[1].

1 Introduction

Face Hallucination (FH) is a domain specific problem which generates high resolution (HR) face images from low resolution (LR) inputs. Different from generic image super resolution (SR) methods, FH exploits specific facial structures and textures. It generates high quality face images compared with generic image SR methods. This activates a series of FH applications ranging from image editing to video surveillance. More generally, FH is taken as a preprocessing step for face related applications.

The state-of-the-art FH methods transfer facial details from HR training images to LR inputs. They aim to exploit the relationship between LR and HR images either globally or locally. One of the solutions is to align face images in pixel-wise precision between the input and training images. So dense correspondences on the training images can be established and HR facial details can be transferred into LR input image in the form of bayesian inference [Tappen and Liu, 2012] or image gradient [Yang *et al.*, 2013]. The transferred result usually contains more details on the facial component compared with the ones generated using generic image SR techniques.

Despite the demonstrated success, the quality of FH results greatly relies on feature matching between training and input images. Because of the limited texture on the LR input (e.g., 60×80), it is difficult to extract handcrafted features such as SIFT [Lowe, 2004] to make a precise description, especially around facial components (i.e., nose, eyes, and mouth). Such a limitation prevents these features to accurately establish the HR correspondence in the training images. It leads to the incorrect detail transfer and the results will be erroneous. As shown in Fig. 1, the nose generated from [Yang *et al.*, 2013] in (b) is in different shape from that of the ground truth in (f).

(a) Input (NN) PSNR / SSIM | (b) SFH 31.13 / 0.85 | (c) SRCNN 32.98 / 0.89

(d) SRResNet 33.47 / 0.89 | (e) Ours **34.40 / 0.90** | (f) Ground Truth $+\infty$ / 1

Figure 1: The performance of FH and image SR methods.

Recently, Convolutional Neural Network (CNN) has been demonstrated effective in image SR [Dong *et al.*, 2015]. It is formulated as a general form of sparsity representation [Yang *et al.*, 2010] and aims to minimize the pixel-wise difference between network output and ground truth. It achieves state-of-the-art performance on natural images where texture patterns uniformly reside in low frequency base and high frequency details. However, direct applying CNN for FH will blur the facial structure because of the uniqueness of component details. As shown in Fig. 1(c) and (d), the results generated using CNN [Dong *et al.*, 2015] or ResNet [Ledig *et al.*, 2017] models cannot enrich the high frequency details around noses. Meanwhile, finetuning their model using face

[1]Complete experimental results and our implementation are provided on the authors' webpage.

images can not make a noticeable improvement. This indicates that CNN based models can not be directly adopted on FH due to the domain specific properties.

In this paper, we Learn to hallucinate face images via Component Generation and Enhancement (**LCGE**). Different from existing end-to-end CNN networks, we propose a two-stage framework for FH. The first stage learns a mapping function to reconstruct the facial structure of the LR input, which benefits the establishment of HR correspondences. This mapping process is formulated via five CNNs. Each CNN corresponds to one facial component (i.e., eyes, eyebrows, noses, mouth and the remaining region). The input face image is thus divided into five subregions and reconstructed independently using CNN. The advantage of the learned facial component is that the texture information is enriched, which alleviates the matching difficulty of LR images. In the second stage, we generate facial components for both training and input images. And a patch-wise K-NN search is performed for each input component. In this way, we can accurately establish HR correspondences without facial alignment. Then we regress to synthesize HR facial structures with fine grained details. However, the regression is conducted on different subjects, which synthesizes HR structures in different illuminations from our desired output. Finally, the details from the HR structures are transferred to the facial components based on edge-aware image filtering. It can successfully recover the missing details to enhance the components. As a result, the output image well approximates the ground-truth image in both global appearance and facial details.

The contributions of this work are summarized as follows:

- We propose to learn deep facial components, which contain basic structure for output and ease the matching difficulty of LR images.
- We propose a component enhancement method. The fine grained facial structures can be effectively extracted from training dataset and their details will be transferred to enhance deep components.
- Quantitative evaluations on the standard benchmarks indicate that the proposed method performs favorably against state-of-the-art approaches.

2 Related Work

Learning based framework is widely adopted in FH methods [Wang *et al.*, 2014; Song *et al.*, 2014; Wang *et al.*, 2017]. They aim to learn the transformation between LR and HR to recover the missing details from the input. In [Gunturk *et al.*, 2003; Wang and Tang, 2005] generalized approaches on eigen domain are proposed to map both LR and HR image spaces. Tensor based approaches are introduced in [Liu *et al.*, 2005; Jia and Gong, 2008]. They can well upsample multiple model face images across different poses and expressions. In [Liu *et al.*, 2007] Principle Component Analysis (PCA) based linear constraints are learned from training images and a patch-based Markov Random Field (MRF) is used to reconstruct the residues. Instead of directly using patch match [Ma *et al.*, 2010] to find correspondence, FH methods adopt image alignment where HR images are matched to LR ones by SIFT flow [Tappen and Liu, 2012] or gradient [Yang *et al.*, 2013]. The quality of output results depends on image alignment, which sometimes fails when poses and expressions are different between training and input images. The convolution neural networks have been adopted in image SR [Dong *et al.*, 2015; Kim *et al.*, 2016] and FH [Zhou *et al.*, 2015; Yu and Porikli, 2016]. Different from existing methods, ours takes the superior performance of CNN to model global appearance and enriches local details through feature matching. It combines the advantage of image SR and FH methods to improve the face image quality.

3 Proposed Algorithm

We present the pipeline of LCGE in Fig. 2. We use CNN to generate deep facial components for the input LR image. They contain basic structure of the output while details are not recovered completely. These components benefit the establishment of LR-HR correspondences and thus fine grained structures can be effectively extracted. The details of these structures are added back to enhance deep facial components to generate the output result.

3.1 Deep Facial Component Generation

We categorize face image into five subregions. Four of them are defined as facial components covering eyes, eyebrows, noses and mouths. The last one is defined as the remaining region. These subregions can be easily obtained using component mask generated by facial landmarks. For an input LR image, we first upsample it to the same resolution as the output using bicubic interpolation and obtain five subregion patches. Then we take each patch as input to the corresponding CNN to generate deep facial component. We have five CNNs in total, each of them contains three convolutional layers. The network structure and training process are similar with those of SRCNN [Dong *et al.*, 2015].

Discussion

We generate the deep facial components for two purposes. First, CNN is effective to minimize the pixel difference between its output and the ground truth. We divide face image into different components and train one CNN for each component independently. Each CNN is set to capture the specific feature of one facial component and generate basic structures of output. Meanwhile, deep facial component is set as an intermediate state between bicubic upsampling of LR input and the ground truth HR image. It is effective to recover the majority of basic structures except some tiny high frequency details. So the remaining work aims to capture such missing details to enhance deep facial component. In this way, the output will approximate ground truth in both global appearance and local details.

Second, deep facial components are able to transform both input and training images into a similar condition, which enables the accurate establishment of HR correspondences so that fine grained facial structures can be effectively extracted. We downsample facial components from HR training images as input. So we can generate deep facial counterpart for each facial component of training images and formulate a training pair with the HR corresponding component. The training

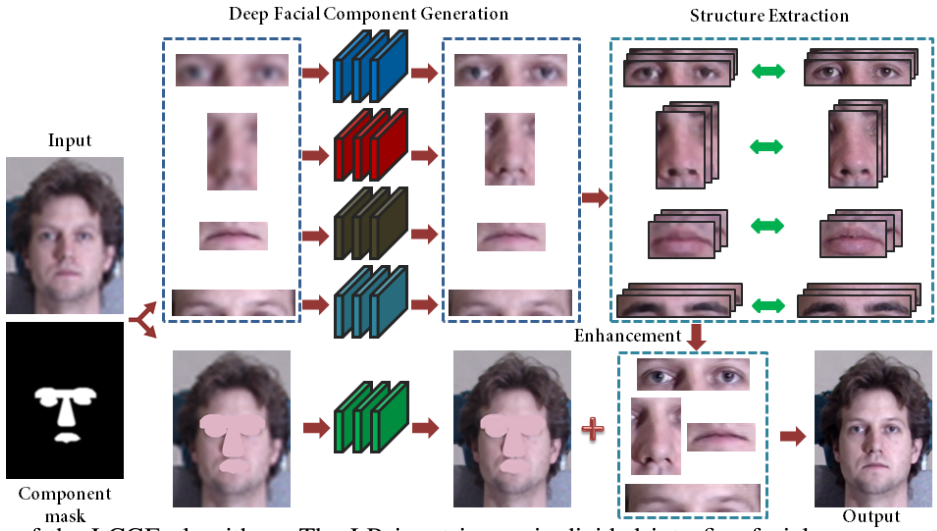


Figure 2: Pipeline of the LCGE algorithm. The LR input image is divided into five facial components. Each of them is upsampled using corresponding CNN to generate deep facial component. Fine grained structures can be extracted from HR training images. We transfer their details to enhance deep facial component to generate output result.

pairs formulation is effective to synthesize HR facial structure through component searching. We compare the similarity of the deep facial component between LR input and LR training images. Once similar components are identified we locate the corresponding HR facial components. Different from the prior art which performs feature matching between LR of input and training images, deep facial component enriches facial texture information and thus can accurately establish the HR correspondences. We use intensity and structure based metric for matching (as shown in Eq. 1) and find it performs well in practice. The main reason is that deep facial component is descriptive enough to distinguish the ambiguity from LR. As such, there is no need to use SIFT [Lowe, 2004] or CNN features [Girshick *et al.*, 2014].

3.2 Component Enhancement

Although deep facial component generation enriches structure information for LR input patches, blur effect still occurs and high frequency details cannot be recovered. Here we propose a component enhancement method to recover high frequency details for the components. It consists of two steps. First, we extract fine grained facial structure from preconstructed training pairs. Then we transfer structure details to enhance deep facial component to generate the output.

Structure Extraction

We aim to extract facial structure from HR training images where the subjects are different from that on an input image. Inspired by [Hertzmann *et al.*, 2001] which involves training image pairs to transfer image style, we construct a training component dataset for facial structure extraction. For each categorized component of the training images, we downsample it into LR and upsample using bicubic interpolation. Then we use the upsampled component as input to obtain deep facial component. As a result training component pairs can be generated which consist of well aligned deep facial components and corresponding HR components.

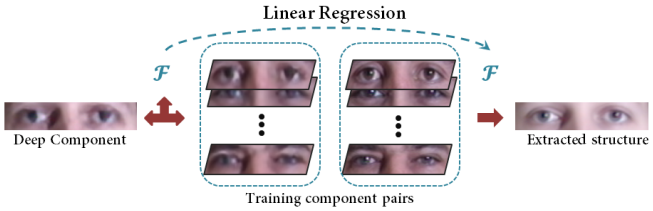


Figure 3: Structure extraction through dataset synthesis. We use CNN to generate deep facial component for both training and input images. Then for each input component patch we establish correspondences from the training dataset. We learn a linear regression function F through the component patches, and use F to map HR training patches to generate extracted result.

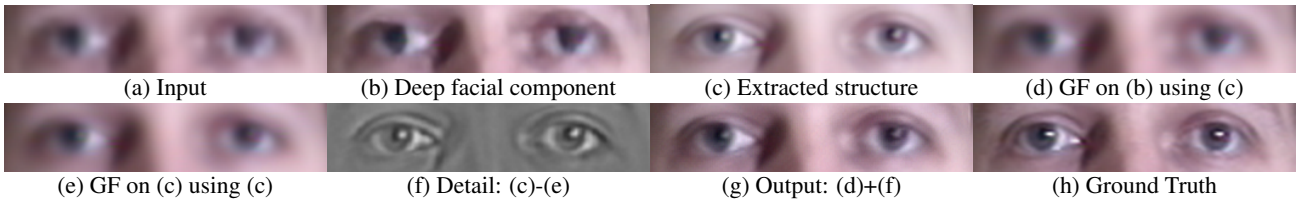
Given an input image we divide it into different components represented by local patches. For one patch centered on pixel p, we perform a K nearest neighbor search (K-NN) on the deep facial component of the training pairs to find the corresponding patches. The patch similarity metric is defined as the combination of normalized cross correlation D_{ncc} and absolute difference D_{abs}:

$$D_p = \alpha \cdot (1 - D_{ncc}) + (1 - \alpha) \cdot D_{abs}, \tag{1}$$

where α is set as 0.2 and K is set as 5 in our experiments. We normalize image pixel value to $[0, 1]$ in order to set two metrics into the same range.

After K-NN search we select K candidate patches from deep components. Let $\bar{\mathrm{T}}_p^i$ $(i \in [1, \cdots, K])$ denote one vector containing all the pixel values of the ith candidate patch, and $\bar{\mathrm{I}}_p$ denote a vector containing the pixel values of the input patch. We also denote the linear regression function as $\mathcal{F}_p = [F_p^1, \cdots, F_p^K]^{\mathrm{T}}$ where F_p^i $(i \in [1, \cdots, K])$ is each coefficient of $\mathcal{F}_p$. The energy function is defined as:

$$E_p^{\text{data}} = ||\bar{\mathrm{T}}_p \cdot \mathcal{F}_p - \bar{\mathrm{I}}_p||^2, \tag{2}$$



(a) Input (b) Deep facial component (c) Extracted structure (d) GF on (b) using (c)

(e) GF on (c) using (c) (f) Detail: (c)-(e) (g) Output: (d)+(f) (h) Ground Truth

Figure 4: Texture transfer. Input LR face image is shown in (a). We generate deep structure shown in (b). Extracted texture is shown in (c). We perform guided filtering on (b) using (c) as guidance shown in (d). We also filter (c) using guided filtering in (e). The lost facial details through filtering can be identified in (f), which is the difference between (c) and (e). We add the details back to (c) to generate the output shown in (g). The ground truth image is shown in (h).

where $\bar{\mathrm{T}}_p = [\bar{\mathrm{T}}_p^1, \bar{\mathrm{T}}_p^2, \cdots, \bar{\mathrm{T}}_p^K]$. It is a linear regression problem and we can compute $\mathcal{F}_p$ as

$$\mathcal{F}_p = (\bar{\mathrm{T}}_p^{\mathrm{T}} \cdot \bar{\mathrm{T}}_p)^{-1} \bar{\mathrm{T}}_p \cdot \bar{\mathrm{I}}_p. \quad (3)$$

We can efficiently compute $\mathcal{F}_p$ when the patches contain texture (i.e., the pixel values in $\bar{\mathrm{T}}_p$ should not be similar with each other). However, in some cases when p is on the smooth region (e.g, nose) $\bar{\mathrm{T}}_p^{\mathrm{T}}$ may be singular and thus $\mathcal{F}_p$ becomes outliers. We resolve the problem by adding a regularization term as:

$$E_p = E_p^{\mathrm{data}} + E_p^{reg} = ||\bar{\mathrm{T}}_p \cdot \mathcal{F}_p - \bar{\mathrm{I}}_p||^2 + \lambda ||\mathcal{F}_p||^2, \quad (4)$$

where λ is the weight controlling the influence of regularization term. It is set as the number of pixels in input patch. We can solve the above energy function as:

$$\mathcal{F}_p = (\bar{\mathrm{T}}_p^{\mathrm{T}} \cdot \bar{\mathrm{T}}_p + \lambda \mathbb{1})^{-1} \bar{\mathrm{T}}_p \cdot \bar{\mathrm{I}}_p, \quad (5)$$

where $\mathbb{1}$ is the identity matrix.

Once we calculate the regression function $\mathcal{F}_p$, we map the HR training patches into the extracted patch. Let T_p^i ($i \in [1, \cdots, K]$) denote one vector containing the pixel values of the corresponding HR training patches. The extracted patch R_p can be computed as:

$$\mathrm{R}_p = \sum_{i=1}^{K} F_p^i \cdot \mathrm{T}_p^i. \quad (6)$$

We compute the extracted patch for each pixel on the input patch. For the overlapping areas between different patches, we perform averaging to generate the result shown in Fig. 3. It can effectively extract fine grained structures through synthesizing from the HR training images.

Detail Transfer

The extracted facial structure contains high frequency details lost in the deep facial component. However, it can not be directly adopted as the output. This is because we extract structure from several training patches which belong to different subjects. The illumination of each subject is different from each other, which results in different grayscale values between extracted structure and ground truth (e.g., Fig. 4 (c) and (h)). We notice that the missing details mostly reside in high frequency (e.g., eyes in Fig. 4). To recover the missing details to enhance deep facial component, we propose a detail transfer method based on edge-preserving filtering [Petschnigg *et al.*, 2004; Eisemann and Durand, 2004]. It can effectively extract the missing details and transfer them back to the deep facial component.

The main steps of detail transfer are shown in Fig. 4. We have a deep facial component patch shown in (b) and a extracted structure shown in (c). We use guided filter [He *et al.*, 2013] to smooth (b) using (c) as guidance. As such, the facial structure of (c) can be transferred into (b). However, the filtered result is likely to be smoothed (as shown in Fig. 4 (d)) through guided filtering process. Nevertheless, we can capture the missing details with the help of (c) to create a similar blurry scenario. First, we smooth (c) using guided filtering with itself as guidance shown in (e). Then missing facial details can be captured through subtracting the smoothed image using (c). As shown in (f), the missing details mainly reside around facial components (e.g, eyes). We add (f) to (d) to recover the missing facial details shown in (g). As a result, both global appearance and facial details of the output component patch is similar to the ground truth shown in (h). After we transfer all the component patches we combine them to generate the output face image.

4 Experiments

We conduct experiments on four datasets: Multi-PIE [Gross *et al.*, 2010] frontal, Multi-PIE pose, PubFig [Kumar *et al.*, 2009] and Multi-PIE HR datasets. In the Multi-PIE pose dataset, face images are taken with pose around 45 degrees while in the other datasets all face images are taken in frontal view. In the PubFig datasets, input images are captured in real world wild condition while in other datasets the inputs are in the lab controlled environment. The resolution of ground truth images in all datasets except Multi-PIE HR is 320×240, and we set the scaling factor as 4. In Multi-PIE HR dataset the resolution of HR images is 800×600, and we set the scaling factor as 10 to evaluate the performance of different algorithms in such an extreme case.

In Multi-PIE frontal dataset, we keep the same setting with that in [Yang *et al.*, 2013] where 2184 images are taken as training and 342 images are taken as input. For Multi-PIE pose and Multi-PIE HR datasets, we adopt leave-one-out strategy for 84 images and 249 images, respectively. For pubFig dataset, we use training images from Multi-PIE frontal to generate 400 output images, which indicates the generality of each method for the real world images. The proposed LCGE method is compared with the state-of-the-art FH methods including FHTP [Liu *et al.*, 2007], SFH [Yang *et al.*, 2013] and four image SR methods including bicubic interpolation,

Table 1: Multi-PIE Frontal Dataset

	Bicubic	FHTP	SRSC	SFH	SRCNN	SRResNet	Ours
PSNR	32.43	30.13	33.54	31.60	33.89	34.10	**35.17**
SSIM	0.89	0.82	0.90	0.86	0.90	0.90	**0.92**

Table 2: Multi-PIE Pose Dataset

	Bicubic	FHTP	SCSR	SFH	SRCNN	SRResNet	Ours
PSNR	33.97	24.59	35.14	32.84	35.45	35.64	**37.55**
SSIM	0.90	0.72	0.91	0.86	0.91	0.92	**0.94**

Table 3: PubFig Dataset

	Bicubic	FHTP	SCSR	SFH	SRCNN	SRResNet	Ours
PSNR	29.55	26.56	30.74	28.51	31.03	31.23	**31.70**
SSIM	0.86	0.71	0.88	0.82	0.88	0.88	**0.89**

Table 4: Multi-PIE HR Dataset

	Bicubic	FHTP	SCSR	SFH	SRCNN	SRResNet	Ours
PSNR	30.38	24.98	30.50	28.71	30.41	30.72	**31.24**
SSIM	0.73	0.66	0.75	0.64	0.75	0.74	**0.76**

Quantitative evaluations on benchmark datasets. Our method performs favorably against state-of-the-art methods in general.

(a) Input (Bic) 32.29 / 0.88 (b) FHTP 29.96 / 0.81 (c) SFH 31.41 / 0.85

(d) SRResNet 34.45 / 0.90 (e) Ours **35.34 / 0.91** (f) Ground Truth PSNR / SSIM

Figure 5: Qualitative evaluation for 4× upsampled face images in Multi-PIE frontal dataset.

SCSR [Yang *et al.*, 2010], SRCNN [Dong *et al.*, 2015] and SRResNet [Ledig *et al.*, 2017]. PSNR and SSIM [Wang *et al.*, 2004] are used to measure image quality.

Table 1 reports the quantitative performance on Multi-PIE frontal dataset under each metric. It shows that bicubic interpolation achieves higher PSNR value than existing FH methods (i.e, FHTP and SFH). This is because FH methods establish HR correspondences through image alignment which is based on hand crafted features such as SIFT flow [Liu *et al.*, 2011]. As the resolution of the input image is low, existing handcrafted features cannot accurately locate HR correspondences. So mismatch occurs and incorrect facial structure will be transferred. As a result, around facial component areas, we will find the distortion of the shape, shifting of the location or change of the lightness, as shown in Fig. 5 (b) and (c). These artifacts deteriorate the image quality. The SCSR, SRCNN and SRResNet methods achieve high PSNR values due to their global optimization scheme. However, blur occurs around high frequency facial components including eyes, noses, and mouth, which limits the image quality as well. The proposed LCGE method recovers the original image content in both low and high frequencies. It enables the similarity of global appearance and local details, which leads to higher numerical values. The remaining datasets indicate similar quantitative performance in Table 2-Table 4. SCSR, SRCNN, and SRResNet are shown to favor better numerical scores than FH methods. But they are still not as good as the performance of proposed LCGE method.

The qualitative evaluation is shown in Fig. 5. The result of FHTP shown in (b) contains noisy and ghosting artifacts (e.g, facial skin) as well as over smoothed facial components (e.g, eyes). The image SR method SRResNet can achieve high numerical scores because of the global optimization scheme. However, they cannot capture high frequency facial details. As shown in (d), the eyeball and eyelid are blurred, as well as noses and mouths. In comparison, SFH can generate high quality facial components shown in (c). This is because SFH selects the most similar component from the dataset and transfer its gradient to recover high frequency details. However, the facial component correspondence can not be well established in LR. In this case, gradient transfer leads to the dissimilar generation of the facial component. The lighting, shape, and position of the left eye in (c) is different from that in (f) in the close ups although they look similar. In addition, noise is included due to incorrect matching around the mouth region. This limitation is solved by the proposed LCGE method where we synthesize from HR images. Through regression we can correctly generate fine grained structures and transfer their details back to the deep facial component. As a result, LCGE will maintain facial details and thus achieve better quantitative values shown in (g). In addition, Fig. 6 and 7 demonstrate similar performance in varying pose and real world conditions, respectively.

The proposed LCGE performs favorably against existing methods in large scaling factors. As shown in Fig. 8 the evaluation is conducted under upscaling factor of 10, which is not conducted by previous FH methods. The visual performance indicates SCSR and SRCNN produce blur on the results shown in (b) and (d). It is because under such a high upscaling factor sparse coding and CNN based methods can not model the relationship between LR and HR well. The result obtained from SFH in (c) contains high frequency details (e.g, eye) when facial components are correctly matched. However, artifacts occur on the mismatched components (e.g., nose and mouth). In comparison, LCGE generates high quality facial structures through HR synthesis and transferring their details to enhance deep component, which maintains

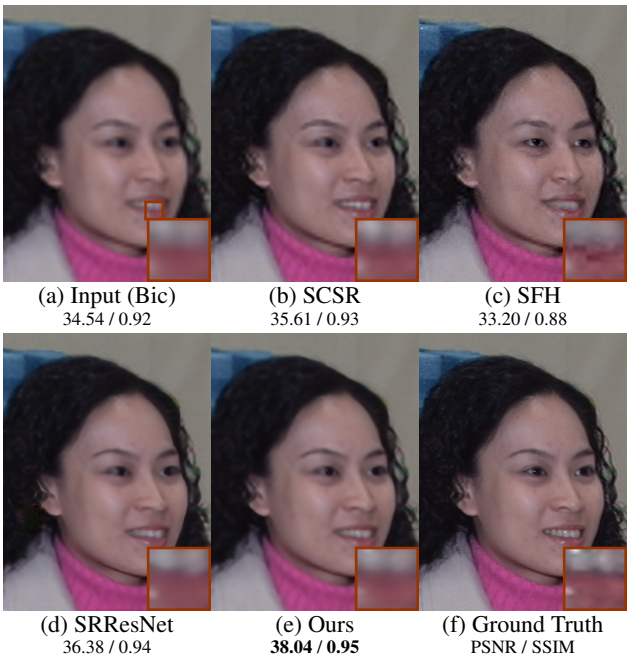


Figure 6: Qualitative evaluation for 4× upsampled face images in Multi-PIE pose dataset.

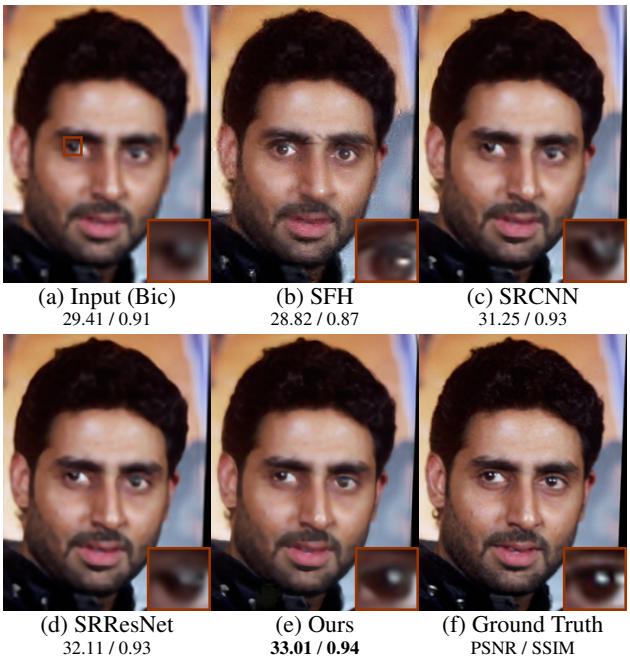


Figure 7: Qualitative evaluation for 4× upsampled face images in PubFig dataset.

high quality global appearance and facial details shown in (e).

5 Concluding Remarks

We propose a FH method named LCGE which integrates global appearance modeling and local feature matching. Different from existing FH methods which adopt handcrafted features for patch matching, LCGE generates deep facial components to narrow down the gap between LR input and HR correspondences. As such, the facial texture is enriched, which eases the matching difficulty. Then fine grained facial structure can be effectively extracted and their details are transferred back to generate the output result. Extensive experiments demonstrate the effectiveness of the proposed LCGE method compared with state-of-the-art approaches.

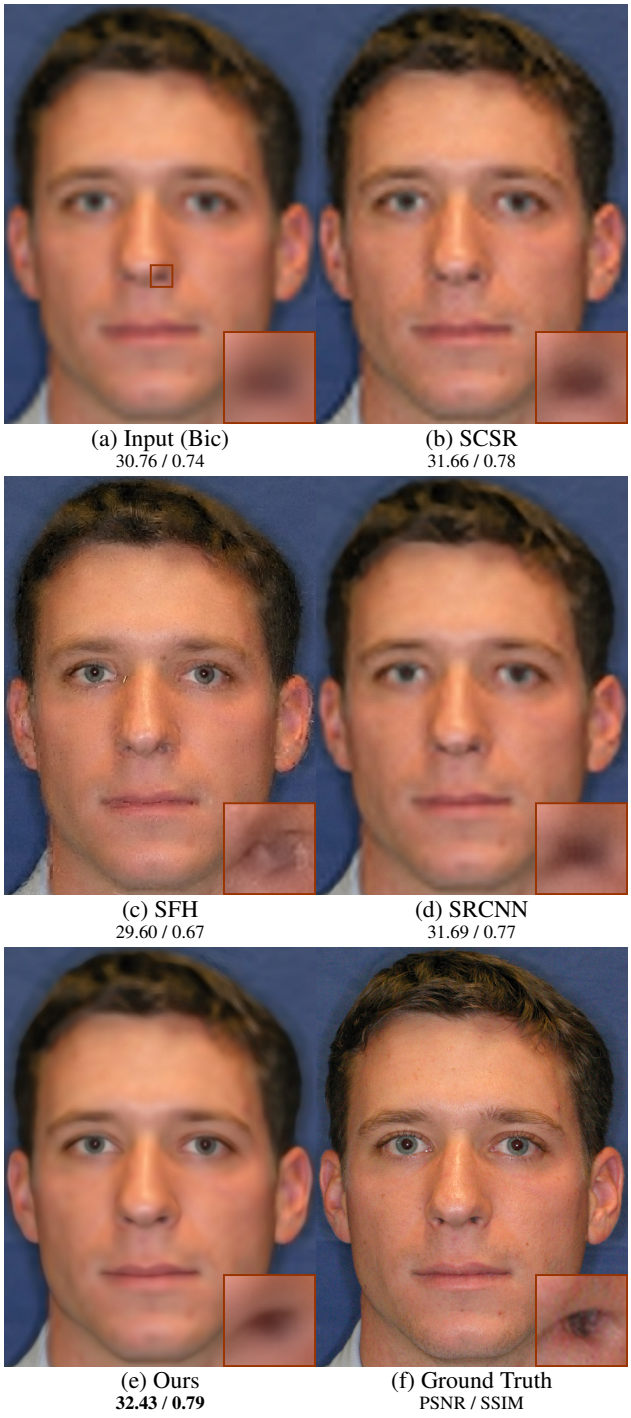


Figure 8: Qualitative evaluation for 10× upsampled face image in Multi-PIE HR dataset.

References

[Dong *et al.*, 2015] Chao Dong, Chen Change Loy, Kaiming He, and Xiaoou Tang. Image super-resolution using deep convolutional networks. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2015.

[Eisemann and Durand, 2004] Elmar Eisemann and Frédo Durand. Flash photography enhancement via intrinsic relighting. *ACM Transactions on Graphics (SIGGRAPH)*, 2004.

[Girshick *et al.*, 2014] Ross Girshick, Jeff Donahue, Trevor Darrell, and Jitendra Malik. Rich feature hierarchies for accurate object detection and semantic segmentation. In *IEEE conference on computer vision and pattern recognition*, 2014.

[Gross *et al.*, 2010] Ralph Gross, Iain Matthews, Jeffrey Cohn, Takeo Kanade, and Simon Baker. Multi-pie. *Image and Vision Computing*, 2010.

[Gunturk *et al.*, 2003] Bahadir K Gunturk, Aziz U Batur, Yucel Altunbasak, Monson H Hayes, and Russell M Mersereau. Eigenface-domain super-resolution for face recognition. *IEEE Transactions on Image Processing*, 2003.

[He *et al.*, 2013] Kaiming He, Jian Sun, and Xiaoou Tang. Guided image filtering. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2013.

[Hertzmann *et al.*, 2001] Aaron Hertzmann, Charles E Jacobs, Nuria Oliver, Brian Curless, and David H Salesin. Image analogies. *ACM Transactions on Graphics (SIGGRAPH)*, 2001.

[Jia and Gong, 2008] Kui Jia and Shaogang Gong. Generalized face super-resolution. *IEEE Transactions on Image Processing*, 2008.

[Kim *et al.*, 2016] Jiwon Kim, Jung Kwon Lee, and Kyoung Mu Lee. Deeply-recursive convolutional network for image super-resolution. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2016.

[Kumar *et al.*, 2009] Neeraj Kumar, Alexander C Berg, Peter N Belhumeur, and Shree K Nayar. Attribute and simile classifiers for face verification. In *IEEE International Conference on Computer Vision*, 2009.

[Ledig *et al.*, 2017] Christian Ledig, Lucas Theis, Ferenc Huszár, Jose Caballero, Andrew Cunningham, Alejandro Acosta, Andrew Aitken, Alykhan Tejani, Johannes Totz, and Zehan Wang. Photo-realistic single image super-resolution using a generative adversarial network. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2017.

[Liu *et al.*, 2005] Wei Liu, Dahua Lin, and Xiaoou Tang. Hallucinating faces: Tensorpatch super-resolution and coupled residue compensation. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2005.

[Liu *et al.*, 2007] Ce Liu, Heung-Yeung Shum, and William T Freeman. Face hallucination: Theory and practice. *International Journal of Computer Vision*, 2007.

[Liu *et al.*, 2011] Ce Liu, Jenny Yuen, and Antonio Torralba. Sift flow: Dense correspondence across scenes and its applications. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2011.

[Lowe, 2004] David G Lowe. Distinctive image features from scale-invariant keypoints. *International Journal of Computer Vision*, 2004.

[Ma *et al.*, 2010] Xiang Ma, Junping Zhang, and Chun Qi. Hallucinating face by position-patch. *Pattern Recognition*, 2010.

[Petschnigg *et al.*, 2004] Georg Petschnigg, Maneesh Agrawala, Hugues Hoppe, Richard Szeliski, Michael Cohen, and Kentaro Toyama. Digital photography with flash and no-flash image pairs. *ACM Transactions on Graphics (SIGGRAPH)*, 2004.

[Song *et al.*, 2014] Yibing Song, Linchao Bao, Qingxiong Yang, and Ming-Hsuan Yang. Real-time exemplar-based face sketch synthesis. In *European Conference on Computer Vision*, 2014.

[Tappen and Liu, 2012] Marshall Tappen and Ce Liu. A bayesian approach to alignment-based image hallucination. In *European Conference on Computer Vision*, 2012.

[Wang and Tang, 2005] Xiaogang Wang and Xiaoou Tang. Hallucinating face by eigentransformation. *IEEE Transactions on Systems, Man, and Cybernetics, Part C: Applications and Reviews*, 2005.

[Wang *et al.*, 2004] Zhou Wang, Alan Conrad Bovik, Hamid Rahim Sheikh, and Eero P Simoncelli. Image quality assessment: from error visibility to structural similarity. *IEEE Transactions on Image Processing*, 2004.

[Wang *et al.*, 2014] Nannan Wang, Dacheng Tao, Xinbo Gao, Xuelong Li, and Jie Li. A comprehensive survey to face hallucination. *International Journal of Computer Vision*, 2014.

[Wang *et al.*, 2017] Nannan Wang, Xinbo Gao, Leiyu Sun, and Jie Li. Bayesian face sketch synthesis. *IEEE Transactions on Image Processing*, 2017.

[Yang *et al.*, 2010] Jianchao Yang, John Wright, Thomas Huang, and Yi Ma. Image super-resolution via sparse representation. *IEEE Transactions on Image Processing*, 2010.

[Yang *et al.*, 2013] Chih-Yuan Yang, Sifei Liu, and Ming-Hsuan Yang. Structured face hallucination. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2013.

[Yu and Porikli, 2016] Xin Yu and Fatih Porikli. Ultra-resolving face images by discriminative generative networks. In *European Conference on Computer Vision*, 2016.

[Zhou *et al.*, 2015] Erjin Zhou, Haoqiang Fan, Zhimin Cao, Yuning Jiang, and Qi Yin. Learning face hallucination in the wild. In *AAAI Conference on Artificial Intelligence*, 2015.