Int J Comput Vis
https://doi.org/10.1007/s11263-017-1044-4

Hallucinating Compressed Face Images

Chih-Yuan Yang[1] · **Sifei Liu**[1] · **Ming-Hsuan Yang**[1]

Received: 2 December 2015 / Accepted: 12 September 2017

Abstract A face hallucination algorithm is proposed to generate high-resolution images from JPEG compressed low-resolution inputs by decomposing a deblocked face image into structural regions such as facial components and non-structural regions like the background. For structural regions, landmarks are used to retrieve adequate high-resolution component exemplars in a large dataset based on the estimated head pose and illumination condition. For non-structural regions, an efficient generic super resolution algorithm is applied to generate high-resolution counterparts. Two sets of gradient maps extracted from these two regions are combined to guide an optimization process of generating the hallucination image. Numerous experimental results demonstrate that the proposed algorithm performs favorably against the state-of-the-art hallucination methods on JPEG compressed face images with different poses, expressions, and illumination conditions.

Keywords Face hallucination · Super resolution · JPEG compression · Image denoising · Landmark points

Communicated by T.E. Boult.

✉ Ming-Hsuan Yang
mhyang@ucmerced.edu

Chih-Yuan Yang
cyang35@ucmerced.edu

Sifei Liu
sliu32@ucmerced.edu

[1] Electrical Engineering and Computer Science, School of Engineering, University of California at Merced, 5200 North Lake Road, Merced, CA 95343, USA

1 Introduction

As a domain-specific single image super resolution problem, face hallucination has been widely studied (Wang et al. 2014; Baker and Kanade 2002; Wang and Tang 2005; Jia and Gong 2005; Liu et al. 2007; Park and Lee 2008; Yang et al. 2010; Ma et al. 2010; Tappen and Liu 2012; Yang et al. 2013; Liang et al. 2014; Jiang et al. 2014; Liu and Yang 2014) but existing methods usually operate on the premise of uncompressed input images without blocky artifacts or noise. However, in real-world scenarios, large amounts of images are stored in compressed formats as trade-offs between visual quality and storage size. As such, existing face hallucination methods are not expected to generate high-quality images from compressed inputs.

Among numerous image compression methods, the widely used JPEG method reduces image data by quantizing high-frequency signals based on human visual perception. The JPEG compression method computes discrete cosine transform (DCT) coefficients of each non-overlapping 8×8 block in an image, and quantizes the DCT coefficients with emphasis on retaining the visual information of high-frequency signals. When a standard quantization table is used, a quality index Q between 1 and 100 indicates the retained information after quantization. For color images, the JPEG compression method operates in the YCbCr color space and quantizes each channel independently.

To restore missing details in JPEG compressed images, numerous schemes have been proposed in the literature (Liu and Bovik 2002; Buades et al. 2005; Figueiredo et al. 2006; Kim and Kwon 2008; Singh et al. 2007; Mairal et al. 2009; Zhai et al. 2008; Choi et al. 2013; Li et al. 2014). Despite being able to reduce JPEG compression noise, existing methods inevitably remove high-frequency details which are important for face hallucination. Thus, a two-step approach

Published online: 08 December 2017

with deblocking and super resolution has to exploit examples to reconstruct the details. Although a number of face hallucination methods (Wang and Tang 2005; Liu et al. 2007; Ma et al. 2010) are able to generate high-resolution (HR) images by averaging HR exemplars globally or locally, the main drawback of those approaches is the lack of flexibility. The test and exemplar images need to be of the same pose, under the same illumination condition, and with similar facial expressions. Otherwise, the generated HR images may contain apparent artifacts because exemplars cannot be effectively matched.

Landmark-based face hallucination methods (Yang et al. 2013; Liu and Yang 2014) are developed to handle various poses and expressions. Typically, landmark points in both exemplar and test images are detected and matched to determine similar low-resolution (LR) facial components. As these methods align exemplar facial components with test ones, the reconstructed HR image tends to be semantically correct and visually appealing. However, most landmark-based face hallucination methods do not take compression noise into account. While a recent method (Liu and Yang 2014) shows that LR compressed inputs can be well reconstructed in controlled environments, it is computationally expensive and limited to a few poses. Thus, it is of great importance to develop efficient and effective hallucination algorithms to process compressed face images with various poses, expressions, and lighting conditions.

In this paper, we propose an effective and efficient hallucination algorithm that deals with compressed face images in various poses, expressions, and illumination conditions. The proposed algorithm differs from the prior work (Liu and Yang 2014) in several aspects. First, face landmarks can be more accurately and robustly located in images. Second, various head orientations and illumination conditions are taken into consideration. Third, to align landmarks, a more efficient approach is proposed in this paper through a close form solution rather than solving an optimization problem in the prior work. Fourth, the proposed method significantly reduces computational load by sorting exemplar images according to the similarity of landmark point coordinates so that adequate exemplar images can be efficiently determined at run-time. As such, the proposed method can effectively reduce JPEG compression noise and efficiently generate better enlarged face images.

While most hallucination methods operate on upright frontal faces with neutral expressions in tightly cropped images, the proposed algorithm is able to generate high-quality results with various poses, illuminations, expressions and background clutter. Numerous experimental results show that the proposed algorithm generates favorable results against the state-of-the-art methods both qualitatively and quantitatively.

2 Related Work and Problem Context

2.1 Face Hallucination

As a domain-specific super resolution problem, numerous face hallucination methods have been proposed based on patch prediction (Baker and Kanade 2002; Ma et al. 2010), constrained subspace reconstruction (Wang and Tang 2005; Liu et al. 2007; Yang et al. 2010), and transfer of exemplars (Tappen and Liu 2012; Yang et al. 2013). Existing methods often assume the input face images are tightly cropped, free of compression noise, and in a fixed pose under controlled lighting conditions (Baker and Kanade 2002; Wang and Tang 2005; Liu et al. 2007; Yang et al. 2010; Ma et al. 2010; Yang et al. 2013).

2.2 Patch Prediction

Face hallucination methods typically extract features from LR patches to predict the HR patch features based on a set of exemplar images (Baker and Kanade 2002). The main limitation of these methods is the ambiguity among similar LR patches, i.e., two similar LR patches generated from two significantly different HR patches. To alleviate the ambiguity problem, location information is exploited as constraints to select patches (Ma et al. 2010). However, the location priors perform best for a homogeneous face dataset containing subjects of the same race and similar ages without significantly changing poses. Otherwise, the exemplar face images are likely to be dissimilar and patches cannot be well matched based on locations for generating HR outputs.

2.3 Constrained Subspace Reconstruction

Numerous methods have been developed to exploit global constraints of face patches for matching LR and HR images based on principle component analysis (Liu et al. 2007) or non-negative matrix factorization (Yang et al. 2008) schemes. These methods transfer high-frequency details contained in HR exemplar patches via a Markov network (Liu et al. 2007) or sparse dictionary (Yang et al. 2008). The global and local constraints help reconstruct high-quality regions (e.g., eyes and eyebrows) as the components can be aligned in the training images and effectively modeled by a linear subspace. However, linear subspace representations are less effective for modeling contours. As training face images are usually aligned at eye locations, facial contours are misaligned due to shape changes caused by head pose variations. Consequently, the contours reconstructed by these methods usually contain significant artifacts.

2.4 Transfer of Exemplars

HR face images can be generated by transferring high-frequency details of HR exemplar images aligned by optical flows (Tappen and Liu 2012) or facial landmarks (Yang et al. 2013). These methods are able to handle various poses and expressions due to alignments, and the reconstructed facial components are of high quality. In addition, because landmark points can be used to segment a face into regions of various types for further processing, this approach exploits statistical priors to predict HR patches for regions of certain types. While facial components are reconstructed by exemplars, contours and textures are typically restored by generic priors.

2.5 Upsampling Compressed Images

To address compression artifacts in generic images, one super resolution method has been proposed to regularize inputs by partial differential equations for reconstructing contours (Xiong et al. 2010). This method upsamples the regularized images by bicubic interpolation, and then sharpens upsampled edges by transferring high-frequency details from exemplar patches. Although the edges on the generated HR images appear sharp with few blocky effects, the textured regions are over-smoothed. As such, this method performs well for images full of sharp edges but without rich textures such as cartoons or sketches.

2.6 Denoising JPEG Images

Numerous algorithms have been developed to remove blocky and ringing artifacts of JPEG compressed images which are caused by reduced high- frequency signals lost in the block-based quantization. Since blocky artifacts occur at fixed positions, an intuitive way to reduce artifacts is to detect and smooth discontinuous intensities along block boundaries (Liu and Bovik 2002; Zhai et al. 2008). However, these approaches may generate over-smoothed results when only considering block boundaries rather than image structures. A method to recover missing JPEG edges is developed by applying learned regression functions on compressed images and suppressing noise along contours (Kim and Kwon 2008). As this method does not separate edges and textures, it tends to generate sharp edges but smooth textures. Therefore, an effective way to reduce JPEG noise but reserve high-frequency details is to adapt the denoising regions to the image structures (Foi et al. 2007): for smooth regions, averaging a group of similar neighboring pixels; for edges, averaging a number of pixels along the contours; for textures, keeping unchanged in order to avoid altering high-frequency details.

3 Proposed Algorithm

We generate HR face images from LR JPEG compressed inputs by exploiting facial structures identified by landmarks. We convert a color image into the YCbCr space and apply the proposed hallucination on the Y channel. The contents in the other two channels are upsampled through bicubic interpolation and merged with the HR results from the Y channel to generate the output image.

In this work, we presuppose that a compressed grayscale LR image L is generated from a HR image H by

$$L = J((H \otimes G_\sigma) \downarrow_s), \tag{1}$$

where $\otimes$ denotes a convolution operator, G_σ is a Gaussian kernel of width σ, $\downarrow_s$ is a downsampling operator of a scaling factor s, and J is a JPEG compression function. To reconstruct a HR image H from a LR input L, we generate a set of gradient maps U to guide the output HR image H^* by

$$H^* = \underset{H}{\operatorname{argmin}} \|\nabla H - U\|^2 \ \text{ s.t. } \ (H \otimes G_\sigma) \downarrow_s = D(L), \tag{2}$$

where ∇ is a gradient operator, and D is a JPEG deblocking function. The operator ∇ generates a set of 8 gradient maps from an input image by computing differences between a central pixel and its 8 adjacent pixels. The guiding gradient map set U is generated by merging two sets U_c and U_g (for structured facial components and non-structured generic regions, respectively) by

$$U^k(i, j) = w(i, j)U_c^k(i, j) + (1 - w(i, j))U_g^k(i, j), \tag{3}$$

where w is a weight map with values between 0 and 1, (i, j) are the coordinate indexes of pixels, and k is the order index of a gradient map in its belonging set. The main steps of the proposed algorithm are shown in Fig. 1, and the methods for generating these gradient maps are presented in the following sections. For ease of understanding, we present the formulations first and leave technical details of how to relax and solve (2) in the appendix.

4 Generating Gradient Maps

To generate the gradient map set U_c and the corresponding weight map w, we exploit landmark points to align and enclose facial components. We use a large set of grayscale HR exemplar face images S_0 with labels for glasses and annotated landmark point coordinates, head orientations, and illumination conditions. Those labels and annotated data are used to select a small subset of S_0 for fast and effective exemplar search of facial components.

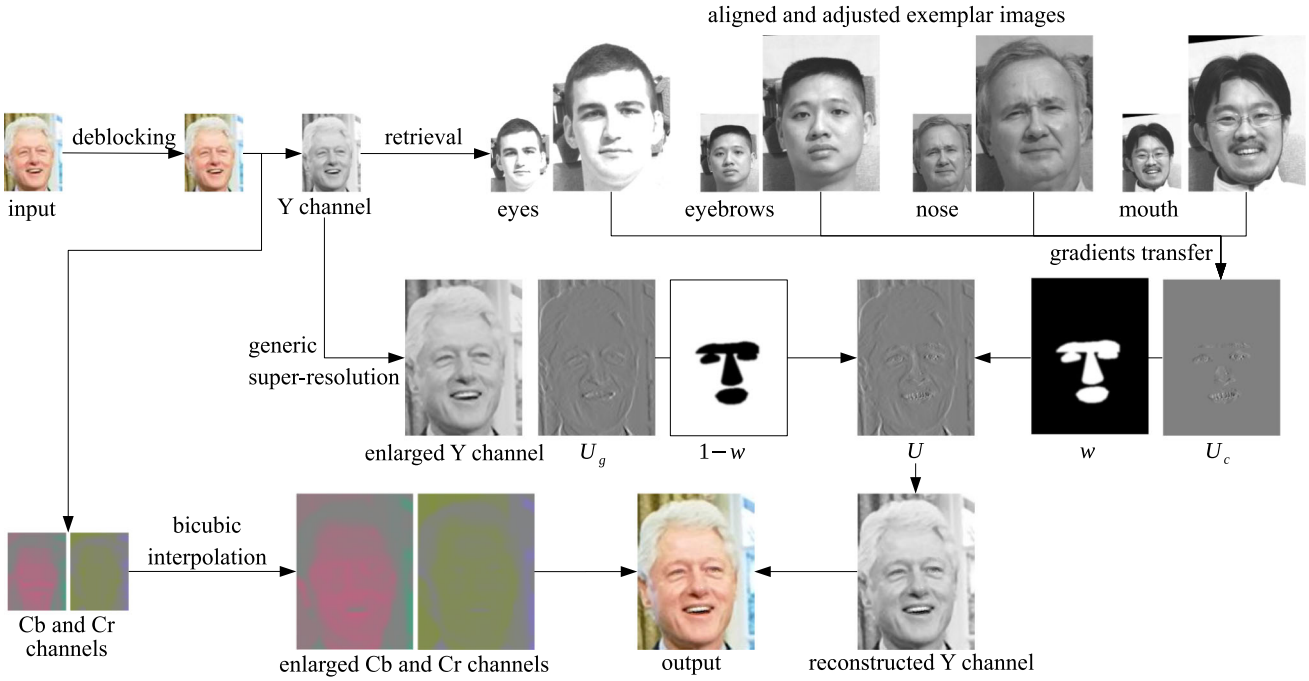

Fig. 1 Flow chart. Given a JPEG low-resolution input, we generate a high-quality high-resolution image while removing artifacts. We estimate the pose and illumination condition to determine the set of exemplars best resembling the input and transfer gradients of the corresponding facial components to a set of high-resolution gradient maps U_c. Non-structured regions—background, hair and skin areas—are upsampled by a generic super resolution algorithm and generate a set of gradient maps U_g. The gradient map sets U_c and U_g are integrated via a weight map w, to guide an optimization process to generate a grayscale high-resolution image, which is then combined with upsampled chromatic channels to generate the output image

4.1 Determining Exemplar Sets via Landmarks

In the training phase, we divide the set of exemplars S_0 into subsets based on the annotated head orientations and illumination conditions. Using (1) without the compression function J, we generate the LR sets and compute the mean of each set where images are aligned on subjects' eyes as shown in Fig. 2. For each head orientation, we set a rectangular region covering eyes and upper lip in which facial appearance is least altered by hair, facial expressions, and glasses. We normalize the intensities in the rectangles as features to describe illumination conditions, denoted as v_i^p where p and i range from 1 to the numbers of head orientations and illumination conditions.

Given a LR image L in the test phase, we first reduce its compression artifacts by applying a deblocking method D (Foi et al. 2007). We detect a face in $D(L)$, localize its facial landmark points, and estimate the head orientation o using the IntraFace algorithm (Xiong and la Torre 2013). Among all head orientation $\{p\}$ in S_0, we pick up the head orientation $\hat{p}$ which is most similar to the test one o, and disregard all other p. Next, we align $D(L)$ as in the training phase to extract features and determine the nearest neighbor

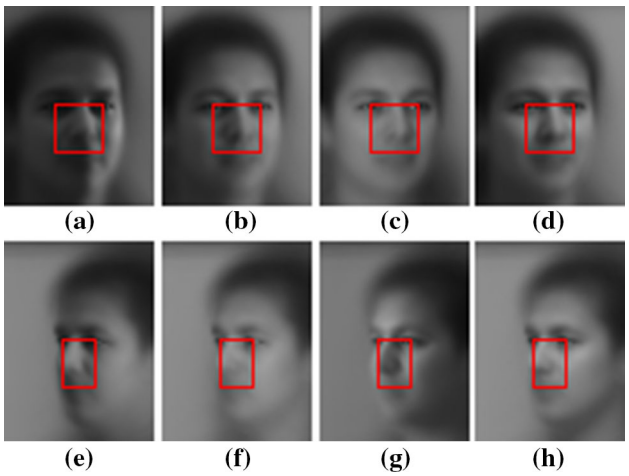

Fig. 2 Mean face images under different poses and illumination conditions. **a**–**d** the pose of roll 0°, yaw 15°, pitch 0°. **e**–**f** the pose of roll 0°, yaw −45°, pitch 0°. The number below each image indicates the illumination condition. The red rectangles show the regions used to compute features. **a** #2. **b** #5. **c** #10. **d** #16. **e** #2. **f** #5. **g** #10. **h** #16 (Color figure online)

$\hat{i}$ in the remaining choice $\{v_i^{\hat{p}}\}$. Let S_1 denote the restricted search space, i.e., all exemplar images contained in the subset used to compute $v_{\hat{i}}^{\hat{p}}$.

Fig. 3 An aligned and adjusted exemplar image. **a** A test image and the landmark points to illustrate the anchors, which are marked in red for the mouth component, and in blue for the other three. **b** An exemplar image shown in a different scale from (**a**)(**c**). **c** The result of (**b**) aligned and adjusted based on the mouth component. Note the intensity values of the aligned components are higher than those in (**b**) and similar to the pixels in (**a**) (Color figure online)

To find effective exemplars efficiently, we propose an align-and-adjust approach which operates on S_1 for four facial components: two eyes, two eyebrows, one nose, and one mouth. The four components are treated independently but in the same manner as described below. First, we define anchors as the two farthest apart landmark points belonging to a component, as shown in Fig. 3. Let (x_i^t, y_i^t) and (x_i^e, y_i^e) denote the coordinates of the anchors on the test and exemplar images, $i \in [1, 2]$ because the number of anchor points are two. We align a group of exemplar landmark points onto the test one using the two anchors through a non-reflective similarity transformation,

$$\begin{bmatrix} x' \\ y' \\ 1 \end{bmatrix} = \begin{bmatrix} \lambda\cos\theta & \lambda\sin\theta & x_1^t \\ -\lambda\sin\theta & \lambda\cos\theta & y_1^t \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} 1 & 0 & -x_1^e \\ 0 & 1 & -y_1^e \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} x \\ y \\ 1 \end{bmatrix} \tag{4}$$

where θ and λ are the angle difference and distance ratio

$$\theta = \arctan\left(\frac{y_2^t - y_1^t}{x_2^t - x_1^t}\right) - \arctan\left(\frac{y_2^e - y_1^e}{x_2^e - x_1^e}\right), \tag{5}$$

$$\lambda = \frac{\sqrt{(y_2^t - y_1^t)^2 + (x_2^t - x_1^t)^2}}{\sqrt{(y_2^e - y_1^e)^2 + (x_2^e - x_1^e)^2}}. \tag{6}$$

Let $(x_j'^e, y_j'^e)$ denote the transformed exemplar landmark points using (4), j from 1 to the number of landmark points belonging to a facial component. We compute the sum of squared distances for each exemplar image

$$s = \sum_j (x_j^t - x_j'^e)^2 + (y_j^t - y_j'^e)^2. \tag{7}$$

We define S_2 as the subset of S_1 containing N exemplar images with the N smallest s values. For the three components influenced by glasses (i.e., eyes, eyebrows, and noses), we use the labels for glasses to prevent images from being included in the set S_2 as the gradients of glasses may cause artifacts in the reconstructed images.

4.2 Adjusting Illumination Variations via Masks

To account for lighting variations between components of test and exemplar images, we create a LR weight map covering each facial component with a polygon. For the eyes which are surrounded by their landmark points, we use 6 landmark points for each eye to create two polygons covering the components. For the mouth, we create the polygon based on 12 surrounding landmark points. For the nose, the polygon is created based on 5 landmark points at the bottom and 1 at the top. Since there are 10 landmark points localized at the top of eyebrows, we connect them as the upper boundary of the mask polygon of eyebrows; the bottom of the polygon is determined based on a line between the rightmost and leftmost landmarks (see Fig. 1). For pixels covered by a polygon, the initial weight values are set to 1, and the other ones are set to 0. To create soft boundaries of the weight map, we apply a low-pass filter on the initial map using a Gaussian filter with a value of kernel width 0.4.

We treat the weight map as a mask in order to adjust pixel intensities of exemplar components. For each individual component of a test image, we use the mask to compute the mean and standard deviation of the covered pixels as μ_t and σ_t. After resampling images of S_2 using parameters computed from (5) and (6), we downsample them into LR using (1) without J. Using the same mask as for μ_t and σ_t, we compute the mean and standard deviation of an aligned LR exemplar image as μ_e and σ_e and adjust its pixel intensities i by

$$i' = (i - \mu_e) \cdot \frac{\sigma_t}{\sigma_e} + \mu_t. \tag{8}$$

Afterwards, we determine the best exemplar component among the N candidates. We extract the gradient features under the component mask from the test image and N adjusted LR exemplar images, among which we find the nearest neighbor under the l^2-norm as the best exemplar. Next, we generate the landmark points of the test image in HR by multiplying the LR coordinates with the scaling factor. The best exemplar in HR is aligned and adjusted in the same manner as for the LR one. Then, we create the HR component mask similarly as the LR one with a large value of the Gaussian kernel width, 1.6, for HR images.

We transfer the gradients of the aligned, adjusted HR exemplar image to U_c under the HR mask. The values of the four component masks are merged to generate the weight map w used in (3). We arrange the priority of the four components from high to low as the eyes, eyebrows, nose, and mouth since the eyes are psychologically more important

than other components. If some masks overlap, we replace the gradients of a lower priority component with a higher one (see Fig. 1).

Different from existing methods (Yang et al. 2013; Liu and Yang 2014) which align all exemplar images, the proposed algorithm requires a lighter computational load for two reasons. First, the proposed method computes the transformation matrix in a close form (4) rather than solving an optimization problem (Yang et al. 2013; Liu and Yang 2014). Second, the proposed method only re-samples N images, which is a small subset of the entire exemplar image set (similar to S_1 in this work) used in prior work (Yang et al. 2013; Liu and Yang 2014). In addition, the proposed method adjusts the brightness of exemplar facial components in order to get better gradients, which is not exploited in prior work (Yang et al. 2013; Liu and Yang 2014).

4.3 Merging Gradient Maps

We generate the weight map w by merging four masks of facial components, the facial component gradients U_c from four exemplar images, and the non-facial component gradients U_g from an image enlarged by a generic super resolution method (Timofte et al. 2014). Figure 1 shows an example of the three types of data. We merge w, U_c, and U_g into the final gradient map set U through (3).

5 Experimental Validation

5.1 Experimental Setups

5.1.1 Exemplar and Test Images

To demonstrate the proposed algorithm's ability to handle generalization, we use the images in the Multi-PIE (Gross et al. 2008) and PubFig (Kumar et al. 2009) datasets as our exemplar and test sets, respectively. The identities in these two sets do not overlap and their appearances differ in numerous aspects—poses, illuminations, expressions, hair styles, and cosmetics. The Multi-PIE dataset contains 2514 image sets in each contained 300 images taken by 15 cameras under 20 controlled lighting conditions. We select a subset as our exemplar set which consists of images taken by 7 frontal cameras (ID 19_0, 04_1, 05_0, 05_1, 14_0, 13_0, and 08_0) and under 18 lighting conditions (01 to 18) to cover a wide range of frontal views (−45 to +45 degrees of horizontal rotation) and illumination conditions. We exclude the other 8 cameras and 2 lighting conditions because their side view angles and low illumination significantly differ from the test images. We manually generate the labels for glasses and refine the landmark points generated by the IntraFace algorithm (Xiong and la Torre 2013). Table 1 lists the IDs of the 7 cameras with the corresponding head poses, and the numbers of images. The total number of exemplar images used in the experiments is 292842, i.e., 16269 (sum of Table 1) multiplied by 18 lighting conditions.

Table 1 Numbers of exemplar images under 7 cameras and head poses in yaw angles (with zero degree in pitch or roll)

ID	19_0	04_1	05_0	05_1	14_0	13_0	08_0
Angle	-45°	-30°	-15°	0°	+15°	+30°	+45°
#	1954	2432	2511	2514	2480	2400	1978

Significantly different from the Multi-PIE dataset created in a controlled environment, the PubFig dataset contains face images of celebrities in various sizes and quality. From it, 835 images are detected using the Adaboost-based detector (Viola and Jones 2004) without further manual processing. We note existing methods typically operate on tightly cropped face images (i.e., only the main facial parts are present without hair or background pixels). In this work, we evaluate the proposed algorithm on both images containing background pixels or not (i.e., images without or with tight cropping). For the former type, each face image is enclosed by a rectangle of 320 × 240 pixels covering the whole head including hair and background regions. The center of the rectangle is set to the middle of the two eyes. If the detected face images do not fit the rectangles (i.e., too large or too small), they are excluded from the test set because two evaluated methods (LSF and MQZ) in the experiments require test images in a fixed size as large as the exemplar images. Although the proposed method is not restricted by the size due to its flexibility to resize exemplar images using facial landmarks, for fair comparison we use the same test images for all evaluated methods. For the latter type in which hair and background regions are excluded, each face image is cropped by a rectangle which encloses only the landmark points of four facial components for evaluation.

The test set contains 100 randomly selected face images from the PubFig dataset. All test images are shown at http://faculty.ucmerced.edu/mhyang/project/FHCI/.

5.1.2 Evaluated Algorithms

We evaluate the proposed algorithm against the state-of-the-art face hallucination methods. The default parameters of the SA-DCT deblocking (Foi et al. 2007) and A+ super resolution (Timofte et al. 2014) methods are used for performance evaluation. We implement the LSF (Liu et al. 2007) and MZQ (Ma et al. 2010) methods since the source code is unavailable. The source code of the YLY (Yang et al. 2013) and LY (Liu and Yang 2014) methods as well as the MATLAB built-in bicubic interpolation function are used for experiments.

Fig. 4 The test image (Barack_Obama_0086_face1) is detected from a frame in the PubFig dataset (Kumar et al. 2009). The input images are enlarged by nearest neighbor interpolation for ease of presentation. The proposed method generates clearer facial details and less noise than other methods. Results are best viewed on a high-resolution display with an adequate zoom level to show each image with at least 320 × 240 pixels for full resolution

The proposed algorithm is implemented in MATLAB with a value of Gaussian kernel width σ of 1.6, a scaling factor s of 4, and JPEG quality indices Q of 100, 75, 50, and 25 to generate LR input images. Images compressed with higher Q values indicate higher visual quality (i.e., lower compression). For each test image, a set of 100 exemplars ($N = 100$) is used for face hallucination as a trade-off between execution time and image quality. In addition, we show that the proposed algorithm is not sensitive to the number of exemplar images on the project web page.

To reconstruct the HR image H^* using (2), we set the value β of the relaxed problem (9) to 0.00625. This parameter plays an important role as it determines the effect of the transferred gradients of facial components from exemplars. A small β means the appearances of generated facial components are likely to be smooth and close to patch-based statistical averages (see also the discussion on the effects of gradients on the A+ method for generic super resolution (Timofte et al. 2014)) with higher PSNR and SSIM values. In contrast, a large β suppresses intensity constraints of test images and forces generated components to be similar to the exemplars although the PSNR and SSIM values may be lower. However, a large β preserves gradients of exemplar components such that the reconstructed faces may be visually sharper and more pleasant. We empirically set β to 6.4 in the experiments of Figs. 9, 10, and 11 to show the ability of the proposed method to reconstruct sharp facial components. The sensitivity anal-

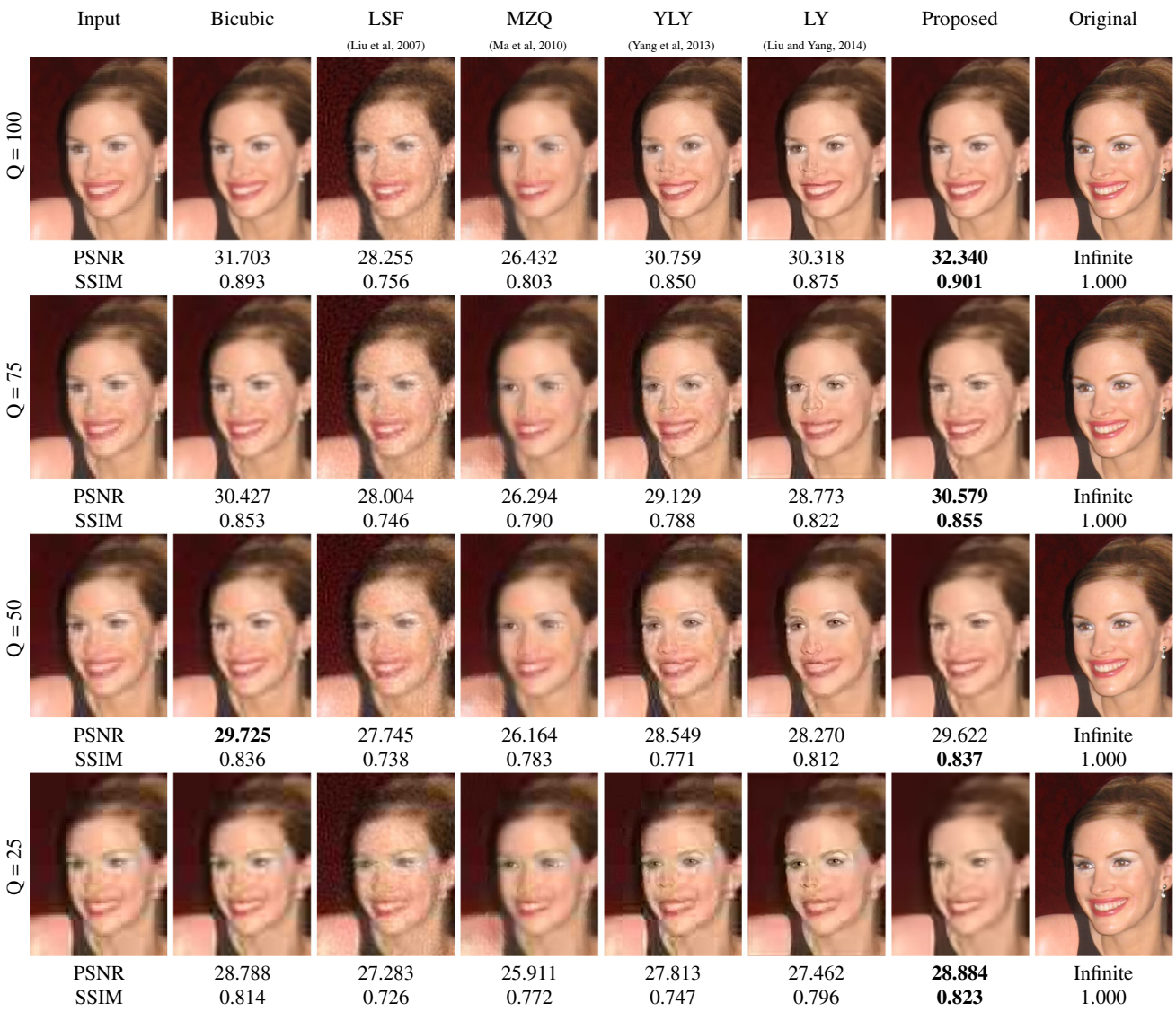

Fig. 5 The test face image (Julia_Roberts_0010_face5) is detected from a frame in the PubFig dataset (Kumar et al. 2009). The input images are enlarged by nearest neighbor interpolation for ease of presentation.The proposed method generates clearer facial details and less noise than other methods. Results are best viewed on a high-resolution display with an adequate zoom level to show each image with at least 320 × 240 pixels for full resolution

ysis on β and generated images are presented on the project web page.

5.1.3 Evaluation Metrics

We evaluate the generated images using PSNR and SSIM values (Wang et al. 2004) on grayscale channels only, as the LSF and MZQ methods only process grayscale images. Since the YLY, LY, and proposed methods all generate HR chromatic channels using bicubic interpolation, we add them to the grayscale images generated by the LSF and MZQ methods for visual comparisons. In addition to evaluation on synthesized LR images, we use real-world images from the Internet for experiments. Since there is no ground truth, images are compared in terms of visual quality. All the code and datasets are available to the public and more results are shown at http://faculty.ucmerced.edu/mhyang/project/FHCI/.

5.2 Synthetic Face Images

Figures 4 to 8 show the results generated by the evaluated methods using face images with different poses, facial expressions, and illumination conditions from the PubFig dataset. Table 2 shows the average results of the evaluated methods using 100 randomly selected face images from the PubFig dataset. Table 3 shows the average results of the central face regions of the same 100 images evaluated in Table 2.

Fig. 6 The test image (Stephen_Colbert_0060_face1) is detected from a frame in the PubFig dataset (Kumar et al. 2009). The input images are enlarged by nearest neighbor interpolation for ease of presentation. The proposed method generates clearer facial details and less noise than other methods. Results are best viewed on a high-resolution display with an adequate zoom level to show each image with at least 320 × 240 pixels for full resolution

We define the central region of a face as the compact rectangle containing all landmark points of the four facial components. The central regions consist mainly of skins and the four facial components rather than upper foreheads, ears, chins, necks, clothing, and background.

At four different compression levels, the LSF method (Liu et al. 2007) generates a few blocky patterns because it models faces in a trained subspace which learns the structure of face images and handles compression noise well. However, this scheme generates noisy and ghost artifacts because the adopted linear subspace is not effective for reconstructing face images under various poses and illumination conditions.

The MZQ scheme (Ma et al. 2010) generates smooth images at four compression levels as the transferred source patches do not contain noise. However, the generated images contain ghost artifacts as shown in Figs. 4, 7, and 8 because the assumption of transferring location-restricted LR patches does not hold. Due to the white background in Figs. 6 and 8 where a test patch is brighter than any exemplar patch at the same location in the Multi-PIE dataset, the reconstructed brightness decreases to gray. Restricted to the fixed locations of face components, the eyes in Fig. 8 are not constructed well.

The YLY approach (Yang et al. 2013) generates apparent blocky artifacts for Q values of 75, 50 and 25 because

Fig. 7 The test image (Miley_Cyrus_0561_face1) is detected from a frame in the PubFig dataset (Kumar et al. 2009). The input images are enlarged by nearest neighbor interpolation for ease of presentation. The proposed method generates clearer facial details and less noise than other methods. Results are best viewed on a high-resolution display with an adequate zoom level to show each image with at least 320 × 240 pixels for full resolution

the adopted back-projection process forces the generated HR images to exactly match the input LR images which contain compression noise. The LY method (Liu and Yang 2014) outperforms the LSF, MZQ, and YLY methods on compressed images because it is developed to deal with JPEG noise. However, this method does not handle images with large changes in poses and illumination conditions. As such, its generated images are likely to contain artifacts because exemplar facial components may not be properly selected and transferred. For example, the nose tips of face images in Fig. 5 are not generated well due to head pose variations.

The proposed algorithm requires less computational loads and generates better images than the LY method for the following reasons. First, the proposed algorithm only compares N aligned exemplar images, but the LY method searches through the entire set. The ratio of N to the entire set, in our experimental setups for the upright frontal pose, is 100:2514 (all the 2514 images reported in Table 1) such that the computational load is significantly reduced. Second, the proposed algorithm uses two anchor points to align an exemplar image rather than solving an optimization problem (as used in the LY method) to compute a transformation matrix. Third, the proposed method utilizes an efficient A+ algorithm (Timofte et al. 2014) to upsample pixels outside face regions, but the LY method uses edge priors and the PatchMatch scheme (Barnes et al. 2010) to upsample edges and textures

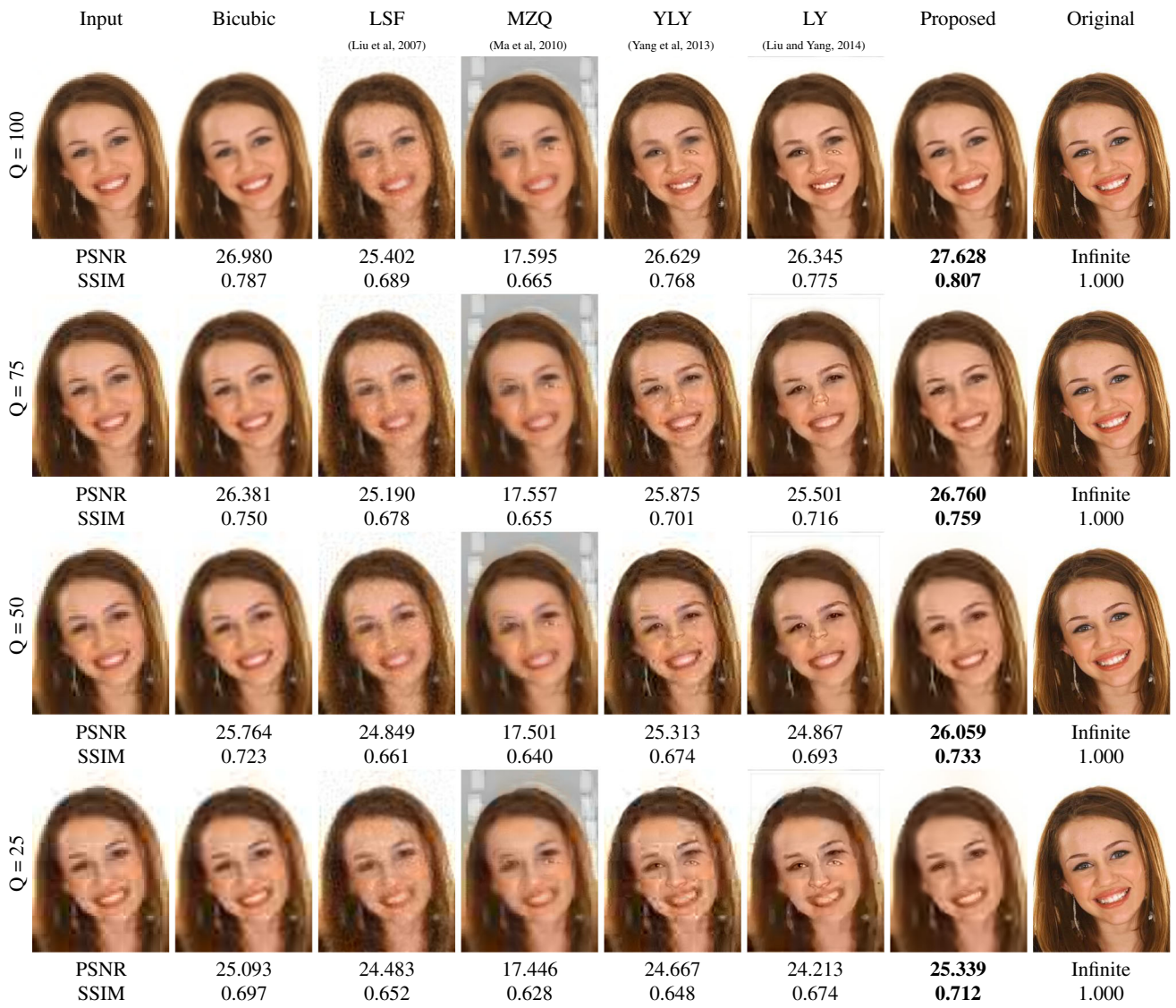

Fig. 8 The test image (Katherine_Heigl_0215_face1) is detected from a frame in the PubFig dataset (Kumar et al. 2009). The input images are enlarged by nearest neighbor interpolation for ease of presentation. The proposed method generates clearer facial details and less noise than other methods. Results are best viewed on a high-resolution display with an adequate zoom level to show each image with at least 320 × 240 pixels for full resolution

separately. As the LY method requires solving an optimization problem for edges and searching through numerous patches using the PatchMatch scheme, the computational load is significantly heavier than the proposed method. On the same machine equipped with a 3.6 GHz Quad-Core processor, it takes 20 and 900 seconds for the proposed and LY methods to enlarge a face image of 60 × 80 pixels with a scaling factor of 4.

We discuss the differences between the proposed and LY methods since both are developed for face hallucination of compressed images. The proposed method uses the state-of-the-art InfraFace (Xiong and la Torre 2013) scheme for landmark detection, which is more robust than the one (Zhu and Ramanan 2012) used in the LY approach. As shown in Fig. 8 (when Q is 100 or 25), the HR images generated by the LY method contains ghost effects around the left eyes because the landmark localization algorithm (Zhu and Ramanan 2012) does not perform well for these images. In addition, the proposed method exploits exemplar images with larger pose and illumination variations than those considered by the LY method. For example, the lighting source of Fig. 4 is from an elevated position and there exist cast shadows on the cheek regions. Similarly, the mouth regions in Figs. 5 and 8 are different from those face images in upright frontal poses. As the proposed algorithm is able to determine exemplars most similar to those in the inputs

Table 2 Quantitative evaluations of 100 PubFig images at four JPEG compression levels

JPEG Q value	Metric	Bicubic interpolation	LSF (Liu et al. 2007)	MZQ (Ma et al. 2010)	YLY (Yang et al. 2013)	LY (Liu and Yang 2014)	Modified LY	Proposed
100	PSNR	28.294	26.117	24.194	28.017	27.918	27.918	**29.590**
	SSIM	0.799	0.678	0.683	0.784	0.789	0.790	**0.830**
75	PSNR	27.356	25.860	24.042	26.761	26.697	26.708	**27.808**
	SSIM	0.750	0.664	0.670	0.704	0.723	0.724	**0.758**
50	PSNR	26.681	25.564	23.868	26.086	26.081	26.085	**26.984**
	SSIM	0.722	0.652	0.659	0.673	0.700	0.702	**0.732**
25	PSNR	25.811	25.067	23.565	25.243	25.223	25.232	**26.099**
	SSIM	0.690	0.635	0.644	0.641	0.673	0.674	**0.706**

We modify the LY method by replacing its landmark localization algorithm with the same one used by the proposed method to show a same-component comparison. The evaluations for Q as 25 are averaged from 99 images rather than 100 because the modified LY method fails to detect a face in one of the 100 images due to its high compression noise. Using the same localization algorithm, the proposed method can process that image because it utilizes a deblocking precessing step in which compression noise is reduced

Table 3 Quantitative evaluations of 100 PubFig images based only on central facial regions at four JPEG compression levels

JPEG Q value	Metric	Bicubic interpolation	LSF (Liu et al. 2007)	MZQ (Ma et al. 2010)	YLY (Yang et al. 2013)	LY (Liu and Yang 2014)	Modified LY	Proposed
100	PSNR	27.960	25.715	24.911	25.716	26.576	26.573	**29.146**
	SSIM	0.813	0.701	0.709	0.747	0.774	0.776	**0.841**
75	PSNR	26.919	25.411	24.654	25.050	25.319	25.354	**27.286**
	SSIM	0.758	0.684	0.691	0.680	0.700	0.707	**0.762**
50	PSNR	26.199	25.087	24.398	24.661	24.750	24.761	**26.423**
	SSIM	0.729	0.669	0.676	0.657	0.675	0.681	**0.734**
25	PSNR	25.320	24.593	24.006	24.055	23.861	23.884	**25.500**
	SSIM	0.695	0.649	0.657	0.627	0.644	0.650	**0.705**

All settings are the same as Table 2 and the evaluated images are shown at http://faculty.ucmerced.edu/mhyang/project/FHCI/

under varying pose and lighting conditions, the enlarged face images tend to be of higher quality. Third, to process skin and texture regions, the proposed method utilizes an effective regression-based generic super resolution algorithm (Timofte et al. 2014) which is more effective than those used in the LY scheme. Quantitative results in Tables 2 and 3 show that the proposed method performs favorably against the evaluated methods at different compression levels.

5.3 Real-World Face Images

Figures 9, 10, and 11 show the qualitative comparisons of three real-world images obtained from the Internet (i.e., the LR images are not downsampled via (1) to conduct experiments). Fig. 9 contains 11 faces of different races in similar poses and illumination conditions at a high compression ratio. While the artifacts generated by the LSF, MZQ, and YLY methods in this figure are similar to those on synthetic images, the differences between the LY and proposed algorithms are discussed here. On faces 1, 4, 8 and 10, eyes and mouths generated by the proposed method are more accurate than the LY method, especially on the eye, eyebrow, and nose regions. The noses generated by the proposed method on faces 4, 8, 9, 10, 11 are more plausible than those generated by the LY method.

Figure 10 shows the generated results from a slightly compressed image. Because landmark points of some face images cannot be detected by the localization algorithm used in the YLY and LY methods, no enlarged results are presented. The images generated by the LSF method contain compression noise, and those by the MZQ method are affected by ghost effects especially on face 2 due to the head pose and background. The YLY and LY methods perform well on faces 1, 3 and 4, but fail on face 2 and generate a fake left eye and right eyebrow on face 5. In contrast, the proposed algorithm performs well especially on the nose of face 1 and eyes of face 3.

Figure 11 contains 17 faces in various poses and illumination conditions. The proposed algorithm performs well

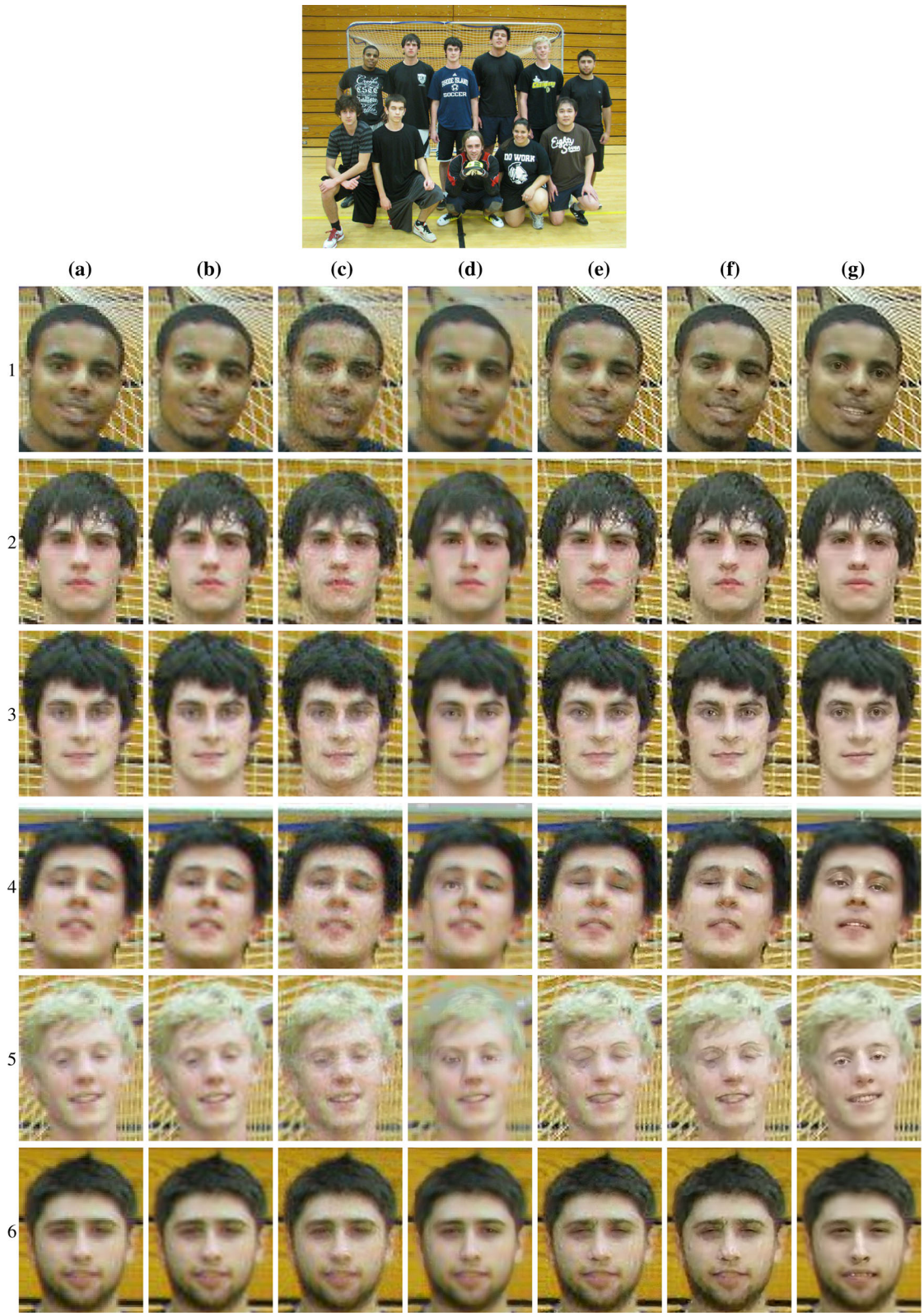

Fig. 9 A JPEG image of 960 × 720 pixels obtained from the Internet (http://img.gawkerassets.com/img/18l4dh51uk3wxjpg/original.jpg). The JPEG Q value is 67 estimated by JpgQ (Voska and Mediachancecom 2001) and the scaling factor is 4 in this experiment. No numerical evaluation is reported because there is no ground truth image. Images are best viewed on a high-resolution display with an adequate zoom level to show each image with at least 320 × 240 pixels for full resolution. **a** Input. **b** Bicubic Interpolation. **c** LSF (Liu et al. 2007). **d** MZQ (Ma et al. 2010). **e** YLY (Yang et al. 2013). **f** LY (Liu and Yang 2014). **g** Proposed

Fig. 9 continued

for all face images without obvious artifacts or amplified compression noise as generated by the LSF, MZQ, and YLY methods. The YLY and LY methods fail to generate images for faces 9 and 16 because the landmark localization algorithm reports face 9 as a profile due to the side light and does not detect the face 16 due to its makeup. The proposed algorithm reconstructs facial components better than the LY method in several regions: the left eyebrow of face 7, eyes of faces 6, 11, 12, 14, and 17, mouth of face 13, and noses of faces 6, 10 and 14.

5.4 Discussion

Although the proposed algorithm performs favorably against the state-of-the-art face hallucination methods, it is less effective in handling certain types of images. First, the success of the generated images depends on the effectiveness of localized landmark points. While they are not well localized, gradients may be incorrectly transferred and the generated images are likely to contain ghost artifacts such as the eyes in Fig. 12a, b. Second, the proposed method does not address other facial components (e.g., ears, beards, wrinkles) and foreign objects (e.g., glasses). As a result, the glasses appear unclear in Fig. 6 and blurry in Fig. 12c, d, and the beards look vague in Fig. 12e, f. Third, the exemplars in the Multi-PIE dataset do not encompass all poses such that some test images cannot be well reconstructed. The proposed method generates artifacts in Fig. 12g, h because there are not sufficient exemplar faces with the same gaze direction available in the Multi-PIE dataset. We

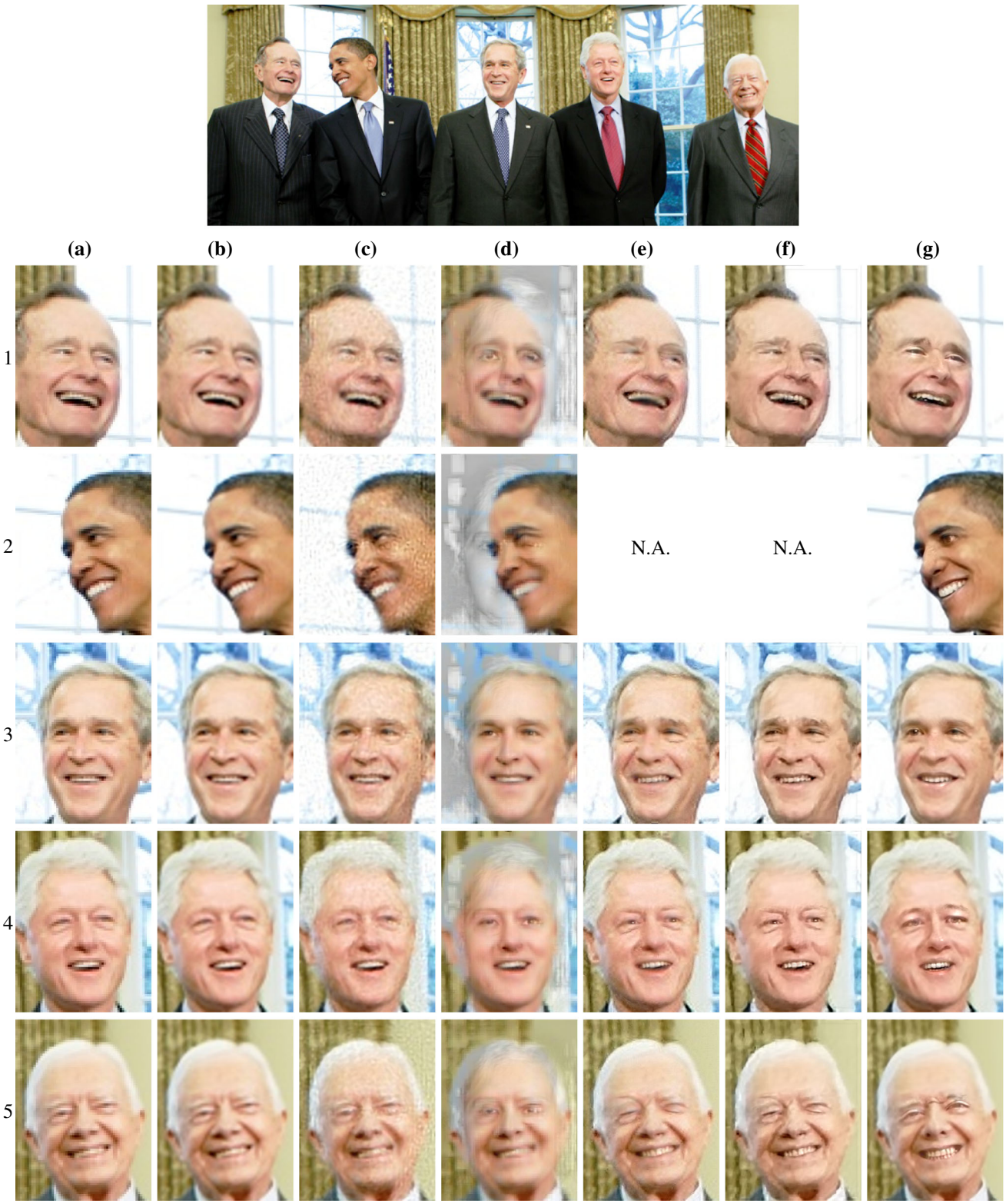

Fig. 10 A JPEG image (698 × 261 pixels) obtained from the Internet (http://blogs.reuters.com/sport/files/2009/07/pres1.jpg). The JPEG Q value is 96 estimated by JpgQ (Voska and Mediachancecom 2001) and the scaling factor is 4 in this experiment. Some results (2(**d**) and 2(**f**)) are not available due to the landmark localization algorithm used in the YLY and LY methods does not detect a face in the test image. No numerical evaluation is reported because there is no ground truth image. Images are best viewed on a high-resolution display with an adequate zoom level to show each image with at least 320 × 240 pixels for full resolution. **a** Input. **b** Bicubic Interpolation. **c** LSF (Liu et al. 2007). **d** MZQ (Ma et al. 2010). **e** YLY (Yang et al. 2013). **f** LY (Liu and Yang 2014). **g** Proposed

Fig. 11 A JPEG image (1539 × 877 pixels) obtained from the Internet (http://thewardrobedoor.com/wp-content/uploads/2014/05/x-men-days-of-future-past-cast.jpg). The JPEG Q value is 84 estimated by JpgQ (Voska and Mediachancecom 2001) and the scaling factor is 4 in this experiment. The evaluated methods are **a** bicubic interpolation **b** LSF (Liu et al. 2007) **c** MZQ (Ma et al. 2010) **d** YLY (Yang et al. 2013) **e** LY (Liu and Yang 2014) **f** the proposed scheme. No numerical evaluation is reported because there is no ground truth image. Images are best viewed on a high-resolution display with an adequate zoom level to show each image with at least 320 × 240 pixels for full resolution. **a** Bicubic Interpolation. **b** LSF **c** MZQ **d** YLY **e** LY **f** Proposed

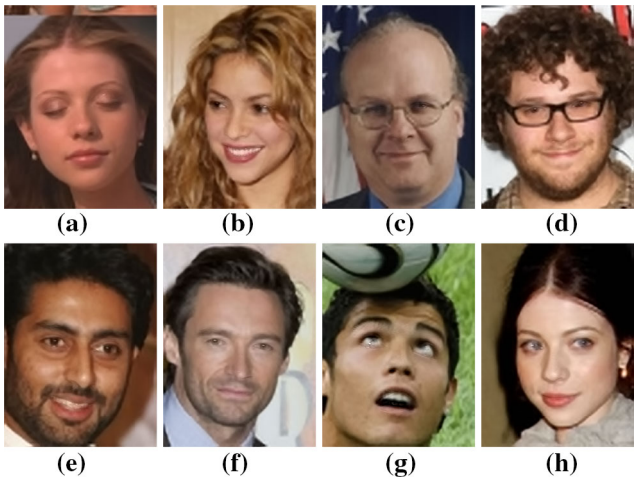

Fig. 12 Failure cases and their image IDs. **a**, **b** Wrong left eyes. **c**, **d** Blurry glasses. **e**, **f** Blurry beards. **g**, **h** Wrong pupils. Images are best viewed on a high-resolution display with an adequate zoom level to show each image with at least 320 × 240 pixels for full resolution. The ground truth images are shown in the project web page. **a** #52. **b** #87. **c** #45. **d** #82. **e** #2. **f** #30. **g** #14. **h** #55

note that all subjects in the Multi-PIE dataset gaze straight forward, but the subjects in Fig. 12g, h stare upward and leftward. Thus the proposed method is unable to find an effective pair of eyes in the Multi-PIE dataset and results in unrealistic pupils in the reconstructed images. Forth, the proposed method focuses on the brightness channel, but simply upsamples the chromatic channels, which may brings on ineffective reconstructed colors. For example, the color of the irises in Fig. 5 is originally brown, but becomes gray after reconstruction.

6 Conclusions

We address the problem of generating high-resolution face images from JPEG compressed LR inputs in this paper. The proposed algorithm integrates a deblocking method and a generic super resolution method, and exploits a large set of exemplar face images to transfer high-frequency details for image reconstruction. Experimental results show that the proposed algorithm generates high-quality images with favorable results against state-of-the-art face hallucination methods qualitatively and quantitatively.

Acknowledgements This work is supported by NSF CAREER Grant 1149783, and gifts from Adobe and Nvidia.

Appendix

Solving (2) Given a LR image $D(L)$, and gradient maps U, we generate a HR image H^* through

$$H^* = \underset{H}{\operatorname{argmin}} \|\nabla H - U\|^2 \ \text{ s.t. } \ (H \otimes G_\sigma) \downarrow_s = D(L).$$

To handle the nonlinear constraint, we relax the problem by

$$H^* = \underset{H}{\operatorname{argmin}} \ \|\nabla H - U\|^2 + \beta \|(H \otimes G_\sigma) \downarrow_s - D(L)\|^2, \tag{9}$$

where β is a weight parameter. We use the gradient descent method to solve the optimization problem.

Algorithm 1 shows the details how (9) is solved. The original energy value e is computed on Line 9, and a descent

Algorithm 1: Generating Hallucination Images

Data: Low-resolution image $D(L)$, gradient maps $\{U^k\}$, Gaussian kernel G, initial image H_0, weight β, tolerance value t, loop number l, line search step number m
Result: High-resolution image H^*

1 $H \leftarrow H_0$ Initialize
2 **for** $i \leftarrow 1$ **to** l **do**
3 $e \leftarrow \|(H \otimes G) \downarrow - D(L)\|^2 + \beta \|\nabla H - U\|^2$ //Compute the original energy value
4 $A \leftarrow \Big(\big((H \otimes G) \downarrow - D(L) \big) \uparrow \Big) \otimes G + \beta \big(\sum_1^K U^k - \mathrm{Div}(H) \big)$ //Compute the descent direction
5 **for** $j \leftarrow 1$ **to** m **do**
6 $\tau \leftarrow 2^{1-j}$ //Set the step length
7 $H' \leftarrow H - \tau A$ //Compute a new image
8 $r_j \leftarrow \beta \|\nabla H' - U\|^2 + \|(H' \otimes G) \downarrow - D(L)\|^2$ //Record the new energy value in an array
9 **end**
10 $j^* \leftarrow \underset{j}{\operatorname{argmin}}\, r_j$ //Find the index whose energy value is minimal
11 **if** $r_{j^*} < e$ **then**
12 $\tau \leftarrow 2^{1-j^*}$ //Compute the step length for updating the image
13 $H \leftarrow H - \tau A$ //Update the image
14 **end**
15 **end**
16 $H^* \leftarrow H$ //Return

direction for generating a new image is computed on Line 10, where the Div$(\cdot)$ is a divergence operator and U^k means the k-th map in U for one of the eight derivative directions. We carry out a line search on Lines 11 to 15 and record the energy values of all step lengths in an array r. We find the best step index j^* and check the energy value $r[j^*]$ on Line 17. If the new energy value $r[j^*]$ is smaller than the original energy value e, the image is updated on Lines 18 to 19.

References

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

Barnes, C., Shechtman, E., Goldman, D. B., & Finkelstein, A. (2010). The generalized PatchMatch correspondence algorithm. In *Proceedings of European conference on computer vision*.

Buades, A., Coll, B., & Morel, J. M. (2005). A non-local algorithm for image denoising. In *Proceedings of IEEE conference on computer vision and pattern recognition*.

Choi, I., Kim, S., Brown, M., & Tai, Y. W. (2013). A learning-based approach to reduce JPEG artifacts in image matting. In *Proceedings of IEEE international conference on computer vision*.

Figueiredo, M. A. T., Dias, J. B., Oliveira, J. P., & Nowak, R. (2006). On total variation denoising: A new majorization-minimization algorithm and an experimental comparison with wavelet denoising. In *Proceedings of IEEE international conference on image processing*.

Foi, A., Katkovnik, V., & Egiazarian, K. (2007). Pointwise shape-adaptive DCT for high-quality denoising and deblocking of grayscale and color images. *IEEE Transactions on Image Processing*, *16*(5), 1395–1411.

Gross, R., Matthews, I., Cohn, J., Kanade, T., & Baker, S. (2008). Multi-PIE. In *Proceedings of IEEE conference on automatic face and gesture recognition*.

Jia, K., & Gong, S. (2005). Multi-modal tensor face for simultaneous super-resolution and recognition. In *Proceedings of IEEE international conference on computer vision*.

Jiang, J., Hu, R., Wang, Z., & Han, Z. (2014). Noise robust face hallucination via locality-constrained representation. *IEEE Transactions on Multimedia*, *16*(5), 1268–1281.

Kim, K. I., & Kwon, Y. (2008). Example-based learning for single-image super-resolution and JPEG artifact removal. Max-Planck-Institut Technical Report.

Kumar, N., Berg, A. C., Belhumeur, P. N., & Nayar, S. K. (2009). Attribute and simile classifiers for face verification. In *Proceedings of IEEE international conference on computer vision*.

Li, Y., Guo, F., Tan, R. T., & Brown, M. S. (2014). A contrast enhancement framework with JPEG artifacts suppression. In *Proceedings of European conference on computer vision*.

Liang, Y., Lai, J. H., Yuen, P. C., Zou, W. W., & Cai, Z. (2014). Face hallucination with imprecise-alignment using iterative sparse representation. *Pattern Recognition*, *47*(10), 3327–3342.

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

Liu, S., & Bovik, A. C. (2002). Efficient DCT-domain blind measurement and reduction of blocking artifacts. *IEEE Transactions on Circuits and Systems for Video Technology*, *12*(12), 1139–1149.

Liu, S., & Yang, M. H. (2014). Compressed face hallucination. In *Proceedings of IEEE international conference on image processing*.

Ma, X., Zhang, J., & Qi, C. (2010). Hallucinating face by position-patch. *Pattern Recognition*, *43*(6), 2224–2236.

Mairal, J., Bach, F., Ponce, J., Sapiro, G., & Zisserman, A. (2009). Non-local sparse models for image restoration. In *Proceedings of IEEE international conference on computer vision*.

Park, J. S., & Lee, S. W. (2008). An example-based face hallucination method for single-frame, low-resolution facial images. *IEEE Transactions on Image Processing*, *17*(10), 1806–1816.

Singh, S., Kumar, V., & Verma, H. K. (2007). Reduction of blocking artifacts in JPEG compressed images. *Digital Signal Processing*, *17*(1), 225–243.

Tappen, M. F., & Liu, C. (2012). A Bayesian approach to alignment-based image hallucination. In *Proceedings of European conference on computer vision*.

Timofte, R., Smet, V. D., & Gool, L. V. (2014). A+: Adjusted anchored neighborhood regression for fast super-resolution. In *Proceedings of Asian conference on computer vision*.

Viola, P., & Jones, M. J. (2004). Robust real-time face detection. *International Journal of Computer Vision*, *57*(2), 137–154.

Voska, R., Mediachancecom. (2001). JpgQ—jpeg quality estimator. www.mediachance.com.

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

Wang, X., & Tang, X. (2005). Hallucinating face by eigentransformation. *IEEE Transactions on Systems, Man, and Cybernetics*, *35*(3), 425–434.

Wang, Z., Bovik, A., Sheikh, H., & Simoncelli, E. (2004). Image quality assessment: From error visibility to structural similarity. *IEEE Transactions on Image Processing*, *13*(4), 600–612.

Xiong, X., & la Torre, F. D. (2013). Supervised descent method and its application to face alignment. In *Proceedings of IEEE conference on computer vision and pattern recognition*.

Xiong, Z., Sun, X., & Wu, F. (2010). Robust web image/video super-resolution. *IEEE Transactions on Image Processing*, *19*(8), 2017–2028.

Yang, CY., Liu, S., & Yang, M. H. (2013). Structured face hallucination. In *Proceedings of IEEE conference on computer vision and pattern recognition*.

Yang, J., Wright, J., Huang, T., & Ma, Y. (2008). Image super-resolution via sparse representation of raw image patches. In *Proceedings of IEEE conference on computer vision and pattern recognition*.

Yang, J., Wright, J., Huang, T., & Ma, Y. (2010). Image super-resolution via sparse representation. *IEEE Transactions on Image Processing*, *19*(11), 2861–2873.

Zhai, G., Zhang, W., Yang, X., Lin, W., & Xu, Y. (2008). Efficient deblocking with coefficient regularization, shape-adaptive filtering, and quantization constraint. *IEEE Transactions on Multimedia*, *10*(5), 735–745.

Zhu, X., & Ramanan, D. (2012). Face detection, pose estimation, and landmark localization in the wild. In *Proceedings of IEEE conference on computer vision and pattern recognition*.