EDFace-Celeb-1M: Benchmarking Face Hallucination With a Million-Scale Dataset

Kaihao Zhang, Dongxu Li, Wenhan Luo, Jingyu Liu, Jiankang Deng, Student Member, IEEE, Wei Liu, Senior Member, IEEE, and Stefanos Zafeiriou, Member, IEEE

Abstract—Recent deep face hallucination methods show stunning performance in super-resolving severely degraded facial images, even surpassing human ability. However, these algorithms are mainly evaluated on non-public synthetic datasets. It is thus unclear how these algorithms perform on public face hallucination datasets. Meanwhile, most of the existing datasets do not well consider the distribution of races, which makes face hallucination methods trained on these datasets biased toward some specific races. To address the above two problems, in this paper, we build a public Ethnically Diverse Face dataset, EDFace-Celeb-1M, and design a benchmark task for face hallucination. Our dataset includes 1.7 million photos that cover different countries, with relatively balanced race composition. To the best of our knowledge, it is the largest-scale and publicly available face hallucination dataset in the wild. Associated with this dataset, this paper also contributes various evaluation protocols and provides comprehensive analysis to benchmark the existing state-of-the-art methods. The benchmark evaluations demonstrate the performance and limitations of state-of-the-art algorithms. https://github.com/HDCVLab/EDFace-Celeb-1M

Index Terms—Benchmarking, EDFace-Celeb-1M, face hallucination, face super-resolution, million-scale dataset

1 INTRODUCTION

HUMAN faces contain important identity information and are central to various vision applications, such as face alignment [1], [2], [3], face parsing [4], [5] and face identification [6], [7]. However, most of these applications require high-quality images as input and the approaches perform less favorably in low-resolution conditions. To alleviate the issue, the task of face hallucination, or face super-resolution, aims to super-resolve low-resolution face images to their high-resolution counterparts, thus facilitating effective face analysis.

As a special case of Single Image Super-Resolution (SISR) [8], [9], [10], [11], [12], [13], [14], [15], [16], [17], face hallucination is a fundamental and challenging problem in face analysis. Different from general SISR, which deals with super-resolving pixels in arbitrary scenes, the face hallucination task tackles only facial images. Therefore, the facial special prior knowledge in face images could help recover accurate face shape and rich facial details. As a result, the face hallucination methods often achieve better performance than general single image super-resolution ones in terms of higher up-scaling factors. Previous face hallucination methods [18], [19], [20], [21], [22], [23], [24], [25], [26] utilize facial priors to restore high-resolution images by approaches which are typically not end-to-end ones. Currently, a number of deep learning based methods [27], [28], [29], [30], [31], [32], [33], [34] are proposed to greatly boost the performance of the task of face hallucination, even surpassing human ability.

Is Face Hallucination Solved? First, many current deep hallucination methods evaluate their methods on non-public synthetic datasets. Especially, they typically download public datasets and synthesize pairs of low-resolution and high-resolution images, and then randomly select some faces as training and testing samples. After that, they evaluate their methods and re-train previous methods on their synthesized datasets. However, using this way for evaluation presents two obvious problems. (1) Given that the division of training/testing groups is random, the following researchers cannot strictly follow and reproduce the previous experiments like [35], [36]. (2) To verify that the new face hallucination methods outperform the previous methods, researchers have to synthesize new datasets by themselves and re-train previous methods again, which greatly increases the unnecessary workload and reduces the credibility of their results like [29], [33]. Therefore, it is not clear how these algorithms would perform on public face hallucination datasets.

On the other hand, though it is popular that the current deep face hallucination methods are evaluated on synthesized face datasets, these datasets face the problem of being biased toward specific races, and other races are significantly ignored. This scheme along with these datasets not only fails to accurately evaluate the performance of face hallucination methods, but also will raise the ethical problem. Therefore, a large-scale face hallucination dataset with relatively balanced race composition is necessary for face analysis in the community.

To address the above problems, in this paper, we introduce an Ethnically Diverse Face dataset called EDFace-Celeb-1M, and design benchmark protocols along with analysis to evaluate and encourage the development of face hallucination algorithms. There exist three key objectives of creating the EDFace-Celeb-1M dataset for face hallucination. (1) It should contain a large-scale set of face images in the wild with unconstrained pose, emotion and exposure. Specially, the proposed EDFace-Celeb-1M dataset includes 1.7 million photos of more than 40,000 unique celebrity subjects. (2) The dataset should include faces from as many different countries and races as possible to mitigate the race bias in the current face hallucination datasets. More specifically, the EDFace-Celeb-1M dataset contains different race groups including White, Black, Latino and Asian with a relatively balanced composition. (3) It should be publicly available, enabling benchmarking current and future face hallucination methods with a unified dataset protocol to ensure fair and effortless evaluation. In this paper, the proposed dataset is public available with fixed training and testing sets for fair comparison.

To investigate the performance of current deep face hallucination algorithms on the constructed dataset with relatively balanced race composition, we design and provide concrete evaluation protocols, and evaluate four publicly available face hallucination methods and four SISR methods. We also introduce in detail our experiment setup and report baseline models to benefit and drive future research for the face hallucination task and inspire other related tasks in the computer vision community.

Our main contributions are summarized as follows.

- First, to the best of our knowledge, we build the first large-scale publicly available face hallucination dataset with relatively balanced race composition. The dataset includes 1.7 million face images collected from different race groups, providing fixed training and testing groups, pairs of low-
resolution and high-resolution images with different scale factors (e.g., $2 \times$, $4 \times$, $8 \times$), and aligned and non-aligned face images, which makes the future comparison more convenient, repeatable and credible.

- Second, we design a benchmark task to evaluate the performance of some current deep face hallucination and SISR methods to super-resolve low-resolution images. By doing so, it is clear how these algorithms perform on public face hallucination datasets (See Section 4).

- Third, we address fundamental questions of face hallucination and obtain several key findings.
  - How well do the current face hallucination and SISR methods perform in the case of different upsampling factors? (See Table 2 and Figs. 8, 9, 10, and 11)
  - How do the noise and blur kernels affect the performance of face hallucination methods? (See Table 2 and Figs. 8, 9, 10, and 11)
  - Is the size of training data important? (See Table 4 and Fig. 7)
  - Using the super-resolved images for the facial analysis tasks like landmark detection and identity preservation, is the gap significant, compared with using the ground truth high-resolution face images? (See Table 3)

## 2 RELATED WORK

### 2.1 Face Hallucination

Many approaches have been proposed for face hallucination, which can be classified into two categories: non-deep learning methods and deep learning methods. For the non-deep learning methods, holistic-based [19], [37], [38] and part-based [22], [39], [40] techniques are two popular models, which upscale face images via representing faces by parameters and extracting facial regions, respectively.

Recently, deep neural networks have been successfully applied to various computer vision tasks including face hallucination. Yu et al. [35] investigate the Generative Adversarial Network (GAN) to super-resolve face images of very low resolution and create perceptually realistic high-resolution face images. Huang et al. [29] introduce wavelet coefficients prediction into deep networks to generate super-resolution face images with different upscaling factors. To train deep face hallucination networks, Zhang et al. [41] propose a super-identity loss function to measure the difference of identity information. Cao et al. [30] design a novel attention-aware face hallucination framework and use deep reinforcement learning to optimize its parameters.

As a domain-specific super-resolution problem, there are also many face hallucination methods that use facial prior knowledge to help super-resolve low-resolution face images. Song et al. [42] propose a two-stage framework, which first generates facial components to represent the basic facial structures and then synthesizes fine-grained facial structures through a component enhancement method. Yu et al. [32] present a multi-task upsampling network to employ the image appearance similarity and exploit the face structure information with the help of the proposed facial component heat maps. Chen et al. [31] introduce a FSRNet model to make use of facial landmark heat maps and parsing maps. In addition, the attention mechanism [43] and bi-network [28] are also applied to make use of facial prior knowledge to train a high-resolution face generator.

### 2.2 Single Image Super-Resolution

The advancement of deep neural networks has achieved great success on image super-resolution, and most of the state-of-the-art SISR methods are based on deep learning [44]. As a pioneer work of deep SISR methods, Dong et al. [8], [45] propose a Super-Resolution Convolutional Neural Network (SRCNN) to super-resolve low-resolution images by first adopting deep learning for SISR. After that, many improvements have been explored. For example, Kim et al. [46] propose a deeply-recursive CNN to make use of skip connections to train their proposed a Deeply-Recurrent Convolutional Network (DRCN). Lim et al. [11] design an Enhanced Deep Super-Resolution Network (EDSR) to remove redundant modules and combine with multi-scale processing. To reduce the computational cost, many efficient SISR methods are proposed [9], [10], [12]. To make the generated images more realistic, GAN based SISR methods are introduced to improve the perceptual quality of HR images [47], [48], [49]. Recently, Hairs et al. [15] develop a Deep Back-Project Network (DBPN) to exploit the mutual dependencies with a feedback mechanism. Zhang et al. [13] introduce dense connections to make use of cues. Residual Channel Attention Networks (RCAN) is proposed in [14] to introduce a residual-in- residual structure and a channel attention module. More recently, there are many SISR methods utilizing novel attention modules [17], [50] or feedback mechanisms [51] to further improve the performance of SISR. Even without considering facial structures, these above methods can also super-resolve low-resolution face images to their corresponding high-resolution versions.

### 2.3 Face Datasets

Currently, there does not exist publicly available face hallucination datasets. In this section, we briefly review some popular face datasets which have been constructed recently. The Labeled Faces in the Wild (LFW) dataset [52] is created in 2007, and it contains 13,000 images. In 2014, the CelebA [53] and CASIA-WebFace [54] datasets are released, including about 20 K and 500 K images, respectively. The VGGface [55] dataset released in 2015 includes 2.6 million images.

More recently, Kemelmacher-Shlizerman et al. [56] assemble a dataset of 4.7 million images to evaluate how face recognition algorithms perform with a very large number of images. Cao et al. [57] release the VGGFace2 dataset. Compared to the VGGface dataset, VGGFace2 has 3.3 million images to cover a larger number of identities. The largest-scale face dataset is MS-Celeb-1 M, which contains 10 million images for training and testing.

Even though there exist many face datasets, none of them can be directly utilized to evaluate the current face hallucination approaches, due to the following reasons. First, none of these datasets provides large-scale pairs of low-resolution and high-resolution face images. However, the current deep face hallucination methods mainly rely on supervised learning and thus pairs of training face images are necessarily required. Second, researchers in the computer vision community have paid increasing attention to the race bias problem. However, these existing datasets are strongly biased toward specific races. Face hallucination models trained on these datasets will generate high-resolution face images with inappropriate race information.

The FairFace [58] contains face images from different races. However, it includes only 10 K images without pairs of face images for the evaluation of face hallucination methods. A summary of current face datasets is listed in Table 1 to give a clear view.

To overcome the problem of lacking face hallucination datasets, current face hallucination researchers synthesize pairs of training and testing samples to evaluate the previous methods and their proposed ones. For example, Yu et al. [32], [36] and Kim et al. [43] train and test their proposed methods on pairs of face images synthesized from CelebA [59]. Zhang et al. [60] synthesize pairs of face images from Multi-PIE [61] and CelebA [59]. WaveletSR [29] is evaluated on synthesized face images from VGGFace2 [57], and Li et al. [62] train and test their methods on the FFHQ [63], VGGFace2 [57] and CelebA [59] datasets. However, most of these synthesized face hallucination datasets are not public. Meanwhile,
3 EDFACE-Celeb-1 M COLLECTION

In this section, we provide an overview of the EDFace-Celeb-1 M dataset and introduce how it is collected in detail. We build the dataset to benchmark the current deep face hallucination methods and drive the development of the face hallucination task in the future. As mentioned above, we aim to build a publicly available large-scale face hallucination dataset, which provides pairs of low-resolution and high-resolution face images with a fixed setting of training and testing samples.

The dataset collection processing includes: how a list of candidate identities is obtained, how the candidate images are collected, how to detect the face in images, and how to synthesize pairs of low-resolution and high-resolution images. In addition, we provide attribute statistics of the proposed dataset such as race and gender.

3.1 Stage I: Obtain a Name List

The first step from scratch is to have a list of subject names whose faces we aim to collect. As mentioned before, race balance is the top priority when we build this dataset. To obtain such a name list with race balance, we split the races into four groups: White, Black, Asian and Latino. Based on this, we collect names for different groups. More specifically, for each group, we collect as many names of celebrities as possible from different countries. The names in our list are from diverse countries. For example, the Asian group includes names of people from East Asian, Southeast Asian, Middle East and India from Asian countries. In addition, it also includes some Asian-Americans. In summary, our EDFace-Celeb-1 M dataset contains more than 20,000 names and the ratio (name lists) of the above four groups is about 31.1% : 19.2% : 19.6% : 18.3%.

3.2 Stage II: Select Images for Each Identity

After obtaining a name list, we use the Google Image Search engine to download 100 ~ 1000 images for each identity. Moreover, to obtain diverse images with age variations, we further add the keyword young to each subject and further download the corresponding images.

3.3 Stage III: Face Detection

We then detect faces in images via the Dlib detector [64]. In this way, we can obtain a facial image dataset, containing faces of different poses/angles, variations of appearance (like glasses, hat). In addition, we manually remove some non-face images.

3.4 Stage IV: Synthesize Pairs of Images

With the above steps, we have obtained about 10 M facial images in total. Given the obtained facial images, we construct two subsets. The first one is composed of real-world low-resolution facial images, which are dedicated to the qualitative study of existing face hallucination methods, as there is no ground truth available. Specifically, we choose images whose resolution is smaller than 50 × 50 to compose this subset.

The second set is dedicated to quantitatively evaluating the existing face hallucination methods, consisting of pairs of high-resolution and low-resolution facial images. To this end, we have to choose high-resolution images and synthesize the corresponding low-resolution images. To be specific, we choose 1.5 M face images whose resolution is larger than 128 × 128 and resize them to 128 × 128. These images serve as high-resolution images. To synthesize the corresponding low-resolution images, we employ the strategies employed in most of the existing face hallucination methods. More specifically, we simulate the degradation process via specific operations like downsampling. The developed subset includes five different degradation settings, named as 2 ×, 4 ×, 4 ×_BD, 4 ×_DN and 8 ×. The numbers indicate the downsampling factor, “D” stands for downsampling, “B” indicates blur operation, and “N “ stands for Gaussian noise that is added to the LR images. The order of letters indicates the order of operations. For example, “BD” means that the blur artifact is applied prior to the downsampling operation. We use bicubic interpolation for downsampling. 2 ×, 4 × and 8 × mean that only bicubic downsampling operation is applied. When creating blurry and noisy face images, Gaussian blur and Gaussian noise are added to images. Among the 1.5 M image pairs, we choose 1.36 M pairs for training and 0.14 M pairs of images for quantitative testing.

In summary, by conducting the above four steps, we derive a dataset including two subsets of non-aligned facial images. The first one contains 200 K real-world low-resolution images for qualitative testing, and the second one includes 1.36 M and 140 K pairs of images for training and quantitative testing.
3.5 The Statistics of EDFace-Celeb-1 M

Lastly, we present the statistics of specific attributes of the developed EDFace-Celeb-1 M dataset as follows.

- First, we show some representative high-resolution face images from the EDFace-Celeb-1 M dataset in Fig. 2. It is obvious that our dataset includes different races including White, Black, Asian and Latino, from different countries. In addition, these faces exhibit evident pose and appearance variations (e.g., glasses). Fig. 1 presents the statistics of the face pose.

- Second, we demonstrate the ratios of races and gender. 64% of the images of our dataset are from males and the rest 36% images are from females. The ratios of White, Black, Asian and Latino are about 31.1%, 19.2%, 19.6% and 18.3%, respectively, which are relatively balanced compared with existing datasets of face images.

- Third, we also analyze the distribution of age of the celebrities we include in our dataset, and show the results in Fig. 3. The age of celebrities is estimated by the model from Insightface. Roughly, the majority of age is between 25 and 55, which aligns well with the property of demography. Notably, our dataset includes also celebrities younger than 20 and older than 60.

- Fourth, Fig. 4 presents the statistics of the resolution of face images before the resizing operation. In general, all the ground-truth high-resolution images are resized from images with resolution greater than $128 \times 128$.

- Fifth, the proposed dataset provides fixed training and testing subsets. Each subsets includes high-quality images and corresponding low-quality images. All images are labeled as HR and LR with factors (e.g., $2 \times$, $4 \times$, $4 \times \text{BD}$, $4 \times \text{DN}$ and $8 \times$). Based on the labels and fixed subsets, the dataset can make the reproducibility and the fair comparison easier.

4 EXPERIMENTS

In this section, we introduce the evaluation protocols and benchmark the existing face hallucination methods and representative SISR methods on the proposed EDFace-Celeb-1 M dataset.

4.1 Evaluated Methods

In this benchmark study, we investigate four state-of-the-art face hallucination methods (i.e., Deep Iterative Collaboration Network (DICNet) [33], Deep Iterative Collaboration GAN (DICGAN) [33], Wavelet-based Super-Resolution Network (WaveletSRNet) [29] and HiFaceGAN [65]), and four SISR methods (i.e., EDSR [11], Holistic Attention Network (HAN) [17], Residual Dense Network (RDN) [13], and Residual Channel Attention Network (RCAN) [14]). All methods are based on deep learning. Specifically, DICGAN [33] is an iterative framework of recurrently estimating landmarks and recovering high-resolution images. Wavelet coefficients are introduced in WaveletSRNet [29] to deal with very-low-resolution facial images. HiFaceGAN [65] is proposed to formulate the face restoration task as a generation problem guided by semantics, and this problem is addressed by a multi-stage framework containing several units of collaborative suppression and replenishment. For the SISR task, EDSR is an enhanced deep super-resolution network containing several Resblocks. HAN is proposed in [17] to model the correlation among different convolution layers with a layer attention module and channel-spatial attention module. RDN [13] is a residual dense network to exploit the hierarchical features from both the local and global perspectives. RCAN [14] is a deep residual channel attention network with both short and long skip connections. Channel attention is also adopted in this network for better performance. In summary, we select the above eight methods for three criteria. First, these methods achieve high values of the commonly used metrics like PSNR and SSIM. Second, the codes of
these methods are publicly available. Third, these methods are proposed for face hallucination or image super-resolution.

### 4.2 Implementation Details

Our dataset has four different degradation settings. Each setting corresponds to pairs of low-resolution and high-resolution face images, which are used to train different models. We use the code released from the original publications. For fair comparisons, the learning rate and epoch number for all methods are set as 0.0001 and 20, respectively. All models are trained using V100 GPUs. We conduct the calculation of different metrics in the RGB space to access the results. During the training stage, all models are trained on the training subset and the testing subset is not used. After training, we evaluate the models on the testing subset.

### 4.3 Face Super-Resolution

To evaluate the four methods on face hallucination, we provide the quantitative results of PSNR and SSIM on the EDFace-Celeb-1 M dataset in Table 2. Based on the PSNR and SSIM values, DICNet achieves the best performance on 4 × BD, 4 × DN and 8 × degradation settings. RCAN achieves the best performance on 2 ×. In terms of SSIM, the best performance values on 4 × BD, 4 × DN, and 8 × are also obtained by WaveletSR. The Table 2 does not provide the results of DICNet and DICGAN on X2 setting because the two methods do not work on this setting. Figs. 8, 9, 10 and 11 show the visual comparison of different methods on the EDFace-Celeb-1 M dataset.

### 4.4 Face Alignment

Facial identity information is important during face hallucination. An ideal face hallucination approach should ensure that the ID information of HR face images and SR face images are the same. However, in the real world, it is difficult for the face hallucination methods to generate SR results that are the same as HR images. In this section, we use a popular face ID information extractor [7] to extract ID information from HR face images and SR face images generated by different face hallucination methods. We then calculate the cosine distance between them. The larger value indicates that the ID information loss is less, which corresponds to a better face hallucination method. Table 3 shows the results of different models. We can see from Table 3 that the best performance on 2 ×, 4 ×, 4 × BD, 4 × DN and 8 × are obtained by HAN, HAN, RCAN, RCAM and RCAN, respectively.

### 4.5 Face Identity Information

Facial identity information is important during face hallucination. An ideal face hallucination approach should ensure that the ID information of HR face images and SR face images are the same. However, in the real world, it is difficult for the face hallucination methods to generate SR results that are the same as HR images. In this section, we use a popular face ID information extractor [7] to extract ID information from HR face images and SR face images generated by different face hallucination methods. We then calculate the cosine distance between them. The larger value indicates that the ID information loss is less, which corresponds to a better face hallucination method. Table 3 shows the results of different models. We can see from Table 3 that the best performance on 2 ×, 4 ×, 4 × BD, 4 × DN and 8 × are obtained by HAN & WaveletSR, RCAN, RCAN, RCAM and RCAN, respectively.

| Scale | Metrics | DICNet [33] | DGCAN [33] | WaveletSR [29] | HiFaceGAN [65] | EDSR [11] | RDN [13] | RCAN [14] | HAN [17] |
|-------|---------|-------------|-------------|----------------|----------------|----------|--------|---------|---------|
| 2×    | PSNR    | 30.60       | 25.68       | 31.23          | 31.37          | 31.42    | 31.41  |
|       | SSIM    | 0.9119      | 0.8836      | 0.8689         | 0.8889         | 0.8892   | 0.8888 |
| 4×    | PSNR    | 29.06       | 28.41       | 26.35          | 25.37          | 26.99    | 27.56  | 27.64   | 27.62   |
|       | SSIM    | 0.8453      | 0.8261      | 0.8211         | 0.7727         | 0.8035   | 0.8153 | 0.8161   | 0.8168   |
| 4 × BD | PSNR    | 29.68       | 28.58       | 26.52          | 24.23          | 27.22    | 27.83  | 27.89   | 27.88   |
|       | SSIM    | 0.8213      | 0.7998      | 0.7940         | 0.7009         | 0.7825   | 0.7955 | 0.7969   | 0.7964   |
| 4 × DN | PSNR    | 27.96       | 27.38       | 25.72          | 22.94          | 26.08    | 26.39  | 26.66   | 26.62   |
|       | SSIM    | 0.8117      | 0.7922      | 0.7815         | 0.6856         | 0.7673   | 0.7817 | 0.7835   | 0.7851   |
| 8×    | PSNR    | 25.29       | 24.64       | 22.33          | 21.88          | 23.24    | 23.74  | 23.77   | 23.73   |
|       | SSIM    | 0.7453      | 0.7134      | 0.6758         | 0.6408         | 0.6890   | 0.7117 | 0.7114   | 0.7114   |

The highest, second highest and lowest results are highlighted in bolded, blue and red, respectively.
about identification information during the training stage. In the future, it may be a meaningful direction for researchers to study face identification information for the task of face hallucination. Finally, we extract similarity scores across different people’s images before hallucination, and compare that with extracted similarity scores across different people’s images after hallucination (by RCAN method) and their high-resolution versions. The Fig. 6 shows the similarity.

4.6 Comparison on the Real LR Face Images
In addition, we also show the performance of the evaluated methods in the case of real-world scenarios based on our EDFace-Celeb-1M dataset. Taking a real-world low-resolution face image from our proposed dataset, we process it by different methods to generate SR face images, and the results are shown in Fig. 5. We can find that most of the current face hallucination methods can improve the quality of low-resolution images.

4.7 Impacts of Training Samples
Given that the proposed EDFace-Celeb-1 M is a large-scale public face hallucination dataset, we conduct an experimental study to explore the effect of the size of training samples for face hallucination. Fig. 7 and Table 4 show that models can achieve better performance with an increasing number of training samples.

5 Discussions
In this section, we discuss the quality of the dataset and the effects we have made to fairly evaluate the current methods.

5.1 The Quality of the Dataset
In summary, we provide an important dataset for the face hallucination community. The literature currently lacks a dataset specific to the face hallucination task and the proposed dataset makes the reproducibility and the fair comparison much easier for further research. To ensure the quality of this dataset, we make several efforts. First, to improve the quality of facial images in our dataset, we use a face detector to obtain facial images. In addition, we also manually remove some low-quality facial images. Second, to improve models’ reproducibility on the dataset, we fixed the two
training and testing subsets. As a comparison, previous methods randomly choose the two subsets. Third, different from previous methods which only provide one or two degraded settings, the provided dataset provides five degraded settings. In this way, the proposed dataset can better evaluate the performance of different face hallucination methods. Fourth, different from previous methods which ignore the problem of ethnicity, the proposed dataset provides a relatively balanced race composition. Fifth, different from some previous methods which evaluate their methods on relatively small-scale datasets, the provided dataset is the largest publicly available face hallucination dataset. In future, we will consider building more datasets to benefit the development of face hallucination, including using stronger face detectors to better detect faces and providing video sequences to help learn spatio-temporal information.

5.2 Fairness

In summary, to fairly evaluate the current methods, this paper makes some significant efforts. First, we build a large-scale publicly available dataset and fix the training and testing sets to conduct experiments, rather than randomly choose samples. Second, we use the popular metrics including PSNR and SSIM to compare different methods. Third, to further evaluate the current methods, we design two task-driven ablation studies including landmark detection and identify preservation.

| Scale | Metrics | 0.3M | 0.7M | 1.35M |
|-------|---------|------|------|-------|
| 4×    | PSNR    | 28.42| 28.78| 29.06 |
|       | SSIM    | 0.8265| 0.8362| 0.8453|
| 4× _BD| PSNR    | 28.93| 29.36| 29.68 |
|       | SSIM    | 0.8053| 0.8162| 0.8213|
| 4× _DN| PSNR    | 27.23| 27.58| 27.96 |
|       | SSIM    | 0.7986| 0.8073| 0.8117|
| 8×    | PSNR    | 24.63| 24.95| 25.29 |
|       | SSIM    | 0.7216| 0.7376| 0.7453|

Results are reported in terms of PSNR and SSIM. 0.3 M, 0.7 M and 1.35 M represent the size of training samples. The highest, second highest and lowest results are highlighted in bolded, Blue and Red, respectively.
Fig. 9. Visual results of BI models (×4) on the EDFace-Celeb-1 M dataset. From left to right: HR, results of bicubic, DICNet, DICGAN, WaveletSR, HiFaceGAN, EDSR, RDN, RCAN, and HAN. “BI” means the bicubic interpolation.

Fig. 10. Visual results of BD models (×4) on the EDFace-Celeb-1 M dataset. From left to right: HR, results of bicubic, DICNet, DICGAN, WaveletSR, HiFaceGAN, EDSR, RDN, RCAN, and HAN. “BD” means that the blur artifact is applied prior to the downsampling operation.
6 Conclusion

In this paper, we first propose the largest publicly available face hallucination dataset with relatively balanced race composition. It contains 1.5 million pairs of LR and HR face images for training and testing, and 140 K real-world tiny face images for quantitative comparisons. Thanks to the proposed EDFace-Celeb-1 M dataset, the following face hallucination can evaluate their methods on a public and fixed division of training and testing samples, which significantly makes the comparison convenient and improves the reliability.

In addition, given that the current face hallucination methods are evaluated on privately synthesized datasets, we benchmark four public available face hallucination methods, and four SISR methods on the proposed EDFace-Celeb-1 M dataset. The proposed dataset will be made publicly available to encourage the development of face hallucination algorithms.

References

[1] Y. Sun, X. Wang, and X. Tang, “Deep convolutional network cascade for facial point detection,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit., 2013, pp. 3476–3483.
[2] X. Zhu, Z. Lei, X. Liu, H. Shi, and S. Z. Li, “Face alignment across large poses: A 3D solution,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit., 2016, pp. 146–155.
[3] A. Bulat and G. Tzimiropoulos, “How far are we from solving the 2D & 3D face alignment problem? (and a dataset of 230,000 3D facial landmarks),” in Proc. IEEE Int. Conf. Comput. Vis., 2017, pp. 1021–1030.
[4] S. Liu, J. Yang, C. Huang, and M.-H. Yang, “Multi-objective convolutional learning for face labeling,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit., 2015, pp. 3451–3459.
[5] J. Lin, H. Yang, D. Chen, M. Zeng, F. Wen, and L. Yuan, “Face parsing with RoI Tanh-warping,” in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit., 2019, pp. 5654–5663.
[6] H. Wang et al., “ConFace: Large margin cosine loss for deep face recognition,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit., 2018, pp. 5265–5274.
[7] J. Deng, J. Guo, N. Xue, and S. Zafeiriou, “ArcFace: Additive angular margin loss for deep face recognition,” in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit., 2019, pp. 4690–4699.
[8] C. Dong, C. C. Loy, K. He, and X. Tang, “Learning a deep convolutional network for image super-resolution,” in Proc. Eur. Conf. Comput. Vis., 2014, pp. 184–199.
[9] C. Dong, C. C. Loy, and X. Tang, “Accelerating the super-resolution convolutional neural network,” in Proc. Eur. Conf. Comput. Vis., 2016, pp. 391–407.
[10] J. Kim, J. K. Lee, and K. Mu Lee, “Accurate image super-resolution using very deep convolutional networks,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit., 2016, pp. 1646–1654.
[11] B. Lim, S. Son, H. Kim, S. Nah, and K. Mu Lee, “Enhanced deep residual networks for single image super-resolution,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit. Workshops, 2017, pp. 136–144.
[12] W.-S. Lai, J.-B. Huang, N. Ahuja, and M.-H. Yang, “Deep laplacian pyramid networks for fast and accurate super-resolution,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit., 2017, pp. 624–632.
[13] Y. Zhang, Y. Tian, Y. Kong, B. Zhong, and Y. Fu, “Residual dense network for image super-resolution,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit., 2018, pp. 2472–2481.
[14] Y. Zhang, K. Li, K. Li, L. Wang, B. Zhong, and Y. Fu, “Image super-resolution using very deep residual channel attention networks,” in Proc. Eur. Conf. Comput. Vis., 2018, pp. 286–301.
[15] M. Haris, G. Shakhnarovich, and N. Ukita, “Deep back-projection networks for super-resolution,” in Proc. IEEE Int. Conf. Comput. Vis. Pattern Recognit., 2018, pp. 1664–1673.
[16] S. Anwar and N. Barnes, “Densely residual laplacian super-resolution,” IEEE Trans. Pattern Anal. Mach. Intell., vol. 44, no. 3, pp. 1192–1204, Mar. 2020.
[17] B. Niu et al., “Single image super-resolution via a holistic attention network,” in Proc. Eur. Conf. Comput. Vis., 2020, pp. 191–207.
[18] A. Chakrabarti, A. Rajagopalan, and R. Chellappa, “Super-resolution of face images using kernel PCA-based prior,” IEEE Trans. Multimedia, vol. 9, no. 4, pp. 888–892, Jun. 2007.
[19] C. Liu, H.-Y. Shum, and W. T. Freeman, “Face hallucination: Theory and practice,” Int. J. Comput. Vis., vol. 75, no. 1, pp. 115–134, 2007.
[20] K. Jia and S. Gong, “Generalized face super-resolution,” IEEE Trans. Image Process., vol. 17, no. 6, pp. 873–886, Jun. 2008.
[21] H. Huang, H. He, X. Fan, and J. Zhang, “Super-resolution of human face image using canonical correlation analysis,” Pattern Recognit., vol. 43, no. 7, pp. 2532–2543, 2010.
[22] X. Ma, J. Zhang, and C. Qi, “Hallucinating face by position-patch,” Pattern Recognit., vol. 43, no. 6, pp. 2224–2236, 2010.
Kaihao Zhang is currently working toward the PhD degree with the College of Engineering and Computer Science, The Australian National University, Canberra, ACT, Australia. His research interests focus on computer vision and deep learning. He has more than 30 referred publications in international conferences and journals, including the CVPR, ECCV, ICML, NeurIPS, AAAI, ACM Multimedia, IEEE Transactions on Pattern Analysis and Machine Intelligence, International Journal of Computer Vision, IEEE Transactions on Image Processing, etc.

Dongxu Li received the bachelor’s degree from Australian National University with first-class honours in computing. He is currently working toward the PhD degree with the Australian National University. His research interests are mainly computer vision and deep learning, including visual sequence representation learning, vision-language learning, and multi-modal learning.

Wenhan Luo received the BE degree from the Huazhong University of Science and Technology, China, 2009, the ME degree from the Institute of Automation, Chinese Academy of Sciences, China, 2012, and the PhD degree from Imperial College London, U.K., 2016. He is currently an associate professor with Sun Yat-Sen University. Prior to that, he worked as a research scientist for Tencent and Amazon. He has published more than 40 papers in top conferences and leading journals, including the ICML, CVPR, ICCV, ECCV, ACL, AAAI, ICLR, IEEE Transactions on Pattern Analysis and Machine Intelligence, International Journal of Computer Vision, IEEE Transactions on Image Processing, etc.
Jingyu Liu received the PhD degree from the Institute of Automation, Chinese Academy of Science, in 2018. He is currently a research engineer with Tencent. His research interests include object recognition and multi-modal learning.

Jiankang Deng (Student Member, IEEE) received the PhD degree from Imperial College London (ICL), supervised by prof. Stefanos Zafeiriou and funded by the Imperial president’s PhD Scholarships. His research topic is deep learning-based face analysis, including detection, alignment, reconstruction, recognition and generation etc. He is a reviewer in prestigious computer vision journals and conferences including IEEE Transactions on Pattern Analysis and Machine Intelligence, International Journal of Computer Vision, CVPR, ICCV, and ECCV. He is one of the main contributors to the widely used open-source platform Insightface.

Wei Liu (Senior Member, IEEE) received the PhD degree in EECS from Columbia University, New York, NY, USA. He is currently a distinguished scientist with Tencent AI Lab and a director with Computer Vision Center. He was a research scientist with IBM T. J. Watson Research Center, Yorktown Heights, NY. He has long been devoted to research and development in the fields of machine learning, computer vision, information retrieval, Big Data, etc. Till now, he has published more than 150 peer-reviewed journal and conference papers, including Proceedings of the IEEE, IEEE Transactions on Pattern Analysis and Machine Intelligence, IEEE Transactions on Knowledge and Data Engineering, International Journal of Computer Vision, NIPS, ICML, CVPR, ICCV, ECCV, KDD, Special Interest Group on Information Retrieval, SIGCHI, WWW, IJCAI, AAAI, etc. His research works won a number of awards and honors, such as the 2011 Facebook Fellowship, the 2013 Jury Award for best thesis of Columbia University, the 2016 and 2017 SIGIR Best Paper Award Honorable Mentions, and the 2018 “AI’s 10 To Watch” honor. He currently serves as an associate editor to several international leading AI journals and an area chair to several international top-tier AI conferences, respectively.

Stefanos Zafeiriou (Member, IEEE) is currently a professor in machine learning and computer vision with the Department of Computing, Imperial College London, London, UK, and an EPSRC Early Career research fellow. He served associate editor and guest editor in various journals including IEEE Transactions on Pattern Analysis and Machine Intelligence, International Journal of Computer Vision, Transactions on Automatic Control, Computer Vision and Image Understanding, and Innovative Veterinary Care Journal. He has been a guest editor of more than eight journal special issues and co-organised more than 16 workshops/special sessions on specialised computer vision topics in top venues. He has co-authored more than 70 journal papers mainly on novel statistical machine learning methodologies applied to computer vision problems, such as 2-/3-D face analysis, deformable object fitting and tracking, shape from shading, and human behaviour analysis, published in the most prestigious journals in his field of research, such as IEEE Transactions on Pattern Analysis and Machine Intelligence, International Journal of Computer Vision, and many papers in top conferences, such as CVPR, ICCV, ECCV, and ICML. His students are frequent recipients of very prestigious and highly competitive fellowships, such as the Google Fellowship x2, the Intel Fellowship, and the Qualcomm Fellowship x4. He has more than 20,000+ citations to his work, h-index 64. He was the General chair of BMVC 2017.

For more information on this or any other computing topic, please visit our Digital Library at www.computer.org/csdl.