Face Hallucination Revisited: An Exploratory Study on Dataset Bias

Preprint · December 2018

CITATIONS
0

READS
62

6 authors, including:

Klemen Grm
University of Ljubljana
17 PUBLICATIONS **237** CITATIONS
SEE PROFILE

Vitomir Štruc
University of Ljubljana
151 PUBLICATIONS **1,934** CITATIONS
SEE PROFILE

Face Hallucination Revisited: An Exploratory Study on Dataset Bias

Klemen Grm[1†], Martin Pernuš[†], Leo Cluzel[‡], Walter J. Scheirer[*], Simon Dobrišek[†], Vitomir Štruc[†]
[†]University of Ljubljana, Tržaška 25, Ljubljana, Slovenia
[‡]ENSEA, 6 avenue du Ponceau, Cergy, France
[*]University of Notre Dame, Notre Dame, IN 46556, USA
[1]Corresponding author, klemen.grm@fe.uni-lj.si

Abstract

Contemporary face hallucination (FH) models exhibit considerable ability to reconstruct high-resolution (HR) details from low-resolution (LR) face images. This ability is commonly learned from examples of corresponding HR-LR image pairs, created by artificially down-sampling the HR ground truth data. This down-sampling (or degradation) procedure not only defines the characteristics of the LR training data, but also determines the type of image degradations the learned FH models are eventually able to handle. If the image characteristics encountered with real-world LR images differ from the ones seen during training, FH models are still expected to perform well, but in practice may not produce the desired results. In this paper we study this problem and explore the bias introduced into FH models by the characteristics of the training data. We systematically analyze the generalization capabilities of several FH models in various scenarios where the degradation function does not match the training setup and conduct experiments with synthetically downgraded as well as real-life low-quality images. We make several interesting findings that provide insight into existing problems with FH models and point to future research directions.

1. Introduction

Face hallucination (FH) refers to the task of recovering high-resolution (HR) facial images from corresponding low-resolution (LR) inputs [2, 6, 11]. Solutions to this task have applications in face-oriented vision problems, such as face editing and alignment, 3D reconstruction or face attribute estimation [3, 6, 19, 23, 24, 25, 31, 43] and are used to mitigate performance degradations caused by input images of insufficient resolution. One particularly popular use of FH models is for LR face recognition tasks[13, 24, 45], where LR probe images are super-resolved to reduce the dissimilarity with HR gallery data.

Formally, face hallucination is defined as an inverse problem described by the following observation model [27]:

$$\mathbf{x} = \mathbf{H}\mathbf{y} + \mathbf{n}, \tag{1}$$

where $\mathbf{x}$ denotes the observed low-resolution face image, $\mathbf{H}$ stands for a composite down-sampling and blurring operator, $\mathbf{n}$ represents an additive i.i.d. Gaussian noise term with standard deviation σ_n, and $\mathbf{y}$ is the latent high-resolution face image that needs to be recovered [27]. Recent techniques increasingly approach the FH problem in (1) using machine learning methods [1, 4, 21, 28, 42, 44] and try to learn a direct (non-linear) mapping f_θ from the LR inputs to the desired HR outputs, i.e., $f_\theta : \mathbf{x} \rightarrow \mathbf{y}$.

This mapping is commonly implemented with a parameterized regression model, e.g., a convolutional neural network (CNN), and the parameters of the model, θ, are learned through an optimization procedure that minimizes a selected training objective (e.g., an L_p loss) over a set of corresponding LR-HR image pairs. Because the learning procedure is supervised, the image pairs needed for training are constructed by artificially degrading HR training images using a selected degradation function, i.e., a known operator $\mathbf{H}$ and noise level σ_n. Such an approach ensures that

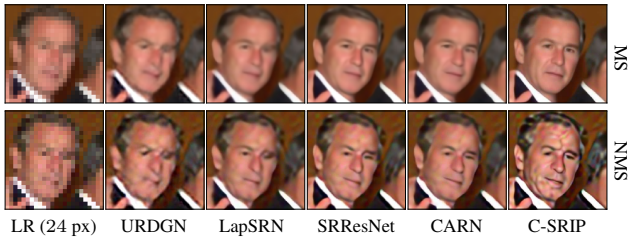


Figure 1. Hallucination examples (8×) for the five FH models used in this work (see Sec. 3 for details). The top row shows results for a LR image generated with a degradation procedure matching (MS) the one used during training and the bottom row shows results for an image produced by a non-matching degradation function (NMS). Note the difference in the reconstruction quality. In this paper, we study the bias introduced into FH models by the training data, which has so far received limited attention in the literature.

all generated LR images have corresponding HR ground truth faces available for training, but also implicitly defines the type of image degradations the learned model is able to handle. If the actual degradation function encountered with (real-world) test data differs from the one used during training, the result of the face hallucination model may be far from optimal - as illustrated in Fig. 1 for five recent state-of-the-art FH models [1, 11, 21, 22, 44].

As can be seen from the presented examples, the HR images recovered from a LR input that matches the characteristics of the training data (Fig. 1, top row) are of significantly better quality than those produced from a non-matching LR input image (Fig. 1, bottom row). While all models are able to convincingly ($8\times$) upscale the example 24×24 face with a matching LR image, the hallucination results exhibit considerable artifacts when a small change in terms of blur and noise is introduced into the degradation procedure. These examples show that the bias introduced into the FH models by the training data has a detrimental effect on the quality of the super-resolved faces and may adversely effect the generalization ability of the trained models to data with unseen characteristics.

Surprisingly, the problem of (face hallucination) model bias has received little attention from the research community so far. Nevertheless, it has important implications for the generalization abilities of FH models as well as for the performance of high-level vision tasks that rely on the generated hallucination results, most notably face recognition. The existing literature on the generalization abilities of FH techniques is typically focused on generalization across different facial characteristics, such as pose, facial expressions, occlusions or alignment, and less so on the mismatch in the degradation functions used to produce the LR test data or qualitative experiments with real-world imagery. Difficulties with model bias are, therefore, rarely observed. Similarly, when used to improve performance of LR face recognition problems, FH models are predominantly applied on artificially degraded images, leaving the question of generalization to real-world LR data unanswered.

In this paper, we aim to address these issues and study the problem of model bias in the field of face hallucination. We try to answer obvious research questions, such as: How do different image characteristics affect the reconstruction quality of FH models? How do FH models trained on artificially degraded images generalize to real-world data? Do FH models ensure improvements in LR face recognition when applied as a preprocessing step? Are there differences in recognition performance when using either artificially generated or real-world LR data? To answer these and related questions we conduct a rigorous analysis using five recent state-of-the-art FH models and examine in detail: *i)* the mismatch between the degradation procedure used to generate the LR-HR training pairs and the actual degradation function encountered with LR data, *ii)* changes in different separability measures before and after the application of FH models, and *iii)* face recognition performance with hallucinated images and a state-of-the-art CNN recognition model. We make interesting findings that point to open and rarely addressed problems in the area of face hallucination and provide insights into future research challenges.

2. Related Work

Bias in computer vision. Machine learning techniques are known to be sensitive to the characteristics of the training data and typically result in models with sub-optimal generalization abilities if the training data is biased towards certain data characteristics. The effect of dataset bias can, for example, be seen in [5], where commercial gender classification systems are shown to have a drop in gender-classification accuracy on darker-skinned subjects compared to lighter-skinned subjects, indicating insufficient training data coverage of the latter. Torralba and Efros [36] demonstrate that image datasets used to train classification models are heavily biased towards specific appearances of object categories, causing poor performance in cross-dataset experiments. Zhao et al. [46] show that datasets for semantic role labeling tasks, contain significant gender bias and introduce strong associations between gender labels and verbs/objects (e.g., *woman* and *cooking*) that lead to biased models for certain labeling tasks. These examples show that understanding dataset bias is paramount for the generalization abilities of machine learning models. Our work is related to these studies, as we also explore dataset bias. However, different from prior work, we focus on the task of face hallucination, which has not been studied from this perspective so far.

Face hallucination for face recognition. Face recognition performance with LR images tends to degrade severely in comparison to HR face data. To mitigate this problem, a significant body of work resorts to FH models and tries to up-sample images during pre-processing [8, 13, 24, 34] or to devise models that jointly learn an upscaling function and recognition procedure [15, 18, 40]. While performance improvements are reported with these works, experiments are commonly limited to artificially down-sampled images, findings are then simply extrapolated to real-world data and potential issues due to dataset bias are often overlooked.

Experiments with real LR images, on the other hand, are scarce in the literature and the usefulness of FH models for face recognition with real-world LR imagery has not received much attention by the research community. As part of our analysis, we study this issue and explore the effect of FH models on data separability and recognition performance on artificially down-sampled and real-world LR data.

3. Methodology

We now describe the methodology used for the analysis. We discuss the selected experimental setup, FH models considered, and the image datasets used in the experiments.

3.1. Experimental setup

We conduct our analysis with several state-of-the-art FH models and LR images of size 24×24 pixels. Since there is no clear distinction on what constitutes a LR image, we select the LR image data to be smaller than 32×32 pixels, which represents an image size, below which most computer vision models are known to deteriorate quickly in performance [12, 37, 39]. Given this rather small size, we use an upscaling factor of $8\times$ with the FH models and generate 192×192 images that are used as the basis for our analysis.

3.2. Face hallucination (FH) models

Using the presented setup, we study the effect of dataset bias using five recent FH (or super-resolution) models, i.e.: the Ultra Resolving Discriminative Generative Network (URDGN, [44]), the Deep Laplacian Super-Resolution Network (LapSRN, [21]), the Super-Resolution Residual Network (SRResNet, [22]), the Cascading Residual Network (CARN, [1]), and the Cascading Super Resolution Network with Identity Priors (C-SRIP, [11]). The selected models differ in the network architecture and training objective, but are all considered to produce state-of-the-art hallucination results as shown in Fig. 1. We also include an interpolation-based method in the experiments to have a baseline for comparisons. A short summary of the models is given below:

- **Bicubic interpolation** [20] is a learning-free approach that up-samples images by interpolating missing pixel values using Lagrange polynomials, cubic splines, or other similar functions. Unlike FH models it does not rely on domain knowledge when generating HR faces.

- **URDGN** consists of a generator and a discriminator network, and is trained using the generative adversarial network (GAN [9]) framework, where the discriminator is trained to tell apart real and generated HR images, whereas the generator is trained to minimize an L_2 reconstruction loss and the accuracy of the discriminator.

- **LapSRN** represents a CNN-based model that progressively up-samples LR images by factors of 2 through bilinear deconvolution and relies on a feature prediction branch to calculate the high-frequency residuals at each scale. Because of the progressive up-sampling, multi-scale supervision signals are used during training.

- **SRResNet** is a variant of the SRGAN [22] model that incorporates many of the recent tweaks used in CNN-based super-resolution, such as adversarial training, pixel shuffle up-sampling, batch normalization and leaky ReLU activations. SRResNet represents the generator network of SRGAN trained with the L_2 loss.

- **CARN** consists of a light-weight CNN, which is able to achieve state-of-the-art performance for the general super-resolution problems using an efficient cascading architecture that combines the design principles of densely connected networks [16] and ResNets [14]. We use the variant with local and global cascading connections, as opposed to the lighter variants of the network.

- **C-SRIP** is a CNN-based FH model that incorporates explicit face identity constraints into the training procedure in addition to the main reconstruction objective. The model has a cascaded architecture that allows it to use supervision signals at multiple scales during training.

To incorporate face-specific domain knowledge into the models and ensure a fair comparison, we train all models on the CASIA Webface [41] dataset using $494,414$ images of $10,575$ subjects. We crop the 192×192 central part of the images and generate the HR-LR data pairs for training by blurring the HR images with a Gaussian kernel of $\sigma_b = \frac{8}{3}$ and then downscaling them $8\times$ using bicubic interpolation.

3.3. Datasets.

We conduct experiments on the Labeled Face in the Wild (LFW [17]) and SCFace [10] datasets. We introduce artificial down-sampling to simulate low image resolutions with LFW and use the SCFace images to explore the effect of training data bias on real-world LR images.

- **LFW** is one of the most popular face dataset available, mainly due to the unconstrained settings in which the images were captured. The dataset [17] consists of $13,233$ face images of size 250×250 pixels belonging to 5749 subjects. For the experiments, we use only the central crop of the images to have faces of similar proportion to the ones used during FH model training.

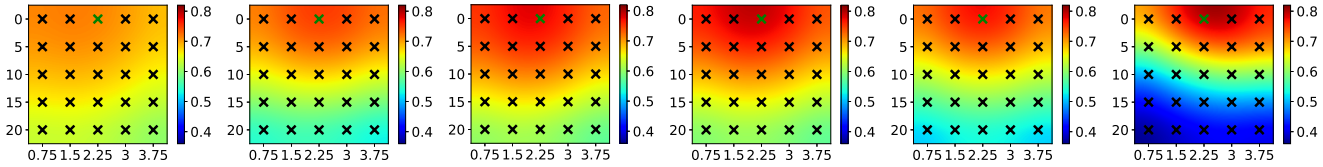
- **SCface** contains images of 130 subjects that are split between a gallery set, containing 130 high-resolution frontal mugshots (1 per subject), and a larger probe set of surveillance-camera images. The daylight camera set, which we use for our experiments, consists of images from 5 different security cameras. Each subject is recorded by each camera at 3 different distances, resulting in a total of $130 \times 5 \times 3 = 1950$ probe set images. We crop facial areas from all images based on the provided facial landmarks prior to the experiments.

4. Experiments and Results

To study the bias introduced into FH models by the particularities of the training-data-generation process, we con-

(a) LR inputs (σ_n vs. σ_b) (b) Bicubic (σ_n vs. σ_b) (f) C-SRIP (σ_n vs. σ_b)

Figure 2. Reconstruction capabilities of the learning-free bicubic interpolation and a selected FH model. The image block on the left (with samples of size 24×24 pixels) illustrates the effect of increasing noise (σ_n, decreases vertically) and blur (σ_b, increases horizontally) for a sample LR LFW image, the second and third block show 192×192 reconstructions generated by bicubic interpolation and C-SRIP, respectively. Images marked green are generated with a degradation function matching the one used during training. For the FH model good HR reconstructions are achieved only with images degraded similarly as the training data, whereas interpolation ensures reasonable reconstructions with all input images. Results for the remaining FH models are shown in the Appendix. Best viewed zoomed in.



(a) Bicubic (b) URDGN (c) LapSRN (d) CARN (d) SRResNet (d) C-SRIP

Figure 3. Reconstruction capabilities with mismatching degradation functions due to different blur and noise levels. The heat maps show the average SSIM values computed over artificially degraded LFW images. The points marked in the heat maps correspond to the sampled levels of noise (σ_n, decreases vertically) and blur (σ_b, increases horizontally). The value of σ_n and σ_b that was used for training is marked green. Note that all FH models achieve good reconstructions only around values that match the training setup. Best viewed in color.

duct a series of rigorous experiments in this section. We first present experiments with artificially degraded images on LFW, where we have complete control over the degradation process, and then report results with real-world surveillance images from SCFace.

4.1. Bias exploration with synthetic LR data

We start our analysis by exploring the sensitivity of FH models to a controlled mismatch in the degradation function. We first crop the (192×192) central part of the LFW images and generate baseline LR test data using the same degradation function as used during training. To simulate the mismatch, we generate additional sets of LR data from LFW by varying the standard deviations of the Gaussian blurring kernel σ_b and Gaussian noise term σ_n, which define $\mathbf{H}$ and $\mathbf{n}$ in (1). We consider five different values for each parameter and select σ_b from $[0.75, 1.5, 2.25, 3, 3.75]$ and σ_n from $[0, 5, 10, 15, 20]$. Because the LR test data is generated artificially, the HR ground truth can be used to evaluate the reconstruction capabilities of the FH models for each combination of σ_b and σ_n. Note that it is in general infeasible to include all possible data variations in the training procedure, so there will always be image characteristics that have not been accounted for by data augmentation. The selected noise and blur levels are therefore as reasonable factors as any to simulate the mismatch.

From the hallucination examples in Fig. 2 we see that visually convincing results for the FH model are produced only for LR images generated with blur and noise levels similar to those used during training (close to the images marked green), and deteriorate quickly as the difference to the training blur and noise levels gets larger (see Appendix for additional results). The interpolation baseline produces less convincing results compared to the best hallucinated image of C-SRIP, but does also introduces lesser distortions with images of other blur and noise levels. A similar observation can also be made for the remaining FH models based on the results in Fig. 3, where average structural similarity (SSIM) values computed over the entire LFW dataset are shown for different levels of noise and blur. Here, the com-

puted SSIM scores are shown in the form of interpolated heat maps for all five FH models and the baseline (bicubic) interpolation procedure. The first thing to notice is that the degradation in reconstruction quality is also visible for the (learning-free) interpolation method. This suggests that the reconstruction problem gets harder with increased noise and blur levels and the worsened reconstruction quality is not linked exclusively to the mismatch in the degradation function. However, the heat maps also clearly show that performance degrades much faster for the FH models than for the interpolation approach and that the degradation is particularly extreme for the C-SRIP model, which otherwise results in the highest peak SSIM score among all models.

In general, all FH models achieve significantly higher SSIM scores with matching degradation functions (see green point in Fig. 3) than the interpolation approach, but their performance falls below bicubic interpolation at the highest noise and blur levels - see lowest heat map part in Fig. 3. This is an important finding and implies that for imaging conditions that are difficult to model and challenging to reproduce using (1), interpolation may still be a better choice for recovering HR faces than FH models, which require representative HR-LR image pairs for training.

The presented results are consistent with recent studies [32, 33], which suggest that the performance of CNN models may come at the expense of robustness and that trying to learn models that are more robust to varying imaging conditions leads to less accurate results. We observe similar behaviour with the tested FH models (compare the heat maps of C-SRIP and URDGN, for example) and hypothesize that the relatively narrow focus of the models on specific degradation functions may be one of the reasons for the convincing performance of recent CNN-based FH models.

4.2. Bias exploration with synthetic and real data

Next, we explore the impact of dataset bias with synthetic LR images from LFW and with real-world surveillance data from SCFace, where the observed image degradations due to the acquisition hardware are not well modelled by the training degradation function. Since there is no HR ground truth available for the SCFace data, measuring the reconstruction quality is not possible with this dataset. We therefore focus on face recognition, which is regularly advocated in the literature as one of the main applications for FH models [8, 13, 24, 34], and use it as a proxy for face hallucination performance. Because this task is different from the reconstruction task studied above, we first run experiments with artificially degraded LFW images to have a baseline for later comparisons with results obtained on real-world SCFace data. We note that recognition experiments add another dimension to our analysis, as we now also explore the impact of the dataset bias on the semantic content of the reconstructed HR data and not only on the perceived quality of the hallucinated faces.

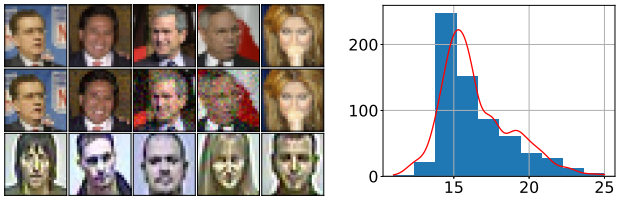


Figure 4. Examples of LR LFW and SCFace images used in the experiments. Left: the first row shows LFW samples degraded using the *matching* scheme (MS), the next row shows LFW images degraded with the *non-matching* scheme (NMS) and the last row shows images from SCFace. Right: distribution of SCFace image widths/heights (in px) for faces captured at the largest distance.

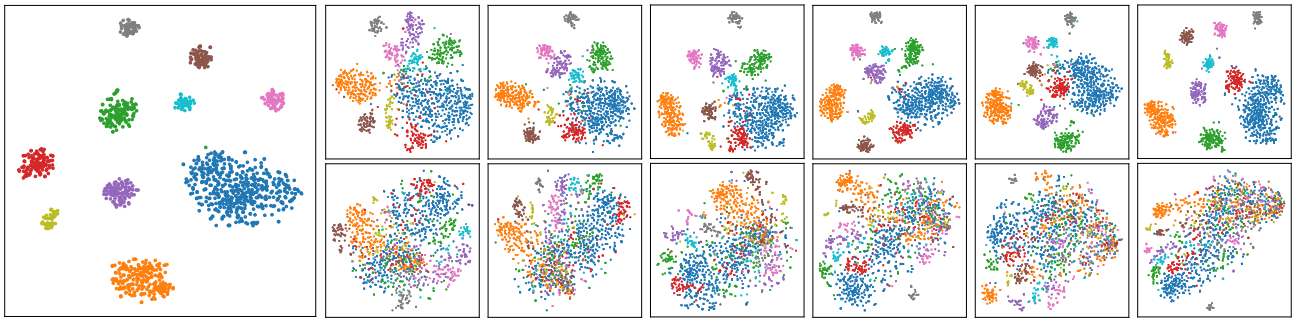
For the experiments, we use a ResNet-101 model [14] and train it for face recognition on a dataset of close to 1.8×10^6 images and 2622 identities [29]. We select the model because of its state-of-the-art performance [26, 30] and the fact that an open-source implementation is readily available. We perform network surgery on the trained ResNet-101 and use the activations from the penultimate network layer as a 512-dimensional descriptor of the input face images.

For the experiments with artificially down-sampled LFW data, we consider two different degradation schemes:

- **A matching scheme (MS),** where each full-resolution LFW image is first filtered with a Gaussian kernel of $\sigma_b = \frac{8}{3}$ and the generated image is then decimated to the target size using bicubic interpolation. No noise is added. This scheme matches the training setup.
- **A non-matching scheme (NMS)**, where σ_b is selected randomly from a uniform distribution, i.e., $\mathcal{U}(0.5, 4)$, for each LFW image. After filtering and down-sampling, images are corrupted through additive Gaussian noise with standard deviation σ_n, drawn randomly from $\mathcal{U}(0, 20)$. This ensures a mismatch between the applied degradation function and the one used during training. Furthermore, it results in a different degradation for every test image.

The two schemes generate LR data of size 24×24 and different characteristics as shown in Fig. 4. The generated images are then fed to the FH models for up-sampling and the HR results are used as inputs for ResNet-101.

For the experiments with the SCFace data, we use a subset of 650 images captured by the five surveillance cameras at the largest of all recorded distances. After removing the interlaced rows from the images as well as a corresponding number of columns to ensure a correct aspect-ratio, we end up with images, where the facial area covers an image region close in size to the 24×24 pixels expected by the FH models - a distribution for the SCFace face widths/heights is shown on the right of Fig. 4. We rescale all images to the correct input size (using bicubic interpolation) and then



(a) HR images (b) Bicubic (c) URDGN (d) LapSRN (e) CARN (f) SRResNet (g) C-SRIP

Figure 5. Visualization of ResNet-101 features extracted from hallucinated HR images using t-SNE [38]. Results are shown for the 10 largest classes of LFW. The plots show distributions for: (a) the original HR images, and (b-g) hallucinated HR face images images down-sampled using the matching (MS) or non-matching (NMS) degradation schemes. Best viewed in color and zoomed in.

Table 1. Average KL divergence for the 10 largest LFW classes with the MS and NMS degradation schemes estimated in the 2D space generated by t-SNE. Arrows indicate an increase or decrease in value compared to the baseline bicubic interpolation method.

Approach	LFW		
	MS	NMS	Change
Bicubic (baseline)	0.5389	0.2135	−0.3254
URDGN	0.5561 ↑	0.2143 ↑	−0.3418
LapSRN	0.6346 ↑	0.2087 ↓	−0.4259
CARN	0.6851 ↑	0.1957 ↓	−0.4894
SRResNet	0.7148 ↑	0.1962 ↓	−0.5222
C-SRIP	0.7676 ↑	0.1972 ↓	−0.5704

Table 2. GSI values achieved by the FH models in the ResNet-101 feature space. Note the decrease in the data separability due to mismatched degradation functions. Arrows indicate an increase or decrease in value compared to the baseline bicubic interpolation.

Approach	LFW			SCFace
	MS	NMS	Change	
Bicubic (baseline)	0.6283	0.5032	−19.9%	0.5963
URDGN	0.6481 ↑	0.4866 ↓	−24, 9%	0.5346
LapSRN	0.6657 ↑	0.4906 ↓	−26.3%	0.6218
CARN	0.7130 ↑	0.4858 ↓	−31.8%	0.5691
SRResNet	0.7084 ↑	0.4927 ↓	−30.4%	0.5840
C-SRIP	0.7104 ↑	0.4893 ↓	−31.1%	0.5712

feed the hallucination results produced by the FH models to ResNet-101 for descriptor computation.

Experiments on data separability. Using the experimental setup described above, we explore whether data separability is improved when facial details are hallucinated and how the separability is affected by the mismatch in the degradation function. To this end, we visualize the distribution of ResNet-101 feature descriptors extracted from hallucinated HR images of the 10 largest LFW classes (i.e., the 10 subjects with the highest number of images) using t-SNE [38] in Fig. 5. In order to quantitatively evaluate the separability of the presented distributions, we also compute a separability measure in the form of the Kullback-Leibler (KL) divergence between the distribution of a given class and joint distribution of all remaining classes in the 2D t-SNE embedding space and report average values calculated over all 10 considered LFW classes in Table 1.

We observe that for the original HR images (before down-sampling) the classes are well separated and show no overlap. After down-sampling with the matching scheme (MS) and subsequent up-sampling (top row in Fig. 5), we see considerable overlap in the class distributions for bicubic interpolation. The FH models, on the other hand, improve the data separability over the interpolation-based baseline and result in significantly higher KL-divergence scores. C-SRIP performs particularly well and generates compact class clusters with very little overlap.

With the non-matching scheme (NMS) all models perform noticeably worse, as shown in the bottom row of Fig. 5. Similarly as with the reconstruction experiments, we again see a drop in performance for bicubic interpolation, which is a learning-free approach and was hence not trained for specific image characteristics. This suggests that ensuring good data separation is a harder task for LR images generated by NMS and that the drop in the KL divergence is not only a result of mismatched degradation functions. However, if we take the performance drop of the interpolation approach as our baseline, we observe that the FH models are much more sensitive to the characteristics of the LR data. The KL divergence of all models drops to a comparable value around 0.2 and for the majority (except for URDGN) even falls slightly behind bicubic interpolation.

To further analyze the separability of the ResNet-101 descriptors of the hallucinated images, we report values for another non-parametric separability measure. i.e., Thornton's Geometric Separability Index (GSI), however, this time for the entire LFW and SCFace datasets and all FH models in Table 2. The index is defined as the fraction of data instances of a given dataset, $\mathcal{S}$, that has the same class-labels as their nearest neighbors, i.e. [35]: $GSI =$

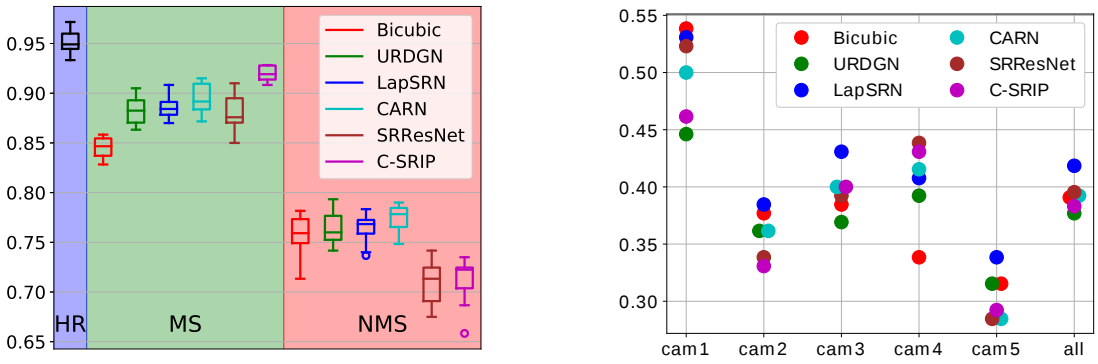


Figure 6. Recognition results on LFW (left) and SCFace (right). With a matching degradation function all models improve upon interpolation. The results are less predictable with image characteristics not seen during training. Best viewed in color.

$\frac{1}{n}\sum_{i=1}^{n} f(\mathbf{z}_i, \mathbf{z}'_i)$, where n stands for the cardinality of $\mathcal{S}$ and f is an indicator function that returns 1 if the i-th ResNet-101 descriptor $\mathbf{z}_i$ and its nearest neighbor $\mathbf{z}'_i$ share the same label and 0 otherwise. GSI is bounded between 0 and 1, where a higher value indicates better separability. We use the cosine similarity to determine nearest neighbors.

The results in Table 2 again show that the data separability is improved with all FH models compared to the baseline with the MS scheme on LFW. With the NMS scheme all models perform worse than the baseline and also exhibit a larger drop in separability than simple bicubic interpolation. On SCFace we see a similar picture. Only LapSRN results in better separability than the interpolation-based baseline, while all other FH models decrease separability. These results again point to the importance of suitable training data, as FH models do not generalize well to unseen image characteristics and perform different than expected when applied on real-world imagery.

Recognition experiments. In our last series of experiments we look at the recognition performance achieved by the FH models and extracted ResNet-101 descriptors on LFW and SCFace. For LFW we follow the so-called "unrestricted outside data" protocol and use the 6000 pre-defined image pairs in verification experiments. We keep one of the images in each pair unchanged (at the original resolution), and down-sample the second one using either the MS or NMS scheme. The LR images are then upscaled with the FH models and used to extract ResNet-101 descriptors. Matching is done with the cosine similarity. We report verification accuracy for the 10 predefined experimental folds. For SCFace we perform a series of identification experiments, where we try to recognize subjects in the upscaled HR probe images based on the HR gallery data.

Fig. 6 shows that on HR LFW images the ResNet-101 model achieves a median verification accuracy of 95.1%. When the image size is reduced to 24×24 pixels with the MS scheme and the LR images are upscaled with bicubic interpolation, the performance drops to 84.5%. The FH models improve on this and achieve significantly better results. The highest median accuracy of 91.8% comes from C-SRIP, which is the top performer in this setting. With the NMS scheme the drop in performance is larger for all methods compared to the HR data. URDGN, LapSRN and CARN are only able to match the performance achieved by bicubic interpolation, while SRResNet and C-SRIP degrade results.

Results for SCFace are shown separately for each of the five cameras and in the form of the overall mean identification accuracy (i.e., rank-1) in Fig. 6. We see that none of the FH models outperforms the bicubic baseline on all cameras. Overall, LapSRN offers a slight improvement over bicubic interpolation considering the average identification accuracy, but the performance gain is modest and in the range of 3%. The ranking of the models is also not consistent across different cameras, which generate LR data with very different characteristics. Observe, for example, C-SRIP, which performs worst with images from camera 2, but is one of the top performers on camera 4, where it gains around 10% in performance over bicubic interpolation. These results show that without suitable mechanisms that are able to compensate for the bias introduced into FH model by the training data, hallucination results with real-world images are unpredictable and findings made with artificially down-sampled images cannot simply be extrapolated to real-world data.

5. Conclusion, discussion and outlook

We have studied the impact of dataset bias on the problem of face hallucination and analyzed five recent CNN-based FH models on artificially degraded as well as real-world LR images. Below is summary of the main findings:

- **Reconstruction and robustness:** FH models achieve better reconstruction performance than the learning-free interpolation baseline on LR images matching the training data in terms of characteristics. However, their supe-

riority fades away quickly as the LR image characteristics diverge from the training setting. The rather sudden drop in reconstruction quality points to an accuracy-robustness trade-off with FH models not present with learning-free approaches, as also observed for other CNN-based models by recent studies [32, 33].

- **Separability and recognition:** We observe statistically significant improvements in data separability and face recognition performance, when LR images are preprocessed with FH models (as opposed to interpolated), but *only* for LR images degraded with the same approach as used during training. For mismatched image characteristics (with real-world data) we found no significant improvements in separability or recognition performance for any of the FH models, which in most cases fall behind simple interpolation.

Overall, our results suggest that despite recent progress, FH models are still very sensitive to the characteristics of the LR input data. We found limited empirical proof of their usefulness for higher-level vision tasks (e.g., recognition) beyond improvements in perceptually quality – which might be important for representation-oriented problems, such as alignment or detection. Our analysis shows that we, as a community, need to move away from the standard evaluation methodology involving artificially degraded LR images and focus on more challenging real-world data when developing FH models for specific vision problems.

A common way to mitigate the effects of dataset bias in CNN-based models from the literature are domain adaption (DA) techniques or ensemble approaches [7]. These have not been explored extensively for the problem of face hallucination yet (see [4] for initial attempts), but seem like an good starting point to improve the generalization abilities of FH models and make them applicable to real-world data.

References

[1] N. Ahn, B. Kang, and K.-A. Sohn. Fast, accurate, and lightweight super-resolution with cascading residual network. In *ECCV*, 2018. 1, 2, 3

[2] S. Baker and T. Kanade. Limits on super-resolution and how to break them. *TPAMI*, 2002. 1

[3] A. Bulat and G. Tzimiropoulos. Super-FAN: Integrated facial landmark localization and SR of real-world LR faces in arbitrary poses with GANs. In *CVPR*, 2018. 1

[4] A. Bulat, J. Yang, and G. Tzimiropoulos. To learn image super-resolution, use a GAN to learn how to do image degradation first. In *ECCV*, 2018. 1, 8

[5] J. Buolamwini and T. Gebru. Gender shades: Intersectional accuracy disparities in commercial gender classification. In *CFATP*, 2018. 2

[6] Y. Chen, Y. Tai, X. Liu, C. Shen, and J. Yang. FSRnet: End-to-end learning face SR with facial priors. In *CVPR*, 2018. 1

[7] G. Csurka. Domain adaptation for visual applications: A comprehensive survey. *arXiv:1702.05374*, 2017. 8

[8] R. A. Farrugia and C. Guillemot. Face hallucination using linear models of coupled sparse support. *TIP*, 2017. 2, 5

[9] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio. Generative adversarial nets. In *NIPS*, 2014. 3

[10] M. Grgic, K. Delac, and S. Grgic. SCface–surveillance cameras face database. *MTA*, 2011. 3, 11, 13

[11] K. Grm, S. Dobrišek, W. J. Scheirer, and V. Štruc. Face hallucination using cascaded super-resolution and identity priors. *arXiv:1805.10938*, 2018. 1, 2, 3

[12] K. Grm, V. Štruc, A. Artiges, M. Caron, and H. K. Ekenel. Strengths and weaknesses of deep learning models for face recognition against image degradations. *IET Biom.*, 2017. 3

[13] B. Gunturk, A. Batur, Y. Altunbasak, M. Hayes, and R. Mersereau. Eigenface-domain super-resolution for face recognition. *TIP*, 2003. 1, 2, 5

[14] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *CVPR*, 2016. 3, 5

[15] P. H. Hennings-Yeomans, S. Baker, and B. V. Kumar. Simultaneous super-resolution and feature extraction for recognition of low-resolution faces. In *CVPR*. IEEE, 2008. 2

[16] G. Huang, Z. Liu, L. Van Der Maaten, and K. Q. Weinberger. Densely connected convolutional networks. In *CVPR*, 2017. 3

[17] G. Huang, M. Ramesh, T. Berg, and E. Learned-Miller. Labeled faces in the wild: A database for studying face recognition in unconstrained environments. In *T. rep.* UMAS, 2007. 3

[18] M. Jian and K.-M. Lam. Simultaneous hallucination and recognition of low-resolution faces based on singular value decomposition. *TCSVT*, 2015. 2

[19] A. Jourabloo, M. Ye, X. Liu, and L. Ren. Pose-invariant face alignment with a single CNN. In *ICCV*, 2017. 1

[20] R. Keys. Cubic convolution interpolation for digital image processing. *TASSP*, 1981. 3

[21] W.-S. Lai, J.-B. Huang, N. Ahuja, and M.-H. Yang. Deep laplacian pyramid networks for fast and accurate superresolution. In *CVPR*, 2017. 1, 2, 3

[22] C. Ledig, L. Theis, F. Huszár, J. Caballero, A. Cunningham, A. Acosta, A. Aitken, A. Tejani, J. Totz, Z. Wang, and W. Shi. Photo-realistic single image super-resolution using a generative adversarial network. In *CVPR*, 2017. 2, 3

[23] Y. Li, S. Liu, J. Yang, and M.-H. Yang. Generative face completion. In *CVPR*, 2017. 1

[24] F. Lin, C. Fookes, V. Chandran, and S. Sridharan. Super-resolved faces for improved face recognition from surveillance video. *ICB*, 2007. 1, 2, 5

[25] L. Liu, S. Li, and C. L. P. Chen. Quaternion locality-constrained coding for color face hallucination. *TC*, 2018. 1

[26] I. Masi, A. Tran, T. Hassner, J. Leksut, and G. Medioni. Do we really need to collect millions of faces for effective face recognition? In *ECCV*, 2016. 5

[27] K. Nasrollahi and T. B. Moeslund. Super-resolution: a comprehensive survey. *MVA*, 2014. 1

[28] K. Nguyen, C. Fookes, S. Sridharan, M. Tistarelli, and M. Nixon. Super-resolution for biometrics: A coprehensive survey. *PR*, 2018. 1

[29] O. Parkhi, A. Vedaldi, and A. Zisserman. Deep face recognition. In *BMVC*, 2015. 5

[30] R. Ranjan, V. Patel, and R. Chellappa. Hyperface: A deep multi-task learning framework for face detection, landmark localization, pose estimation, and gender recognition. *TPAMI*, 2017. 5

[31] J. Roth, Y. Tong, and X. Liu. Adaptive 3d face reconstruction from unconstrained photo collections. In *CVPR*, 2016. 1

[32] D. Stutz, M. Hein, and B. Schiele. Disentangling adversarial robustness and generalization. *arXiv:1812.00740*, 2018. 5, 8, 11

[33] D. Su, H. Zhang, H. Chen, J. Yi, P. Chen, and Y. Gao. Is robustness the cost of accuracy? A comprehensive study on the robustness of 18 deep image classification models. In *ECCV*, 2018. 5, 8, 11

[34] W.-T. Su, C.-C. Hsu, C.-W. Lin, and W. Lin. Supervised-learning based face hallucination for enhancing face recognition. In *ICASSP*, 2016. 2, 5

[35] C. Thornton. Separability is a learners best friend. In *NCP Workshop*, 1998. 6

[36] A. Torralba and A. Efros. Unbiased look at dataset bias. In *CVPR*, 2011. 2

[37] A. Torralba, R. Fergus, and W. T. Freeman. 80 million tiny images: A large data set for nonparametric object and scene recognition. *TPAMI*, 2008. 3

[38] L. van der Maaten and G. Hinton. Visualizing data using t-SNE. *JMLR*, 2008. 6

[39] Z. Wang, S. Chang, Y. Yang, D. Liu, and T. S. Huang. Studying very lr recognition using deep networks. In *CVPR*, 2016. 3

[40] J. Wu, S. Ding, W. Xu, and H. Chao. Deep joint face hallucination and recognition. *arXiv:1611.08091*, 2016. 2

[41] D. Yi, Z. Lei, S. Liao, and S. Z. Li. Learning face representation from scratch. *arXiv:1411.7923*, 2014. 3

[42] X. Yu, B. Fernando, B. Ghanem, F. Porikli, and R. Hartley. Face super-resolution guided by facial component heatmaps. In *ECCV*, 2018. 1

[43] X. Yu, B. Fernando, R. Hartley, and F. Porikli. Super-resolving very low-resolution face images with supplementary attributes. In *CVPR*, 2018. 1

[44] X. Yu and F. Porikli. Ultra-resolving face images by discriminative generative networks. In *ECCV*, 2016. 1, 2, 3

[45] J. Zhao, Y. Mao, Q. Fang, Z. Liang, F. Yang, and S. Zhan. Heterogeneous face recognition based on SR reconstruction by adaptive multi-dictionary learning. In *CCBR*, 2015. 1

[46] J. Zhao, T. Wang, M. Yatskar, V. Ordonez, and K.-W. Chang. Men also like shopping: Reducing gender bias amplification using corpus-level constraints. *arXiv:1707.09457*, 2017. 2