A. Appendix

In this section we show some additional results and present additional findings made during our analysis.

A.1. Reconstruction vs. recognition

The literature on face hallucination typically focuses on designing FH models that ensure visually convincing super-resolution results and then presents recognition experiments with artificially down-sampled images to demonstrate how the models aid recognition. The main assumption here is that better reconstruction capabilities (in terms of average PSNR and SSIM scores) translate into better recognition performance. While this may be true for human perception, machine learning models do not necessarily behave in the same way, especially if the bias introduced into the models by the training data is taken into account.

To analyze the relationship between reconstruction capabilities and the recognition performance ensured by the FH models, we present a number of results in Tables 3, 4 and 5. Table 3 shows the average PSNR and SSIM scores for the matching (MS) and non-matching degradation (NMS) schemes achieved by the models on the LFW dataset. Table 4 presents the recognition accuracy for both degradation schemes on LFW and separately for all five cameras on SCFace. Here, results are reported in terms of the average verification accuracy computed over 10 experimental folds for LFW and as the rank-1 recognition rate for SCFace. Tables 5 summarizes the results from Tables 3 and 4 in terms of relative ranking for the given task.

Table 3. Average SSIM and PSNR values achieved by the tested FH models with the matching (MS) and non-matching (NMS) degradation schemes on LFW.

Approach	MS		NMS	
	PSNR	SSIM	PSNR	SSIM
Bicubic	24.401	0.7129	23.459	0.6745
URDGN	25.594	0.7539	23.434	0.6566
LapSRN	26.417	0.7792	23.967	0.6929
CARN	26.894	0.7938	23.934	0.6923
SRResNet	27.176	0.8013	23.589	0.6760
C-SRIP	27.233	0.8202	23.138	0.6145

Table 4. Average recognition accuracy achieved by the tested FH models and ResNet-101 with the matching (MS) and non-matching (NMS) degradation schemes on LFW and all cameras of SCFace at the largest 2.6m distance.

Approach	LFW		SCFace				
	MS	NMS	C1	C2	C3	C4	C5
Bicubic	0.846	0.759	0.538	0.377	0.384	0.338	0.315
URDGN	0.882	0.759	0.445	0.361	0.369	0.392	0.315
LapSRN	0.884	0.768	0.531	0.384	0.431	0.407	0.338
CARN	0.891	0.778	0.500	0.361	0.400	0.415	0.284
SRResNet	0.876	0.713	0.523	0.338	0.392	0.438	0.284
C-SRIP	0.919	0.722	0.461	0.330	0.400	0.431	0.292

From the presented results we see that the reconstruction quality with matching LR image characteristics is not a good indicator of the reconstruction quality with mismatched characteristics nor of the recognition performance ensured by the FH models. Models that performed well in one aspect do not necessarily generalize well to other tasks and image characteristics. C-SRIP, for example, achieves the highest PSNR and SSIM scores with the MS scheme on LFW and also leads to the best recognition performance in this setting, but performs worst in terms of reconstruction and recognition with the NMS scheme. Moreover, it also performs poorly on the SCFace data. LapSRN, on the other hand, is among the bottom three performers with the MS scheme on LFW in terms of reconstruction quality, but does better in the reconstruction experiments with the NMS scheme - the ranking here should be interpreted with reservation, as all tested models achieve very similar average PSNR scores. In terms of recognition performance, LapSRN still ensures only average results with the MS scheme, but does somewhat better with the NMS scheme. However, in recogition experiments on real-world SCFace data, LapSRN is overall the top performer, ensuring slight (statistically non-significant) improvements over the interpolation baseline and doing relatively well with LR images of all five cameras. If we look at the results for bicubic interpolation, we see that with matching LR image characteristics this baseline exhibits the weakest reconstruction capabilities, but is more competitive with non-matching data characteristics. In terms of recognition performance, it comes in last with the MS scheme, third (out of six) with the NMS scheme, and is very competitive on SCFace .

Overall, we observe that the peak reconstruction performance achieved with matching image characteristics typically reported in the literature, does not correlate well with the recognition performance ensured by the FH models - when used as a preprocessing step for face recognition. Instead, we notice that the recognition performance seems to be related more to the robustness of the models and their ability to handle image characteristics not seen during training. If we look at the heat maps in Fig. 7, where SSIM and PSNR scores, computed over all LFW images, are presented for different noise and blur levels, we see that techniques that degrade the least across different settings, e.g., bicubic interpolation and LapSRN, in terms of reconstruction quality (even if their reconstruction performance is average), also result in competitive recognition accuracy with real-world images. Models, sensitive to image characteristics, such as C-SRIP, on the other hand, deteriorate quickly as the degradation function deviates from the training setup (see Fig. 8), and perform relatively worse in the recognition task.

While we do not explore these findings further, our results suggest that the common way of optimizing the peak reconstruction performance of FH models may not be the most optimal choice of approaching the hallucination prob-

Table 5. Relative ranking of the FH models for the reconstruction (based on PSNR) and recognition tasks with matched and mismatched LR image characteristics. Note that good reconstruction performance does not necessarily translate into good recognition performance.

Approach	LFW - Reconstruction		LFW - Recognition		SCFace - Recognition				
	MS	NMS	MS	NMS	C1	C2	C3	C4	C5
Bicubic	6th	4th	6th	3rd	1st	2nd	5th	6th	2nd
URDGN	5th	5th	4th	4th	6th	3rd	6th	5th	3rd
LapSRN	4th	1st	3rd	2nd	2nd	1st	1st	4th	1st
CARN	3rd	2nd	2nd	1st	4th	4th	2nd	3rd	5th
SRResNet	2nd	3rd	5th	5th	3rd	5th	4th	1st	6th
C-SRIP	1st	6th	1st	6th	5th	6th	3rd	2nd	4th

lem if the target application is face recognition. In the recently observed accuracy-robustness trade-off of CNN models [32, 33], the robustness aspect seems more important, even, when it comes at the expense of performance. This is an interesting observation and suggests that we need to rethink the standard methodologies used in the field of face hallucination, especially if the hallucination task is paired with a higher-level vision problem.

A.2. Mismatch due to blur and noise

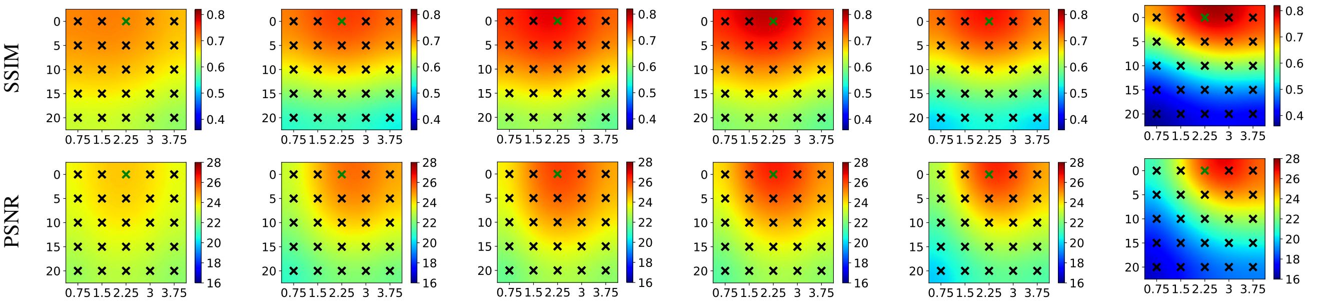
In Fig. 8 we show visual examples of the reconstruction capabilities of all tested FH model, as opposed to the main part of the paper, where only results for the bicubic interpolation approach and the top-performer in terms of reconstruction quality with matching image characteristics, i.e., the C-SRIP model, were presented. We see a consistent behaviour with all models - they are able to generate convincing reconstructions with LR images matching the characteristics of the data used during training, but introduce considerable artifacts as soon as the degradation function starts to deviate from the training degradation function. Also, we observe that the visual quality of the reconstructions produced by C-SRIP, which produces the highest quality HR face images with a matching function, is most affected by the mismatch. The remaining models still deteriorate in performance, but visually, the results appear less distorted.

More objective results evaluating the effect of mismatched noise and blur levels over the entire LFW dataset are presented in the heat maps in Fig. 7. Here, we show results for PSNR and SSIM, while in the main part of the paper only heat maps for SSIM were presented. The point that corresponds to the training setting is again marked green in the figures. We see that PSNR behaves similar to SSIM. the FH models are able to outperform bicubic interpolation significantly around conditions similar to the ones seen during training, but are less robust than interpolation and degrade faster once the degradation function used to generate the LR data starts to deviate from the functions used during training.

A.3. Reconstructing real-world LR images

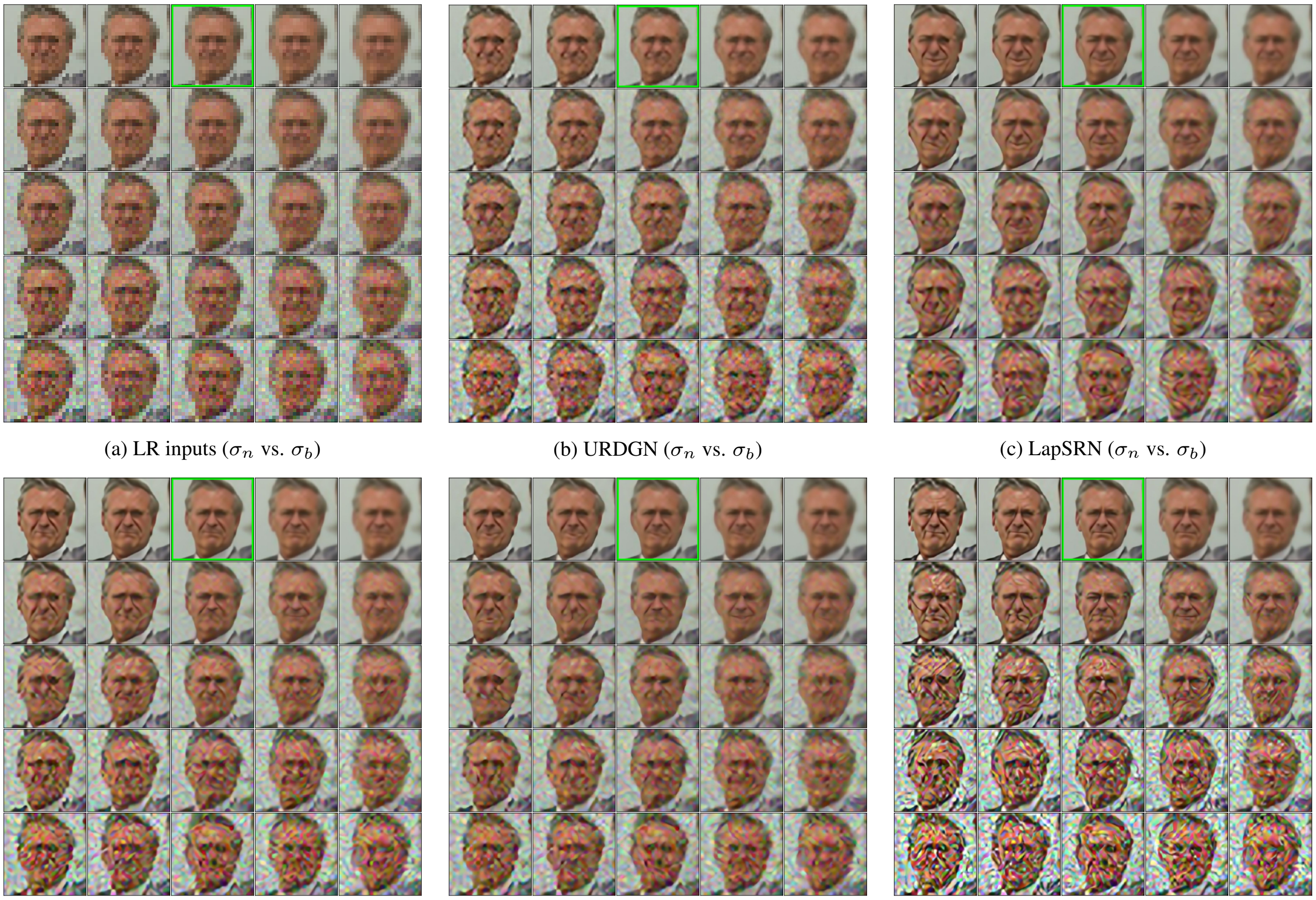
In Fig. 9, we present outputs of the tested face hallucination models on the SCFace [10] dataset. As in our recognition experiment, we only consider the images captured from a 2.6 meter distance (i.e., the distance 1 series of the dataset), because these images most closely match the training input size of the face hallucination models. As this dataset contains real-world images from several different surveillance cameras, we don't have a high-resolution ground truth images available for the LR images, only a high-resolution gallery for every subject. We therefore compare the the outputs of the face hallucination models to these HR galleries qualitatively.

Fig. 9 shows that the characteristics of the LR images differ considerably from camera to camera and affect not only the perceived quality of the LR data, but also other aspects, such as color scheme, saturation, image contrast, etc. The FH models are able to improve upon the visual quality of the HR reconstructions compared to the interpolation baseline for certain cameras and less so for others. For example, we see considerably more facial details in the C-SRIP reconstructions on cameras 2, 3, and 4 compared to the baseline. On cameras 1 and 5, however, the images still appear crisper and less blurred, but image artifacts are also present and contribute to the perception of low-quality HR reconstructions. Similar observations can also be made for other FH models, which follow the outlined trend and behave similarly to C-SRIP. Considering the three selected examples, LapSRN seems to strike a good balance between reconstruction quality and the amount of introduced image distortions - observe the HR reconstructions from cameras 1 and 5 for all three subjects.



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

Figure 7. Image reconstruction capabilities with mismatching degradation functions due to different blur and noise levels. The heat maps show the average SSIM (top row) and PSNR (bottom row) values computed over artificially degraded LFW images. The points marked in the heat maps correspond to the different 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 and zoomed in.



(a) LR inputs (σ_n vs. σ_b) (b) URDGN (σ_n vs. σ_b) (c) LapSRN (σ_n vs. σ_b)

(d) SRResNet (σ_n vs. σ_b) (e) CARN (σ_n vs. σ_b) (f) C-SRIP (σ_n vs. σ_b)

Figure 8. Reconstruction capabilities of all tested FH models with mismatching degradation functions due to different blur and noise levels. The LR image block (with samples of size 24 × 24 pixels) in the top left corner illustrates the effect of increasing noise (σ_n, decreases vertically) and blur (σ_b, increases horizontally) levels for a sample LFW image, the remaining image blocks show the 192×192 reconstructions generated by the tested models. Images marked green are generated with a degradation function matching the one used during training. Note that good HR reconstructions are achieved only with images degraded similarly as the training data. Best viewed zoomed in.

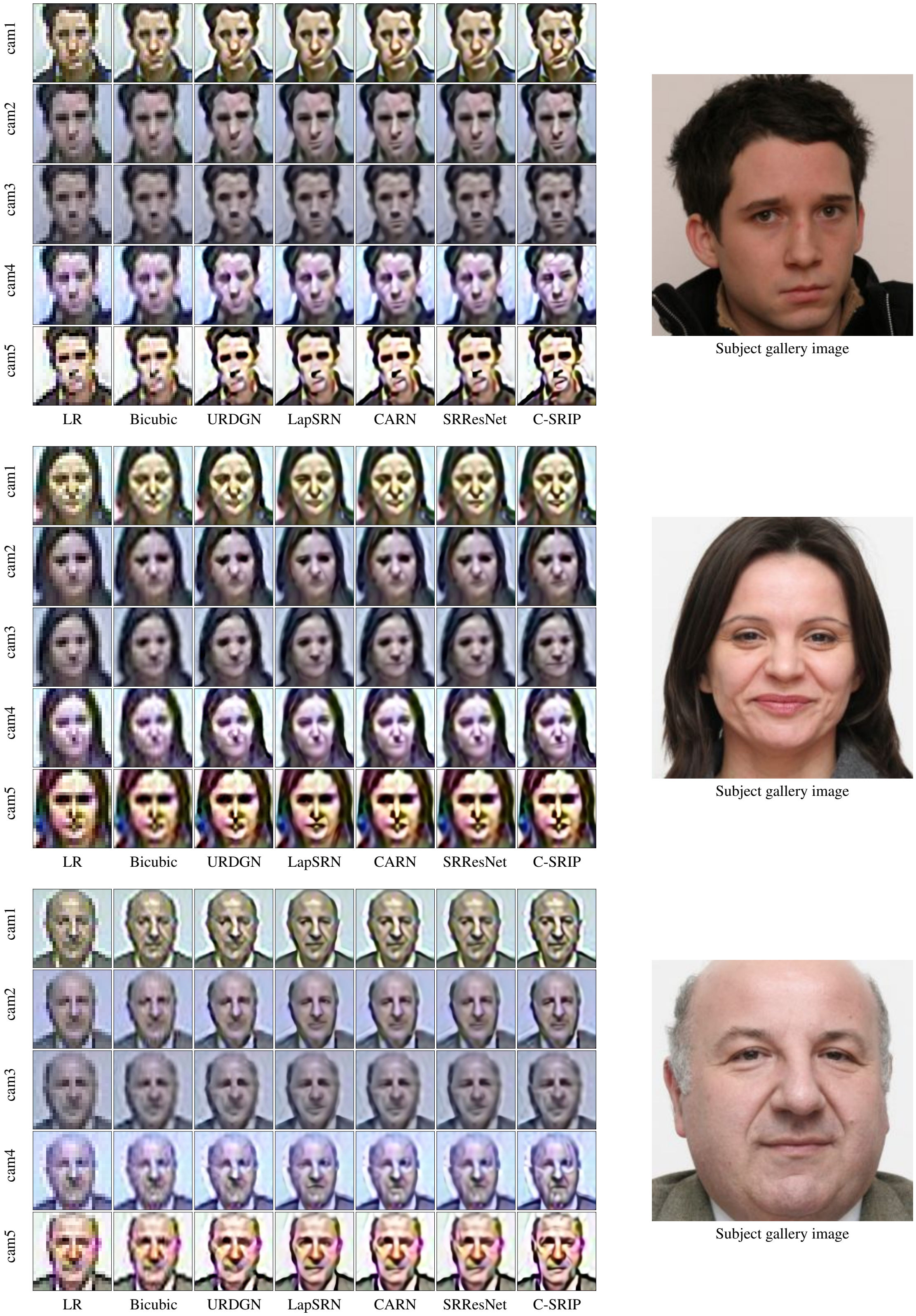


Figure 9. Examples of hallucinated HR faces from the SCFace [10] dataset. Results are presented for three subjects and separately for all five surveillance cameras. The FH models offer improvements in discernible facial details over the interpolation baseline, but introduce considerable distortions for some of the cameras (i.e., cameras 1 and 5). The figure is best viewed zoomed in.