APPENDIX

In this section, we present some additional results to further highlight the characteristics of the C-SRIP model. Similarly to the main part of the paper, we use LR images from the LFW [68], HELEN and CelebA datasets generated by smoothing and sub-sampling the original HR images. The inputs for all experiments are all of size 24×24 pixels.

A. Comparison to the state-of-the-art - additional results

In the main part of the paper, we present numerical result and visual examples of $8\times$ super-resolved images when comparing C-SRIP with competing models. Here, we show additional hallucination results (in Fig. 16) for all 9 FH models tested in the main part of paper. We again observe that the proposed C-SRIP model ensures the most convincing results among the tested models.

To get additional insight into the performance of the evaluated FH models we present in Fig. 17 Cumulative Score Distribution (CSD) curves of the PSNR, SSIM and VIF scores generated during the comparative experiments. Since SR models are increasingly focusing on learning-based techniques, which are expected to perform inconsistently across images of different characteristics, CSD curves provide a reasonable way of visualizing this performance variability. From the curves in Fig. 17 we see that all tested methods vary significantly in PSNR, SSIM and VIF scores across the LFW, Helen and CelebA datasets, with a large fraction of images producing sub-average performance scores. The ℓ_p and the proposed C-SRIP models are superior to other models and appear to have very similar performance in terms of the CSD curve for the PSNR score. However, the difference becomes significantly more apparent on the CSD curve for the SSIM and especially the VIF scores, where C-SRIP is clearly the top performer.

To further highlight the performance of C-SRIP compared to competing SR models, we show in Fig. 18 a couple of visual examples of the SR results for the top three performing SR models from our comparative assessment. As can be seen, the perceptual-loss-based SR model, ℓ_p, amplifies high-frequency noise, while the CARN model generates overly smooth results. C-SRIP, on the other hand, results in sharp images, but as expected is not able the recover all of the high frequency information (e.g., hair strains, wrinkles, beard details, etc.). Consequently, the subjects appear younger in the super-resolved images compared to the HR ground truths.

B. Generalization to smaller faces

Our model has a fully convolutional structure and, while it was trained to super-resolve 24×24 pixel images, it can in general process images of arbitrary input size. In the next series of experiments we, therefore, evaluate the ability of C-SRIP to upsample low-resolution facial images smaller than the 24×24 pixel images used for training. Specifically, we explore input image sizes of 20×20, 16×16, 12×12 and 10×10 pixels. We conduct experiments on the LFW data and down-sample the ground-truth images to $8\times$ the size of the query images to be able to quantify performance. We compare our model against those capable of accepting input images of arbitrary size - i.e., SRCNN, VDSR and CARN.

From the results in Fig. 19 and Table XI we see that the C-SRIP model is only able to generalize well at the 20×20 pixel input size. Below this size, it works similarly to other models - only super-resolving general geometric features in the image (as shown in Fig. 19), although it is still the top performer in terms of the average PSNR, SSIM and VIF scores.

C. Results for intermediate magnification factors

Because of space constraints in the main part of the paper, we show here additional results generated by the C-SRIP model for lower magnification factors, i.e., $2\times$ and $4\times$, that produce images of size 48×48 pixels and 96×96 pixels, respectively, given 24×24 pixel LR inputs. Note again that these images correspond to the intermediate results of the C-SRIP model and are generated by the first and second SR module of C-SRIP. A few illustrative SR examples generated for the $2\times$ and $4\times$ the input scale are presented in Fig. 20.

We observe that our model achieves realistic SR results even for small magnification factors. That is, even when the images are upscaled to a (still modest) size of 48×48 or 96×96 pixels, the hallucinated images preserve the identity of the subjects reasonably well, despite the limited performance of the SqueezNet models at these scales and, consequently, the relatively weak identity constraint applied during training. It needs to be noted that none of the presented subjects has been included in our training data.

D. Improving the visual quality of the hallucinated images

It is possible to further improve on the (perceived) visual quality of the SR images produced by the C-SRIP model (for large magnification factors of $8\times$) by utilizing simple image enhancement techniques. In Fig. 21 and Fig. 22 we show some examples, where a standard 3×3 sharpening filter (i.e., $[0, -1, 0; -1, 5, -1; 0, -1, 0]$) is applied on the SR outputs to amplify the high frequency components of the generated images. The result of applying such post-processing steps are significantly sharper and crisper SR images. However, in terms of summary statistics (i.e., average PSNR, SSIM and VIF scores) these are not competitive to the results reported in the main part of the paper - the sharpening operation deteriorates (quantitatively measured) performance. These results are in line with recent findings that suggest that there is a trade-off between the capability of SR models to either minimize distortion measures (i.e., maximize SSIM, PSNR or VIF scores) or to produce perceptually convincing results [76]. In Fig. 21 and Fig. 22 we show some sample images post-processed with a sharpening filter and include results for a couple of example images that were already presented in the main part of the paper to facilitate implicit comparisons with competing methods.

Interestingly, after the post-processing some of the SR images appear sharper than the original HR targets. This can be partially explained by the presence of noise in the target images that is not present in the SR reconstructions and the higher image contrast after enhancement that contributes towards the perception of higher-quality images.

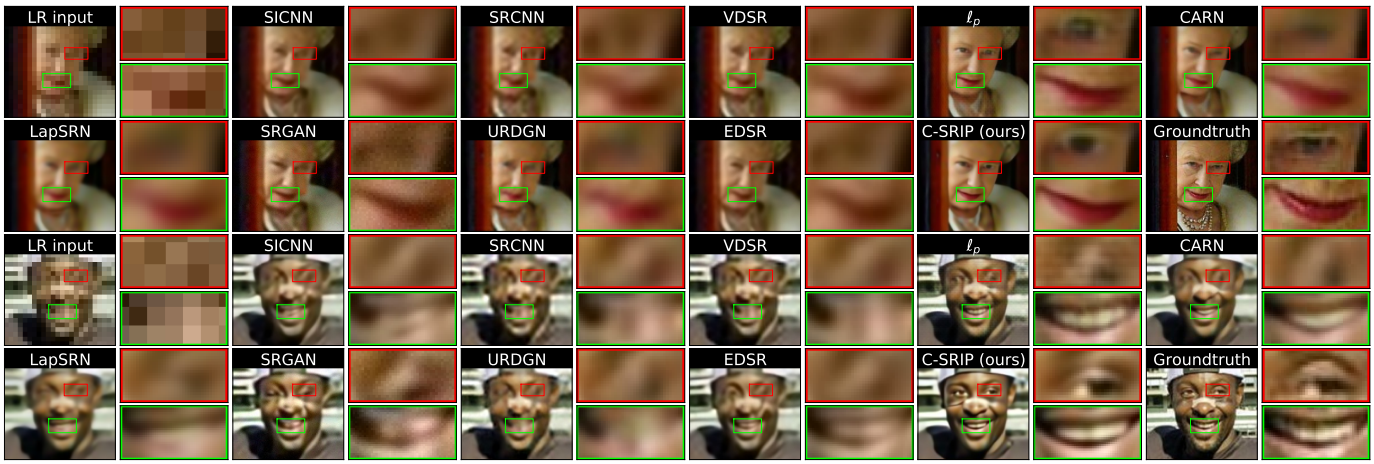

Fig. 16. Qualitative comparison of the evaluated SR models on two sample images with highlighted image details. Note the image details C-SRIP is able to recover compared to the competing models. The figure is best viewed zoomed in.

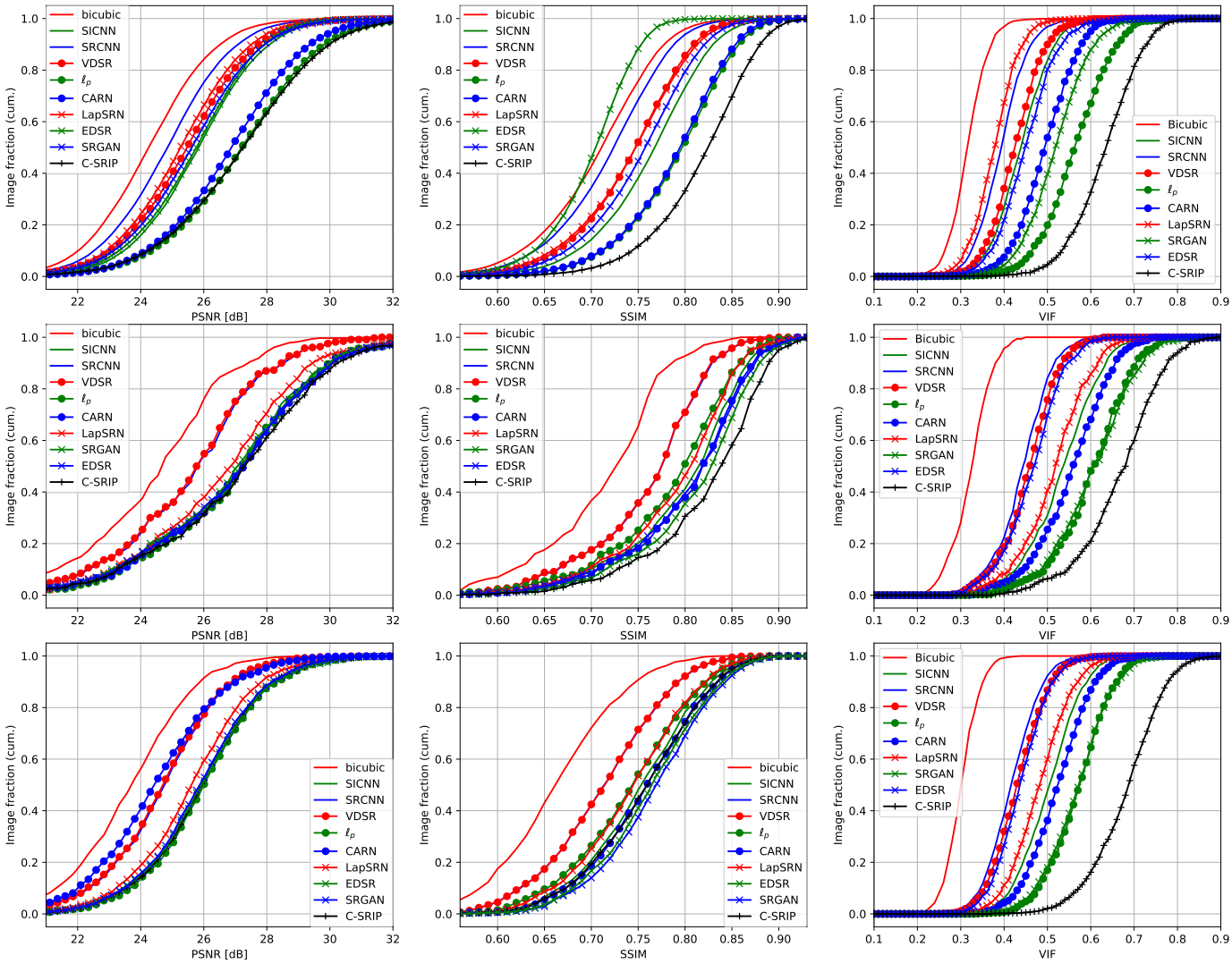

Fig. 17. Cumulative Score Distribution (CSD) curves for the PSNR (left), SSIM (middle) and VIF (right) scores over the LFW (top), Helen (middle) and CelebA (bottom) datasets generated using a magnification factor of $8\times$. Curves further to the right represent better performance on the given dataset. Note that C-SRIP is the top performer considering any of the performance measures and achieves by far the best VIF scores on all three datasets. The distribution of the performance measures (PSNR, SSIM and VIF) is relatively consistent across the datasets and across the tested super-resolution models. While all methods exhibit considerable score variability, the graphs still show that C-SRIP is able to achieve the highest performance for the majority of test images.

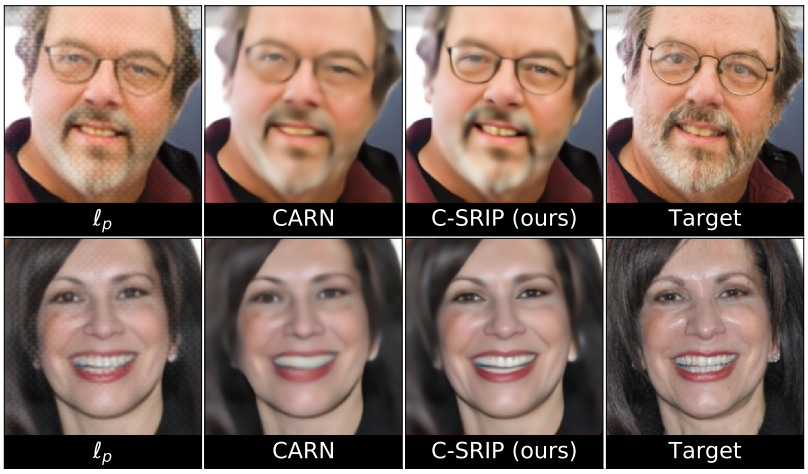

Fig. 18. Comparison of super-resolution results produced by the three best performing models of our assessment at a magnification factor of $8\times$. Bigger images are shown to better highlight the reconstructed image details. Best viewed zoomed in.

Fig. 19. Sample SR results generated with smaller input images. The left part of the figure shows the HR ground truth and the images on the right represent results for $8\times$ uscaling from (left to right): 10×10, 12×12, 16×16, 20×20 and 24×24 pixel images. Note that none of the models generalizes well to image sizes different from 24×24 pixels that was used for training.

TABLE XI
RESULTS FOR DIFFERENT INPUT IMAGE SIZES. THE BEST AND SECOND-BEST RESULTS ARE SHOWN IN RED AND BLUE, RESPECTIVELY.

Method	Input size [px]	PSNR	SSIM	VIF
SRCNN [15]	20×20	23.658	0.6438	0.2791
VDSR [12]	20×20	24.072	0.6642	0.2845
CARN [38]	20×20	24.174	0.7291	0.3127
C-SRIP (ours)	20×20	25.498	0.7751	0.3325
SRCNN [15]	16×16	22.088	0.6074	0.2659
VDSR [12]	16×16	22.315	0.6266	0.2705
CARN [38]	16×16	23.326	0.6854	0.2843
C-SRIP (ours)	16×16	23.674	0.7170	0.3206
SRCNN [15]	12×12	20.765	0.5351	0.2236
VDSR [12]	12×12	20.835	0.5297	0.2258
CARN [38]	12×12	21.931	0.6178	0.2631
C-SRIP (ours)	12×12	22.002	0.6540	0.2587
SRCNN [15]	10×10	19.947	0.4889	0.2414
VDSR [12]	10×10	20.041	0.5017	0.2128
CARN [38]	10×10	20.127	0.5624	0.2545
C-SRIP (ours)	10×10	20.935	0.6115	0.2387

TABLE XII
PSNR AND SSIM SCORES OBTAINED ON THE TRAINING DATA WITH THE MSE- AND SSIM-BASED LOSSES.

	MSE-based loss	SSIM-based loss
PNSR $[dB]$	28.3275	29.0227
SSIM	0.9189	0.9325

TABLE XIII
COMPARISON OF THE PSNR AND SSIM SCORES ON THE TEST DATA OBTAINED WITH THE MSE- AND SSIM-BASED LOSSES.

	MSE-based loss	SSIM-based loss
PSNR $[dB]$	26.1748	26.0251
SSIM	0.7547	0.7579

E. Quantitative results on the impact of the SSIM loss

Next, we present some (additional) quantitative results related to the proposed SSIM loss. Our SSIM formulation uses convolutions with a discrete Gaussian kernel, $\mathbf{g}$ - see Eq. (3), to approximate the local averages used with the original SSIM and is, therefore, easily implementable using standard deep learning frameworks. As emphasized in the main part of the paper, the result of using the proposed SSIM-based loss instead of the MSE-based loss are significantly better training characteristics in terms of faster convergence and lower PSNR and SSIM scores on the training data as shown in Table XII. Here, the results are presented for the simplest architecture from the ablation study (Section 4.3), where *i)* the images are processed through a series of 21 residual blocks, *ii)* all three upscaling layers are placed at the end of the SR network, and *iii)* supervision is applied only at the output of the model.

The proposed SSIM-based loss ensures significantly better performance scores during training. Even though the MSE-based loss is directly proportional to the PSNR score, our SSIM-based loss results in a lower average PSNR score on the training data, which suggests that a better optimum is found by the backpropagation-based learning procedure. On the test data the proposed loss still improves on the average SSIM and

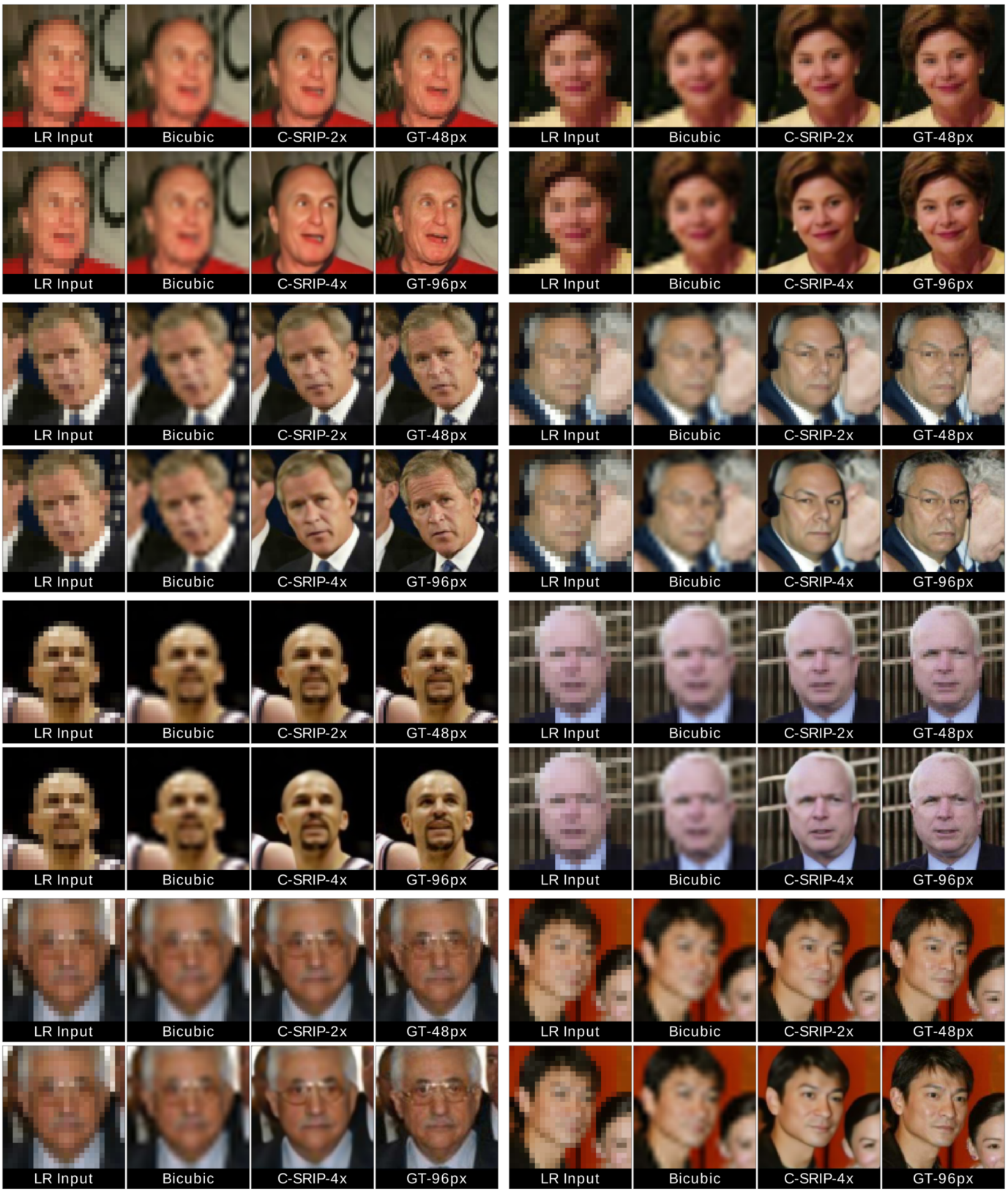

Fig. 20. Qualitative results for the intermediate scales generated by the C-SRIP model. The columns correspond to (from left to right): the 24×24 pixel input image, bicubic interpolation, results generated by C-SRIP (at a $2\times$ or $4\times$ upscaling factor) and the ground truth (GT) at either 48×48 or 96×96 pixels. Note how more detail is added as the upscaling factor gets larger.

VIF scores on all three experimental dataset, LFW, HELEN and CelebA, but offers no improvements in terms of PSNR value on LFW and HELEN, as shown in Table XIII - this fact is already highlighted in the ablation study of the main part

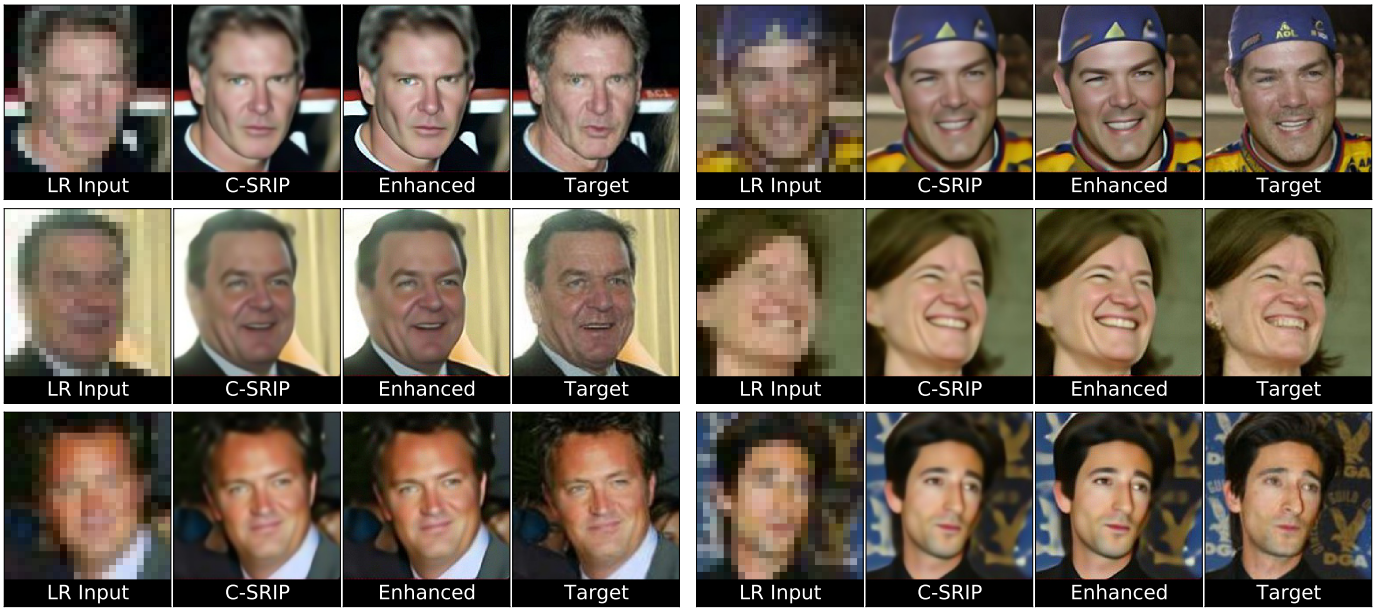

Fig. 21. Qualitative results for SR outputs post-processed with a standard image enhancement technique (i.e., with a sharpening filter). For each 24×24 LR input image (on the far left of each quadruplet) the following columns correspond to (from left to right): C-SRIP, C-SRIP with image enhancement, and the target HR image. Best viewed in high resolution.

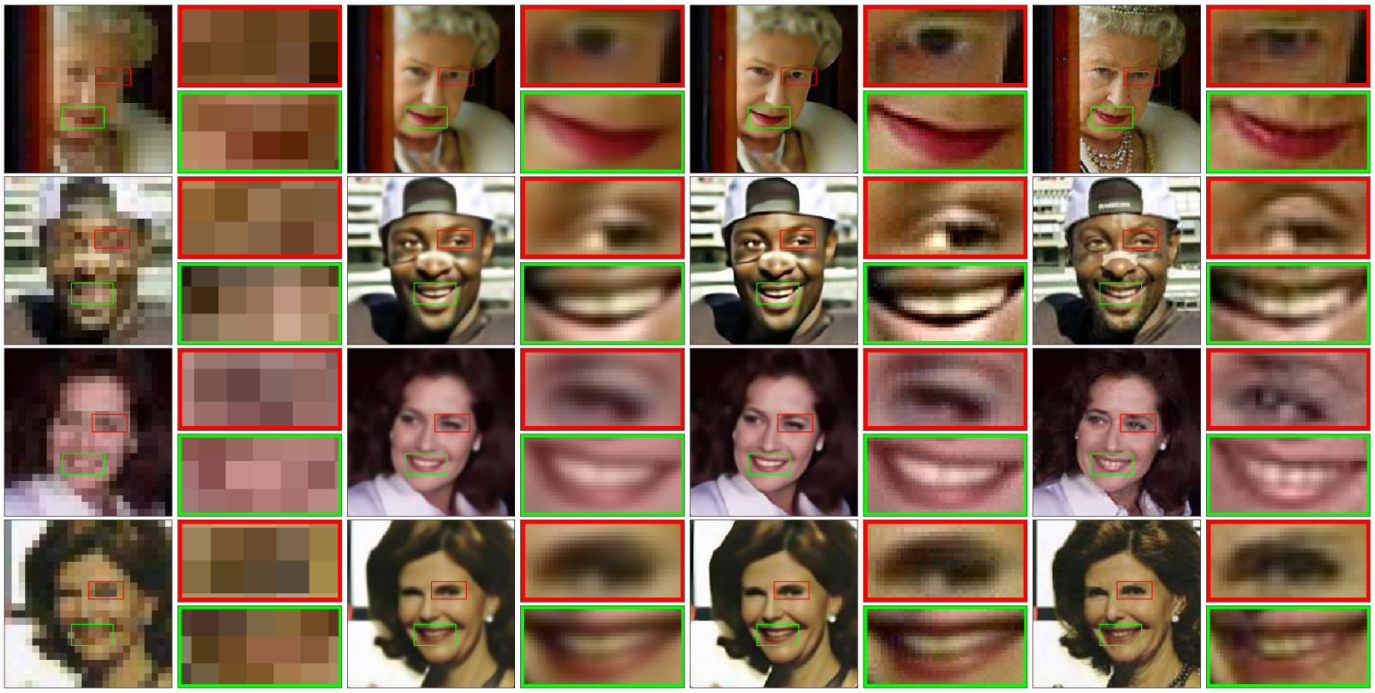

Fig. 22. Qualitative results for SR outputs post-processed with a standard image enhancement technique (i.e., with a sharpening filter) with highlighted image details. For each 24×24 LR input image (on the far left of each quadruplet) the following columns correspond to (from left to right): C-SRIP, C-SRIP with image enhancement and the target HR image. Best viewed in high resolution.

of the paper.

F. Reconstruction vs. recognition loss

To evaluate the importance of using both learning objectives (reconstruction and recognition) when training the SR network of C-SRIP, we train the SR network of C-SRIP in this section without the data-fidelity term and use only the recognition loss. The goal of this experiment is to assess whether good quality reconstruction could be generated by the supervision with the recognition networks alone. From the example results in Fig. 23 we see that the optimization procedure finds an optimum for the SR network parameters that does not result in meaningful HR reconstruction. We therefore conclude that the both learning objectives are important and are needed to

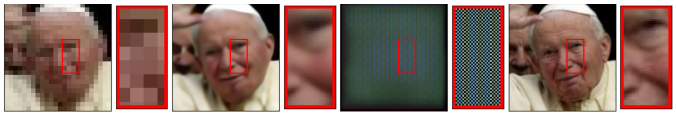

Fig. 23. Importance of using the reconstruction and recognition losses when training the SR network of C-SRIP. The figure shows (from left to right): the input LR image, the HR reconstruction generated by the SR network trained with both losses, the HR reconstruction generated by the SR network trained only with the recongition loss, the target HR image.

generate good quality HR images with C-SRIP.

G. Face hallucination performance on training identities

As described in Section IV-C, the parameters of the SR network of C-SRIP are learned with the help of a number of recognition networks, which are trained using the identities from the CASIA WebFace dataset. All experiments in the main part of the paper use images from datasets that have no overlap in terms of identities with the training data and, hence, demonstrate how the model generalizes to unseen identities. Nonetheless, these experiments leave an interesting research question unanswered, i.e.: *Has the model learned to better upsample identities included in the training data compared to identities not seen during training?*

To explore this question we collect a small dataset of 100 images (corresponding to 10 subjects) from the internet and make sure the images come from subjects also present in the CASIA WebFace dataset. We avoid duplicates with the training data by collecting only images that were captured and posted on the web after the WebFace data has been published. With this collection procedure we ensure that the collected dataset features the same identities as our training data, but not the same exact images. We denote this set of images as TRI when presenting results. Next, we randomly select a set of 100 images (of 10 subjects) from the LFW dataset and a set of 100 images (of 10 subjects) from the training data itself and denote these test sets as TRS and LFW, respectively. The created test sets exhibit different characteristics that allow us to evaluate the difference in face-hallucination performance when using images of subjects included in the training data and images of subjects that were not used during training, i.e.: *i)* TRS has been part of the training material, *ii)* TRI has the same subjects, but not the same images as used for training, and *iii)* LFW has no overlap in terms of images or subjects with the training data. We again perform experiments with 24×24 pixels inputs and the $8\times$ upscaling task.

From Table XIV we observe that images that were part of the training data (TRS) result in the best performance scores. This result is expected, as these images were directly involved in the optimization of the parameters of the SR network of C-SRIP. Images from the TRI set are reconstructed slightly worse, but still better than images of subjects that were not included in the training data. While the results for all three test sets are relatively close there is a consistent trend across the PSNR, SSIM and VIF scores that suggests that the performance of C-SRIP is somewhat better for images of identities that were part of the training data as opposed to images of subjects not seen during training.

TABLE XIV
MEAN PSNR, SSIM AND VIF SCORES GENERATED FOR THREE TEST SETS: *i)* A SET OF IMAGES THAT WAS PART OF THE TRAINING DATA (TRS), *ii)* A SET OF IMAGES THAT FEATURE THE SAME IDENTITIES AS THE TRAINING DATA, BUT NOT THE SAME SAMPLES/IMAGES (TRI), AND *iii)* A SET OF IMAGES FROM LFW (LFW) THAT HAS NO OVERLAP WITH THE TRAINING DATA IN TERMS OF IDENTITIES. C-SRIP SUPER-RESOLVES IMAGES OF TRAINING IDENTITIES SLIGHTLY BETTER THAN IMAGES OF IDENTITIES NOT SEEN DRUING TRAINING.

Method	PSNR	SSIM	VIF
Training samples (TRS)	27.565	0.8525	0.6503
Training identities (TRI)	27.382	0.8250	0.6419
LFW images (LFW)	27.091	0.8136	0.6245

A few visual examples of the face hallucination results for the three test sets are shown in Fig. 24. Here, the first row presents images from TRS, the second row shows images from TRI and the third row shows images from LFW. Note again how the quality of the reconstructions decreases slightly from the top to the bottom row examples.

H. Usefulness for recognition

The C-SRIP model is trained using a learning objective that combines (multi-scale) data-reconstruction and recognition-oriented losses. While we show in the main part of the paper that this contributes to better HR reconstructions, it should intuitively also contribute to improved recognition performance when the C-SRIP super-resolved images are used for recognition purposes.

To evaluate this hypothesis, we perform recognition experiments using the Labeled Faces in the Wild (LFW) dataset. We use the hallucinated images generated for the comparative assessment in Table III (see Fig. 6) in the main part of the paper for this experiment. Note that these images were generated from small 24×24 pixel inputs by upscaling them using a magnification factor of $8\times$. This setup allows us to directly evaluate the impact of the SR models on the recognition performance and to compare the performance achieved with the HR reconstruction with that ensured by the original HR images. The setup is also in line with standard evaluation methodology used with SR models [33].

We perform the recognition tests according to the standard LFW experimental protocol [68], which defines a 10-fold cross-validation experimental setup with 600 identity comparisons in each fold - equally balanced between genuine and impostor comparisons. We report the results in terms of verification accuracy in the form: $\mu \pm \sigma$, where μ is the average accuracy computed over the 10 experimental folds and σ is the corresponding standard deviation. We use the state-of-the-art ResNet-101 face recognition model trained with the large-margin cosine loss to extract 512-dimensional descriptors from each image and compare descriptors using the cosine similarity.

From the results in Table XV we see that the recognition model achieves competitive recognition performance with an average accuracy of 0.9806. The baseline bicubic interpolation is the worst performer among all tested methods with an average recognition accuracy of 0.8355, which shows that basic interpolation methods cannot recover much of the identity

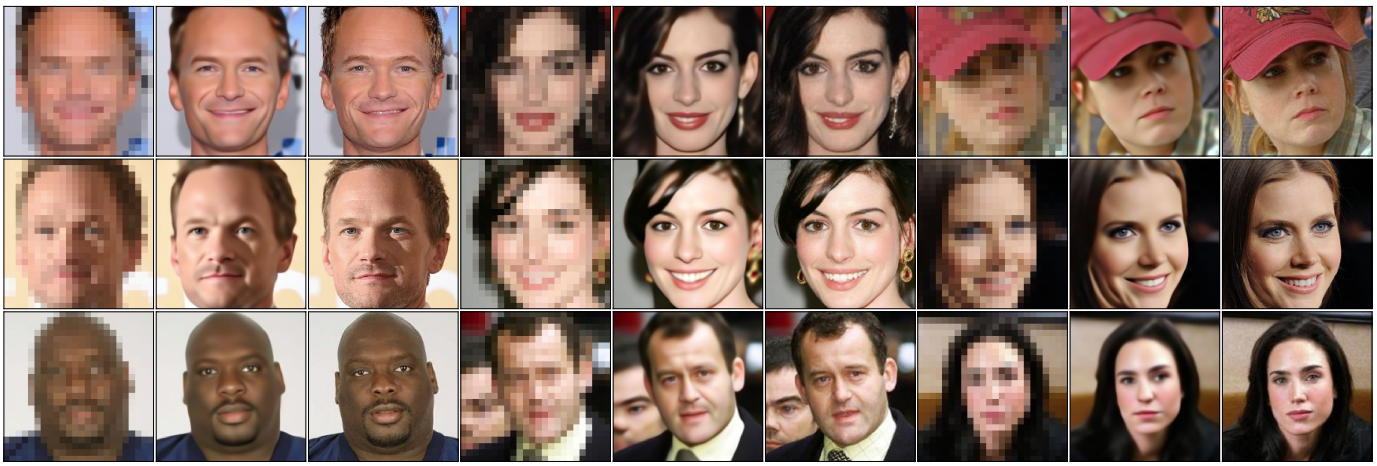

Fig. 24. Visual examples of hallucination results for the three test images sets. The figure shows face hallucination results for *i)* images from the training data (TRS, top row), *ii)* images that were not part of the training, but belong to subjects seen during training (TRI, middle row), and *iii)* images that have no overlap in terms of subjects with the training material (LFW, bottom row). Observe the slight decrease in hallucination quality from top to bottom.

TABLE XV
RESULTS OF THE LFW RECOGNITION EXPERIMENT. IMAGES SUPER-RESOLVED WITH C-SRIP ACHIEVE THE BEST OVERALL RESULT, SIGNIFICANTLY OUTPERFORMING THE NINE COMPETING MODELS. THE SR MODELS ARE ORDERED IN TERM OF INCREASING RECOGNITION PERFORMANCE.

Method	Verification accuracy ($\mu \pm \sigma$)
Bicubic	0.8355 ± 0.0077
LapSRN	0.8513 ± 0.0138
VDSR	0.8625 ± 0.0110
SRCNN	0.8627 ± 0.0134
SICNN	0.8802 ± 0.0107
URDGN	0.8875 ± 0.0116
EDSR	0.8904 ± 0.0129
ℓ_p	0.8917 ± 0.0105
CARN	0.8952 ± 0.0107
SRGAN	0.8990 ± 0.0107
C-SRIP	0.9217 ± 0.0099
HR images	0.9806 ± 0.0066

information from the LR input images. The super-resolution models, on the other hand, improve on this by a significant margin. Especially the SICNN, URDGN, EDSR, ℓ_p, CARN, SRGAN and C-SRIP model seem to be particularly effective. Interestingly, SICNN does not seem to have an advantage over competing face hallucination models, such as URDGN, EDSR, ℓ_p, CARN or SRGAN, despite the fact that it relies on identity information when learning to super-resolve faces. Overall, C-SRIP is the top performer in this experiment and ensures the highest recognition performance with an average verification accuracy of 0.9217. Nevertheless, a considerable gap still remains to the performance achieved with the original HR images, which suggests that not all of the useful identity information is recovered by the best performing model, C-SRIP.

I. Usefulness for facial landmarking

Another useful application of face hallucination models often advocated in the literature is facial landmarking (or alignment) of low-resolution facial data [4], [21], [77]–[79]. The idea here is to enhance the semantic content of the LR face images using face hallucination models with the goal of enabling more effective localization of salient facial features.

To demonstrate the usefulness of C-SRIP for this task, we perform a series of landmarking experiments using the landmarker from [79]. The landmarker aims to locate the standard set of 68 fiducial points in the face images and is trained on the training part of the Helen dataset that contains 2000 images with labelled locations of facial features. We use the 300 images from the Helen test set for the evaluation and first apply the landmarker on the original HR images to have a baseline for later comparisons. Next, we down-sample the HR images to a size of 24×24 pixels and finally upsample them using C-SRIP. To put the generated results into perspective, we repeat this procedure for all competing FH models already included in our previous experiments. We report all results in terms of the standard point-to-point error between the predicted and ground truth facial feature locations normalized by the inter-ocular distance [78].

As the results in Table XVI show, all hallucination models improve upon the baseline bicubic interpolation. Overall, C-SRIP again results in the best overall performance, followed closely by ℓ_p, SRGAN, EDSR and CARN. The ramaining models are less competitive. Interestingly, the order of the models is slightly different from the order in the recognition experiments in the previous section, which suggests that different aspects of the super-resolved images are important for the recognition and landmarking tasks.

In Fig. 25 some landmarking results are presented for images upsampled with different face hallucination models as well as for the baseline HR face images. Here, the ground truth facial feature locations are shown in green and the predicted landmarks are shown in red. The examples show that bicubic upsampling often leads to misdetected facial features, especially around the mouth area and facial outline, which are not clearly visible in the LR images. The face hallucination models, on the other hand, provide more semantic content and produce sharper edges around specific facial components, which is beneficial for the landmarking procedure.

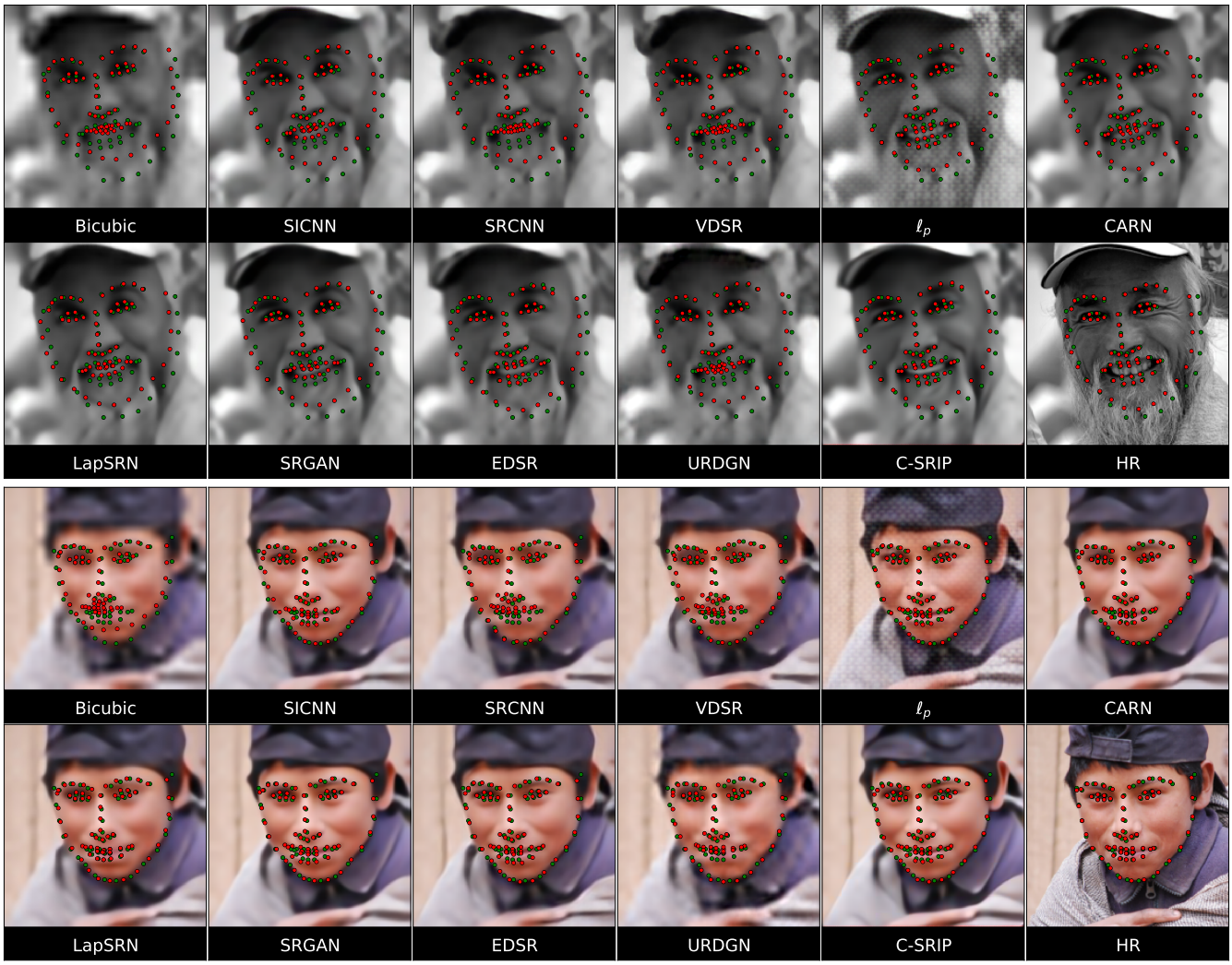

Fig. 25. Example landmarking results generated with super-resolved images produced by different face hallucination models. The ground truth landmarks are marked green and the predicted landmarks are shown in red. Observe how upsampling with bicubic interpolation often leads to misdetected fiducial points, especially along the facial outline and around the mouth area. The face hallucination models improve on this by recovering more facial details which helps with the landmarking performance. The figure is best viewed electronically.

TABLE XVI
RESULTS OF THE LANDMARKING EXPERIMENT ON THE HELEN DATASET. C-SRIP ENSURES THE OVERALL BEST LANDMARKING PERFORMANCE AMONG THE TESTED FACE HALLUCINATION MODELS. THE SR MODELS ARE ORDERED IN TERM OF DECREASING LANDMARKING ERROR.

Method	Error
Bicubic	0.0531
SRCNN	0.0502
VDSR	0.0502
URDGN	0.0487
LapSRN	0.0449
SICNN	0.0431
CARN	0.0417
EDSR	0.0409
SRGAN	0.0405
ℓ_p	0.0396
C-SRIP	0.0380
HR images	0.0344

J. More real-life examples

In Fig. 26 we show an additional example of faces super-resolved from a real-word image from the internet. The image presents a comparison with nearest neighbor and bicubic interpolation techniques and shows the added level of detail that can be recovered from the LR input images when using the proposed C-SRIP model.

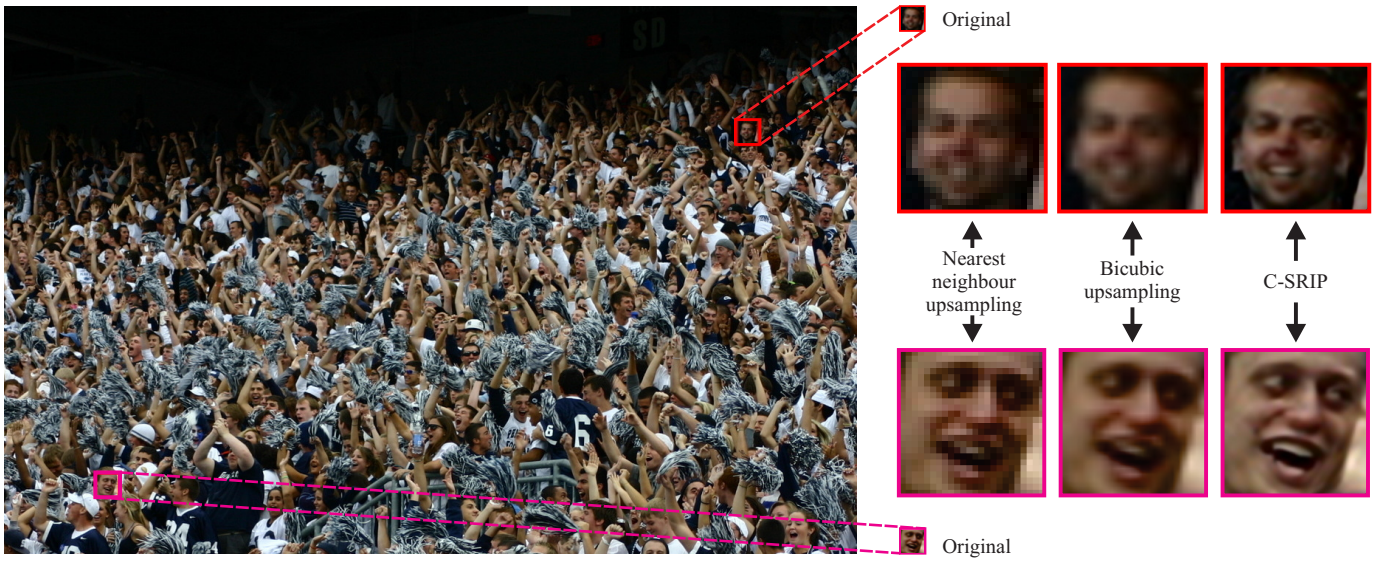

Fig. 26. Application of C-SRIP on a real-world image taken from the web. The image shows a crowd with several real-life LR faces. On the right side are super-resolution results generated with C-SRIP and two interpolation baselines for an upsacling factor of $8\times$. To illustrate the difficulty of the task, the LR input faces are also shown in the original size (marked "Original"). Note that C-SRIP is able to recover significantly more detail from the input LR images than the nearest neighbour and bicubic interpolation-based upsampling methods.