

Face Hallucination Using Cascaded Super-Resolution and Identity Priors

Klemen Grm^{id}, *Student Member, IEEE*, Walter J. Scheirer^{id}, *Senior Member, IEEE*,
and Vitomir Štruc^{id}, *Member, IEEE*

Abstract—In this paper we address the problem of hallucinating high-resolution facial images from low-resolution inputs at high magnification factors. We approach this task with convolutional neural networks (CNNs) and propose a novel (deep) face hallucination model that incorporates identity priors into the learning procedure. The model consists of two main parts: *i*) a cascaded super-resolution network that upscales the low-resolution facial images, and *ii*) an ensemble of face recognition models that act as identity priors for the super-resolution network during training. Different from most competing super-resolution techniques that rely on a single model for upscaling (even with large magnification factors), our network uses a cascade of multiple SR models that progressively upscale the low-resolution images using steps of $2\times$. This characteristic allows us to apply supervision signals (target appearances) at different resolutions and incorporate identity constraints at multiple-scales. The proposed C-SRIP model (Cascaded Super Resolution with Identity Priors) is able to upscale (tiny) low-resolution images captured in unconstrained conditions and produce visually convincing results for diverse low-resolution inputs. We rigorously evaluate the proposed model on the Labeled Faces in the Wild (LFW), Helen and CelebA datasets and report superior performance compared to the existing state-of-the-art.

Index Terms—Face hallucination, deep learning, CNN, identity.

I. INTRODUCTION

FACE hallucination (FH) represents a domain-specific super-resolution (SR) problem where the goal is to recover high-resolution (HR) facial images from low-resolution (LR) inputs [1]. Face hallucination techniques have important applications in various face-related vision tasks, such as face editing, face detection, 3D face reconstruction or face recognition [2]–[10], where they are used to counteract performance degradations caused by low-resolution input images.

Similarly to general single-image super-resolution tasks, face hallucination is inherently ill-posed. Given a fixed image-degradation model, every LR facial image can be shown to

Manuscript received February 12, 2019; revised June 26, 2019 and August 23, 2019; accepted September 18, 2019. Date of publication October 11, 2019; date of current version December 30, 2019. This work was supported in part by the ARRS Research Programme P2-0250 (B) Metrology and Biometric Systems and NVIDIA through a GPU Grant. The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Chia-Kai Liang. (*Corresponding author: Klemen Grm.*)

K. Grm and V. Štruc are with the Faculty of Electrical Engineering, University of Ljubljana, SI-1000 Ljubljana, Slovenia (e-mail: klemen.grm@fe.uni-lj.si; vitomir.struc@fe.uni-lj.si).

W. J. Scheirer is with the Department of Computer Science and Engineering, University of Notre Dame, South Bend, IN 46556, USA.

This article has supplementary downloadable material available at <http://ieeexplore.ieee.org>, provided by the authors.

Digital Object Identifier 10.1109/TIP.2019.2945835



Fig. 1. Face hallucination results generated with the C-SRIP model. The figure shows (from left to right): a 24×24 low-resolution (LR) input face image, the $8\times$ super-resolved (SR) image, and the high-resolution (HR) ground truth. Note that C-SRIP is able to ensure visually convincing super-resolution results.

have many possible HR counterparts. Thus, the solution space for FH problems is extremely large and recent models typically try to produce plausible SR results by learning to “hallucinate” high-frequency information using relationships between corresponding HR and LR images from a training dataset. While significant progress has been made in the area of learning-based (face) super-resolution over recent years [11]–[24], super-resolving facial images of arbitrary characteristics in a convincing manner, especially at high magnification factors, is still an unsolved problem, mainly due to:

- The ambiguous nature of the face hallucination task, where the solution space is known to grow exponentially with an increase in the desired magnification factor [9]. Despite strong reconstruction constraints it is exceptionally difficult to find good solutions and devise methods that work well for a broad range of LR facial images. Furthermore, even for domain-specific SR problems, such as face hallucination, where the solution space is already constrained by facial appearances, there is still an overwhelming number of plausible HR solutions that explain the observed LR input equally well.
- The difficulty of integrating strong priors into FH models that sufficiently constrain the solution space beyond solely the visual quality of the reconstructions. Most of the existing priors utilized for super-resolution relate to specific image characteristics, such as gradient distribution [25], total variation [26], smoothness [27] and the like, and hence focus on the perceptual quality of the super-resolved results. If discernibility of the semantic content (e.g., facial features) is the goal of the SR procedure, such priors may not be the most optimal choice, as they are not sufficiently task-oriented.

The outlined limitations are most evident for challenging face hallucination problems where tiny low-resolution images (e.g., of size 24×24 pixels) of arbitrary characteristics need to be super-resolved at high magnification factors (e.g., $8\times$). In this paper, we try to address some of these limitations with a new

hallucination model build around deep convolutional neural networks (CNNs).

Our model, called C-SRIP, uses a Cascade of simple Super-Resolution models (referred to as SR modules hereafter) for image upscaling and Identity Priors in the form of pretrained recognition networks as constraints for the training procedure. Thus, it combines a powerful (general-purpose) super-resolution network with prior domain knowledge related to face recognition. Specifically, our model uses multiple SR modules to super-resolve LR input images in magnification increments of $2\times$ and, consequently, allows for intermediate supervision at every scale. This intermediate supervision confines the explosion of the solution-space size and contributes towards more accurate hallucination results. To preserve identity-related features in the SR images, we incorporate pretrained recognition models into the training procedure, which act as identity constraints for the face hallucination problem. The recognition models are trained to respond only to the hallucinated high-frequency parts of the SR images and ensure that the added facial details are not only plausible, but as close to the true details as possible. Due to availability of intermediate SR results, we incorporate the identity constraints at multiple scales in C-SRIP. For data fidelity, we use a multi-scale loss derived from the structural similarity index (SSIM, [28]) that provides a stronger error signal for model training than the L_p -norm-based loss functions commonly used in this area. As we show through extensive experiments on the Labeled Faces in the Wild (LFW), Helen and CelebA datasets, the combination of reconstruction-oriented and identity-related losses results in visually convincing super-resolved face images that compare favourably with state-of-the-art FH models from literature.

The main motivation for using identity information in C-SRIP is to exploit high-level cues that relate to facial appearance (i.e., identity) in addition to commonly used pixel-level cues when learning to super-resolve facial images. By relying on an optimization objective that combines a data-reconstruction loss for data fidelity and a recognition loss for identity preservation we are able to use the best of both worlds and infuse the model with domain-knowledge that would be difficult to learn from pixel-comparisons alone.

In summary, we make the following contributions in this paper:

- 1) We introduce C-SRIP, a new CNN-based face hallucination model, that integrates identity priors at multiple scales into the training procedure of a SR network, and ensures state-of-the-art FH results. To the best of our knowledge, the model represents the first attempt to exploit *multi-scale identity information* to constrain the solution space of deep-learning based SR models.
- 2) We introduce a *cascaded SR network* architecture that super-resolves images in magnification steps of $2\times$ and offers a convenient and transparent way of incorporating supervision signals at multiple scales. Once trained, the SR network is able to hallucinate tiny unaligned 24×24 pixel LR images at magnification factors of $8\times$ and produce realistic and visually convincing hallucination results as illustrated in Fig. 1.

- 3) We propose a mechanism for integrating identity priors into FH models, which constrain the appearance of the hallucinated (high-frequency) facial details.
- 4) We make all models, weights and source code used in the experiments publicly available and provide the community with strong baselines for future FH research.

II. RELATED WORK

In this section we discuss recent work related to the C-SRIP model. The reader is referred to some of the existing surveys on super-resolution and face hallucination for a more comprehensive coverage of the field, e.g. [29]–[32].

A. Super-Resolution Models

Recent (single-image) super-resolution (SR) solutions are dominated by learning-based techniques that use pairs of corresponding HR and LR images to train machine learning models capable of predicting HR outputs from LR evidence [11]–[16]. The learning procedures used with these models typically aim to minimize an objective function that quantifies the error between the ground truth HR images and the SR predictions. Common objectives include the L_p , Huber or Lorentzian error-norm losses and more recent error measures that are closer to human image quality perception, such as structural similarity or CCN-based perceptual losses [17], [33], [34]. Our SR model follows the outlined learning paradigm, but incorporates a novel learning objective related to the concept of structural similarity [35] (SSIM). Specifically, it enforces a SSIM loss on the output of every SR module (i.e., on $2\times$, $4\times$ and $8\times$ super-resolved images) and naturally extends the loss to a multi-scale form.

The C-SRIP model is based on convolutional neural networks (CNNs) and in this sense is related to contemporary SR techniques that exploit CNNs for image upscaling, e.g., [12], [15], [17]–[24]. While these methods are capable of producing impressive SR results, the majority relies only on LR-HR image pairs for training and super-resolves images in a single step. However, recent developments [22], [36], [37] have shown the prospect of so-called cascaded models, where image upsampling is performed progressively using smaller steps (e.g., of $2\times$) to reach the overall magnification factor (e.g., $8\times$). This progressive upsampling strategy significantly constrains the solution space of the ill-posed super-resolution problem and contributes toward higher quality results. The reason for this, as argued by the authors [22], [36], [37], is the possibility of including intermediate supervision signals that help to find a better optimum during model learning. Similarly to these and related approaches [22], [36]–[39], C-SRIP also upscales LR inputs in a cascaded manner using carefully designed SR modules that increase the spatial dimension of the input images in steps of $2\times$, which in turn allows us to incorporate reconstruction and identity-related objectives at multiple scales into the training procedure.

Recent CNN-based SR models, (e.g., [12], [18]) exploit contemporary network architectures, such as ResNets [40] and Generative Adversarial Networks (GANs, [41]). These models are closely related to our work, as we also make heavy use of residual connections and incorporate a generative and a

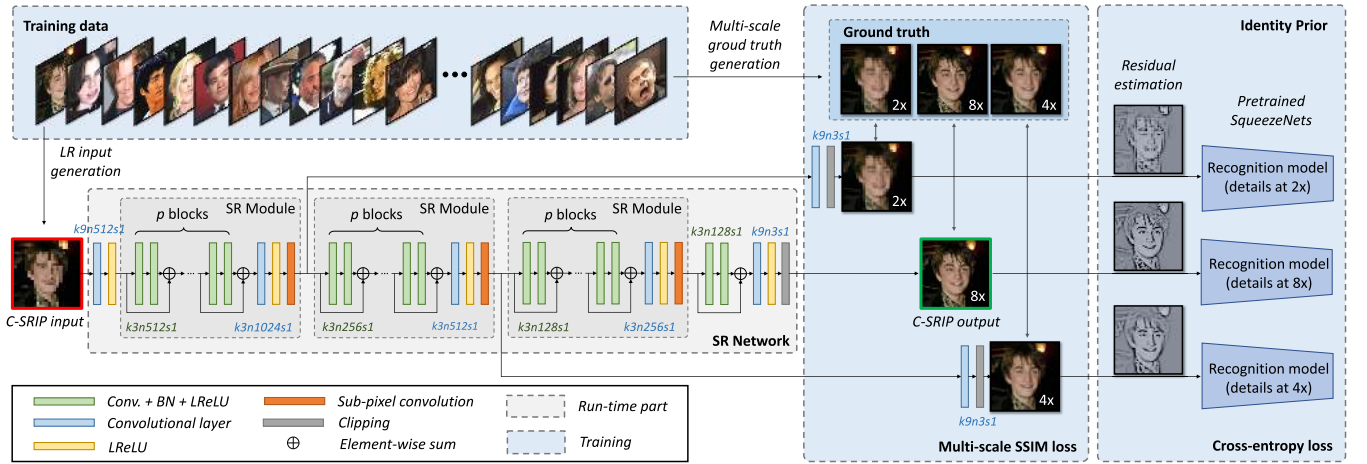


Fig. 2. Illustration of the C-SRIP model. The model consists of a generative SR network and an ensemble of face recognition models that serve as identity priors during training. The figure shows all architectural details and is best viewed electronically. The $kNnNs1$ notation introduced in [18] is used in the figure to denote a convolutional layer with N convolutional filters with $K \times K$ support, applied with stride S in the spatial dimensions.

discriminative network in our model. While we do not rely on GANs per se, our model does include a discriminative (classification) model that constrains the solution space of the generative SR network. However, our discriminative model is pre-trained and frozen and is not optimized alternatively with the generator, which (according to our preliminary experiments) greatly improves training stability and still results in realistic SR outputs. Finally, our work can also be seen as an extreme case of the perceptual-loss (ℓ_p) image transformation model from [17], which relies on comparisons of high-level features extracted from a pre-trained secondary network as the learning objective for SR, instead of comparisons at the pixel level. Our model follows a similar idea, but uses identity (i.e., information at the highest possible semantic level) to constrain the solution space of the generative SR network. Thus, instead of network features, C-SRIP considers the outputs of pre-trained recognition networks during training.

B. Face Hallucination and Identity Constraints

Different from general single-image SR tasks, the solution space of face hallucination (or face super-resolution) models is typically constrained to a set of plausible HR facial appearances. As a result, much better performance has been achieved with FH models at high magnification factors than with domain-agnostic SR models [42]. Similarly to other vision problems, research in FH is moving increasingly towards deep learning and numerous CNN-based FH models have been presented recently in the literature, e.g., [36], [42]–[52]. Here, we contribute to this body of work with a novel deep face hallucination model. While the SR network of our model is general and applicable to arbitrary input images, we infuse domain-specific knowledge into the model through the use of face images during training as well as through the pre-trained face recognition models that act as a source of prior information for the SR-model learning procedure.

Note that using identity information as a prior (or constraint) for SR models has been examined before [53], [54]. Henning-Yeomans et al. [55], for example, formulated a joint

optimization approach that maximized for super-resolution and face recognition performance simultaneously. This approach is conceptually similar to our work, but our approach is more general in the sense that it can be applied with any differentiable classification model. The approach from [55], on the other hand, is focused only on linear feature extraction techniques, e.g., PCA [56]. A CNN-based approach relying on identity information was recently proposed in [57]. Here, the authors proposed several different approaches for joint training of a face recognition and face hallucination network. However, they all involve separate loss functions for the two separate models. C-SRIP, on the other hand, tries to maximize the recognition performance of multiple pre-trained recognition models during training (via a cross-entropy loss) by propagating it through the super-resolution network, and, while pursuing a similar idea, is conceptually very different from the procedure in [57].

III. PROPOSED METHOD

In this section we describe the proposed C-SRIP face hallucination model and discuss its characteristics.

A. Overview of C-SRIP

As illustrated in Fig. 2, C-SRIP consists of two main components: *i*) a generative SR network for image upscaling, build around a powerful cascaded residual architecture, and *ii*) an ensemble of face recognition models that serve as a source of identity information during training.

Formally, C-SRIP aims to define a mapping, $f_{\theta_{SR}}$, from a LR input face image \mathbf{x} to a HR counterpart \mathbf{y} , i.e.

$$f_{\theta_{SR}} : \mathbf{x} \rightarrow \mathbf{y}, \quad (1)$$

where θ_{SR} denotes the set of C-SRIP parameters that need to be learned. To learn this mapping (i.e., the parameters θ_{SR} of the SR network), C-SRIP uses a combination of multi-scale SSIM and cross-entropy losses that jointly drive the training procedure. Details on the C-SRIP model and its training procedure are discussed in the following sections.

B. The Cascaded SR Network

The generative part of the C-SRIP model, the cascaded SR network, is a 52-layer CNN that takes a LR facial image as input and super-resolves it at a magnification factor of $8\times$. The network progressively upscales the LR input image using a cascaded series of so-called *SR modules*, where each module upscales the image only by a factor of $2\times$ (see Fig. 2). This progressive upscaling makes it possible to apply a loss function on the intermediate SR results and ensures better control of the training procedure in comparison to competing solutions that exploit supervision only at the final scale. Furthermore, the cascaded architecture allows us to solve a series of easier and better conditioned problems using repeated bottom-up inference with top-down supervision instead of one complex problem with an overwhelming amount of possible solutions.

We design the SR network around a fully-convolutional architecture that relies heavily on residual blocks [40] for all processing within one SR module and sub-pixel convolutions (i.e., expanding convolutional layers followed by pixel shuffle operations, [58]) for image upscaling. Our design choices are motivated by the success of fully-convolutional CNN models in various vision problems [40], [59], [60] and the state-of-the-art performance ensured by the sub-pixel convolutions in prior SR work [18], [58]. Similarly to [18], the residual blocks of the SR modules consist of two convolution–batch–norm–activation sub-blocks, followed by a post-activation element-wise sum. We ensure a constant memory footprint of all SR modules by decreasing the number of filters in the convolutional layers by a factor of 2 with every upscaling step. This maximizes the capacity of the network and balances the computational complexity across the SR modules. To upscale the feature maps at the output of each SR module, we rely on sub-pixel convolution layers proposed in [58]. These layers increase the spatial dimensions of the feature maps by reshuffling and aggregating pixels from multiple LR feature maps and, thus, for every upscaling step of $2\times$ reduce the number of available feature maps by a factor of 4. We counteract this effect by doubling the number of filters in the convolutional layer preceding the sub-pixel convolutions and, consequently, ensure that the capacity of the SR modules is not compromised due to the upscaling procedure. After reaching the target resolution, the feature maps are passed through one last residual block and a final convolutional layer (with 3 output channels) that produce the $8\times$ super-resolved output RGB image.

The network branches off after each SR module to allow for intermediate top-down supervision during training. Each branch applies a series of large-filter convolutions to produce intermediate SR resolution results at different scales (i.e., at $2\times$ and $4\times$ the initial scale) that are incorporated into the loss functions discussed in Section III-D. The large filter (9×9) convolutions are also applied at the beginning of the network to increase the model’s receptive field size. Details on the SR network architecture are given in Fig. 2 and Table I.

C. The Identity Prior

Using prior information to constrain the solution space of SR models during training is a key mechanism in the area of super-resolution [8], [25]–[27], [61], [62]. The main

TABLE I

ARCHITECTURE OF THE SR NETWORK. THE NETWORK CONSISTS OF A SERIES OF SR MODULES. THE STRUCTURE OF THE MODULES IS SHOWN IN BRACKETS IN THE FORM “[FILTER SIZE, NUMBER OF FILTERS, STRIDE]”

Layer No.	Network part	Output size	Layer type	Architectural details
1	Initial layer	24×24	Conv+LReLU	$[9 \times 9, 512, \text{stride } 1] \times 1$
		24×24	Conv+BN+LReLU	$[3 \times 3, 512, \text{stride } 1] \times 7$
2 – 17	SR module 1	24×24	Conv+BN+LReLU	$[3 \times 3, 512, \text{stride } 1] \times 1$
		24×24	Conv+LReLU	$[3 \times 3, 1024, \text{stride } 1] \times 1$
		48×48	Upsampling $2\times$	Sub-pixel convolution
		48×48	Conv+BN+LReLU	$[3 \times 3, 256, \text{stride } 1] \times 7$
18 – 33	SR module 2	48×48	Conv+BN+LReLU	$[3 \times 3, 256, \text{stride } 1] \times 1$
		48×48	Conv+LReLU	$[3 \times 3, 512, \text{stride } 1] \times 1$
		96×96	Upsampling $2\times$	Sub-pixel convolution
		96×96	Conv+BN+LReLU	$[3 \times 3, 128, \text{stride } 1] \times 7$
34 – 49	SR module 3	96×96	Conv+BN+LReLU	$[3 \times 3, 128, \text{stride } 1] \times 1$
		96×96	Conv+LReLU	$[3 \times 3, 256, \text{stride } 1] \times 1$
		192×192	Upsampling $2\times$	Sub-pixel convolution
		192×192	Conv+BN+LReLU	$[3 \times 3, 128, \text{stride } 1] \times 2$
40 – 52	Final layers	192×192	Conv+BN+LReLU	$[3 \times 3, 128, \text{stride } 1] \times 1$
		192×192	Conv+Clip	$[9 \times 9, 3, \text{stride } 1] \times 1$

*BN stands for batch normalization.

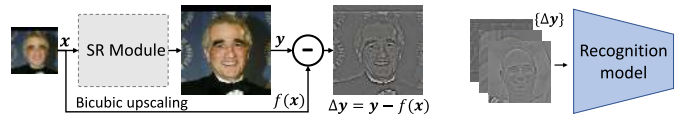


Fig. 3. Each SR module adds high-frequency facial details during upscaling (left). The recognition models are pretrained to respond to these details only (right) and are, therefore, used as identity priors during training.

motivation for incorporating priors into SR models is to provide a source of additional information for the learning procedure that supplements the data-fidelity objectives and contributes towards sharper and more accurate SR results.

An exceptionally strong prior in this context is identity. Because identity information relates to the semantic content (i.e.: Who is in the image?) and not the perceptual quality (i.e.: How visually convincing is the image?) of the SR images, it represents a natural choice for constraining the solution space of FH models. In fact, it seem intuitive to think about FH from both *i) an image-enhancement* as well as a *ii) content-preservation* perspective and to incorporate both views into the FH model for optimal results. While the image enhancement perspective is covered in C-SRIP by a reconstruction-based loss (see Section III-D), the content-preservation aspect is addressed through an ensemble of face recognition models that ensure that identity information is not altered during upscaling.

For C-SRIP we associate each recognition model with one of the SR modules and use it as an identity prior for the corresponding SR output, as illustrated in Fig. 2. Since each SR module can be shown to add only high-frequency details to the input images (see Fig. 3 left), we pre-train all recognition models to respond only to the hallucinated details and ignore the low-resolution content that is shared by the input and SR images (see Fig. 3 right). By focusing exclusively on the added details, we are able to directly link the recognition models to the desired SR outputs and penalize the results in case they alter the facial identity. This mechanism allows us to learn the parameters of the SR network by considering an identity-dependent loss in the overall learning objective.



Fig. 4. Training data generation. The figure shows (from left to right): an example of the training image quadruplets generated with Guassing blurring and decimation, residual images at three different spatial resolutions (corresponding to the residuals added by the $8\times$, $4\times$ and $2\times$ super-resolution steps). Note that the residuals are generated by subtracting a blurred version of the reference image at the given resolution from the original reference image.

While in principle any differentiable recognition model could be used as the identity prior for the FH model, we select SqueezeNets for this work [63]. The main reason for our choice is the lightweight architecture of SqueezeNet, which does not impose significant runtime slowdowns due to its relatively small memory and FLOPS footprint.

D. Training Details and SSIM Loss

We train the C-SRIP model in two stages. In the first stage, we learn the parameters of the SqueezeNet models for all three SR outputs (i.e., at $2\times$, $4\times$ and $8\times$ upscaling). In the second stage, we freeze the the weights of the recognition models and train the SR network with the combined (reconstruction and identity) loss. Details of both stages are presented next.

1) *Recognition-Model Training*: Next to LR and HR image pairs, we also require two intermediate reference images between the lowest and the highest resolution to learn the parameters of the recognition models and SR modules, as illustrated in Fig. 2. To this end, we take a training set of N high-resolution facial images $\{\mathbf{y}_i\}_{i=1}^N$ and apply a simple degradation model on the images to generate N image quadruplets for training, i.e., $\{\mathbf{x}_i, \mathbf{y}_i^{2\times}, \mathbf{y}_i^{4\times}, \mathbf{y}_i^{8\times}\}_{i=1}^N$, where \mathbf{x}_i represents the LR input image, $\mathbf{y}_i^{2\times}$ and $\mathbf{y}_i^{4\times}$ stand for the intermediate SR reference images at $2\times$ and $4\times$ the initial scale, respectively, and the HR image $\mathbf{y}_i^{8\times} = \mathbf{y}_i$ corresponds to the ground truth for the final $8\times$ super-resolved output. The degradation model uses Gaussian blurring and image decimation for down-sampling and produces training data (i.e., image quadruplets at different scales) as shown in Fig. 4 (left).

To train the recognition models, we construct residual images that reflect the facial details that need to be learned by the SR modules. The residual images, shown on the right side of Fig. 4, are computed by smoothing the ground truth images by a Gaussian kernel and subtracting the smoothed image from the original, i.e., $\Delta\mathbf{y}_i^j = \mathbf{y}_i^j - \mathbf{g} * \mathbf{y}_i^j$, for $j \in \{2\times, 4\times, 8\times\}$, where σ values of $\sigma_{2\times} = 1/3$, $\sigma_{4\times} = 1$ and $\sigma_{8\times} = 7/3$ are used with images at $2\times$, $4\times$, and $8\times$ the LR image size, respectively. We train the SqueezeNet models for classification based on the generated residual images using the categorical cross-entropy loss function L_{CE} :

$$L_{CE}(\theta_{SN}, \Delta\mathbf{y}) = - \sum_{k=1}^K p_{\Delta\mathbf{y}}(k) \log \hat{p}_{\Delta\mathbf{y}}(k), \quad (2)$$

where $p_{\Delta\mathbf{y}}$ denotes the ground truth class probability distribution of the residual image $\Delta\mathbf{y}$ (i.e., $p_{\Delta\mathbf{y}} \in \{0, 1\}^K$ is a class-encoded one-hot vector), $\hat{p}_{\Delta\mathbf{y}} \in \mathbb{R}^K$ stands for the output probability distribution produced by SqueezeNet's softmax layer based on $\Delta\mathbf{y}$, K stands for the number of classes in

the training data and θ_{SN} represents the parameters of the network. We learn the parameters of all three recognition models through backpropagation by minimizing the L_{CE} loss over the training dataset, i.e.:

$$\hat{\theta}_{SN}^j = \arg \min_{\theta_{SN}^j} \mathbb{E}_{\Delta\mathbf{y}^j} [L_{CE}(\theta_{SN}^j, \Delta\mathbf{y}^j)]. \quad (3)$$

The results of this first training stage are three SqueezeNet face recognition models (parameterized with $\hat{\theta}_{SN}^{2\times}$, $\hat{\theta}_{SN}^{4\times}$, $\hat{\theta}_{SN}^{hr}$), one for each image resolution, that respond only to the hallucinated facial details. These trained models are then frozen and serve as identity priors for the SR network.

2) *SR Network Training*: Standard reconstruction-oriented loss functions used for learning SR models, such as L_p error norms, are known to produce overly smooth and often blurry SR results [18]. We therefore design a new loss function for our SR network around the structural similarity index (SSIM, [28], [35]), and integrate it directly into our learning algorithm. Specifically, we use a novel multi-scale version of SSIM as a learning objective for the C-SRIP hallucination model.

Given a ground truth HR image $\mathbf{y}^{8\times}$ and the corresponding SR network prediction $\hat{\mathbf{y}}^{8\times} = f_{\theta_{SR}}(\mathbf{x})$, we first define a (single-scale) SSIM-based loss over the $8\times$ super-resolved image. Different from the original patch-based SSIM formulation from [28], we formulate SSIM using Gaussian kernels and convolutional operations that are easily implemented using common deep learning frameworks. Note that we drop the $8\times$ superscript in the equations to keep the notation uncluttered:

$$L_{SSIM}(\theta_{SR}, \mathbf{y}) = \frac{1}{2} \left(1 - \mathbb{E}_x [SS\hat{I}M(\mathbf{y}, \hat{\mathbf{y}})] \right), \quad (4)$$

where the SR network f is parameterized by θ_{SR} , $\mathbb{E}_x [\cdot]$ stands for the expectation operator over the spatial coordinates and $SS\hat{I}M(\mathbf{y}, \hat{\mathbf{y}})$ is a spatial similarity map between \mathbf{y} and $\hat{\mathbf{y}}$, i.e.:

$$SS\hat{I}M(\mathbf{y}, \hat{\mathbf{y}}) = \frac{(2\mu_{12} + C_1) \odot (2\sigma_{12} + C_2)}{(\mu_1^2 + \mu_2^2 + C_1) \odot (\sigma_1^2 + \sigma_2^2 + C_2)}, \quad (5)$$

where

$$\begin{aligned} \mu_1 &= \mathbf{y} * \mathbf{g}, & \mu_1^2 &= \mu_1 \odot \mu_1, \\ \mu_2 &= \hat{\mathbf{y}} * \mathbf{g}, & \mu_2^2 &= \mu_2 \odot \mu_2, \\ \sigma_1^2 &= (\mathbf{y} \odot \mathbf{y}) * \mathbf{g} - \mu_1^2, \\ \sigma_2^2 &= (\hat{\mathbf{y}} \odot \hat{\mathbf{y}}) * \mathbf{g} - \mu_2^2, \\ \mu_{12} &= \mu_1 \odot \mu_2, & \sigma_{12} &= (\mathbf{y} \odot \hat{\mathbf{y}}) * \mathbf{g} - \mu_{12}. \end{aligned}$$

In the above equations $*$ denotes the convolution operator, \odot denotes the Hadamard product, and the open parameters, \mathbf{g} , C_1 and C_2 , are defined as per the SSIM reference implementation (given in [28]), i.e., \mathbf{g} is an 11×11 Gaussian kernel with $\sigma = 1.5$ and $C_1 \approx 6.55$, $C_2 \approx 58.98$. If we define a similar

loss for the intermediate SR results at $2\times$ and $4\times$ the LR image size, we arrive at the final multi-scale form of SSIM that we use to learn the parameters of the SR network of C-SRIP, i.e.:

$$L_{MSSIM}(\theta_{SR}, \{\mathbf{y}^j\}) = \sum_{j \in \mathcal{D}} L_{SSIM}(\theta_{SR}, \mathbf{y}^j), \quad (6)$$

where $\mathcal{D} = \{2\times, 4\times, 8\times\}$. It needs to be noted that this multi-scale form of SSIM is different from existing multi-scale formulations of structural similarity (e.g., [35]), where images are down-sampled to capture image statistics at multiple resolutions. With our multi-scale SSIM formulation, structural similarity is measured between the ground truth images of different resolutions and the progressively upsampled LR images. As we discuss in the experimental section, the proposed loss results in better training characteristics compared to standard L_p norm based losses, which makes it easier to train (very) deep SR networks, such as the one devised for C-SRIP.

Based on the pre-trained SqueezeNet models and the loss introduced above, we define the overall loss of the C-SRIP model as follows:

$$L(\theta_{SR}, \{\mathbf{y}^j\}) = \sum_{j \in \mathcal{D}} L_{SSIM}(\theta_{SR}, \mathbf{y}^j) + \alpha L_{CE}(\theta_{SN}^j, \Delta \mathbf{y}^j), \quad (7)$$

where $\mathcal{D} = \{2\times, 4\times, hr\}$, α is a weight parameter that balances the relative impact of the reconstruction- and recognition-based losses and θ_{SR} stands for the parameters of the SR network that we aim to learn. The residual images $\Delta \mathbf{y}^j$ are constructed during training as illustrated in Fig. 4 (right). We use backpropagation to minimize the loss over our training data and find the parameters of the SR network $\hat{\theta}_{SR}$, i.e., $\hat{\theta}_{SR} = \arg \min_{\theta_{SR}} \mathbb{E}_{\mathbf{y}^j} [L(\theta_{SR}, \{\mathbf{y}^j\})]$.

Once the training is complete, we remove the recognition models and network branches used to generate the intermediate SR results at $2\times$ and $4\times$ magnification factors and use only the main output of the SR network for face hallucination. The final SR network takes a LR image \mathbf{x} of size 24×24 pixels as input and returns an $8\times$ upscaled 192×192 facial image $\mathbf{y}^{8\times}$ at the output.

E. Implementation Details

1) *Recognition Model*: The recognition models for all three output scales are implemented in accordance with the so-called *complex SqueezeNet* architecture from [63]. The models consist of 9 fire modules with intermediate shortcut connections, followed by a global average pooling layer and a softmax classifier on top. We train the first recognition model to classify residual images at $2\times$ the initial LR scale, i.e., 48×48 pixels, the second to classify images at $4\times$ the initial scale, i.e., 96×96 pixels, and the last for recognition of residual images of 192×192 pixels in size. To learn the model parameters we use backpropagation and the Adam [64] minibatch gradient descent algorithm, with a batch size of 128 and an initial learning rate of 10^{-4} . The learning rate is multiplied by a factor of $\frac{1}{3}$ every 20 epochs. To avoid over-fitting, we resort to data augmentation in the form of random horizontal flipping and random crops. We employ an

early stopping criterion based on accuracy improvements on the validation set. If no improvements are observed over 10 consecutive training epochs we stop the learning procedure and assume the recognition model has converged.

2) *The SR Network*: The SR network consist of three SR modules that are preceded by a convolutional layer with 512 large-scale filters of size 9×9 pixels. The SR modules are implemented with $p = 7$ residual blocks that contain 512 filters in the first SR module, 256 filters in the second SR module, and 128 filters in the last SR module, as shown in Fig. 2. We set the number of filters for the final convolutional layer of the SR modules, to 1024 for the first, 512 for the second and 256 for the third module. All filters are of size 3×3 pixels. For the activations, we use Leaky Rectified Linear Units (LReLU). The last residual block of the SR network has 128 filters 3×3 pixels in size. Before generating SR results at the output of the network and in the off-branches, a convolutional layer with three 9×9 filters is used followed by a clipping layer to ensure that the SR RGB images are within the valid intensity range of $[0, 255]$. A summary of the architecture is given in Table I.

IV. EXPERIMENTS

In this section, we present extensive experiments to validate the performance of our model. We start the section with a description of the datasets and performance metrics used for the evaluation. Next, we report comparative results with the state-of-the-art, conduct a fine-grained ablation study to highlight the impact of our contributions and finally explore the robustness and limitations of the proposed FH model.

A. Experimental Datasets

We select four popular face datasets for the experiments, i.e., CASIA WebFace [65], Labeled Faces in the Wild (LFW) [66], HELEN [66] and CelebA [67].

We use the CASIA WebFace dataset to learn the parameters of C-SRIP. The dataset contains a total of 494, 414 images of 10, 575 distinct identities, (i.e., $N = 494, 414$; $K = 10, 575$) and represents a mid-sized dataset very well suited for learning CNNs for various face-related vision tasks. Because the dataset ships with images of size 250×250 pixels that are relatively loosely cropped around the face, we take only the central 192×192 pixel patches of the images and use these as the basis for the experiments. Finally, we smooth the images using a Gaussian kernel (with $\sigma = 0.25 \times$ down-sampling factor) and sub-sample the images using bicubic interpolation. We do this several times for each image to produce the image quadruplets needed for training of the recognition models and the SR network of C-SRIP - see Fig. 4 for an illustration.

For testing, we use the complete Labeled Faces in the Wild (LFW) [66] dataset with 13, 233 facial images and 5, 749 subjects as well as images from HELEN [68] CelebA [67]. We select LFW for the experiments because it features images of variable quality captured in unconstrained conditions and thus represent a significant challenge for SR models. More importantly, it contains no overlap with CASIA WebFace in terms of identity, which is paramount to ensure a fair and unbiased evaluation of the C-SRIP model. The HELEN and



(a) CASIA WebFace examples

(b) LFW examples

(c) HELEN examples

(d) CelebA examples

Fig. 5. Visual examples of the pre-processed images used in the experiments: (a) CASIA WebFace, (b) LFW, (c) HELEN and (d) CelebA. The presented images are of size 192×192 pixels and represent the high-resolution ground truth. All images are cropped to contain only (or mostly) the facial area.

CelebA datasets, on the other hand, are selected to test the performance of C-SRIP on images of different characteristics than LFW and, hence, assess the generalization capabilities of our FH model.

We observe that the HELEN and CelebA datasets contain images of high resolutions, but also considerable amounts of JPEG-compression artifacts. Therefore, we take the following steps to preprocess the datasets. We first crop the facial regions using the provided landmark coordinates to achieve similar crops to the images present in the LFW and CASIA WebFace datasets. Next, we take the highest-resolution images from both datasets and down-sample them to 192×192 pixels using Gaussian blur and bicubic interpolation. We then treat the resulting square 192×192 pixel images as the target high-resolution images. With this procedure we process a total of 330 images from HELEN and 1126 images from CelebA that form the test set for our experiments. A comparison of the face images from the four datasets is shown in Fig. 5.

B. Performance Metrics

To measure the performance of the tested SR techniques we follow standard methodology from the literature [12], [16], [18], [37], [69] and report our results using:

- The *Peak Signal-to-Noise Ratio (PSNR)*, which is defined as follows:

$$PSNR(\mathbf{y}, \hat{\mathbf{y}}) = 20 \log_{10} \left(\frac{L}{\sqrt{MSE(\mathbf{y}, \hat{\mathbf{y}})}} \right) [dB], \quad (8)$$

where L is the maximum possible pixel value of an image (i.e., 255 for images stored with 8 bits per channel) and MSE is the mean squared error between the original high-resolution ground truth \mathbf{y} and the super-resolved image $\hat{\mathbf{y}}$. PSNR transforms the squared-error measure into the logarithmic space (in decibels) and considers only errors between individual pixels. It is defined in the range of $(0, \infty]$, where higher values indicate better resemblance between the ground-truth and the SR images.

- The *Structural Similarity (SSIM) index* given by [28]:

$$SSIM(\mathbf{y}, \hat{\mathbf{y}}) = \frac{1}{M} \sum_{i=1}^M s(\mathbf{y}_i, \hat{\mathbf{y}}_i), \quad (9)$$

where the local similarity function $s(\cdot, \cdot)$ that measures the structural similarity between the M image patches \mathbf{y}_i and $\hat{\mathbf{y}}_i$ (sampled from \mathbf{y} and $\hat{\mathbf{y}}$), is defined as

$$s(\mathbf{y}_i, \hat{\mathbf{y}}_i) = \frac{(2\mu_1\mu_2 + C_1)(2\sigma_{12} + C_2)}{(\mu_1^2 + \mu_2^2 + C_1)(\sigma_1^2 + \sigma_2^2 + C_2)}. \quad (10)$$

TABLE II

RECOGNITION PERFORMANCE OF THE SQUEEZE NET MODELS FOR DIFFERENT IMAGE SIZES OF THE TRAINING AND VALIDATION DATA. RESULTS ARE REPORTED IN TERMS OF RANK-1 RECOGNITION RATES

Model [†]	Image size [px]	Training data	Validation data
SqueezeNet at $2\times$	48×48	0.5138	0.2974
SqueezeNet at $4\times$	96×96	0.7215	0.4266
SqueezeNet at $8\times$	192×192	0.8569	0.5713

[†] Note that the models are trained to classify residual images.

In the above equation μ_1 and μ_2 denote the means of the local patches \mathbf{y}_i and $\hat{\mathbf{y}}_i$, σ_1^2 and σ_2^2 stand for their local variances, σ_{12} represents the local covariance of \mathbf{y}_i and $\hat{\mathbf{y}}_i$, and C_1 and C_2 are hyperparameters that are set based on the reference implementation of the SSIM authors, i.e., $C_1 = 6.55$, $C_2 = 58.98$. The valid range of the SSIM index is $(0, 1]$, where 1 indicates that the ground truth \mathbf{y} and the super-resolved image $\hat{\mathbf{y}}$ are identical.

- The *Visual Information Fidelity (VIF)* [70] which quantifies the fraction of the Shannon information in the wavelet domain that is shared between the ground truth face image \mathbf{y} and the super-resolution result $\hat{\mathbf{y}}$ relative to the information contained in \mathbf{y} . The range of output values for VIF is $(0, 1]$, where the value is 1 for identical ground truth and super-resolved images \mathbf{y} and $\hat{\mathbf{y}}$. It needs to be noted that most of the SR models included in the experimental evaluation are (implicitly) trained to maximize either PSNR or SSIM (by minimizing MSE or our SSIM-derived loss). Hence, we select VIF as a (third) unbiased performance measure for the experiments, as it is not directly related to PSNR or SSIM.

C. Model Training

Training of the C-SRIP model involves two sequential stages: *i*) training of the three SqueezeNet face recognition models that act as constraints for the C-SRIP learning procedure, and *ii*) training of the actual SR-network.

For the first stage (i.e., the SqueezeNet training) we randomly sample identities from CASIA WebFace, utilizing 90% of the images for training and 10% for validation. We use the standard cross-entropy loss and the Adam [64] optimization algorithm with an initial learning rate of 10^{-3} and an annealing factor of 10 every 50 epochs for the learning procedure. As shown by the results in Table II, the recognition models converge to the rank one recognition rate of 0.5138 (0.2974[†]) with 48×48 px residual images, 0.7215 (0.4266[†]) with 96×96 px residual images and 0.8569 (0.5713[†]) with



Fig. 6. Qualitative comparison with nine state-of-the-art SR models from the literature. The first two rows show sample results from LFW, the second two rows show results from HELEN and the last two rows show results from the CelebA dataset. The first column of each row shows the input 24×24 pixel LR image, upscaled with nearest neighbor interpolation for display purposes. The figure is best viewed zoomed in.

192×192 px residual images on the training (\dagger validation) data. As expected, the performance decreases with a decreasing size of the residual images and is adversely affected by the lack of low-frequency information during training (see, e.g., [72] for the expected performance of SqueezeNet for face recognition). Nevertheless, the models contribute towards accurate and visually convincing SR results, as evidenced by the results in the following sections.

In the second training stage, we fix the weights of the SqueezeNet models and learn the parameters of the SR-network of C-SRIP. Because we need identity labels in this stage as well, we again use the 90% vs. 10% data split per identity for training and validation. With this setup we train the SR network on 494,414 CASIA WebFace images using the objective in Eq. (7) that includes the SSIM-based image-reconstruction loss and the recognition performance of the SqueezeNet models. We balance the contribution of both loss terms with a value of $\alpha = 0.001$ and use backpropagation with the Adam [64] minibatch gradient descent algorithm for training. Due to the large memory footprint of the SR network and the face recognition models, we select a relatively small batch size of 8. The initial learning rate is set to $\frac{10}{3} \times 10^{-3}$ and is multiplied by $\frac{1}{3}$ at the end of epochs 10, 25, 50 and 80. The learning procedure is stopped early if both the SSIM and MSE values exhibit no improvements over 10 epochs.

We train all models on a workstation with two Nvidia GTX Titan Xp GPUs. On this hardware, the SqueezeNet training takes 1, 2, and 5 days, respectively, for the $2\times$, $4\times$ and $8\times$ scale models. The training of the SR network with the identity constraints included takes around 8 days. Once trained, the SR network is capable of processing images at an average speed

of 19 ms/image on GPU in batch mode, or 30 ms/image in real-time (i.e., single-sample batch) mode.

D. Comparison to the State-of-the-Art

We compare the C-SRIP model with 9 state-of-the-art SR and FH models, i.e.: the Super-identity convolutional neural network (SICNN) from [57], the Super-Resolution Convolutional Neural Network (SRCNN) from [15], the Very Deep Super Resolution Network (VDSR) from [12], the perceptual-loss based SR model (ℓ_p) from [17], the Cascading Residual Network (CARN) from [37], the Deep Laplacian Pyramid Super-Resolution Network (LapSRN) from [22], the Super-Resolution Generative Adversarial Network (SRGAN) from [18], the Enhanced Deep Residual Network (EDSR) from [71] and the Ultra Resolving Discriminative Generative Network (URDGN) from [42]. Since some of these models were introduced for general super-resolution problems, we re-train all models on the 494,414 CASIA WebFace dataset and use open-source implementations of the authors (where available) for a fair comparison. For ℓ_p we use features from the fire2, fire3 and fire4 layers of our full-scale SqueezeNet recognition network for the perceptual loss during training. We include results for bicubic interpolation, a standard image processing technique, as a baseline for the lower bound of the image reconstruction performance. To make our results reproducible, we make all code, model definitions and weights publicly available from <https://lmi.fe.uni-lj.si/en/research/fh/>.

1) *Comparison at the Highest Magnification Factor:* In our first series of experiments, we compare the performance of all SR models in the most challenging setting, i.e., with

TABLE III

COMPARISON OF C-SRIP WITH NINE STATE-OF-THE-ART SR MODELS ON THE MOST CHALLENGING TASKS, WHERE 24×24 PIXEL IMAGES ARE UPSCALED TO THE FINAL SIZE OF 192×192 PIXELS USING A MAGNIFICATION FACTOR OF $8\times$. THE BEST AND SECOND-BEST RESULTS ARE HIGHLIGHTED IN RED AND BLUE, RESPECTIVELY. OUR C-SRIP MODEL ATTAINS HIGHLY COMPETITIVE PERFORMANCE ON ALL THREE DATASETS

SR Model	LFW			HELEN			CelebA		
	PSNR	SSIM	VIF	PSNR	SSIM	VIF	PSNR	SSIM	VIF
Bicubic	24.256	0.7060	0.3157	24.410	0.7077	0.3231	23.215	0.6642	0.3020
SICNN [57]	25.857	0.7610	0.4387	26.855	0.7973	0.5232	25.916	0.7505	0.5028
SRCNN [15]	24.793	0.7211	0.3923	24.821	0.7146	0.4413	24.487	0.7019	0.4177
VDSR [12]	25.285	0.7361	0.4246	25.173	0.7406	0.4550	24.241	0.7265	0.4310
ℓ_p [17]	26.985	0.7897	0.5641	26.915	0.7951	0.6016	26.136	0.7735	0.5707
CARN [37]	26.811	0.7873	0.4938	26.618	0.7761	0.5537	25.972	0.7862	0.5228
LapSRN [22]	25.216	0.7330	0.4777	25.417	0.7513	0.5139	25.103	0.7365	0.4820
SRGAN [18]	25.669	0.6993	0.5181	26.047	0.7263	0.5682	25.830	0.7193	0.5684
URDGN [42]	25.575	0.7516	0.4494	26.882	0.7916	0.4639	25.136	0.7411	0.4379
EDSR [71]	25.648	0.7559	0.5381	25.317	0.7480	0.5396	25.909	0.7554	0.5418
C-SRIP (ours)	27.164	0.8171	0.6323	27.074	0.8235	0.6263	26.028	0.7945	0.6306

upsampling factors of $8\times$. The input to the models are 24×24 pixel images and the task is to generate 192×192 pixel outputs. In this experiment, a staggering amount of 98.43% of image pixels need to be hallucinated from the low-resolution inputs.

From the visual results in Figs. 6 we see that with such high magnification factors general SR models, such as SRCNN and VDSR, do not manage to generate convincing face hallucination results and amplify noise present in the LR images. These models fail to make use of the available facial context due to their relatively low receptive fields. The LapSRN and CARN models, which use a cascaded model topology similarly to C-SRIP, produce better results, but still struggle to produce crisp high-resolution face images. The EDSR [71] model is able to generate more facial details despite not including any priors or face-specific modifications, which is likely due to its deeper structure and higher model capacity. The SRGAN, URDGN, SICNN and ℓ_p models further improve on this by including secondary networks as constraints during SR training. ℓ_p is consistently the best-performing model included in our comparison, only slightly behind C-SRIP. However, we notice it often adds high-frequency noise when trying to minimize the perceptual loss of the convolutional maps of the secondary network. We speculate the reason our model is not susceptible to these errors is because it uses a global cross-entropy loss defined over the secondary recognition networks as opposed to the loss defined over local convolutional features exploited by ℓ_p . We also observe competitive performance for CARN, which performs slightly worse than C-SRIP and ℓ_p .

The findings made based on the visual results are also supported by the average PSNR, SSIM and VIF values reported in Table III. C-SRIP results in the best overall performance in terms of PSNR, SSIM and VIF values across all three datasets, followed by ℓ_p , CARN, SICNN, EDSR, and URDGN, which all produce strong performance metrics on the test datasets. While providing reasonably convincing visual results, SRGAN produces only average PSNR, SSIM and VIF scores and even results in the lowest SSIM score among all tested models on LFW. This result is expected and is observed regularly in the literature [18] with GAN-based SR methods. SRCNN, VDSR and LapSRN improve upon the Bicubic baseline in terms of

performance scores, but are less competitive in comparison to the top performers of this experiment.

2) *Comparison at Smaller Magnification Factors:* The architecture of C-SRIP allows us to super-resolve images at several magnification factors (i.e., at $2\times$, $4\times$, and $8\times$) in one forward pass through the model. To put the quality of the generated upscaling results at the smaller magnification factors into perspective, we re-train the nine competing models for the $2\times$ and $4\times$ upscaling tasks and report average PSNR, SSIM and VIF scores for the three datasets in Table IV. While we again use 24×24 pixel images as input, this problem is still easier than the one explored in the previous section, as less image content needs to be filled in by the SR models.

If we compare the reported results to the results in Table III we see that most methods achieve consistently higher performance scores as the magnification factor gets smaller. C-SRIP is again very competitive and achieves clearly the best performance among all tested methods for the $4\times$ upscaling task. For $2\times$ upscaling, the C-SRIP model never ranks worse than second, but is overall close to the runner up, the SRGAN model, in this experiment. Interestingly, while the SRGAN model was among the worst performers (in terms of performance metrics, not visual quality) on the more challenging $8\times$ upscaling problem, it is very competitive in these simpler tasks. However, we already see a collapse of the VIF score when going from the $2\times$ to the $4\times$ upscaling tasks for SRGAN - a trend that is even more evident in the transition from the $4\times$ to the $8\times$ upscaling problem. We also observe considerable performance from the ℓ_p , CARN, EDSR and SICNN models, which produce relatively competitive performance scores and result in visually solid HR reconstructions. SRCNN, VDSR, URDGN and LapSRN clearly outperform the baseline interpolation procedure, but produce lower average PSNR, SSIM and VIF scores on all three datasets compared to the best performing models. A visual comparison of all SR models with upscaling factors of $2\times$ and $4\times$ is shown in Fig. 7.

E. Ablation Study

In the next series of experiments, we perform an ablation study (for the $8\times$ upscaling problem) to assess the contribution of the individual components of the proposed C-SRIP model.

TABLE IV

COMPARISON WITH STATE-OF-THE-ART SR MODELS IN TERMS OF AVERAGE PSNR, SSIM AND VIF ACHIEVED ON LFW, HELEN AND CELEBA. THE TABLE SHOWS RESULTS FOR UPSCALING FACTORS OF 2 \times AND 4 \times WITH LOW-RESOLUTION 24 \times 24 PIXEL INPUT IMAGES. THE BEST AND SECOND-BEST RESULTS FOR EACH UPSCALING FACTOR ARE HIGHLIGHTED IN RED AND BLUE, RESPECTIVELY

SR Model	Scale	LFW			HELEN			CelebA		
		PSNR	SSIM	VIF	PSNR	SSIM	VIF	PSNR	SSIM	VIF
Bicubic	2 \times	27.275	0.8775	0.4851	27.511	0.8835	0.4867	26.684	0.8780	0.4895
SICNN [57]	2 \times	29.107	0.9275	0.6451	29.350	0.9245	0.6684	30.086	0.9317	0.6488
SRCNN [15]	2 \times	28.584	0.9142	0.5314	28.428	0.9133	0.5281	28.369	0.9164	0.5377
VDSR [12]	2 \times	28.717	0.9216	0.5688	28.993	0.9191	0.5745	28.832	0.9237	0.5843
ℓ_p [17]	2 \times	29.716	0.9306	0.6574	30.830	0.9481	0.6702	30.714	0.9358	0.6575
URDGN [42]	2 \times	28.616	0.9178	0.6173	28.727	0.9131	0.6348	28.769	0.9215	0.6283
LapSRN [22]	2 \times	28.611	0.9180	0.6195	28.503	0.9153	0.6035	29.017	0.9199	0.6207
CARN [37]	2 \times	29.862	0.9351	0.6309	30.141	0.9318	0.6447	30.311	0.9248	0.6514
SRGAN [18]	2 \times	30.254	0.9587	0.6831	31.412	0.9416	0.6720	30.572	0.9413	0.6685
EDSR [71]	2 \times	29.219	0.9251	0.6267	29.572	0.9106	0.6492	30.184	0.9335	0.6418
C-SRIP (ours)	2 \times	30.831	0.9459	0.6704	31.271	0.9551	0.6803	30.891	0.9525	0.6591
Bicubic	4 \times	24.829	0.7619	0.4086	24.950	0.7666	0.4128	23.956	0.7449	0.4018
SICNN [57]	4 \times	27.638	0.8398	0.6270	27.914	0.8425	0.5990	26.138	0.8340	0.6142
SRCNN [15]	4 \times	25.762	0.7941	0.4407	25.318	0.7865	0.4355	25.479	0.8084	0.4381
VDSR [12]	4 \times	25.875	0.8167	0.4385	25.576	0.7891	0.4236	25.816	0.8137	0.4413
ℓ_p [17]	4 \times	27.716	0.8553	0.6153	27.931	0.8745	0.6285	27.362	0.8691	0.6092
URDGN [42]	4 \times	25.989	0.8407	0.6420	25.678	0.8222	0.6471	25.958	0.8176	0.6379
LapSRN [22]	4 \times	25.897	0.8255	0.4931	25.725	0.8164	0.5132	26.016	0.8309	0.5518
CARN [37]	4 \times	27.734	0.8691	0.5538	28.018	0.8710	0.5843	27.460	0.8608	0.5476
SRGAN [18]	4 \times	27.839	0.8567	0.5216	28.052	0.8697	0.5396	27.454	0.8569	0.5405
EDSR [71]	4 \times	27.509	0.8621	0.6038	27.957	0.8679	0.6233	27.283	0.8471	0.5675
C-SRIP (ours)	4 \times	27.995	0.8769	0.6503	28.226	0.8880	0.6434	27.635	0.8777	0.6425



Fig. 7. Visual comparison of the SR results for magnification factors of 2 \times and 4 \times . The left block of images shows results for the magnification factor of 2 \times and the right block of images shows results for 4 \times . Note that C-SRIP achieves the most convincing visual results. GT-48px and GT-96px stands for ground truth images of size 48 \times 48 and 96 \times 96 pixels, respectively. The figure is best viewed electronically.

Towards this end, we train the following models using the methodology and data described in Section IV-C and evaluate their performance:

- *Baseline*: A baseline SR model without the cascaded SR modules and intermediate supervision. The model consist of 21 residual blocks similarly to the C-SRIP model, but the three sub-pixel convolution layers for upscaling are all placed at the end of the model. The model is trained using standard MSE loss. This model is in essence equivalent to the generator of the SRGAN approach from [18] and is included here to demonstrate the importance of the loss-functions and cascaded architecture used in C-SRIP.
- *B+SSIM*: The baseline SR model (Baseline), but trained with the proposed SSIM-based loss. This model is again equivalent to the SRGAN generator from [18] in terms of topology and is included in the here to show the impact of the loss-functions and cascaded architecture of C-SRIP.
- *C+SSIM*: The cascaded SR model, trained with the proposed SSIM-based loss, but without the identity priors and without multi-scale supervision i.e., the loss function is only applied at the output of the model. This model is used to demonstrate the effect of the cascaded architecture and the importance of multi-scale supervision.

TABLE V
SUMMARY OF MODEL CHARACTERISTICS USED IN THE ABLATION STUDY

Component	SSIM Loss	Cascaded	Multi-scale supervision	Identity prior
Baseline	\times	\times	\times	\times
B-SSIM	\checkmark	\times	\times	\times
C-SSIM	\checkmark	\checkmark	\times	\times
C-SSIM-M	\checkmark	\checkmark	\checkmark	\times
C-SRIP	\checkmark	\checkmark	\checkmark	\checkmark

- *C+SSIM+M*: The cascaded SR model, trained with multi-scale supervision and the proposed SSIM-based loss function, but without the identity priors. C+SSIM+M is included in the ablation study to highlight the importance of the of the multi-scale supervision, but also the identity prior used during C-SRIP training.
- *C-SRIP*: The C-SRIP model with multi-scale SSIM and identity supervision. The complete C-SRIP model shows the effect of putting all components together and, specifically, demonstrates the impact of the identity prior - when compared to C+SSIM+M.

The main model characteristics of the models used in the ablation study are summarized in Table V.

TABLE VI

RESULTS OF THE C-SRIP ABLATION STUDY. THE BEST AND SECOND-BEST RESULTS ARE HIGHLIGHTED IN RED AND BLUE, RESPECTIVELY. WE SEE THAT WITH EACH ADDITIONAL COMPONENT THE PERFORMANCE OF THE MODEL INCREASES ON AVERAGE. THE OVERALL BEST PERFORMANCE ACROSS ALL THREE DATASETS IS OBSERVED FOR THE COMPLETE C-SRIP MODEL

SR Model	LFW			HELEN			CelebA		
	PSNR	SSIM	VIF	PSNR	SSIM	VIF	PSNR	SSIM	VIF
Baseline	26.175	0.7547	0.5527	26.229	0.7597	0.5462	25.873	0.7345	0.5435
B-SSIM	26.025 ↓	0.7597 ↑	0.5764 ↑	25.964 ↓	0.7624 ↑	0.5714 ↑	25.924 ↑	0.7669 ↑	0.5692 ↑
C-SSIM	26.414 ↑	0.7731 ↑	0.6334 ↑	26.577 ↑	0.7638 ↑	0.6408 ↑	26.525 ↑	0.7719 ↑	0.6366 ↑
C-SSIM-M	26.451 ↑	0.7841 ↑	0.6575 ↑	26.669 ↑	0.7694 ↑	0.6613 ↑	26.313 ↑	0.7755 ↑	0.6632 ↑
C-SRIP	27.164 ↑	0.8171 ↑	0.6617 ↑	27.073 ↑	0.8235 ↑	0.6659 ↑	26.028 ↓	0.7945 ↑	0.6674 ↑

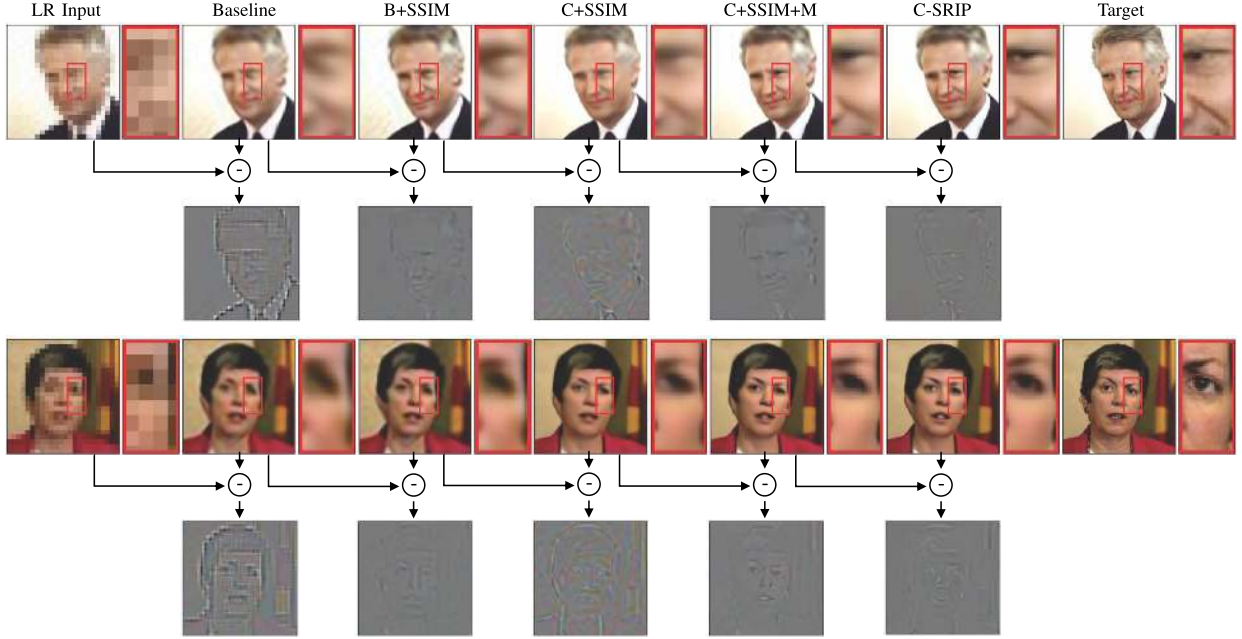


Fig. 8. Visual results of the ablation study. The figure shows examples of super-resolved images generated by the models included in the experiments (the top row of each example) and the details that are added by each model compared to the previous one (bottom row in each example). The images on the left (marked LR Input) show the low-resolution inputs upscaled using nearest neighbor (NN) interpolation. We see that the Baseline model already ensures better visual characteristics than the NN interpolation. We also observe significant jumps in visual quality when switching to the cascaded architecture (observe the increase in image sharpness in the zoomed in images) and when adding identity information (see, for example, the eye details in the first example image). The impact of including identity information is also clearly visible in the bottom row of each of the two examples, where the high-frequency details that are added when going from C+SSIM-M to C-SRIP are presented. Best viewed zoomed in.

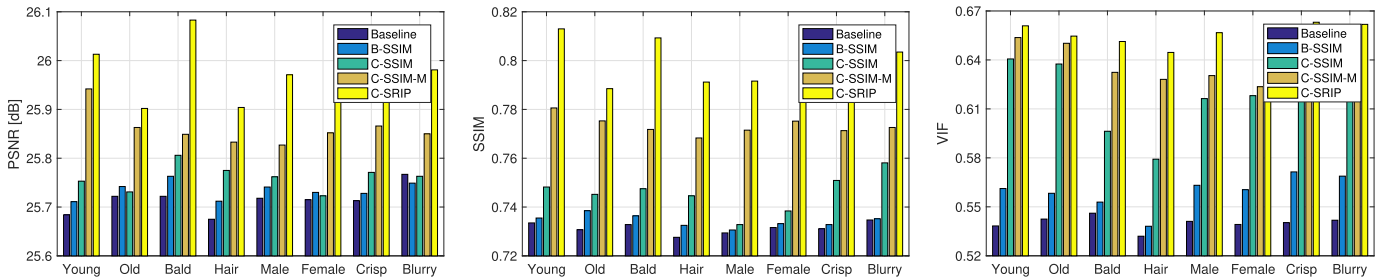


Fig. 9. Fine-grained ablation study. The graphs show (from left to right): average PSNR, SSIM, and VIF scores achieved across attribute-dependent subsets of the CelebA dataset. The individual subsets have different numbers of images, i.e.: young (437), old (319), bald (51), hair (83), male (212), female (326), crisp (247), blurry (52). Results show that the cascaded architecture, the multi-scale supervision and identity prior have the biggest impact on performance.

1) *Impact of C-SRIP Components:* The first thing to notice from the results in Table VI is that with each additional component, the performance of the model increases for the majority of performance metrics - as indicated by the arrow next to the performance scores. One performance decrease we see is when switching from the MSE loss (Baseline) to the SSIM-based loss (B-SSIM), which slightly lowers the average PSNR score on LFW and HELEN, but results in higher SSIM and VIF scores on all three datasets.

This result is expected, as PSNR is directly proportional to MSE and, thus, SR models optimizing for MSE typically achieve lower PSNR values than models using other loss functions. Nevertheless, overall the SSIM-based loss contributes towards improved performance and results in much better training characteristics, since our models converged faster and achieved significantly better SSIM and MSE scores on the training and validation data than the MSE-based models.

TABLE VII
EFFECT OF USING DIFFERENT LOSS FUNCTIONS TO TRAIN THE SR NETWORK OF C-SRIP

Model/Loss	LFW			HELEN			CelebA		
	PSNR	SSIM	VIF	PSNR	SSIM	VIF	PSNR	SSIM	VIF
Perceptual Loss [17] (PL)	25.695	0.7387	0.4472	25.602	0.7415	0.4352	25.738	0.7496	0.4673
Super-identity Loss [57] (SL)	26.944	0.7935	0.5860	26.726	0.8013	0.5602	26.503	0.7791	0.5585
Adversarial Loss [18] (AL)	26.164	0.7654	0.5531	25.947	0.7785	0.5496	25.408	0.7655	0.5712
C-SRIP - Proposed (P)	27.164	0.8171	0.6617	27.073	0.8235	0.6263	26.028	0.7945	0.6306
C-SRIP - No Residuals (NR)	27.213	0.8064	0.6605	27.018	0.8196	0.6308	26.071	0.7895	0.6258

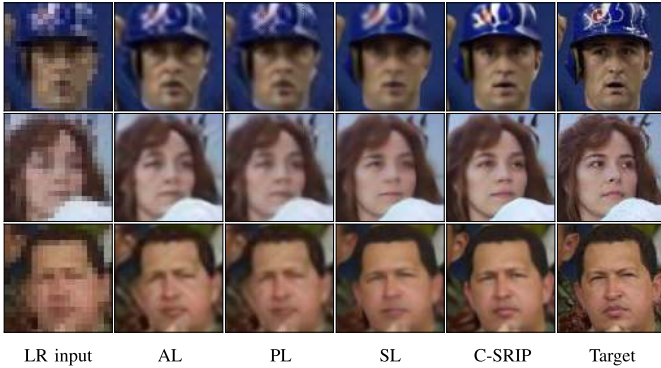


Fig. 10. Qualitative comparison of the effect of using different loss functions to train the SR network of the C-SRIP model. The recognition loss of C-SRIP ensures the most convincing results, followed closely by the super-identity loss from [57]. Best viewed zoomed in.

When looking at the impact of the cascaded architecture and multi-scale supervision (going from B-SSIM to C-SSIM and C-SSIM-M), we again observe considerable performance improvements for all performance scores on LFW, HELEN and CelebA. In fact, the change to the cascaded architecture has the biggest impact of the average VIF scores among all contributions on all three datasets.

On LFW and HELEN we see the biggest increase in the average PSNR and SSIM scores when adding the multi-scale identity supervision - see comparison between C-SSIM-M and C-SRIP. This addition also results in one of the biggest visual improvements of the SR images as seen in Fig. 8 - compare details (e.g., details around the eyes, etc.) in the zoomed in regions between C-SRIP and C+SSIM+M.

2) *Fine-Grained Ablation Study*: In order to investigate the performance of C-SRIP with respect to specific image characteristics and further assess the impact of the model components, we perform a fine-grained ablation study using the attribute labels of the CelebA dataset. Each label in CelebA is binary and indicates the presence or absence of attribute in the given image. We conduct our fine-grained ablation study using the following attributes:

- *Age or gender bias*: We are interested in whether C-SRIP performs differently on facial images of different age or gender groups and how the individual model components contribute towards the overall performance. To this end, we run experiments on subsets of the data based on the “male/female” and “young/old” attributes.
- *Image quality*: We are interested in how image reconstruction quality is affected when the ground truth image is of low quality. We, therefore, evaluate reconstruction performance on subsets of the dataset using the “blurry/crisp” attribute, respectively.

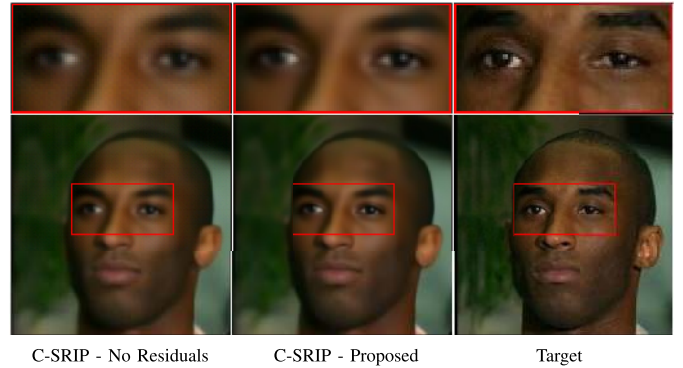


Fig. 11. Comparison of hallucination results generated by: i) applying the recognition networks directly on the ($2\times$, $4\times$ and $8\times$) hallucination outputs of the SR modules (left), and ii) using the recognition loss over the hallucinated residuals (middle column). Both approaches generated similar results, but the proposed C-SRIP variant with residuals produces slightly less noisy HR reconstructions - see zoomed in regions for details. Best viewed electronically.

- *Hair*: Hair is an obvious source of high-frequency details in face images. We aim at investigating how reconstruction performance is affected by its absence. To this end, we split the dataset using the “bald/hair” attribute, and evaluate each subset separately.

From the results in Fig. 9 we see that our model performs better on images of young people than the old, which is likely a consequence of smoother facial features with the young. We observe no significant gender bias in our model and interestingly also no significant difference between the performance with crisp and blurry ground truth images. As expected, our model performs slightly better on images of bald people than it does on images that contain hair, although in this case the number of samples in each class is again fairly small - see caption of Fig. 9.

In terms of contribution of the individual model components, the results are similar as in the previous section: the cascaded architecture results in the biggest performance increase in terms of the average VIF score across all image subsets, while the multi-scale supervision and identity constraints contribute towards the biggest performance increase when measured through the average PSNR and SSIM values.

3) *Evaluation of the Identity Loss*: We now evaluate the proposed identity loss in detail and compare it to other alternatives from the literature.

For the first experiment, we train multiple SR models using our cascaded SR network architecture and replace the C-SRIP recognition loss defined by Eq. (7) with competing losses from the literature. Specifically, we compare our loss with the following loss functions:

TABLE VIII

IMPACT OF THE IDENTITY PRIOR ON FACE HALLUCINATION PERFORMANCE WITH UPSAMPLING FACTORS OF $2\times$ AND $4\times$. RESULTS SHOW THAT USING IDENTITY INFORMATION DURING TRAINING IS LESS IMPORTANT FOR SIMPLER FH PROBLEMS WITH LOWER UPSAMPLING FACTORS

SR Model	Scale	LFW			HELEN			CelebA		
		PSNR	SSIM	VIF	PSNR	SSIM	VIF	PSNR	SSIM	VIF
C-SSIM-M	$2\times$	30.845	0.9437	0.6682	31.248	0.9572	0.6779	30.905	0.9538	0.6506
C-SRIP	$2\times$	30.831	0.9459	0.6704	31.271	0.9551	0.6803	30.891	0.9525	0.6591
C-SSIM-M	$4\times$	27.819	0.8673	0.6402	27.995	0.8764	0.6307	27.691	0.8753	0.6481
C-SRIP	$4\times$	27.995	0.8709	0.6503	28.226	0.8880	0.6434	27.635	0.8777	0.6425

TABLE IX

ROBUSTNESS OF C-SRIP TO FACIAL SCALE CHANGES - REPORTED IN TERMS OF AVERAGE PSNR, SSIM AND VIF SCORES

Crop scale	250px	240px	230px	220px	210px	200px	192px
PSNR	23.766	24.039	24.334	24.650	25.017	25.750	27.164
SSIM	0.7667	0.7725	0.7788	0.7857	0.7926	0.8063	0.8171
VIF	0.3194	0.3308	0.3587	0.3812	0.4018	0.4475	0.6323

TABLE X

ROBUSTNESS OF C-SRIP TO FACIAL ROTATIONS - REPORTED IN TERMS OF AVERAGE PSNR, SSIM AND VIF SCORES

Rotation	0°	15°	30°	45°	60°	75°	90°
PSNR	27.164	25.557	24.850	24.627	24.539	24.794	24.786
SSIM	0.8171	0.8044	0.7937	0.7825	0.7793	0.7768	0.7728
VIF	0.6617	0.4631	0.3995	0.3917	0.3834	0.3800	0.3816

- *Perceptual loss (PL)*: This loss penalizes the difference between low-level (fire2 and fire3 layer) feature representations of the super-resolved and reference HR images within the pretrained SqueezeNet face recognition model.
- *Super-identity loss (SL)*: Here, we adopt the super-identity training framework from [57]. Specifically, we train the super-resolution and recognition networks from scratch and learn them concurrently with the so-called super-identity learning objective, which is a combination of a pixel-wise MSE loss, a MSE loss between normalized high-level embeddings, and a face recognition loss. We use the authors' method of training the hallucination and recognition methods interchangeably in each iteration.
- *Adversarial loss (AL)*: We also train our super-resolution network using the GAN framework proposed for super-resolution by [18]. Here, we use a shallow 8 layer CNN model as the discriminator for the adversarial training to improve the training stability in the adversarial setting.

From the results in Table VII and Fig. 10 we see that the proposed recognition loss is best suited for our SR network architecture, as the C-SRIP again produces the highest quality HR reconstructions. Similarly to the original ℓ_p model, our SR network trained with the perceptual loss learns to resolve some facial details, but again results in a high-frequency pattern that overlays the HR reconstructions. The super-identity loss generates visually convincing HR reconstructions, but performs somewhat worse than C-SRIP. The model trained with the adversarial loss performs slightly better than the model trained with the perceptual loss and the SRGAN model used in the comparative experiments in Section IV-D.

In our second experiment, we examine the impact of feeding the hallucinated residuals instead of complete super-resolution output to the pretrained recognition models when learning



Fig. 12. Robustness to changes in facial scale: HR reconstructions generated from 24×24 LR faces of different size (top), HR ground truth (bottom). The figure on the left corresponds to the training setting (i.e., 192px crop).



Fig. 13. Robustness of C-SRIP to image rotations. The top row shows the HR reconstructions generated from rotated 24×24 LR faces. The bottom row shows the HR ground truth.

the C-SRIP SR network. To this end, we retrain all three face recognition networks (for $2\times$, $4\times$ and $8\times$ magnification factors) on complete face images (instead of using only the hallucinated high-frequency residuals) and use them to train the C-SRIP model from scratch. We again use the multi-scale SSIM loss as our data fidelity term.

The comparison of both C-SRIP variants is presented in Table VII and Fig. 11. We observe that both C-SRIP variants performs similarly well both in terms of performance scores on all three test datasets, as well as in terms of visual comparison. We do notice, however, that the C-SRIP variant trained with complete images (i.e., without penalizing the residuals) produces slightly noisier results on average, which can be seen from the zoomed in region at the top of Fig. 11.

So far, we have evaluated the impact of the identity prior only for the $8\times$ upsampling task. In the third experiment of this series, we examine the impact of the recognition loss for smaller upsampling factors, i.e., $2\times$ and $4\times$. In Table VIII we show a comparison of the performance scores achieved when using the C-SSIM-M (cascaded architecture + multi scale SSIM supervision) and C-SRIP (all components including the identity prior) models. Interestingly, adding identity information for the $2\times$ upsampling tasks does not seem to help much over a pure reconstruction loss. For the $4\times$ upsampling tasks results do improve, but not as much as observed in the most challenging scenario - the $8\times$ upsampling problems. These results suggest that the identity prior becomes important as the hallucination problem gets harder. For these challenging problems the identity information provides additional cues that contribute to higher-quality FH results, for easier problems,



Fig. 14. Examples of poor SR results produced by the C-SRIP model considering PSNR, SSIM and VIF scores. The four columns of each image group correspond to (from left to right): the input LR image, bicubic interpolation, C-SRIP and the target HR image. The captions provide information on the possible reason for the weak performance.

on the other hand, the identity prior is not as effective. The reported performance scores also provide insight into the results from Table IV, where we found the C-SRIP model to be less dominant compared to other models for smaller magnification factors.

F. Robustness of C-SRIP

Face detection is a necessary first pre-processing step when considering face hallucination in a real-life deployment scenario. In this step, the face images may be detected at different scales and under different rotations. In our next experiments, we are interested in how well our model can handle variations in scale and orientation of the input faces. To this end, we perform two experiments, where we systematically vary the scale of the faces within the image and where we rotate the cropped low-resolution faces around their center. We use images from the LFW dataset for this experiment. For the scale experiment, we use differently sized crops from the LFW dataset. We start with the training setting of 192×192 pixels and gradually increase the size of the crops to the final size of 250×250 pixels. We then rescale the images to a fixed input size for the C-SRIP model of 24×24 pixels. For the rotation experiment, we rotate images counter-clockwise from 0° to 90° with a step size of 15° and observe differences in performance. The generated scale and rotation variations clearly exceed the variability typically induced by a face detector, but help to demonstrate the behaviour of C-SRIP under extreme scale and rotation changes.

From the results in Tables IX and X we observe that C-SRIP performs relatively well for settings that are close to the training setup, but start to degrade in performance when larger deviations from the training setting are present. Nonetheless, after an initial drop in performance additional scale and rotation changes have only a limited effect on performance. If we look at the example hallucination results in Figs. 12 and 13, we see that relatively convincing reconstructions are achieved for the first two or three scale and rotation variations, but the results clearly (visually) deteriorate as the difference to the training setup gets larger.

The reason for the performance drop, we believe, can be found in the characteristics of the training data, which contains mostly frontal upright faces with minor scale and

rotation variations. Our model naturally learns to best super-resolve images matching the training setup and deteriorates in performance with major deviations from the training characteristics. However, note that the robustness to variations in scale and rotations could be improved, e.g., by incorporating additional alignment procedures into the model, similarly to [51], [73].

G. Limitations of C-SRIP

To evaluate the weaknesses of the C-SRIP model, we examine in Fig. 14 a few example images that result in the worst SR results on the datasets used in our experiments. We identify a few potential reasons for the poor SR performance:

- *High-frequency details not related to the face.* Image 14(a), contains a great amount of high-frequency details (background, hair). Our SR network is guided by face-recognition models that ignore non-face regions.
- *Significant occlusion.* In images 14(a) and 14(d), the face is partially occluded by a foreground object. The occlusion changes the global facial appearance, which adversely affects C-SRIP's reconstruction capabilities.
- *Significant pose variations.* In 14(c), the subject's face is partially obscured due to the profile pose. Few samples in our training dataset feature profile poses, which deteriorates performance on this type of facial images.
- *Low-quality HR image.* Image 14(b) has a significant amount of noise, which is reduced during down-sampling and cannot be reconstructed.

H. Qualitative Results on Real-World Images

The results presented so far have focused on images that were artificially down-sampled using Gaussian blurring and image sub-sampling. This is a standard approach used in the super-resolution literature needed to quantify the performance of the trained upsampling models. In this last section, we use a few example images from the web and upscale selected faces using C-SRIP and a couple of baseline techniques. Note that this task is significantly more challenging than the experiments presented in the previous sections, as the degradation function that generated the LR images has not been used to train the SR models. Since no ground truth HR images are available, it is not possible to report performance



Fig. 15. Application of C-SRIP on real-world images taken from the web. The images show crowds with several real-life LR faces. On the right side of each image are super-resolution results generated with C-SRIP (bottom) and two interpolation baselines for an upscaling factor of $8\times$. C-SRIP is able to recover significantly more detail from the input LR images than the nearest neighbour (top) and bicubic interpolation-based upsampling methods (middle).

scores for this experiments and we only show qualitative results in Fig. 15. We super-resolve images using an upscaling factor of $8\times$ for the presented examples. As can be seen, C-SRIP is able to recover more facial detail from the tiny input images than the nearest neighbour and bicubic interpolation-based baselines and produces considerably crisper results.

V. CONCLUSION

We have presented a novel CNN-based model for face hallucination from very low-resolution images (i.e., 24×24 pixels) at high magnification factors. We have shown that the proposed model improves SR results on face images compared to both existing general super-resolution and face hallucination models. In terms of future work, we see the possibility of adapting our model to other modalities, e.g., to video sequences via recurrent attention models.

REFERENCES

- [1] S. Baker and T. Kanade, "Hallucinating faces," in *Proc. Autom. Face Gesture Recognit. (FG)*, 2000, pp. 83–88.
- [2] 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.
- [3] A. Bulat and G. Tzimiropoulos, "Super-FAN: Integrated facial landmark localization and super-resolution of real-world low resolution faces in arbitrary poses with GANs," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2018, pp. 109–117.
- [4] Y. Chen, Y. Tai, X. Liu, C. Shen, and J. Yang, "FSRNet: End-to-end learning face super-resolution with facial priors," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2018, pp. 2492–2501.
- [5] A. Jourabloo, M. Ye, X. Liu, and L. Ren, "Pose-invariant face alignment with a single CNN," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 3200–3209.
- [6] Y. Li, S. Liu, J. Yang, and M.-H. Yang, "Generative face completion," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 3911–3919.
- [7] J. Roth, Y. Tong, and X. Liu, "Adaptive 3D face reconstruction from unconstrained photo collections," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 4197–4206.
- [8] X. Yu, B. Fernando, R. Hartley, and F. Porikli, "Super-resolving very low-resolution face images with supplementary attributes," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2018, pp. 908–917.
- [9] S. Baker and T. Kanade, "Limits on super-resolution and how to break them," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 24, no. 9, pp. 1167–1183, Sep. 2002.
- [10] B. K. Gunturk, A. U. Batur, Y. Altunbasak, M. H. Hayes, and R. M. Mersereau, "Eigenface-domain super-resolution for face recognition," *IEEE Trans. Image Process.*, vol. 12, no. 5, pp. 597–606, May 2003.
- [11] J. Yang, J. Wright, T. S. Huang, and Y. Ma, "Image super-resolution via sparse representation," *IEEE Trans. Image Process.*, vol. 19, no. 11, pp. 2861–2873, Nov. 2010.
- [12] J. Kim, J. K. Lee, and K. M. Lee, "Accurate image super-resolution using very deep convolutional networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 1646–1654.
- [13] S. Yang, M. Wang, Y. Chen, and Y. Sun, "Single-image super-resolution reconstruction via learned geometric dictionaries and clustered sparse coding," *IEEE Trans. Image Process.*, vol. 21, no. 9, pp. 4016–4028, Sep. 2012.
- [14] R. Timofte, V. De Smet, and L. Van Gool, "Anchored neighborhood regression for fast example-based super-resolution," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Dec. 2013, pp. 1920–1927.
- [15] 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. (ECCV)*, 2014, pp. 184–199.
- [16] J. Salvador and E. Perez-Pellitero, "Naive bayes super-resolution forest," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Dec. 2015, pp. 325–333.
- [17] J. Johnson, A. Alahi, and L. Fei-Fei, "Perceptual losses for real-time style transfer and super-resolution," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2016, pp. 694–711.
- [18] C. Ledig *et al.*, "Photo-realistic single image super-resolution using a generative adversarial network," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 4681–4690.
- [19] X. Xu, D. Sun, J. Pan, Y. Zhang, H. Pfister, and M.-H. Yang, "Learning to super-resolve blurry face and text images," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 251–260.
- [20] M. S. M. Sajjadi, B. Schölkopf, and M. Hirsch, "Enhancenet: Single image super-resolution through automated texture synthesis," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 4501–4510.
- [21] T. Tong, G. Li, X. Liu, and Q. Gao, "Image super-resolution using dense skip connections," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 4809–4817.
- [22] 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. (CVPR)*, vol. 2, Jul. 2017, pp. 5835–5843.
- [23] Y. Tai, J. Yang, and X. Liu, "Image super-resolution via deep recursive residual network," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 2790–2798.
- [24] Y. Huang, L. Shao, and A. F. Frangi, "Simultaneous super-resolution and cross-modality synthesis of 3D medical images using weakly-supervised joint convolutional sparse coding," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 5787–5796.
- [25] S. B. Kang, N. Joshi, C. L. Zitnick, T. S. Cho, R. Szeliski, and W. T. Freeman, "Image restoration by matching gradient distributions," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 34, no. 4, pp. 683–694, Apr. 2012.
- [26] Y. Wang, W. Yin, and Y. Zhang, "A fast algorithm for image deblurring with total variation regularization," CAAM, Putrajaya, Malaysia, Tech. Rep. TR07-10, 2007.
- [27] S. Dai, M. Han, W. Xu, Y. Wu, and Y. Gong, "Soft edge smoothness prior for alpha channel super resolution," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2007, pp. 1–8.

- [28] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli, "Image quality assessment: From error visibility to structural similarity," *IEEE Trans. Image Process.*, vol. 13, no. 4, pp. 600–612, Apr. 2004.
- [29] J. Tian and K.-K. Ma, "A survey on super-resolution imaging," *Signal, Image Video Process.*, vol. 5, no. 3, pp. 329–342, 2011.
- [30] K. Nasrollahi and T. B. Moeslund, "Super-resolution: A comprehensive survey," *Mach. Vis. Appl.*, vol. 25, no. 6, pp. 1423–1468, 2014.
- [31] N. Wang, D. Tao, X. Gao, X. Li, and J. Li, "A comprehensive survey to face hallucination," *Int. J. Comput. Vis.*, vol. 106, no. 1, pp. 9–30, 2014.
- [32] K. Nguyen, C. Fookes, S. Sridharan, M. Tistarelli, and M. Nixon, "Super-resolution for biometrics: A comprehensive survey," *Pattern Recognit.*, vol. 78, pp. 23–42, Jun. 2018.
- [33] A. Panagiotopoulos and V. Anastassopoulos, "Super-resolution image reconstruction techniques: Trade-offs between the data-fidelity and regularization terms," *Inf. Fusion*, vol. 13, no. 3, pp. 185–195, 2012.
- [34] H. Zhao, O. Gallo, I. Frosio, and J. Kautz, "Loss functions for image restoration with neural networks," *IEEE Trans. Comput. Imag.*, vol. 3, no. 1, pp. 47–57, Mar. 2017.
- [35] Z. Wang, E. P. Simoncelli, and A. C. Bovik, "Multiscale structural similarity for image quality assessment," in *Proc. 37th Asilomar Conf. Signals, Syst. Comput.*, vol. 2, Nov. 2003, pp. 1398–1402.
- [36] S. Zhu, S. Liu, C. C. Loy, and X. Tang, "Deep cascaded bi-network for face hallucination," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2016, pp. 614–630.
- [37] N. Ahn, B. Kang, and K.-A. Sohn, "Fast, accurate, and lightweight super-resolution with cascading residual network," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, Sep. 2018, pp. 252–268.
- [38] Z. Shen, W.-S. Lai, T. Xu, J. Kautz, and M.-H. Yang, "Deep semantic face deblurring," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2018, pp. 8260–8269.
- [39] S. Nah, K. H. Kim, and K. M. Lee, "Deep multi-scale convolutional neural network for dynamic scene deblurring," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 3883–3891.
- [40] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 770–778.
- [41] I. Goodfellow *et al.*, "Generative adversarial nets," in *Proc. Adv. Neural Inf. Process. Syst. (NIPS)*, 2014, pp. 1–9.
- [42] X. Yu and F. Porikli, "Ultra-resolving face images by discriminative generative networks," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2016, pp. 318–333.
- [43] K. Jia and S. Gong, "Generalized face super-resolution," *IEEE Trans. Image Process.*, vol. 17, no. 6, pp. 873–886, Jun. 2008.
- [44] Y. Jin and C. S. Bouganis, "Robust multi-image based blind face hallucination," in *Proc. IEEE Conf. Comput. Vision Pattern Recognit. (CVPR)*, Jun. 2015, pp. 5252–5260.
- [45] C.-Y. Yang, S. Liu, and M.-H. Yang, "Structured face hallucination," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2013, pp. 1099–1106.
- [46] E. Zhou, H. Fan, Z. Cao, Y. Jiang, and Q. Yin, "Learning face hallucination in the wild," in *Proc. AAAI Conf. Artif. Intell. (AI)*, 2015, pp. 3871–3877.
- [47] R. A. Farrugia and C. Guillemot, "Face hallucination using linear models of coupled sparse support," *IEEE Trans. Image Process.*, vol. 26, no. 9, pp. 4562–4577, Sep. 2017.
- [48] X. Yu and F. Porikli, "Face hallucination with tiny unaligned images by transformative discriminative neural networks," in *Proc. AAAI Conf. Artif. Intell. (AI)*, 2017, pp. 1–7.
- [49] X. Yu and F. Porikli, "Imagining the unimaginable faces by deconvolutional networks," *IEEE Trans. Image Process.*, vol. 27, no. 6, pp. 2747–2761, Jun. 2018.
- [50] Y. Song *et al.*, "Joint face hallucination and deblurring via structure generation and detail enhancement," *Int. J. Comput. Vis.*, vol. 127, nos. 6–7, pp. 785–800, 2018.
- [51] X. Yu, B. Fernando, B. Ghanem, F. Porikli, and R. Hartley, "Face super-resolution guided by facial component heatmaps," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2018, pp. 219–235.
- [52] A. Bulat, J. Yang, and G. Tzimiropoulos, "To learn image super-resolution, use a gan to learn how to do image degradation first," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2018, pp. 187–202.
- [53] W. Liu, D. Lin, and X. Tang, "Neighbor combination and transformation for hallucinating faces," in *Proc. Int. Conf. Multimedia Expo (ICME)*, Jul. 2005, p. 4.
- [54] B. Li, H. Chang, S. Shan, and X. Chen, "Aligning coupled manifolds for face hallucination," *IEEE Signal Process. Lett.*, vol. 16, no. 11, pp. 957–960, Nov. 2009.
- [55] P. H. Hennings-Yeoman, S. Baker, and B. V. K. V. Kumar, "Simultaneous super-resolution and feature extraction for recognition of low-resolution faces," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2008, pp. 1–8.
- [56] M. Turk and A. Pentland, "Eigenfaces for recognition," *J. Cognit. Neurosci.*, vol. 3, no. 1, pp. 71–86, 1991.
- [57] K. Zhang *et al.*, "Super-identity convolutional neural network for face hallucination," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2018, pp. 183–198.
- [58] W. Shi *et al.*, "Real-time single image and video super-resolution using an efficient sub-pixel convolutional neural network," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 1874–1883.
- [59] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet classification with deep convolutional neural networks," in *Proc. Adv. Neural Inf. Process. Syst. (NIPS)*, 2012, pp. 1097–1105.
- [60] O. M. Parkhi, A. Vedaldi, and A. Zisserman, "Deep face recognition," in *Proc. Brit. Mach. Vis. Conf. (BMVC)*, 2015, pp. 1–12.
- [61] D. C. Lee, M. Hebert, and T. Kanade, "Geometric reasoning for single image structure recovery," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2009, pp. 2136–2143.
- [62] J. Sun, Z. Xu, and H.-Y. Shum, "Image super-resolution using gradient profile prior," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2008, pp. 1–8.
- [63] F. N. Iandola, S. Han, M. W. Moskewicz, K. Ashraf, W. J. Dally, and K. Keutzer, "SqueezeNet: AlexNet-level accuracy with 50x fewer parameters and <0.5 MB model size," Feb. 2016, *arXiv:1602.07360*. [Online]. Available: <https://arxiv.org/abs/1602.07360>
- [64] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," in *Proc. Int. Conf. Learn. Represent. (ICLR)*, 2015, pp. 1–15.
- [65] D. Yi, Z. Lei, S. Liao, and S. Z. Li, "Learning face representation from scratch," Nov. 2014, *arXiv:1411.7923*. [Online]. Available: <https://arxiv.org/abs/1411.7923>
- [66] G. B. Huang, M. Mattar, T. Berg, and E. Learned-Miller, "Labeled faces in the wild: A database for studying face recognition in unconstrained environments," in *Proc. Workshop Faces 'Real-Life' Images, Detection, Alignment, Recognit.*, 2008, pp. 1–11.
- [67] Z. Liu, P. Luo, X. Wang, and X. Tang, "Deep learning face attributes in the wild," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Dec. 2015, pp. 3730–3738.
- [68] V. Le, J. Brandt, Z. Lin, L. Bourdev, and T. S. Huang, "Interactive facial feature localization," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2012, pp. 679–692.
- [69] G. Mu, X. Gao, K. Zhang, X. Li, and D. Tao, "Single image super resolution with high resolution dictionary," in *Proc. 18th IEEE Int. Conf. Image Process. (ICIP)*, Sep. 2011, pp. 1141–1144.
- [70] H. R. Sheikh and A. C. Bovik, "A visual information fidelity approach to video quality assessment," in *Proc. Int. Workshop Video Process. Qual. Metrics Consum. Electron. (VPQMCE-W)*, 2005, pp. 23–25.
- [71] B. Lim, S. Son, H. Kim, S. Nah, and K. M. Lee, "Enhanced deep residual networks for single image super-resolution," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. Workshops (CVPR-W)*, vol. 2, Jul. 2017, pp. 136–144.
- [72] 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 Biometrics*, vol. 7, no. 1, pp. 81–89, Jan. 2018.
- [73] X. Yu and F. Porikli, "Hallucinating very low-resolution unaligned and noisy face images by transformative discriminative autoencoders," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 3760–3768.
- [74] K. Grm, M. Pernuš, L. Cluzel, W. Scheirer, S. Dobrišek, and V. Štruc, "Face hallucination revisited: An exploratory study on dataset bias," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR) Workshops*, 2019, pp. 1–13.
- [75] Y. Blau and T. Michaeli, "The perception-distortion tradeoff," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2018, pp. 6228–6237.
- [76] C. Sagonas, G. Tzimiropoulos, S. Zafeiriou, and M. Pantic, "300 faces in-the-wild challenge: The first facial landmark localization challenge," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV) Workshops*, Jun. 2013, pp. 397–403.
- [77] V. Kazemi and J. Sullivan, "One millisecond face alignment with an ensemble of regression trees," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2014, pp. 1867–1874.