Learning to super-resolve faces through recognition

Klemen Grm, Simon Dobrišek, Vitomir Štruc

University of Ljubljana, Faculty of Electrical Engineering, Tržaška 25, Ljubljana, Slovenia
klemen.grm@fe.uni-lj.si

In this paper, we propose a novel high magnification factor face hallucination model that incorporates identity priors into the learning procedure. The model consists of i) a cascaded super-resolution network that upscales the low-resolution images, and ii) an ensemble of face recognition models that act as identity priors during training. Our network uses a cascade of SR models that progressively upscale the low-resolution images using steps of $2\times$. This allows us to apply appearance and recognition supervision at different resolutions. Our model is able to upscale (very) low-resolution images captured in unconstrained conditions and produce visually convincing results. We evaluate the proposed model on a large dataset of facial images and report superior performance compared to the state-of-the-art.

1 Introduction

Face hallucination represents a domain specific super-resolution (SR) problem where the goal is to recover high-resolution (HR) face images from low-resolution (LR) inputs [1]. It has important applications in image enhancement, compression and face recognition [2], but also surveillance and security [3, 4].

Like other 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 have many possible HR counterparts. Thus, the solution space for SR problems is extremely large and existing solutions commonly try to produce plausible reconstructions by "hallucinating" high-frequency information based on the provided LR evidence. While significant progress has been made in recent years in the area of super-resolution and face hallucination [5, 6, 7, 8, 9, 10, 11, 12], super-resolving arbitrary facial images, especially at high magnification factors, is still an open and challenging problem, mainly due to:

- The ill-posed nature of the face hallucination problem, where the solution space is known to grow exponentially with an increase in the desired magnification factor [3].
- The difficulty of integrating priors beyond solely the visual quality of the reconstructions. Most of the existing priors utilized for super-resolution relate to image characteristics such as gradient distribution [13], total variation [14] and smoothness [15]. If discernibility of the semantic content is the goal of the SR procedure, such priors may not be the most optimal choice.

Figure 1: Results generated with the proposed method.

The outlined limitation are most evident for challenging face hallucination problems where very low-resolution images (e.g., 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 using convolutional neural networks (CNNs). Our model 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. The SR models super-resolve the LR input images in magnification increments of $2\times$ and, consequently, allow for intermediate supervision at every scale. This intermediate supervision confines the explosion of the solution space size and contributes towards more accurate hallucination results. 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 the proposed model.

Overall, we make the following main contributions:

1. We propose a new face hallucination model, that integrates identity priors at multiple scales into the training procedure of a super-resolution network. To the best of our knowledge, this is 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 an multiple scales.

Once trained, the SR network is able to upsample 24×24 pixel LR images at magnification factors of $8\times$ and produce realistic and visually convincing hallucination results as illustrated in Fig. 1.

2 Proposed method

Our face hallucination model consists of two main components: *i) a generative SR network* for image upscaling, and *ii)* an ensemble of face recognition models that serve as identity priors. In the following sections we describe all components of the proposed model in detail and elaborate on the training procedure used to learn the model parameters.

2.1 The cascaded SR network

The generative part of our proposed model is a 53-layer deep convolutional neural network (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 images using a cascade of *SR modules*. Each module upscales the image by a factor of $2\times$, which allows us to apply a loss function on the intermediate results. The cascaded architecture allows us to solve a series of 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 our SR network around a fully-convolutional architecture that relies heavily on residual blocks [16] for all processing within one SR module and sub-pixel convolutions [17] for image upscaling. Our design choices are motivated by the success of fully-convolutional CNN models in various vision problems [16, 18, 19] and the state-of-the-art performance ensured by the sub-pixel convolutions in prior SR work [17, 10]. Similarly to [10], the residual blocks of the SR modules consist of two (convolution, batch-norm, activation) sub-blocks, followed by a post-activation element-wise sum.

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., $2\times$ and $4\times$ the initial scale) that are incorporated into the loss functions discussed in Section 2.3. However, these branches are not used at test time. The entire architecture of our network is illustrated in detail in Fig. 2.

2.2 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 [14, 15, 13]. The main motivation for incorporating priors into SR models is to provide a source of additional information for the learning procedure that complements the commonly used reconstruction-oriented objectives and contributes towards sharper and more accurate SR results.

Identity is an exceptionally strong prior in this context. In fact, it seems intuitive to think about SR in terms of both *i) an image enhancement* and *ii) content preservation*. While the image enhancement perspective is covered in our model by a reconstruction-based loss, the content preservation aspect is addressed through an ensemble of face recognition models that ensure that identity information is preserved.

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, we 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. 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.

We use SqueezeNet [20] models for this work. 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.

2.3 Training details and SSIM loss

We train the model in two stages. In the first stage, we learn the parameters of the SqueezeNet models for all three SR outputs. In the second stage, we freeze the the weights of the recognition models and train the SR network with a combined loss.

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. To this end, we apply a simple degradation model on the available HR images $\mathbf{x}_i^{hr}$ and generate N image quadruplets for training, i.e., $\{\mathbf{x}_i^{lr}, \mathbf{x}_i^{2\times}, \mathbf{x}_i^{4\times}, \mathbf{x}_i^{hr}\}$, where $\mathbf{x}_i^{lr}$ represents the LR input image, $\mathbf{x}_i^{2\times}$ and $\mathbf{x}_i^{4\times}$ stand for the intermediate SR results at $2\times$ and $4\times$ magnification factors, respectively, and the HR image $\mathbf{x}_i^{hr}$ corresponds to the ground truth for the magnification factor of $8\times$. Our degradation model uses Gaussian blurring followed by image decimation for down-sampling to produce training data.

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 are computed by smoothing the ground truth images by a Gaussian kernel and subtracting the smoothed image from the original, i.e., $\Delta\mathbf{x}_i^j = \mathbf{x}_i^j - \mathbf{g} * \mathbf{x}_i^j$, for $j \in \{2\times, 4\times, hr\}$, where σ values of $\sigma_{2\times} = 1/3$, $\sigma_{2\times} = 1$ 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 based on the generated residual images using categorical cross-entropy L_{CE}:

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

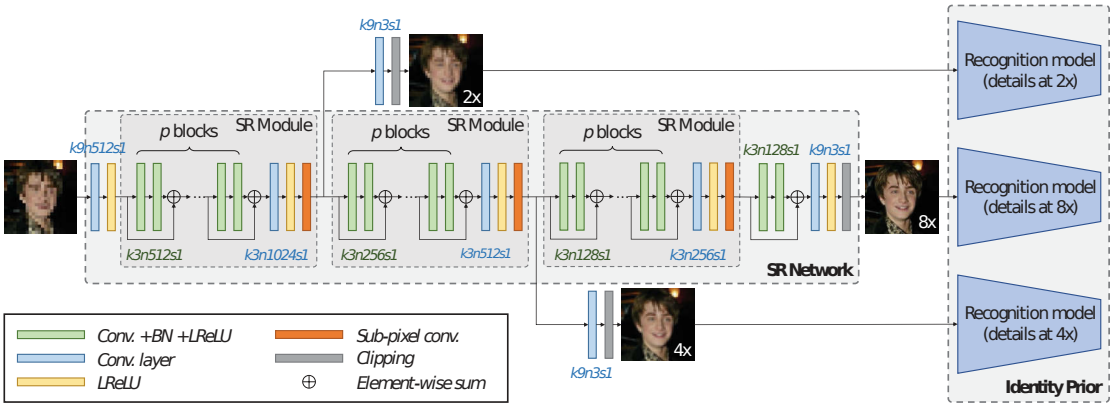

Figure 2: The proposed model consists of a generative SR network and face recognition models that serve as identity priors.

where $p_{\Delta\mathbf{x}}$ denotes the ground truth class probability distribution of the residual image $\Delta\mathbf{x}$ (i.e., $p_{\Delta\mathbf{x}} \in \{0,1\}^K$ is a class-encoded one-hot vector), $\hat{p}_{\Delta\mathbf{x}} \in \mathbb{R}^K$ stands for the output probability distribution produced by SqueezeNet's softmax layer based on $\Delta\mathbf{x}$, i.e., K stands for the number of classes in the training data and θ_{SN} represents the parameters of the network. We train the parameters of all three recognition models by minimizing the L_{CE} loss over the training dataset. The result of this training stage are three face recognition models $\hat{\theta}^{2\times}_{SN}, \hat{\theta}^{4\times}_{SN}, \hat{\theta}^{hr}_{SN}$, that serve as identity constraints for the SR network. We use the Adam [21] algorithm for training, 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, use random horizontal flipping and random crops for data augmentation.

SR network training. Standard reconstruction oriented loss functions used for training SR models, such as MSE or MAE, are known to produce overly smooth and often blurry SR results [10]. We therefore design a new loss function for our SR network around the structural similarity index (SSIM, [22]), and integrate it directly into our learning algorithm. Specifically, we use our SSIM approximation as a loss function for the hallucination model.

Given a ground truth image $\mathbf{x}$ and the corresponding SR network prediction $\hat{\mathbf{x}} = f_{\theta_{SR}}(\mathbf{x})$, we compute the SSIM-based loss as follows:

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

where the SR network f is parametrized by θ_{SR}, $\mathbb{E}_x[\cdot]$ stands for the expectation operator over the spatial coordinates and $\hat{SSIM}(\mathbf{x}, \hat{\mathbf{x}})$ is a spatial similarity map between $\mathbf{x}$ and $\hat{\mathbf{x}}$ defined as:

$$\hat{SSIM}(\mathbf{x}, \hat{\mathbf{x}}) = \frac{(2\mu_{12} + C_1) \odot (2\sigma_{12} + C_2)}{(\mu_1^2 + \mu_2^2 + C_1) \odot (\sigma_1^2 + \sigma_2^2 + C_2)}, \quad (3)$$

where we model the means μ_1 and μ_2 as convolutions of local patches with a Gaussian kernel, μ_{12} as their entry-wise product, and σ_1^2, σ_2^2 and σ_{12} as $\mathbf{xx}$, $\hat{\mathbf{x}}\hat{\mathbf{x}}$ and $\mathbf{x}\hat{\mathbf{x}}$ convolved with the same kernel, respectively.

The open parameters C_1 and C_2 are defined as per the SSIM reference implementation provided by the authors of [23], i.e., $C_1 \approx 6.55$, $C_2 \approx 58.98$.

Based on the pre-trained SqueezeNet models and the loss introduced above, we defined the overall loss of our face hallucination model as follows:

$$L(\theta_{SR}, \{\mathbf{x}^j\}) = \sum_{j \in \mathcal{D}} L_{SSIM}(\theta_{SR}, \mathbf{x}^j) + \alpha L_{CE}(\theta^j_{SN}, \Delta\mathbf{x}^j), \quad (4)$$

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.

We again use the Adam [21] algorithm for training, minimizing (4) with $\alpha = 0.001$. Due to the large memory footprint of the SR network and the face recognition models, we use a relatively small batch size of 8. We set the initial learning rate to $\frac{10}{3} \times 10^{-3}$ and multiply it by $\frac{1}{3}$ at the end of epochs 10, 25, 50 and 80.

Once the training is complete, we remove the recognition models and network branches used to generate the intermediate SR results. The final SR network takes an image $\mathbf{x}_{lr}$ of size 24×24 pixels as input and returns a 192×192 facial image $\mathbf{x}_{hr}$.

3 Experiments

We select two datasets for our experiments. To train the model we use the CASIA WebFace dataset [24] which features $494,414$ images of $10,575$ identities, (i.e., $N = 494,414$; $K = 10,575$). The CASIA WebFace images are blurred and sub-sampled to produce the necessary image quadruplets for training the recognition models and the SR network. For testing, we use the Labeled Faces in the Wild (LFW) [25] dataset with $13,233$ facial images and $5,749$ subjects. The two datasets are selected for the experiments because they feature images of variable quality captured in unconstrained conditions and thus represent a significant challenge for SR models. More importantly, they are designed to contains zero overlap in terms of identity, which is paramount to ensure a fair and unbiased evaluation of the model.

We compare our proposed model with 6 SR and face hallucination models, i.e.: the Naive Bayes Super-Resolution

Figure 3: Qualitative comparison of state-of-the-art SR models on sample images from the LFW dataset.

Table 1: PSNR and SSIM scores over the LFW dataset.

Model	Bicubic	NBSRF [8]	SRCNN [7]	VDSR [5]
PSNR	24.256	25.092	24.812	25.415
SSIM	0.7060	0.7268	0.7187	0.7411
Model	**ℓ_p [9]**	**SRGAN [10]**	**URDGN [26]**	**our model**
PSNR	26.985	25.669	25.575	**27.164**
SSIM	0.7903	0.6993	0.7516	**0.8171**

Forest (NBSRF) from [8], the Super-Resolution Convolutional Neural Network (SRCNN) from [7], the Very Deep Super Resolution Network (VDSR) from [5], the perceptual-loss based SR model (ℓ_p) from [9], the Super-Resolution Generative Adversarial Network from [10], and the Ultra Resolving Discriminative Generative Network (URDGN) from [26]. We train all models with the same data 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 SqueezeNet for the learning criterion. We include results for bicubic interpolation as a baseline. We present the quantitative results in terms of average PSNR and SSIM scores in the table 1. A few sample SR images are presented in Fig. 3. We see that with magnification factors of $8\times$, interpolation methods are insufficient and result in the loss of facial details. Furthermore, general SR models, such as NBSRF, SRCNN and VDSR, fail to provide substantial improvements and are seen to 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 SRGAN, URDGN and ℓ_p models 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 our model. 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 the global cross-entropy loss of the secondary networks as opposed to the local conv features exploited by ℓ_p.

References

[1] S. Baker and T. Kanade, "Hallucinating faces," in *FG*, 2000.

[2] C. Liu, H. Shum, and W. T. Freeman, "Face hallucination: Theory and practice," *IJCV*, 2007.

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

[4] B. K. Gunturk, A. U. Batur, Y. Altunbasak, M. H. Hayes, and R. M. Mersereau, "Eigenface-domain super-resolution for face recognition," *TIP*, 2003.

[5] J. Kim, J. Kwon Lee, and K. Mu Lee, "Accurate image super-resolution using very deep convolutional networks," in *CVPR*, 2016.

[6] R. Timofte, V. De Smet, and L. Van Gool, "Anchored neighborhood regression for fast example-based super-resolution," in *ICCV*, 2013.

[7] C. Dong, C. Loy, K. He, and X. Tang, "Learning a deep convolutional network for image super-resolution," in *ECCV*, 2014.

[8] J. Salvador and E. Perez-Pellitero, "Naive Bayes super-resolution forest," in *ICCV*, 2015.

[9] J. Johnson, A. Alahi, and L. Fei-Fei, "Perceptual losses for real-time style transfer and super-resolution," in *ECCV*, 2016.

[10] C. Ledig, L. Theis, F. Huszar, J. Caballero, A. Cunningham, A. Acosta, A. Aitken, A. Tejani, J. Totz, Z. Wang, et al., "Photo-realistic single image super-resolution using a generative adversarial network," in *CVPR*, 2017.

[11] M. Sajjadi, B. Schölkopf, and M. Hirsch, "Enhancenet: Single image super-resolution through automated texture synthesis," in *ICCV*, 2017.

[12] Y. Tai, J Yang, and X. Liu, "Image super-resolution via deep recursive residual network," in *CVPR*, 2017.

[13] T. Cho, C. Zitnick, N. Joshi, SB Kang, R. Szeliski, and WT Freeman, "Image restoration by matching gradient distributions," *TPAMI*, 2012.

[14] Y. Wang, W. Yin, and Y. Zhang, "A fast algorithm for image deblurring with total variation regularization," 2007.

[15] S. Dai, M. Han, W. Xu, Y. Wu, and Y. Gong, "Soft edge smoothness prior for alpha channel super resolution," in *CVPR*, 2007.

[16] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *CVPR*, 2016.

[17] S. Shi, J. Caballero, F. Huszár, J. Totz, A. Aitken, R. Bishop, D. Rueckert, and Z. Wang, "Real-time single image and video super-resolution using an efficient sub-pixel convolutional neural network," in *CVPR*, 2016.

[18] A. Krizhevsky, I. Sutskever, and G. Hinton, "Imagenet classification with deep convolutional neural networks," in *NIPS*, 2012.

[19] O. Parkhi, A. Vedaldi, A. Zisserman, et al., "Deep face recognition.," in *BMVC*, 2015.

[20] F. Iandola, S. Han, M. Moskewicz, K. Ashraf, W. Dally, and K. Keutzer, "SqueezeNet: AlexNet-level accuracy with 50x fewer parameters and 0.5 MB model size," *arXiv preprint arXiv:1602.07360*, 2016.

[21] D. Kingma and J. Ba, "Adam: A method for stochastic optimization," in *ICLR*, 2015.

[22] Z. Wang, E. Simoncelli, and AC Bovik, "Multiscale structural similarity for image quality assessment," in *ACSSC*, 2003.

[23] Z. Wang, AC Bovik, HR Sheikh, and E. Simoncelli, "Image quality assessment: from error visibility to structural similarity," *TIP*, 2004.

[24] D. Yi, Z. Lei, S. Liao, and SZ Li, "Learning face representation from scratch," *arXiv preprint arXiv:1411.7923*, 2014.

[25] GB Huang, M. Ramesh, T. Berg, and E. Learned-Miller, "Labeled faces in the wild: A database for studying face recognition in unconstrained environments," Tech. Rep., Technical Report 07-49, University of Massachusetts, Amherst, 2007.

[26] X. Yu and F. Porikli, "Ultra-resolving face images by discriminative generative networks," in *ECCV*, 2016.