Verification of Very Low-Resolution Faces Using An Identity-Preserving Deep Face Super-resolution Network

Ataer-Cansizoglu, E.; Jones, M.J.; Zhang, Z.; Sullivan, A.

TR2018-116 August 24, 2018

Abstract

Face super-resolution methods usually aim at producing visually appealing results rather than preserving distinctive features for further face identification. In this work, we propose a deep learning method for face verification on very low-resolution face images that involves identity-preserving face super-resolution with an extreme upscaling factor of 8. Our framework includes a super-resolution network and a feature extraction network. We train a VGG-based deep face recognition network [1] to be used as feature extractor. Our super-resolution network is trained to minimize the feature distance between the high resolution ground truth image and the super-resolved image, where features are extracted using our pre-trained feature extraction network. We carry out experiments on FRGC, Multi-PIE, LFW-a, and MegaFace datasets to evaluate our method in controlled and uncontrolled settings. The results show that the presented method outperforms conventional superresolution methods in low-resolution face verification.

arXiv

Verification of Very Low-Resolution Faces Using An Identity-Preserving Deep Face Super-resolution Network

Esra Ataer-Cansizoglu, Michael Jones, Ziming Zhang and Alan Sullivan

Mitsubishi Electric Research Labs (MERL)
Cambridge, MA USA

Abstract. Face super-resolution methods usually aim at producing visually appealing results rather than preserving distinctive features for further face identification. In this work, we propose a deep learning method for face verification on very low-resolution face images that involves identity-preserving face super-resolution with an extreme upscaling factor of 8. Our framework includes a super-resolution network and a feature extraction network. We train a VGG-based deep face recognition network [1] to be used as feature extractor. Our super-resolution network is trained to minimize the feature distance between the high resolution ground truth image and the super-resolved image, where features are extracted using our pre-trained feature extraction network. We carry out experiments on FRGC, Multi-PIE, LFW-a, and MegaFace datasets to evaluate our method in controlled and uncontrolled settings. The results show that the presented method outperforms conventional super-resolution methods in low-resolution face verification.

1 Introduction

Face images appear in various platforms and are vital for many applications ranging from forensics to health monitoring. In most cases, these images are in low-resolution, making face identification difficult. Although many algorithms have been developed for face recognition from high-quality images, few studies focus on the problem of very low-resolution face recognition. The performance of the traditional face recognition algorithms developed for high quality images, degrades considerably on low-resolution faces.

There exists a tremendous amount of work in image enhancement and upsampling. Recently, high magnification factors greater than 4 times have gained more attention for targeted objects such as faces with the rise in deep learning methods. Existing methods provide an upsampling of the image that is as close as possible to "*a* face image". Since resulting upsampled images are meant to be used in face identification task, recovering "*the* face" is essential. We present a face super-resolution method that preserves the identity of the person during super-resolution by minimizing the distance in feature space as opposed to the traditional face super-resolution methods designed to minimize the distance in high-resolution image space.

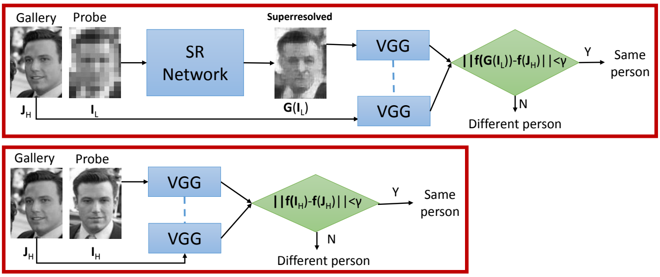


Fig. 1. System overview for (top) low-resolution and (bottom) high-resolution-resolution face verification. Dashed lines indicate weight sharing between networks.

The goal of this paper is to verify whether a given low-resolution face image is the same person as in a high-resolution gallery image. Our focus is very low-resolution face images with a tiny visible facial area (as low as 6×6 pixels). We represent a face image with its VGG face descriptor. Since our goal is verification, the face descriptor of a super-resolved face image should be as close as possible to the face descriptor of its ground truth high-resolution version. Thus, we train a super-resolution network by minimizing the feature distance between them. Contrary to the conventional face hallucination methods, we consider face descriptor similarity instead of appearance similarity during super-resolution. Moreover, we also perform detailed experiments in order to investigate the effect of various losses in training super-resolution for the task of low-resolution face verification.

The performance of super-resolution methods are evaluated by using image quality assessment measures such as peak signal-to-noise ratio (PSNR). These measures account for the visual similarity of two images by equally paying attention to every pixel in intensity domain. However, face identification relies on discriminative features. In this work, we present face descriptor similarity as an evaluation measure to assess the capacity of a method in preserving identity.

The main contributions of our study include: (1) a novel loss term to be used for face super-resolution in order to preserve identity for aggressive scaling factors as big as 8, (2) an evaluation measure to account for identity preservation on the super-resolved faces, and (3) a thorough analysis of various loss terms in training a face super-resolution network for low-resolution face verification.

1.1 Related Work

To solve the problem of low-resolution face verification (in which a low-resolution probe face is compared to a high-resolution gallery face) the main problem is how to handle the mismatch in resolutions. There are two basic approaches to solve this problem. The first is to map the low-resolution probe face and the high-resolution gallery face to a common feature space. Methods such as coupled locality preserving mappings [2], coupled kernel embeddings [3], and multidimensional scaling [4] follow this approach. Unfortunately, finding a resolution-robust feature space is very hard especially for high magnification factors. The second approach is to upsample the low-resolution face image (using a super-resolution algorithm) and compare to the high-resolution gallery face using a standard face recognition method. Our method falls into the second category.

Baker and Kanade [5] showed how to greatly improve super-resolution quality specifically for faces using pairs of high and low-resolution examples of faces. Since their work there have been many papers on face-specific techniques for super-resolution [6–11]. All of these methods try to produce visually pleasing high-resolution face images given the low-resolution input and face-specific models. They are not directly concerned with improving face recognition accuracy on the upsampled faces. However, since the main application of face super-resolution is face recognition, it makes sense to optimize a face-specific super-resolution algorithm explicitly to improve face recognition accuracy. This idea has been explored in a number of papers [12–14]. The basic idea is to find a high-resolution face image that optimizes both reconstruction and recognition costs simultaneously. They mainly use linear models for extracting face recognition feature vectors (e.g. PCA and LDA). These papers were all written before the recent era of deep neural networks which now dominate the face recognition field because of their high accuracy. The older recognition-optimizing face-specific super-resolution algorithms work well for frontal faces taken in controlled environments, but do not work nearly as well in typical “in-the-wild” settings for which deep networks are so effective.

Recently, a few papers have used convolutional neural networks (CNNs) or generative adversarial networks (GANs) for face-specific super-resolution. These methods better handle uncontrolled input faces with variations in lighting, pose and expression and only rough alignment. Zhou et al. [15]’s bi-channel approach used a CNN to extract features from the low-resolution input face and then mapped these using fully connected network layers to an intermediate upsampled face image, which is linearly combined with a bicubicly interpolated upsampling of the input face image to create a final high-resolution output face. In other work, Yu and Porikli [16, 17] used GANs for 8 times super-resolution of face images. Their method provides visually appealing results, but the resulting images can distort the identity of the person, which is a critical issue for face recognition applications. Cao et al. [18] presented a deep reinforcement learning approach that sequentially discovers attended patches followed by facial part enhancement. Zhu et al. [19] proposed a bi-network architecture to solve super-resolution in a cascaded way. Each of these methods is intended to produce visually pleasing face images and do not consider face recognition accuracy. They do not test their methods on a face recognition task.

A recent paper by Ledig et al. [20] that use a GAN for general image super-resolution is similar in spirit to ours in the sense that they also use the feature vector from a pre-trained CNN in the loss function used to optimize their network. In their case, in addition to reconstruction and adversarial losses they use many feature maps from a VGG-19 network [21] trained on ImageNet [22] to compare the similarity of upsampled and reference images (with any content) and achieve at most 4 times super-resolution. In our case we use the single penultimate feature vector from a VGG Deep Face network [1] to compare upsampled and reference face images in order to train a face-specific super-resolution network that maintains identity for 8 times magnification.

To the best of our knowledge, our paper is the first to use a deep neural network for face-specific super-resolution that is optimized not for visual quality, but for face recognition accuracy. We show that our super-resolution algorithm improves over other state-of-the-art super-resolution algorithms in terms of face verification accuracy on both controlled datasets (FRGC [23] and Multi-PIE [24]) as well as an in-the-wild datasets (LFW-a [25, 26] and MegaFace [27, 28]).

2 Method

2.1 Notation

We denote $\mathbf{x}_L^i \in \mathbb{R}^{N\times M}$ and $\mathbf{x}_H^i \in \mathbb{R}^{dN\times dM}$ as a pair of low-resolution and high-resolution (i.e. d times larger) versions of the i-th face image, function $\mathbf{G} : \mathbb{R}^{N\times M} \to \mathbb{R}^{dN\times dM}$ as a high-resolution image generator from low-resolution images, function $\mathbf{f} : \mathbb{R}^{dN\times dM} \to \mathbb{R}^D$ as a D-dim feature extractor from high-resolution images, $\|\cdot\|$ and $\|\cdot\|_F$ as the ℓ_2 norm of a vector and the Frobenius norm of a matrix, respectively.

2.2 Face Verification Problem Setup

Training: We are provided with a set of K pairs as well as their identities, i.e. $\left\{\left(\mathbf{x}_L^i, \mathbf{x}_H^i, y_i\right)\right\}_{i=1}^K$, where $y_i \in \mathcal{Y}$ denotes the identity of the i-th image pair. We would like to learn face verification models based on such training data.
Testing: We are provided with a new pair of low-resolution (as probe) and high-resolution (as gallery) face images, and asked whether these two images share the same identity based on the learned models.

2.3 Algorithm

An overview of our proposed approach is shown in Figure 1. We first super-resolve the given low-resolution face image using a deep convolutional network. Next, we extract features from the super-resolved image and a high-resolution gallery image using the VGG deep face network [1]. The similarity of the two images is decided based on the Euclidean distance between their feature vectors. Finally the verification is performed with a thresholding on the feature distance.

Training Objective We start our explanation from the objective function for training our model. Inspired by conventional methods, we propose optimizing the following objective function:

$$\min_{\mathbf{G},\mathbf{f}} \mathfrak{L}\left(\left\{\left(\mathbf{x}_L^i, \mathbf{x}_H^i, y_i\right)\right\}_{i=1}^K, \mathbf{G}, \mathbf{f}\right) + \lambda_1 \Omega_1(\mathbf{f}) + \lambda_2 \Omega_2(\mathbf{G}), \qquad (1)$$

where $\mathfrak{L}$ denotes the loss function, Ω_1, Ω_2 denote two regularizers on $\mathbf{f}$ and $\mathbf{G}$ (e.g. weight decay), respectively, and $\lambda_1 \geq 0, \lambda_2 \geq 0$ denote the predefined constants. In particular, we decompose the loss function as follows:

$$\mathfrak{L} \overset{\text{def}}{=} \mathfrak{L}_f \left(\left\{ \left(\mathbf{x}_H^i, y_i \right) \right\}_{i=1}^K, \mathbf{f} \right) + \mathfrak{L}_{recog} \tag{2}$$

where $\mathfrak{L}_f$ denotes the classification loss (e.g. least square) used in conventional recognition approaches for measuring the performance of $\mathbf{f}$, $\mathfrak{L}_{recog}$ denotes the *recognition loss* to measure the performance of $\mathbf{G}$ given $\mathbf{f}$

$$\mathfrak{L}_{recog} = \sum_i \omega_i \left\| \mathbf{f}(\mathbf{G}(\mathbf{x}_L^i)) - \mathbf{f}(\mathbf{x}_H^i) \right\|, \tag{3}$$

where $\omega_i \geq 0, \forall i$ denotes a weighting constant, in general, and in our current implementation we simply set $\omega_i = \frac{1}{K}$. Further investigation on the effect of varying ω_i's will be conducted in our future work.

Since this loss term computes the similarity of super-resolved and ground truth high-resolution faces, it can be used as an evaluation measure to assess the capacity of a method to preserve identity during super-resolution.

Discussion: There are two alternative loss functions to recognition loss that are widely used in image super resolution or restoration [29].

(1) Reconstruction Loss: It measures the difference (in Euclidean space) between the reconstructed image from a low-resolution image and its corresponding high-resolution image, defined as follows:

$$\mathfrak{L}_{recon} = \frac{1}{K} \sum_i \left\| \mathbf{G}(\mathbf{x}_L^i) - \mathbf{x}_H^i \right\|_F. \tag{4}$$

Minimizing this loss usually introduces a large amount of smoothing and averaging artifacts in reconstruction images that helps improve visual appearance but not verification accuracy necessarily.

(2) Structural Similarity (SSIM) Loss: SSIM is a popular measure that accounts for humans perception of image quality. On a local patch of two images $\mathbf{x}$ and $\mathbf{y}$ around pixel $\mathbf{p}$, SSIM is computed as

$$SSIM(\mathbf{x}, \mathbf{y}, \mathbf{p}) = \frac{2\mu_x\mu_y + c_1}{\mu_x^2 + \mu_y^2 + c_1} \frac{2\sigma_{xy} + c_2}{\sigma_x^2 + \sigma_y^2 + c_2}, \tag{5}$$

where μ_x, μ_y and σ_x, σ_y denote the mean and variance, respectively, of the intensities in the local patch around pixel $\mathbf{p}$ in $\mathbf{x}$ and $\mathbf{y}$ and σ_{xy} is the covariance of intensities in the two local patches. c_1 and c_2 are constant factors to stabilize the division with weak denominator. We divide the image into $h \times h$ pixel grids and compute the mean of SSIM sum over all patches as the similarity between two images

$$r(\mathbf{x}, \mathbf{y}) = \frac{1}{T} \sum_{k=1}^T \sum_{\mathbf{p}} SSIM(\mathbf{x}^k, \mathbf{y}^k, \mathbf{p}) \tag{6}$$

where T is the total number of patches and $\mathbf{x}^k, \mathbf{y}^k$ are kth corresponding patch pair. Consequently, our SSIM loss is formulated as

$$\mathfrak{L}_{ssim} = \frac{1}{K} \sum_i \left[h^2 - r\left(\mathbf{G}(\mathbf{x}_L^i), \mathbf{x}_H^i\right)\right]. \tag{7}$$

Compared with reconstruction loss, minimizing SSIM loss helps recover the information at high frequency visually, but still unnecessarily improves the verification performance.

In our experiments we conduct comprehensive comparison on these three loss functions to demonstrate the correct usage of recognition loss for the task of face verification.

Two-Stage Minimization To optimize Eq. 1 we propose using two-stage minimization technique. Precisely, we first learn the feature extraction function $\mathbf{f}$ *supervisedly* by optimizing

$$\min_{\mathbf{f}} \mathfrak{L}_f \left(\left\{\left(\mathbf{x}_H^i, y_i\right)\right\}_{i=1}^K, \mathbf{f}\right) + \lambda_1 \Omega_1(\mathbf{f}). \tag{8}$$

Then we learn the high-resolution image generator function $\mathbf{G}$ *unsupervisedly* based on the learned $\mathbf{f}$ by optimizing

$$\min_{\mathbf{G}} \frac{1}{K} \sum_i \left\|\mathbf{f}(\mathbf{G}(\mathbf{x}_L^i)) - \mathbf{f}(\mathbf{x}_H^i)\right\| + \lambda_2 \Omega_2(\mathbf{G}). \tag{9}$$

Two-Stage vs. End-to-End: We implement end-to-end training algorithm as well, but find that the performance is much worse than our current two-stage minimization algorithm. We hypothesize that the end-to-end training involves many more parameters that need to be optimized, leading to overfitting on training data due to higher model complexity with respect to limited data samples. In contrast, our two-stage training strategy serves as regularization similar to the early stopping criterion used in deep learning.

Network Architecture: We use a similar architecture to Yu et al. [16] for our face super-resolution network (i.e. function $\mathbf{G}$), except our super-resolution network is trained on gray-scale images. The network is a deconvolutional network with 3 deconvolutional layers with stride 2 and 2 additional convolutional layers. Our face recognition network (i.e. function $\mathbf{f}$) has a VGG architecture with 19 layers as reported in [1]. Note that face descriptors from other deep networks for face recognition could be used instead of VGG, but this was chosen because the implementation of VGG deep face network is publicly available, and achieves near state-of-the-art performance on face recognition [1].

Face Verification at Test Time After learning the super-resolution network, each super-resolved image is represented with its VGG face descriptor. The decision of whether a low-resolution face image $\mathbf{x}_L$ and a high-resolution gallery image $\mathbf{x}_H$ contain the same person is given based on the indicator function

$$I(\mathbf{x}_L, \mathbf{x}_H, \gamma) = \begin{cases} 1 \text{ if } \|\mathbf{f}(\mathbf{G}(\mathbf{x}_L)) - \mathbf{f}(\mathbf{x}_H)\| < \gamma, \\ 0 \qquad\qquad \text{otherwise,} \end{cases} \tag{10}$$

where γ is a threshold that can be determined using cross-validation.

3 Experiments and Results

3.1 Datasets and Experimental Setup

We carried out two sets of experiments under controlled and uncontrolled, i.e. in the wild, settings. For controlled settings, we used Face Recognition Grand Challenge (FRGC) [23] and Multi-PIE [24] datasets. For uncontrolled setting, we used an aligned version of the Labeled Faces in the Wild dataset [25], called Labeled Faces in the Wild-a (LFW-a) [26] and MegaFace dataset [27, 28].
FRGC: The FRGC dataset contains frontal face images taken in a studio setting under two lighting conditions with only two facial expressions (smiling and neutral). We generated training and test splits, where we kept the identities in each set disjoint. The training set consisted of 20,000 images from 409 subjects and the test set consisted of 2,149 images from 142 subjects.
Multi-PIE: The Multi-PIE dataset consists of face images of 337 subjects captured from various viewpoints and illumination conditions over multiple sessions. Our goal in using this set was to better evaluate the peformance of face verification under different facial poses and illumination conditions in a controlled setting. We use the three most frontal views (05_1, 05_0, 14_0) and the four most frontal lighting conditions (06, 07, 08, 09) from each data collection session. We randomly generated training and test splits, where we kept the identities in each set disjoint. The training set consisted of 9,091 images from 252 subjects and the test set consisted of 3,000 images from 85 subjects. For both FRGC and Multi-PIE datasets, we carried out face alignment [30].
LFW-a: This dataset consists of faces captured in an uncontrolled setting with several poses, lightings and expressions. We used the training and test splits as indicated in the LFW development benchmark, which also contains a set of image pairs to be tested in the verification task. The benchmark contains 9,525 training images, 3,708 test images and 1,000 image pairs from the test set to be verified.
MegaFace: MegaFace consists of 4.7M images from 672K identities collected from Flickr users. We followed the experimental protocol described for face verification in MegaFace challenge 2. The training set is provioded along with facial landmark points. We performed face alignment using provided landmarks. We discarded the images with resolution smaller than our high resolution images and the images with high registration error during alignment. As a result we used 2.8M images from MegaFace for training. Following the verification protocol, we used FaceScrub dataset [31] as our probe images during testing. FaceScrub comprises a total of 106,863 face images of male and female 530 celebrities. Negative

pairs for verification are constructed using 10K distractor images provided by MegaFace challenge. Since facial landmarks are not provided for original high resolution images of FaceScrub and MegaFace distractor images, we carry out face alignment following [32].

Data preparation

High-resolution (HR) images had 128×128 pixel resolution, where the faces occupied approximately 50×50 pixels area. We generated low-resolution (LR) images with 8 times downsampling by following the approach in [33]. Namely, we filtered the high-resolution images with a Gaussian blur kernel $\sigma = 2.4$ followed by downsampling. As a result, the low-resolution images contained a facial area of approximately 6×6 pixels. For the task of low-resolution face verification, we tested whether a given low-resolution probe image contains the same person as a given high-resolution gallery image. In the FRGC and Multi-PIE datasets, we considered verification of all test image pairs. For LFW-a dataset, we tested all $1,000$ image pairs in the benchmark, where each pair was tested two times by switching probe and gallery. For MegaFace dataset, we followed the verification protocol given in the MegaFace challenge. More specifically, negative pairs consisted of all pairs between FaceScrub as probe images and MegaFace distractor dataset as gallery images. As for positive pairs, we tested each probe image from FaceScrub with each of the other images of the same identity as gallery.

In terms of a baseline, we compare against face verification using the original high-resolution images for both gallery and probe (See Figure 1). In this case, our "baseline" algorithm provides an upper bound on the accuracy for low-resolution face recognition.

3.2 Algorithm Details

We trained the VGG deep face network using the VGG face dataset [1] for 256×256 pixel gray-scale face images. [1] The network outputs 4096 dimensional feature vectors. Next, we trained the super-resolution network for each dataset separately by minimizing each of the loss terms stated in equations (3), (4) and (7). We also analyzed how joint optimization of all the terms affects the verification performance by minimizing weighted sum of all loss terms

$$\hat{\mathbf{G}} = \underset{\mathbf{G}}{argmin} \left(\mathfrak{L}_{recon} + \alpha \mathfrak{L}_{ssim} + \beta \mathfrak{L}_{recog} \right). \qquad (11)$$

We set weights α and β following a greedy search procedure. The code is implemented using the Caffe deep learning framework [34]. Optimization was performed using the RMSProp algorithm [35] with a learning rate of 0.001 and a decay rate of 0.01. The training took 3 days on a NVIDIA GeForce GTX TITAN X GPU with Maxwell architecture and 12GB memory. Average running time for verification of a probe and gallery image pair is 0.1 second. We used a patch size of $h = 8$ for SSIM loss.

[1] Note that, the output of super-resolution network is upsampled 2× with linear interpolation to be input to the VGG network.

Method	FRGC				Multi-PIE			
	$\mathfrak{L}_{recog}$	AUC	PSNR	SSIM	$\mathfrak{L}_{recog}$	AUC	PSNR	SSIM
Bicubic	1.135	0.767	23.812	0.606	1.170	0.850	18.923	0.443
SRCNN [36]	1.115	0.773	23.733	0.619	1.153	0.875	23.249	0.610
URDGN [16]	1.025	0.780	17.738	0.512	1.143	0.896	23.845	0.632
VDSR [37]	1.088	0.796	24.794	0.646	1.192	0.685	13.366	0.285
MZQ [38]	0.909	0.806	25.287	0.758	1.170	0.850	18.923	0.443
Only $\mathfrak{L}_{recon}$	0.834	0.818	**26.485**	**0.797**	0.912	0.958	**24.718**	**0.724**
Only $\mathfrak{L}_{ssim}$	1.185	0.618	14.075	0.068	0.994	0.929	22.504	0.640
Only $\mathfrak{L}_{recog}$	0.794	0.831	15.730	0.247	**0.879**	0.963	18.176	0.431
Joint	**0.788**	**0.833**	26.169	0.747	0.887	**0.968**	24.428	0.684

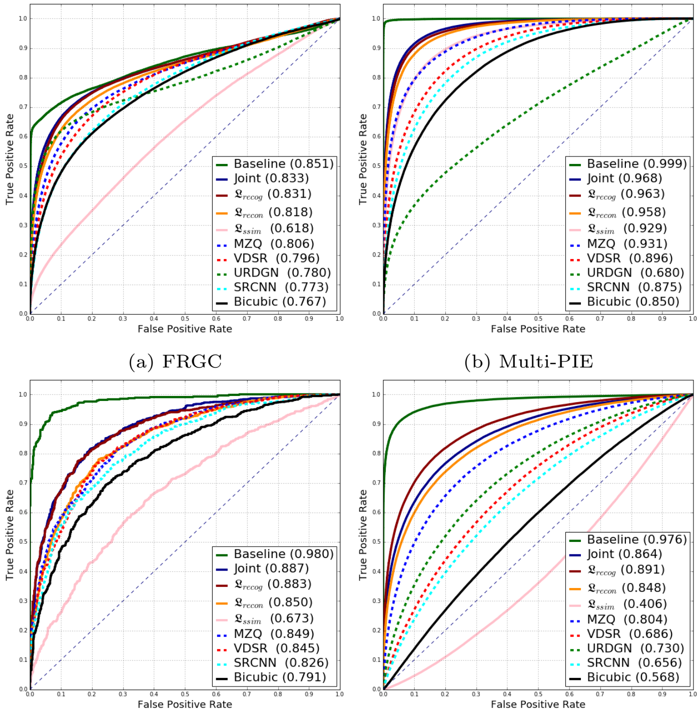
Table 1. Quantitative results on FRGC and Multi-PIE, where baseline (HR face verification) AUC between original image pairs are computed as 0.851 and 0.998 respectively.

Method	LFW-a				MegaFace			
	$\mathfrak{L}_{recog}$	AUC	PSNR	SSIM	$\mathfrak{L}_{recog}$	AUC	PSNR	SSIM
Bicubic	1.131	0.791	22.273	0.566	1.174	0.568	20.615	0.514
SRCNN [36]	1.067	0.826	22.833	0.601	1.126	0.656	21.177	0.546
URDGN [16]	-	-	-	-	1.107	0.730	16.804	0.401
VDSR [37]	1.074	0.845	**23.408**	0.621	1.109	0.686	21.680	0.573
MZQ [38]	0.992	0.849	22.660	0.623	1.033	0.804	21.496	0.604
Only $\mathfrak{L}_{recon}$	1.018	0.850	22.655	**0.625**	0.981	0.848	**22.674**	**0.672**
Only $\mathfrak{L}_{ssim}$	1.159	0.673	13.417	0.171	1.253	0.406	11.305	0.071
Only $\mathfrak{L}_{recog}$	0.974	0.883	16.600	0.336	**0.900**	**0.891**	15.354	0.376
Joint	**0.963**	**0.887**	22.055	0.537	0.949	0.864	21.218	0.555

Table 2. Quantitative results on LFW-a and MegaFace datasets, where baseline (HR face verification) AUC between original image pairs are computed as 0.980 and 0.976 respectively. Best value for each column is shown in bold.

3.3 Quantitative Results

We report quantitative results in Table 1 and Table 2 for the experiments on controlled and uncontrolled datasets, respectively. In order to evaluate face verification performance, we computed a receiver operating characteristic (ROC) curve (plotting true positive versus false positive face verifications) by varying the threshold γ and reported area under curve (AUC). Since an important contribution of our method is minimizing the distance of low-resolution and high-resolution images in feature space, we also report recognition loss $\mathfrak{L}_{recog}$ on the test set as a means of quantifying identity preservation. For evaluating appearance quality, we report peak signal-to-noise ratio (PSNR) and SSIM. As seen in the tables, the network trained by minimizing recognition loss outperforms all other methods and the other individual loss functions in terms of face verification, although the others yield higher PSNR and SSIM values. Note that SSIM loss is computed on each image patch independently as opposed to the SSIM value that we compute for the whole image during evaluation. Low SSIM value



(a) FRGC

(b) Multi-PIE

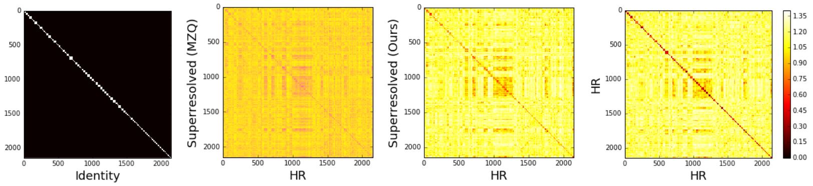
(c) LFW-a

(d) MegaFace

Fig. 2. ROC curves for the results on (a) FRGC, (b) Multi-PIE, (c) LFW-a, and (d) MegaFace datasets. The numbers in parantheses indicate area under curve (AUC).

using $\mathfrak{L}_{ssim}$ loss on FRGC is due to the fact that the images are all frontal, yielding filters learned independently for each facial patch (Please see block affects on visual results in Figure 4).

After a greedy procedure, we set weights of joint optimization for FRGC and Multi-PIE as $\alpha = 1,000, \beta = 300$ and for LFW-a and MegaFace as $\alpha = 10,000, \beta = 3,000$. Note that the magnitude of $\mathfrak{L}_{recon}$ and $\mathfrak{L}_{ssim}$ is in the order of number of pixels in the high-resolution image and in the patch respectively, while the magnitude of $\mathfrak{L}_{recog}$ is in the order of feature vector dimension, which is 4096. In Table 1 the results for joint optimization of all terms are comparable to the results with recognition loss. However, adjusting the weights of the terms in the joint optimization is a tedious process. Therefore, using only recognition loss is sufficient, if the final goal of the super-resolution is face identification. During greedy selection procedure, we also trained the network without SSIM loss, but the verification performance was slightly worse compared to using all three terms together. In Table 2, verification accuracy is comparable for joint loss and recognition loss on LFW-a, while on MegaFace dataset recognition loss outperforms with a difference of around 0.03 in AUC.



(a) Labels (b) MZQ method (c) Our method (d) Baseline

Fig. 3. Colormap visualization of distance matrices: (a) facial identity matrix (white pixels indicate the images of the same person) and the distance matrices in feature domain between (b) super-resolved images with MZQ technique [38] and HR images, (c) super-resolved images by our method and HR images, and (d) among HR image pairs, i.e. baseline.

We compared the performance of our method with the state-of-the-art generic and face-specific super-resolution methods as reported in the table. For general object super-resolution methods, we compared with two deep learning-based methods: SRCNN by Dong et al. [36] and VDSR by Kim et al. [37]. Since both methods handle at most 4× magnification, we performed 4× magnification followed by 2× magnification in order to achieve the same upsampling with our method. For face-specific super-resolution methods, we compared against URDGN by Yu et al. [16] and MZQ by Ma et al. [38]. We also performed experiments with structured face hallucination (SFH) technique [39], but we omitted its results since it was not successful for most of the images for 8× magnification due to its dependence on facial landmark detection. For the MZQ method we used the implementation of Yang et al. [39] and employed the same training set as ours for training except MegaFace dataset. Since MZQ is a dictionary-based approach and the implementation had memory constraints, for MegaFace dataset we randomly selected 100K images from the training set and used them for training MZQ method. For the other methods, we used the provided pre-trained models by the respective studies. URDGN method was trained on colored images with a different face alignment than ours. Therefore, we tested their method after aligning the test images according to their settings. Also, URDGN experiments were carried out on colored images, but final evaluation measures were computed on grayscale images. Note that all datasets consist of colored images except LFW-a, which has only grayscale images.

Figure 2 shows the ROC curves for all methods in each of the datasets. The difference between recognition and other losses is more visible in uncontrolled datasets, while recognition loss still outperforms the other losses in both controlled setting datasets, FRGC and Multi-PIE. The super-resolution network trained with recognition loss has accuracy very close to the high-resolution (optimal) baseline for FRGC and Multi-PIE (around 0.03 in terms of AUC), although for LFW-a and MegaFace there is a larger gap. Compared to other state-of-the-art super-resolution algorithms, our network trained on $\mathfrak{L}_{recog}$ does significantly better. The network trained on joint loss is slightly better than $\mathfrak{L}_{recog}$ for FRGC

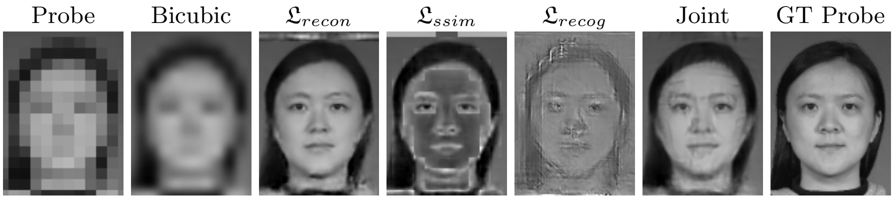


Fig. 4. Example probe images and super-resolution results on FRGC dataset. From left to right: LR probe image, bicubic interpolation, super-resolution results with reconstruction, SSIM, recognition and joint losses, ground truth (GT) HR probe image.

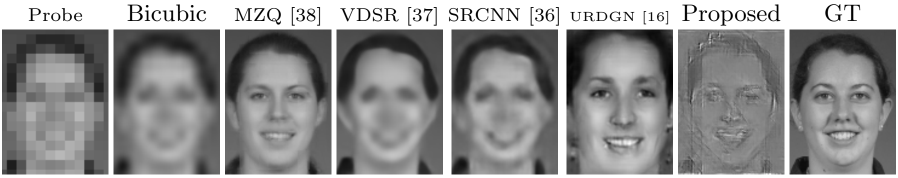


Fig. 5. Comparative results on FRGC dataset.

and LFW-a and slightly worse for Multi-PIE, at the expense of more time consuming training. Proposed method gives the best performance on large-scale MegaFace dataset. Our network trained on $\mathfrak{L}_{recon}$ is better in terms of AUC on all test sets than competing super-resolution methods, although not as good as when using $\mathfrak{L}_{recog}$ or joint loss.

Our main goal in this study is to obtain super-resolved images that are similar to the corresponding ground truth high-resolution (HR) images in the feature domain. Thus, we also provide a colormap visualization of the pairwise distance matrices in the feature domain for the FRGC dataset in Figure 3. As can be seen, the distance matrix of HR-to-HR (i.e. baseline) pairs is very similar to the distance matrix of super-resolved-to-HR pairs. Figure 3b also shows the distance matrix for MZQ algorithm [38] that gives the best performance among the comparative methods in terms of face verification. The difference between matching and unmatching pairs is more visible in the distance matrix obtained from our method.

3.4 Qualitative Results

The goal of this study is not to obtain good looking super-resolved images, but rather to improve face verification accuracy for low-resolution images. We display visual results from our super-resolution network in order to elaborate what kind of features are retained and input to the VGG network for better face identification.

Face verification was performed between a low-resolution probe image and high-resolution gallery image. Figure 4 shows example low-resolution probe images along with the super-resolved images from the minimization of various loss

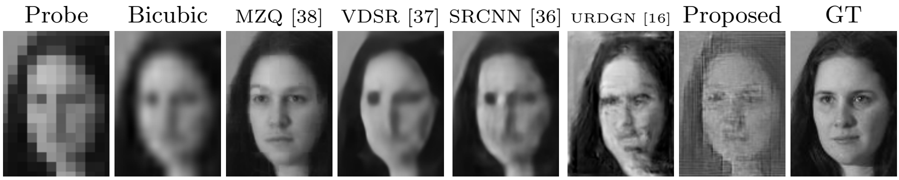


Fig. 6. Comparative results on Multi-PIE dataset.

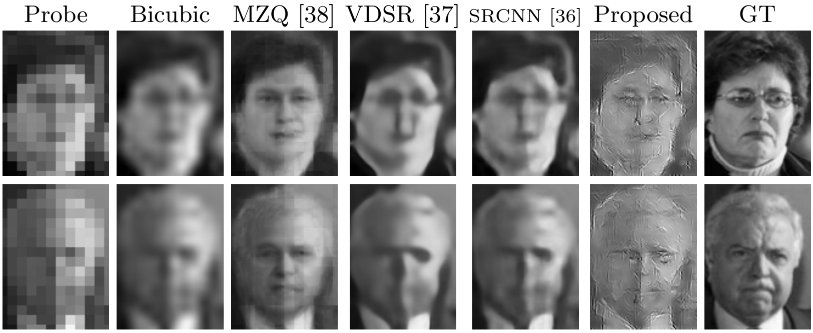


Fig. 7. Comparative results on LFW-a dataset.

functions on the FRGC dataset. Reconstruction loss and SSIM loss smooth out facial details, while recognition-based loss can yield more details around important facial regions such as eyes and nose.

Figure 5 shows comparative super-resolution results from all tested methods on FRGC dataset. The column labeled "Proposed" is our super-resolution network trained using recognition loss only. As can be seen face-specific super-resolution methods yield visually appealing results. However, their face verification performance is poor.

Figure 6 shows visual results using various methods on the Multi-PIE dataset. The proposed method is not affected from uneven brightness and shadows on the face that occur due to lighting direction as seen in the figure. Comparative results on LFW-a and MegaFace datasets are displayed in Figures 7 and 8 respectively.

An interesting observation about faces upsampled using our proposed super-resolution network is that the lighting effects are largely removed, which is very useful if face recognition accuracy is the goal, although it is not what should happen if the goal is to create a high-resolution face image that looks like the input low-resolution face when downsampled. Figure 7 and Figure 8 illustrate this point the best. In Figure 7, the input faces have strong lighting on the right side of the image, but our proposed super-resolved faces are much more evenly lit. Similar effects are visible on the results in Figure 8. Removing the effects of lighting is well-known to improve face recognition accuracy.

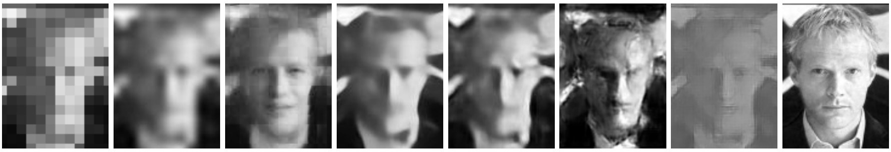


Fig. 8. Comparative results on Megaface dataset.

4 Conclusion and Discussion

We presented a super-resolution-based method for verification of very low-resolution faces. Our method exploited a VGG network for feature extraction. We trained a deep neural network for 8 times super-resolution of face images by minimizing the distances of high-resolution and super-resolved images of the same person in terms of their face descriptors computed by the VGG Deep Face network. The results on controlled and uncontrolled settings showed that the presented method provides better verification accuracy compared to conventional super-resolution techniques. This work demonstrates that generating visually appealing super-resolved images is not necessary if the final goal is improving face recognition accuracy. Instead, the super-resolution network should directly optimize the distance to the desired face descriptor.

In this work, our aim was to learn a high-resolution image generator given a feature extractor, assuming that the feature extractor already performs well in high-resolution face verification. In other words, we trained the generator to find the best mapping function that will transform a given low-resolution image to its high-resolution version in feature space. Thus, feature extractor and generator were trained in alternating steps rather than an end-to-end fashion. Moreover, simultaneous learning of feature extractor and generator is a cumbersome task as it will require learning of a larger number of parameters.

Low-resolution face verification is a challenging task that involves detection, super-resolution and verification. In this work, we focused on super-resolution and verification components by assuming the faces are already detected. Although low-resolution face detection is hard, it can be done with current algorithms, especially using upper torso or full pedestrian detection to localize the face. Face alignment carried out in our experiments are very rough (especially LFW-a), hence a face bounding box will be sufficient for our verification module. Moreover, the architecture of our super-resolution network is more robust to transformations on the input image as explained in [16]. Detection and super-resolution problems can benefit from each other. We would like to work on low-resolution face detection incorporating our super-resolution approach as a future extension.

Using video inputs rather than single images is an important future extension of our work. Also, we would like to incorporate more constraints to our loss function to increase the feature distance between face images with different identities.

References

1. Parkhi, O.M., Vedaldi, A., Zisserman, A.: Deep face recognition. In: Proc. British Machine Vision Conf. (BMVC). (2015)
2. Li, B., Shan, S., Chen, X.: Low-resolution face recognition via coupled locality preserving mappings. IEEE Signal Processing Letters **17**(1) (2010) 20–23
3. Ren, C., Dai, D., Yan, H.: Coupled kernel embeddings for low-resolution face recognition. IEEE Trans. Image Process. **21**(8) (2012) 3770–3783
4. Biswas, S., Aggarwal, G., Flynn, P.J., Bowyer, K.W.: Pose-robust recognition of low-resolution face images. IEEE Trans. Pattern Anal. Mach. Intell. **35**(12) (2013) 3037–3049
5. Baker, S., Kanade, T.: Hallucinating faces. In: Inter. Conf. on Automatic Face and Gesture Recognition. (2000)
6. Wang, N., Tao, D., Gao, X., Li, X., Li, J.: A comprehensive survey to face hallucination. Int'l J. Computer Vision **106**(1) (2014) 9–30
7. Baker, S., Kanade, T.: Limits on super-resolution and how to break them. IEEE Trans. Pattern Anal. Mach. Intell. **24**(9) (2002) 1167–1183
8. Liu, C., Shum, H.Y., Zhang, C.: A two-step approach to hallucinating faces: global parametric model and local nonparametric model. In: Proc. IEEE Conf. Computer Vision and Pattern Recognition (CVPR). (2001)
9. Capel, D., Zisserman, A.: Super-resolution from multiple views using learnt image models. In: Proc. IEEE Conf. Computer Vision and Pattern Recognition (CVPR). (2001)
10. Li, Y., Lin, X.: Face hallucination with pose variation. In: Inter. Conf. on Automatic Face and Gesture Recognition. (2004)
11. Liu, C., Shum, H.Y., Freeman, W.T.: Face hallucination: Theory and practice. Int'l J. Computer Vision **75**(1) (2007) 115–134
12. Jia, J., Gong, S.: Multi-modal tensor face for simultaneous super-resolution and recognition. In: Proc. IEEE Int'l Conf. Computer Vision (ICCV). (2005)
13. Hennings-Yeomans, P.H., Baker, S., Kumar, B.V.: Simultaneous super-resolution and feature extraction for recognition of low-resolution faces. In: Proc. IEEE Conf. Computer Vision and Pattern Recognition (CVPR). (2008)
14. Zou, W.W., Yuen, P.C.: Very low resolution face recognition problem. IEEE Trans. Image Proc. **21**(1) (2012) 327–340
15. Zhou, E., Fan, H., Cao, Z., Jiang, Y., Yin, Q.: Learning face hallucination in the wild. In: Proc. of the AAAI Conf. on Artificial Intelligence. (2015)
16. Yu, X., Porikli, F.: Ultra-resolving face images by discriminative generative networks. In: Proc. European Conf. Computer Vision (ECCV). (2016) 318–333
17. Yu, X., Porikli, F.: Hallucinating very low-resolution unaligned and noisy face images by transformative discriminative autoencoders. In: Proc. IEEE Conf. Computer Vision and Pattern Recognition (CVPR). (2017) 3760–3768
18. Cao, Q., Lin, L., Shi, Y., Liang, X., Li, G.: Attention-aware face hallucination via deep reinforcement learning. In: Proc. IEEE Conf. Computer Vision and Pattern Recognition (CVPR). (2017)
19. Zhu, S., Liu, S., Loy, C.C., Tang, X.: Deep cascaded bi-network for face hallucination. In: European Conference on Computer Vision, Springer (2016) 614–630
20. Ledig, C., Theis, L., Huszár, F., Caballero, J., Cunningham, A., Acosta, A., Aitken, A., Tejani, A., Totz, J., Wang, Z., et al.: Photo-realistic single image super-resolution using a generative adversarial network. In: Proc. IEEE Conf. Computer Vision and Pattern Recognition (CVPR). (2017)

21. Simonyan, K., Zisserman, A.: Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:1409.1556 (2014)
22. Deng, J., Dong, W., Socher, R., Li, L.J., Li, K., Fei-Fei, L.: ImageNet: A large-scale hierarchical image database. In: Proc. IEEE Conf. Computer Vision and Pattern Recognition (CVPR). (2009)
23. Phillips, P.J., Flynn, P.J., Scruggs, T., Bowyer, K.W., Chang, J., Hoffman, K., Marques, J., Min, J., Worek, W.: Overview of the face recognition grand challenge. In: Proc. IEEE Conf. Computer Vision and Pattern Recognition (CVPR). Volume 1. (2005) 947–954
24. Gross, R., Matthews, I., Cohn, J., Kanade, T., Baker, S.: Multi-PIE. Image and Vision Computing **28**(5) (2010) 807–813
25. Huang, G.B., Ramesh, M., Berg, T., Learned-Miller, E.: Labeled faces in the wild: A database for studying face recognition in unconstrained environments. Technical report, ECCV Workshop on Faces in Real-life Images (2008)
26. Wolf, L., Hassner, T., Taigman, Y.: Effective unconstrained face recognition by combining multiple descriptors and learned background statistics. IEEE Trans. Pattern Anal. Mach. Intell. **33**(10) (2011) 1978–1990
27. Kemelmacher-Shlizerman, I., Seitz, S.M., Miller, D., Brossard, E.: The MegaFace benchmark: 1 million faces for recognition at scale. In: Proc. IEEE Conf. Computer Vision and Pattern Recognition (CVPR). (2016) 4873–4882
28. Nech, A., Kemelmacher-Shlizerman, I.: Level playing field for million scale face recognition. In: Proc. IEEE Conf. Computer Vision and Pattern Recognition (CVPR). (2017)
29. Zhao, H., Gallo, O., Frosio, I., Kautz, J.: Loss functions for image restoration with neural networks. IEEE Transactions on Computational Imaging **3**(1) (2017) 47–57
30. Xiong, X., De la Torre, F.: Supervised descent method and its applications to face alignment. In: Proc. IEEE Conf. Computer Vision and Pattern Recognition (CVPR). (2013) 532–539
31. Ng, H.W., Winkler, S.: A data-driven approach to cleaning large face datasets. In: Proc. IEEE Int'l Conf. Image Processing (ICIP), IEEE (2014) 343–347
32. Tuzel, O., Marks, T.K., Tambe, S.: Robust face alignment using a mixture of invariant experts. In: Proc. European Conf. Computer Vision (ECCV), Springer (2016) 825–841
33. Yang, C.Y., Ma, C., Yang, M.H.: Single-image super-resolution: A benchmark. In: Proc. European Conf. Computer Vision (ECCV). (2014) 372–386
34. Jia, Y., Shelhamer, E., Donahue, J., Karayev, S., Long, J., Girshick, R., Guadarrama, S., Darrell, T.: Caffe: Convolutional architecture for fast feature embedding. arXiv preprint arXiv:1408.5093 (2014)
35. Hinton, G., Srivastava, N., Swersky, K.: RMSProp: Divide the gradient by a running average of its recent magnitude. Neural networks for machine learning, Coursera lecture 6e (2012)
36. Dong, C., Loy, C.C., He, K., Tang, X.: Learning a deep convolutional network for image super-resolution. In: Proc. European Conf. Computer Vision (ECCV). (2014) 184–199
37. Kim, J., Kwon Lee, J., Mu Lee, K.: Accurate image super-resolution using very deep convolutional networks. In: Proc. IEEE Conf. Computer Vision and Pattern Recognition (CVPR). (2016) 1646–1654
38. Ma, X., Zhang, J., Qi, C.: Hallucinating face by position-patch. Pattern Recognition **43**(6) (2010) 2224–2236
39. Yang, C.Y., Liu, S., Yang, M.H.: Structured face hallucination. In: Proc. IEEE Conf. Computer Vision and Pattern Recognition (CVPR). (2013) 1099–1106