Attribute-Guided Coupled GAN for Cross-Resolution Face Recognition

Veeru Talreja*, Fariborz Taherkhani*, Matthew C. Valenti, and Nasser M. Nasrabadi
West Virginia University
Morgantown, WV, USA
vtalreja@mix.wvu.edu,fariborztaherkhani@gmail.com,valenti@ieee.org,nasser.nasrabadi@mail.wvu.edu

Abstract

In this paper, we propose a novel attribute-guided cross-resolution (low-resolution to high-resolution) face recognition framework that leverages a coupled generative adversarial network (GAN) structure with adversarial training to find the hidden relationship between the low-resolution and high-resolution images in a latent common embedding subspace. The coupled GAN framework consists of two sub-networks, one dedicated to the low-resolution domain and the other dedicated to the high-resolution domain. Each sub-network aims to find a projection that maximizes the pair-wise correlation between the two feature domains in a common embedding subspace. In addition to projecting the images into a common subspace, the coupled network also predicts facial attributes to improve the cross-resolution face recognition. Specifically, our proposed coupled framework exploits facial attributes to further maximize the pair-wise correlation by implicitly matching facial attributes of the low and high-resolution images during the training, which leads to a more discriminative embedding subspace resulting in performance enhancement for cross-resolution face recognition. The efficacy of our approach compared with the state-of-the-art is demonstrated using the LFWA, Celeb-A, SCFace and UCCS datasets.

1. Introduction

Facial biometrics is used in a variety of modern recognition and surveillance applications ranging from stand-alone camera applications in banks and supermarkets to multiple networked closed-circuit televisions in law enforcement applications or even in cloud-based authentication applications. [40–45]. The large distance between surveillance cameras and the subjects leads to low-resolution (LR) face regions in the captured images. Usually, the discriminant properties of the face are degraded in the LR images, which leads to a significant drop in the accuracy of traditional face recognition algorithms developed for high-resolution (HR) images. An efficient face-recognition algorithm should perform well even for LR faces without significantly reducing recognition accuracy.

In comparison to HR face images, LR faces have their own unique visual properties. Although many visual features are missing in LR face images, humans are still able to notice similarities between the LR and HR face images of a given subject. This implies that the neural systems of the human brain is able to recover missing visual properties of LR faces if the human brain is familiar with the high-resolution image of that subject or a given identity [5]. Inspired by this fact, several LR face recognition models have been introduced that can be generally divided into two categories: the hallucination category and the embedding category. The models in the hallucination category reconstruct HR faces from LR faces before recognition [3, 11, 18, 19, 35, 46, 51]. The hallucination category of super-resolution is also used for other applications [4]. For instance, Kolouri and Rohde [18] introduced a method based on optimal transport for single frame super-resolution to automatically build a nonlinear Lagrangian model of HR facial appearance. Thereafter, the LR facial image is improved by exploring the parameters of the model which perfectly fit the given LR data.

Methods based on hallucination usually achieve promising results in recognizing the reconstructed HR face images. However, the super-resolution operation in hallucination models usually requires significant additional computation that often translates to a reduction in the recognition speed. In contrast to methods based on the hallucination, methods in the embedding category extract features from LR faces by leveraging various external face contexts. Ren et al. [29] introduce a coupled kernel embedding to implicitly map face images with different resolutions into an infinite space. The recognition task is then performed in this new space to minimize the dissimilarities obtained by their kernel Gram matrices in the low and high-resolution spaces, respectively. Intuitively, the main step in the embedding

*Authors Contributed Equally

978-1-7281-1522-1/19/$31.00

method is to transfer knowledge from HR to LR face images. However, in these methods, one must be careful to transfer only the desired knowledge instead of transferring all knowledge from a HR domain to a LR domain.

In addition to knowledge sharing between the high and low-resolution images, soft biometric traits such as facial attributes can also be used as complimentary information to improve the cross-resolution face recognition model. Facial attributes have been previously used jointly with face biometrics in different face recognition applications [2].

In this paper, we present an embedding model for cross-resolution face recognition based on novel attribute-guided deep coupled learning framework using generative adversarial network (GAN) to find the hidden relationship between the features of high-resolution and low-resolution images in a latent common embedding subspace. The framework also utilizes convolutional neural network (CNN) weight sharing followed by dedicated weights for learning the representative features for each specific face attribute. Specifically, our coupled framework exploits the facial attribute to further maximize the correlation between the low-resolution and high-resolution domains, which leads to a more discriminative embedding subspace to enhance the performance of the main task, which is the cross-resolution face recognition. Additionally, in our approach, we also predict the attributes for low-resolution images along with cross-resolution face recognition in a multi-tasking paradigm. Multi-task learning attempts to solve correlated tasks simultaneously by leveraging the knowledge sharing between the two tasks [37–39]. To summarize, our main contributions are:

- A novel attribute guided cross-resolution (low-resolution to high-resolution) face recognition model using coupled GAN and multiple loss functions.
- A mutli-task learning framework to predict facial attributes for low-resolution facial images.
- Extensive experiments using four different datasets and a comparison of the proposed method with state-of-the-art methods.

2. Related Work

There are two categories of methods for low-resolution face recognition. The hallucination based methods [3, 19, 35, 46, 51] reconstruct high-resolution faces before the face recognition, while embedding based methods extract latent features directly from low-resolution and high-resolution faces by using the embedding techniques. Yang et al. [51] leverage sparse representation to simultaneously perform recognition and hallucination to synthesize person-specific versions of low-resolution faces without a significant drop in recognition. In [46], an algorithm is proposed to recognize faces via sparse representation with a specific dictionary that includes many natural and facial images. Furthermore, deep models such as the ones presented in [3] and [19] can generate intensely realistic high-resolution images from low-resolution faces. Nevertheless, the speed of such super-resolution or hallucination methods might be a bit slow because of the complex high-resolution face reconstruction process, which restricts their direct use in applications where computational resources are limited.

Rather than reconstructing high-resolution faces, a more direct method is embedding low-resolution faces into different external contexts to retrieve the missing information during resolution deterioration [5]. Inspired by this, embedding methods have been proposed to transform both the high and low-resolution faces into a integrated feature domain for matching [5, 8, 12, 20, 23, 24, 33, 47, 48, 50, 52]. In [23], the multi-dimensional scaling is used to learn a common transformation matrix to jointly transform the facial features of low and high-resolution training face images. On the other hand, Wang et al. [48] solve very low-resolution recognition problem via deep learning approaches. In [9], CNNs are used with a manifold-based track comparison technique for low-resolution face recognition in videos. It is worth mentioning that the core idea of the embedding-based models is to transfer the knowledge from high-resolution face images, which is also the main idea of our proposed method.

3. Generative Adversarial Network

Generative adversarial networks (GANs) have been widely used in different computer vision application such as style transfer, sketch to photo synthesis, and also in military applications [14–16, 25, 26, 36]. GANs consist of two competing networks, namely a generator G and discriminator D. The goal of GAN is to train the generator G to produce samples from training noise distribution $p_z(z)$ such that the discriminator D cannot distinguish the synthesized samples from actual data y with distribution p_{data}. Generator $G(z;\theta_g)$ is a differentiable function which maps the noise variable z to a data space using the parameters θ_g. On the other hand, discriminator $D(.;\theta_d)$ is also a differentiable function, which tries to discriminate using a binary classification between the real data y and $G(z)$. Specifically, the generator and discriminator compete with each other in a two-player minimax game to minimize the Jenson-Shannon divergence [6]. The loss function $L(D,G)$ for GAN is given as:

$$\begin{aligned} L(D,G) = {} & E_{y\sim P_{data}(y)}[\log D(y)] \\ & + E_{z\sim P_z(z)}[\log(1-D(G(z)))] \end{aligned} \tag{1}$$

The objective (two player minimax game) for GAN is given by:

$$\begin{aligned} \min_G \max_D L(D,G) = \min_G \max_D [& E_{y\sim P_{data}(y)}[\log D(y)] \\ & + E_{z\sim P_z(z)}[\log(1-D(G(z)))]] \end{aligned} \tag{2}$$

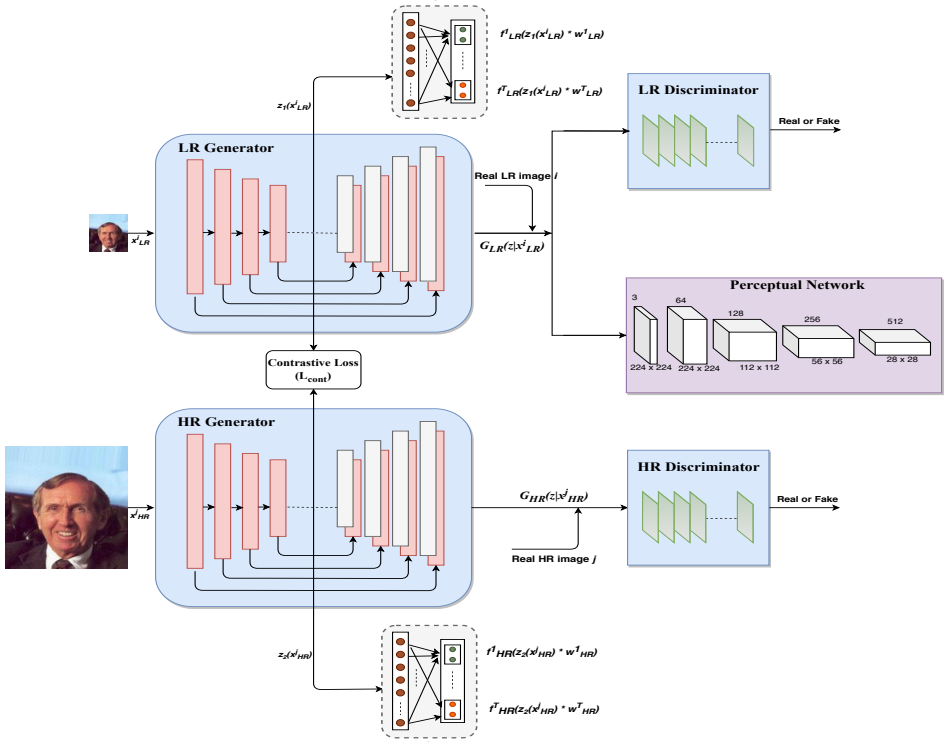


Figure 1: Block diagram of the proposed framework.

Conditional GAN is another variant of GAN where both the generator and the discriminator are conditioned on an additional variable x. This additional variable could be any kind of auxilary information such as discrete labels [22], text [28], or images [10]. The loss function for conditional GAN is given as:

$$L_c(D,G) = E_{y\sim P_{data}(y)}[\log D(y|x)] + E_{z\sim P_z(z)}[\log(1 - D(G(z|x)))]. \quad (3)$$

The objective for the conditional GAN is the same two player minimax game as in (2) with loss function as $L_c(D,G)$. Hereafter, we will denote the objective for conditional GAN as $O_{cGAN}(D,G,y,x)$, which is given by:

$$O_{cGAN}(D,G,y,x) = \min_G \max_D [E_{y\sim P_{data}(y)}[\log D(y|x)] + E_{z\sim P_z(z)}[\log(1 - D(G(z|x)))]]. \quad (4)$$

4. Proposed Method

In this section, we describe the proposed method for cross-resolution face recognition. In contrast to the hallucination approach, we do not up-sample each low-resolution image to the high-resolution domain before matching. Instead, we seek to project the high and low-resolution images to a common latent low-dimensional embedding subspace using generative modeling. Inspired by the success of GANs [6], we explore adversarial networks in a multi-tasking paradigm to project low and high-resolution images to a common subspace for recognition, and also predict facial attributes from low recognition images.

As shown in Fig. 1, the proposed method consists of a coupled framework made of two sub-networks, where each sub-network is a GAN architecture made of a generator and a discriminator. The generators are coupled together using a contrastive loss function. Each generator is also responsible to predict attributes in a multi-tasking paradigm. In addition to the adversarial loss, and contrastive loss, we propose to guide the sub-networks using a perceptual loss based on the VGG 16 architecture and also an L_2 reconstruction error. This is because the perceptual loss in optimization helps to achieve a realistic image reconstruction [13].

4.1. Deep Coupled Framework

The objective of our method is the recognition of low-resolution face images with respect to a gallery of high-resolution images, which have not been seen during the training. The matching of the low-resolution and the high-resolution images is performed in a common embedding subspace. For this reason, we use a coupled framework which contains two sub-networks: low-resolution (LR) network and high-resolution (HR) network. The LR network consists of a GAN (generator + discriminator), attribute predictor, and a perceptual network based on VGG-16, while the HR network consists of a GAN (generator + discrimina-

tor) and an attribute predictor.

For the generators, we have used a U-Net network [30] to better capture the low-level features and overcome the vanishing gradient problem due to deep network. Motivated by [10], we have used patch-based discriminators, which are trained iteratively along with the respective generators. Patch-based discriminator ensures preserving of high-frequency details which are usually lost when only L_1 loss is used. The final objective of our proposed method is to find the global deep latent features in a common embedding subspace representing the relationship between the low-resolution and their corresponding high-resolution face images. To find this common subspace between the two domains, we couple the two generators via a contrastive loss function L_{cont} [1].

This loss function (L_{cont}) is minimized so as to drive the genuine pairs (i.e., a LR image with its own corresponding HR image) towards each other in a common embedding subspace, and at the same time, push the impostor pairs (i.e., a LR image of a subject with another subject's HR image) away from each other. Let x^i_{LR} denote the input LR face image, and x^j_{HR} denote the input HR image. $c(i,j)$ is a binary label, which is equal to 0 if x^i_{LR} and x^j_{HR} belong to the same class (i.e., genuine pair), and equal to 1 if x^i_{LR} and x^j_{HR} belong to the different class (i.e., impostor pair). Let $z_1(.)$ and $z_2(.)$ denote the deep convolutional neural network (CNN)-based embedding functions to transform x^i_{LR} and x^j_{HR}, respectively into a common latent embedding subspace. Then, contrastive loss function (L_{cont}) if $c(i,j) = 0$ (i.e., genuine pair) is given as:

$$L_{cont}(z_1(x^i_{LR}), z_2(x^j_{HR}), c(i,j)) = \frac{1}{2}\left\|z_1(x^i_{LR}) - z_2(x^j_{HR})\right\|^2_2. \tag{5}$$

Similarly if $c(i,j) = 1$ (i.e., impostor pair), then contrastive loss function (L_{cont}) is :

$$L_{cont}(z_1(x^i_{LR}),z_2(x^j_{HR}), c(i,j)) = \frac{1}{2}\max\left(0, m - \left\|z_1(x^i_{LR}) - z_2(x^j_{HR})\right\|^2_2\right), \tag{6}$$

where m is the contrastive margin and is used to "tighten" the constraint. Therefore, the total loss function for coupling the sub-networks is denoted by L_{cpl} and is given as:

$$L_{cpl} = \frac{1}{N^2}\sum_{i=1}^{N}\sum_{j=1}^{N} L_{cont}(z_1(x^i_{LR}), z_2(x^j_{HR}), c(i,j)), \tag{7}$$

where N is the number of training samples. The main motivation for using the coupling loss is that it has the capacity to find the discriminative embedding subspace because it uses the class labels implicitly, which may not be the case with some other metric such as Euclidean distance. This discriminative embedding subspace would be useful for matching of the LR images with the HR images and also for attribute prediction task.

4.2. Attribute Prediction Task

In addition to cross-resolution face recognition, another important objective of our proposed method is prediction of attributes using a LR or HR face image. However, separating these two objectives by learning multiple CNNs individually is not optimal since different objectives may share common features and have hidden relationship, which can be leveraged to jointly optimize the objectives. This notion of joint optimization has been used in [53], where they train a CNN for face recognition, and utilize the features for attribute prediction. Therefore, for this task, we use the respective feature set (i.e., $z_1(x^i_{LR})$ for LR, or $z_1(x^j_{HR})$ for HR) from the common embedding subspace to also predict the attributes for a given image. Also, our network shares a large portion of its parameters among different attribute prediction tasks in order to enhance the performance of the recognition task in a mutli-task paradigm.

For the attribute prediction task, a LR or HR image is given as input to the network to predict a set of attributes. Consider that the input is a LR image denoted by x^i_{LR}, where the class label for the image is given by $\ell^i \in L$ for $i = 1, \cdots, N$ where N is the number of training samples. Let's consider T to be the number of different facial attributes and $a^{i,t}$ denotes the ground truth attribute label for training sample i and attribute t for $t = 1 \cdots T$. In this case, using the feature set from the common embedding subspace, the attribute prediction loss function is given as:

$$L_{aLR} = \frac{1}{N}\sum_{i=1}^{N}\sum_{t=1}^{T} l(f^t_{LR}(z_1(x^i_{LR}) \times w^t_{LR}), a^{i,t}), \tag{8}$$

where $f^t_{LR}(.)$ is a binary classifier for the attribute t operated on the bottle-neck of LR generator as shown in Fig. 1. The classifier is learned by using a loss function l (e.g., cross entropy) and w^t_{LR} represents the weight parameters for the classifier and these parameters are learned separately for each facial attribute task.

Similarly, we can consider the other sub-network (HR network) and perform the same procedure with a HR image. The HR network also predicts the set of facial of attributes using the features $z_2(x^j_{HR})$ from the common embedding subspace for a given HR image. Therefore, the loss function for facial attribute prediction for HR image is given as:

$$L_{aHR} = \frac{1}{N}\sum_{j=1}^{N}\sum_{t=1}^{T} l(f^t_{HR}(z_2(x^j_{HR}) \times w^t_{HR}), a^{j,t}), \tag{9}$$

where the notations are similar to (8) but correspond to the HR network. The total attribute prediction loss given as:

$$L_a = L_{aLR} + L_{aHR}. \tag{10}$$

4.3. Generative Adversarial Loss

Let G_{LR} and G_{HR} denote the generators that synthesize the corresponding LR and HR images from the input LR and HR image, respectively. Let D_{LR} and D_{HR} denote the discriminators for LR and HR GANs respectively. We have utilized the GAN loss function [6] to train the generators and the corresponding discriminators in order to ensure that the discriminators cannot distinguish the synthesized images by the generators from the corresponding ground truth images. Also, it can be observed from Fig. 1, that the generators G_{LR} and G_{HR} try to generate a LR and HR image with the network conditioned on the input LR and HR image, respectively. The total loss for the coupled GAN is given by:

$$L_{GAN} = L_{LR} + L_{HR}, \tag{11}$$

where L_{LR} and L_{HR} denote the GAN loss functions for the LR and the HR network, respectively and are given as:

$$L_{LR} = O_{cGAN}(D_{LR}, G_{LR}, y^i, x^i_{LR}) \tag{12}$$

$$L_{HR} = O_{cGAN}(D_{HR}, G_{HR}, y^j, x^j_{HR}), \tag{13}$$

where function O_{cGAN} is given by (4). x^i_{LR} (x^j_{HR}) is the LR (HR) image used as a condition for the LR (HR) GAN and y^i (y^j) denotes the real LR (HR) data. Note that the real LR (HR) data y^i (y^j) and the network condition given by x^i_{LR} (x^j_{HR}) are the same.

4.4. Perceptual Loss

Perceptual loss function was introduced in [13] for style transfer and super-resolution. In [13], instead of relying only on L_1 or L_2 reconstruction error, the network parameters are learned using errors between high-level image feature representations extracted from a pre-trained convolutional neural network. Similarly, in our proposed approach, perceptual loss is added only to the LR network using a pre-trained VGG-16 [34] network to extract high-level features (ReLU3-3 layer) and the L_1 distance between these features of real and synthesized images is used to guide the generator G_{LR}. The perceptual loss for features for only the LR network is:

$$\begin{aligned} L_{P_{LR}} = & \frac{1}{C_p W_p H_p} \sum_{c=1}^{C_p} \sum_{w=1}^{W_p} \sum_{h=1}^{H_p} \\ & \left\| V(G_{LR}(z|x^i_{LR}))^{c,w,h} - V(y^i)^{c,w,h} \right\|, \end{aligned} \tag{14}$$

where y^i is the ground truth LR image, $G_{LR}(z|x^i_{LR})$ is the output of the LR generator. $V(.)$ represents a particular layer of the VGG-16 network, where the layer dimensions are given by C_p, W_p, and H_p. We have applied the perceptual loss only for the LR network to generate more sharper LR images helpful for recognition.

Similarly, we utilized perceptual loss for attribute prediction as well to measure the difference between the facial attributes of the synthesized and the real image. We applied the perceptual loss for attributes for both the LR and the HR network. To extract the attributes from a given HR image, we fine-tune the pre-trained VGG-Face [27] on 12 annotated facial attributes, which are shown in Table 2. After this, we utilize this attribute predictor to measure the attribute perceptual loss for both the LR and HR networks and the respective losses are given below:

$$L_{pa_{LR}} = \left\| A(G_{LR}(z|x^i_{LR})) - A(y^i) \right\|_2^2, \tag{15}$$

$$L_{pa_{HR}} = \left\| A(G_{HR}(z|x^j_{HR})) - A(y^j) \right\|_2^2, \tag{16}$$

where A(.) is the fine-tuned VGG-Face attribute predictor network. The total attribute perceptual loss is the sum of the perceptual attribute loss for the LR network ($L_{pa_{LR}}$) and the HR network ($L_{pa_{HR}}$):

$$L_{pa} = L_{pa_{LR}} + L_{pa_{HR}}. \tag{17}$$

4.5. L_2 Reconstruction Loss

L_2 reconstruction loss measures the reconstruction error in terms of Euclidean distance between the synthesized image and the corresponding real image and is defined for the LR and the HR network as follows:

$$L_{2_{LR}} = \left\| G_{LR}(z|x^i_{LR}) - y^i \right\|_2^2 \tag{18}$$

$$L_{2_{HR}} = \left\| G_{HR}(z|x^j_{HR}) - y^j \right\|_2^2. \tag{19}$$

The total L_2 reconstruction loss function is given by:

$$L_2 = L_{2_{LR}} + L_{2_{HR}}. \tag{20}$$

4.6. Overall Objective Function

The overall objective function for learning the network parameters in the proposed method is given as the sum of all the above defined loss functions:

$$\begin{aligned} L_{tot} = & L_{cpl} + \lambda_1 L_a + \lambda_2 L_{GAN} \\ & + \lambda_3 L_{P_{LR}} + \lambda_4 L_{pa} + \lambda_5 L_2, \end{aligned} \tag{21}$$

where L_{cpl} is the coupling loss function, L_a is the total attribute prediction loss function, L_{GAN} is the total generative adversarial loss function, $L_{P_{LR}}$ is the perceptual loss for the LR network, L_{pa} is the total perceptual attribute loss function and L_2 is the total reconstruction error. $\lambda_1, \lambda_2, \lambda_3, \lambda_4, \lambda_5$ are the adjustable hyper-parameters to weigh the different loss terms.

5. Experiments and Results

In this section, we demonstrate the effectiveness of the proposed approach by conducting various experiments on four datasets: Labeled Faces in the Wild-a (LFWA) [49], CelebFaces Attributes Dataset (CelebA) [21], Surveillance Camera Face (SCFace) [7], and UnConstrained College Students (UCCS) Dataset [32]. We have compared our proposed method with six state-of-the-art methods on different datasets: VLRR [48], DCA [8], LRFRW [20], D-align [24], SHSR [35] and SKD [5]. In addition, we conduct an ablation study to demonstrate the effectiveness of each loss function of our network.

5.1. Datasets

CelebA consists of 202,599 images with training, validation and test splits of approximately 162,000, 20,000 and 20,000 images, respectively. The total dataset corresponds to about 10,000 identities (20 images per identity) with no identity overlap. Images are annotated with 40 facial attributes such as, "wavy hair", "chubby", "bald", "male", etc. However, we only use 12 attributes (shown in Table 2) for our proposed method. We use the pre-cropped version of the dataset, where the face images aligned using the hand-labeled key points. The image size in regular HR resolution is equal to 178×212. We downsample the images to low-resolutions of 88×108, 68×84, 48×58.

LFWA has a total of 13,232 images of 5,749 identities with pre-defined train and test splits dividing the entire dataset into approximately two equal partitions. Each image is annotated with the same 40 attributes used in CelebA dataset. The images are normalized to 224×224 for HR image and downsampled to 96×96, 64×64, 32×32.

The SCface Dataset consists of 130 subjects, each having one HR frontal face image and multiple HR images, captured from three distances (4.2m, 2.6m and 1.0m, respectively) using different quality surveillance cameras. For fair comparison with previous methods [8], 50 subjects are randomly selected for training and the rest 80 subjects for testing. As in [8], we fix the HR image at 128×128 and downsample to 64×64, 32×32 and 16×16 for LR images.

The UCCS dataset is a very challenging dataset taken under unconstrained conditions. Following the experimental setting in [48], we perform evaluations on a 180-subject subset, where each subject has 25 or more images. We get a total of 5,220 images and and use 4,200 images for training, and the remaining 1,020 images for testing. For fair comparison, we normalize the cropped face regions to 80×80 as HR, and downsample them for LR images of 16×16. For the datasets SCFace and UCCS, which are not annotated with attributes, we use the state-of-the-art mixed-objective optimization network (MOON) [31] for generating the ground truth attributes.

5.2. Training Details

As mentioned, we have implemented the U-Net network as generators and patch based discriminators for the LR and HR networks. The entire architecture has been been implemented in Pytorch. For convergence, all the hyper-parameters are set to 1 except λ_3, and λ_4, which are set to 0.5. We have used a batch size of 6 for Adam optimizer [17] with first-order momentum of 0.5, and learning at a rate of 0.0004. We have used ReLU activation for the generator and Leaky ReLU with a slope of 0.25 for the discriminator. For fine-tuning the attribute predictor network VGG-Face for attribute perceptual loss, we have chosen 12 attributes (shown in Table 2 from the LFWA dataset).

The complex loss in (21) makes it difficult to train the whole network directly as the gradient diffusion caused by different tasks will lead to slow network convergence. To address this issue, we have employed a stage-wise learning strategy, where the information in the training data is presented to the network gradually. Specifically, we first optimize each task greedily by not updating the other task simultaneously. After the initialization for each task, we fine-tune the whole network all together by optimizing all the tasks jointly.

For training, we require genuine and impostor pairs. The genuine/impostor pairs are constructed using LR and HR images of same/different subject. We balance the training set by using same number of genuine and impostor pairs.

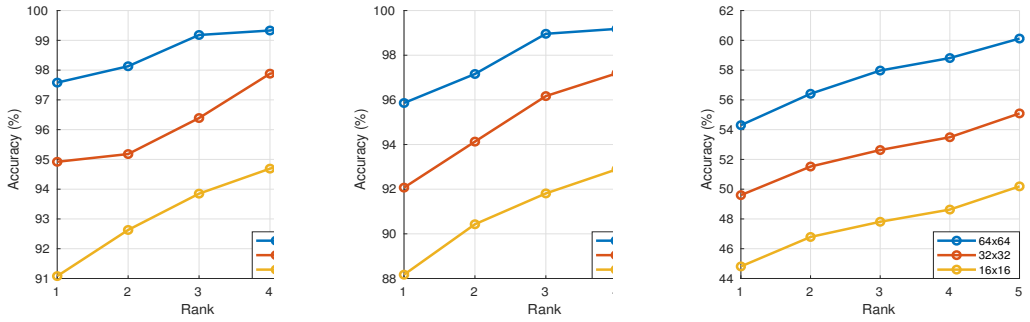
5.3. Testing of the Proposed Method

The main objective of our proposed method is to match a test LR image against a gallery of HR images using the corresponding feature set from the common latent embedding subspace. During testing, a given probe LR image x^p_{LR} is passed through the LR network, and $z_1(x^p_{LR})$ is generated from the common embedding subspace. Similarly, the HR images from the gallery are passed through the HR network and $z_2(x^j_{HR})$ is measured for each image x^j_{HR}. Eventually, the face recognition is performed by calculating the minimum Euclidean distance between $z_1(x^p_{LR})$ and $z_2(x^j_{HR})$ for all the gallery HR images:

$$\hat{j} = \arg\min_j \left\| z_1(x^p_{LR}) - z_2(x^j_{HR}) \right\|_2^2 . \tag{22}$$

Therefore, $x^{\hat{j}}_{HR}$ is the matching HR image from the gallery for the given probe LR image x^p_{LR}. The ratio of the number of correctly classified probes to the total number of probes is computed as the identification rate.

Additionally, the LR network can also be used for facial attribute prediction of a given LR probe image by passing the feature set $z_1(x^p_{LR})$ through the attribute predictor of the LR network. The predicted facial attribute can be used to narrow down the search for identification in a large gallery of HR images.



(a) LFWA (b) CelebA (c) SCFace

Figure 2: CMC curves for rank-n recognition accuracy for different low-resolution images for different datasets.

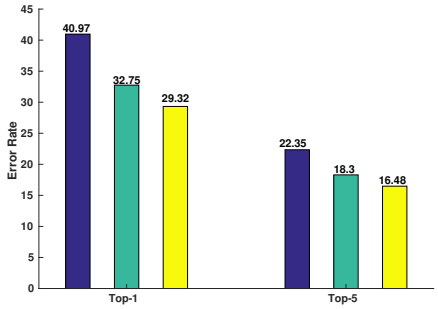


Figure 3: Top-1 and Top-5 Error rate comparison for VLLR (blue), SKD (Green), and our method (Yellow) using the UCCS dataset.

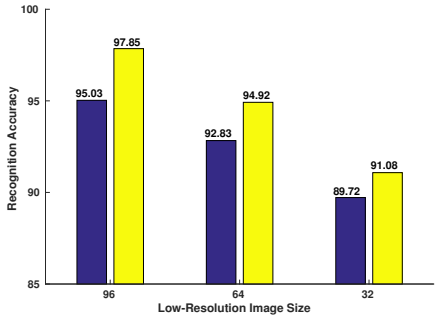


Figure 4: Recognition accuracy (%) comparison for SKD (Blue), and our method (Yellow) using the LFWA dataset for different size of LR image.

5.4. Performance Evaluation

We have evaluated the proposed method and compared with other state-of-the-art methods on four different datasets using different low-resolution images. Fig. 2 provides the recognition accuracy of our proposed method from rank-1 to rank-5 for different resolution images using the LFWA, CelebA and SCFace dataset. We can clearly see that the proposed method gives very good performance for the LFWA and CelebA dataset. However, the SCFace has more challenging face variations than the LFWA and CelebA and the SCFace images have been taken in a typical commercial surveillance environment, which leads to lower recognition performance for the SCFace dataset when compared to LFWA and CelebA as seen in Fig. 2.

Table 1: Rank-1 recognition accuracy (%) on SCFace.

Model	Dist-1	Dist-2	Dist-3	Average
SHSR	14.70	15.70	19.10	16.50
DCA	12.19	18.44	25.53	18.72
LRFRW	20.40	20.80	31.71	18.72
D-Align	34.37	39.38	49.37	24.30
SKD	43.50	48.00	53.50	48.33
Our Method	44.81 ± 0.36	49.60 ± 0.41	54.30 ± 0.23	49.57 ± 0.39

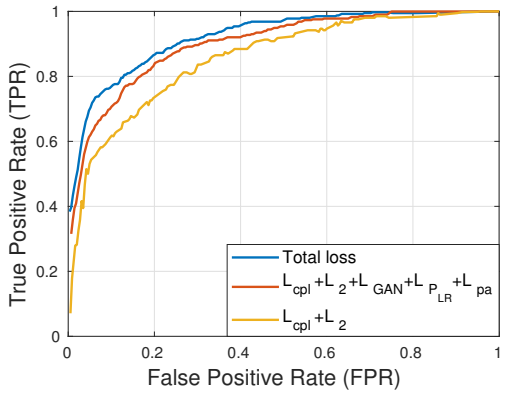


Figure 5: ROC curves corresponding to the ablation study.

Table 1 tabulates the comparison of Rank-1 recognition rates with different state-of-the-art methods for the SCFace dataset using different distances from the camera, which corresponds to different resolution images. We can observe that our proposed method outperforms the state-of-the-art embedded method of cross-resolution face recognition model SKD [5] by 1.31%,1.60%,0.80%, and 1.22% on Dist-1, Dist-2, Dist-3, and average, respectively. We can also observe that our model even outperforms the state-of-the-art hallucination model SHSR [35] by approximately 33% an average for all the three distances. We have also compared proposed method with VLRR [48], and SKD [5]

Table 2: Attribute prediction accuracy (%) comparison using CelebA dataset.

	Double Chin	Chubby	Eye Glasses	Male	Pale Skin	Moustache	Mouth Slightly Open	Young	Smiling	Goatee	Bald	Blond Hair
HR Input (Net A)	90.8	90.1	96.5	97.5	89.2	92.8	90.1	85.3	91.3	93.8	97.4	92.7
LR Input (Net A)	51.3	50.1	56.5	55.3	46.4	53.1	51.6	49.2	55.4	56.9	62.8	59.7
LR Input (fine-tuned Net A)	71.6	69.3	74.6	72.9	68.6	71.4	70.3	69.5	75.8	76.0	88.5	81.4
LR Input (Proposed Method)	83.2	82.6	89.3	91	83.3	88.1	86.6	78.9	87.2	88.9	93.6	88.4

using the UCCS dataset. Top-1 and Top-5 error rate comparison has been shown in Fig. 3. UCCS is also a very challenging dataset, where the faces have been captured in completely unconstrained conditions. Due to this reason, the error rates for this dataset are very high. However, our proposed method performs better than the other two compared method by giving a lower error rate of at least 3.4% and 1.8% for Top-1 and Top-5 recognition. We can also notice from Fig. 4, that our proposed method outperforms SKD even for the LFWA dataset for different resolutions.

From performance evaluation, we observe that our proposed coupled framework with the contrastive loss function and leveraging facial attributes to transform different domains (LR and HR) into a common discriminative embedding subspace is superior than the other embedding techniques such as SKD and D-Align. It also shows the efficacy of exploiting multiple loss functions for cross-resolution face recognition. The relative importance of the loss functions has been covered in detail in ablation study (Sec. 5.6).

5.5. Attribute Prediction for Low-Resolution

One of the advantages of the proposed method is that it can be used for attribute prediction for LR face images. To illustrate the efficacy of our proposed approach for attribute prediction of LR face images, we have performed attribute prediction for 4 different scenarios: 1) Attribute prediction for HR images with the VGG-Face based attribute predictor, which is represented as Net A in Sec. 4.4. 2) Attribute prediction for LR images using attribute predictor Net A. 3) In this scenario, we first fine-tune the attribute predictor A with annotated LR images and then use it for attribute prediction of LR test images. This will be called "fine-tuned Net A" 4) In this final case, we test our attribute predictor from LR network for attribute prediction of the LR test images (see Fig. 1). We have performed this experiment for the Celeb-A dataset using 68×88 as LR images.

The attribute prediction results for the above 4 scenarios for 12 attributes using the Celeb-A have been tabulated in Table 2. We can notice from Table 2 that our approach shows the best performance in predicting the attributes of LR images for both datasets. Fine-tuning the Net A with LR images helps in improving its performance, however it does not perform as well as our method. Additionally, the performance of our LR network attribute predictor is comparable to the Net A performance for HR images.

5.6. Ablation Study

The objective function defined in (21) contains multiple loss functions: coupling loss (L_{cpl}), attribute prediction loss (L_a), perceptual loss ($L_{P_{LR}}$, L_{pa}), L_2 reconstruction loss (L_2), and GAN loss (L_{GAN}). In this section, we study the relative importance of different loss functions and the benefit of using them in our proposed method. For this experiment, we use different variations of our proposed approach and perform the evaluation using the LFWA dataset (64×64 LR images). The variations are: 1) cross-resolution face verification using the coupled framework with only coupling loss and L_2 reconstruction loss ($L_{cpl} + L_2$); 2) cross-resolution face verification using the coupled framework with coupling loss, L_2 reconstruction loss, GAN loss and perceptual loss ($L_{cpl} + L_2 + L_{GAN} + L_{P_{LR}} + L_{pa}$); 3) cross-resolution face verification using our framework with all the loss functions ($L_{cpl}+L_2+L_{GAN}+L_{P_{LR}}+L_{pa}+L_a$).

We use the above three variations of our framework and plot the receiver operating characteristic (ROC) curve for the task of cross-resolution face verification using the features from the common embedding subspace. We can see from Fig. 5 that the generative adversarial loss and the perceptual loss (red curve) help in improving the cross-resolution verification performance, and adding the attribute prediction loss (blue curve) helps in more performance improvement. The reason for this improvement is that using facial attribute loss along with the contrastive loss leads to a more discriminative embedding subspace leading to a better face recognition performance. This also shows that multi-task learning of attribute prediction and face recognition is useful and helps in cross-resolution face recognition task.

6. Conclusion

We have proposed a novel framework which adopts a coupled GAN and exploits facial attributes for cross-resolution face recognition. The coupled GAN includes two sub-networks which project the low and high-resolution images into a common embedding subspace, where the goal of each sub-network is to maximize the pair-wise correlation between low and high-resolution images during the projection process. Moreover, we leverage facial attributes to further maximize the pair-wise correlation by implicitly matching facial attributes of the low and high-resolution images during the training. We comprehensively evaluated our model on four standard datasets and the results indicate that our model significantly outperforms other state-of-the-art models for cross resolution face recognition. Addi-

tionally, the enhancement obtained by different losses in the proposed method has been considered in an ablation study.

References

[1] S. Chopra, R. Hadsell, and Y. LeCun. Learning a similarity metric discriminatively, with application to face verification. In *Proc. IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 539–546, 2005.

[2] A. Dantcheva, P. Elia, and A. Ross. What else does your biometric data reveal? a survey on soft biometrics. *IEEE Transactions on Information Forensics and Security*, 11(3):441–467, March 2016.

[3] C. Dong, C. C. Loy, K. He, X. Tang, D. Fleet, T. Pajdla, B. Schiele, and T. Tuytelaars. Learning a deep convolutional network for image super-resolution. In *Proc. European Conference on Computer Vision (ECCV)*, pages 184–199, 2014.

[4] S. N. Ferdous, M. Mostofa, and N. M. Nasrabadi. Super resolution-assisted deep aerial vehicle detection. In *Artificial Intelligence and Machine Learning for Multi-Domain Operations Applications*, 2019.

[5] S. Ge, S. Zhao, C. Li, and J. Li. Low-resolution face recognition in the wild via selective knowledge distillation. *IEEE Transactions on Image Processing*, 28:2051–2062, 2018.

[6] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio. Generative adversarial nets. In *Proc. Neural Information Processing Systems*, pages 2672–2680. 2014.

[7] M. Grgic, K. Delac, and S. Grgic. Scface surveillance cameras face database. *Multimedia Tools and Applications*, 51:863–879, 2009.

[8] M. Haghighat and M. Abdel-Mottaleb. Low resolution face recognition in surveillance systems using discriminant correlation analysis. In *Proc. IEEE International Conference on Automatic Face & Gesture Recognition*, pages 912–917, 2017.

[9] C. Herrmann, D. Willersinn, and J. Beyerer. Low-resolution convolutional neural networks for video face recognition. In *Proc. IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS)*, pages 221–227, 2016.

[10] P. Isola, J.-Y. Zhu, T. Zhou, and A. A. Efros. Image-to-image translation with conditional adversarial networks. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5967–5976, 2017.

[11] M. Jian and K.-M. Lam. Simultaneous hallucination and recognition of low-resolution faces based on singular value decomposition. *IEEE Transactions on Circuits and Systems for Video Technology*, 25(11):1761–1772, 2015.

[12] J. Jiang, R. Hu, Z. Wang, and Z. Cai. Cdmma: Coupled discriminant multi-manifold analysis for matching low-resolution face images. *Signal Processing*, 124:162–172, 2016.

[13] J. Johnson, A. Alahi, and L. Fei-Fei. Perceptual losses for real-time style transfer and super-resolution. In *Proc. European Conference on Computer Vision (ECCV)*, 2016.

[14] H. Kazemi, S. M. Iranmanesh, and N. Nasrabadi. Style and content disentanglement in generative adversarial networks. In *Proc. IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 848–856, 2019.

[15] H. Kazemi, S. Soleymani, F. Taherkhani, S. Iranmanesh, and N. Nasrabadi. Unsupervised image-to-image translation using domain-specific variational information bound. In *Advances in Neural Information Processing Systems*, pages 10348–10358, 2018.

[16] H. Kazemi, F. Taherkhani, and N. M. Nasrabadi. Unsupervised facial geometry learning for sketch to photo synthesis. In *International Conference of the Biometrics Special Interest Group (BIOSIG)*. IEEE, 2018.

[17] D. P. Kingma and J. Ba. Adam: A method for stochastic optimization. *CoRR*, abs/1412.6980, 2015.

[18] S. Kolouri and G. K. Rohde. Transport-based single frame super resolution of very low resolution face images. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4876–4884, 2015.

[19] C. Ledig, L. Theis, F. Huszar, J. Caballero, A. P. Aitken, A. Tejani, J. Totz, Z. Wang, and W. Shi. Photo-realistic single image super-resolution using a generative adversarial network. *CoRR*, abs/1609.04802, 2016.

[20] P. Li, L. Prieto, D. Mery, and P. J. Flynn. On low-resolution face recognition in the wild: Comparisons and new techniques. *IEEE Transactions on Information Forensics and Security*, page 11, 2019.

[21] Z. Liu, P. Luo, X. Wang, and X. Tang. Deep learning face attributes in the wild. In *Proceedings of International Conference on Computer Vision (ICCV)*, 2015.

[22] M. Mirza and S. Osindero. Conditional generative adversarial nets. *arXiv preprint arXiv:1411.1784*, 2014.

[23] S. P. Mudunuri and S. Biswas. Low resolution face recognition across variations in pose and illumination. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 38(5):1034–1040, 2016.

[24] S. P. Mudunuri, S. Venkataramanan, and S. Biswas. Dictionary alignment with re-ranking for low-resolution nir-vis face recognition. *IEEE Transactions on Information Forensics and Security*, 14(4):886–896, April 2019.

[25] U. M. Osahor and N. M. Nasrabadi. Deep adversarial attack on target detection systems. In *Artificial Intelligence and Machine Learning for Multi-Domain Operations Applications*, 2019.

[26] U. M. Osahor and N. M. Nasrabadi. Design of adversarial targets: fooling deep atr systems. In *Automatic Target Recognition XXIX*, 2019.

[27] O. M. Parkhi, A. Vedaldi, A. Zisserman, et al. Deep face recognition. In *Proc. British Machine Vision Conference (BMVC)*, volume 1, page 6, 2015.

[28] S. Reed, Z. Akata, X. Yan, L. Logeswaran, B. Schiele, and H. Lee. Generative adversarial text to image synthesis. In *International Conference on Machine Learning (ICML)*, 2016.

[29] C.-X. Ren, D.-Q. Dai, and H. Yan. Coupled kernel embedding for low-resolution face image recognition. *IEEE Transactions on Image Processing*, 21(8):3770–3783, 2012.

[30] O. Ronneberger, P. Fischer, and T. Brox. U-net: Convolutional networks for biomedical image segmentation. In *Proc. International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 234–241. Springer, 2015.

[31] E. M. Rudd, M. Günther, and T. E. Boult. Moon: A mixed objective optimization network for the recognition of facial attributes. In *Proc. European Conference on Cmputer Vision (ECCV)*, 2016.

[32] A. Sapkota and T. E. Boult. Large scale unconstrained open set face database. In *2013 IEEE Sixth International Conference on Biometrics: Theory, Applications and Systems (BTAS)*, Sep. 2013.

[33] S. Shekhar, V. M. Patel, and R. Chellappa. Synthesis-based robust low resolution face recognition. *arXiv preprint arXiv:1707.02733*, 2017.

[34] K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.

[35] M. Singh, S. Nagpal, M. Vatsa, R. Singh, A. Majumdar, and IIIT-Delhi. Identity aware synthesis for cross resolution face recognition. In *Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 479–488, 2018.

[36] S. Soleymani, A. Dabouei, J. Dawson, and N. M. Nasrabadi. Adversarial examples to fool iris recognition systems. *CoRR*, abs/1906.09300, 2019.

[37] S. Soleymani, A. Dabouei, S. M. Iranmanesh, H. Kazemi, J. Dawson, and N. M. Nasrabadi. Prosodic-enhanced siamese convolutional neural networks for cross-device text-independent speaker verification. In *Proc. IEEE International Conference on Biometrics Theory, Applications and Systems (BTAS)*, Oct 2018.

[38] F. Taherkhani and M. Jamzad. Restoring highly corrupted images by impulse noise using radial basis functions interpolation. *IET Image Processing*, 12(1):20–30, 2017.

[39] F. Taherkhani, H. Kazemi, and N. M. Nasrabadi. Matrix completion for graph-based deep semi-supervised learning. In *Thirty-Third AAAI Conference on Artificial Intelligence*, 2019.

[40] F. Taherkhani, N. M. Nasrabadi, and J. Dawson. A deep face identification network enhanced by facial attributes prediction. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 553–560, 2018.

[41] F. Taherkhani, V. Talreja, H. Kazemi, and N. Nasrabadi. Facial attribute guided deep cross-modal hashing for face image retrieval. In *International Conference of the Biometrics Special Interest Group (BIOSIG)*. IEEE, 2018.

[42] V. Talreja, T. Ferrett, M. C. Valenti, and A. Ross. Biometrics-as-a-service: A framework to promote innovative biometric recognition in the cloud. In *IEEE International Conference on Consumer Electronics (ICCE)*, 2018.

[43] V. Talreja, S. Soleymani, M. C. Valenti, and N. M. Nasrabadi. Learning to authenticate with deep multibiometric hashing and neural network decoding. *CoRR*, abs/1902.04149, 2019.

[44] V. Talreja, F. Taherkhani, M. C. Valenti, and N. M. Nasrabadi. Using deep cross modal hashing and error correcting codes for improving the efficiency of attribute guided facial image retrieval. In *Proc. IEEE Global Conference on Signal and Information Processing (GlobalSIP)*, pages 564–568, 2018.

[45] V. Talreja, M. C. Valenti, and N. M. Nasrabadi. Multibiometric secure system based on deep learning. In *Proc. IEEE Global Conference on Signal and Information Processing*, pages 298–302, Nov. 2017.

[46] T. Uiboupin, P. Rasti, G. Anbarjafari, and H. Demirel. Facial image super resolution using sparse representation for improving face recognition in surveillance monitoring. In *Proc. IEEE Signal Processing and Communication Application Conference (SIU)*, pages 437–440, 2016.

[47] X. Wang, H. Hu, and J. Gu. Pose robust low-resolution face recognition via coupled kernel-based enhanced discriminant analysis. *IEEE/CAA Journal of Automatica Sinica*, 3(2):203–212, 2016.

[48] Z. Wang, S. Chang, Y. Yang, D. Liu, and T. S. Huang. Studying very low resolution recognition using deep networks. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, pages 4792–4800, 2016.

[49] L. Wolf, T. Hassner, and Y. Taigman. Effective unconstrained face recognition by combining multiple descriptors and learned background statistics. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 33(10):1978–1990, Oct 2011.

[50] X. Xing and K. Wang. Couple manifold discriminant analysis with bipartite graph embedding for low-resolution face recognition. *Signal Processing*, 125:329–335, 2016.

[51] M.-C. Yang, C.-P. Wei, Y.-R. Yeh, and Y.-C. F. Wang. Recognition at a long distance: Very low resolution face recognition and hallucination. In *Proc. International Conference on Biometrics (ICB)*, pages 237–242, 2015.

[52] P. Zhang, X. Ben, W. Jiang, R. Yan, and Y. Zhang. Coupled marginal discriminant mappings for low-resolution face recognition. *Optik*, 126(23):4352–4357, 2015.

[53] Y. Zhong, J. Sullivan, and H. Li. Face attribute prediction using off-the-shelf cnn features. In *Proc. International Conference on Biometrics (ICB)*, 2016.