Subclass Heterogeneity Aware Loss for Cross-Spectral Cross-Resolution Face Recognition

Soumyadeep Ghosh *Student Member, IEEE,* Richa Singh *Senior Member, IEEE,* Mayank Vatsa *Senior Member, IEEE.*

Abstract—One of the most challenging scenarios of face recognition is matching images in presence of multiple covariates such as cross-spectral and cross-resolution. Law enforcement agencies across the world face this arduous task for which the existing face recognition algorithms do not yield the desired level of performance. In this paper, we propose a Subclass Heterogeneity Aware Loss (SHEAL) to train a deep convolutional neural network model such that it produces embeddings suitable for heterogeneous face recognition. The performance of the proposed SHEAL function is evaluated on four databases in terms of the recognition performance as well as convergence in time and epochs. We observe that SHEAL not only yields state-of-the-art results for the most challenging case of Cross-Spectral Cross-Resolution face recognition, it also achieves excellent performance on homogeneous face recognition.

Index Terms—Face recognition, Cross-spectral cross-resolution matching, Deep metric learning.

1 Introduction

The increasing effectiveness of Deep Convolutional Neural Networks (Deep-CNNs) has led to the emergence of very efficient face recognition algorithms [1], [2], [3], [4]. With this development, various applications ranging from unlocking of mobile phones and laptops to monitoring of public places are now using face recognition technology. These images are usually captured in controlled scenarios and constrained settings. However, the query images may be captured in unconstrained environment by any kind of camera; for instance, surveillance cameras. These cameras are generally placed at a high standoff distance from the subjects and have a large field-of-view [5], [6]. As a result, the effective resolution and quality of the captured face image may be low. In addition to that, when sufficient visible illumination is not available, these cameras operate in the Near-Infrared (NIR) mode and the probe images are captured in NIR spectrum. This results in a heterogeneous image/face matching (recognition) problem between the high resolution visible spectrum gallery and low resolution NIR spectrum probes (Fig. 1). The combination of the acquisition environment and the position of the user in relation to the camera location leads to three possible scenarios of heterogeneous face matching.

- Cross-Spectral matching where the visible spectrum face image (gallery) is matched with the NIR spectrum images (probes).
- Cross-Resolution matching where the high resolution face images (gallery) is matched with the low resolution images (probes) obtained from surveillance cameras.
- Cross-Spectral Cross-Resolution matching where low resolution NIR images (probe) are matched with high resolution visible spectrum mugshot images (gallery).

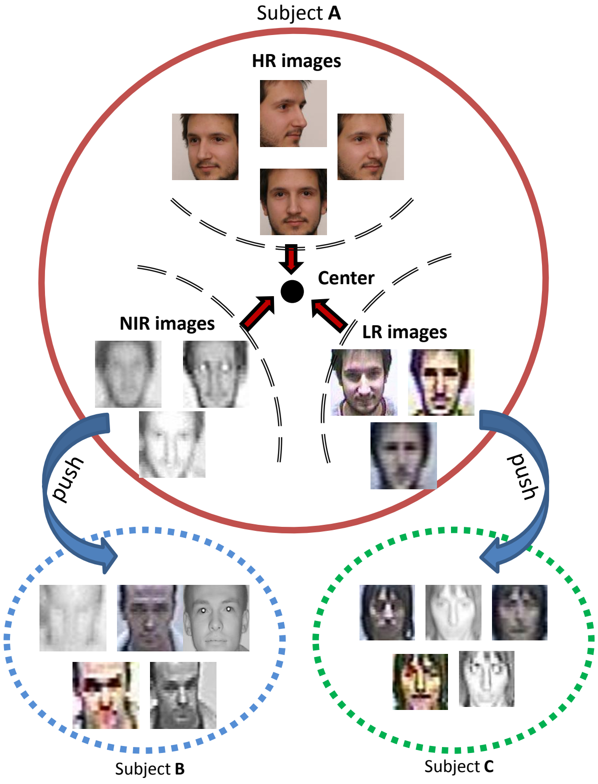


Fig. 1: Visual abstract of the proposed Subclass Heterogeneity Aware Loss (SHEAL). The intraclass distance between the different subclasses, each represented by a particular covariate such as high resolution (HR), low resolution (LR) and NIR images is minimized, while pushing other impostor classes away, in the embedding space of the model. (best viewed in color)

- *S. Ghosh is with the Department of Computer Science and Engineering, IIIT-Delhi, India. R. Singh and M. Vatsa are with the Department of Computer Science and Engineering, IIT Jodhpur, India.*
Email: soumyadeepg@iiitd.ac.in; richa@iitj.ac.in; mvatsa@iitj.ac.in

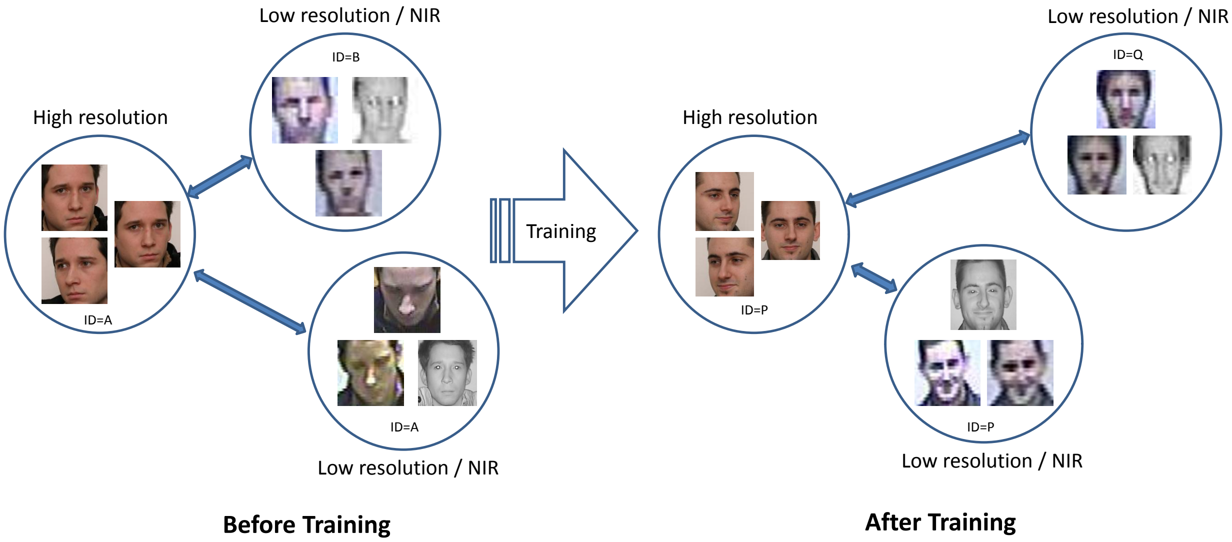


Fig. 2: Illustrating the effect of training with the proposed loss function. The proposed loss function attempts to minimize the distance between the intra-class embeddings compared to the distance between the embeddings of the images belonging to different classes.

Several researchers have proposed solutions for heterogeneous face recognition. At the core of many of these solutions lies the most fundamental concept of training a face recognition model, which is, train the model such that the intra-class is minimized and the inter-class distance is maximized, both for with intra-view (homogeneous) and inter-view (heterogeneous) data variations [7], [8], [9]. However, most of the existing algorithms focus on only one covariate at a time, either cross-resolution or cross-spectral variations, not both together. Given the increasing use of surveillance cameras for security, it is important to address both the covariates together.

In this research, we propose a unified Subclass Heterogeneity Aware Loss (SHEAL) to train a discriminative model which produces face embeddings for accurate classification in the presence of multiple face recognition covariates. A novel subclass based optimization approach is presented, which optimizes the clusters based on different subclasses in the data. As shown in Fig. 2, the proposed model learns discriminative embeddings for both high resolution and visible spectrum gallery images and low resolution, NIR spectrum probe images. These learnt embeddings are then matched using Euclidean distance. Experiments on four challenging databases, namely SCface [10], FaceSurv [11], CASIA NIR-VIS 2.0 [12], and Labeled Faces in the Wild [13], demonstrate the efficacy of the proposed approach, not only in the identification performance but also with respect to convergence in terms of training time and epochs.

2 Related Work

This paper addresses the problem of cross-spectral cross-resolution face recognition with a novel deep metric learning algorithm. Therefore, the review section first outlines the related work performed on cross-spectral and cross resolution face recognition especially using deep learning methods, followed by the literature on deep metric learning methods for face recognition.

Prior to the emergence of deep learning based face recognition algorithms, several discriminative learning and transfer learning based approaches were proposed for cross-spectral [7], [9], [12], [14], [15], [16], [17] and cross-resolution [18], [19], [20], [21] face recognition. Deep learning based algorithms have also been proposed for these tasks. Lu *et al.* [22] learned binary descriptors for heterogeneous face recognition. Yi *et al.* [23] used a shared representation learning based approach using Restricted Boltzman Machines for cross-spectral face recognition. Saxena *et al.* [24] used a metric learning based algorithm to learn a Mahalanobis distance based embedding space for the same. Lezama et *al.* [25] used a low rank embedding based approach for hallucination of NIR to visible spectrum face images for cross-spectral face matching. He et *al.* [26] proposed an algorithm to learn a deep-CNN model where the high level layer is divided into two orthogonal subspaces that learn modality-invariant representation for cross-spectral face recognition. Wu et *al.* [27] used an approximate variational formulation in a coupled deep learning framework for matching NIR face images to a gallery of visible-spectrum face images. Song et *al.* [28] proposed an adversarial discriminative learning algorithm for the same, using an integration of cross-spectral face hallucination and discriminative feature learning. Pereira et *al.* [29] proposed a deep learning approach using a framework that learns domain specific feature detectors for cross-spectral face recognition. Recently, Peng et *al.* [30] proposed a locally linear re-ranking (LLRe-Rank) approach for the same problem. He et *al.* [31] performed face completion by texture inpainting and pose correction using generative modelling for translating NIR face images for efficient matching with visible spectrum images.

Singh *et al.* [32] proposed a Synthesis via Hierarchical Sparse Representation for generating a high resolution face image from a low resolution one, for cross-resolution face recognition. Lu et *al.* [1] utilized a deep coupled end to end CNN consisting of a trunk network and two branch networks for cross-resolution face matching. Lu et *al.* [33] utilized a discriminative multidimensional scaling approach for face recognition from low resolution images. Ge et *al.* [34] proposed an approach using a two-stream CNN for low resolution face recognition. Li et *al.* [35] used a supervised discriminative learning approach for low resolution face recognition. Zangeneh et *al.* [36] proposed a novel nonlinear coupled

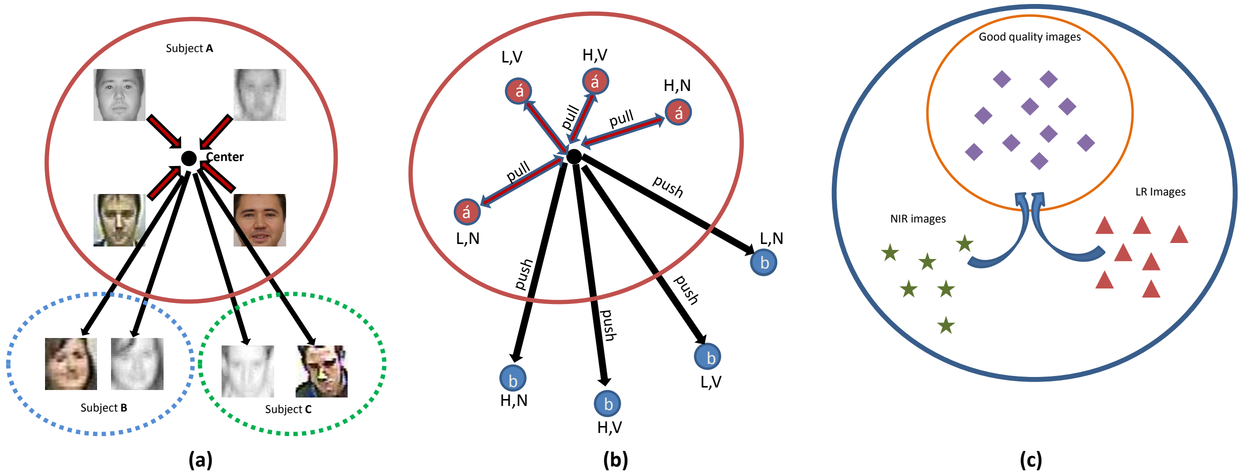


Fig. 3: Representation of the proposed method: (a) Overall motivation of the problem, (b) Illustration of the proposed loss metric which minimizes the intra-class distance and maximizes the inter-class distance (including intra-view and inter-view variations) and (c) Subclass based cluster optimization.

mapping architecture for face recognition from low resolution images. Abdollahi et *al.* [37] proposed a modified finetuning approach using different variations of the training data for low resolution face recognition. Recently Singh et *al.* [38] utilized a dual directed capsule network for very low resolution face recognition.

The popularity of deep metric learning methods has led to the development of several loss functions [39], [40], [41], [42], [43], [44], [45], [46], [47], [48] to train deep neural network models for face recognition. Schroff *et al.* [49] introduced the triplet loss based training method for face verification. Quadruplet loss [50], an extension of triplet loss, adds an extra negative sample to the loss function. This loss function enforces a stricter inter-class distance on the output embedding space of the model being trained. However, both these techniques do not consider any heterogeneity in the data during training. They also require extensive hard-sample mining for effective training. In order to account for heterogeneity in data, Liu *et al.* [51] have proposed a heterogeneous variant of triplet loss. This loss function can take at most one heterogeneity (e.g. cross-resolution) at a time and is not suitable for handling more than one covariate (e.g. cross-resolution and cross-spectral both). In addition, it required exhaustive hard mining prior to the training process. Several modifications [44], [50], [52] to the triplet loss have been proposed for a diverse range of applications such as person-re-identification, matching images of cars, object recognition, patch matching and so on. However, none of these methods addressed scenarios where matching of images with multiple heterogeneity is involved.

3 Proposed Algorithm

In this section, we illustrate the proposed algorithm which is utilized to learn a model for face recognition invariant to both spectrum and resolution. First, the framework for a heterogeneous matching problem is illustrated with only one covariate/heterogeneity (resolution and spectrum) across probes and gallery images. The formulation is then extended to include invariance to two covariates, namely resolution and spectrum. It is important to note that while the proposed loss function SHEAL, L_{SHEAL}, optimizes for heterogeneous matching with one or two covariates, it also optimizes for homogeneous matching (no covariates). The first subsection presents the formulation of SHEAL followed by the sub-class based cluster optimization. Finally, the heterogeneous face recognition algorithm is presented which learns a model with a highly discriminative output embedding space for cross-spectral cross-resolution face recognition. Fig. 3 illustrates the concept of the proposed Subclass Heterogeneity Aware Loss (SHEAL).

3.1 SHEAL: Subclass Heterogeneity Aware Loss

For a heterogeneous face matching problem, the gallery contains images with high resolution visible spectrum while the probe images are captured with different covariates present (for instance low resolution and/or NIR). For simplicity, let us assume only one kind of heterogeneity, e.g. resolution, is available in the data (i.e. gallery of high resolution images and probes are low resolution images). In order to learn a discriminative model for such a task, the loss metric needs to perform two tasks, minimizing (pulling together) and maximizing (pushing away) the intra-class and inter-class distances, respectively in intra-view[1] (homogeneous) settings, and performing the same in inter-view[2] (heterogeneous) settings. The proposed heterogeneous loss function is expressed as,

$$\begin{aligned} L = {} & [||g(X_i^H) - g({X'}_i^H)||_2^2 - ||g(X_i^H) - g(X_j^H)||_2^2 + \alpha_1]_+ \\ & + [||g(X_i^H) - g(X_i^L)||_2^2 - ||g(X_i^H) - g(X_k^L)||_2^2 + \alpha_2]_+ \quad (1) \end{aligned}$$

$$\forall (X_i^H, X_j^H, X_i^L, X_k^L) \in \tau$$

where, H and L signify high and low resolution, respectively. X_i^H is the high resolution anchor image of subject i, ${X'}_i^H$ is another high resolution image of the same subject i, X_i^L is a low resolution image of the subject i, X_j^H is the high resolution image of subject j, X_k^L is a low resolution image of another subject k where, $i \neq j \neq k$ and $[\cdot]_+ = max(\cdot, 0)$.

[1] Intra-view settings refer to the scenario when the gallery and probe are homogeneous in nature, for example, same resolution and spectrum.

[2] Inter-view settings refer to the scenario when both gallery and probe images are heterogeneous in nature, for example, different resolution or spectrum.

In a complex (more realistic) scenario, the heterogeneity may be due to two different views, namely resolution and spectrum. For example, the gallery images are in visible spectrum and high resolution, while the probes are in NIR and low resolution. Let the visible spectrum and NIR spectrum be denoted as V and N, respectively, and subscripts i, j, k, l, m represent different subjects/classes. Let the high resolution visible spectrum image of the i^{th} subject (class) be $X_i^{H,V}$. Another image of the same subject in the same setting is denoted as ${X'}_i^{H,V}$. Similarly, $X_i^{H,N}$, $X_i^{L,V}$ and $X_i^{L,N}$ represent the high resolution NIR spectrum image, low resolution visible spectrum image, and low resolution NIR spectrum image of the i^{th} subject, respectively. To accommodate both cross-resolution cross-spectral variations, the proposed loss function is formulated with two cross-views, hence require four separate terms. The first term takes care of the homogeneous matching scenario, the next two terms accommodates for cross-resolution and cross-spectral matching respectively, followed by the last term for cross-spectral cross-resolution matching.

The homogeneous loss term (L_{Ho}) is computed as,

$$L_{Ho} = [||g(X_i^{H,V}) - g({X'}_i^{H,V})||_2^2 - \\ ||g(X_i^{H,V}) - g(X_j^{H,V})||_2^2 + \alpha_1]_+ \quad (2)$$

This loss expression is composed of two parts, the former $\|g(X_i^{H,V}) - g({X'}_i^{H,V})\|_2^2$ minimizes the intra-class distance between the embedding of the anchor image $g(X_i^{H,V})$ and $g({X'}_i^{H,V})$, which is another image of the same subject captured in the same condition. The later part of the expression, $\|g(X_i^{H,V}) - g(X_j^{H,V})\|_2^2$ maximizes the inter-class distance between $g(X_i^{H,V})$ and $g(X_j^{H,V})$. However, in order to calculate the intra-class loss, we can replace the embedding of the anchor image $g(X_i^{H,V})$ by the center embedding of the i^{th} class (subject) given by $g_c(X_i^{H,V})$. In addition to that, the inter-class distances are also computed from $g_c(X_i^{H,V})$ instead of $g(X_i^{H,V})$. Therefore, for SHEAL, the loss function for the homogeneous component (L_{Ho}^c) can be written as:

$$L_{Ho}^c = [||g_c(X_i^{H,V}) - g({X'}_i^{H,V})||_2^2 - \\ ||g_c(X_i^{H,V})) - g(X_j^{H,V})||_2^2 + \alpha_1]_+ \quad (3)$$

Next, the cross-resolution loss term is expressed as:

$$L_{CR}^c = [||g_c(X_i^{H,V}) - g(X_i^{L,V})||_2^2 - \\ ||g_c(X_i^{H,V}) - g(X_k^{L,V})||_2^2 + \alpha_2]_+ \quad (4)$$

This loss expression contains two parts, the former $\|g_c(X_i^{H,V}) - g(X_i^{L,V})\|_2^2$ pertains to the distance between the center embedding of the images of the same subject i in visible spectrum and low resolution. The later term $\|g_c(X_i^{H,V}) - g(X_k^{L,V})\|_2^2$ focuses on maximizing the inter-class distance between the center embedding of another subject k in visible spectrum and low resolution. Similarly, the cross-spectral loss is expressed as,

$$L_{CS}^c = [||g_c(X_i^{H,V}) - g(X_i^{H,N})||_2^2 - \\ ||g_c(X_i^{H,V}) - g(X_l^{H,N})||_2^2 + \alpha_3]_+ \quad (5)$$

Along the same lines, the **cross-spectral cross-resolution loss** is computed as,

$$L_{CS-CR}^c = [||g_c(X_i^{H,V}) - g(X_i^{L,N})||_2^2 - \\ ||g_c(X_i^{H,V}) - g(X_m^{L,N})||_2^2 + \alpha_4]_+ \quad (6)$$

Equation 6 models the most challenging scenario where the intra-class and inter-class distances are evaluated between the high resolution visible spectrum images and images captured in low resolution and NIR. Such probe images differ from the gallery images with respect to both resolution and spectrum. The final loss function combines the homogeneous and heterogeneous losses as follows:

$$L_{SHEAL} = \lambda_1.L_{Ho}^c + \lambda_2.L_{CR}^c + \lambda_3.L_{CS}^c + \lambda_4.L_{CS-CR}^c \quad (7)$$

$$\forall({X'}_i^{H,V}, X_j^{H,V}, X_i^{L,V}, X_k^{L,V}, \\ X_i^{H,N}, X_l^{H,N}, X_i^{L,N}, X_m^{L,N}) \in \tau$$

where, τ is the set of 8-tuples. Each such 8-tuple is considered as a training sample and the coefficients $\{\lambda_1, \lambda_2, \lambda_3, \lambda_4\}$ may be used to adjust the weights of each component of the loss function. The gradient of this loss can be utilized to train the parameters of a model using representation learning (e.g. CNN model).

The traditional triplet loss works by *pulling* the embeddings of all samples of the same class towards the anchor and *pushing* the same for the impostor classes away from the anchor. However, this loss is unable to handle a heterogeneous matching problem where a pair of images of different views/modalities are to be matched during testing. In order to approach this problem we are required to train a discriminative model which can generate heterogeneity aware embeddings. To train such a model, the loss function should incorporate different matching scenarios, i.e. both homogeneous and heterogeneous. The proposed loss function (Equation 7) has been formulated by combining multiple heterogeneous variations for face matching. To summarize, the salient contributions/novelty of this work are as follows:

- We propose a method to train a discriminative model which can be utilized to match images belonging to more than one covariate. Equation 7 has four loss terms, viz L_{Ho}^c, L_{CR}^c, L_{CS}^c and L_{CS-CR}^c. Each of them contributes a gradient which is used to update the weights of the model $g(.)$ being trained.
- In Equation 7, different terms are weighed by adjustable λ parameters. If we want the model to be more often used for cross-spectral-cross-resolution matching then the coefficient of L_{CS-CR}^c can be given a higher value. This allows the model to be tuned for a specific application scenario as well.

The model $g(.)$ is trained using Equation 7 which results in disjoint clusters for each class in the output embedding space of the model. These clusters are further optimized using subclass based cluster optimization as illustrated in the next subsection.

3.2 Subclass based Cluster Optimization

We optimize the clusters (learned using L_{SHEAL}) in the embedding space of the model using a subclass based loss formulation. As shown in Fig. 3(c), in each cluster which contains embeddings of the images of a particular subject, the embeddings of the good quality images (high resolution visible spectrum) of the respective subject are expected to be very close to each other. On the other hand, the images different from the good quality ones (i.e. low resolution and NIR) are expected to be farther away in the same cluster. Using this as a hypothesis, each cluster is expected to contain two subclasses, one representing the good quality images (homogeneous) and the other for the heterogeneous images. An optimization stage is utilized to create a more compact cluster by bringing these two subclasses closer to each other. The loss function for the cluster optimization stage is expressed as,

$$\begin{aligned} L_{PP} = \beta_1.[||g_{c_1}(X_i^{H,V}) - g({X'}_i^{H,V})||_2^2 \\ - ||g_{c_1}(X_i^{H,V}) - g(X_j^{H,V})||_2^2 + \alpha_1]_+ + \\ \beta_2.[||g_{c_2}(X_i^{L,N}) - g(X_i^{L,N})||_2^2 - \\ ||g_{c_2}(X_i^{L,N}) - g(X_k^{L,N})||_2^2 + \alpha_2]_+ \\ + \beta_3.[||g_{c_1}(X_i^{H,V}) - g_{c_2}(X_i^{L,N})||_2^2]_+ \quad (8) \\ \forall(X_i^{H,V}, X_i^{L,N}, X_j^{H,V}, X_k^{L,N}) \in \tau \end{aligned}$$

where, $g_{c_1}(X_i^{H,V})$ and $g_{c_2}(X_i^{L,N})$ are the centers of the subclasses pertaining to the homogeneous (good quality) and the heterogeneous (low resolution and NIR) images, respectively. $\beta_1, \beta_2, \beta_3$ are weights for each component. The first term in Equation 8 is similar to the first term of Equation 3, which brings the embedding of the good quality (homogeneous) images closer in the output embedding space of the model. The second term brings the embedding of the heterogeneous images closer thus making the subclass of the heterogeneous images (low resolution and NIR) more compact. The third term brings the centers of the two subclasses of the cluster closer to each other. The coefficients $\{\beta_1, \beta_2, \beta_3\}$ are used to adjust the strength of each component of the loss function. At the end of this cluster optimization phase, it is expected that all the images (heterogeneous and homogeneous) of each class must make a compact cluster, thereby enhancing heterogeneous matching performance of the trained model.

3.3 Heterogeneous Face Recognition using SHEAL

In order to train a heterogeneity aware model for face recognition, Equation 7 followed by Equation 8 is utilized. Once the model is trained, the test data is partitioned into probe and gallery according to the protocol of the testing database. For cross-spectral cross-resolution face recognition, the probes are NIR images of low resolution and the gallery images are high resolution visible spectrum images. A probe is given as input to the trained discriminative model to extract the embeddings, and the same is performed to generate the embeddings of the gallery images. The Euclidean distance is used to calculate match scores between the probe and gallery embeddings, which is finally used for face recognition.

3.4 Implementation Details

In this section, we outline the implementation details required to reproduce the results.

3.4.1 CNN Model

The proposed SHEAL is utilized to train a deep-CNN model for heterogeneous face recognition. The CNN model used is Light-CNN-29 [3] which is one of the popular models for face recognition with 29 convolutional layers and 4 pooling layers. After every convolutional layer a Max-Feature-Map operation is performed. The network is built using 6 blocks and each block contains convolution and Max-Feature-Map layers. The final layer is a Max-Feature-Map layer which gives an embedding of size 256.

3.4.2 Preparing Data for Training

In order to prepare training data for the SHEAL metric, each training sample is represented by an 8-tuple. Unlike existing approaches [49], [50], [52] we do not perform any hard mining on the set of 8-tuples, rather, we randomly prepare 500 8-tuples for every epoch. In order to prepare each 8-tuple, the images of a randomly selected subject/class (high resolution and visible spectrum) are used to calculate the center embedding. Other images for the 8-tuples are then chosen randomly from the training set of the database accordingly. We train only one epoch on each set of 500 8-tuples that are created in every iteration. We also performed experiments by running multiple epochs on each set of 8-tuples, but there is a tendency of the model to overfit on the samples that are generated. Thus, each epoch constitutes generating the set of 500 8-tuples and running one iteration of training on it. This keeps the pace of learning stable and effective.

3.4.3 Loss Function Parameters

The deep-CNN model is trained by back-propagating the gradient of the proposed SHEAL. The optimization is performed using Adam with a batch size of 20. The learning rate is initially kept at 10^{-3} which is then gradually decreased to 10^{-7}. The criteria for decreasing the learning rate was non-increment of validation accuracy for 20 epochs. The learning rate was decreased in steps of 0.05. The values of the margin variables α_1, α_2, α_3, α_4 are set differently for different databases during training. For the SCface database, we keep $\alpha_1 = 0.2$, $\alpha_2 = 0.4$, $\alpha_3 = 0.4$, and $\alpha_4 = 0.6$. For the FaceSurv database, we keep $\alpha_1 = 0.2$, $\alpha_2 = 0.4$, $\alpha_3 = 0.4$, and $\alpha_4 = 0.8$. For the CASIA NIR-VIS 2.0 database, we keep $\alpha_1 = 0.3$, $\alpha_2 = 0.4$, $\alpha_3 = 0.4$, and $\alpha_4 = 0.8$.

The parameters for the loss function coefficients $\lambda_1, \lambda_2, \lambda_3, \lambda_4$ for training are set as follows. For the SCface database, $\lambda_1 = 0.1$, $\lambda_2 = 0.2$, $\lambda_3 = 0.4$, and $\lambda_4 = 0.7$. For the FaceSurv database, $\lambda_1 = 0.4$, $\lambda_2 = 0.5$, $\lambda_3 = 0.5$, and $\lambda_4 = 0.9$. For the CASIA NIR-VIS 2.0 database, $\lambda_1 = 0.1$, $\lambda_2 = 0.4$, $\lambda_3 = 0.6$ and $\lambda_4 = 0.6$. Experiments are performed on a machine with Intel Core i7 CPU, with 32GB of RAM and NVIDIA GTX 1080Ti GPU with a PyTorch implementation.

3.4.4 Weights (β) for Subclass Cluster Optimization

The β parameters are used to assign weight of different components (Equation 8) in the subclass optimization step of SHEAL. Although these parameters are chosen empirically, we have followed a strategy while selecting the β parameters. As illustrated in Section 3.2, β_1 and β_2 are the weights of the subclasses for the visible spectrum high resolution images and the NIR low resolution images, respectively. On the other hand β_3 are weights for bringing the two subclasses closer into a single compact cluster. Since the later is the main motive of this step, β_3 is given a higher value than β_1 and β_2. Using this guideline and some empirical observations, the best β parameters are obtained for the subclass based cluster optimization.

4 Experiments and Analysis

To show the efficacy of the proposed approach, we have performed three different heterogeneous experiments on four challenging face

TABLE 1: Experimental details to evaluate the performance of the proposed SHEAL.

Experiment	Databases	Spectrum		Resolution	
		Gallery	Probe	Gallery	Probe
Cross-Resolution Face Recognition (CR-FR)	SCface	Visible	Visible	128 x 128	24 x 24, 32 x 32, 48 x 48
	FaceSurv	Visible	Visible	128 x 128	48 x 48, 64 x 64
	LFW	Visible	Visible	128 x 128	32 x 32, 48 x 48
Cross-Spectral Face Recognition (CS-FR)	SCface	Visible	NIR	128 x 128	128 x 128
	CASIA NIR-VIS 2.0	Visible	NIR	128 x 128	128 x 128
Cross-Spectral Cross-Resolution Face Recognition (CSCR-FR)	SCface	Visible	NIR	128 x 128	24 x 24, 32 x 32, 48 x 48
	FaceSurv	Visible	NIR	128 x 128	48 x 48, 64 x 64
	CASIA NIR-VIS 2.0	Visible	NIR	128 x 128	48 x 48, 64 x 64

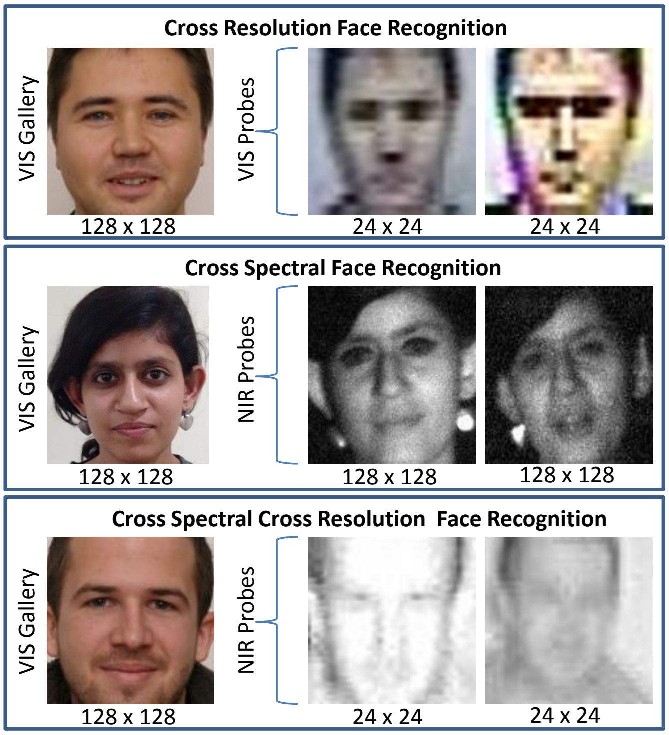


Fig. 4: Three different cases of heterogeneous face recognition considered in this work, including the most challenging case of cross-spectral cross-resolution matching. Images are taken from the SCface [10] and FaceSurv [11] databases.

databases. Very few papers in the literature have analyzed all three heterogeneous scenarios (a typical scenario of face recognition for video surveillance) using a single algorithm.

4.1 Databases and Protocol

As shown in Fig. 4 three experiments are performed, namely Cross-Resolution Face Recognition (CR-FR), Cross-Spectral Face Recognition (CS-FR), and Cross-Spectral Cross-Resolution Face Recognition (CSCR-FR). Details of experimental protocol are illustrated in Table 1. The details of the databases used for the experiments are as follows.

SCface Database [10] is one of the most popular face datasets that contains real world surveillance quality images. It contains 4160 images of 130 subjects captured using 8 surveillance cameras from three standoff distances namely 1 mt, 2.6 mts and 4.2 mts. The effective resolution of the face images detected from these surveillance images are 24×24, 32×32 and 48×48 for these three distances, respectively. Out of the 8 cameras, 5 operate in the visible spectrum and the remaining capture images in the NIR mode. The gallery images are captured using high resolution cameras and are sub-sampled to a resolution of 128×128. For CSCR-FR, NIR probe images pertaining to the three different resolutions have been matched with the high resolution visible spectrum gallery. For CR-FR, the same matching has been performed with low resolution visible spectrum probe images.

CASIA NIR-VIS 2.0 Database [12] is the largest publicly available dataset for CS-FR. It contains a total of 17,415 visible spectrum and NIR images pertaining to 725 subjects. The images in the training and testing sets are fixed and contain non-overlapping subjects. The database is divided into two views, namely view 1 and 2. The former is a development set and the later is for reporting the results. The gallery set contains one high resolution visible spectrum image for each subject. In order to train the deep CNN model using the proposed loss metric, we need low resolution visible and NIR images, in addition to the high resolution (128×128) visible and NIR images that are already present in the database. The images (both visible and NIR) are subsampled to a resolution of 32×32 to synthetically create low resolution versions of the same. To perform testing for CSCR-FR, the probe images are subsampled to a resolution of 48×48 and 64×64. For CS-FR, the usual protocol of the database (128×128 NIR probes) is utilized.

FaceSurv Database [11] contains videos captured under surveillance conditions in both day-time (in visible spectrum) and night-time (in NIR). The videos contain subjects walking at a standoff distance of 1-10 mts from the camera. The night-time videos have been captured in a completely dark environment using NIR illumination, while the day-time videos have been captured in outdoor settings. Both day-time and night-time videos are captured under uncontrolled illumination, pose and expression variations. The gallery images contain three high resolution (subsampled to 128×128) visible spectrum images for every subject. Images pertaining to 30 subjects are used for training and images of the remaining subjects are used for testing. In terms of the number of images, the training and testing sets have 13,617 and 109,131 video frames (of non-overlapping subjects), respectively. In order to perform CSCR-FR, night-time (NIR) probe videos have been divided into two subsets, video frames that are captured at a distance to 5-10 mts (48×48 resolution) and frames that are captured at a distance to 1-5 mts (64×64 resolution) from the camera. For CR-FR, the same matching has been performed with day-time video frames.

Labeled Faces in the Wild (LFW) Database [13] contains 13,233 images of 5,749 subjects, out of which 1,680 subjects have

TABLE 2: Rank 1 identification accuracies on the SCface database.

Algorithm	**Cross-Spectral Cross-Resolution**			**Cross-Resolution**		
	24 x 24	**32 x 32**	**48 x 48**	**24 x 24**	**32 x 32**	**48 x 48**
Biswas *et al.* (2013) (Multi-Dimensional Scaling) [18]	-	-	-	64.8	70.4	76.1
Bhatt *et al.* (2014) (Co-Transfer Learning) [20]	-	-	-	70.1	76.2	83.4
Wu *et al.* (2015) (LightCNN29) [3]	8.4	23.7	69.0	33.1	85.5	97.8
COTS (2016) (FaceVacs) [53]	1.7	2.9	6.5	10.3	18.5	35.7
Ghosh *et al.* (2016) (Autoencoder+SIFT) [53]	-	37.0	53.8	-	-	-
Schroff *et al.* (2015) (Triplet loss) [49]	11.1	37.9	67.8	35.3	87.5	97.4
Chen *et al.* (2017) (Quadruplet Loss) [50]	10.6	25.6	70.7	33.0	86.0	97.7
Hermans *et al.* (2017) (Hard Triplet Loss) [54]	11.2	28.9	71.7	35.6	87.7	97.2
He *et al.* (2018) (Triplet Center Loss) [55]	14.8	29.4	72.0	34.6	89.1	97.9
Yang *et al.* (2018) (DMDS) [33]	-	-	-	61.5	67.2	62.9
Yang *et al.* (2018) (LDMDS) [33]	-	-	-	62.7	70.7	65.5
Talreja *et al.* (2019) [56]	-	-	-	44.8	49.6	54.3
Li *et al.* (2019) [35]	-	-	-	20.4	20.8	31.7
Proposed SHEAL	**43.9**	**73.0**	**87.6**	**72.8**	**97.6**	**99.1**

TABLE 3: Rank 1 identification accuracies on the CASIA NIR-VIS 2.0 database.

Algorithm	**Cross-Spectral Cross-Resolution**		**Cross-Spectral**
	48 x 48	**64 x 64**	**128 x 128**
Wu *et al.* (2015) (LightCNN29) [3]	62.9	77.4	79.1
Schroff *et al.* (2015) (Triplet loss) [49]	67.3	81.2	82.5
Liu *et al.* (2016) (Transferable Triplet Loss) [51]	-	-	95.7
Lezama *et al.* (2017) (Face Hallucination) [25]	-	-	96.4
He *et al.* (2017) (Invariant Deep Representation) [26]	-	-	97.3
Chen *et al.* (2017) (Quadruplet Loss) [50]	68.5	81.7	83.1
Hermans *et al.* (2017) (Hard Triplet Loss) [54]	70.4	83.8	86.0
Lu *et al.* (2018) (C-SLBFLE) [57]	-	-	86.9
Huo *et al.* (2018) (K-MCMML) [2]	-	-	96.5
Proposed SHEAL	**93.8**	**97.5**	**97.6**

more than 2 images. The database is divided into views 1 and 2, where view 1 is the development set. View 2, which is the set on which results are reported has 10 folds, each of which contains 300 genuine and 300 impostor pairs. In order to perform cross-resolution face recognition experiments, low resolution images (32×32 and 48×48) are synthetically prepared (similar to the experiment on the CASIA NIR-VIS 2.0 database) for both training and testing.

4.2 Experimental Results and Analysis

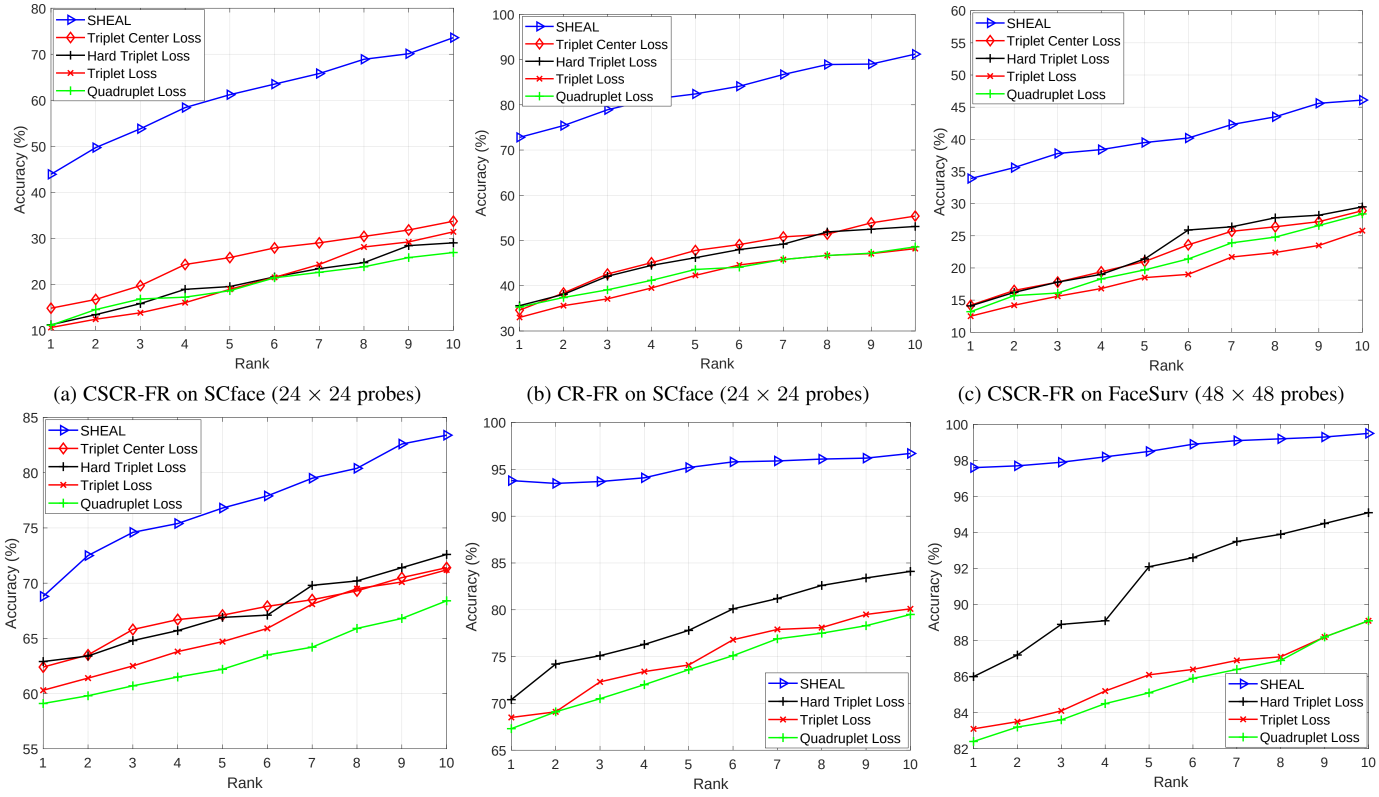
The proposed method is evaluated on four datasets and the results are outlined in Tables 2, 3, 4, and 5[3] and Figures 5 to 8. The experiments are performed to analyze the accuracies along with convergence analysis, ablation study, and visual inspection of results. The results are also compared with without pretraining and comparison with recent state-of-the-art algorithms.

4.2.1 Comparison with State-of-the-Art Methods

For the SCface [10], CASIA NIR-VIS 2.0 [12] and FaceSurv [11] databases, extensive comparisons have been performed with recent deep metric learning methods and state-of-the-art heterogeneous face recognition methods. For the SCface [10] database, CS-FR, CR-FR, and CSCR-FR experiments are performed on three resolution variations of the probes, 24×24, 32×32, and 48×48. As shown in Table 2, the proposed method achieves state-of-the-art results and outperforms popular deep metric learning and recent heterogeneous face recognition methods on all the resolutions. It can be observed that SHEAL yields larger improvement with low resolution probe images. As shown in Table 3, on the CASIA NIR-VIS 2.0 [12] database as well, the proposed SHEAL metric outperforms popular deep metric learning and recent cross-spectral face recognition methods on both CS-FR and CSCR-FR. On the FaceSurv [11] database we have performed CSCR-FR and CR-FR on two different probe resolutions, namely 48×48 and 64×64. As shown in Table 4, the proposed algorithm outperforms both Triplet [49] and Quadruplet loss [50] based methods (along with their variants), and achieves state-of-the-art results on both CSCR-FR and CR-FR experiments. As outlined in Tables 2, 3, and 4, the proposed SHEAL is among the top performing algorithms on all the probe resolutions. It should be noted that for lower resolutions, such as 24×24 in Table 2, the accuracy of SHEAL is much higher compared to existing algorithms. The CMC curves showcasing the identification accuracies are shown in Fig. 5. In addition, **homogeneous face recognition** experiment is performed on the SCface database using the same model (that is trained using the SHEAL metric). The proposed model achieves rank 1 accuracy of 97.29% on CR-FR for 32×32 probes on the SCFace database.

Finally, on the LFW face database [13], CR-FR experiment is performed and the results are documented in Table 5. The results show that with gallery images of size 128×128 and probe images of 32×32 or 48×48, the proposed algorithm is at least 1.5% better than other deep metric learning algorithms. On the LFW database, comparisons have been performed with popular deep metric learning methods, and the proposed method outperform them for CR-FR scenario on this database on two different probe resolutions. Note that due to image size variations to conduct CR-

[3] For CR-FR or CS-FR, wherever applicable, published results are reported. On the other hand, for CSCR-FR, we have performed the comparisons with existing algorithms using publicly available codes.



(a) CSCR-FR on SCface (24 × 24 probes)

(b) CR-FR on SCface (24 × 24 probes)

(c) CSCR-FR on FaceSurv (48 × 48 probes)

(d) CR-FR on FaceSurv (48 × 48 probes)

(e) CSCR-FR on CASIA (48 × 48 probes)

(f) CS-FR on CASIA (128 × 128 probes)

Fig. 5: CMC curves for Cross-Resolution Face Recognition (CR-FR), Cross-Spectral Face Recognition (CS-FR) and Cross-Spectral Cross-Resolution Face Recognition (CSCR-FR) on the SCface [10], FaceSurv [11] and CASIA NIR-VIS 2.0 [12] databases.

TABLE 4: Rank 1 identification accuracies on the FaceSurv database.

Algorithm	**Cross-Spectral Cross-Resolution**		**Cross-Resolution**	
	48 x 48	**64 x 64**	**48 x 48**	**64 x 64**
Wu *et al.* (2015) (LightCNN29) [3]	14.0	62.3	62.6	90.4
Schroff *et al.* (2015) (Triplet loss) [49]	13.2	62.5	59.1	90.1
Chen *et al.* (2017) (Quadruplet Loss) [50]	12.5	59.0	60.3	90.2
Hermans *et al.* (2017) (Hard Triplet Loss) [54]	14.1	61.7	62.9	90.0
He *et al.* (2018) (Triplet Center loss) [55]	14.2	59.8	62.4	90.5
Proposed SHEAL	**33.9**	**74.8**	**68.8**	**90.7**

TABLE 5: Verification accuracies at 1% False accept rate (FAR) for cross-resolution face recognition on the LFW database, with unrestricted no-outside labeled data protocol.

Algorithm	**Cross-Resolution**	
	32 x 32	**48 x 48**
Schroff *et al.* (2015) (Triplet Loss) [49]	58.2	87.6
Chen *et al.* (2017) (Quadruplet Loss) [50]	60.5	91.1
Hermans *et al.* (2017) (Hard Triplet Loss) [54]	62.9	92.4
He *et al.* (2018) (Triplet Center Loss) [55]	61.7	90.3
Proposed SHEAL	**64.4**	**94.2**

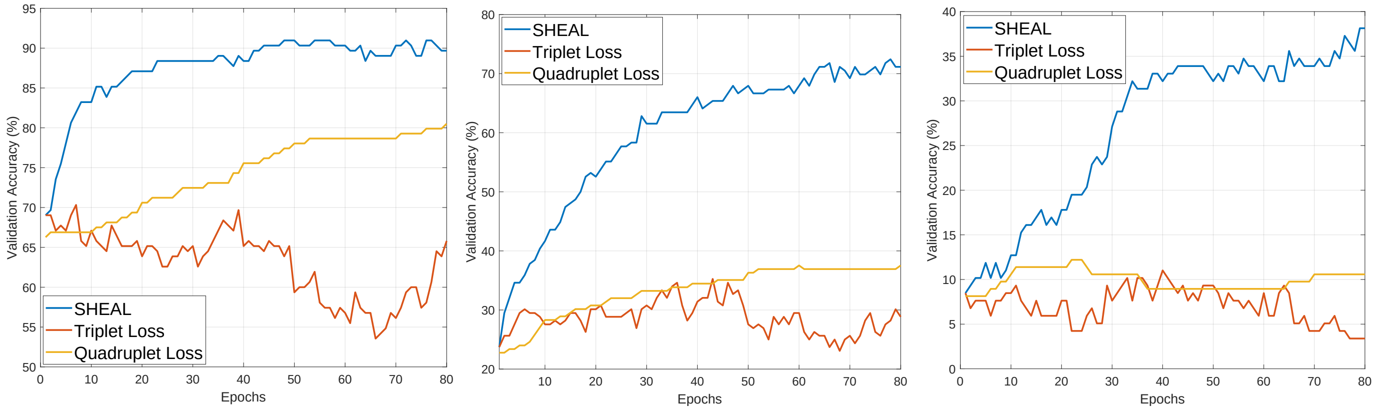
FR experiments, we cannot directly compare with reported results on the LFW database.

4.2.2 Convergence Analysis

Figure 6 shows the rate of convergence of SHEAL, triplet loss, and quadruplet loss on the SCface dataset. It can be observed that the convergence of SHEAL is significantly fast and effective. The validation accuracy of the model trained using SHEAL reaches to 83.23% from 69.03% (on 48 × 48 probes) in just 10 epochs (Fig. 6(a)). Compared to SHEAL, the quadruplet and triplet losses converge slowly. Figures 7(a) and (b) show the time taken and the number of epochs required to converge, respectively. The number of epochs required by SHEAL to converge is only 48 compared to 95 and 118 epochs required by triplet and quadruplet loss for the same. In terms of total time, SHEAL takes 115.3 seconds against 158.2, 171.3, 140.4 and 122.3 seconds required by triplet loss, quadruplet loss, hard triplet loss and triplet center loss respectively for convergence. These results suggest that the proposed SHEAL converges rapidly, takes lesser time, and exhibits significantly higher face recognition accuracies in heterogeneous settings.

4.2.3 Ablation Study

We have performed two separate ablation studies for a thorough understanding of the effect of the loss functions (Equations 7 and 8) on the trained model's performance on CR-FR and CSCR-FR scenarios. As illustrated in Section 3.1, Equation 7 is composed of four separate terms: L^c_{Ho}, L^c_{CR}, L^c_{CS} and L^c_{CS-CR}. We have performed an ablation study on Equation 7, where we have utilized these specific terms for training the models separately.

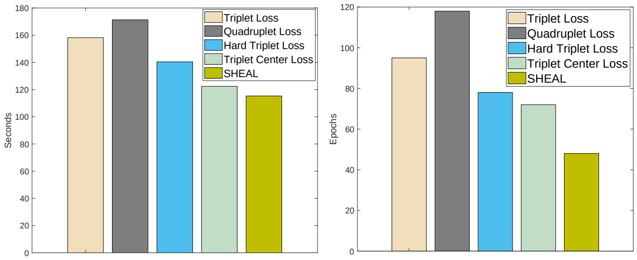


(a) 48 × 48

(b) 32 × 32

(c) 24 × 24

Fig. 6: Convergence analysis of the proposed method on different probe resolutions of the SCface database [10]. It can be observed that the convergence of the proposed method is significantly better than the triplet [49] and the quadruplet loss [50] methods.



(a) Training time

(b) Epochs for convergence

Fig. 7: Performance analysis of the proposed method: (a) Time taken to converge (training) and (b) Number of epochs for convergence. It can be observed that the proposed algorithm not only converges rapidly, but also takes much lesser time and epochs for the same. Training is performed on the SCface database.

Each of these terms have a disjoint effect on the trained model which is evident in the results obtained on the SCface database (Table 6). The model, when trained only with the L^c_{Ho} loss term yields the worst performance. However, when L^c_{CR} and L^c_{CS} loss terms are used separately for training, the corresponding testing performance (eg. when L^c_{CR} term is used for training the cross-resolution performance is improved during testing) is improved. It can be observed that when training is performed with L^c_{CS-CR}, the results, during testing, are improved considerably for both CS-CR and CR face recognition.

In addition to the above, we have also performed an ablation study on Equation 8 (subclass based cluster optimization). As illustrated in Section 3.2, Equation 8 is composed of three terms. The first term is a homogeneous matching term, the second term makes subclass containing the low resolution and NIR images more compact, and the third terms brings the subclasses closer into one compact cluster. As shown in Table 6, we observe that the third term is the major contributing factor in the subclass optimization stage. The value of β_3 is also kept higher during this optimization stage, to give more weight to the third term of Equation 8.

4.2.4 Loss Function Coefficients

For training using SHEAL and the cluster optimization phase (Equations 7 and 8), we have the loss function coefficients (λ_1, λ_2, λ_3 and λ_4), which can be used to adjust the weight of each component of the loss function. Since homogeneous matching is a less challenging problem compared to CSCR-FR, CS-FR and CR-FR, we kept a considerably lower value for λ_1 than the other weight terms. For the SCface [10] database, we kept a much higher value for λ_4 since CSCR-FR matching is an extremely challenging problem.

TABLE 6: Rank 1 identification accuracies (%) for the ablation study on Equations 7 and 8 performed on the SCface database.

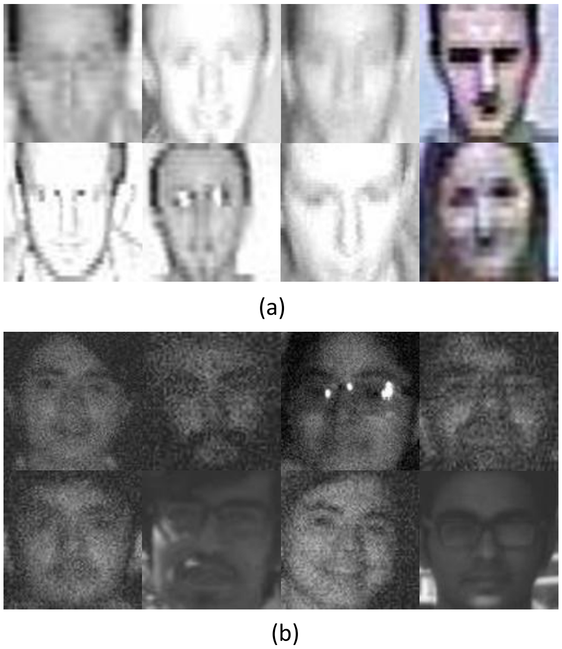
Loss Term		CS-CR			CR		
		24 x 24	**32 x 32**	**48 x 48**	**24 x 24**	**32 x 32**	**48 x 48**
Eq. 7	L^c_{Ho}	9.5	29.4	65.1	37.0	58.9	63.4
	L^c_{CR}	18.6	39.4	80.4	70.9	96.4	98.7
	L^c_{CR}	12.5	31.2	69.0	40.5	51.2	64.3
	L^c_{CS-CR}	41.3	72.6	85.9	68.7	94.3	99.0
	$L^c_{Ho} + L^c_{CR} + L^c_{CR}$	38.4	68.4	83.9	64.2	92.6	98.2
Eq. 8	$1^{st} term + 2^{nd} term$	37.2	67.1	78.9	64.3	92.1	85.4
	$1^{st} term + 3^{rd} term$	42.1	72.5	87.3	70.2	96.4	98.4
	Proposed	**43.9**	**73.0**	**87.6**	**72.8**	**97.6**	**99.1**

4.2.5 Without Pretraining

In order to make a fair comparison, the other deep metric learning methods with which we have compared in the paper ([50], [49] and their variants) have been trained on the same data using the weights of the same pretrained model. In addition to this, we trained our method from scratch (on a randomly initialized model), and achieved 90.71% accuracy wheras those obtained by Chen *et al.* [50] and Schroff *et al.* [49] are 82.13% and 80.34% respectively, on 64 × 64 probes of the CASIA NIR-VIS 2.0 database. It shows that even without pretraining, the proposed method outperforms the most popular deep metric learning algorithms.

4.2.6 Visual Inspection of the Results

We also performed visual inspection of the results and some cases are presented in Fig. 8. It can be observed that images of the SCface database which have extremely low resolution and quality (Fig. 8(a)) are correctly classified by the proposed algorithm. On the other hand, the images in the FaceSurv database, which in addition to low resolution suffer from heavy motion blur and poor illumination, are also correctly classified by the proposed algorithm. These results showcase the potential applicability of the proposed algorithm to real world surveillance scenarios.



(a)

(b)

Fig. 8: Sample images of some extremely noisy and poor quality images of (a) SCface [10] and (b) FaceSurv [11] databases that are correctly classified by the model trained with SHEAL, but were incorrectly classified using triplet [49], quadruplet [50] and triplet center loss [55] based methods.

5 Conclusion

The problem of heterogeneous face recognition is compounded when test data shows multiple heterogeneity. Current deep metric learning approaches generally do not handle such heterogeneous problems and yield poor recognition accuracies. This paper introduces a subclass heterogeneity aware loss function which is utilized to train a discriminative model to generate heterogeneity invariant embeddings. This helps to project a pair of face images of different covariates into an embedding space where matching can be performed efficiently irrespective of the images being captured in very different scenarios. This can be applied in a surveillance application where the data may encompass multiple covariates. In future, we plan to extend the proposed algorithm to include other covariates of face recognition such as disguise and aging along with multiple heterogeneous variations.

6 Acknowledgements

This research was partly supported by MEITY, Government of India. M. Vatsa is also supported through the Swarnajayanti Fellowship by the Government of India. S. Ghosh is supported by TCS Research Fellowship.

References

[1] Z. Lu, X. Jiang, and A. Kot, “Deep coupled resnet for low-resolution face recognition,” *IEEE SPL*, vol. 25, no. 4, pp. 526–530, 2018.

[2] J. Huo, Y. Gao, Y. Shi, W. Yang, and H. Yin, “Heterogeneous face recognition by margin-based cross-modality metric learning,” *IEEE TC*, vol. 48, no. 6, pp. 1814–1826, 2018.

[3] X. Wu, R. He, Z. Sun, and T. Tan, “A light CNN for deep face representation with noisy labels,” *arXiv preprint arXiv:1511.02683*, 2015.

[4] O. M. Parkhi, A. Vedaldi, and A. Zisserman, “Deep face recognition.” in *BMVC*, vol. 1, 2015, pp. 6–19.

[5] S. OBeirne, “What cctv users need to know about gdpr,” 2017, uRL: http://www.fmj.co.uk/cctv-users-need-know-gdpr/.

[6] O. S. Shop, “Do facial recognition cameras in public places infringe on our privacy?” 2017, uRL: https://tinyurl.com/y7lju5c5.

[7] J.-Y. Zhu, W.-S. Zheng, J.-H. Lai, and S. Z. Li, “Matching NIR face to VIS face using transduction,” *IEEE TIFS*, vol. 9, no. 3, pp. 501–514, 2014.

[8] S. Nagpal, M. Singh, R. Singh, M. Vatsa, A. Noore, and A. Majumdar, “Face sketch matching via coupled deep transform learning,” in *IEEE ICCV*, 2017, pp. 5429–5438.

[9] T. I. Dhamecha, P. Sharma, R. Singh, and M. Vatsa, “On effectiveness of histogram of oriented gradient features for visible to near infrared face matching,” in *IAPR ICPR*, 2014, pp. 1788–1793.

[10] M. Grgic, K. Delac, and S. Grgic, “SCface–surveillance cameras face database,” *Springer MTA*, vol. 51, no. 3, pp. 863–879, 2011.

[11] S. Gupta, N. Gupta, S. Ghosh, M. Singh, S. Nagpal, R. Singh, and M. Vatsa, “A benchmark video dataset for face detection and recognition across spectra and resolutions,” in *IEEE FG*, 2019.

[12] S. Z. Li, D. Yi, Z. Lei, and S. Liao, “The CASIA NIR-VIS 2.0 face database,” in *IEEE CVPRW*, 2013, pp. 348–353.

[13] E. Learned-Miller, G. B. Huang, A. RoyChowdhury, H. Li, and G. Hua, “Labeled faces in the wild: A survey,” in *Springer AFDFIA*, 2016, pp. 189–248.

[14] M. Kan, S. Shan, H. Zhang, S. Lao, and X. Chen, “Multi-view discriminant analysis,” *IEEE TPAMI*, vol. 38, no. 1, pp. 188–194, 2016.

[15] S. P. Mudunuri, S. Venkataramanan, and S. Biswas, “Dictionary alignment with re-ranking for low-resolution NIR-VIS face recognition,” *IEEE TIFS*, vol. 14, no. 4, pp. 886–896, 2018.

[16] F. Juefei-Xu, D. K. Pal, and M. Savvides, “NIR-VIS heterogeneous face recognition via cross-spectral joint dictionary learning and reconstruction,” in *IEEE CVPRW*, 2015, pp. 141–150.

[17] Z. Lei and S. Z. Li, “Coupled spectral regression for matching heterogeneous faces,” in *IEEE CVPR*, 2009, pp. 1123–1128.

[18] S. Biswas, G. Aggarwal, P. J. Flynn, and K. W. Bowyer, “Pose-robust recognition of low-resolution face images,” *IEEE TPAMI*, vol. 35, no. 12, pp. 3037–3049, 2013.

[19] S. P. Mudunuri and S. Biswas, “Low resolution face recognition across variations in pose and illumination,” *IEEE TPAMI*, vol. 38, no. 5, pp. 1034–1040, 2016.

[20] H. S. Bhatt, R. Singh, M. Vatsa, and N. K. Ratha, “Improving cross-resolution face matching using ensemble-based co-transfer learning,” *IEEE TIP*, vol. 23, no. 12, pp. 5654–5669, 2014.

[21] S. P. Mudunuri and S. Biswas, “A coupled discriminative dictionary and transformation learning approach with applications to cross domain matching,” *Elsevier PRL*, vol. 71, pp. 38–44, 2016.

[22] J. Lu, V. E. Liong, X. Zhou, and J. Zhou, “Learning compact binary face descriptor for face recognition,” *IEEE TPAMI*, vol. 37, no. 10, pp. 2041–2056, 2015.

[23] D. Yi, Z. Lei, and S. Z. Li, “Shared representation learning for heterogenous face recognition,” in *IEEE FG*, vol. 1, 2015, pp. 1–7.

[24] S. Saxena and J. Verbeek, “Heterogeneous face recognition with CNNs,” in *Springer ECCV*, 2016, pp. 483–491.

[25] J. Lezama, Q. Qiu, and G. Sapiro, “Not afraid of the dark: NIR-VIS face recognition via cross-spectral hallucination and low-rank embedding,” in *IEEE CVPR*, 2017, pp. 6807–6816.

[26] R. He, X. Wu, Z. Sun, and T. Tan, “Learning invariant deep representation for NIR-VIS face recognition.” in *AAAI*, vol. 4, 2017, pp. 7–16.

[27] X. Wu, L. Song, R. He, and T. Tan, “Coupled deep learning for heterogeneous face recognition,” in *AAAI*, 2018, pp. 1679–1686.

[28] L. Song, M. Zhang, X. Wu, and R. He, “Adversarial discriminative heterogeneous face recognition,” in *AAAI*, 2018, pp. 7355–7362.

[29] T. de Freitas Pereira, A. Anjos, and S. Marcel, “Heterogeneous face recognition using domain specific units,” *IEEE TIFS*, vol. 14, no. 7, pp. 1803–1816, 2018.

[30] C. Peng, N. Wang, J. Li, and X. Gao, “Re-ranking high-dimensional deep local representation for NIR-VIS face recognition,” *IEEE TIP*, vol. 28, no. 9, pp. 4553–4565, 2019.

[31] R. He, J. Cao, L. Song, Z. Sun, and T. Tan, “Cross-spectral face completion for NIR-VIS heterogeneous face recognition,” *arXiv preprint arXiv:1902.03565*, 2019.

[32] M. Singh, S. Nagpal, M. Vatsa, R. Singh, and A. Majumdar, “Identity aware synthesis for cross resolution face recognition,” in *IEEE CVPR Workshops*, 2018, pp. 479–488.

[33] F. Yang, W. Yang, R. Gao, and Q. Liao, “Discriminative multidimensional scaling for low-resolution face recognition,” *IEEE SPL*, vol. 25, no. 3, pp. 388–392, 2018.

[34] S. Ge, S. Zhao, C. Li, and J. Li, "Low-resolution face recognition in the wild via selective knowledge distillation," *IEEE TIP*, vol. 28, no. 4, pp. 2051–2062, 2018.

[35] P. Li, L. Prieto, D. Mery, and P. J. Flynn, "On low-resolution face recognition in the wild: Comparisons and new techniques," *IEEE TIFS*, vol. 14, no. 8, pp. 2000–2012, 2019.

[36] E. Zangeneh, M. Rahmati, and Y. Mohsenzadeh, "Low resolution face recognition using a two-branch deep convolutional neural network architecture," *Expert Systems with Applications*, pp. 112–121, 2019.

[37] O. Abdollahi Aghdam, B. Bozorgtabar, H. Kemal Ekenel, and J.-P. Thiran, "Exploring factors for improving low resolution face recognition," in *IEEE CVPRW*, 2019, pp. 1158–1166.

[38] M. Singh, S. Nagpal, R. Singh, and M. Vatsa, "Dual directed capsule network for very low resolution image recognition," in *IEEE ICCV*, 2019, pp. 340–349.

[39] S. Chopra, R. Hadsell, and Y. LeCun, "Learning a similarity metric discriminatively, with application to face verification," in *IEEE CVPR*, vol. 1, 2005, pp. 539–546.

[40] Y. Duan, J. Lu, J. Feng, and J. Zhou, "Deep localized metric learning," *IEEE TCSVT*, pp. 1212–1222, 2017.

[41] J. Hu, J. Lu, and Y.-P. Tan, "Discriminative deep metric learning for face verification in the wild," in *IEEE CVPR*, 2014, pp. 1875–1882.

[42] J. Hu, J. Lu, Y.-P. Tan, J. Yuan, and J. Zhou, "Local large-margin multi-metric learning for face and kinship verification," *IEEE TCSVT*, pp. 1254–1265, 2017.

[43] J. Lu, J. Hu, and Y.-P. Tan, "Discriminative deep metric learning for face and kinship verification," *IEEE TIP*, vol. 26, no. 9, pp. 4269–4282, 2017.

[44] Y. Wen, K. Zhang, Z. Li, and Y. Qiao, "A discriminative feature learning approach for deep face recognition," in *Springer ECCV*, 2016, pp. 499–515.

[45] X. Zhang, Z. Fang, Y. Wen, Z. Li, and Y. Qiao, "Range loss for deep face recognition with long-tailed training data," in *IEEE CVPR*, 2017, pp. 5409–5418.

[46] W. Liu, Y. Wen, Z. Yu, M. Li, B. Raj, and L. Song, "Sphereface: Deep hypersphere embedding for face recognition," in *IEEE CVPR*, 2017, pp. 212–220.

[47] Y. Zheng, D. K. Pal, and M. Savvides, "Ring loss: Convex feature normalization for face recognition," in *IEEE CVPR*, 2018, pp. 5089–5097.

[48] H. Wang, Y. Wang, Z. Zhou, X. Ji, D. Gong, J. Zhou, Z. Li, and W. Liu, "Cosface: Large margin cosine loss for deep face recognition," in *IEEE CVPR*, 2018, pp. 5265–5274.

[49] F. Schroff, D. Kalenichenko, and J. Philbin, "FaceNet: A unified embedding for face recognition and clustering," in *IEEE CVPR*, 2015, pp. 815–823.

[50] W. Chen, X. Chen, J. Zhang, and K. Huang, "Beyond triplet loss: a deep quadruplet network for person re-identification," in *IEEE CVPR*, 2017, pp. 403–412.

[51] X. Liu, L. Song, X. Wu, and T. Tan, "Transferring deep representation for NIR-VIS heterogeneous face recognition," in *IAPR ICB*, 2016.

[52] D. Cheng, Y. Gong, S. Zhou, J. Wang, and N. Zheng, "Person re-identification by multi-channel parts-based CNN with improved triplet loss function," in *IEEE CVPR*, 2016, pp. 1335–1344.

[53] S. Ghosh, R. Keshari, R. Singh, and M. Vatsa, "Face identification from low resolution near-infrared images," in *IEEE ICIP*, 2016, pp. 938–942.

[54] A. Hermans, L. Beyer, and B. Leibe, "In defense of the triplet loss for person re-identification," *arXiv preprint arXiv:1703.07737*, 2017.

[55] X. He, Y. Zhou, Z. Zhou, S. Bai, and X. Bai, "Triplet-center loss for multi-view 3D object retrieval," in *IEEE CVPR*, 2018, pp. 1945–1954.

[56] V. Talreja, F. Taherkhani, M. C. Valenti, and N. M. Nasrabadi, "Attribute-guided coupled gan for cross-resolution face recognition," *arXiv preprint arXiv:1908.01790*, 2019.

[57] J. Lu, V. E. Liong, and J. Zhou, "Simultaneous local binary feature learning and encoding for homogeneous and heterogeneous face recognition," *IEEE TPAMI*, vol. 40, no. 8, pp. 1979–1993, 2018.

Soumyadeep Ghosh received B.Sc (Hons), and M.Sc degrees in Computer Science from the University of Calcutta, India and M.Tech degree in Computer Science and Engineering From IIIT Bhubaneswar India. He is pursuing Ph.D. from IIIT-Delhi, India where he works with the Image Analysis and Biometrics Lab and the Infosys Center for Artificial Intelligence. He was awarded the prestigious TCS Research Fellowship in 2016 and received the Best Poster Award at IEEE BTAS 2016. His area of interests are Deep Learning, Surveillance Biometrics and Computer Vision.

Richa Singh Richa Singh received the Ph.D. degree in computer science from West Virginia University, Morgantown, USA, in 2008. She is currently a Professor at IIT-Jodhpur, India, and an Adjunct Professor with IIIT-Delhi and West Virginia University, USA. She has co-edited book Deep Learning in Biometrics and has delivered tutorials on deep learning and domain adaptation in ICCV 2017, AFGR 2017, and IJCNN 2017. Her areas of interest are pattern recognition, machine learning, and biometrics. She is a fellow of IAPR and a Senior Member of IEEE and ACM. She was a recipient of the Kusum and Mohandas Pai Faculty Research Fellowship at the IIIT-Delhi, the FAST Award by the Department of Science and Technology, India, and several best paper and best poster awards in international conferences. She has also served as the Program Co-Chair of AFGR2019 and BTAS 2016, and a General Co-Chair of ISBA 2017. She is currently serving as a Program Co-Chair of IJCB 2020. She is also the Vice President (Publications) of the IEEE Biometrics Council. She is an Associate Editor-in-Chief of Pattern Recognition, and Area/Associate Editor of several journals.

Mayank Vatsa Mayank Vatsa received the M.S. and Ph.D. degrees in computer science from West Virginia University, USA, in 2005 and 2008, respectively. He is currently a Professor at IIT-Jodhpur, India, and an Adjunct Professor with IIIT-Delhi and West Virginia University, USA. He has co-edited a book Deep learning in Biometrics and co-authored over 250 research papers. His areas of interest are biometrics, image processing, machine learning, computer vision, and information fusion. He is the recipient of the prestigious Swarnajayanti fellowship award from Government of India, A. R. Krishnaswamy Faculty Research Fellowship at the IIIT-Delhi, the FAST Award Project by DST, India, and several Best Paper and Best Poster Awards at international conferences. He is an Area Chair of Information Fusion and Pattern Recognition Journals, General Co-Chair of IJCB 2020, and the PC Co-Chair of the ICB 2013 and IJCB 2014. He has served as the Vice President (Publications) of the IEEE Biometrics Council where he started the IEEE Transactions on Biometrics, Behavior, And Identity Science.