Quality-Aware Multimodal Biometric Recognition

Sobhan Soleymani, Ali Dabouei, Fariborz Taherkhani, Seyed Mehdi Iranmanesh,
Jeremy Dawson and Nasser M. Nasrabadi, *Fellow, IEEE*
Lane Department of Computer Science and Electrical Engineering
West Virginia University

Abstract—We present a quality-aware multimodal recognition framework that combines representations from multiple biometric traits with varying quality and number of samples to achieve increased recognition accuracy by extracting complimentary identification information based on the quality of the samples. We develop a quality-aware framework for fusing representations of input modalities by weighting their importance using quality scores estimated in a weakly-supervised fashion. This framework utilizes two fusion blocks, each represented by a set of quality-aware and aggregation networks. In addition to architecture modifications, we propose two task-specific loss functions: multimodal separability loss and multimodal compactness loss. The first loss assures that the representations of modalities for a class have comparable magnitudes to provide a better quality estimation, while the multimodal representations of different classes are distributed to achieve maximum discrimination in the embedding space. The second loss, which is considered to regularize the network weights, improves the generalization performance by regularizing the framework. We evaluate the performance by considering three multimodal datasets consisting of face, iris, and fingerprint modalities. The efficacy of the framework is demonstrated through comparison with the state-of-the-art algorithms. In particular, our framework outperforms the rank- and score-level fusion of modalities of BIOMDATA [1] by more than 30% for true acceptance rate at false acceptance rate of 10^{-4}.

Index Terms—Multimodal biometrics, quality-aware multimodal fusion, representation separability, network compactness.

1 Introduction

Biometrics research explores the possibility of automatically recognizing individuals based on their unique physical or behavioral traits such as face, fingerprint, voice, iris, or handwriting, which are referred to as biometric modalities. The uniqueness of the biometric features extracted from these traits, have allowed unimodal biometric systems to be widely used for identification and verification applications in a wide variety of scenarios [2], [3], [4], [4], [5]. Beyond unimodal systems, a major merit of multimodal biometric recognition is its robustness to noisy data, non-universality, and category-based variations. Indeed, fusing multiple instances of biometric information lowers recognition error rates for low-quality and unreliable biometric samples, such as latent fingerprints, face images captured at a distance, and low-resolution iris images [6], [7], [8]. Hence, we argue that the multimodal framework should be able to leverage the quality information of the input samples to incorporate all identification information within these samples while discarding the distorted information in low-quality samples that may negatively affect the identification.

The most commonly deployed feature fusion methods for multimodal frameworks presented in the literature are feature concatenation [9], [10], bilinear multiplication [11], [12], and compact bilinear pooling [13], [14], [15]. However, these methods treat all samples equally, and do not take their reliability and usefulness into account. A multimodal recognition algorithm requires selecting the discriminative and informative features from each sample, as well as exploring the dependencies between features extracted from different samples in a multimodal sample set. This framework should also determine the priority of features according to their usefulness and reliability for the recognition task. In addition, the information provided by different samples in a multimodal biometric sample set may or may not be independent. For instance, consider the case where the multimodal biometric recognition framework has access to videos captured by a surveillance camera and fingerprints collected using a sensor. In this situation, face images captured with pose variations are correlated, while the face images and fingerprint samples are independent. Therefore, compared to other recognition frameworks, information fusion for multimodal biometric systems has remained a challenging task.

A biometric sample is of good quality if it is suitable for automated matching. This quality can be quantified as a measure of how properly the biometric sample can be processed within the matching algorithm, including feature extraction and accurate recognition with a high confidence score [6]. For a multimodal sample set consisting of different modalities and a varying number of samples from each modality, the recognition framework should investigate both intra-modality and inter-modality information fusion. The intra-modality and inter-modality usefulness can be interpreted as the intra-modality quality of the samples and the inter-modality quality of different modalities, respectively.

Recent multi-sample recognition frameworks [16], [17], [18], [19] aim to solve this problem when all samples in the multimodal sample set are of the same modality. However, multimodal multi-sample recognition requires the consideration of independent samples in the set, where samples represent different modalities. Our proposed framework seeks to formalize a learning framework that automatically identifies the usefulness of the samples in a multimodal sample set through the loss defined by the underlying recognition task, where this usefulness is due to the intra-modality and inter-modality quality of the samples. This quality-aware framework aims to improve the representation of a multimodal sample set in the embedding space by estimating the quality of each of its samples in a weakly-supervised fashion.

As presented in Fig. 1, our framework employs two weakly-supervised quality-aware fusion blocks. This framework represents each sample in the multimodal sample set with a feature

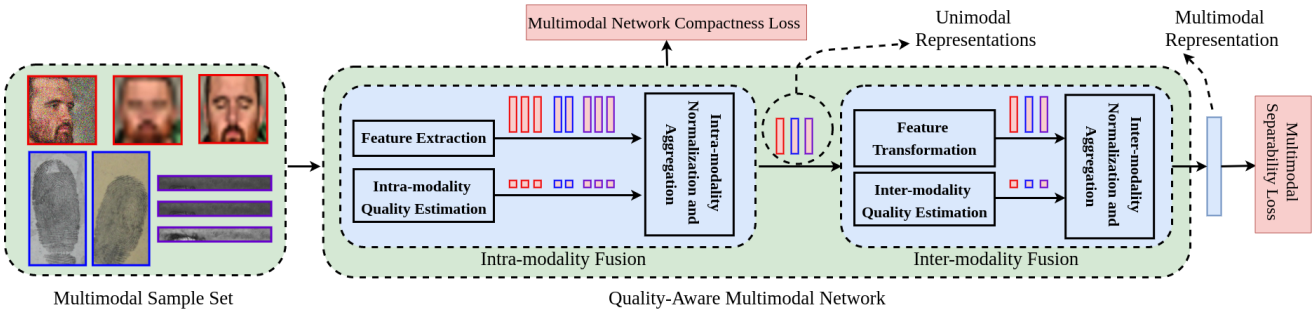

Fig. 1: A multimodal biometric sample set consists of samples from different modalities and varying quality. The quality-aware multimodal network, which consists of two fusion blocks, represents the sample set as a multimodal representation. The intra-modality and inter-modality qualities are estimated in a weakly-supervised fashion by minimizing the multimodal separability training loss, while the multimodal network compactness loss regularizes the network to provide better generalization.

vector and an intra-modality quality score. The feature vectors and quality scores associated with samples from each modality are utilized to construct a unimodal feature representation for each modality, while the inter-modality quality scores are utilized for credit assignment among the unimodal feature vectors in the fusion of the modalities. No quality scores are explicitly provided to the framework, and quality estimation allows different features from different samples to dynamically come to the forefront as needed by re-weighting the features when constructing the multimodal embedding space [20].

Our proposed multimodal recognition model includes two quality-aware fusion blocks for adaptive feature-level re-weighting [16], [17]. We jointly train these quality-aware fusion blocks through minimizing *multimodal separability loss* and *multimodal network compactness losses*. The first loss imposes an equi-distributed multimodal embedding in which the inter-class distance of multimodal representations is maximized and the intra-class variance is minimized. In addition, this loss function considers other constraints on the unimodal embeddings, linking them to the multimodal embedding. Our trained multimodal network is utilized during the test phase to incorporate the usefulness of each sample for the recognition task. Therefore, we propose multimodal network compactness loss to improves the generalization capability of the network by minimizing the hyperspherical energy for the layers of the network.

In the experimental setup, we focus on three multimodal recognition scenarios. In the first scenario, which can represent a traditional biometric framework, each modality in the multimodal sample set consists of a single sample. Here, the first quality-aware fusion block acts as a feature extraction block while the second one aims to satisfy the recognition task by learning the inter-modality quality of each modality. This scenario enables us to analyze the performance of the second fusion block. The second recognition scenario characterizes a multi-biometric capture system used for criminal booking. Due to variations in sensor type, training level of the booking officer, and overall human error, the biometric samples collected during booking and subsequently entered into an Electronic Biometric Transmission Specification (EBTS) can vary wildly in quality. In this scenario, for each subject a set of low-quality face images and varying number of latent fingerprints are considered for the identification. This scenario mainly provides the possibility of studying the performance of the intra-modality fusion block. The third scenario, which can model an access control security system, is focused on representing a set of samples per modality with a single embedding space representation while the first fusion block considers the quality of each sample. Then, the second block aggregates the representations corresponding to different modalities through their inter-modality quality in the recognition task. This setup allows a more comprehensive evaluation of the joint performance of the two quality-aware fusion blocks.

The contributions of this paper in the field of multimodal biometrics are as follows:

- We propose a quality-aware fusion framework for multimodal biometrics applications which is optimized through learning in a weakly-supervised fashion without direct supervision of the quality of the samples or modalities.
- We formalize the multi-sample multimodal recognition problem by learning two consecutive embeddings dedicated to extract discriminative features by exploiting the intra- and inter-modality information.
- An end-to-end training framework is proposed consisting of two novel loss functions for training the quality-aware fusion blocks.
- Three specific multi-sample multimodal person recognition scenarios are designed to carefully evaluate the performance of the proposed framework. These scenarios consider two chimeric multimodal and one real-world multimodal datasets.

2 Background

2.1 Feature extraction and fusion

Convolutional neural networks (CNNs) are efficient tools that can be employed to extract and represent discriminative features from raw data. Compared to hand-crafted features, the use of CNNs as domain feature extractors has been demonstrated to be more promising when facing different biometric modalities such as face [21], iris [22], and fingerprint [23]. One of the major challenges in multimodal fusion is managing the large dimensionality of the fused feature representations, which highlights the importance of the fusion algorithm. In comparison to score- [24], [25], rank- [26], [27], [28], and decision-level [29], [30], [31] fusion schemes, feature-level fusion results in a better discriminative classifier [32], [33], [34] due to preservation of raw information [4]. Feature-level fusion integrates different features

extracted from different modalities into a more abstract and compact feature representation, which can be further used for verification or identification [2], [35]. Several frameworks have exploited feature-level fusion for multimodal biometric identification. Among them, serial feature fusion [36], parallel feature fusion [37], Canonical Correlation Analysis (CCA)-based feature fusion [38], Joint Sparse Representation Classifier (JSRC) [4], Supervised Multimodal Dictionary Learning (SMDL) [3], and Multiset Discriminant Correlation Analysis (MDCA) [2] are the most prominent techniques.

The prevalent feature fusion method in the deep learning literature is feature concatenation, which becomes very inefficient as the dimensionality of the feature space increases [9], [10], [39], [40]. Bilinear feature multiplication [11], [12] is effective since all elements of different modalities interact with each other through multiplication. The main issue in bilinear operation is the high dimensionality of its output regarding the cardinality of the inputs. Recently, to overcome this shortcoming, compact bilinear pooling has been proposed [13], [14], [15], [21]. This pooling algorithm mimics results close to bilinear pooling while the dimensionality of the embedding space is relatively small.

2.2 Multi-sample recognition

Multi-sample recognition has been recently utilized in recognition frameworks. The authors in [17] have considered a neural aggregation network, in which a set of face images is represented by a vector in the embedding space. In their proposed framework, a CNN block maps each face image into a feature vector in the embedding space. Their aggregation module consists of two blocks. These blocks adaptively aggregate the feature vectors and form a single fixed-sized feature vector to represent a set of face images. Their framework is trained with a classification loss function without direct supervision. They concluded that their proposed framework can learn to differentiate between high-quality and low-quality face images in a set of images by solely minimizing this loss function. The authors in [16] have considered the problem where a set of face images are aggregated to be presented by a vector in the embedding space. Their proposed network consists of two branches. The first branch constructs a feature vector in the embedding space for each image sample, while the second branch computes the quality score for each image sample. Finally, all of the feature vectors and quality scores in one set are aggregated through the loss function to construct the feature vector in the embedding space and represent the set of face images.

2.3 Quality-aware fusion

The quality of a biometric sample is defined as its suitability for feature extraction, correct recognition, and automated matching with a high confidence score [6]. Multimodal quality-based fusion frameworks give higher weights to the more reliable modalities. On the other hand, fusion algorithms which do not consider the quality of modality samples provide a fixed weighting scheme. Therefore, these frameworks do not present the optimal decision when sample quality varies. The quality-based fusion frameworks should receive an effective set of quality measures. These frameworks should also present an effective fusion mechanism to consider these quality scores from all samples and make an optimal decision [41].

One of the very first works considering the quality of the samples in biometric fusion is presented in [42]. In this work, the authors have manually deployed quality measures generated by human experts. The authors in [43] have proposed a framework to minimize cross-device matching performance degradation by device-specific quality-dependent score normalization. In this framework, each device score is normalized independently. To fuse the outputs of different devices, these scores are combined using a naive Bayes approach. A user-quality-based fusion of biometric modalities is proposed in [44]. This work quantifies the quality of biometric data by using user templates to incorporate the quality of the sensor data in order to generate a more reliable estimate on the matching scores, while a score-level fusion of the matching scores in the multimodal setting is considered.

A unified framework for quality-based fusion from a Bayesian perspective is proposed in [41]. In this work, the authors have investigated feature-based and cluster-based fusion algorithms for their quality-based framework. The authors in QFuse [45] present an adaptive context switching algorithm coupled with online learning to address uncontrolled noisy conditions and scalability. A probabilistic logic to explicitly take uncertainty and trust into consideration is proposed in [46]. The authors in [47] have proposed a dynamic weighted sum fusion quality metric while combining unimodal scores. This work proposes a single quality metric for each gallery-probe comparison, instead of incorporating the quality of the gallery and probe images separately. The context weighted majority algorithm presented in [48] introduced score-level and decision-level context-aware biometric fusion methods to consider the context in which biometric inputs are acquired.

In contrast to the works mentioned in this section, the framework proposed here provides a multi-sample multimodal framework which benefits from quality-aware fusion. Our framework provides a unimodal representation for each modality considering intra-modality quality of the samples in that modality and aggregates these representations using their inter-modality quality. The proposed framework is trained in a weakly-supervised fashion by minimizing the proposed multimodal separability loss to uniformly spread the centers of class representations in the embedding space. The proposed multimodal network compactness loss regularizes the multimodal network by minimizing the hyperspherical energy for different layers of the network. The performance of the proposed framework is compared with several state-of-the-art methods mentioned in this section.

3 Quality-Aware Multimodal Network

Here, we describe our methodology to provide a framework for multi-sample multimodal recognition for inputs consisting of different modalities with different quality and varying number of samples per modality. We describe the notations and the proposed quality-aware fusion in Sections 3.1 and 3.2, respectively. This fusion framework consists of two quality-aware fusion blocks. Then, as discussed in Section 3.3, we present our training criteria consisting of two loss functions. The multimodal separability loss aims to construct an embedding which provides separability of the representations by uniformly distributing the multimodal class centers in the embedding space. Due to over-parametrization, the multimodal networks suffer from a lack of generalization on unseen samples. Hence, we propose a second loss function as multimodal network compactness loss to improve the generalization of the framework by minimizing the hyperspherical

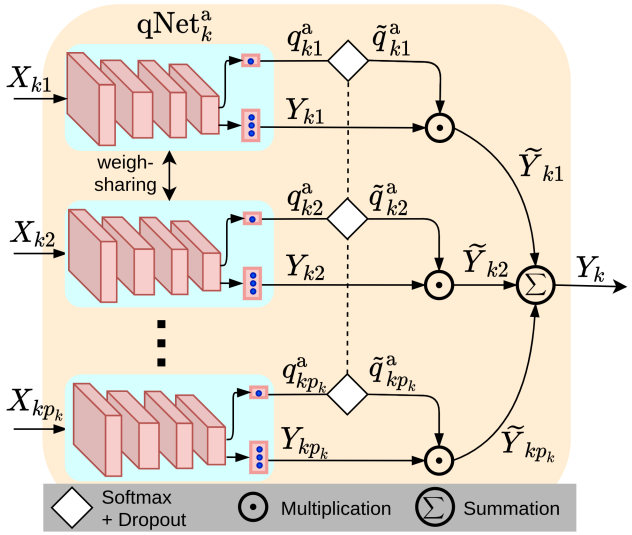

Fig. 2: Fusion$^{a}_{k}$ consists of a multi-task CNN block, qNet$^{a}_{k}$, and intra-modality aggregation to deliver a unimodal embedding space representation.

energy for different layers of the network. The two proposed loss functions are customized for the multimodal multi-sample settings. In Section 4, we study the effect of these loss functions on the performance of the proposed framework.

3.1 Notations

The following notations are used throughout this paper:

- Multimodal sample set, X, consists of one or a set of a varying number of samples from each modality.
- We consider K modalities, N training samples, and M training classes in the proposed framework.
- X_k represents the set of samples from the k^{th} modality in multimodal sample set X.
- X_{ki} represents one sample from modality k in multimodal sample set X.
- L represents the number of layers in the architecture and N_j represents the number of kernels in the j^{th} layer.
- For simplicity, in this section, we use the notation k, $1 \leq k \leq K$, for the different modalities. However, in the next section, we replace them with the actual modality names.

3.2 Quality-aware fusion mechanism

We denote each multimodal sample set by $X = \{X_k\}_{k=1}^{K}$, which consists of samples from K modalities, and X_k represents samples from the k^{th} modality. Our quality-aware framework consists of two fusion blocks. The first quality-aware block converts each X_k to a unimodal representation, Y_k, while considering the quality of each sample in X_k. This fusion block is applied on the samples from each modality and extracts the features that are the best representation of the corresponding modality: $Y_k = \text{Fusion}^{a}_{k}(X_k)$. The second quality-aware fusion block constructs the multimodal embedding space representation, Z, from the unimodal representations of modalities, $Y_k, k = 1, 2, ..., K$, considering the quality of information across the modalities. This block determines the relative credit assignment to the feature vectors in the unimodal constructed embedding spaces: $Z = \text{Fusion}^{b}(Y_1, Y_2, ...Y_K)$. In the following, we describe each of these fusion blocks in greater detail.

Intra-modality fusion (Fusion$^{a}_{k}$)**:** When multiple samples per modality are available, one can consider the average representation of samples to combine their identification information for the corresponding modality. This is equivalent to assigning equal quality scores to all of these samples. However, a better choice is to incorporate the quality of the samples to combine their representation. Fig. 2 presents this fusion block consisting of feature extraction, intra-modality quality estimation, and feature aggregation. The inputs to this block are samples from the k^{th} modality, $X_{k1}, X_{k2}, ..., X_{kp_k}$, where p_k is the number of samples for this modality in the multimodal sample set, and can vary from one sample set to the other. To utilize the quality of samples in X_k and construct a richer unimodal representation, Y_k, this block consists of a quality-aware modality dedicated network, qNet$^{a}_{k}$, and the intra-modality feature aggregation.

The first fusion block aims to provide a discriminative unimodal representation for a set of samples X_k. This block constructs a fixed-size vector representation, Y_{ki}, and an intra-modality quality score, q_{ki}, for sample X_{ki}. These representations construct the embedding space representation for the k^{th} modality through softmax normalization of the quality scores:

$$\tilde{q}^{a}_{ki} = \frac{e^{q^{a}_{ki} d_{ki}}}{\sum_j e^{q^{a}_{kj} d_{kj}}}, \tag{1}$$

where $i = 1, ..., p_k$, and $\tilde{q}^{a}_{ki}$ represents the normalized intra-modality quality score for the i^{th} sample. These quality scores are utilized to construct the representation of the k^{th} modality in the embedding space, where $q^{a}_{ki} Y_{ki}$ represents the normalized quality-aware embedding space vector representation for sample X_{ki}:

$$Y_k = \sum_i \tilde{q}^{a}_{ki} Y_{ki}. \tag{2}$$

One of the main concerns for the proposed framework is the possibility of high-quality samples dominating the other samples during the training, in which the quality scores corresponding to these samples tend to be significantly high, forcing the scores corresponding to the other samples to be very small. To resolve this issue, we perform the dropout technique on samples of the modality during the training and randomly set some of the quality scores to zero. Here, d_{ki} is a binary value, and takes values based upon the modality-specific dropout probability, μ_k. As presented in Fig. 2, when a modality consists of one sample, $X_k = X_{k1}$, the first fusion block acts as a feature extraction block, and provides a discriminative unimodal representation for this sample.

Inter-modality fusion (Fusion^{b})**:** As presented in Fig. 1, the second fusion block consists of feature transformation, inter-modality quality estimation and inter-modality feature aggregation. The feature transformation aims to provide the flexibility for the unimodal representations, Y_k, utilized for inter-modality quality estimation and feature aggregation trained by loss functions described in Section 3.3. This fusion block includes inter-modality networks, qNet$^{b}_{k}$, and a fully-connected block FNet^{b} to estimate the inter-modality quality scores.[1] This fusion block constructs a multimodal representation for the multimodal sample set, X,

1. qNet$^{a}_{k}$ networks have different architectures. However, although parameters for qNet$^{b}_{k}$ networks differ, we consider their architecture to be the same.

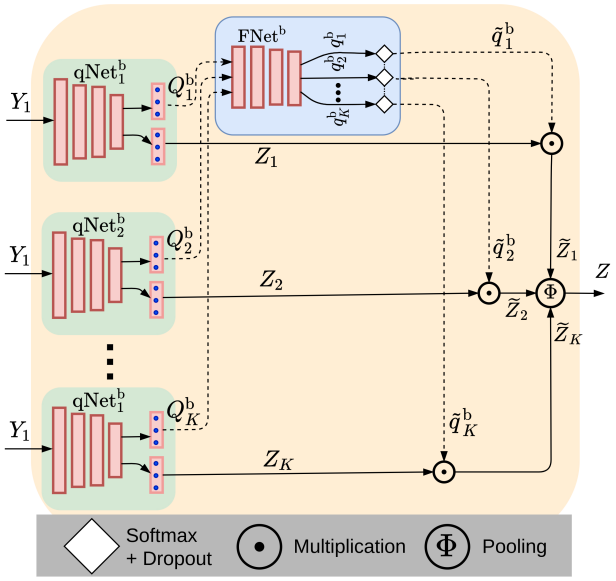

Fig. 3: $\text{Fusion}_k^{\text{b}}$ consists of K modality-dedicated networks, qNet_k^{b}, and a fully-connected network, FNet^{b}. Each modality-dedicated network, qNet_k^{b}, presents a modality with a representation and a modality-dedicated quality vector. Quality vectors are concatenated and fed into FNet^{b} to provide the inter-modality quality scores.

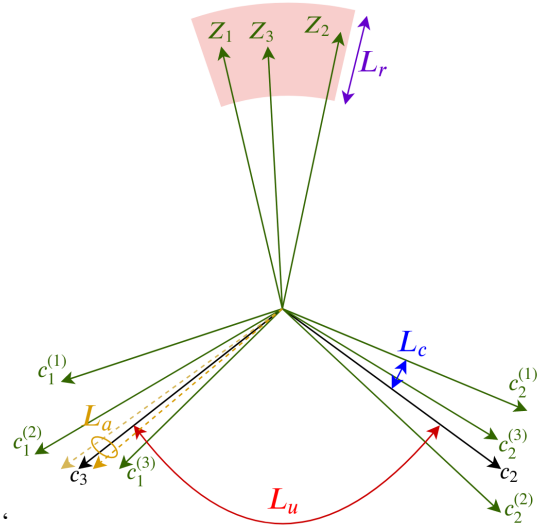

Fig. 4: Our multimodal separability training loss consists of four multimodal losses. The angular loss, L_a, provides the compactness between different multimodal sample sets of a given class. The uniform loss, L_u, aims to guarantee that the centers for different multimodal classes are uniformly distributed in the embedding space. L_c forces the centers of different modalities for a class to follow the same direction. For each class, L_r aims to provide Z_k representations that are comparable.

through learning the inter-modality quality of the unimodal representations. To this aim, the unimodal representations are fed into inter-modality networks (qNet_k^{b}, $1 \leq k \leq K$). This fusion block constructs an embedding space representation, Z_k, as well as a modality-dedicated quality vector, Q_k^{b}. Each representation and the quality vector, interacting with the corresponding unimodal feature vectors and quality vectors from the other modalities, build Z as the multimodal embedding space representation of X. To this aim, as presented in Fig. 3, the quality vectors are concatenated and fed into a fully-connected block of layers, FNet^{b}. The quality vectors from all of the modalities interact through this network, and the inter-modality quality scores corresponding to each modality, as q_k^{b}, $k = 1, 2, ..., K$ are estimated. These quality scores are normalized through softmax normalization:

$$\tilde{q}_k^{\text{b}} = \frac{e^{q_k^{\text{b}} d_k}}{\sum_j e^{q_j^{\text{b}} d_j}}, \quad k = 1, ..., K, \tag{3}$$

where $\tilde{q}_k^{\text{b}}$ represents the normalized inter-modality quality score for the k^{th} modality in X and present the relative importance of this modality in the recognition of X. To avoid the possibility of one modality dominating the other modalities during the training, we randomly set some of these quality scores to zero. This approach is implemented utilizing binary values, d_k, which take values based upon the dropout probability, μ. In the case of single-sample multimodal recognition, each normalized inter-modality quality score interprets the quality of the sample as well as the inter-modality quality of the modality. These normalized quality scores interact with the embedding space representations of the modality samples, Z_k vectors, to present a quality-aware representation of X, where $\tilde{q}_k^{\text{b}} Z_k$ represents the normalized embedding space representation for X_k. We aggregate the representations of all the modalities as:

$$Z = \Phi_{k=1}^{K}(\tilde{q}_k^{\text{b}} Z_k), \tag{4}$$

where Φ represents the aggregation method applied to the normalized embedding space vectors, such as addition [2], concatenation [10], bilinear multiplication [12], or compact bilinear pooling [13]. In all experiments presented in this paper, we consider addition of the feature vectors as the aggregation method, which results in $Z = \sum_{k=1}^{K} \tilde{q}_k^{\text{b}} Z_k$. Equivalently, this operation represents re-weighting the last layers of the modality-dedicated networks. The learned representations, Z and Y_k, are considered for multimodal and unimodal frameworks, respectively. These representations are utilized through the loss functions described in the next section to construct the decision.

The proposed framework learns inter-modality quality scores by minimizing the recognition loss function. These scores represent both the quality of the samples of one modality assigned to each multimodal sample and the inter-modality quality of the modality compared to the other modalities for a multimodal sample set. Therefore, the inter-modality quality of the k^{th} modality over the dataset for the underlying recognition task is computed as:

$$P_k^{\text{b}} = \mathbb{E}_X\{\tilde{q}_k^{\text{b}}\}, \tag{5}$$

where $\mathbb{E}_X$ represents the expectation over the multimodal sample sets in the dataset.

3.3 Multimodal separability and network compactness

Inspired by the recent advances in metric learning for deep biometric recognition such as SphereFace [49], ArcFace [50], and UniformFace [51], we design an equi-distributed multimodal

embedding in which the inter-class distance of multimodal representations is maximized and the intra-class variance is minimized. The equi-distributed constraint is applied on the centers of each class representations by uniformly spreading them in the embedding space. However, simple consideration of this constraint is not sufficient in a multimodal framework in practice. Thus, we consider two additional constraints on the unimodal embeddings, linking them to the multimodal embedding, which subsequently results in unifying the unimodal representations. In addition, to boost the generalization capability of the proposed framework, we regularize the weights in each layer of the architecture. The proposed regularization method benefits from considering similar, but not necessarily the same, architectures for different networks.

Separability of representations: The softmax loss is the common loss used for training CNN-based classifiers in the literature. This loss is defined as a combination of the last fully-connected layer, a softmax function, and a cross-entropy loss [52]. The radial properties of features learned by the softmax loss do not contribute to the discrimination of samples, thereby the angular similarity should be preferred, leading to normalized features [53].

Let us assume that N is the number of training samples, x_i is the learned feature representation corresponding to the i^{th} training sample with label y_i, and v_j and b_j are the weights and bias of the last fully connected layer corresponding to j^{th} class, respectively. To impose the angular similarity to the softmax loss, we assume that $||v_j|| = 1$ and $b_j = 0$. These assumptions result in the classification to depend entirely on the angles between x_i and v_j, $\theta_{j,i}$. Therefore, the modified softmax loss can be defined only based on $\theta_{j,i}$ [49]. Several works have provided more general assumptions on the modified softmax loss for different recognition tasks:

$$L_a = -\frac{1}{N}\sum_i \log \frac{e^{||x_i||(\cos(m_1\theta_{y_i,i}+m_2)-m_3)}}{e^{||x_i||(\cos(m_1\theta_{y_i,i}+m_2)-m_3)} + \sum\limits_{j\neq i} e^{||x_i||\cos(\theta_{j,i})}}, \tag{6}$$

where L_a represents the angular similarity loss. The effect of m_1, m_2, and m_3 are studied in SphereFace [49], ArcFace [50], and CosFace [54], respectively. Inspired by the UniformFace [51], we maintain the discriminative nature of the framework considering the angular loss, while forcing the embedding space representations to follow a uniform distribution. Here, during the training, centers c_{j_1} and c_{j_2} are assigned to the ${j_1}^{th}$ and ${j_2}^{th}$ classes [51]. Then, the uniform loss is defined as:

$$L_u = \frac{1}{M(M-1)}\sum_{j_1=1}^{M}\sum_{\substack{j_2=1\\ j_2\neq j_1}}^{M}\frac{1}{||c_{j_1}-c_{j_2}||_2+1}, \tag{7}$$

where M is the number of classes during the training phase. We combine the uniform loss and angular softmax loss as uniform angular loss:

$$L_1 = L_a + \lambda_u L_u, \tag{8}$$

where λ_u is the regularization parameter.

As described in Equation 11, we train the multimodal representation, Z and each of the unimodal representations, Y_k, using Equation 8. However, to enforce the unimodal representations, Z_k, for different modalities of a class to be comparable for both estimating the inter-modality quality scores and multimodal fusion, we define a representation loss between these representations. In particular, we want the magnitude and phase of $\widetilde{q}_k^{\mathrm{b}} Z_k$ to capture the inter-modality quality of the corresponding modality and its recognition information, respectively. Therefore, we constrain the magnitude of the modality representations to depend solely on $\widetilde{q}_k^{\mathrm{b}}$:

$$L_r = \frac{1}{NK(K-1)}\sum_{i=1}^{N}\sum_{k_1=1}^{K}\sum_{\substack{k_2=1\\ k_2\neq k_1}}^{K}\frac{(||Z_{k_1}||_2 - ||Z_{k_2}||_2)^2}{\sum_{k=1}^{K}||Z_k||_2}, \tag{9}$$

where the first summation represents multimodal sample sets in the training set and the denominator represents the summation of the norms of Z_k representations for a multimodal sample set.

We consider that for each class, Z_k representations share the same direction. This assumption provides a better separability between the classes since small variations of unimodal representations for each modality result in a minimal multimodal intra-class variance. Equivalently, the unimodal representations for different modalities of the same class should represent the same directions as the multimodal embedding space representation of that class. We define a similarity loss between the directions of the centers of the modalities of the same class as:

$$L_c = \frac{1}{KM}\sum_{k=1}^{K}\sum_{j=1}^{M}||\frac{c_j}{||c_j||_2} - \frac{c_j^{(k)}}{||c_j^{(k)}||_2}||^2, \tag{10}$$

where $c_j^{(k)}$ represents the representation center for k^{th} modality of the j^{th} class. We define multimodal separability loss as:

$$L_{ms} = \underbrace{L_a + \lambda_u L_u}_{L_1} + \underbrace{\lambda_c L_c + \lambda_r L_r}_{L_2} + \frac{1}{K}\sum_{k=1}^{K}(\lambda_{ak}L_{ak} + \lambda_{uk}L_{uk}), \tag{11}$$

where L_{ak} and L_{uk} represent the unimodal uniform angular loss function for the k^{th} modality and λ_r, λ_c, λ_{ak}, and λ_{uk} are the regularization parameters. In verification setups, $\lambda_c = 0$. In this training loss, L_2 represents the inter-modality training loss, while L_1 represents multimodal uniform angular training loss. Fig. 4 highlights the effect of the four defined losses on the separability of our multimodal framework. It is worth mentioning that the last term in the above equation represents the unimodal uniform angular training loss [51]. Therefore, while computing it, the unimodal centers of classes are considered.

Network compactness: Although deep neural networks are powerful nonlinear functions that can be trained end-to-end to extract the features and satisfy the underlying recognition task simultaneously, their over-parametrization results in highly correlated neurons that can hurt the generalization ability and incur unnecessary computation cost [55]. Multimodal deep neural networks suffer from this shortcoming the most, since they require a vast number of parameters and training multimodal sample sets. Regularization of the deep neural networks aims to avoid the representation redundancy. Regularization of these networks can roughly be categorized into implicit and explicit methods [56].

Implicit methods do not directly impose constraints on the weights, but instead, regularize the networks in order to prevent over-fitting and stabilize the training dynamics. Batch normalization [59], dropout [60], weight normalization [61], and group normalization [62] are examples of the implicit regularization methods. Explicit models, such as orthonormal regularization [63], [64], diversification [65], [66], uncorrelation [67], [68], and minimizing the hyperspherical energy (MHE) [55], aim to impose di-

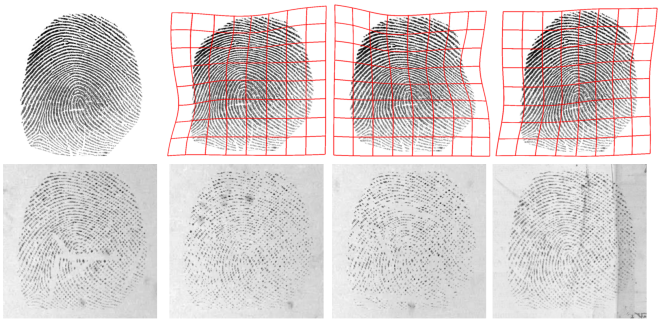

Fig. 5: Corrupted fingerprint samples from the BioCOP dataset in the training set generated from the clean sample (top left) by warping the clean fingerprint [57] (first row) and fading the fingerprint ridges at random points and adding backgrounds [58] (second row).

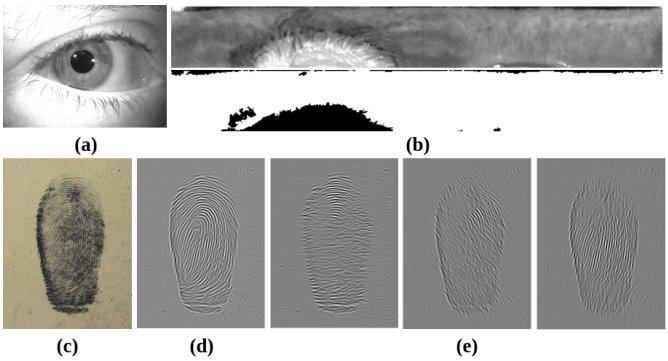

Fig. 6: (a) Eye image, (b) normalized iris and mask images, (c) latent fingerprint image, (d) enhanced fingerprint image using [70], and (e) three enhanced fingerprint images using constant Gabor angles of $0°, 60°$, and $100°$ for the whole image.

rect constraints on the weights of the network. However, the high-dimensionality of the kernels in convolutional neural networks, in addition to the vast number of kernels in multimodal frameworks, makes it difficult to regularize our multimodal framework using explicit methods [69]. Here, we expand *Compressive Hyperspherical Energy Minimization* (CoMHE) [56], which projects the kernels and neurons of the network to a low-dimensional space and minimizes the energy in the projected space, to apply it in our multimodal setting. We define the hyperspherical energy for the j^{th} convolutional layer which consists of N_j kernels, $\mathbf{W}_j = \{w_1, w_2, ..., w_{Nj}\}$ as [55]:

$$E_s(\mathbf{W}_j) = \sum_{i=1}^{N_j} \sum_{l=1, l\neq i}^{N_j} (||g(\hat{w}_i) - g(\hat{w}_l)||_2)^{-2}, \qquad (12)$$

where $\hat{w}_i = \frac{w_i}{||w_i||_2}$. However, the proposed energy minimization problem can result in colinear kernels in opposite directions. Therefore, we consider MHE in half space, in which both $\hat{w}_i$ and $-\hat{w}_i$ are utilized in the energy function above. The same energy function can be applied to the fully-connected layers where the vector w_i represents the weights going to the i^{th} neuron.

The compression function is defined as $g(\hat{w}_i) = \frac{P^* \hat{w}_i}{||P^* \hat{w}_i||}$, where P^* is the optimized projection matrix using unrolled optimization [56]. Then, the hyperspherical loss can be defined as:

$$L_{mc} = \lambda_h \sum_{j=1}^{L-1} \frac{1}{N_j(N_j-1)} \{E_s\}_j + \lambda_{h_0} \frac{1}{N_L(N_L-1)} E_s(\hat{w}_i^{out}|_{i=1}^{M}), \qquad (13)$$

where L is the number of layers. For our multimodal framework, the P^* matrices for different modalities are shared when the kernel size is the same. We find it beneficial to share projection matrices for different modalities. Sharing the projection basis can effectively reduce the number of projection parameters, also reducing the inconsistency within the hyperspherical energy minimization of projected neurons for different modalities, and further improves the generalization. To implement this loss function, we consider the dimension of the projected space to be equal to 30 for all layers. For the rest of this paper we refer to this loss function as multimodal network compactness loss, L_{mc}. Then, the overall training loss is defined as:

$$L = L_{ms} + L_{mc}. \qquad (14)$$

4 Experiments and Discussions

In this section, we present performance metrics, the data representation for different modalities, training setup, experimental scenarios, and results. We conclude this section with the discussions and comparisons with the state-of-the-art classical and deep learning algorithms to address multi-sample multimodal recognition problem. To evaluate the performance of the proposed framework, we follow the evaluation metrics and protocols presented in [71]. For the identification setup, the *Recall* metric is used. This metric computes the probability that a subject is correctly classified at least at the specified rank, while the candidate classes are sorted by their similarity score to the query samples. The performance metrics for the verification setup are the area under the curve (AUC), equal error rate (EER), and true acceptance rate (TAR) at different false acceptance rates (FAR). In addition, cumulative match curve (CMC) and receiver operating characteristic (ROC) are considered to present the performance for identification and verification setups, respectively.

Data representation: To preprocess the samples, the face images are aligned through five landmarks (two eyes, two mouth corners and nose) [72], and cropped to 112×112 resolution images. As presented in Fig. 6(b), iris images are segmented, normalized using OSIRIS [73], and transformed into 64×512 strips. In addition, each iris image is concatenated in depth with its mask image. Fingerprint images are enhanced using the method described in [74], in which the core point is detected from the enhanced image [70], and a 224×224 region centered by the core point is cropped for recognition. We follow the conventional Gabor filtering for enhancing fingerprint ridge information [70].

This approach identifies the locally optimal Gabor filter using the estimated ridge frequency and orientation maps. However, since in our problem fingerprints are assumed to be of different quality, these maps and the subsequent filtering can be significantly deteriorated. Hence, instead of estimating the best local Gabor filter, which is unreliable for low-quality samples, we feed the network with the response of several major Gabor filters with varying angles. We assume that the network learns to select the appropriate response through minimizing the recognition loss. Each fingerprint image is concatenated in depth with nine other images. The algorithm described in [70] computes the direction of the Gabor filter to estimate the ridge maps locally. Each of these additional nine images is the response of Gabor filtering with the

Fig. 7: The estimated quality of the corrupted samples from the BioCOP dataset in a multi-sample unimodal framework.

angels in $[0°–160°]$ with steps of $20°$. Figs 6(d) and 6(e) visualize several Gabor responses obtained from the latent fingerprint in Fig. 6(c).

4.1 Training setup

Training datasets: The BioCOP multimodal dataset[2] is one of the few datasets that allows training of multi-sample multimodal fusion since it contains a vast number of samples from different modalities from the same individual. This dataset consists of four sub-collections acquired over the course of 5 years, labeled by the year when each sub-collection was initiated; 2008, 2009, 2012, and 2013. There are 3,990 distinct subjects in these four sub-collections, while there are subjects common in these sub-collections, e.g., 294 in sub-collections 2012 and 2013. We consider face, iris, and fingerprint samples from this dataset in our training phase. There are a total number of 254,660, 264,821, and 338,912 samples for face, iris, and fingerprint modalities in this dataset, respectively.

The face modality contains both constrained and unconstrained images with different expressions, camera angles, and camera models. The constrained face images are acquired in head pose angles of $\pm 90°$, $\pm 45°$, and $0°$, with open and closed eyes. The fingerprint modality consists of all ten fingers captured using the *CrossMatch Verifier 300LC*, *CrossMatch Verifier 310*, and *UPEK EikonTouch 700* sensors. The iris samples contain both left and right irises and are acquired using the *Aoptix Insight*, *CrossMatch I SCAN 2*, and *LG ICAM 4000* near-infra-red sensors. Although, the interval of data acquisition in BioCOP dataset can vary up to five years, we also utilize the VGGFace2 [75] dataset to consider age-progression during the training phase. The VGGFace2 dataset also provides the training setup with more pose variations. This dataset consists of 9,131 subjects with 3.31 million face images. These images include different pose, quality, and resolution variations. As described in the **Training** section in more details, for half of the subjects which have the least number of face samples, the face samples in BioCOP dataset are replaced with face samples from the VGGFace2 dataset.

To provide real-world scenarios for our training setup, we augment the described datasets with corruptions that reduce image quality. The face images are corrupted using motion blur, JPEG compression, additive Gaussian noise, scaling (width to height ratio $\sim 0.9 - 1.1$), down-sampling and smoothing. To corrupt the iris images, blurring matrix, B, warping, W, downsampling, D, and additive noise $\bar{n}$ are considered: $\bar{X} = DBWX + \bar{n}$ as described in [76].

2. This dataset is available upon request: Jeremy.Dawson@mail.wvu.edu.

TABLE 1: The input size and network architectures. The first row for each network represents the main branch which delivers the embedding space representation and the second row represents the quality score branch.

network	input	architecture	
$\text{qNet}^{\text{a}}_{\text{Face}}$	$112 \times 112 \times 3$	C64-3×RES64-3×RES128	-3×RES256-FC512 -M-FC1
$\text{qNet}^{\text{a}}_{\text{Iris}}$	$64 \times 512 \times 2$	2×C64-3×RES64-3×RES128	-3×RES256-FC512 -M-FC1
$\text{qNet}^{\text{a}}_{\text{Fing}}$	$224 \times 224 \times 10$	2×C64-3×RES64-3×RES128	-3×RES256-FC512 -M-FC1
$\text{qNet}^{\text{b}}_{k}$	$512 \times 1 \times 1$	FC512	-FC512 -FC64-FC1
FNet^{b}	$16K \times 1 \times 1$	FC16-FC16-FCK	

Similarly, as presented in Fig. 5, the fingerprint images are degraded using two corruptions [57], [58]. The first corruption consists of warping the clean fingerprints [57] by randomly sampling the first two principal warp components extracted from the Tsinghua Distorted Fingerprint Database [57], [77]. The other corruption considers fading the fingerprint ridges at random points [58]. Data augmentation is also performed on the fingerprint images, where 20 samples are generated for each fingerprint image by translating the core point both vertically and horizontally using distances coming from Gaussian distributions [39]. Here, ten translated images are generated using a Gaussian distribution with parameters $\mu = 0$ and $\sigma = 2.5$. The remaining ten augmented images are generated with $\mu = 0$ and $\sigma = 5$.

Architecture: As presented in Fig. 2, the main architecture of each $\text{qNet}^{\text{a}}_{k}$ is a multi-task CNN that delivers a unimodal embedding space representation and a scalar quality score. This architecture consists of a ResNet network [78] and a fully-connected modality-dedicated embedding layer of size 512 to deliver Y_{ki}. The quality estimation branch of this network delivers a scalar quality score, q_{ki}. As presented in Fig. 3, the $\text{qNet}^{\text{b}}_{k}$ networks are also multi-task networks consisting of fully-connected layers which deliver an embedding space representation, Z_k, of size 512 and a modality-dedicated quality vector, Q^{b}_{k}. The score vectors dedicated to modalities are concatenated and fed into FNet^{b} to estimate the inter-modality quality scores for each modality.

Table 1 lists the architectures for these networks. We use (M) as max-pooling of size 2×2 with stride 2, (C[i]) as convolutional layers with i kernels of spatial size 3×3 followed by an M, and

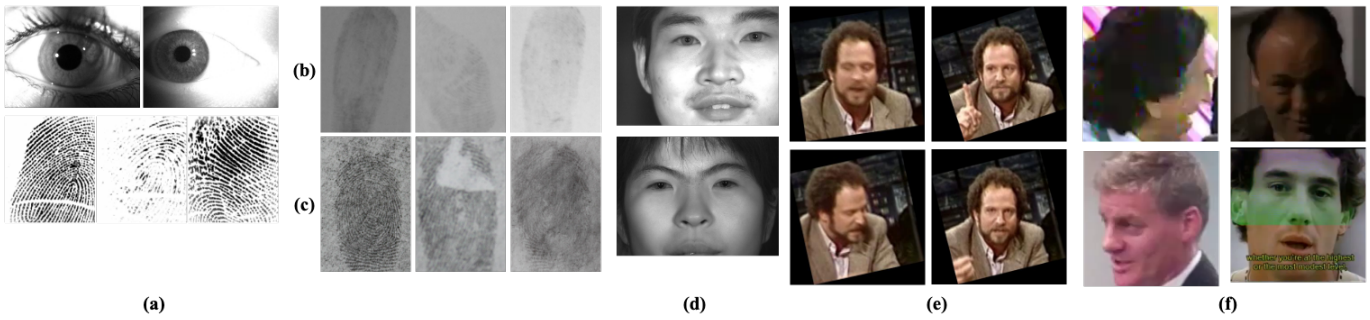

Fig. 8: Samples from test datasets before preprocessing: (a) BIOMDATA non-ideal, (b) IIIT-Delhi MOLF-Latent (D4), (c) IIIT-Delhi Latent, (d) CASIA Iris V4-Distance, (e) YouTube Face, and (f) IJB-A.

TABLE 2: Test datasets description, fine-tuning of the test datasets, and training parameters. For the verification setups, $\lambda_c = 0$.

	# classes	Modalities	Fine-tuning Datasets	Unimodal m_1, m_2, m_3	Unimodal λ_h, λ_{h_0}	Unimodal $\lambda_{ak}, \lambda_{uk}$	Unimodal μ_k	Multimodal m_1, m_2, m_3	Multimodal λ_h, λ_{h_0}	Multimodal λ_u	Multimodal λ_r, λ_c	Multimodal μ
i	219	BIOMDATA L iris	BIOMDATA R Iris	1.2, 0.3, 0.2	1.5,1	0.3,0.3	0.1	1.1,0.4,0.2	2.5,1	1	0.2,0.2	0.2
		BIOMDATA L index	BIOMDATA R Index	1.2, 0.4, 0.2	1.5,1	0.3,0.3	0.1					
		BIOMDATA L thumb	BIOMDATA R Thumb	1.2, 0.4, 0.2	1.5,1	0.3,0.3	0.1					
ii	500	IJB-A	VGGFACE2	1.35, 0.4, 0.15	1.5,1	0.4,0.4	0.2	1.1,0.4,0.2	2,1	1	0.2,0.2	0.3
		MOLF-Latent	MOLF-D4 (remain. cla.)	1.2, 0.4, 0.2	1.5,1	0.4,0.4	0.1					
iii	142	YouTube Face	VGGFACE2	1.35, 0.4, 0.15	1.5,1	0.3,0.3	0.2	1.1,0.4,0.2	2.5,1	1	0.2,0	0.3
		CASIA-distance L iris	CASIA-distance R iris	1.2, 0.3, 0.2	1.5,1	0.3,0.3	0.1					
		IIIT-Dehli Latent	IIIT-Lat. (remain. cla.)	1.2, 0.4, 0.2	1.5,1	0.3,0.3	0.1					

(FC[i]) as fully-connected layers with i nodes. Element $j\times$RES[i] consists of $2j$ residual blocks with skip connections after two convolutional layers with i kernels followed by an M. For each $\mathrm{qNet}_k^{\mathrm{a}}$, the quality estimation branch diverges from the main branch at RES128 and delivers the quality score. This branch contains M and FC layers. ReLU is used as the non-linearity after each layer for all networks, except for the final layer of score estimation branches for the $\mathrm{qNet}_k^{\mathrm{a}}$ networks and $\mathrm{FNet}^{\mathrm{b}}$ network where $sigmoid$ function is considered to limit the scores in the range [0,1].

Training: We initially train each $\mathrm{qNet}_k^{\mathrm{a}}$ for the classification setup with a varying number of modality samples per multimodal sample set, where a feature vector of size 512 is trained using uniform angular loss and network compactness loss as defined in Equations 8 and 13, respectively. Iris and fingerprint unimodal networks are trained on their respective BioCOP modalities, while the face network is trained on the combination of BioCOP and VGGFace2 datasets. The estimated normalized quality scores for degraded samples in the BioCOP dataset can be found in Fig. 7. Each row in this figure presents eight samples of the same subject to construct the unimodal multi-sample set. The number of samples from a modality in a multimodal sample set is chosen to represent the test datasets. Therefore, up to 30 samples are considered for the face modality, while for the other two modalities up to five samples are considered.

As described in Table 2, each multimodal network, is trained for the multimodal setup, while the multimodal separability loss and multimodal network compactness losses are enforced. This setup is trained for 3,990 subjects in the BioCop dataset, where, for half of the subjects which have the least number of face samples, the face samples are replaced with face samples from the VGGFace2 dataset. To study the effect of data augmentation on the $\mathrm{qNet}_{\mathrm{Fing}}^{\mathrm{a}}$, we compare the rank-25 recognition rate, with and without data augmentation, with NBIS software [79]. Data augmentation improves the performance of the proposed framework from 14.29% to 17.87%, while the NBIS software results in 12.72%.

The main branch of $\mathrm{qNet}_k^{\mathrm{a}}$ networks are initialized with weights pre-trained on Imagenet [80]. The other parameters are initialized using Kaiming initialization [81]. The preprocessing algorithm consists of the channel-wise mean subtraction. The five-fold cross-validation method is considered to estimate the best hyperparameters during the training phase. The training algorithm is deployed using mini-batch stochastic gradient descent with momentum of 0.9. The training is regularized by weight decay of 5×10^{-4} and 50% dropout for the fully-connected layers, except for the last layer of each network where the representations are considered for recognition. The moving average decay is set to 0.99 for all the networks except the iris modality, for which it is set to 0.9. Batch size is set to 32 and 16 for unimodal and multimodal frameworks, respectively. The initial learning rate is set to 0.1. The learning rate decreases exponentially by a factor of 0.1 after 10^5 iterations, and then every 5×10^4 iterations, with the final learning rate of 10^{-6}.

4.2 Results

Datasets: In our experiments, we consider multimodal dataset BIOMDATA [1], face datasets IJB-A [71] and YouTube Face (YTF) [82], iris dataset CASIA-Distance [83], and fingerprint datasets IIIT-Delhi MOLF-Latent (D4) [84] and IIIT-Delhi Latent fingerprint (D4) [85]. Samples from these datasets are presented in Fig. 8. To evaluate the performance of the proposed framework, we consider three multimodal datasets. In the first dataset, which can represent a traditional biometric framework, the recognition framework has access to only one sample from each modality. The second experimental scenario characterizes a multi-biometric identification framework where multiple samples extracted from

TABLE 3: The performance for the BIOMDATA non-ideal multimodal dataset.

Method	@10^{-4}	Rank-1	EER
JSRC [4]	27.13	97.15	7.12
GJSRC [89]	28.54	97.24	6.84
SMDL [3]	28.14	97.12	6.15
MDCA [2]	30.41	98.51	5.94
VeriFinger+OSIRIS-Sum	30.85	98.28	5.87
VeriFinger+OSIRIS-Major	29.54	97.86	6.15
CNN-Major	32.96	99.12	5.65
CNN-Sum	38.33	99.22	5.21
Weighted feature fusion [39]	43.83	99.40	4.84
Multi-abstract fusion [39]	54.06	99.57	2.97
Generalized compact bilinear [21]	58.30	99.74	2.45
Ours w/o L_2	89.06	99.80	0.86
Ours w/o L_c	**90.13**	99.84	**0.82**
Ours with weight sharing	86.82	99.80	0.90
Ours	90.11	**99.92**	**0.82**

TABLE 4: The performance for the second multimodal dataset for varying number of latent fingerprint samples per multimodal sample set.

		Verification		Identification		
		@10^{-2}	@10^{-3}	Rank-1	@10^{-2}	@10^{-1}
VeriFinger +(qNet[a] -Face) -Sum	1	96.31	92.35	96.12	91.64	94.57
	2	96.74	92.59	96.57	92.06	94.76
	3	97.12	93.42	97.67	92.54	95.14
	4	97.64	94.12	97.93	92.75	95.78
Weighted feature fusion [39]	1	97.64	93.48	98.37	92.97	96.48
	2	97.88	93.72	98.39	93.12	96.55
	3	97.93	94.01	98.41	93.21	96.58
	4	98.01	94.12	98.42	93.25	96.61
General. compact bilinear [21]	1	97.67	93.52	98.38	93.02	96.51
	2	98.90	93.75	98.40	93.14	96.57
	3	97.92	94.07	98.42	93.21	96.59
	4	98.02	94.18	98.43	93.25	96.60
Ours	0	97.34	93.14	98.37	92.71	96.36
	1	97.92	94.22	98.48	93.22	96.64
	2	98.16	94.72	98.53	93.43	96.75
	3	98.31	95.03	98.56	93.52	96.79
	4	**98.40**	**95.18**	**98.57**	**93.54**	**96.81**

a low-quality video footage and a varying number of latent fingerprints are available. The third experimental scenario, which can model a access control security system, studies the possibility of improved recognition when multiple samples are available for multiple modalities, e.g., face, iris and fingerprint. To evaluate the performance of the proposed framework for these scenarios, we consider three datasets corresponding to these three scenarios.

There are few multimodal datasets captured in real-world circumstances where each modality consists of multiple samples. Therefore, in our multimodal test setup, except for the BIOMDATA multimodal dataset, we create virtual subjects by assigning real-world biometric samples from subjects in one dataset to the subjects in other dataset *i.e.*, chimeric pairing. For instance, we consider the face samples from IJB-A dataset and fingerprint samples from IIIT-Dehli MOLF fingerprint dataset to create our second multimodal dataset. It might be noted that this procedure is feasible since modalities considered in this work are intrinsically independent [86], [87], [88]. For each dataset, the number of samples per subject and per modality may vary. Therefore, for each subject, up to 25 multimodal sample sets are randomly constructed. A brief description of each multimodal dataset, the fine-tuning, and the hyper-parameters for fine-tuning the architecture for each test dataset are presented in Table 2. It is worth mentioning that, although virtual subjects inherently can provide different recognition performances because of the quality of the samples assigned to them e.g., thumb compared to index fingerprints, the same virtual subjects are considered for all the baselines, which results in a fair comparison. In addition, to provide a better performance assessment for chimeric datasets, we expanded our experiments by evaluating the standard deviation of the multimodal recognition performance over five different sets of virtual subjects from the unimodal datasets.

Baselines: Unimodal matching algorithms considered as baselines for the iris modality are OSIRIS (Version 4.1) [73], Sun *et al.* [90], and Zhao *et al.* [91]. The performance of the fingerprint modality is compared to NBIS (Release 5.0.0) [79] and VeriFinger (Version 10.0) [92]. In addition, angular decision-making algorithms such as SphereFace [49] and UniformFace [51] as well as aggregation algorithms such as Neural Aggregation Network (NAN) [17] are considered to build baselines for face recognition performance. The performance of the proposed framework is compared with the decision-level and score-level fusion of the mentioned algorithms as well as the unimodal performance of our proposed framework. To achieve score-level and decision-level fusion we train independent classifiers for each modality. Then, we aggregate the outputs by adding the corresponding scores of each modality or using the majority voting among the independent decisions. These approaches are abbreviated with *Sum* and *Major*, respectively [3]. We also consider the element-wise averaging of the feature vectors representing samples, equivalent to assigning similar quality scores to all the samples in one modality. This approach is abbreviated as *Avg*. The performance of the proposed framework is compared with the performance of Joint Sparse Representation Classifier (JSRC) [4], Generalized Joint Sparse Representation Classifier (GJSRC) [89], Supervised Multimodal Dictionary Learning (SMDL) [3], and Multiset Discriminant Correlation Analysis (MDCA) [2]. Multi-abstract fusion and generalized compact bilinear fusion are adopted from [39] and [21], respectively.

First multimodal dataset: BIOMDATA non-ideal multimodal database-Release 1 [1] is a challenging dataset, since many of the samples are damaged with blur, occlusion, sensor noise and shadows [2]. Six biometric modalities are considered in our experiments: left and right irises, and thumb and index fingerprints from both hands. Our experiments are conducted on 219 subjects that have samples in all six modalities. Following the protocol in [93], we fine-tune the network on the right index and thumb fingerprints and the right iris samples. The performance of the framework is tested on the left thumb and index fingerprints and the left iris samples as three modalities constructing the multimodal sample sets. In this dataset, there are 1458, 1519 and 1504 images for left iris, left thumb, and left index, respectively. In the identification setup, for each modality, four randomly chosen samples are used as the gallery and the remaining samples are used for the test set. For any modality in which the number of the samples is less than five, one sample is used as the probe and the remaining samples are used as the gallery. Then, the multimodal sample sets are generated as described in the **Datasets** section. In the verification setup, the pairs of multimodal sample sets are generated from these described disjoint sub-sets.

Table 3 and Fig. 9 present the results for this dataset. Here,

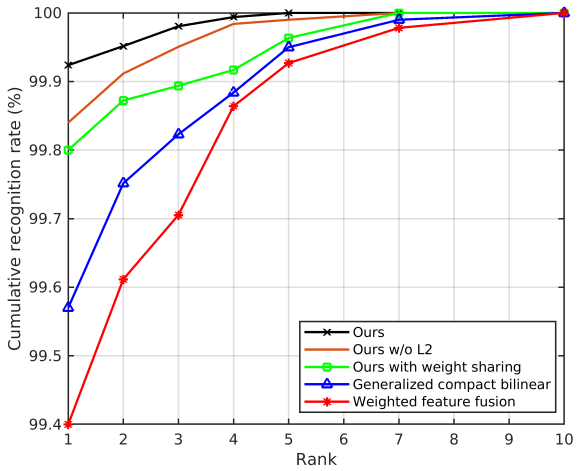

Fig. 9: The CMC curves for identification performance on the first multimodal dataset.

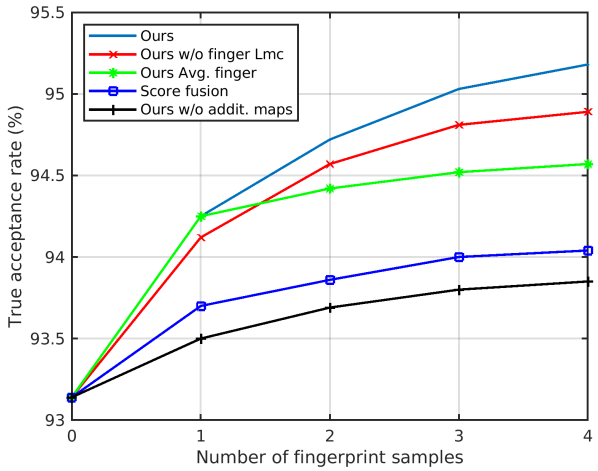

Fig. 10: The verification TAR at FAR $= 10^{-3}$ for the second multimodal dataset when the number of latent fingerprint samples per subject increases.

CNN-Major and CNN-Sum represent the rank-level and score-level fusion of the outputs of each modality network when quality scores for the second fusion block are not considered. As presented in Table 3, the proposed framework outperforms these two frameworks by more than 30% for TAR at FAR $= 10^{-4}$. We also observe that the effect of weight-sharing between $\mathrm{qNet}_k^{\mathrm{b}}$ networks, which decreases the number of parameters in the inter-modality quality score estimation networks by 66% and results in 0.08 drop of the performance in terms of EER. Fig. 9 presents the CMC curve for the proposed framework in comparison with mentioned frameworks. As presented in this figure, the proposed framework consistently outperforms the baseline frameworks.

For this dataset, since one sample per modality is considered, the first quality-aware fusion block acts as a feature extraction framework, feeding the features to the second fusion block to estimate the inter-modality-quality of each modality. In our experiments, we observed that for this dataset the expectation of inter-modality quality score, as defined in Equation 5, for the left iris, index, and thumb are 0.44, 0.32, and 0.24, respectively. This observation is aligned with our expectation since the unimodal performance of these three modalities, which can represent the overall quality of these modalities, follows the same sequence. Since the considered modalities are independent, we expect that the alignment of the embedding space representations of different modalities during the training should only affect the identification and not the verification performance. These expectations are consistent with the performance observed in Table 3.

Second multimodal dataset: This dataset consists of face samples from IJB-A [71] and latent fingerprints from IIIT-Delhi MOLF-Latent (D4) [84]. IJB-A is a challenging face recognition dataset consisting of unconstrained images. This dataset contains 500 individuals, 5,397 images, and 20,412 video frames split from 2,042 videos. These images are captured with extreme pose, illumination, and expression conditions. The testing protocol for the dataset consists of 10 folds, where each fold is represented by a different random collection of 333 subjects for training and 167 for testing. IIIT-Delhi MOLF-Latent (D4) contains 4,400 fingerprint samples from 1,000 classes (10 fingers of 100 individuals). The latent fingerprints are captured using a black powder dusting process. In this experimental setup, we consider two modalities, where we assign 500 fingerprint classes with multiple samples to the IJB-A subjects. The network is fine-tuned on multimodal sample sets consisting of the remaining 500 fingerprint classes and 500 classes of VGGFace2. As presented in Table 4, for our multimodal experiments, we follow the setup presented for the IJB-A dataset. In addition, Fig. 10 and Fig. 11 study the effect of increasing the number of fingerprint samples per subject.

We compare the performance of the proposed framework with the same framework when the network compactness loss, L_{mc}, is not considered for the fingerprint modality. We observe that the AUC performance gap widens from 0.13 to 0.29, when the number of fingerprint samples increases from one to four. We believe this improvement is due to the more reliable fingerprint representation in the embedding space when four fingerprints are considered. We also observe that the performance when the feature vectors representing the fingerprints are averaged, i.e., the same quality score is considered for all the fingerprint samples, drops by 0.61 when considering four fingerprint samples. We study the effect of the auxiliary ridge maps when we compare the performance with the framework in which no additional map is concatenated to the original map. We observe that the variation of our framework with these additional maps in which equal quality scores are assigned to the fingerprints outperforms the previously mentioned framework with a margin of 0.72 for four fingerprints. In addition, we compare the performance of the proposed framework with the score-based fusion of VeriFinger and $\mathrm{qNet}_{\mathrm{Face}}^{\mathrm{a}}$. As presented in Table 4, inter-modality quality estimation of the proposed framework can improve the performance of the $\mathrm{qNet}_{\mathrm{Face}}^{\mathrm{a}}$ when combined with $\mathrm{qNet}_{\mathrm{Fing}}^{\mathrm{a}}$ compared to its combination with VeriFinger.

To compare the performance of the proposed method with other fusion methods, we average the feature vectors representing each fingerprint sample and consider the score-level fusion of the fingerprint and face modalities. As presented in Fig. 10, the proposed framework consistently outperforms the score-level fusion of the modalities. Table 4 studies the effect of the inter-modality-quality score estimation by comparing the performance of the

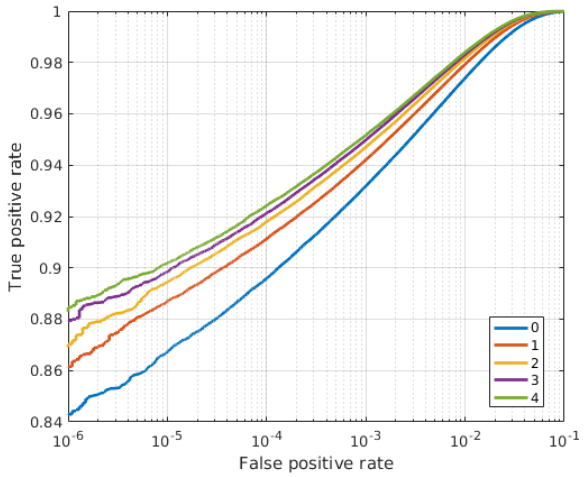

Fig. 11: The verification performance for the second multimodal dataset when the number of latent fingerprint samples per subject varies.

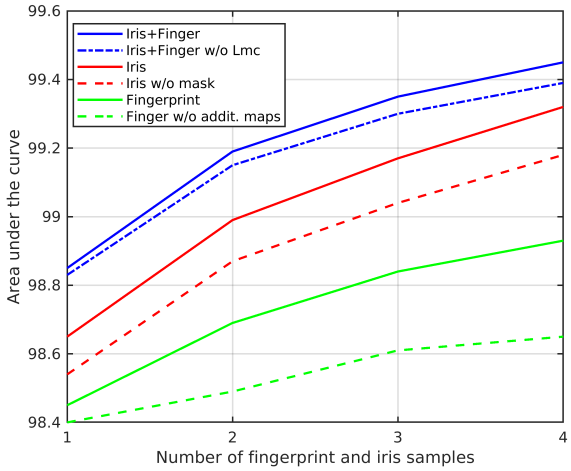

Fig. 12: The verification AUC for the third multimodal dataset when the number of iris and latent fingerprint samples per subject increases.

proposed framework with score-level fusion of and $\mathrm{qNet}^{\mathrm{a}}_{\mathrm{Face}}$, and weighted feature fusion [39] and generalized compact bilinear pooling [21] of the outputs of $\mathrm{qNet}^{\mathrm{a}}_{k}$ networks. As presented in this table, the conventional aggregation of the feature vectors, when one of the modalities is significantly more informative than the other modality and samples in the dataset are of varying quality, does not prove to be beneficial. On the other hand, the estimation of the inter-modality-quality scores for samples in the multimodal sample set can provide better recognition performance. As presented in this table, our proposed framework outperforms these fusion methods. However, when we study the rate of the recognition improvement by adding fingerprint samples, we observe that the gap in this rate shrinks when the number of fingerprint samples increases. For instance, adding the first fingerprint sample results in a TAR at a FAR of 10^{-3} improvement of 0.81 and 0.34 for the proposed framework and weighted feature fusion, respectively. However, the improvement, when adding the fourth sample, is equal to 0.15 and 0.11, respectively.

For this dataset which consists of face image sets and latent fingerprint samples of low quality, the expectations for inter-modality quality scores, when one latent fingerprint is considered, are 0.72 and 0.28 respectively. The expectation for inter-modality-quality scores of the fingerprint modality scores increases from 0.28 to 0.33 when the number of fingerprints per multimodal sample set increases from one to four. These inter-modality quality scores, which can represent both the importance of a modality in the joint multimodal framework as well as the overall quality of the samples, are aligned with our expectation about inter-modality-quality scores.

Third multimodal dataset: We construct this multimodal dataset using three datasets. YouTube Face dataset (YTF) [82] is designed for unconstrained face verification in videos. It contains 3,425 videos of 1,595 different people, and the video lengths vary from 48 to 6,070 frames with an average length of 181.3 frames. In this dataset, ten folds of 500 video pairs are available. CASIA Iris V4-Distance dataset is a subset of database [83] and contains 2,446 instances from 142 different subjects. IIIT-Delhi Latent fingerprint (D4) dataset [85] consists of 1046 latent fingerprint samples pertaining to 15 subjects with all 10 fingerprints, thus the dataset has 150 classes. The latent fingerprints are captured under semi-controlled environment the black powder dusting process. The dataset is prepared in multiple sessions with variations in background, and captures the effect of dryness, wetness, and moisture. This provides sample variation in the quality, noise, and information content of latent fingerprint samples. The samples of lifted latent fingerprints are digitized using a Canon EOS 500D camera. For our third dataset, we consider three modalities and 142 classes. Here, we select 142 random classes from YouTube Face, assign iris classes and randomly selected fingerprint classes to them, and follow the protocol described in YouTube Face for our verification setup. The network is fine-tuned using the VGGFACE2, the right iris samples from CASIA Iris V4-Distance, and the remaining subjects from IIIT-Delhi Latent (D4).

Table 5 and Fig. 12 present the verification performance for this dataset. In Table 5a, we study the impact of adding the latent fingerprint and iris samples to the YouTube Face dataset classes. As expected, adding iris samples improves the recognition performance better than adding the latent fingerprint samples. In addition, the performance improves drastically after adding the first few fingerprint and iris samples. However, this improvement lessens when adding several samples. Furthermore, Fig. 12 presents results related to the effect of different modifications in the multimodal framework. In this figure, the blue curves present the results when up to four pairs of iris and latent fingerprint samples are added to each subject. Similarly, red and green curves represent the performance when adding only iris or fingerprint samples to each class, respectively. The dashed blue curve represents the performance when L_{mc} is not considered for iris and fingerprint networks. As presented in this figure, the iris modality outperforms the latent fingerprint modality by a wide margin. In addition, the gap between the improvement resulted by using iris samples and latent fingerprint samples widens as the number of considered samples increases. On the other hand, the performance of the framework in which only iris samples are added to YouTube Face dataset classes approaches the performance of framework using all three modalities as the number of iris samples increases.

TABLE 5: The verification AUC for the third multimodal dataset when the number of latent fingerprint and iris samples per subject vary.

Fingerprint \ Iris	0	1	2	3	4
0	98.12	98.65	98.99	99.17	99.32
1	98.45	98.85	99.10	99.25	99.39
2	98.69	99.01	99.19	99.31	99.42
3	98.84	99.12	99.25	99.35	99.44
4	98.93	99.17	99.28	99.38	99.45

(a) Ours

Fingerprint \ Iris	0	1	2	3	4
0	98.12	98.23	98.44	98.53	98.61
1	98.24	98.42	98.53	98.61	98.75
2	98.43	98.61	98.82	98.72	98.94
3	98.67	98.82	98.94	99.03	99.12
4	98.72	98.93	99.13	99.26	99.36

(b) $\mathrm{qNet}^{\mathrm{a}}_{\mathrm{Face}}$+VeriFinger+OSIRIS-Sum.

For this multimodal dataset, we observe that the inter-modality-quality score for face, iris and fingerprint modalities, when one fingerprint and iris samples are included in the multimodal sample set, are 0.51, 0.30, and 0.19, respectively. However, by increasing the number of iris and fingerprint samples from one to four, their corresponding scores improve to 0.37 and 0.24, respectively. As studied in [50], the combination of inter-class loss functions cannot improve the angular loss defined in Equation 6 for face recognition. However, we observe that the multimodal network compactness loss, L_{mc}, although not very efficient for the unimodal networks, can improve the multimodal performance. We believe this is due to the over-parametrization of the multimodal network compared to the unimodal networks. In addition, we compare the performance of the proposed framework with the score-level fusion of VeriFinger and OSIRIS in Table 5b. As presented in this table, the proposed framework outperforms the performance of this score-level fusion with a wide margin when considering a fewer number of samples. However, when the number of samples is increased, the performance of the score-level fusion becomes closer to the proposed framework.

Chimeric pairing and statistical analysis: Due to the lack of multimodal datasets for evaluating the current work, we manually constructed two chimeric multimodal sets. Multimodal samples generated randomly in our experiments can possess varying recognition potentials which consequently affects the matching performance and comparisons. To provide a better performance assessment for chimeric datasets, we expanded our experiments by evaluating the standard deviation of the multimodal recognition performance over five different sets of virtual subjects from the unimodal datasets. The standard deviation of the performance for the second and the third datasets, when considering four samples from each modality, is observed to be less than 0.05% and 0.02%, respectively. This validates the effectiveness of random pairing for constructing multimodal datasets using independent single modalities, and advocates that the same criterion can be used to alleviate the scarcity of real-world multimodal datasets. It is also worth mentioning that, although the improvement in the performance resulted from multimodal settings compared to the unimodal settings, may be considered as small, these improvements can be beneficial when large-scale applications such as passport control are considered [82], [94], [95], [96].

TABLE 6: The unimodal performance of $\mathrm{qNet}^{\mathrm{a}}_{\mathrm{Iris}}$ on single-samples. For Sun *et al.* and Zhao *et al.*, results are reported from [91].

	BIOMDATA Left			CASIA-Dist. Left		
Method	@10^{-3}	Rank-1	EER	@10^{-3}	Rank-1	EER
OSIRIS [73]	86.30	95.17	4.43	80.07	88.68	6.39
Sun *et al.* [90]	90.11	–	5.19	83.07	–	7.89
Zhao *et al.* [91]	94.30	–	2.63	84.10	–	5.50
Ours w/o mask	92.84	99.38	3.14	83.51	94.52	6.47
Ours	**95.48**	**99.51**	**2.15**	**86.12**	**96.14**	**5.09**

4.3 Ablation study

Here, we study the feature extraction performance of the first quality-aware fusion block when a single sample is fed to the network. On the other hand, the aggregation capability and quality estimation performance of this block is studied when we feed multiple samples to the unimodal network.

Single-sample and single-modality: For the IJB-A [71] and YouTube Face [82] datasets, we follow the protocol presented in the corresponding papers. For the identification setup for iris datasets contained in BIOMDATA [1] and CASIA-Distance [83], we consider the protocol presented in [2], while for the verification setup we follow [91]. Here, in the identification setup, for each subject, four randomly selected samples are considered as the gallery and the remaining samples are considered as the probes. For fingerprint datasets present in BIOMDATA, IIIT-Delhi MOLF-Latent (D4) [84], and the IIIT-Delhi Latent fingerprint dataset [85], we follow [2], [84], and [85], respectively. It should be noted that, for experiments performed on the IIIT-Delhi MOLF-Latent (D4) dataset, we consider the latent-to-sensor framework, while for the IIIT-Delhi Latent fingerprint dataset, latent-to-latent recognition is considered.

Tables 6 and 7 present the unimodal results for the single-sample setup on iris and fingerprint modalities. The performance of the $\mathrm{qNet}^{\mathrm{a}}_{\mathrm{Iris}}$ is compared to OSIRIS [73], Sun *et al.* [90], and Zhao *et al.* [91]. We also compare the performance of the proposed framework with the same framework when not concatenating the mask images to the iris images. As presented in Table 6, concatenating the mask image with the iris image can improve the recognition performance, which outperforms the state-of-the-art framework [91] by 0.48 and 0.41 in terms of EER on BIOMDATA left and CASIA-Distance left, respectively.

Table 7 presents the performance of $\mathrm{qNet}^{\mathrm{a}}_{\mathrm{Fing}}$ compared to NBIS [79], VeriFinger [92], and $\mathrm{qNet}^{\mathrm{a}}_{\mathrm{Fing}}$ fed with only the original fingerprint image processed with Gabor filters with locally estimated angles and not concatenated with the remaining nine maps explained in the **Data representation** section. As presented in this table, the additional ridge maps are mostly beneficial for latent fingerprints since these maps are constructed using constant directions for the whole fingerprint image. These auxiliary maps are considered according to the fact that ridge orientation estimation for latent fingerprints is not accurate due to the photometric and geometric distortions, and therefore, using that orientation map for enhancement reduces the reliability of ridge information. However, by considering a constant global direction for each auxiliary map, there is a chance that the ridge information from unreliable regions can be enhanced in at least one of the additional maps and the deep model can exploit these maps to construct better representations.

TABLE 7: The unimodal performance of $\mathrm{qNet}^{\mathrm{a}}_{\mathrm{Fing}}$ for single-samples.

	BIOMDATA L Thumb			BIOMDATA L Index			MOLF-Latant			IIIT-Latant			
Method	$@10^{-3}$	EER	Rank-1	$@10^{-3}$	EER	Rank-1	$@10^{-2}$	Rank-25	Rank-50	$@10^{-2}$	Rank-1	Rank-10	Rank-25
NBIS [79]	57.49	14.45	71.23	68.68	6.48	87.17	12.21	5.01	8.63	48.33	52.31	58.90	63.42
VeriFinger [92]	68.26	12.03	76.16	76.12	6.18	90.41	8.14	2.87	6.56	55.19	61.02	74.00	77.44
Ours w/o 9 maps	61.23	13.16	74.78	73.12	6.35	88.51	5.21	4.51	7.73	52.89	57.57	68.21	72.35
Ours w/o L_{mc}	74.83	8.71	83.43	81.84	5.87	94.87	22.12	37.14	64.94	60.56	69.47	81.54	86.81
Ours	**76.16**	**7.52**	**84.71**	**87.63**	**4.66**	**95.12**	**25.14**	**40.41**	**67.11**	**65.53**	**73.81**	**85.21**	**89.74**

TABLE 8: Identification performance of $\mathrm{qNet}^{\mathrm{a}}_{\mathrm{Fing}}$ for multi-sample setup consisting of one to four samples.

	BIOMDATA L Thumb Rank-1				BIOMDATA L Index Rank-1				MOLF-Latant Rank-25				IIIT-Latant Rank-1			
Method	1	2	3	4	1	2	3	4	1	2	3	4	1	2	3	4
VeriFinger-Major	76.16	76.74	82.45	84.23	90.41	91.22	94.92	95.58	2.87	3.54	9.10	13.41	61.02	62.81	62.84	76.20
VeriFinger-Sum	76.16	80.10	83.87	85.41	90.41	93.98	95.15	95.89	2.87	5.49	10.83	14.45	61.02	68.82	74.43	78.56
Ours-w/o 9 maps-Sum	74.78	76.45	77.34	78.12	88.51	90.38	91.79	92.24	4.51	5.81	6.76	7.13	57.57	61.79	63.64	65.42
Ours-w/o 9 maps	74.78	78.51	81.64	83.34	88.51	91.56	93.58	94.19	4.51	6.03	8.42	9.05	57.57	64.80	69.28	63.82
Ours-Major	84.71	85.12	91.53	93.85	95.12	95.47	97.63	98.51	40.41	40.92	56.52	60.33	73.81	73.91	80.95	82.49
Ours-Sum	84.71	87.74	91.63	93.59	95.12	97.96	98.01	98.94	40.41	50.48	56.33	60.07	73.81	77.23	80.18	82.41
Ours-Avg.	84.71	88.57	92.46	94.12	95.12	98.17	98.87	99.22	40.41	51.39	57.27	62.76	73.81	77.40	81.42	83.83
Ours	**84.71**	**89.41**	**93.74**	**95.64**	**95.12**	**98.53**	**99.23**	**99.70**	**40.41**	**52.34**	**58.79**	**64.88**	**73.81**	**78.78**	**82.64**	**85.12**

TABLE 9: Recognition performance of $\mathrm{qNet}^{\mathrm{a}}_{\mathrm{Iris}}$ for multi-sample setup consisting of one to four samples.

	BIOMDATA Left								CASIA-Dist. Left							
	$@10^{-3}$				Rank-1				$@10^{-3}$				Rank-1			
Method	1	2	3	4	1	2	3	4	1	2	3	4	1	2	3	4
OSIRIS-Major	86.30	86.41	88.41	90.17	95.17	95.21	95.83	95.97	80.07	80.23	85.68	86.35	88.68	88.70	89.01	89.12
OSIRIS-Sum	86.30	89.10	90.92	91.53	95.17	95.94	96.43	96.78	80.07	84.49	86.83	88.45	88.68	88.88	89.15	89.35
Ours-Major	95.48	97.02	97.12	97.34	99.51	99.54	99.55	99.55	86.12	86.22	86.32	86.38	96.14	96.63	96.93	97.13
Ours-Sum	95.48	97.74	97.63	97.59	99.51	99.55	99.07	99.59	86.12	88.33	89.41	90.45	96.14	96.75	96.82	97.25
Ours-Avg.	95.48	97.49	98.17	98.55	99.51	99.60	99.63	99.67	86.12	88.87	91.24	93.85	96.14	96.87	97.57	97.92
Ours	**95.48**	**98.67**	**99.75**	**99.82**	**99.51**	**99.62**	**99.67**	**99.71**	**86.12**	**89.58**	**92.81**	**94.32**	**96.14**	**97.35**	**98.51**	**98.89**

Multi-sample and single-modality: Tables 10, 8, and 9 present the results for the unimodal multi-sample setup for face, fingerprint, and iris modalities, respectively. The unimodal networks are trained with uniform angular loss as defined by Equation 8 and network compactness loss. In addition to comparison with other frameworks, in these tables, we study the effect of these loss functions on the unimodal performance. As presented in Table 10, the performance of the $\mathrm{qNet}^{\mathrm{a}}_{\mathrm{Face}}$ is compared to frameworks such as SphereFace [49] and UniformFace [51], and it is shown to achieve state-of-the-art performance with fewer training samples; 3M compared to 6M for UniformFace. Here, on the face recognition task, we clearly observe the effect of the network compactness loss as well as the separability loss which improve the performance of the proposed framework to 93.1% and 98.12% on IJB-A and YouTube Face datasets for verification at a FAR of 10^{-3}, respectively.

Table 8 presents the multi-sample recognition results on four fingerprint datasets when up to four samples are considered. The performance of the proposed quality-aware framework is compared to the rank-level and score-level fusions of the single-sample representations as well as element-wise averaging of these representations. On the other hand, the fusion of the VeriFinger scores outperforms our framework when auxiliary ridge maps are not considered. However, the proposed framework outperforms VeriFinger by a large margin on the latent fingerprints. Table 9 presents the results for BIOMDATA left and CASIA-Distance left iris datasets. As presented in this table, the proposed framework consistently outperforms different score-, rank-, and feature-level variations of the multi-sample frameworks as well as score-level and rank-level fusion of the OSIRIS verifier.

Quality measures: To analyze the quality scores estimated by the proposed framework, we focus on our single-sample single-modality frameworks and the comparison of the distribution of learned scores with no-reference image quality metrics Blind/Referenceless Image Spatial Quality Evaluator (BRISQUE) [102] and NIST Fingerprint Image Quality (NFIQ 2.0) [103]. The BRISQUE score is predicted by a support vector regression model trained on a set of images with their corresponding differential mean opinion score as the target. The set of images contain original images along with their distorted versions corrupted by known distortion effects such as compression artifacts, blurring, and noise. The BRISQUE score is a scalar value typically in the range $[0, 100]$. Lower values of BRISQUE score reflects better perceptual quality. NFIQ 2.0 assigns each fingerprint a score from zero (low quality) to 100 (high quality). In Figs 13 and 14, to provide a better comparison, all the quality scores are normalized to $[0, 1]$, with higher values representing better quality. It is worth mentioning that in the original BRISQUE algorithm lower values represent higher qualities. Therefore, in Fig 13, the BRISQUE scores are reversed as, *1-normalized score*, to provide comparison with the scores estimated by other frameworks.

Fig. 13 compares the distribution of the quality scores mea-

TABLE 10: The performance of $\mathrm{qNet}^{\mathrm{a}}_{\mathrm{Face}}$ for multi-sample face recognition.

	IJB-A				YTF	
	Verification		Identification		Verification	
Method	$@10^{-2}$	$@10^{-3}$	Rank-1	Rank-5	$@10^{-3}$	AUC
DR-GAN [97]	77.4 ± 2.7	53.9 ± 4.3	85.5 ± 1.5	94.7 ± 1.1	–	–
Triplet Similarity [98]	79.0 ± 3.0	59.0 ± 5.0	88.0 ± 1.5	95.0 ± 0.7	–	–
Template Adaptation [99]	93.9 ± 1.3	83.6 ± 2.7	92.8 ± 1.0	97.7 ± 0.4	–	–
NAN [17]	93.3 ± 0.9	86.0 ± 1.2	95.4 ± 0.7	97.8 ± 0.4	95.72	98.8
SphereFace [49]	92.3 ± 1.6	88.4 ± 4.2	93.2 ± 1.3	96.5 ± 1.1	95.0	–
DeepFace [100]	–	–	–	–	91.4	96.3
CosFace [54]	–	–	–	–	97.6	–
ArcFace [50]	–	–	–	–	98.02	–
PRN [101]	96.5 ± 0.4	91.9 ± 1.3	98.2 ± 0.4	$\mathbf{99.2 \pm 0.2}$	95.8	–
UniformFace [51]	96.9 ± 0.8	92.3 ± 1.7	97.9 ± 0.5	98.8 ± 0.2	97.7	–
Ours-Avg.	92.5 ± 0.8	82.7 ± 2.3	97.9 ± 0.5	98.8 ± 0.2	97.51	99.0
Ours w/o L_{mc}	97.0 ± 0.7	92.5 ± 1.9	98.0 ± 0.5	98.8 ± 0.3	98.06	99.0
Ours	$\mathbf{97.3 \pm 0.7}$	$\mathbf{93.1 \pm 1.7}$	$\mathbf{98.4 \pm 0.4}$	$\mathbf{99.2 \pm 0.2}$	**98.12**	**99.1**

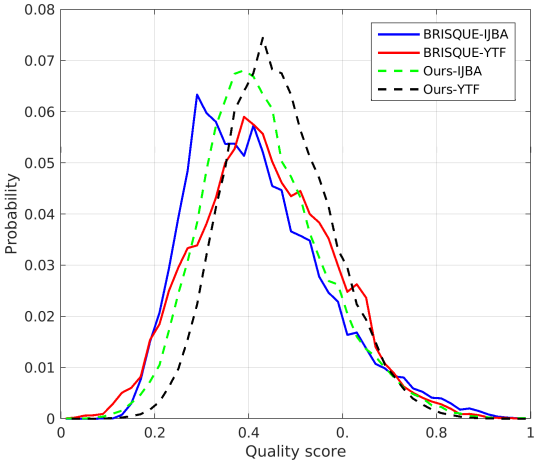

Fig. 13: The probability of estimated quality scores for the proposed framework compared to BRISQUE on IJB-A and Youtube Face datasets.

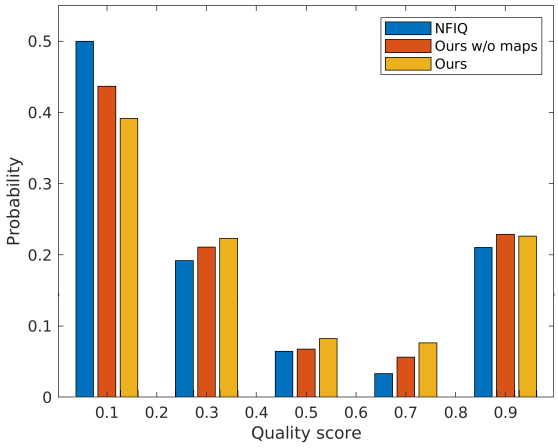

Fig. 14: The probability of *quantized* quality scores for the proposed framework compared to NFIQ 2.0 on fingerprint datasets.

sured by the BRISQUE with our multi-sample single-modality framework on IJB-A and Youtube Face datasets. As presented in this figure, the distribution of the quality scores estimated by the proposed framework closely follow the BRISQUE scores. On the other hand, Fig. 14 compares the quality scores estimated by the proposed framework with NFIQ 2.0. In this figure, all the fingerprint samples from three test multimodal datasets are considered and the scores are *quantized* to five levels. As presented in these figures, our estimated scores are closer to one compared to no-reference measures since our proposed networks are trained mostly on corrupted samples, while NFIQ 2.0 and BRISQUE are trained on a large number of images with various qualities including high-quality samples. Therefore, our estimated scores provide a relative quality measure for low-quality samples, while each mentioned no-reference measure computes a more global quality metric. It should be noted that our quality scores do not aim to serve as stand-alone quality measures, and the current comparisons seek to showcase the positive correlation of our sample weighting scores with actual quality scores.

Failed verification of sample sets: To study the performance of the proposed framework from another perspective, we investigate the multimodal sample sets that our framework fails to correctly verify. The first multimodal dataset which consists of BIOMDATA samples, includes images corrupted with blur, occlusion, shadows, and sensor noise. In Fig. 15, we present four pairs of sample sets, consisting of left iris, left thumb, and left index, that the framework fails to correctly verify. For better presentation, we also present the left eye, although the eye image is not used as an input to the framework. In Figs 15(a) and 15(b), we investigate positive multimodal sample pairs that are rejected by our framework for false acceptance rate (FAR) of 10^{-4}. To further investigate the performance of the framework, we increase the FAR for these sample sets and observe that these pairs are correctly verified at FAR = 2×10^{-4} and 5×10^{-4}, respectively. Then, in Figs 15(c) and 15(d), we illustrate two negative multimodal sample pairs that are accepted by the framework at false rejection rates (FRR) of 10^{-4}. Similarly, we increase the FRR for these sample sets and these pairs are rejected at FRR = 5×10^{-3} and 10^{-2}, respectively.

For the third multimodal dataset which consists of multiple samples of faces, irises captured from distance, and latent fingerprints, we study the performance when adding iris and fingerprint samples to a pair of incorrectly verified sample sets of face images.

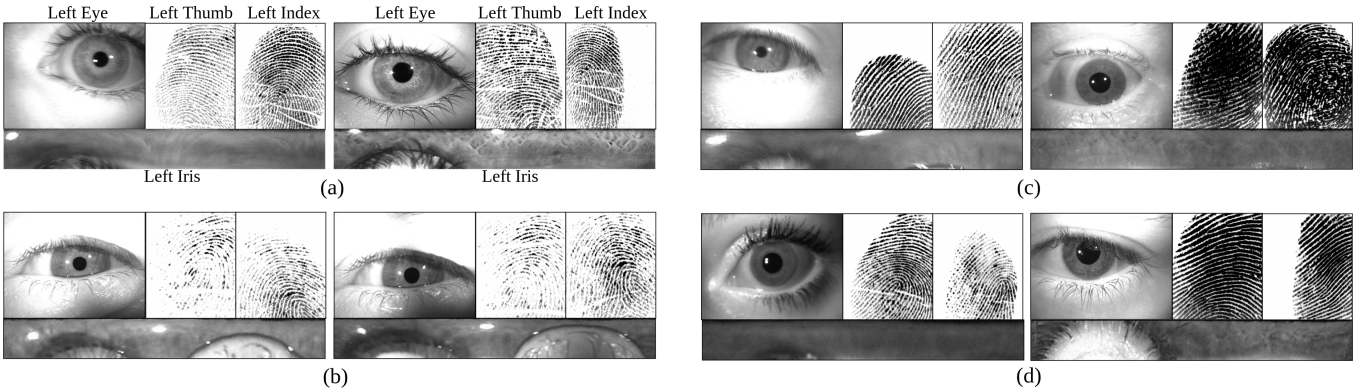

Fig. 15: Examples of failed verification for the first multimodal dataset, (a-b) false rejection at FAR = 10^{-4} and (c-d) false acceptance at FRR = 10^{-3}.

As presented in Fig. 16, the pairs of samples that we study include two sets of frames from the same class that are rejected by the framework at FAR = 5×10^{-4}. However, by adding two latent fingerprint and two iris samples to each of the multimodal face sample sets, the framework successfully accepts the sample set at the same mentioned FAR. It might be noted that these fingerprint and iris samples are not independently verified correctly by the framework at the same threshold.

One observation through our experiments is that, as discussed earlier in this section, although some of the fingerprint and iris samples are visually considered of very bad quality, the preprocessing plays a crucial role in their verification. We can also observe that the majority of the samples our framework fails to verify correctly are of very low quality or far from their class center. Another observation in these experiments is that the incorrectly accepted pairs of multimodal sample sets which pass the hardest FRR values are the pairs of very low quality. This is aligned with the common observation in image recognition that, in the representation domain, very low quality samples are relatively closer to the mass centers of the training set compared to the mass center of their own class [104], [105].

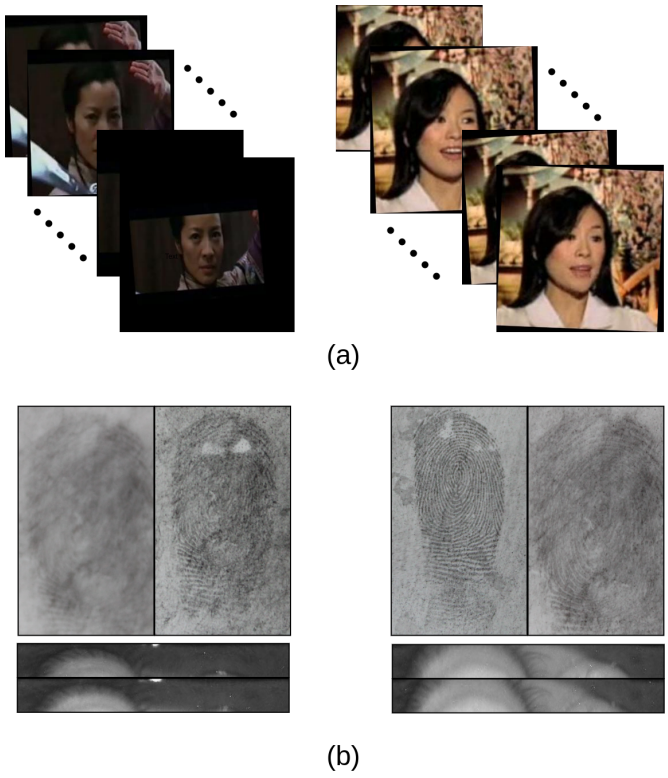

Fig. 16: Failed verification for the third multimodal dataset, (a) false rejection at FAR = 5×10^{-4} for the pairs of video frames and (b) correct verification at the same threshold when the iris and fingerprint samples are added to the multimodal sample sets.

5 Conclusions

In this paper, we presented a quality-aware fusion with feature-level fusion for multi-sample multimodal recognition using modalities of face, iris, and fingerprint consisting of two blocks to handle different multi-sample multimodal scenarios. The first block provides high-level representation of each modality considering the quality of the samples in that modality. The second block provides a multimodal representation for the input multimodal sample set, while utilizing the inter-modality quality of modalities. The network is trained using the proposed multimodal separability loss, while the multimodal network compactness loss alleviates the over-fitting caused by the over-parametrization of multimodal networks. To study the performance of the proposed framework, as well as loss functions proposed in this paper, we consider three real-world multimodal and eight unimodal datasets. The proposed framework is trained end-to-end in a weakly-supervised fashion without any quality score supervision. The expectation of the inter-modality quality on each multimodal test dataset represents the importance of the modalities in the recognition task and is aligned with the unimodal performance of the framework. Compared to state-of-the-art algorithms, we demonstrated that the proposed architecture significantly improves the multimodal performance by estimating the quality of the samples. The proposed quality-aware fusion can be adapted to different preprocessing algorithms, feature extraction methods, and loss functions for recognition and several other multimodal applications such as access control security system, passport control, and unlocking the smartphones.

ACKNOWLEDGEMENT

This material is based upon a work supported by the Center for Identification Technology Research and the National Science

Foundation under Grant #1650474.

REFERENCES

[1] S. Crihalmeanu, A. Ross, S. Schuckers, and L. Hornak, "A protocol for multibiometric data acquisition, storage and dissemination," *Technical Report, WVU, Lane Department of Computer Science and Electrical Engineering*, 2007.

[2] M. Haghighat, M. Abdel-Mottaleb, and W. Alhalabi, "Discriminant correlation analysis: Real-time feature level fusion for multimodal biometric recognition," *IEEE Transactions on Information Forensics and Security*, vol. 11, no. 9, pp. 1984–1996, 2016.

[3] S. Bahrampour, N. M. Nasrabadi, A. Ray, and W. K. Jenkins, "Multimodal task-driven dictionary learning for image classification," *IEEE transactions on Image Processing*, vol. 25, no. 1, pp. 24–38, 2016.

[4] S. Shekhar, V. M. Patel, N. M. Nasrabadi, and R. Chellappa, "Joint sparse representation for robust multimodal biometrics recognition," *IEEE transactions on Pattern Analysis and Machine Intelligence*, vol. 36, no. 1, pp. 113–126, 2014.

[5] K. Sundararajan and D. L. Woodard, "Deep learning for biometrics: A survey," *ACM Computing Surveys (CSUR)*, vol. 51, no. 3, pp. 1–34, 2018.

[6] M. Singh, R. Singh, and A. Ross, "A comprehensive overview of biometric fusion," *Information Fusion*, 2019.

[7] H. Jaafar and D. A. Ramli, "A review of multibiometric system with fusion strategies and weighting factor," *International Journal of Computer Science Engineering (IJCSE)*, vol. 2, no. 4, pp. 158–165, 2013.

[8] C.-A. Toli and B. Preneel, "A survey on multimodal biometrics and the protection of their templates," in *IFIP International Summer School on Privacy and Identity Management*, 2014, pp. 169–184.

[9] A. Nagar, K. Nandakumar, and A. K. Jain, "Multibiometric cryptosystems based on feature-level fusion," *IEEE transactions on information forensics and security*, vol. 7, no. 1, pp. 255–268, 2012.

[10] G. Goswami, P. Mittal, A. Majumdar, M. Vatsa, and R. Singh, "Group sparse representation based classification for multi-feature multimodal biometrics," *Information Fusion*, vol. 32, pp. 3–12, 2016.

[11] T.-Y. Lin, A. RoyChowdhury, and S. Maji, "Bilinear CNN models for fine-grained visual recognition," in *Proceedings of the IEEE International Conference on Computer Vision*, 2015, pp. 1449–1457.

[12] A. R. Chowdhury, T.-Y. Lin, S. Maji, and E. Learned-Miller, "One-to-many face recognition with bilinear cnns," in *IEEE Winter Conference on Applications of Computer Vision (WACV)*. IEEE, 2016, pp. 1–9.

[13] Y. Gao, O. Beijbom, N. Zhang, and T. Darrell, "Compact bilinear pooling," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 317–326.

[14] J.-B. Delbrouck and S. Dupont, "Multimodal compact bilinear pooling for multimodal neural machine translation," *arXiv preprint arXiv:1703.08084*, 2017.

[15] A. Fukui, D. H. Park, D. Yang, A. Rohrbach, T. Darrell, and M. Rohrbach, "Multimodal compact bilinear pooling for visual question answering and visual grounding," *arXiv preprint arXiv:1606.01847*, 2016.

[16] Y. Liu, J. Yan, and W. Ouyang, "Quality aware network for set to set recognition," in *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017, pp. 4694–4703.

[17] J. Yang, P. Ren, D. Zhang, D. Chen, F. Wen, H. Li, and G. Hua, "Neural aggregation network for video face recognition," in *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017, pp. 5216–5225.

[18] J. Zhao, J. Han, and L. Shao, "Unconstrained face recognition using a set-to-set distance measure on deep learned features," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 28, no. 10, pp. 2679–2689, 2017.

[19] J. Si, H. Zhang, C.-G. Li, J. Kuen, X. Kong, A. C. Kot, and G. Wang, "Dual attention matching network for context-aware feature sequence based person re-identification," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 5363–5372.

[20] K. Xu, J. Ba, R. Kiros, K. Cho, A. Courville, R. Salakhudinov, R. Zemel, and Y. Bengio, "Show, attend and tell: Neural image caption generation with visual attention," in *International conference on machine learning*, 2015, pp. 2048–2057.

[21] S. Soleymani, A. Torfi, J. Dawson, and N. M. Nasrabadi, "Generalized bilinear deep convolutional neural networks for multimodal biometric identification," in *2018 25th IEEE International Conference on Image Processing (ICIP)*, 2018, pp. 763–767.

[22] A. Gangwar and A. Joshi, "Deepirisnet: Deep iris representation with applications in iris recognition and cross-sensor iris recognition," in *IEEE International Conference on Image Processing (ICIP)*, 2016, pp. 2301–2305.

[23] R. F. Nogueira, R. de Alencar Lotufo, and R. C. Machado, "Fingerprint liveness detection using convolutional neural networks," *IEEE Transactions on Information Forensics and Security*, vol. 11, no. 6, pp. 1206–1213, 2016.

[24] H. Mehrotra, R. Singh, M. Vatsa, and B. Majhi, "Incremental granular relevance vector machine: A case study in multimodal biometrics," *Pattern Recognition*, vol. 56, pp. 63–76, 2016.

[25] N. Poh, A. Ross, W. Lee, and J. Kittler, "A user-specific and selective multimodal biometric fusion strategy by ranking subjects," *Pattern Recognition*, vol. 46, no. 12, pp. 3341–3357, 2013.

[26] R. Connaughton, K. W. Bowyer, and P. J. Flynn, "Fusion of face and iris biometrics," in *Handbook of Iris Recognition*, 2013, pp. 219–237.

[27] M. M. Monwar and M. L. Gavrilova, "Multimodal biometric system using rank-level fusion approach," *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics)*, vol. 39, no. 4, pp. 867–878, 2009.

[28] Z. Liu, S. Wang, L. Zheng, and Q. Tian, "Robust imagegraph: Rank-level feature fusion for image search," *IEEE Transactions on Image Processing*, vol. 26, no. 7, pp. 3128–3141, 2017.

[29] A. Kumar and S. Shekhar, "Personal identification using multibiometrics rank-level fusion," *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)*, vol. 41, no. 5, pp. 743–752, 2011.

[30] H. S. Bhatt, R. Singh, and M. Vatsa, "On recognizing faces in videos using clustering-based re-ranking and fusion," *IEEE Transactions on Information Forensics and Security*, vol. 9, no. 7, pp. 1056–1068, 2014.

[31] S. Kamyab and M. Eftekhari, "Feature selection using multimodal optimization techniques," *Neurocomputing*, vol. 171, pp. 586–597, 2016.

[32] L. Nie, X. Wang, J. Zhang, X. He, H. Zhang, R. Hong, and Q. Tian, "Enhancing micro-video understanding by harnessing external sounds," in *Proceedings of the 25th ACM international conference on Multimedia*, 2017, pp. 1192–1200.

[33] M. Faundez-Zanuy, "Data fusion in biometrics," *IEEE Aerospace and Electronic Systems Magazine*, vol. 20, no. 1, pp. 34–38, 2005.

[34] A. A. Ross and R. Govindarajan, "Feature level fusion of hand and face biometrics," in *Defense and Security*. International Society for Optics and Photonics, 2005, pp. 196–204.

[35] M. Eshwarappa and M. V. Latte, "Multimodal biometric person authentication using speech, signature and handwriting features," *International Journal of Advanced Computer Science and Applications, Special Issue on Artificial Intelligence*, 2011.

[36] C. Liu and H. Wechsler, "A shape-and texture-based enhanced fisher classifier for face recognition," *IEEE transactions on image processing*, vol. 10, no. 4, pp. 598–608, 2001.

[37] J. Yang, J.-y. Yang, D. Zhang, and J.-f. Lu, "Feature fusion: parallel strategy vs. serial strategy," *Pattern recognition*, vol. 36, no. 6, pp. 1369–1381, 2003.

[38] Q.-S. Sun, S.-G. Zeng, Y. Liu, P.-A. Heng, and D.-S. Xia, "A new method of feature fusion and its application in image recognition," *Pattern Recognition*, vol. 38, no. 12, pp. 2437–2448, 2005.

[39] S. Soleymani, A. Dabouei, H. Kazemi, J. Dawson, and N. M. Nasrabadi, "Multi-level feature abstraction from convolutional neural networks for multimodal biometric identification," in *2018 24th International Conference on Pattern Recognition (ICPR)*, 2018, pp. 3469–3476.

[40] Y. Shi and R. Hu, "Rule-based feasibility decision method for big data structure fusion: Control method." *International Journal of Simulation–Systems, Science & Technology*, vol. 17, no. 31, 2016.

[41] N. Poh and J. Kittler, "A unified framework for biometric expert fusion incorporating quality measures," *IEEE transactions on pattern analysis and machine intelligence*, vol. 34, no. 1, pp. 3–18, 2012.

[42] J. Bigun, J. Fiérrez-Aguilar, J. Ortega-Garcia, and J. Gonzalez-Rodriguez, "Multimodal biometric authentication using quality signals in mobile communications," in *12th Int. Conf. Image Analysis and Processing*, 2003, p. 2.

[43] N. Poh, J. Kittler, and T. Bourlai, "Quality-based score normalization with device qualitative information for multimodal biometric fusion," *IEEE Transactions on Systems, Man, and Cybernetics-Part A: Systems and Humans*, vol. 40, no. 3, pp. 539–554, 2010.

[44] A. Kumar and D. Zhang, "Improving biometric authentication performance from the user quality," *IEEE transactions on instrumentation and measurement*, vol. 59, no. 3, pp. 730–735, 2010.

[45] S. Bharadwaj, H. S. Bhatt, R. Singh, M. Vatsa, and A. Noore, "Qfuse: Online learning framework for adaptive biometric system," *Pattern Recognition*, vol. 48, no. 11, pp. 3428–3439, 2015.

[46] K. Vishi and A. Jøsang, "A new approach for multi-biometric fusion based on subjective logic," in *Proceedings of the 1st International Conference on Internet of Things and Machine Learning*, 2017, p. 68.

[47] N. Khiari-Hili, C. Montagne, S. Lelandais, and K. Hamrouni, "Quality dependent multimodal fusion of face and iris biometrics," in *Image Processing Theory Tools and Applications (IPTA), 2016 6th International Conference on*, 2016, pp. 1–6.

[48] D. Sivasankaran, M. Ragab, T. Sim, and Y. Zick, "Context-aware fusion for continuous biometric authentication," in *2018 International Conference on Biometrics (ICB)*, 2018, pp. 233–240.

[49] W. Liu, Y. Wen, Z. Yu, M. Li, B. Raj, and L. Song, "Sphereface: Deep hypersphere embedding for face recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 212–220.

[50] J. Deng, J. Guo, N. Xue, and S. Zafeiriou, "Arcface: Additive angular margin loss for deep face recognition," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2019, pp. 4690–4699.

[51] Y. Duan, J. Lu, and J. Zhou, "Uniformface: Learning deep equidistributed representation for face recognition," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2019, pp. 3415–3424.

[52] W. Liu, Y. Wen, Z. Yu, and M. Yang, "Large-margin softmax loss for convolutional neural networks." in *ICML*, vol. 2, no. 3, 2016, p. 7.

[53] Y. Zheng, D. K. Pal, and M. Savvides, "Ring loss: Convex feature normalization for face recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 5089–5097.

[54] 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 *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 5265–5274.

[55] W. Liu, R. Lin, Z. Liu, L. Liu, Z. Yu, B. Dai, and L. Song, "Learning towards minimum hyperspherical energy," in *Advances in Neural Information Processing Systems*, 2018, pp. 6222–6233.

[56] R. Lin, W. Liu, Z. Liu, C. Feng, Z. Yu, J. M. Rehg, L. Xiong, and L. Song, "Compressive hyperspherical energy minimization," *arXiv preprint arXiv:1906.04892*, 2019.

[57] A. Dabouei, H. Kazemi, S. M. Iranmanesh, J. Dawson, and N. M. Nasrabadi, "Fingerprint distortion rectification using deep convolutional neural networks," in *International Conference on Biometrics*, 2018.

[58] A. Dabouei, S. Soleymani, H. Kazemi, S. M. Iranmanesh, J. Dawson, and N. M. Nasrabadi, "ID preserving generative adversarial network for partial latent fingerprint reconstruction," in *2018 IEEE 9th International Conference on Biometrics Theory, Applications and Systems (BTAS)*. IEEE, 2018, pp. 1–10.

[59] S. Ioffe and C. Szegedy, "Batch normalization: Accelerating deep network training by reducing internal covariate shift," *arXiv preprint arXiv:1502.03167*, 2015.

[60] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov, "Dropout: a simple way to prevent neural networks from overfitting," *The Journal of Machine Learning Research*, vol. 15, no. 1, pp. 1929–1958, 2014.

[61] T. Salimans and D. P. Kingma, "Weight normalization: A simple reparameterization to accelerate training of deep neural networks," in *Advances in Neural Information Processing Systems*, 2016, pp. 901–909.

[62] Y. Wu and K. He, "Group normalization," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018, pp. 3–19.

[63] A. Brock, T. Lim, J. M. Ritchie, and N. J. Weston, "Neural photo editing with introspective adversarial networks," in *5th International Conference on Learning Representations 2017*, 2017.

[64] W. Liu, Y.-M. Zhang, X. Li, Z. Yu, B. Dai, T. Zhao, and L. Song, "Deep hyperspherical learning," in *Advances in neural information processing systems*, 2017, pp. 3950–3960.

[65] P. Xie, J. Zhu, and E. Xing, "Diversity-promoting Bayesian learning of latent variable models," in *International Conference on Machine Learning*, 2016, pp. 59–68.

[66] W. Liu, Z. Liu, Z. Yu, B. Dai, R. Lin, Y. Wang, J. M. Rehg, and L. Song, "Decoupled networks," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 2771–2779.

[67] P. Xie, A. Singh, and E. P. Xing, "Uncorrelation and evenness: a new diversity-promoting regularizer," in *Proceedings of the 34th International Conference on Machine Learning-Volume 70*, 2017, pp. 3811–3820.

[68] M. Cogswell, F. Ahmed, R. Girshick, L. Zitnick, and D. Batra, "Reducing overfitting in deep networks by decorrelating representations," in *International Conference on Learning Representations*, 2016.

[69] W. Wang, D. Tran, and M. Feiszli, "What makes training multi-modal networks hard?" *arXiv preprint arXiv:1905.12681*, 2019.

[70] A. K. Jain, S. Prabhakar, L. Hong, and S. Pankanti, "Filterbank-based fingerprint matching," *IEEE transactions on Image Processing*, vol. 9, no. 5, pp. 846–859, 2000.

[71] B. F. Klare, B. Klein, E. Taborsky, A. Blanton, J. Cheney, K. Allen, P. Grother, A. Mah, and A. K. Jain, "Pushing the frontiers of unconstrained face detection and recognition: IARPS Janus benchmark A," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 1931–1939.

[72] K. Zhang, Z. Zhang, Z. Li, and Y. Qiao, "Joint face detection and alignment using multitask cascaded convolutional networks," *IEEE Signal Processing Letters*, vol. 23, no. 10, pp. 1499–1503, 2016.

[73] N. Othman, B. Dorizzi, and S. Garcia-Salicetti, "Osiris: An open source iris recognition software," *Pattern Recognition Letters*, vol. 82, pp. 124–131, 2016.

[74] S. Chikkerur, C. Wu, and V. Govindaraju, "A systematic approach for feature extraction in fingerprint images," *Biometric Authentication*, pp. 1–23, 2004.

[75] Q. Cao, L. Shen, W. Xie, O. M. Parkhi, and A. Zisserman, "Vggface2: A dataset for recognising faces across pose and age," in *2018 13th IEEE International Conference on Automatic Face & Gesture Recognition (FG)*. IEEE, 2018, pp. 67–74.

[76] F. Alonso-Fernandez, R. A. Farrugia, and J. Bigun, "Very low-resolution iris recognition via eigen-patch super-resolution and matcher fusion," in *2016 IEEE 8th International Conference on Biometrics Theory, Applications and Systems (BTAS)*, 2016, pp. 1–8.

[77] X. Si, J. Feng, J. Zhou, and Y. Luo, "Detection and rectification of distorted fingerprints," *IEEE transactions on pattern analysis and machine intelligence*, vol. 37, no. 3, pp. 555–568, 2015.

[78] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.

[79] K. Ko, "Users guide to export controlled distribution of nist biometric image software (nbis-ec)," *NIST Interagency/Internal Report (NISTIR)-7391*, 2007.

[80] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "Imagenet: A large-scale hierarchical image database," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2009, pp. 248–255.

[81] K. He, X. Zhang, S. Ren, and J. Sun, "Delving deep into rectifiers: Surpassing human-level performance on imagenet classification," in *Proceedings of the IEEE international conference on computer vision*, 2015, pp. 1026–1034.

[82] L. Wolf, T. Hassner, and I. Maoz, "Face recognition in unconstrained videos with matched background similarity," in *Proceedings of the IEEE conference on computer vision and pattern recognition (CVPR)*, 2011, pp. 529—534.

[83] "Casia v4 iris database, http://biometrics.idealtest.org/dbdetailforuser.do?id=4." [Online]. Available: http://biometrics.idealtest.org/dbDetailForUser.do?id=4

[84] A. Sankaran, M. Vatsa, and R. Singh, "Multisensor optical and latent fingerprint database," *IEEE access*, vol. 3, pp. 653–665, 2015.

[85] A. Sankaran, T. I. Dhamecha, M. Vatsa, and R. Singh, "On matching latent to latent fingerprints," in *2011 international joint conference on biometrics (IJCB)*, 2011, pp. 1–6.

[86] C. Velardo and J.-L. Dugelay, "Improving identification by pruning: A case study on face recognition and body soft biometric," in *2012 13th International Workshop on Image Analysis for Multimedia Interactive Services*, 2012, pp. 1–4.

[87] P. Lopes Silva, E. Luz, G. Moreira, L. Moraes, and D. Menotti, "Chimerical dataset creation protocol based on doddington zoo: A biometric application with face, eye, and ecg," *Sensors*, vol. 19, no. 13, p. 2968, 2019.

[88] S. C. Dass, K. Nandakumar, and A. K. Jain, "A principled approach to score level fusion in multimodal biometric systems," in *International conference on audio-and video-based biometric person authentication*, 2005, pp. 1049–1058.

[89] R. Primorac, R. Togneri, M. Bennamoun, and F. Sohel, "Generalized joint sparse representation for multimodal biometric fusion of heterogeneous features," in *2018 IEEE International Workshop on Information Forensics and Security (WIFS)*, 2018, pp. 1–6.

[90] Z. Sun and T. Tan, "Ordinal measures for iris recognition," *IEEE Transactions on pattern analysis and machine intelligence*, vol. 31, no. 12, pp. 2211–2226, 2008.

[91] Z. Zhao and A. Kumar, "Towards more accurate iris recognition using deeply learned spatially corresponding features," in *Proceedings of the IEEE International Conference on Computer Vision*, 2017, pp. 3809–3818.
[92] "Verifinger, www.neurotechnology.com/verifinger.html." [Online]. Available: VeriFinger.www.neurotechnology.com/verifinger.html
[93] K. Wang and A. Kumar, "Towards more accurate iris recognition using dilated residual features," *IEEE Transactions on Information Forensics and Security*, 2019.
[94] I. Kemelmacher-Shlizerman, S. M. Seitz, D. Miller, and E. Brossard, "The MegaFace benchmark: 1 million faces for recognition at scale," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 4873–4882.
[95] G. B. Huang, M. Mattar, T. Berg, and E. Learned-Miller, "Labeled faces in the wild: A database forstudying face recognition in unconstrained environments," in *Workshop on faces in'Real-Life'Images: detection, alignment, and recognition*, 2008.
[96] M. Sultana, P. P. Paul, and M. L. Gavrilova, "Social behavioral information fusion in multimodal biometrics," *IEEE Transactions on Systems, Man, and Cybernetics: Systems*, vol. 48, no. 12, pp. 2176–2187, 2017.
[97] L. Tran, X. Yin, and X. Liu, "Disentangled representation learning GAN for pose-invariant face recognition," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 1415–1424.
[98] S. Sankaranarayanan, A. Alavi, C. D. Castillo, and R. Chellappa, "Triplet probabilistic embedding for face verification and clustering," in *2016 IEEE 8th international conference on biometrics theory, applications and systems (BTAS)*, 2016, pp. 1–8.
[99] N. Crosswhite, J. Byrne, C. Stauffer, O. Parkhi, Q. Cao, and A. Zisserman, "Template adaptation for face verification and identification," *Image and Vision Computing*, vol. 79, pp. 35–48, 2018.
[100] Y. Taigman, M. Yang, M. Ranzato, and L. Wolf, "Deepface: Closing the gap to human-level performance in face verification," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2014, pp. 1701–1708.
[101] B.-N. Kang, Y. Kim, and D. Kim, "Pairwise relational networks for face recognition," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018, pp. 628–645.
[102] A. Mittal, A. K. Moorthy, and A. C. Bovik, "Blind/referenceless image spatial quality evaluator," in *2011 conference record of the forty fifth asilomar conference on signals, systems and computers (ASILOMAR)*, 2011, pp. 723–727.
[103] O. Bausinger and E. Tabassi, "Fingerprint sample quality metric nfiq 2.0," *BIOSIG 2011–Proceedings of the Biometrics Special Interest Group*, 2011.
[104] Y. Tokozume, Y. Ushiku, and T. Harada, "Between-class learning for image classification," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 5486–5494.
[105] H. Zhang, M. Cisse, Y. N. Dauphin, and D. Lopez-Paz, "mixup: Beyond empirical risk minimization," *arXiv preprint arXiv:1710.09412*, 2017.

Ali Dabouei received his master of science in electrical engineering from Sharif University of Technology, Tehran, Iran. He has been pursuing his doctoral studies since 2017 at West Virginia University, USA. His area of research includes machine learning, deep learning, and their applications in computer vision and biometrics. His research received best student paper awards in IEEE International Conference on Biometrics (BTAS), 2018, and IAPR International Conference on Biometrics (ICB) 2018. He has authored over 20 publications including journals and peer-reviewed conferences. He has also served as the reviewer for prestigious venues such as IEEE Transactions on Neural Networks and Learning Systems (TNNLS), International Conference on Computer Vision and Pattern Recognition (CVPR), and International Conference on Computer Vision (ICCV).

Fariborz Taherkhani is currently a postdoctoral researcher at the Robotics Institute in Carnegie Mellon University, Pittsburgh, PA, USA. He received his PhD in computer science from West Virginia University, Morgantown, WV, USA. He received his BSc. degree in computer engineering from National University of Iran, Tehran, Iran, and his MSc. degree in computer engineering from Sharif University of Technology, Tehran, Iran. His research interests include machine learning, computer vision, biometrics, and image retrieval.

Seyed Mehdi Iranmanesh received B.Sc. and M.Sc. degrees in Computer engineering and Artificial Intelligence from Iran University of Science and Technology, Tehran, Iran, in 2008 and 2011, respectively. He graduated in computer science from West Virginia University in 2021. His research interests include computer vision and machine learning.

Sobhan Soleymani received his B.Sc. degree in Electrical Engineering from University of Tehran and his M.Sc. degree in Electrical Engineering from École Polytechnique Fédérale de Lausanne (EPFL). He is graduated with the Ph.D. degree in Electrical Engineering from the Lane Department of Computer Science and Electrical Engineering, West Virginia University in 2021. He has served as reviewer for IEEE Transactions on Biometrics, Behavior, and Identity Science (TBIOM), IEEE Transactions on Neural Networks and Learning Systems (TNNLS), Neural Information Processing Systems (NeurIPS), International Conference on Learning Representations (ICLR), International Joint Conference on Biometrics (IJCB), and Winter Conference on Applications of Computer Vision (WACV). His areas of interest are computer vision, machine learning, signal and image processing, and their application in biometrics.

Jeremy Dawson, Associate Professor, joined the Lane Department of Computer Science and Electrical Engineering in the Fall of 2015. His background is in microelectronics and nanophotonics, and he has extensive experience in complex, multi-domain system integration and hardware system implementation. His current biometrics research efforts are focused on the creation of large-scale biometrics datasets that can be used to evaluate sensor operation and other human factors, as well as apply novel machine learning algorithms to solve human identification challenges. Dr. Dawson also has extensive experience in micro and nanophotonic sensor platforms. His research in biosensors led to the identification of a need for new signal processing methodologies for DNA systems, which resulted in the first molecular biometrics (DNA) project funded through the WVU Center for Identification Technology Research (CITeR), an NSF IUCRC.

Nasser M. Nasrabadi (S'80 – M'84 – SM'92 – F'01) received the B.Sc. (Eng.) and Ph.D. degrees in electrical engineering from the Imperial College of Science and Technology, University of London, London, U.K., in 1980 and 1984, respectively. In 1984, he was with IBM, U.K., as a Senior Programmer. From 1985 to 1986, he was with the Philips Research Laboratory, New York, NY, USA, as a member of the Technical Staff. From 1986 to 1991, he was an Assistant Professor with the Department of Electrical Engineering, Worcester Polytechnic Institute, Worcester, MA, USA. From 1991 to 1996, he was an Associate Professor with the Department of Electrical and Computer Engineering, State University of New York at Buffalo, Buffalo, NY, USA. From 1996 to 2015, he was a Senior Research Scientist with the U.S. Army Research Laboratory. Since 2015, he has been a Professor with the Lane Department of Computer Science and Electrical Engineering. His current research interests are in image processing, computer vision, biometrics, statistical machine learning theory, sparsity, robotics, and neural networks applications to image processing. He is a fellow of ARL and SPIE and has served as an Associate Editor for the IEEE Transactions on Image Processing, the IEEE Transactions on Circuits, Systems and Video Technology, and the IEEE Transactions on Neural Networks.