IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received 28 June 2023, accepted 28 July 2023, date of publication 2 August 2023, date of current version 9 August 2023.

Digital Object Identifier 10.1109/ACCESS.2023.3301150

RESEARCH ARTICLE

On the Representation Learning of Conditional Biometrics for Flexible Deployment

TIONG-SIK NG[1], CHENG-YAW LOW[2], JACKY CHEN LONG CHAI[1], AND ANDREW BENG JIN TEOH[1], (Senior Member, IEEE)

[1]School of Electrical and Electronics Engineering, College of Engineering, Yonsei University, Seoul 03722, South Korea
[2]Center for Mathematical and Computational Sciences, Data Science Group, Institute for Basic Science, Daejeon 34126, South Korea

Corresponding author: Andrew Beng Jin Teoh (bjteoh@yonsei.ac.kr)

This work was supported in part by the National Research Foundation of Korea (NRF) funded by the Korean Government (MSIP) under Grant NRF-2022R1A2C1010710, and in part by the Hyundai Motor Chung Mong-Koo Global Scholarship.

ABSTRACT Unimodal biometric systems are commonplace nowadays. However, there remains room for performance improvement. Multimodal biometrics, i.e., the combination of more than one biometric modality, is one of the promising remedies; yet, there lie various limitations in deployment, e.g., availability, template management, deployment cost, etc. In this paper, we propose a new notion dubbed Conditional Biometrics representation for flexible biometrics deployment, whereby a biometric modality is utilized to condition another for representation learning. We demonstrate the proposed conditioned representation learning on the face and periocular biometrics via a deep network dubbed the Conditional Biometrics Network. Our proposed Conditional Biometrics Network is a representation extractor for unimodal, multimodal, and cross-modal matching during deployment. Our experimental results on five in-the-wild periocular-face datasets demonstrate that the network outperforms their respective baselines for identification and verification tasks in all deployment scenarios.

INDEX TERMS Conditional biometrics, face, flexible matching, periocular, representation learning.

I. INTRODUCTION

Biometrics are associated with a subject's identity, pertaining to how these biological traits are unique to each person. For instance, fingerprint, face, and iris are the three most popular biometric traits for commercial deployment [1]. Biometric systems consisting of only a single modality for deployment are better known as unimodal biometrics. Despite its convenience, unimodal biometrics usually suffers from under-performance [2]. This is due to the stochastic nature of biometric signals, [3], which is attributed to different angles, lighting conditions, noisy environments, and whatnot; implementing a biometric system is not a simple feat.

To address this issue, multimodal biometrics that utilizes multiple biometric modalities emerged. Multiple biometric modalities can be fused at representation, score, decision, or rank-level [4], [5]. Though it is well-proven that multimodal biometrics can drastically improve accuracy performance, various issues exist during its deployment. One main issue would be the storage of multiple biometric templates, requiring higher management costs. Another major issue would be the biometrics availability during the query stage, e.g., voice may affect the multimodal biometric system's performance if the subject's voice is muffled due to the subject's well-being. Additionally, using multiple biometric modalities may require a subject's active cooperation, while certain combinations may not be realistically available, such as gait and iris.

Another possible deployment mode of biometrics is cross-modal biometrics, a new notion that balances the pros and cons of unimodal and multimodal biometrics. For instance, suppose that the face template is enrolled, and another biometric modality, such as voice, is used for the query. Therefore, unlike unimodal or multimodal biometrics, which require the presence of the same modality for matching,

The associate editor coordinating the review of this manuscript and approving it for publication was Zahid Akhtar.

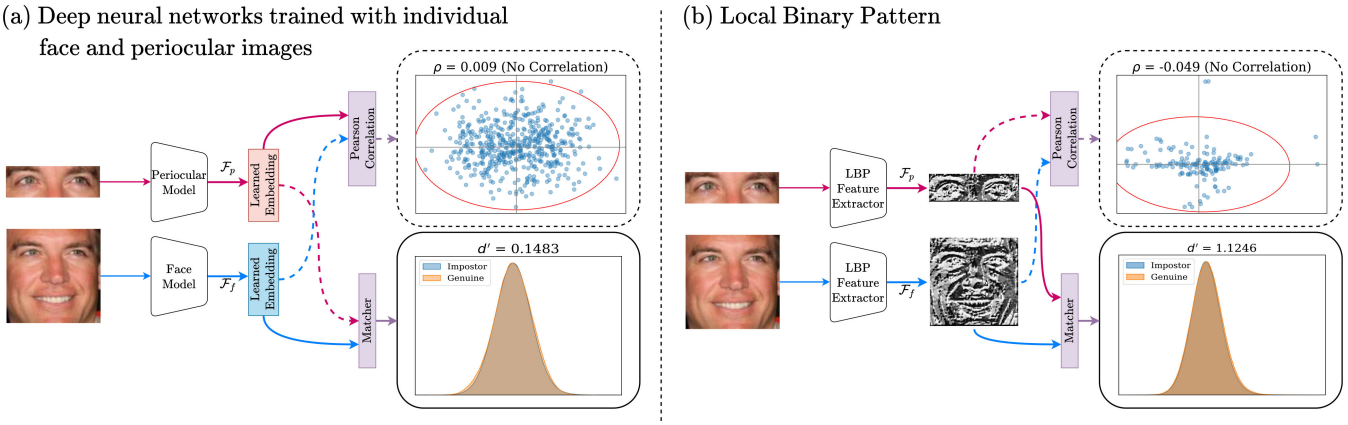


FIGURE 1. Evaluation of cross-modal biometrics matching based on (a) learning-based representation (deep network) and (b) hand-crafted feature extractor (LBP) in terms of genuine-impostor score distribution [7] and Pearson correlation [8] plot. (a) For the separately-trained CNNs with individual face and periocular images, periocular representation ($\mathcal{F}_p$) shows almost zero correlation ($\rho = 0.01$) with face representation ($\mathcal{F}_f$). Furthermore, a low genuine-impostor separation score, d' of 0.15, is elicited. (b) It is noted that despite the LBP-extracted features being visually similar (i.e., periocular feature visually being a subset of the face), a near-zero correlation ($\rho = -0.05$) indicates that both faces and periocular are of different modalities. Also, the genuine-impostor separation score d' is as low as 1.12.

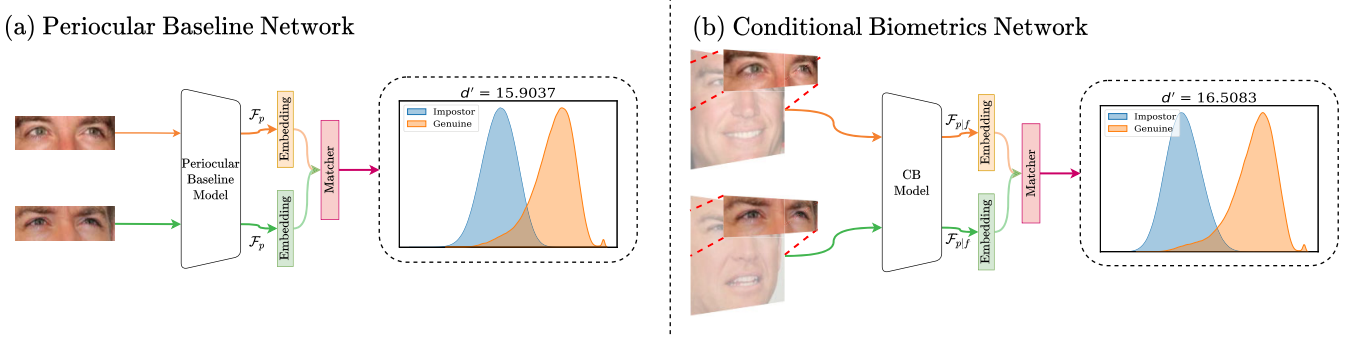


FIGURE 2. Evaluation of periocular biometrics performance with CB-enabled instance vs. without counterpart (baseline) in the genuine-impostor matching score distributions. (a) Periocular representations from its baseline network have a lower genuine-impostor separation score, d', of 15.90. (b) Periocular representations conditioned by face biometrics have a higher genuine-impostor separation score, d', of 16.51.

cross-modal biometrics is flexible. Unfortunately, the accuracy would be far from satisfactory in directly matching two different biometric modalities despite being the same identity.

This paper considers the face and periocular biometrics as study subjects. Periocular, a peripheral region of the ocular area, is a somewhat weaker biometric modality [6] since periocular biometrics only contain information surrounding the peripheral area of a subject, as opposed to face biometrics which includes the complete facial representations. Periocular biometrics is helpful when a subject's complete facial representations are not available, such as the subject has make-up on, has performed facial surgery, or the subject's face is occluded [9], [10].

In Fig. 1, we depict the cross-matching between face and periocular biometrics that are processed by convolutional networks (CNN) and Local Binary Pattern (LBP) [11], a representative learning-based and hand-crafted feature extractor, respectively. We note the almost zero correlation between face and periocular features of the same identity i.e., no relation between two modalities, and the strong overlapping of genuine and imposter matching score distribution, which implies poor matching performance. These suggest that despite periocular images being considered to be a subset of the face i.e., is made up of the facial and ocular area while being of the same RGB domain; it is a distinctive biometric modality from the face. In addition, the results also indicate that cross-modal matching between face and periocular is unlikely.

In this paper, we propose a notion coined as Conditional Biometrics (CB) that strives to achieve performance gain in three biometrics deployment modes i.e., unimodal, multimodal, and cross-modal. The CB utilizes a biometric modality to condition another for representation learning. We demonstrate that the performance of periocular biometrics can be elevated remarkably when conditioned by the face. In Fig. 2, we depict that the CB-enabled periocular representation reveals better genuine and imposter matching score distribution separation compared to its sole periocular counterpart, which suggests the performance gain of the former. Parallelly, face recognition conditioned by periocular is equally helpful.

The CB-enabled multimodal and cross-modal biometric representation can also contribute to performance gain. As opposed to the low correlation shown in Fig. 1,

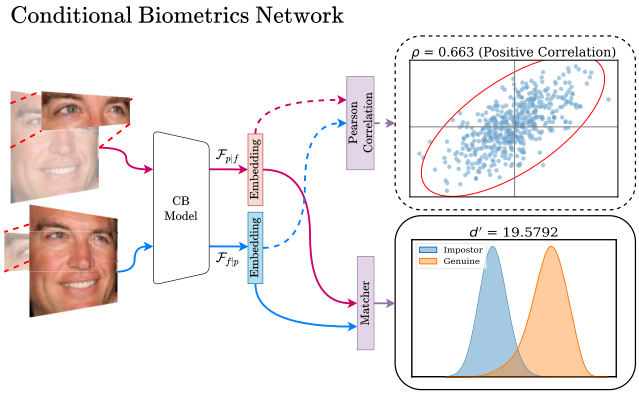


FIGURE 3. Evaluation of cross-modal matching on CB-enabled representation in genuine-impostor matching score distribution and Pearson correlation. CB-enabled periocular ($\mathcal{F}_{p|f}$) and face ($\mathcal{F}_{f|p}$) representations have high correlations ($\rho = 0.66$), while also having a much higher genuine-impostor score, d' of 19.58, compared to the ones shown in Fig. 1.

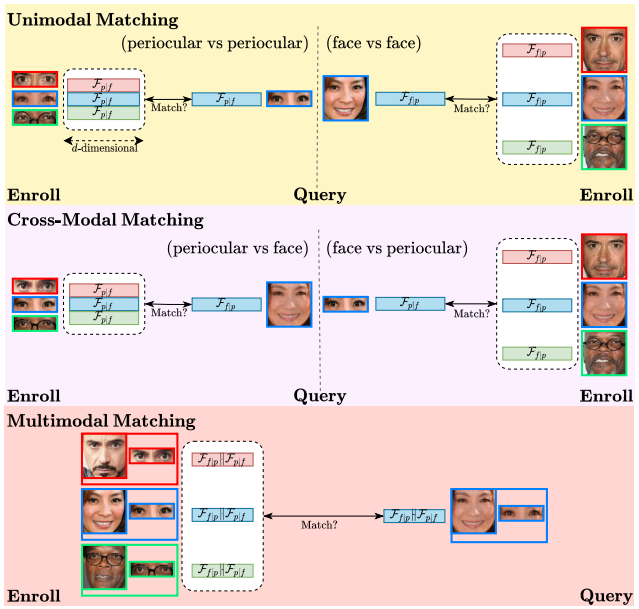


FIGURE 4. Matching modes supported by the CB regimen, including unimodal, cross-modal, and multimodal matching.

we demonstrate that the CB-enabled periocular and face representations have a high correlation with one another in Fig. 3, which implies the possibility of cross-modal matching. We demonstrate that the CB representation enables various flexible deployment modes, as shown in Fig. 4. In Fig. 4, the CB representation is deployable not only for the unimodal matching mode, including CB periocular vs. CB periocular and CB face vs. CB face, but also for cross-modal CB face vs. CB periocular matching. Besides, multimodal matching is also permitted by aggregating CB face and CB periocular representations.

The CB representation learning is substantiated by a CNN comprising a shared-parameter backbone encoder, appended with two classification heads for each face and periocular. In addition to classification losses, a regularized CB loss is devised to pull inter-modality and intra-subject examples closer and push intra-modality and inter-subject examples far apart in the embedding space.

TABLE 1. Matching setting in practical applications.

Matching Setting	Application Scenario
Unimodal	Suppose that a subject's representation (either face or periocular) is enrolled to the system, during the query stage, the subject's representation is extracted (*corresponding to the same modality* of the stored representation), and matching is performed between the queried and enrolled representations. The system then accepts or rejects the subject's query based on the matching score obtained.
Cross-Modal	Given that the subject's *face representation is enrolled* in the database, the subject's *periocular representation is then extracted during the query* stage. To reach a decision, the matching is done between the representations of different modalities (enrolled face and queried periocular). Cross-modal matching is also possible with the *enrolled periocular and queried face representations*, as long as they are of different modalities.
Multimodal	Both a subject's *face and periocular* representations are fused and stored together under the subject's identity during the enrollment stage. During the query stage, both the subject's *face and periocular* representations are fused in a similar manner, in which the matching is performed to obtain a decision.

We summarize the contributions of this paper as follows:

1) We introduce CB representation learning - a new means of representation learning mechanism by conditioning a biometric modality on another for performance gain and flexible deployment.
2) We propose the Conditional Biometrics Network (CB-Net) to realize the CB notion alongside a regularized CB loss. The CB-Net is a representation learning model that attracts examples of the same identity but different modalities while enabling the correlation between the learned representations, allowing the cross-modal matching task.
3) We benchmark the performance of CB-Net deployed in unimodal, cross-modal, and multimodal based upon five periocular-face in the wild datasets. We demonstrate that the CB-Net enhances the face and periocular discriminability for identification and verification tasks.

II. RELATED WORKS

This section presents the remarkable state-of-the-art approaches for periocular biometrics, multimodal biometrics, cross-modal biometrics, and other works relevant to the proposed CB notion.

A. PERIOCULAR BIOMETRICS

Early research on periocular biometrics relies on hand-crafted feature extraction methods. Among these techniques include

masking and filtering [12], Histogram of Orientation and Gradient (HOG) [13], Local Binary Pattern (LBP) [11], and Scale Invariant Feature Transform (SIFT) [14]. Though it is shown that these methods can achieve a good performance, the datasets collected through these methods are under a constrained environment, wherein there are no pose invariances nor illumination differences for all of the subjects in question [15], [16].

As deep learning has become the norm in the last decade, [17], the recent works for periocular recognition rely on CNNs for representation learning instead of hand-crafted feature descriptors. For instance, [18] focuses on extracting near-infrared (NIR) periocular representation, particularly mid-level ones. On the other hand, [19] uses a different approach to infrared, whereby cross-spectrum between visible light and infrared matching was performed using a twin shared-parameter CNN with attention. Reference [20] also utilized a shared-parameter CNN for periocular images. In this case, a subject's left and right eyes were considered instead to fuse the RGB periocular images. Also, different from the previously mentioned works, Tiong et al. attempted to solve the periocular in the wild representation problem, which is more challenging compared to periocular datasets in a controlled environment.

B. MULTIMODAL BIOMETRICS WITH PERIOCULAR

Due to the lack of discrimination power, the periocular trait is practically deployed in conjunction with other biometrics. In particular, the iris is usually fused with periocular biometrics [21], [22], [23], [24]. On the other hand, some other biometric modalities have also been considered, such as face [25], [26] and soft biometrics from facial representations [27].

In [25], a combination of periocular, face, and hand-crafted iris representations obtained from the mobile phone are fused at the score level. On the other hand, [26] used a multi-representation deep learning network in addition to texture descriptors to combine face and periocular modalities. Different from [25], representation level fusion was adopted instead, such that the correlation between the descriptor and the raw data renders a new representation.

In a more recent work [24], the combination of periocular and iris scores was learned via a hierarchical fusion network. The network is used to perform the fusion and can search for the best method for score fusion. Similarly, [22] also utilized the fusion of periocular and iris. However, an end-to-end neural network with a co-attention module was used to fuse the representations adaptively.

C. CROSS-MODAL MATCHING BIOMETRICS

Most works involving cross-modal matching of distinct biometric modalities typically revolve around matching the face and voice biometrics. For instance, the work by Nagrani et al. [28] explores the possibility of cross-modal verification to determine if the given face and voice inputs are from the same subject. In [29], the work considers information sharing between face and voice, different from the work above. Specifically, a Siamese network with contrastive loss was proposed so that both representations are learned in a shared space. In addition, it is shown that the projection of the representation was sufficient to perform matching.

Unlike the previous works, [30] proposed a work that performs cross-modal matching between visible light (VIS) face and near-infrared (NIR) face images. Notably, a subspace projection hashing was designed so that both VIS face and NIR face images are projected to a common subspace. The generated hashed codes via this projection method enable the network to perform matching in different domains with a vast performance improvement.

D. CONDITIONAL BIOMETRICS RELEVANT WORKS

The existing works close to the CB notion are usually associated with soft biometrics, i.e., attributes that barely capture a person's identity credentials, such as gender, age, skin color, etc. This is mainly attributed to soft biometrics being considered a weaker representation than typical biometric modalities such as the face. Reference [31] introduced attribute-aware loss such that the representation mapping with soft biometrics contributes to the performance gain of face biometrics. On the other hand, [32] proposed an adaptive margin-based angular loss that functions to leverage face recognition via soft biometrics. Soft biometrics is embedded into the margin of the loss function. Despite both works using soft biometrics to condition face biometrics, the reverse, i.e., using face to condition soft biometrics, is undoable, and so for cross-modal and CB multimodal matching.

In [33], the knowledge distillation [34] with label smoothing was used to enhance the performance of periocular biometrics via face. More specifically, a teacher-student network was utilized, whereby the teacher network was pre-trained with face images. Then, the teacher network functions to leverage the student network trained with periocular images. One drawback of this method would be that two distinct networks must be trained separately. In addition, the periocular network may not be used reversely to leverage and enhance the performance of the face network.

III. PROPOSED WORK

We deliberate in this section on the CB-Net network architecture and the regularized CB loss. Subsequently, we disclose the realistic deployment modes of the proposed CB systems.

A. CB-NET NETWORK ARCHITECTURE

The proposed CB-Net is comprised of a shared-parameter encoder $\mathcal{F}_\star(\boldsymbol{x};\phi)$ such that $\star = \{p|f, f|p\}$ where $\boldsymbol{x}$ is the input image, ϕ denotes the encoder parameter, and $\{p|f, f|p\}$ represent the periocular conditioned by face, and face conditioned by periocular notations, respectively. In this work,

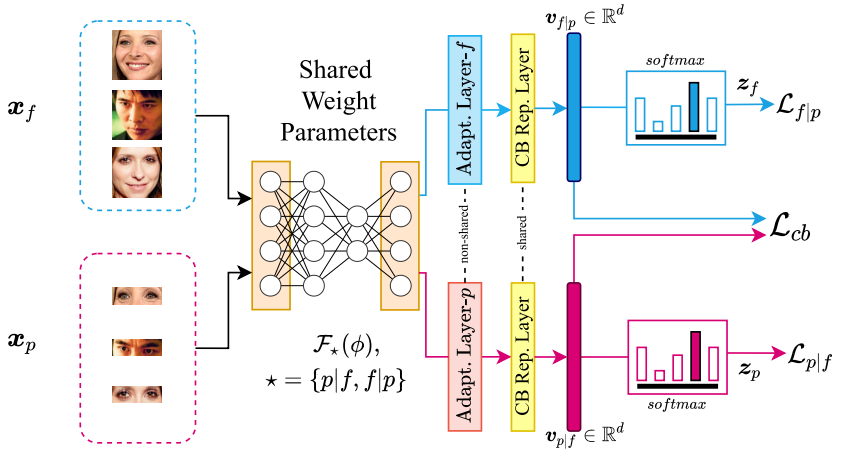


FIGURE 5. CB-Net architecture is a composition of a shared-parameter convolutional neural network, interleaved with non-shared adaptation layers (as face and periocular vary in input resolutions) and a shared CB representation layer. This is followed by two softmax heads for each face and periocular. The training loss comprises a CB-based loss term, a face conditioned by periocular, and a periocular conditioned by face classification losses.

we adopt MobileFaceNet [35] as the representation encoder. On the other hand, we append the network encoder with two softmax-based predictors. Each conditioned representation is associated with the corresponding prediction head as illustrated in Fig. 5.

This paper's image resolutions for periocular and face are 37 × 112 pixels and 112 × 112 pixels, respectively. We, therefore, interleave the network backbone with an individual (non-shared) adaptation layer for each modality to yield fixed-dimension vectors for learning the corresponding conditioned representation, $\boldsymbol{v}_\star \in \mathbb{R}^d$. It is noted that we set $d = 512$ throughout our experiments.

B. LOSS FUNCTIONS

1) NOTATION

Given a set of N face (f) and periocular (p) images with shared identity labels $\{(\boldsymbol{x}_{*i}, \boldsymbol{y}_i)|i = 1, \ldots, N\}$, $* = \{p, f\}$ of C identities, the softmax predictions can be computed via $\boldsymbol{z}_* = softmax(\mathbf{W}_*{}^T \mathcal{F}_\star(\boldsymbol{x}_*; \phi))$ where $\star = \{p|f, f|p\}$ and $\mathbf{W}_* \in \mathbb{R}^{C\times d}$ is the prototype weight matrix.

2) CLASSIFICATION LOSS FOR CONDITIONED FACE AND PERIOCULAR PREDICTORS

For each predictor $\boldsymbol{z}_*$, the CB-Net is trained with respect to the margin-based angular softmax loss e.g., CosFace [36] to be specific, as it is well-proven to enhance the inter-subject separation and reduce the intra-subject variations.

Given B batch samples of $\boldsymbol{x}_{*i}$ with its corresponding identity label $\boldsymbol{y}_i$, the margin-based angular loss, $\mathcal{L}_{\star_i}$ is defined as follows:

$$\mathcal{L}_{\star_i} = \frac{1}{B}\sum_{i=1} -\log \frac{e^{s(\cos(\theta_{y_{i,i}})-m)}}{e^{s(\cos(\theta_{y_{i,i}})-m)} + \sum_i e^{s\cos\theta_i}} \tag{1}$$

where θ_{*i} represent the angles between the L_2 normalized prototype weight vector $\hat{\mathbf{w}}_{*j} \in \mathbb{R}^d$ and the L_2 normalized $\hat{\boldsymbol{v}}_{\star_i}$ which are distributed on a hypersphere with radius s, and m represents the margin penalty, such that $\cos(\theta_j, i) = \hat{\mathbf{w}}_{*j}^T \hat{\boldsymbol{v}}_{\star_i}$. In this case, we set an equal value for both face and periocular scales and margins, so neither learning dominates one other during the training process.

3) CB LOSS WITH REGULARIZATION

The two classification losses attempt to learn an identity-wise representation for each face and periocular but neglect the modality gap between them. The CB loss is devised to reduce inter-modality and intra-subject discrepancies (specifically, face and periocular belonging to the same subject) and enhance intra-modality and inter-subject separation (e.g., periocular examples from different subjects).

Let $(\boldsymbol{v}_{\star_i}, \boldsymbol{y}_i)$ be an anchor example, such that $(\boldsymbol{v}^+_{\star_i}, \boldsymbol{y}^+_i)$ is an intra-subject example $\boldsymbol{v}^+_{\star_i}$ with an identity label $\boldsymbol{y}^+_i$, and $(\boldsymbol{v}^-_{\star_i}, \boldsymbol{y}^-_i)$ represents an inter-subject example $\boldsymbol{v}^-_{\star_i}$ labeled with $\boldsymbol{y}^-_i$, where $\boldsymbol{y}^+_i = \boldsymbol{y}_i$ and $\boldsymbol{y}^-_i \neq \boldsymbol{y}_i$. Given $\mathcal{B}$ samples of intra-subject and inter-subject periocular-face pairs for a mini-batch of the face and periocular representation. We define the conditional biometric (CB) loss $\mathcal{L}_{cb}$, using an angular contrastive loss, i.e., a special case of the supervised contrastive loss [37] as follows:

$$\mathcal{L}_{cb} = -\frac{1}{\mathcal{B}}\sum_{i=1}^{\mathcal{B}} \log \frac{e^{\left(\boldsymbol{v}_{\star_i}\boldsymbol{v}^+_{\star_i}/\tau\right)}}{e^{\left(\boldsymbol{v}_{\star_i}\boldsymbol{v}^+_{\star_i}/\tau\right)} + \Sigma_{\boldsymbol{y}^-_i \neq \boldsymbol{y}_i} e^{\left(\boldsymbol{v}_{\star_i}\boldsymbol{v}^-_{\star_i}/\tau\right)}} \tag{2}$$

, where $\boldsymbol{v}_{\star_i}\boldsymbol{v}^+_{\star_i}$ and $\boldsymbol{v}_{\star_i}\boldsymbol{v}^-_{\star_i}$ computes the Cosine similarity for the intra-subject and the inter-subject periocular-face representation pairs, respectively, and τ, is a temperature term. We aim to render a discriminative embedding space consisting of multiple representation modalities by rectifying the modality discrepancy between periocular and face (see Fig. 6). In other words, $\mathcal{L}_{cb}$ operates by attracting the intra-subject representation pairs close to each other while repelling the inter-subject representation pairs to be as far apart as possible, particularly those intra-subject and

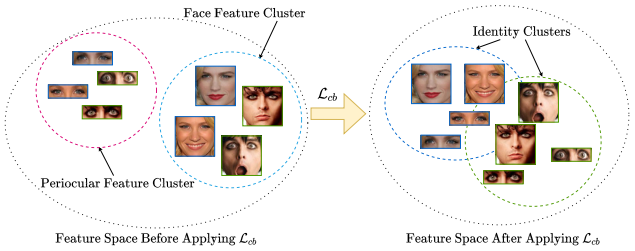


FIGURE 6. Learning paradigm of regularized CB loss $\mathcal{L}_{rcb}$, where modality discrepancy between face and periocular is remedied by an aggressive sampling of inter-modality intra-class and inter-class representation pairs, aside from intra-modality counterparts.

inter-subject pairs constituted by periocular and face (inter-modality). To this end, we introduce a regularization term to $\mathcal{L}_{cb}$ as follows:

$$\mathcal{L}_{rcb} = \mathcal{L}_{cb} + \gamma \sum_{i=1}^{\mathcal{B}} \left(e^{\left(1.0 - \boldsymbol{v}_{\star_i} \boldsymbol{v}^{+}_{\star_i} / \tau\right)^{\frac{1}{2}}} - e^{\left(1.0 - \boldsymbol{v}_{\star_i} \boldsymbol{v}^{-}_{\star_i}\right)} \right) \quad (3)$$

where γ is set to 0.001 in our experiments unless otherwise stated. We demonstrate in Section IV that $\mathcal{L}_{rcb}$ improves the CB-Net performance using optimizing the inter-modality intra-subject and inter-subject variabilities. This is mainly reflected in the Equal Error Rate (EER) and cross-modal matching. We single out the hardest inter-subject representation pairs in our formulation using hard negative mining [38].

Given only the classification losses, the CB-Net exclusively learns two representation clusters, i.e., one each for face and periocular, despite the shared-parameter encoder. Specifically, training CB-Net without $\mathcal{L}_{rcb}$ results in two embedding spaces with a minimal intersection, despite both modalities sharing the same identity labels. Therefore, we introduce $\mathcal{L}_{rcb}$ to resolve the inherent modality gap between periocular and face, as depicted in Fig. 6. The CB-Net instance trained with $\mathcal{L}_{rcb}$ leverages the inter-modality representation pairs to elicit a joint embedding space, whereby the angular distances for the intra-subject periocular and face pairs are explicitly minimized with respect to the annotated identity labels. This results in more effective cross-modal matching.

4) TOTAL LOSS

Given two modalities, i.e., periocular conditioned by face and face conditioned by periocular, the proposed CB-Net is learned with respect to three loss terms as follows:

$$\mathcal{L} = \mathcal{L}_{p|f} + \mathcal{L}_{f|p} + \alpha \mathcal{L}_{rcb} \quad (4)$$

where α is a weighting factor governing the contribution of $\mathcal{L}_{rcb}$ to $\mathcal{L}$.

IV. EXPERIMENTAL ANALYSIS AND DISCUSSIONS

A. DATASETS

Our training set consists of the face and periocular images sampled from VGGFace [39] and Ethnic [20] datasets. It is assembled with 166,737 examples of 1,054 identities for each periocular and face modality. We evaluate the generalization performance of CB-Net on five testing datasets, namely Ethnic, Pubfig [40], FaceScrub (FS) [41], IMDb Wiki (IMDb) [42], and AR [43]. Notably, these testing datasets are *completely disjoint from that of training*, i.e., no redundant identities. Except for the AR dataset, the others are challenging in-the-wild datasets. However, the AR dataset is mainly probed against occlusion, a common distracting factor in real-world deployment scenarios. We summarize the data distribution for these testing datasets in Table 2.

TABLE 2. Summary of testing datasets.

	Testing Datasets				
	Ethnic	Pubfig	FS	IMDb	AR
# ID	328	200	530	2,129	100
# gallery	1,645	9,221	31,066	40,241	700
# probe1	24,171	7,680	21,518	17,658	2,800
# probe2	-	6,138	27,292	15,252	1,400
# probe3	-	6,101	-	16,273	3,500
# probe4	-	-	-	-	600

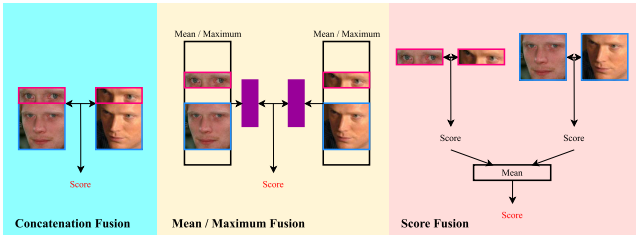


FIGURE 7. Multimodal matching scenarios through representation-level fusion (including direct concatenation and mean/maximum-pooling), and score-level fusion.

B. ENROLLMENT AND QUERY STAGES

With reference to Fig. 4, the CB-Net-trained representations apply to four operational matching modes at both enrollment and query stages, namely, unimodal, cross-modal, and multimodal. We detail each deployment mode as follows:

- **Unimodal Matching**: A CB template, either periocular or face-conditioned representation, is stored during enrollment as a gallery set. Direct matching is performed for verification or identification in the presence of the corresponding instance during the query stage as a probe/test set.
- **Cross-Modal Matching**: This mode necessitates only a single modality to be enrolled as a gallery set. On the other hand, another biometric modality is probed during the query stage, whereby the matching between the representations of the two modalities is performed to reach a decision. Under the assumption that periocular representations are enrolled, and face representations are queried, the matching score is computed between face and periocular. The same assumption is also possible for vice versa.
- **Multimodal Matching**: Multimodal matching is performed between the aggregated representation of the

TABLE 3. Hyperparameter configuration for experiments.

Hyperparameters	Details
Mini Batch Size (Face / Periocular)	64 / 64
# Epochs	40
Dropout	0.3
Learning Rate (LR)	0.001
LR Scheduler	0.1 every 12th epoch
Weight Decay	1.0×10^{-5}
s, m	128.0, 0.35
α, γ, τ	10.0, 0.001, 0.7

conditioned face and periocular. Our experiments consider (1) representation-level fusion methods, i.e., concatenation, and element-wise mean and maximum fusion; and (2) score-level fusion, i.e., score averaging, as shown in Fig. 7.

C. EXPERIMENTAL SETUP

We compare the performance of the CB-Net with the baseline networks over the five testing above datasets summarized in Table 2. Our experimental results for identification and verification tasks are reported [2] in terms of *rank-1 identification rate (IR)* and *verification equal error rate (EER)*.

For the identification task, we adopt the k-fold cross-validation such that the gallery and probe sets are alternated to compute an average IR. For example, as the Pubfig contains a gallery and three probes, each set is selected as a gallery, while the remaining are used as probes. This results in a total of $3 \times 4 = 12$ matches, wherein the accuracy is obtained via averaging. It is noted that this protocol is only applicable for unimodal matching and multimodal matching, as we fix a single gallery set for cross-modal matching.

On the other hand, the verification task involves the selection of positive and negative pairs via the random sampling of 4 images from each identity in the gallery set. This elicits $4 \times (4 - 1) = 12$ positive samples and $(4 \times 4) = 16$ negative samples per identity, in which the matching scores are calculated. We pursue this evaluation protocol for a fair comparison across different matching modes. Table 3 summarizes our configurations for all the empirical parameters.

We apply aggressive data augmentation during the training, namely random plane rotation within the range of $(-10, 10)$ degrees, random horizontal flipping, and random scaling within (1.0, 1.2) ranges.

D. PERFORMANCE ANALYSIS AND DISCUSSIONS

This section summarizes our empirical results accordingly. We denote the unconditioned face and periocular representations learned by the single-modality baselines as $\boldsymbol{v}_p$ and $\boldsymbol{v}_f$ in Tables 4, 5, and 6. On the contrary, the conditioned CB-Net representations, including periocular conditioned by face, and face conditioned by periocular, are referred to as $\boldsymbol{v}_{p|f}$ and $\boldsymbol{v}_{f|p}$, respectively.

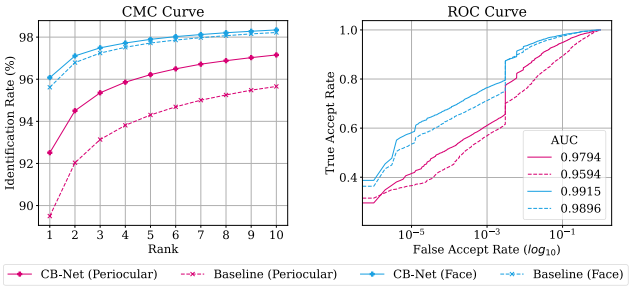


FIGURE 8. Periocular and face identification CMC and verification ROC curves for baseline and CB-Net averaged among 5 datasets (Ethnic, Pubfig, FaceScrub, IMDb Wiki, AR).

1) UNIMODAL MATCHING

We observe from Table 4 that CB-Net outperforms the baselines consistently. In particular, periocular conditioning by face reports a remarkable performance improvement (3.01% for rank-1 IR and 4.08% for EER), compared with face conditioning by periocular (0.45% for rank-1 IR and 0.70% for EER). The critical reason is that the face is a stronger attribute than the periocular; therefore, conditioning the periocular on the face leads to significant performance gain. On the contrary, the merit of CB is not entirely revealed for face conditioned by periocular as the periocular attributes are essentially dominated by the face.

In the meantime, Fig. 8 illustrates the Cumulative Matching Characteristic (CMC) and Receiver Operating Characteristic (ROC) curves for the unimodal matching of both periocular and face modalities. In both curves, the superiority of CB-Net is demonstrated in the performance difference compared to the baseline networks, particularly for the periocular. In Fig. 8, we also disclose the Area Under the Curve (AUC), wherein the AUC for CB-Net is remarkably higher than the baseline (2.00% for periocular, 0.19% for face).

2) CROSS-MODAL MATCHING

We perform cross-modal matching in two different settings: (1) probing periocular (test) against face (gallery) and (2) probing face (test) against periocular (gallery). Our experimental results are summarized in Table 5.

In Table 5, notice that the performance for the baseline is poor, regardless of the task being performed. On the contrary, the CB-Net has a vast performance improvement compared to the baseline, wherein a performance improvement of 82.28%, 82.11%, and 40.61% is observed for the rank-1 IR (using periocular and face as a gallery, respectively) and EER respectively. We illustrate these values in the CMC and ROC curves in Fig. 9.

In Fig. 9, it is observed that though the performances of the baseline and CB-Net are gradually increasing throughout the ranks, the performance of the baseline is still not as significant as the rank-1 IR for both periocular and face galleries are still less than 5% even after rank-10 IR is considered. On the other hand, the CB-Net saw a more significant IR performance

TABLE 4. Performance summary in terms of Rank-1 IR (%) and EER (%) for CB-Net on five testing datasets, including Ethnic, Pubfig, FaceScrub, IMDb Wiki, and AR.

Datasets	Rank-1 IR (%)				EER (%)			
	Periocular		Face		Periocular		Face	
	Baseline, v_p	CB-Net, $v_{p\|f}$	Baseline, v_f	CB-Net, $v_{f\|p}$	Baseline, v_p	CB-Net, $v_{p\|f}$	Baseline, v_f	CB-Net, $v_{f\|p}$
Ethnic	89.30	92.86	97.62	97.77	10.94	4.56	3.34	3.04
Pubfig	95.85	96.91	99.30	99.45	9.43	6.78	4.27	2.44
FS	94.73	96.42	98.57	98.69	6.23	3.98	2.74	2.19
IMDb	77.63	82.74	91.43	92.48	10.54	7.64	5.62	4.75
AR	89.98	93.62	91.17	91.99	14.50	8.27	4.10	4.11
Average	89.50	**92.51**	95.62	**96.07**	10.32	**6.24**	4.01	**3.31**

TABLE 5. Performance summary for cross-modal matching in terms of Rank-1 IR (%) and EER (%) for CB-Net on five testing datasets, including Ethnic, Pubfig, FaceScrub, IMDb Wiki, and AR.

Datasets	Rank-1 IR (%)				EER (%)	
	Periocular Gallery		Face Gallery			
	Baseline, v_p	CB-Net, $v_{p\|f}$	Baseline, v_f	CB-Net, $v_{f\|p}$	Baseline	CB-Net
Ethnic	0.24	78.45	0.41	77.35	49.96	10.75
Pubfig	1.57	93.70	0.71	92.96	50.25	8.57
FS	0.13	94.01	0.09	93.00	49.46	5.12
IMDb	0.06	70.35	0.06	69.70	49.12	9.34
AR	0.83	76.95	0.33	79.13	47.86	9.83
Average	0.41	**82.69**	0.32	**82.43**	49.33	**8.72**

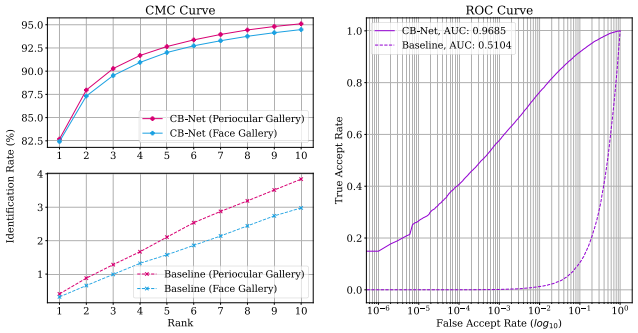


FIGURE 9. Cross-modal identification CMC and verification ROC curves for baseline and CB-Net averaged among 5 datasets (Ethnic, Pubfig, FaceScrub, IMDb Wiki, AR).

TABLE 6. Performance analysis for different multimodal fusion strategies in terms of Rank-1 IR (%) and EER (%) averaged over all testing datasets, namely Ethnic, Pubfig, FaceScrub, IMDb Wiki, and AR.

Fusion Strategies	Rank-1 IR		EER	
	Baseline, v_p, v_f	CB-Net, $v_{p\|f}, v_{f\|p}$	Baseline, v_p, v_f	CB-Net, $v_{p\|f}, v_{f\|p}$
Concatenation$^\sharp$	95.62	**96.52**	4.71	**4.33**
Mean$^\sharp$	**96.27**	95.81	4.87	**4.50**
Maximum$^\sharp$	**96.03**	96.01	5.70	**4.46**
Score$^\flat$	**96.62**	96.52	4.71	**4.33**

$\sharp$ Representation-level Fusion
$\flat$ Score-level Fusion

increase throughout the ranks. This case is also observed for the ROC curve in Fig. 9, whereby CB-Net achieves an AUC of 96.85%, while the baseline only achieves an AUC of 51.04%.

3) MULTIMODAL MATCHING

As illustrated in Fig. 7, we opt for representation-level (through direct concatenation, mean-pooling, and maximum-pooling) and score-level fusion strategies to facilitate multimodal matching as follows:

- **Direct Concatenation:** Given two periocular representations conditioned by face $\boldsymbol{v}_{p|f}, \boldsymbol{v}'_{p|f} \in \mathbb{R}^d$ and two face representations conditioned by periocular $\boldsymbol{v}_{f|p}, \boldsymbol{v}'_{f|p} \in \mathbb{R}^d$, we aggregate these representations into $(\boldsymbol{v}_{p|f}|\boldsymbol{v}_{f|p}) \in \mathbb{R}^{2d}$ and $(\boldsymbol{v}'_{p|f}|\boldsymbol{v}'_{f|p}) \in \mathbb{R}^{2d}$ for matching purposes. Also, given the baseline periocular and face representations, i.e., $\boldsymbol{v}_p, \boldsymbol{v}'_p \in \mathbb{R}^d$ and $\boldsymbol{v}_f, \boldsymbol{v}'_f \in \mathbb{R}^d$, we perform the multimodal matching between $(\boldsymbol{v}_p|\boldsymbol{v}_f) \in \mathbb{R}^{2d}$ and $(\boldsymbol{v}'_p|\boldsymbol{v}'_f) \in \mathbb{R}^{2d}$.
- **Mean-Pooling:** Accordingly, we compute the mean representations for $\boldsymbol{v}$ and $\boldsymbol{v}'$ to elicit the averaged representations for matching between $\mu(\boldsymbol{v}_{p|f}, \boldsymbol{v}_{f|p})$ and $\mu(\boldsymbol{v}'_{p|f}, \boldsymbol{v}'_{f|p})$. As for the baseline models, $\mu(\boldsymbol{v}_p, \boldsymbol{v}_f)$ is matched against $\mu(\boldsymbol{v}'_p, \boldsymbol{v}'_f)$ during evaluation.
- **Maximum-Pooling:** In lieu of mean-pooling, we exercise maximum-pooling on $\boldsymbol{v}$ and $\boldsymbol{v}'$ to compose the maximum representations for matching between $\max(\boldsymbol{v}_{p|f}, \boldsymbol{v}_{f|p}) \in \mathbb{R}^d$ and $\max(\boldsymbol{v}'_{p|f}, \boldsymbol{v}'_{f|p}) \in \mathbb{R}^d$, and $\max(\boldsymbol{v}_p, \boldsymbol{v}_f) \in \mathbb{R}^d$ and $\max(\boldsymbol{v}'_p, \boldsymbol{v}'_f) \in \mathbb{R}^d$.
- **Score Fusion:** For score-level fusion, we first obtain the matching scores for $(\boldsymbol{v}_{p|f}, \boldsymbol{v}'_{p|f})$, and $(\boldsymbol{v}_{f|p}, \boldsymbol{v}'_{f|p})$. We then compute the mean of both for decision-making.

For simplicity, we only exhibit the averaged values among the five testing datasets in Table 6 for rank-1 IR and EER.

Table 6 shows that the CB-Net consistently outperforms the baseline for verification. However, its rank-1 IR pales

TABLE 7. Ablation analysis for different CB-Net configurations in terms of Rank-1 IR (%) and EER (%) averaged over all testing datasets, namely Ethnic, Pubfig, FaceScrub, IMDb Wiki, and AR.

Network	$\mathcal{L}_p$	$\mathcal{L}_f$	$\mathcal{L}_{cb}$	$\mathcal{L}_{rcb}$	γ	Periocular		Face		Cross-Modal	
						Rank-1	EER	Rank-1	EER	Rank-1	EER
Baseline (Periocular)	✓					89.50	10.32	-	-	0.37	49.33
Baseline (Face)		✓				-	-	95.62	4.01		
CB-Net	✓	✓				92.33	7.05	95.16	3.52	1.02	46.63
	✓	✓	✓			92.49	6.80	**96.11**	3.68	78.71	10.13
	✓	✓		✓	0.1	91.30	6.47	94.22	3.98	87.68	**6.98**
	✓	✓		✓	0.01	92.00	6.49	95.75	3.66	**88.73**	7.41
	✓	✓		✓	0.001	**92.51**	**6.24**	96.07	**3.31**	82.56	8.72

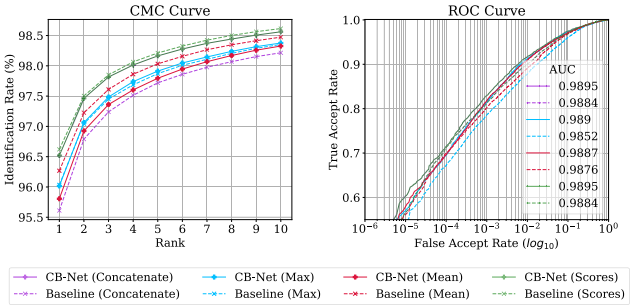


FIGURE 10. Multimodal identification CMC and verification ROC curves for baseline and CB-Net averaged among 5 datasets (Ethnic, Pubfig, FaceScrub, IMDb Wiki, AR).

compared to the baseline network for maximum, mean, and score fusion. In addition, notice that the performance of the score fusion is strikingly similar to the concatenation, particularly for the CB-Net. One plausible explanation for this is due to the saturation of the network, in which the performance will not improve any further for rank-1 IR or EER. We include the CMC and ROC in Fig. 10 to depict the similarities between the score fusion and concatenation fusion, alongside the maximum and mean fusion methods.

E. ABLATION STUDY

We conduct an ablation analysis in Table 7 with respect to different loss configurations. We report the average performance over all the testing datasets for CB-Net deployed under the unimodal and the cross-modal (switching periocular and face as gallery and test alternately) scenarios.

It is evident that the presence of $\mathcal{L}_{cb}$ results in a vast performance improvement, particularly for the cross-modal matching scenario. Notably, there is a performance gain of 77.70% and 36.5% in rank-1 IR and EER, respectively. In addition, the regularizer γ causes an even increased performance gain for cross-modal matching, whereby $\gamma = 0.1$ and $\gamma = 0.01$ resulted in an increase of 8.97% and 10.02% for rank-1 IR and 3.15% and 2.72% for EER respectively.

In summary, despite CB-Net in the presence of $\mathcal{L}_{rcb}$ where $\gamma = 0.001$ reports the best performance for unimodal deployment, we conclude that setting $\gamma = 0.01$ achieves the most balanced performance. Furthermore, though setting $\gamma = 0.01$ leads to marginal performance degradation for unimodal compared to $\gamma = 0.001$, it is indispensable owing to the significant performance enhancement for cross-modal matching. More specifically, with this setting, we discern a performance gain of 6.17% and 1.31% in rank-1 IR and EER, respectively. Therefore, we deduce that γ plays a vital role in inter-modality, and intra-class attraction, reflected by the cross-modal matching performance.

TABLE 8. Performance comparison with other SoTA periocular networks in terms of Rank-1 IR (%) and EER (%) averaged over all testing datasets, except for AR.

Networks	Periocular	
	Rank-1	EER
RGB-OCLBCP [20]	80.83	11.50
PF-GLSR [33]	88.70	10.47
CB-Net	92.25	5.73

F. COMPARISON WITH OTHER WORKS

We compare in Table 8 the proposed CB-Net and two relevant periocular networks, specifically [20] and [33]. These works comply with the CB regimen, which evaluates the same unconstrained periocular datasets. We exclude the AR dataset in this section for a fair comparison. We discern that the CB-Net shows a significant performance improvement over [20] and [33] - at least 3.55% in rank-1 IR and 4.74% in terms of EER. The important reasons are: (1) [20] encodes the periocular representations based upon the pre-extracted descriptors, whereas the CB-Net learns directly from the raw periocular images. (2) [33] involves knowledge distillation (KD) from a teacher (face) network to facilitate embedding learning for a student (periocular) network. This restricts its overall performance to that of the teacher, and KD prohibits the teacher network from subsequent learning.

G. DISCUSSIONS

1) HOW DISCRIMINATIVE IS THE CONDITIONED REPRESENTATION?

In Figs. 11 and 12, we evaluate the discriminability of the baseline and the CB representations in terms of decidability index d' [7] with respect to the intra-subject and the inter-subject Cosine similarity scores for both periocular and face,

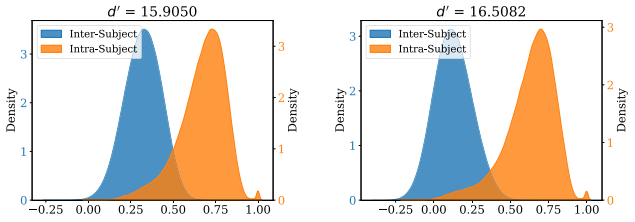


FIGURE 11. Intra-modality (Periocular) for baseline and CB-Net (with $\mathcal{L}_{cb}$).

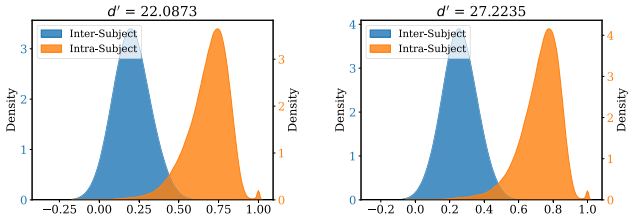


FIGURE 12. Intra-modality (Face) for baseline and CB-Net (with $\mathcal{L}_{cb}$).

respectively. Mathematically, we define d' as follows:

$$d' = \frac{|\mu_+ - \mu_-|}{\sqrt{(\sigma_+^2 + \sigma_-^2)}} \tag{5}$$

where μ_+ and σ_+ indicate the mean and the variance of the intra-subject similarity scores; μ_- and σ_-, in contrast, denote the mean and the variance of the inter-subject similarity scores. By definition, d' is a performance indicator, whereby a greater d' signifies higher representation discrimination.

Following (5), we investigate the *intra-modality matching* with respect to d' i.e., comparing periocular to periocular, face to face, for a more thorough analysis. Figs. 11 and 12 show that CB-Net has a higher d' than the baseline by 0.60 and 5.14, respectively. The increased d' values conform to the performance improvement for both periocular and face in rank-1 IR and EER.

We also conduct an *inter-modality matching* analysis on d' for a more thorough analysis. In Fig. 13, we depict the superiority of CB-Net over the baseline network, as well as the importance of $\mathcal{L}_{cb}$ in our network. Our analyses are conducted for both inter-class and intra-class for the baseline networks and the CB-Net, with and without the presence of $\mathcal{L}_{cb}$.

In Fig. 13, the intra-subject and inter-subject similarity score distribution for both the baseline and CB-Net without $\mathcal{L}_{cb}$ networks overlap one another around the Cosine similarity score of 0.0, whereby the d' values are 0.15 and 1.14 respectively. Notably, the intra-subject similarity scores are deemed to be the same as the inter-subject, despite being of the same identity. On the contrary, deploying CB-Net with $\mathcal{L}_{cb}$ shows its flexibility via the well-separated intra-subject and inter-subject histograms, despite being of different modalities. This is computed in an estimated d' value of 19.58, whereby the intra-class distance of different modalities has significantly higher similarity scores. This signifies that $\mathcal{L}_{cb}$ plays an essential role in improving the performance of cross-modal matching via representation learning.

TABLE 9. Performance summary for cross-modal matching in terms of Rank-1 IR (%) and EER (%) for CB-Net on five testing datasets in the absence of $\mathcal{L}_{cb}$.

Datasets	Rank-1 IR (%)		EER (%)
	Periocular Gallery	Face Gallery	
Ethnic	1.82	1.04	47.34
Pubfig	1.49	0.84	47.22
FS	0.64	0.49	46.34
IMDb Wiki	0.02	0.09	46.23
AR	1.15	2.31	46.01
Average	1.02	0.95	46.63

2) HOW STRONG IS THE CORRELATION BETWEEN FACE AND PERIOCULAR REPRESENTATIONS?

The correlation between face and periocular representations should be reasonably high for decent cross-modal matching. We attempt to answer this question via Pearson correlation defined [8] as follows:

$$\rho = \frac{\sum_i (x_i - \mu_x)(y_i - \mu_y)}{\left(\sum_i (x_i - \mu_x)^2 \sum_i (y_i - \mu_y)^2\right)^{\frac{1}{2}}} \tag{6}$$

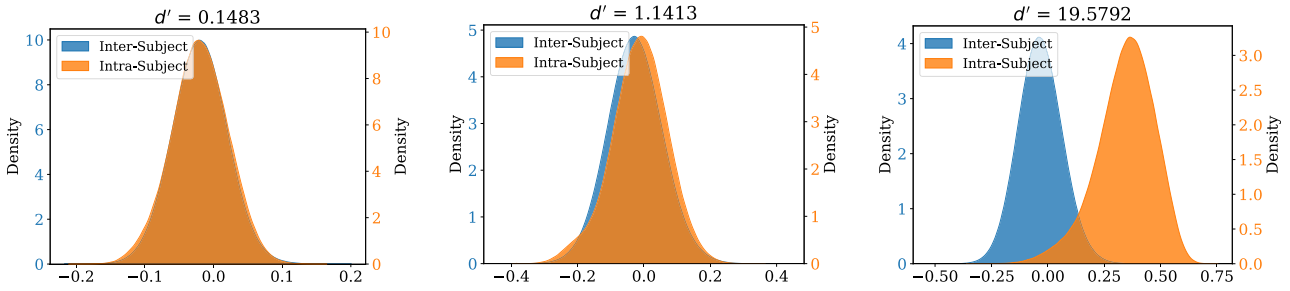
where μ_x and μ_y denote the mean values for x and y respectively. $\rho = 1$ signifies a strong linear dependency between x and y, implying a perfect positive correlation. On the contrary, $\rho = 0$ indicates no linear dependency between the two variables, while $\rho = -1$ implies a perfect negative correlation.

With reference to (6), we compute ρ as the inter-modality correlation between the CB representation, i.e., face x and periocular y, based upon the softmax-learned prototypes, indicated by μ_x and μ_y, respectively. For simplicity, we subsample a toy image subset of only six classes in this pilot study. As shown in Fig. 14, both baseline and the CB-Net instance without $\mathcal{L}_{cb}$ render two representation sets of no inter-modality correlation, i.e., with ρ values of only -0.02 and -0.01. On the contrary, CB-Net with $\mathcal{L}_{cb}$ shows a positive inter-modality correlation of 0.48. This is proportional to the performance in Table 5. For a complete analysis, we provide in Table 9 the cross-modal matching performance for CB-Net without $\mathcal{L}_{cb}$.

In summary, the usage of CB-Net alongside $\mathcal{L}_{cb}$ with a positive correlation showed an improved cross-modal matching performance of at least 82.56%, in contrast to CB-Net without $\mathcal{L}_{cb}$ and the baseline networks with negative correlations, which showed cross-modal matching of 1.02% and 0.37% respectively (see Table 7).

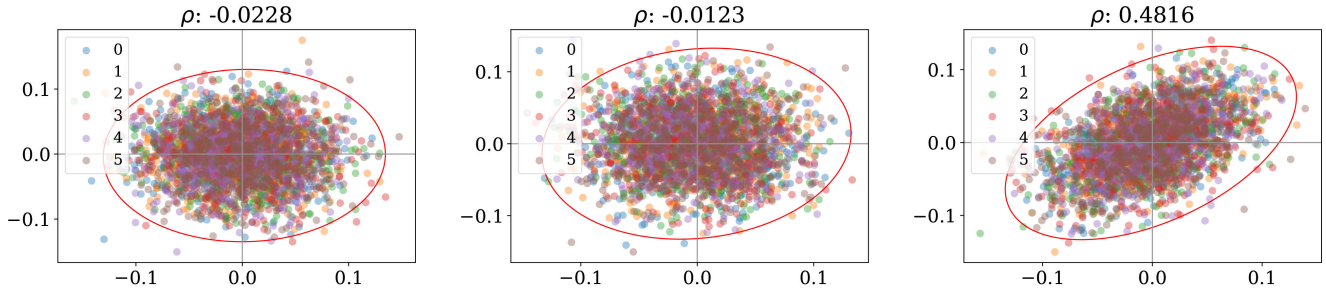
3) HOW WELL DO HAND-CRAFTED DESCRIPTORS WORK FOR CROSS-MATCHING?

One may ask whether cross-matching the non-learning-based face and periocular features is feasible, as the periocular is part of the face. This section addresses this query with one of the most classical hand-crafted texture descriptors - Local



(a) Baseline (b) CB-Net (without $\mathcal{L}_{cb}$) (c) CB-Net (with $\mathcal{L}_{cb}$)

FIGURE 13. Intra-subject and inter-subject cosine similarity distribution for inter-modalities for Baseline and CB-Net (with and without the presence of $\mathcal{L}_{cb}$).



(a) Baseline (b) CB-Net (without $\mathcal{L}_{cb}$) (c) CB-Net (with $\mathcal{L}_{cb}$)

FIGURE 14. Pearson correlation of face and periocular representations for baseline and CB-Net (with and without $\mathcal{L}_{cb}$).

TABLE 10. Performance analysis of LBP descriptors for various matching modes in terms of Rank-1 IR (%) for all testing datasets, namely Ethnic, Pubfig, FaceScrub, IMDb Wiki, and AR.

Datasets	Rank-1 IR (%)		
	Periocular	Face	Cross-Modal
Ethnic	12.15	19.78	2.45
Pubfig	32.71	46.58	7.22
FaceScrub	16.50	34.65	2.13
IMDb	2.75	9.71	1.15
AR	34.81	42.33	6.18
Average	19.78	30.61	3.83

Binary Pattern (LBP) [11], which we use as a feature extractor of both face and periocular. We present the experimental results amongst the five datasets for unimodal and cross-modal matching in Table 10. For simplicity, the rank-1 IR for cross-modal matching is obtained from the averaged periocular and face galleries.

As expected, it is disclosed in Table 10 that the LBP face descriptors perform better than periocular with a discrepancy of 10.83% in rank-1 IR, as periocular suffers from severe under-representation issues. Overall, the LBP descriptors perform poorly across all testing datasets for unimodal and cross-modal matching, compared to CB-Net (refer to Table 7). We depict the CMC curve in Fig. 15. For comparison, we also include the averaged curve for multimodal matching.

In Fig. 15, it is observed that the face has the best performance, even more than the matching of multimodal and also periocular. On the contrary, cross-modal matching

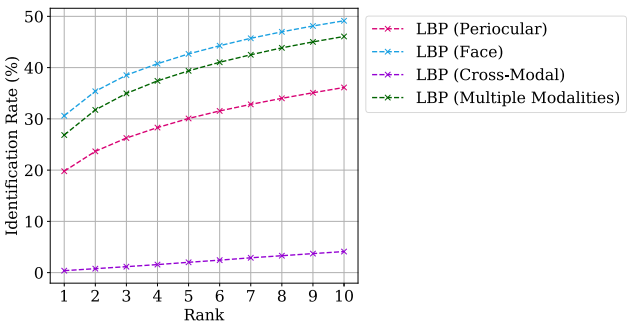


FIGURE 15. CMC curve for various matching modes for LBP descriptors averaged over 5 testing datasets (Ethnic, Pubfig, FaceScrub, IMDb Wiki, and AR).

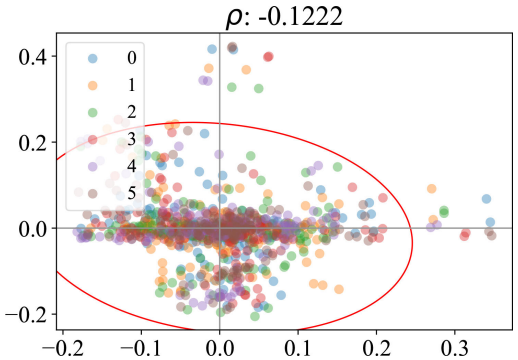


FIGURE 16. Pearson correlation for face and periocular descriptors handcrafted using LBP.

also performs poorly, despite LBP being a hand-crafted descriptor dependent on the image's pixel-wise features.

Furthermore, despite the periocular being a sub-region of a face, the LBP descriptors perform poorly in cross-modal matching. This is because the face and periocular are captured in different input resolutions, causing the facial landmarks to be misaligned. We also demonstrate in Fig. 16 that the Pearson correlation for periocular and face features extracted by LBP is estimated to be only $\rho = -0.12$, i.e., close to no inter-modality correlation. This reaffirms that face and periocular are unique biometric modalities, although both share a common peripheral ocular region.

V. CONCLUSION

This paper introduced the notion of Conditional Biometrics (CB), a framework designed to enhance the performance of various biometric systems, with a particular focus on periocular and facial modalities in our study. We have developed CB-Net, a deep neural network architecture that facilitates representation learning to achieve this. The CB-Net enabled mutual conditioning between periocular and facial biometric information, resulting in improved performance in both unimodal and multimodal matching scenarios. Additionally, CB-Net training established correlations between the two biometric modalities to enable cross-modal matching.

As a demonstration, the experiments were conducted using challenging in-the-wild face and periocular datasets to evaluate the learning aptitude of CB-Net on unimodal, cross-modal, and multimodal matching modes. Our experimental results disclosed that the CB-Net shows a consistent performance improvement in periocular recognition, particularly in the cross-modal deployment scenario.

In future research, we plan to explore other combinations of biometric modalities, such as face and iris, periocular and iris, etc. We hypothesize that using conditioned representation learning based on the CB framework can enhance baseline performances for these biometric combinations.

REFERENCES

[1] A. K. Jain, D. Deb, and J. J. Engelsma, "Biometrics: Trust, but verify," 2021, *arXiv:2105.06625*.

[2] M. O. Oloyede and G. P. Hancke, "Unimodal and multimodal biometric sensing systems: A review," *IEEE Access*, vol. 4, pp. 7532–7555, 2016.

[3] M. Singh, R. Singh, and A. Ross, "A comprehensive overview of biometric fusion," *Inf. Fusion*, vol. 52, pp. 187–205, Dec. 2019.

[4] J. Fierrez, A. Morales, R. Vera-Rodriguez, and D. Camacho, "Multiple classifiers in biometrics—Part 2: Trends and challenges," *Inf. Fusion*, vol. 44, pp. 103–112, Nov. 2018.

[5] J. Fierrez, A. Morales, R. Vera-Rodriguez, and D. Camacho, "Multiple classifiers in biometrics—Part 1: Fundamentals and review," *Inf. Fusion*, vol. 44, pp. 57–64, Nov. 2018.

[6] U. Park, A. Ross, and A. K. Jain, "Periocular biometrics in the visible spectrum: A feasibility study," in *Proc. IEEE 3rd Int. Conf. Biometrics, Theory, Appl., Syst.*, Sep. 2009, pp. 1–6.

[7] M. Lee, H. Hong, and K. Park, "Noisy ocular recognition based on three convolutional neural networks," *Sensors*, vol. 17, no. 12, p. 2933, Dec. 2017.

[8] J. L. Rodgers and W. A. Nicewander, "Thirteen ways to look at the correlation coefficient," *Amer. Statistician*, vol. 42, no. 1, pp. 59–66, Feb. 1988.

[9] P. Kumari and K. R. Seeja, "A novel periocular biometrics solution for authentication during COVID-19 pandemic situation," *J. Ambient Intell. Humanized Comput.*, vol. 12, pp. 10321–10337, Jan. 2021.

[10] Y. Martínez-Díaz, H. Méndez-Vázquez, L. S. Luevano, M. Nicolás-Díaz, L. Chang, and M. González-Mendoza, "Towards accurate and lightweight masked face recognition: An experimental evaluation," *IEEE Access*, vol. 10, pp. 7341–7353, 2022.

[11] G. Mahalingam and K. Ricanek, "LBP-based periocular recognition on challenging face datasets," *EURASIP J. Image Video Process.*, vol. 2013, no. 1, pp. 1–13, Dec. 2013.

[12] J. Xu, M. Cha, J. L. Heyman, S. Venugopalan, R. Abiantun, and M. Savvides, "Robust local binary pattern feature sets for periocular biometric identification," in *Proc. 4th IEEE Int. Conf. Biometrics, Theory, Appl. Syst. (BTAS)*, Sep. 2010, pp. 1–8.

[13] U. Park, R. R. Jillela, A. Ross, and A. K. Jain, "Periocular biometrics in the visible spectrum," *IEEE Trans. Inf. Forensics Security*, vol. 6, no. 1, pp. 96–106, Mar. 2011.

[14] S. Karahan, A. Karaöz, Ö. F. Özdemir, A. G. Gü, and U. Uludag, "On identification from periocular region utilizing SIFT and SURF," in *Proc. 22nd Eur. Signal Process. Conf. (EUSIPCO)*, Sep. 2014, pp. 1392–1396.

[15] A. Rattani and R. Derakhshani, "Ocular biometrics in the visible spectrum: A survey," *Image Vis. Comput.*, vol. 59, pp. 1–16, Mar. 2017.

[16] I. Nigam, M. Vatsa, and R. Singh, "Ocular biometrics: A survey of modalities and fusion approaches," *Inf. Fusion*, vol. 26, pp. 1–35, Nov. 2015.

[17] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet classification with deep convolutional neural networks," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 25, 2012, pp. 1097–1105.

[18] H. Hwang and E. C. Lee, "Near-infrared image-based periocular biometric method using convolutional neural network," *IEEE Access*, vol. 8, pp. 158612–158621, 2020.

[19] S. S. Behera, S. S. Mishra, B. Mandal, and N. B. Puhan, "Variance-guided attention-based twin deep network for cross-spectral periocular recognition," *Image Vis. Comput.*, vol. 104, Dec. 2020, Art. no. 104016.

[20] L. C. O. Tiong, A. B. J. Teoh, and Y. Lee, "Periocular recognition in the wild with orthogonal combination of local binary coded pattern in dual-stream convolutional neural network," in *Proc. Int. Conf. Biometrics (ICB)*, Jun. 2019, pp. 1–6.

[21] R. Jillela and A. Ross, "Mitigating effects of plastic surgery: Fusing face and ocular biometrics," in *Proc. IEEE 5th Int. Conf. Biometrics, Theory, Appl. Syst. (BTAS)*, Sep. 2012, pp. 402–411.

[22] Z. Luo, J. Li, and Y. Zhu, "A deep feature fusion network based on multiple attention mechanisms for joint iris-periocular biometric recognition," *IEEE Signal Process. Lett.*, vol. 28, pp. 1060–1064, 2021.

[23] M. Karakaya, "Iris-ocular-periocular: Toward more accurate biometrics for off-angle images," *J. Electron. Imag.*, vol. 30, no. 3, Jun. 2021, Art. no. 033035.

[24] F. Algashaam, K. Nguyen, J. Banks, V. Chandran, T.-A. Do, and M. Alkanhal, "Hierarchical fusion network for periocular and iris by neural network approximation and sparse autoencoder," *Mach. Vis. Appl.*, vol. 32, no. 1, p. 15, Jan. 2021.

[25] K. B. Raja, R. Raghavendra, M. Stokkenes, and C. Busch, "Multi-modal authentication system for smartphones using face, iris and periocular," in *Proc. Int. Conf. Biometrics (ICB)*, May 2015, pp. 143–150.

[26] L. C. O. Tiong, S. T. Kim, and Y. M. Ro, "Implementation of multimodal biometric recognition via multi-feature deep learning networks and feature fusion," *Multimedia Tools Appl.*, vol. 78, no. 16, pp. 22743–22772, Aug. 2019, doi: 10.1007/s11042-019-7618-0.

[27] V. Talreja, N. M. Nasrabadi, and M. C. Valenti, "Attribute-based deep periocular recognition: Leveraging soft biometrics to improve periocular recognition," in *Proc. IEEE/CVF Winter Conf. Appl. Comput. Vis. (WACV)*, Jan. 2022, pp. 1141–1150.

[28] A. Nagrani, S. Albanie, and A. Zisserman, "Learnable PINs: Cross-modal embeddings for person identity," *Proc. Eur. Conf. Comput. Vis. (ECCV)*, Sep. 2018, pp. 71–88.

[29] C. Kim, H.V. Shin, T.H. Oh, A. Kaspar, M. Elgharib, and W. Matusik, "On learning associations of faces and voices," in *Proc. Asian Conf. Comput. Vis.* Cham, Switzerland: Springer, 2018, pp. 276–292.

[30] H. Wang, X. Dong, Z. Jin, J.-L. Dugelay, and M. Tistarelli, "Cross-spectrum face recognition using subspace projection hashing," in *Proc. 25th Int. Conf. Pattern Recognit. (ICPR)*, Jan. 2021, pp. 615–622.

[31] L. Jiang, J. Zhang, and B. Deng, "Robust RGB-D face recognition using attribute-aware loss," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 42, no. 10, pp. 2552–2566, Oct. 2020.

[32] S. M. Iranmanesh, A. Dabouei, and N. M. Nasrabadi, "Attribute adaptive margin softmax loss using privileged information," 2020, *arXiv:2009.01972*.
[33] Y. G. Jung, C. Y. Low, J. Park, and A. B. J. Teoh, "Periocular recognition in the wild with generalized label smoothing regularization," *IEEE Signal Process. Lett.*, vol. 27, pp. 1455–1459, 2020.
[34] G. Hinton, O. Vinyals, and J. Dean, "Distilling the knowledge in a neural network," 2015, *arXiv:1503.02531*.
[35] S. Chen, Y. Liu, X. Gao, and Z. Han, "MobileFaceNets: Efficient CNNs for accurate real-time face verification on mobile devices," in *Proc. Chin. Conf. Biometric Recognit.* Cham, Switzerland: Springer, 2018, pp. 428–438.
[36] 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 *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 5265–5274.
[37] P. Khosla, P. Teterwak, C. Wang, A. Sarna, Y. Tian, P. Isola, A. Maschinot, C. Liu, and D. Krishnan, "Supervised contrastive learning," 2020, *arXiv:2004.11362*.
[38] F. Schroff, D. Kalenichenko, and J. Philbin, "FaceNet: A unified embedding for face recognition and clustering," 2015, *arXiv:1503.03832*.
[39] O. M. Parkhi, A. Vedaldi, and A. Zisserman, "Deep face recognition," in *Proc. BMVC*, 2015, pp. 1–12.
[40] N. Kumar, A. C. Berg, P. N. Belhumeur, and S. K. Nayar, "Attribute and simile classifiers for face verification," in *Proc. IEEE 12th Int. Conf. Comput. Vis. (ICCV)*, Sep. 2009, pp. 365–372.
[41] H.-W. Ng and S. Winkler, "A data-driven approach to cleaning large face datasets," in *Proc. IEEE Int. Conf. Image Process. (ICIP)*, Oct. 2014, pp. 343–347.
[42] R. Rothe, R. Timofte, and L. Van Gool, "DEX: Deep EXpectation of apparent age from a single image," in *Proc. IEEE Int. Conf. Comput. Vis. Workshop (ICCVW)*, Dec. 2015, pp. 252–257.
[43] A. Martinez and R. Benavente, "The AR face database," CVC, New Delhi, India: Tech. Rep. #24, 1998.

TIONG-SIK NG received the B.Eng. degree in electronics majoring in computer and the M.Sc. degree in information technology from Multimedia University, Malaysia, in 2016 and 2019, respectively. He is currently pursuing the Ph.D. degree in electrical and electronic engineering with Yonsei University, South Korea. His research interests include biometric security and deep learning.

CHENG-YAW LOW received the Ph.D. degree in electrical and electronic engineering from Yonsei University, Seoul, South Korea, in 2018. He is currently a Research Associate with the Data Science Group, Institute for Basic Science, South Korea. His research interests include biometric recognition and deep learning.

JACKY CHEN LONG CHAI received the M.Eng. degree (Hons.) in electrical and electronic engineering from the University of Nottingham Malaysia, Selangor, Malaysia, in 2019, and the M.S. degree from Yonsei University under the supervision of Prof. Andrew Beng Jin Teoh. He was a System-on-a-Chip (SoC) Design Engineer with Intel Corporation Malaysia and was selected as a scholarship candidate by the National Institute of International Education Korea (NIIED), in 2019. His current research interests include computer vision and biometric recognition.

ANDREW BENG JIN TEOH (Senior Member, IEEE) received the B.Eng. degree in electronic and the Ph.D. degree from the National University of Malaysia, in 1999 and 2003, respectively. He is currently a Full Professor with the Electrical and Electronic Engineering Department, College Engineering, Yonsei University, South Korea. His research for which he has received funding focuses on biometric applications and biometric security. He has published more than 350 international refereed journal articles, conference articles, edited several book chapters, and edited book volumes. His current research interests include machine learning and information security. He served/serving as a Guest Editor for the *IEEE Signal Processing Magazine* and an Associate Editor for IEEE TRANSACTIONS ON INFORMATION FORENSIC AND SECURITY, IEEE BIOMETRICS COMPENDIUM, and *Machine Learning with Applications* (Elsevier).