Periocular in the Wild Embedding Learning with Cross-Modal Consistent Knowledge Distillation

Yoon Gyo Jung, Jaewoo Park, Cheng Yaw Low, Leslie Ching Ow Tong, Andrew Beng Jin Teoh

a School of Electrical and Electronic Engineering, Yonsei University, Seoul 120-749, Republic of Korea
b Computational Science Research Center, Korea Institute of Science and Technology, Seoul 02792, Republic of Korea
E-mail: jungyg@yonsei.ac.kr, julypraise@yonsei.ac.kr, chengyawlow@yonsei.ac.kr, tiongleslie@kist.re.kr, bjteoh@yonsei.ac.kr

*Corresponding Author:
E-mail address: bjteoh@yonsei.ac.kr
Tel: (82-2) 2123 5772
Postal address: School of Electrical and Electronic Engineering, College of Engineering, Yonsei University, 50 Yonsei-ro, Seodaemun-gu, Seoul 120-749, Republic of Korea

Abstract
Periocular biometric, or peripheral area of ocular, is a collaborative alternative to face, especially if a face is occluded or masked. In practice, sole periocular biometric captures least salient facial features, thereby suffering from intra-class compactness and inter-class dispersion issues particularly in the wild environment. To address these problems, we transfer useful information from face to support periocular modality by means of knowledge distillation (KD) for embedding learning. However, applying typical KD techniques to heterogeneous modalities directly is suboptimal. We put forward in this paper a deep face-to-periocular distillation networks, coined as cross-modal consistent knowledge distillation (CM-CKD) henceforward. The three key ingredients of CM-CKD are (1) shared-weight networks, (2) consistent batch normalization, and (3) a bidirectional consistency distillation for face and periocular through an effectual CKD loss. To be more specific, we leverage face modality for periocular embedding learning, but only periocular images are targeted for identification or verification tasks. Extensive experiments on six constrained and unconstrained periocular datasets disclose that the CM-CKD-learned periocular embeddings extend identification and verification performance by 50% in terms of relative performance gain computed based upon face and periocular baselines. The experiments also reveal that the CM-CKD-learned periocular features enjoy better subject-wise cluster separation, thereby refining the overall accuracy performance.

Keywords: periocular biometric, cross-modal distillation, shared-weight networks, embedding learning
1. Introduction

Biometrics, either be physiological or behavioral such as fingerprint, face, iris, retina, voice, signature, to name just a few, are biological identifiers unique to individuals. Apart from the universally accepted fingerprint, biometric systems exploiting ocular traits, including iris and retina, have also accomplished a significant breakthrough in the past decades. Recently, periocular-based biometrics, or peripheral area of ocular, encapsulating eyebrow, eyelid, eyelash, eye shape, tear duct and skin texture (as depicted in Figure 1) has gained increasing attention as one of the substitutional modality to face [1], [2], [3].

![Examples of paired face (left) and periocular (right) images employed in training stage. Benchmarking datasets (unknown testing datasets) are paired in the same manner, but only periocular images are targeted for identification and verification tasks.](image)

Since periocular biometrics encloses only the immediate vicinity of eyes, i.e., a sub-region of a face, it captures relatively less information compared with that of the face. Despite that, it is a collaborative alternative to face particularly in cases of occlusion, aging, makeup, and plastic surgery [3]. The primary reason is that the periocular images also embed meaningful details related to an individual. Moreover, given a face dataset, a large-scale of periocular images are easily acquired by cropping the periocular region from face images with respect to the detected facial landmarks [4].

The pioneering works of periocular biometrics emphasize the handcrafted texture descriptors [3]-[9]. Until recently, deep neural networks are explored to learn a powerful periocular feature embedding [10]-[13]. Unfortunately, most of the present periocular datasets are restricted to the controlled environments and hence the overall performance for identification and verification tasks is deteriorated applied to the wild setting [10]. This is attributed to the poor intra-class compactness and inter-class dispersion due to external distractors, e.g. sensor location, illumination conditions, pose alignments, and other real-world disturbances. In a nutshell, the co-modulation of external distractors and insufficient identity-relevant information are the key problems of periocular recognition in the wild.

This paper attempts to enhance the unconstraint periocular embedding learning by distilling the face information to the periocular biometrics. We leverage the strong correlation of face and periocular for face-to-periocular distillation in the training stage, but only the periocular modality is demanded during the inference stage. Therefore, this sets our work apart from the conventional multimodal biometrics.
where all biometric modalities are required for training and testing. To be specific, we put forward a cross-modal consistent knowledge distillation (CM-CKD) model in this paper that is substantiated by a shared-weight network (SWN) enabled with consistent batch normalization (CBN) and a cross-modal consistent knowledge distillation loss (CKD).

1.1 Related Work
In this section, we first review a collection of periocular recognition literature that covers both hand-crafted and deep learning-based features. Then, we present various knowledge distillation techniques relevant to our proposed method.

Periocular Biometric
Early works in periocular biometric focus on hand-crafted texture descriptors, such as Histogram of Orientation and Gradient (HOG), Scale Invariant Feature Transform (SIFT), and Local Binary Pattern (LBP) [5]-[7] are dominant for periocular recognition under constraint environment. Further studies pursued by combining several hand-crafted texture descriptors, such as applying the Gabor Filter followed by HOG and LBP [8]. [9] examines several hand-crafted texture descriptors such as Walsh Hadamard product LBP (WLBP), Laws Masks, Discrete Cosine Transform (DCT), Discrete Wavelet Transform (DWT), Force Field transform, Speed-Up Robust Features (SURF), Gabor filters, and Laplacian of Gaussian (LoG) filters. Although some accuracy performance gain is observed, they are suffered from aging degradation, pose invariance and illumination changes. These studies reveal clearly that hand-crafted features are not well-performed due to the lack of high-level semantic information.

The recent works learn robust representations by deep Convolutional Neural Networks (CNNs) for periocular recognition. For instance, [12] interchanges the ocular area containing iris and sclera from the periocular images of different identities to synthesize artificial samples. These samples are classified with respect to the identity of the periocular region but not the ocular. Such an approach forces the CNNs to learn from the periocular region and pay less attention to the ocular area. [13] mimics human visual attention mechanism that focuses on the eyebrow and ocular area of the periocular image. The authors first train a Fully Convolutional Network (FCN) to segment the eyebrow and ocular area. The FCN yields a region of interest (ROI) that guides the periocular recognition network. The periocular network is trained at the second fold via sigmoid cross-entropy loss function. [15] adopted generative adversarial networks (GAN) [16] to remove the noisy attributes such as glasses and eye gaze, leaving only the discriminative attributes meaningful for recognition purposes.

Interestingly, near infrared (NIR) periocular recognition also has started to gain attention. [17] concatenates the global average pooled mid-level feature maps of CNN to generate the feature vector, by assuming mid-level feature maps capture global information while high-level feature maps offer
local information, hence using both would provide better results. [18] trains an RGB-NIR cross-modal CNN which can match across modalities by using shared weights networks in the high-level equipped with attention module with a variance guided loss function. However, these works do not perform well under difficult scenarios, such as misalignment of training-testing data or cases where eyebrows are occluded.

Considering the periocular in the wild challenge, [10] outlines a dual-stream CNN that encodes the Orthogonal Combination-Local Binary Coded Pattern (OC-LBCP) descriptor – a composition of the well-established Local Binary Pattern and Local Ternary Patterns (LTP). The OC-LBCP descriptor for both left and right eyes is probed on a newly collected periocular repository set under the unconstrained environment. On the contrary, [19] tackles this problem by regularizing the ground truth labels with a pre-task periocular network, whereas [20] stands out from the state of the arts by learning from the face modality on top of periocular.

**Knowledge Distillation**

Knowledge distillation (KD) [21] has been widely explored for network compression. In comparison to the vanilla deep networks, the KD instance exercises a lightweight network design, network pruning and network quantization to build a more compact and efficient network. In general, KD for network compression transfer the knowledge embedded in logits and/or attention maps [21]-[23] of the teacher network to supervise the student network by inducing similar relational responses among mini-batches [24]-[26].

To be specific, [21] generates soft target prediction from a large and cumbersome teacher network logits with a temperature term in order to supervise the small student network’s logits with KL divergence. [22] distills the teacher network's attention map of intermediate features to the student network to induce similar attention patterns. Similarly, [27] distills the flow of solution procedure matrix which is a response between two intermediate features. On the other hand, [24] distills the Euclidian statistics of the intermediate feature maps from the teacher to the student network. [25] generalizes [24] utilizes $p$-distance or cosine-distance instead of Euclidean distance. Similarly, [26] models the teacher’s knowledge and transfers it to the student network by minimizing the divergence of teacher and student network probability distributions.

Recent studies explore knowledge distillation without a teacher network, which is known as self-distillation. [27] appends classification branches of CNN to the intermediate layers and distill the feature maps and logits from the last classification branch to the others. [28] distills itself using augmented (distorted) input images by minimizing the maximum mean discrepancy and bi-directional KL-divergence term between the image and its augmentation. [29] approaches self-distillation in a similar way where they use different samples of the same class instead of data augmentation.
However, these contemporary works are restricted to only network compression with respect to the same modality but not in the cross-modal setting. A general rule of thumb of KD is that the inputs for the teacher and the student networks are to be homogenous strictly. When the inputs are of heterogeneous, such as images with different sizes, or with different modalities such as image, sound or text, plain KD does not work.

Several studies [30]-[32] attempt to solve cross-modal distillation problem by transferring the knowledge from the teacher penultimate activations or logits to the student network with paired inputs. [30] uses paired image and audio cues from videos to train an audio network in an unsupervised manner. The authors generate pseudo-labels for the training data and minimize the KL-divergence between each networks’ predictions. [31] reports a cross-modal knowledge distillation from a teacher network trained on high-resolution images to a student network that trained on the low-resolution images with their inputs paired up. The knowledge transfer learning paradigm is utilized to initialize the student network and the KL-divergence of the soft target distributions for the teacher-student networks is minimized. In the meantime, the RGB images are paired to the corresponding depth images in [32] for KD learning through intermediate feature maps seeing that both are strongly correlated and similar in terms of image dimension. Apart from [31], another recent contribution to KD for face is presented in [33]. It is reported that the vanilla KD is in favor of improving the face recognition tasks by feature realignment, whereby the L2 distance for the normalized penultimate activations rendered by the teacher-student networks is minimized.

1.2 Motivations and Contributions

In this paper, we particularly address the deep embedding learning problem for periocular identification or verification in the wild. We attempt to mine and transfer the discriminative information from the face modality that captures better identity-specific features to the periocular modality for embedding learning. The notion is based upon the fact that face and periocular biometrics share considerable common information, e.g. identity, skin color, the shape of eyes, etc. This motivates us to explore knowledge distillation (KD) for cross-modal KD, i.e. from face to periocular. Unfortunately, the direct application of the prevailing KD models for cross-modal KD appears to be suboptimal owning to the cross-modality nature.

For an effectual face-to-periocular distillation, we introduce in this paper the cross-modal consistent knowledge distillation network (CM-CKD). The CM-CKD backbone is composed of a shared-weight network (SWN) affixed with two non-shared projection heads – each for face and periocular. Sharing the early weights (through SWN) and setting the later weights independent (through projection heads) permits the two biometric modalities to learn the common low-level features while rendering for each
projection head the modality-specific features. In principle, the CM-CKD exercises knowledge transfer from face to periocular in both implicit and explicit manners. In place of the typical batch normalization, the SWN trunk is interleaved by the consistent batch normalization (CBN) to synchronize the statistics discrepancies for an implicit distillation.

On the other hand, the explicit distillation is substantiated by means of a bidirectional distillation loss term, denoted by consistent knowledge distillation (CKD) loss in this paper. Unlike other unidirectional KD models whereby only the teacher networks guide the student networks, the bidirectional CM-CKD also distills student knowledge in response to the teacher. This is to cultivate a consistent state between the two modalities. Our experiments disclose that the CM-CKD-learned periocular features enjoy better subject-wise cluster separation, thereby refining the overall accuracy performance. Furthermore, we show a connection exists in between CM-CKD and the label smoothing regularization (LSR) [34], which justifies why CM-CKD offers an improved generalization.

In summary, our contributions are as follows:

1. We exploit the stronger face biometrics to enhance the periocular embedding learning in the wild. However, only the periocular modality is demanded during the identification or verification stage.

2. We propose a novel deep learning-based CM-CKD for embedding learning, of which the cross-modal distillation is substantiated by the CBN-enabled SWN and the CKD loss. We justify that CM-CKD generalizes better than the conventional KD in view of LSR.

3. Extensive experiments are conducted on six periocular datasets for identification and verification tasks. We demonstrate that the CM-CKD-learned features achieve the best performance in terms of rank-1 identification rate (%) and equal error rate (%). In the meantime, we disclose that CM-CKD improves the face and periocular baselines by 50% with respect to the relative gain index.

2. Preliminary
In this section, we briefly explain the knowledge distillation (KD) introduced in [20]. The primitive goal of KD is for network compression by means of distilling the pre-learned knowledge from a large cumbersome teacher network, which possesses high performance, to a compact student network of in terms of network architecture.

To be specific, given a labelled data set \( \{x_i, y_i\}, i=1,\ldots,n \) where \( x_i \) is the data and \( y_i \) is the label, the knowledge from the teacher’s prediction \( p_t^r \) is distilled to the student’s prediction \( p_r \) by minimizing the loss function as follows:
\[ L = (1 - \lambda)L_{CE} + \tau^2\lambda L_{KD} \]  
where \( L_{CE} = -y_i \sum \log p(y_i|x_i) \) is the cross-entropy loss with student network prediction \( p(y_i|x_i) = \text{softmax}(f(x_i)) \) and \( f(\cdot) \) is the logit of the network. \( L_{KD} \) is the KD loss defined as:

\[ L_{KD} = \sum_i KL(p_T^i(y_i|x_i) \| p_T(y|x_i)) \]  
Here, \( p_T(y|x) = \text{softmax}(f(x)/\tau) \) and \( p_T^i(y|x) = \text{softmax}(f^T(x)/\tau) \) are the smoothen student’s prediction and teacher’s prediction, respectively. The smoothness of the prediction is regulated by the temperature term \( \tau \) and \( KL(\cdot \| \cdot) \) is defined as the KL divergence. The hyperparameter \( \lambda \) decides the contribution of the KD loss.

\begin{figure}[h]
\centering
\includegraphics[width=\textwidth]{image}
\caption{Generic network architecture for CM-CKD. A paired inputs contains a face and a periocular region feedforwarded to a shared-weights networks, followed by respective fully-connected projection head.}
\end{figure}

### 3. Cross-Modal Consistent Knowledge Distillation for Embedding Learning

In this section, the proposed cross-modal consistent knowledge distillation (CM-CKD) is detailed. In particular, we elaborate the three key ingredients of CM-CKD, i.e. the shared-weight networks (SWN), the consistent batch normalization (CBN), and the CKD loss. We also justify the relation of KD and label smooth regularization, which reveals how the CM-CKD can generalize better. Finally, we show how periocular identification and verification are carried out.

#### 3.1 Cross-Modal Shared-Weight Networks

The CM-CKD network is built upon the convolutional neural networks (CNN). As portrayed in Figure
2, it is composed of the SWN $h_\varphi$ parameterized by $\varphi$, i.e. a CNN trunk shared by face and periorcular, followed two projection heads for each face and periorcular modality. To be more precise, the face branch network is stacked with $h_\varphi$ and a face projection head $g_{\varphi}$, where $g_{\varphi,\varphi} = g_{\varphi} \circ h_\varphi$. On the contrary, the periorcular branch network consists of $h_\varphi$ and a periorcular projection head $f_\theta$, where $f_{\theta,\varphi} = f_\theta \circ h_\varphi$. We summarize the CM-CKD network construction in Table 2. Other relevant settings are presented in Section 4.2.

In the embedding learning stage, every labeled face-periorcular image pair is feedforwarded to the SWN. Accordingly, the SWN output for each face and periorcular, i.e. intermediate face and periorcular tensors, is dispatched to the corresponding modality-specific projection head. Our intuition is to first learn the low-level joint features from the two modalities, followed by the high-level modality-specific features. Note that the face and periorcular images are in different spatial dimension, yielding intermediate tensors of dissimilar sizes to each modality-specific projection head.

In fact, the CM-CKD construction resembles the one applied to multi-task learning (MTL). In general, MTL resolves multiple tasks concurrently based on the commonalities and differences between the tasks [35]-[37]. Similar to CM-CKS, MTL shares the hidden layers amongst all tasks, but equipping multiple task-specific output layers. An important difference is that CM-CKD is supplied with two inputs, whilst MTL accepts only one.

Since face and periorcular modalities are highly correlated, the weight sharing property of the SWN in favor of embedding learning by distillation. With this property, the SWN learns the common features along with regularity from both face and periorcular modalities. The feature embedding learned by the non-shared weight networks is likely biased to a specific modality and therefore prone to overfitting. Our ablation study in Section 4.3.1 attests that the SWN offers better generalizability by weight sharing.

### 3.2 Consistent Batch Normalization

On top of the cross-modal SWN, we introduce a modified batch normalization (BN) layer, known as consistent BN (CBN) in this paper. Typically, BN [38] is a well-known activation re-scaling technique formulated to normalize the channel-wise activations with respect to the batch mean $\mu_B$ and variance $\sigma_B^2$ over $m$ samples as follows:

$$\hat{x}_i = \frac{x_i - \mu_B}{\sqrt{\sigma_B^2}},$$

where $x_i$ is the feature activation in the hidden layer, $\mu_B = \frac{1}{m} \sum_{i=1}^{m} x_i$ is the per-dimension mean, and $\sigma_B^2 = \frac{1}{m} \sum_{i=1}^{m} (x_i - \mu_B)^2$ is the spatial dimension-wise variance. The $\hat{x}_i$ is scaled and shifted with the
learned parameters $\gamma$ and $\beta$ by $y_t = \gamma x_t + \beta$.

When dealing with multiple modalities, handling of BN turns out to be an important issue. This is because the input streams of different modalities are likely to have different running statistics during training.

![Figure 3](image)

**Figure 3.** Batch normalization operated by BN and CBN. In case of BN (a), each face and periocular input is normalized individually yielding a distinct statistic set for each face and periocular. On the contrary, CBN (b) first concatenates the two inputs along spatial dimensions computing only a single statistic set. Note that $H$, $W$, and $C$ refers to height, width and channel capacity for each face and periocular modality.

In our disposition, we train the SWN with two different input modalities simultaneously. When each face and periocular modality navigate a BN layer separately, each mini-batch computes a distinct statistic set of $\mu_B$ and $\sigma_B^2$. To address this issue, we concatenate two inputs in the spatial dimension-wise manner in order to share the channel-wise statistics so that the BN layer computes only a single batch statistics during the forward and backward propagation passes as illustrated in Figure 3. We coin this modified BN as consistent batch normalization (CBN) as it collects a consistent statistic set between the two modalities. The effectiveness of the CBN is validated in Section 4.3.1 (Configuration 4).

In brief, the proposed CBN enabled shared-weight network is devised for two different modalities, where one has more discriminative information than the other to achieve implicit distillation. After embedding learning, we only adopt $f_{\theta, \varphi}$, which is the periocular network for periocular feature extraction, so we can dispose of the face branch $g_{\varphi}$.
3.3 Consistent Knowledge Distillation

The cross-modal consistent knowledge distillation (CKD) proposed in this paper attempts to enhance the weak periocular attributes (target) by distilling in an explicit manner the face information beneficial to periocular identification and verification in the wild. To accomplish this goal, the CKD permits both face and periocular networks to co-learn from each another, thereby imposing a strong consistency between the two network prediction sets.

Let \( \{(x^p_i, x^f_i, y_i)\}, i=1,\ldots,n \) be a training dataset consisting of \( n \) labelled face-periocular pairs where \( x^f_i \) is a face image, \( x^p_i \) is the corresponding periocular image cropped from \( x^f_i \), and \( y_i \) indicates the annotated identity label. Mathematically, the prediction output for the face branch is expressed as:

\[
p^f_i(\cdot| x^f_i) = \text{softmax}(g_{\phi,\theta}(x^f_i)/\tau).
\]

whereas the prediction output for the periocular branch is defined as:

\[
p^p_i(\cdot| x^p_i) = \text{softmax}(f_{\theta,\phi}(x^p_i)/\tau).
\]

The proposed CKD is realized by minimizing \( L_{\text{total}} \) as follows:

\[
L_{\text{total}} = L^f_{CE} + L^p_{CE} + L_{CKD}
\]

Accordingly, \( L^f_{CE} \) and \( L^p_{CE} \) formulate the cross-entropy losses for each face and periocular branch as follows:

\[
L^f_{CE} = -\frac{1}{n} \sum_{i=1}^{n} \log p^f(y_i|x^f_i),
\]

\[
L^p_{CE} = -\frac{1}{n} \sum_{i=1}^{n} \log p^p(y_i|x^p_i)
\]

On the contrary, the \( L_{CKD} \) in (6) imposes consistency between the cross-modal predictions as follows:

\[
L_{CKD} = L_{KD-F2P} + L_{KD-P2F}.
\]

In principle, the \( L_{CKD} \) minimizes the sum of two distillation losses, namely \( L_{KD-F2P} \) and \( L_{KD-P2F} \), for bi-directional distillation as follows:

\[
L_{KD-F2P} = \tau^2 \frac{1}{N} \sum_{i=1}^{N} KL \left( \hat{p}^f_i(\cdot | x^f_i) \parallel p^f_i(\cdot | x^f_i) \right),
\]

\[
L_{KD-P2F} = \tau^2 \frac{1}{N} \sum_{i=1}^{N} KL \left( \hat{p}^p_i(\cdot | x^p_i) \parallel p^p_i(\cdot | x^p_i) \right),
\]

where \( \hat{p}^f_i(\cdot | x^f_i) \) and \( \hat{p}^p_i(\cdot | x^p_i) \) are the fixed copies of \( p^f_i(\cdot | x^f_i) \) and \( p^p_i(\cdot | x^p_i) \), respectively.
Based upon the face-to-periocular distillation loss $L_{KDF2P}$, the prediction from the face branch $p^F$ also learns from the periocular prediction $p^P$. We remark that the teacher network for the conventional knowledge distillation (KD), typically pre-trained, is halted from learning throughout the entire training session. On the assumption that the pre-learned teacher is incompetent at untangling the target problem, a well-generalized student network is hardly trained. On the contrary, the face network (teacher) in CM-CKD is adaptive, learning by minimizing $L_{KDF2P}$. Moreover, since CM-CKD learns both teacher and student networks jointly, the two-stage KD training phase is degenerated into a single-stage learning. This in turn leads the CM-CKD be more efficient than the conventional KD. We provide justifications in Section 3.3.

In a nutshell, the co-learning phase of face and periocular results in a strong similarity between $p^F$ and $p^P$. The consistency between the face and periocular predictions allows the periocular network to learn the well-generalized and discriminative embedding. We disclose in Section 4.4 that the periocular embedding learned from CM-CKD improves intra-class compactness and inter-class separation.

### 3.4 Relation with Label Smoothing Regularization

In this section, we justify the effectiveness of our CKD by analyzing it in the light of label smoothing regularization (LSR) [34]. To this end, we first show any type of KD is equivalent to LSR. The equivalence implies that any KD improves the generalization of the student as in LSR. Then, we reason that CKD offers better generalization than the conventional KD and label smoothing with uniform noise. In particular, we argue that, since CKD is not fixed but adaptively learns during the distillation process, it allows the student to find better local minima of the cross-entropy loss and thus better generalize than the counterparts from the conventional KD and label smoothing.

Here, face and periocular networks are called as teacher and student networks, respectively for convention and consistency. Suppose $H(y, p^S)$ denotes the cross-entropy loss with the label $y$ and student network prediction $p^S$. Let $p^T$ denotes the teacher’s prediction. $p^T_\tau$ and $p^S_\tau$ denote the re-scaled predictions of $p^T$ and $p^S$, respectively with temperature $\tau > 0$.

**Proposition 1.** Any KD is equivalent to LSR with a regularizer: namely, for $\lambda \in (0, 1)$ and $\tau > 0$

$$\left(1 - \lambda\right)H(y, p^S) + \lambda \tau^2 KL(p^T_\tau, p^S_\tau)$$  \hspace{1cm} (10)

is equivalent to

$$\alpha H(y', p^S) + \beta R(p^S)$$  \hspace{1cm} (11)

for some $\alpha, \beta > 0$ where $y'_s$ is the smoothed label of $y$. 

\[ y' = (1 - \gamma) y + \gamma p^T \]
with \( \gamma \in (0, 1) \). \( R(p) = \log \sum_j p_j^{(1/\tau)} \) is a regularizer.

**Proof.** Let \( z = (1 - \lambda) + \lambda \tau \), and set \( \alpha = Z, \gamma = \lambda \tau / Z, \beta = \lambda \tau^2 \). Then, simple algebra gives the equivalence.

Based on Prop 1, thus, we can view any KD is indistinguishable from label smoothing mechanism and vice versa. [34] shows that label smoothing, under certain conditions, improves generalization of \( p^S \). For example, label smoothing with uniform noise allows \( p^S \) to generalize better in classification [34]. [20] shows that KD with a fixed teacher as a label smoothing mechanism improves the generalization of \( p^S \) and is more effective than the label smoothing with uniform noise. In particular, label smoothing by KD allows the student network to learn features that are better class-wise separated. However, we argue that since the teacher prediction is fixed in KD, it may be too hard for the student to follow; in other words, the distillation loss term \( KL(p^T_j, p^S_j) \) may not reach an effective local minima.

Accordingly, we claim that allowing the teacher’s prediction \( p^T_j \) to adaptably adjust during the distillation and imposing bidirectional distillation by \( KL(p^T_j, p^S_j) + KL(p^S_j, p^T_j) \), is helpful at improving the generalization of student’s prediction \( p^S \) than KD. Figure 7 shows that our CKD indeed finds a better local minima of the distillation loss term \( KL(p^T_j, p^S_j) \), and the result in Table 12 verifies that CKD is more effective at class-wise (subject-wise) feature separation and generalization, thus better embedding learning for periocular biometric.

### 3.5 Embedding for Identification and Verification

Upon training completion, the non-target face projection head is simply dispensed. In another word, the inference stage involves only on the shared-weight networks (SWN) and the periocular projection head. As depicted in Figure 4, we extract from the deep embedding layer the periocular feature vector for the subsequent identification and verification tasks.

The identification task requires gallery and probe sets. The gallery is collected by forwarding the gallery images to the trained periocular network and extracting the feature vectors \( v_G \) paired with each of their identities. Then, the probe images are forwarded to the network and their feature vectors \( v_P \) are compared with every \( v_G \) by the cosine similarity measure, \( \cos(\cdot) \):

\[
\cos(v_G, v_P) = \frac{v_G \cdot v_P}{\|v_G\|_2 \|v_P\|_2},
\]

where \( \|\cdot\|_2 \) is defined as L2 norm. The identity with the highest cosine similarity score is chosen as the identity.
Different from 1-to-N matching in the identification task, the 1-to-1 verification task determines if the two images of a pair belong to the same (positive) or different (negative) identities. The generic pipeline for both identification and verification tasks are illustrated in Figure 4.

![Figure 4: Generic CM-CKD pipeline for identification and verification tasks.](image)

4. Evaluation and Analysis

A series of experiments are performed to evaluate the cross-modal CKD-learned features for periocular identification or verification. This section describes the benchmarking datasets, the empirical settings and configurations, the ablation analyses, followed by a thorough comparison and discussion.

4.1 Dataset Description

Our experiments employ a summation of six datasets, including Ethnic [10], PubFig [40], FaceScrub [41], Imdb/wiki [42], AR [43], and YTF [44]. For training and validation purposes, a random subset is sampled from VGGFace [45] and Ethnic. The unseen testing set (composing of the gallery and the probe images) is configured in the open-world setting for a generally more challenging problem, where some
identities are non-existent in the training set. We summarize the sample distribution for each individual training/validation/testing set in Table 1.

Table 1: Sample distribution for training/validation/testing sets.

| Datasets         | # Iden. | Learning Set # | Testing Set # |
|------------------|---------|----------------|---------------|
| VGGFace/Ethnic   | 1,054   | 166,737        | # Gallery      |
|                  | 328     | 47,372         | # Probe 1      |
|                  |         |                | # Probe 2      |
|                  |         |                | # Probe 3      |
|                  |         |                | # Probe 4      |
| Pubfig           | 200     |                |               |
| FaceScrub        | 530     |                |               |
| IMDB Wiki        | 2129    |                |               |
| AR               | 100     |                |               |
| YTF              | 225     |                |               |

Training Dataset consists of 1054 subjects (577 subjects from the VGG face dataset and 477 subjects from the Ethnic dataset [10]) with 238,919 images as a whole for both face and periocular images. We crop the periocular region from each face image to yield the periocular images. We randomly split all images into 166,737 for training, 47,372 for validation.

Ethnic Dataset [10] contains five ethnic groups: Asian, African, White, Middle Eastern, and Latin American. The subjects are mostly sportsman, celebrities, and politicians; and the images are acquired under unconstrained environments. There are 329 subjects in total (disjoint with the 477 subjects in the training dataset); 1,645 gallery images which are 5 images for each subject, and 24,171 probe images.

PubFig Dataset [40] composed of 200 subjects which come from real-life images collected from the internet. Images are acquired in the wild without any other settings. In our work, there are 3 probe sets. The gallery set has 9,221 images, and the probe sets have 6,138, 6,101, and 7,680 images respectively.

FaceScrub Dataset [41] consists of 530 subjects where images are acquired in the unconstrained environments. Images were acquired with various poses, illuminations, facial expressions and backgrounds. This dataset has two probe sets. The gallery set has 31,066 images, the first probe set has 21,518 images, and the second probe set has 27,292 images.

Imdb_wiki Dataset [42] collects 2,129 subjects where images are taken in the wild environment and have three probe datasets. The gallery set has 40,241 images, the first probe has 17,658 images, the second one has 15,252 images, and the third one has 16,273 images.
AR Dataset [43] consists of faces with varying illumination, expression, blur, and occlusion conditions. The set up for distance from the subject to camera and illumination conditions are controlled during acquisition, which means this dataset is not in the wild, but several distracting factors are added, therefore we can measure the robustness of the proposed method. The detail description of AR dataset can be found in [43]. Periocular images are acquired by cropping the face images from this dataset. Both gallery and probe set composed of 100 subjects. The gallery set contains 7 images per subject resulting in 700 images as a whole. We have done the experiments on four probe sets. The first probe set is blurred images with 4 different extents of Gaussian filters applied resulting in 2,800 images in total. The second probe set is conducted with various expression, illumination and blurring resulting in 1,400 images in total. The third probe set is occlusion with square zero values with 5 various sizes from 10×10 - 50×50 resulting in 3500 images in total. The last probe set is occlusion with scarf resulting in 600 images in total.

YouTube Face (YTF) Dataset [44] is an online YouTube video repository for 1,595 identities capturing a wide range of awful visual variations, including low-resolution and motion blurred footages. The YTF evaluation protocol is originally set up for face verification. For periocular identification, we single out 225 subjects with at least four videos per subject in our experiments where three videos are designated for gallery set and the remaining for probe set. This results in 150,259 gallery images and 36,995 probe images.

4.2 Experiment Settings
We implement the CM-CKD model by PyTorch [46] and it is trained with a single NVidia GeForce
RTX2080Ti GPU. The input dimension for face and periocular images is affixed to 128×128 pixels and 48×128 pixels, respectively. The CM-CKD backbone is ResNet-18 [47] built on top of a shared-weight network (SWN), followed by a (non-shared) modality-specific projection head for each individual face and periocular. The SWN is constituted by three residual groups (indicated by Res-1 to Res-3), each of which contains two squeeze-and-excitation (SE) residual blocks [48]. On the other hand, each modality projection head is interleaved by a single SE residual block (Res-4), a deep embedding layer, and a logit layer. The complete network construction is detailed in Table 2.

Table 2. Network construction for cross-modal CKD, including shared-weight networks and non-shared modality-specific projection heads built on top of multiple squeeze-to-excitation residual groups. Note that k, s, p, and c indicate kernel size, stride, padding and output channels, respectively.

| Shared-Weight Networks (ResNet-18) | *Residual Block |
|-----------------------------------|-----------------|
| Module/Layer | Descriptions | Module/Layer | Descriptions |
| Conv2d | k=3×3, s=1, p=1, c=64 | CBN |
| CBN |
| ReLU |
| Residual Group 1 (Res-1) | 2×Residual Blocks*, c=64 | ReLU |
| Residual Group 2 (Res-2) | 2×Residual Blocks*, c=128 | Conv2d | k=3×3, s=1, p=1 |
| Residual Group 3 (Res-3) | 2×Residual Blocks*, c=256 | SE Block |

| Face Projection Head (Non-Shared) | Periocular Projection Head (Non-Shared) |
|-----------------------------------|----------------------------------------|
| Module/Layer | Descriptions | Module/Layer | Descriptions |
| Residual Group 4 (Res-4) | 2×Residual Blocks* | 2×Residual Blocks* |
| BN |
| Flattening |
| FC-1, BN |
| (Deep Embedding Layer) | in=3×8×512, out=512 | in=8×8×512, out=512 |
| (Logit Layer) | in=512, out=# of subjects | in=512, out=# of subjects |

The entire training phase is optimized by Stochastic Gradient Descent (SGD) for 90 epochs in total, where the learning rate is initialized to 0.1 to be decayed by a factor of 0.10 at epoch 30, 60, and 80. Other hyper-parameters include 0.9 for momentum, and 5.0×10⁻⁴ for weight decay, and the temperature τ is set to 2.5. On the other hand, the SWN is initialized with Kaiming initialization [49], the batch normalization layers are initialized with 1 for the weights and 0 for the bias terms, and the fully connected layers’ weights standard deviation Gaussian distribution are initialized with 0 mean and 0.01.

For the identification task, we report the average cumulative match curve (CMC) [48] for rank-1 up to rank-10 identification rate (%). In general, CMC illustrates the retrieval rate of probe samples from the gallery images. To be specific, with N samples of gallery set, M probe sets are tested by matching
with every gallery samples. If the ground truth identity is found at rank $k$, the scores of rank-$k$ and below increase by 1. After testing for every probe samples, the whole scores are divided by the number of probe set images $M$ and we display from the identification rate at rank-1 to rank-10 of CMC. For each dataset, there is one gallery set and one or more probe sets which are described in Table 1. However, to test as much as we can, we consider every case which can be considered and average the result. For example, for the PubFig dataset, it has 1 gallery set and 3 probe sets, and we tested 12 cases of gallery-probe combinations from the given 4 sets (1 gallery+3 probe) and average the 12 results.

To probe into the verification task, we randomly select 4 samples per subject from the gallery set for every dataset. Each sample compares with every other one - resulting in $4 \times 3 \times$ (number of identities)/2 positive cases and $4 \times 4 \times$ number of identities x (number of identities – 1)/2 negative cases in total. For example, in case of PubFig, there are 200 subjects so we can extract 800 images. There would be 1,200 positive cases and 318,400 negative cases in total. We show the Equal Error Rate (EER) and the interpolated average ROC curve over all 6 datasets.

4.3 Experimental Results and Discussion

4.3.1 Ablation Study

Recall the CM-CKD model integrates three key components – the shared-weight networks (SWN), the consistency batch normalization (CBN), the CKD loss (encapsulating both face-to-periocular KL loss and periocular-to-face KL loss). For an extensive ablation analysis, the CM-CKD model is degenerated into five basic configurations as follows:

1) excluding CKD loss, i.e. with cross-entropy (CE) loss only.
2) replacing the bi-directional CKD loss with a one-way face-to-periocular distillation $L_{KD-F2P}$ alone,
3) unfolding the SWN trunk with two separate (non-shared) networks,
4) replacing the original batch normalization instead of the proposed CBN,
5) replacing the original face projection head by a new periocular projection head – analogous to deep mutual learning [50] that trains two independent networks of the same modality.

These basic configurations are compared to face baseline, periocular baseline, and the complete CM-CKD configuration in Table 3.
Firstly, we present the individual face and periocular networks learned with cross-entropy loss only, i.e. $L_{CE}^{F}$ and $L_{CE}^{P}$, respectively. In the following sections, these two instances are dubbed face baseline and periocular baseline. To explore the degree of periocular performance improvement with the supervision of face modality, we quantify the ratio of relative performance gain with respect to both face and periocular baselines in terms of Relative Gain (RG). We express the RG formulation as follows:

$$\text{Relative Gain, RG} = \frac{\text{avg}_C - \text{avg}_P}{\text{avg}_F - \text{avg}_P} \times 100\%$$ (14)

Note that $\text{avg}_C$ in (14) indicates the average performance for each basic configuration in Table 3. For comparisons, the average ROC and CMC are portrayed in Figure 6.

Table 4 Rank-1 identification rates for six benchmarking datasets and their average. The highest accuracy are highlighted in bold font.

| Method          | Ethnic | Pubfig | Facescrub | Imdb_wiki | AR  | YTF  | Avg. | RG  |
|-----------------|--------|--------|-----------|-----------|-----|------|------|-----|
| Face Baseline   | 97.48  | 99.19  | 98.66     | 90.18     | 91.74| 78.52| 92.63|     |
| Periocular Baseline | 92.82  | 94.75  | 96.44     | 77.63     | 93.50| 56.78| 85.49|     |
| Basic Config. (1) | 93.21  | 96.14  | 96.18     | 78.63     | 94.59| 57.57| 86.05|  8  |
| Basic Config. (2) | 93.09  | 95.92  | 96.48     | 77.81     | 94.39| 59.52| 86.20| 10  |
| Basic Config. (3) | 95.67  | 97.43  | 97.07     | 83.17     | 96.46| 63.42| 88.87|  47 |
| Basic Config. (4) | 95.50  | 97.06  | 97.20     | 82.92     | 95.63| 62.85| 88.53|  43 |
| Basic Config. (5) | 94.11  | 96.36  | 96.26     | 79.63     | 94.83| 57.89| 86.51| 14  |
| CM-CKD          | 95.75  | 97.45  | 97.32     | 83.93     | 96.11| 63.18| 88.96| 49  |

Table 5 EER verification rates with six benchmarking datasets and their average.

| Method          | Ethnic | Pubfig | Facescrub | Imdb_wiki | AR  | YTF  | Avg. | RG  |
|-----------------|--------|--------|-----------|-----------|-----|------|------|-----|
| Face Baseline   | 4.06   | 3.50   | 1.58      | 4.79      | 4.23| 11.84| 5.00 |     |
| Periocular Baseline | 7.33   | 7.26   | 3.80      | 8.37      | 12.23| 18.34| 9.55 |     |
| Basic Config. (1) |  7.21 |  7.47 |  4.12 |  8.40 |  8.62 | 18.00 |  8.97 |  13 |
|-------------------|-------|-------|-------|-------|-------|-------|-------|-----|
| Basic Config. (2) |  7.06 |  7.12 |  3.63 |  8.11 |  8.61 | 16.89 |  8.57 |  22 |
| Basic Config. (4) |  6.03 |  6.03 |  3.02 |  6.76 |  7.50 | 15.73 |  7.51 |  45 |
| Basic Config. (5) |  6.88 |  6.93 |  4.08 |  8.36 |  9.84 | 18.39 |  9.08 |  10 |
| CM-CKD            |  5.61 |  5.48 |  3.13 |  6.53 |  6.77 | 15.16 |  7.11 |  54 |

As summarized in Table 4 and Table 5, the proposed CM-CKD model improves the average rank-1 identification rate across all benchmarking datasets by 3.47%. In the meantime, decreases the average EER across six datasets by 2.44% comparing with the periocular baseline. The RGs are 49% for identification and 54% for verification, which are the considerable performance gains over the baselines.

Secondly, ablation studies analyze how each of the components contributes to our proposed method. Configuration (1) uses a shared-weight network with CBN without CKD loss. The RGs for this configuration are 8% for identification and 13% for verification. The experiment validates the merit of CBN enabled shared-weight network.

![Figure 6. Average CMC and Average ROC curve of our method and ablation studies of the 6 datasets.](image-url)

In comparison to the periocular baseline, we discern that Configuration (2) with only unidirectional KD from face-to-periocular intensifies the average rank-1 identification rate by 0.71% and the average EER by 0.98%. However, the complete CM-CKD configuration shows a more significant improvement by 2.76% and 1.46% respectively for both tasks. For further investigation, we examine the training and the validation distribution for each bidirectional and unidirectional KD (with and without the CKD loss term). We observe from Figure 7 that the CM-CKD model trained from the CKD loss term (black line) convergences better compared to the unidirectional counterpart (orange line). In another word, the KL divergence term from periocular-to-face contributes largely by learning consistency between the face and periocular modalities that extends the degree of KD. This demonstrates the aptitude of bidirectional KD.
For Configuration (3), we learn two non-shared face and periocular networks from scratch, each with a modality-specific cross-entropy loss alongside of the CKD loss. The RG for identification is reported to be 47%, specifically 2% below the RG accomplished by CM-CKD. However, the RG for EER is 41% for Configuration (3) whereas ours achieves 54%. This implies that SWN is indispensable for the CM-CKD model.

We accentuate in Configuration (4) the effect of CBN where the typical batch normalization is trained for each periocular and face independently. On average, we substantiate that CBN improves the rank-1 identification rate and EER by 0.43% and 0.4%, respectively. Without CBN activated, we observe that this configuration is also outperformed by Configuration (3) (with two non-shared networks). Thus, to maximize the benefits stemmed from the SWN construction, the proposed CBN is necessary to stabilize the face and periocular statistics in a consistent manner.

**Figure 7.** The difference of $L_{KD-F2P}$ when the network is trained with one-directional KL divergence (orange), and our proposed consistent distillation loss (bidirectional KL divergence) (black) while training each network. (Left) shows the training $L_{KD-F2P}$ loss and (right) shows the validation $L_{KD-F2P}$ loss. The rest of the settings are same. Training with consistent distillation loss explores a better minimum point than training with one-directional KL divergence.

Last of all, Configuration (5) targets to examine the importance of the face modality in the CM-CKD training stage. Specifically, we demonstrate if the overall performance gain is attributed to SWN, or due to the employment of face as the KD source. Hence, we substitute the face projection head with that of periocular-based. Unlike deep mutual learning [50] with multiple unshared networks, Configuration (5) pursues SWN. With the absence of face modality for KD, Configuration (5) reports a poor RG, i.e. 14% for identification and 10% for verification, compared with 49% and 54% by CM-CKD. This attests our claim that the face modality is essential for periocular embedding learning by KD.
4.3.2 Comparison to Relevant State-of-the-Art

This section compares CM-CKD with other state-of-the-art (SOTA) targeting periocular identification or verification under the same evaluation protocol, including Attnet [13], OCLBCP [10], NIRPR [17], L2SR [19], and GLSR [20]. Similar to CM-CKD, we re-implemented these deep learning-based SOTA with ResNet-18 based on our training dataset.

We summarize in Table 6, Table 7 and Figure 8 that CM-CKD outperforms all relevant SOTA by a remarkable margin. It is noteworthy that Attnet requires a segmentation network on top of ResNet-18, and OC-LBCP is appended a sub-stream network to encode the texture descriptors.

Table 6 Performance comparison of CM-CKD and other periocular-based KD models, where the best rank-1 identification rate (%) is shown in bold.

| KD Models       | Ethnic | Pubfig | Facescrub | Imdb_wiki | AR | YTF | Avg.  |
|-----------------|--------|--------|-----------|-----------|----|-----|-------|
| Attnet [13]     | 87.29  | 90.93  | 91.07     | 65.35     | 82.38 | 45.78 | 77.13 |
| OC-LBCP [10]    | 92.43  | 95.48  | 95.64     | 76.21     | 94.92 | 56.51 | 85.20 |
| NIRPR [17]      | 83.70  | 87.95  | 87.80     | 58.24     | 78.25 | 43.00 | 73.16 |
| L2SR [19]       | 94.02  | 96.42  | 96.76     | 80.34     | 94.30 | 59.20 | 86.84 |
| GLSR [20]       | 94.24  | 96.62  | 96.96     | 80.82     | 94.67 | 59.77 | 87.18 |
| CM-CKD          | 95.75  | 97.45  | 97.32     | 83.93     | 96.11 | 63.18 | 88.96 |

Table 7 Performance comparison of CM-CKD and other periocular-based KD models, where the best EER (%) is shown in bold.

| KD Models       | Ethnic | Pubfig | Facescrub | Imdb_wiki | AR | YTF | Avg.  |
|-----------------|--------|--------|-----------|-----------|----|-----|-------|
| Attnet [13]     | 8.93   | 8.61   | 5.26      | 9.51      | 14.85 | 19.12 | 11.05 |
| OC-LBCP [10]    | 7.98   | 8.38   | 4.83      | 9.59      | 9.60  | 19.22 | 9.93  |
| NIRPR [17]      | 9.69   | 9.25   | 5.50      | 10.22     | 15.28 | 19.67 | 11.60 |
| GLSR [20]       | 6.54   | 6.39   | 3.20      | 7.34      | 9.24  | 18.06 | 8.46  |
| CM-CKD          | 5.61   | 5.48   | 3.13      | 6.53      | 6.77  | 15.16 | 7.11  |
4.3.3 Comparison to Other Knowledge Distillation Models

In this section, we compare CM-CKD to other prevailing KD models ([21], [22], [24]-[26], [27]-[29]) designated neither for periocular nor other biometrics, and we summarize these models in Table 8. We first learn the periocular-to-periocular distillation from a large cumbersome networks to a compressed network (training KD for network compression [21], [22], [24]-[26]), or within itself (also referred to as self-distillation [27]-[29]). Subsequent to that, the face-to-periocular distillation is analyzed. As most of these models exercise KD for only a single modality, we twist them into the cross-modality mode in our experiments. This is to demonstrate that the naïve execution of KD for cross-modal distillation does not perform well due to the difference of modality.

Table 8 State-of-the-art KD models designated neither for periocular nor other biometrics.

| KD Models | Descriptions |
|-----------|--------------|
| KD [21]   | The first instance of KD originated in 2015. |
| AT [22]   | The KD is carried out via attention (AT) maps instead of the predictive distribution of the teacher network. |
| SPKD [24] | Similarity-preserving KD (SPKD) guides the training of a student network such that input pairs that produce similar (or dissimilar) activations in the teacher network produce similar (or dissimilar) activations in the student network. |
| RKD [25]  | Relational KD (RKD) transfers mutual relations of data instances via distance-wise and angle-wise distillation losses that penalize structural differences in relations. |
| PKT [26]  | Probabilistic Knowledge Transfer (PKT) performs KD by matching the probability distribution of the data in the feature space instead of their actual representation. |
| BYOT [27] | Be Your Own Teacher (BYOT) transfers the knowledge in the deeper portion of the networks into the shallow networks. |
| DDGSD [28] | Data-distortion guided self-distillation (DDGSD) is one of the regularization techniques, forcing the outputs across different augmented data instances. |
| CWSD [29] | Class-wise Self Distillation (CWSD) penalizes the predictive distribution between different samples of the same label by minimizing the KL divergence. |
Periocular-to-Periocular Distillation

For exploration, ResNet-50 serves as the teacher network trained for a summation of 50 epochs by SGD. The initial learning rate is set to 0.1 to be decayed at epoch 25, 40, 45 by a ratio of 0.1. Other parameters include a weight decay of $5e^{-4}$ and a momentum of 0.9. The student network, on the other hand, recruits the ResNe-18 backbone trained according to Section 4.2. For the self-distillation models [27]-[29], only ResNet-18 is considered.

Table 9 Performance comparison of CM-CKD and other KD models for periocular-to-periocular distillation, where the best rank-1 identification rates (%) are shown in bold.

| KD Models       | Ethnic | Pubfig | Facescrub | Imdb_wiki | AR    | YTF    | avg     | RG    |
|-----------------|--------|--------|-----------|-----------|-------|--------|---------|-------|
| Periocular Baseline | 92.82  | 95.78  | 96.44     | 77.63     | 93.50 | 56.78  | 85.49   | 0     |
| AT Ocular [22]   | 93.68  | 96.24  | 96.41     | 79.03     | 94.24 | 58.87  | 86.41   | 13    |
| KD Ocular [21]   | 94.12  | 96.55  | 96.94     | 80.50     | 94.38 | 57.97  | 86.74   | 18    |
| PKT Ocular [26]  | 94.11  | 96.49  | 96.76     | 80.36     | 94.42 | 59.13  | 86.88   | 19    |
| SPKD Ocular [24] | 94.21  | 96.42  | 96.79     | 80.53     | 94.85 | 58.90  | 86.95   | 20    |
| RKD Ocular [25]  | 94.45  | 96.74  | 96.79     | 81.05     | 95.29 | 59.10  | 87.23   | 24    |
| DDDGSD Ocular [28] | 95.15  | 97.00  | 96.68     | 82.10     | 94.58 | 60.49  | 87.67   | 30    |
| BYOT Ocular [27] | 93.30  | 96.02  | 96.64     | 78.69     | 93.41 | 57.00  | 85.84   | 5     |
| SPKD Ocular [24] | 94.21  | 96.42  | 96.79     | 80.53     | 94.85 | 58.90  | 86.95   | 20    |
| RKD Ocular [25]  | 94.45  | 96.74  | 96.79     | 81.05     | 95.29 | 59.10  | 87.23   | 24    |
| DDDGSD Ocular [28] | 95.15  | 97.00  | 96.68     | 82.10     | 94.58 | 60.49  | 87.67   | 30    |
| BYOT Ocular [27] | 93.30  | 96.02  | 96.64     | 78.69     | 93.41 | 57.00  | 85.84   | 5     |

Table 10 Performance comparison of CM-CKD and other KD models for periocular-to-periocular distillation. The best EERs (%) are shown in bold.

| KD Models       | Ethnic | Pubfig | Facescrub | Imdb_wiki | AR    | YTF | Avg. | RG    |
|-----------------|--------|--------|-----------|-----------|-------|-----|------|-------|
| Periocular Baseline | 7.33   | 7.26   | 3.80      | 8.37      | 12.23 | 18.34 | 9.55 | 0     |
| AT Ocular [22]   | 7.36   | 7.18   | 3.94      | 7.99      | 11.00 | 17.22 | 9.12 | 10    |
| KD Ocular [21]   | 6.84   | 6.32   | 3.18      | 7.47      | 9.11  | 17.01 | 8.32 | 27    |
| PKT Ocular [26]  | 6.88   | 7.00   | 3.33      | 7.57      | 10.24 | 17.50 | 8.75 | 18    |
| SPKD Ocular [24] | 7.11   | 6.58   | 3.40      | 7.79      | 10.63 | 17.60 | 8.85 | 15    |
| RKD Ocular [25]  | 6.54   | 6.18   | 3.37      | 7.55      | 10.39 | 17.33 | 8.56 | 22    |
| DDGSD Ocular [28] | 6.54   | 6.45   | 3.52      | 6.92      | 8.90  | 17.77 | 8.35 | 27    |
| BYOT Ocular [27] | 7.22   | 7.50   | 3.30      | 8.25      | 10.38 | 17.65 | 9.05 | 11    |
| CWSD Ocular [29] | 7.79   | 6.74   | 2.90      | 7.78      | 9.73  | 18.77 | 8.95 | 13    |
| CM-CKD          | **5.61** | **5.48** | **3.13** | **6.53** | **6.77** | **15.16** | **7.11** | **54** |
According to Table 9, Table 10, and Figure 9, the generic KD models is shown extending the periocular baseline performance in varying degrees. In addition to CM-CKD, the two outperforming KD models are DDGSD [28] and RKD [25]. However, the proposed CM-CKD stands out by a remarkable margin. In a nutshell, we unveil in this experiment the shortcoming of the same-modal KD whilst accentuating the necessity of supplying a stronger modality for cross-modal KD.

**Face-to-Periocular Distillation**

We examine in this section the face-to-periocular distillation. For a fair comparison, the teacher is ResNet-18 which is the same as our face network (teacher). The teacher network of [21], [24] and [26] is realized by a pre-trained face model. However, the spatial dimension for the face attention map of [22] has to be adjusted since its size is different from the periocular attention map. We therefore downsample them with adaptive average pooling to have the same size as periocular. What we want to show in this section is the novelty of how CM-CKD distills from face to periocular and how the resulted model differs from the other KD variants.
Table 1 Performance comparison of CM-CKD and other KD models for face-to-periocular, where the best rank-1 identification rates (%) are shown in bold.

| KD Models       | Ethnic | Pubfig | Facescrub | Imdb_wiki | AR   | YTF  | Avg. | RG |
|-----------------|--------|--------|-----------|-----------|------|------|------|----|
| Periocular Baseline | 92.82  | 95.78  | 96.44     | 77.63     | 93.50| 56.78| 85.49| 0  |
| AT Face [22]    | 93.65  | 96.24  | 96.21     | 78.67     | 94.75| 58.88| 86.40| 13 |
| KD Face [21]    | 93.92  | 96.40  | 96.69     | 79.47     | 95.13| 59.07| 86.78| 18 |
| PKT Face [26]   | 92.45  | 95.13  | 95.30     | 74.60     | 93.74| 57.51| 84.79| 10 |
| SPKD Face [24]  | 93.47  | 96.17  | 96.08     | 78.68     | 94.44| 57.53| 86.06| 8  |
| CM-CKD          | 95.75  | 97.45  | 97.32     | 83.93     | 96.11| 63.18| 88.96| 49 |

Table 2 Performance comparison of CM-CKD and other KD models for face-to-periocular, where the best EER (%) are shown in bold.

| Method          | Ethnic | Pubfig | Facescrub | Imdb_wiki | AR   | YTF  | Avg. | RG |
|-----------------|--------|--------|-----------|-----------|------|------|------|----|
| Periocular Baseline | 7.33  | 7.26   | 3.80      | 8.37      | 12.23| 18.34| 9.55 | 0  |
| AT Face [22]    | 7.10   | 7.52   | 3.99      | 8.30      | 9.73 | 18.00| 9.11 | 10 |
| KD Face [21]    | 6.76   | 6.95   | 3.75      | 8.18      | 9.13 | 18.16| 8.82 | 16 |
| PKT Face [26]   | 8.24   | 8.25   | 4.46      | 8.87      | 9.12 | 17.72| 9.44 | 2  |
| SPKD Face [24]  | 7.41   | 7.89   | 4.17      | 8.37      | 9.47 | 17.71| 9.17 | 8  |
| CM-CKD          | 5.61   | 5.48   | 3.13      | 6.53      | 6.77 | 15.16| 7.11 | 54 |

Our revelation in Table 11, Table 12 and Figure 10 show that all comparing KD models improve the base performance, except PKT [26]. In view of the RG indexes, we observe that the performance gain is relatively marginal compared with the periocular-to-periocular distillation discussed in the preceding section. This suggests that these models are restricted to only self-distillation. In contrast, the proposed CM-CKD is again demonstrated superior to other comparing counterparts for cross-modal distillation.

4.4 Embedding Learning Evaluation

This section evaluates the aptitude of CM-CKD in embedding learning. In what follows, we measure the subject-wise cluster separation. We apply Davies Bouldin Index (DBI) [39], which is a measure for clustering algorithms. The DBI quantifies the intra-class and inter-class variations simultaneously as follows:

$$DBI = \sum_{i=1}^{c} D_i,$$

$$D_i = \max_{i \neq j} \left( \frac{S_i^2 + S_j^2}{\|\mathbf{c}_i - \mathbf{c}_j\|_2^2} \right), S_i = \frac{1}{N_i} \sum_{k=1}^{N_i} (\mathbf{v}_{ik} - \mathbf{c}_i)^2$$

where \(c\) refers to the number of subjects, \(\mathbf{c}_i\) and \(\mathbf{c}_j\) are the means for each cluster \(i\) and \(j\), provided
that \( i \neq j \). In the meantime, \( \| \| \) denotes the L2 distance, \( v_{ik} \) is the \( k \)-th component of \( v_i \) for intra-class cluster \( i \), and \( N_i \) represents the number of examples within \( v_i \) belonging to the \( i \)-th cluster. In general, a lower DBI value indicates higher inter-class variation and a lower intra-class variation, which signifies the class-wise cluster separation in our exposition.

For DBI, we use the training and validation dataset from learning set, the gallery set from Pubfig, Facescrub, Imdb_wiki and AR datasets, and the probe set from the Ethnic and the YTF datasets. We compare the DBIs for periocular embedding with respect to three configurations as follows:

1. Periocular embedding trained by CM-CKD.
2. Periocular embedding trained by conventional KD.
3. Periocular embedding trained by cross-entropy loss only (baseline)

As discussed in Section 3.4, face network is adaptively trained by the periocular network during distillation for Configuration (1) but not for Configuration (2).

As shown in Table 13, the periocular embedding with Configuration (1) shows lower DBI for all train, validation and average of benchmarking datasets over Configuration (2). Comparing to the baseline, we found that the training dataset has lower DBI, but higher DBI for the validation and benchmarking datasets. This validates that our model generalizes better in the embedding space.

| Network Configuration                               | Davies Bouldin Index (DBI) |
|----------------------------------------------------|----------------------------|
|                                                     | Training  | Validation | Testing |
| (1) Periocular Embedding Trained w/ Cross-Modal CKD | 1.46      | 1.69       | 1.82    |
| (2) Periocular Embedding Trained w/ Conventional KD | 1.55      | 1.87       | 2.04    |
| (3) Periocular Embedding Trained w/ CE loss         | 1.38      | 1.77       | 2.04    |

5. Conclusion
We presented in this paper a deep learning-based periocular embedding learning paradigm by means of cross-modal consistent knowledge distillation (CM-CKD) for periocular identification or verification in the wild. We distilled and transferred important knowledge from the more informative face to the relatively less powerful periocular during the training stage, whereas the inference stage demanded no face inputs but only the periocular counterparts. The CM-CKD model was constituted by a consistent batch normalized shared-weight network and an effectual CKD loss that enabled bidirectional consistency distillation between the face and periocular modalities. We associated the CKD loss to the label smoothing regularization and justified that CM-CKD could generalize better than the conventional KD. Our ablation experiments on six periocular in the wild datasets revealed that the proposed CM-
CKD rendered feature embedding of better subject-wise cluster separation.

Acknowledgment
This work was supported by the National Research Foundation of Korea (NRF) grant funded by the Korea government (MSIP) (NO. NRF-2019R1A2C1003306).

Reference
[1] Nigam, Ishan, Mayank Vatsa, and Richa Singh, Ocular biometrics: A survey of modalities and fusion approaches, Information Fusion 26 (2015) 1-35.
[2] Rattani, Ajita, and Reza Derakhshani, Ocular biometrics in the visible spectrum: A survey." Image and Vision Computing 59 (2017) 1-16.
[3] Alonso-Fernandez, Fernando, and Josef Bigun, A survey on periocular biometrics research, Pattern Recognition Letters 82 (2016) 92-105.
[4] Sun, Yi, Xiaogang Wang, and Xiaoou Tang, Deep convolutional network cascade for facial point detection, in: IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2013, pp. 3476-3483.
[5] Park, Unsang, Raghavender Reddy Jillela, Arun Ross, and Anil K. Jain, Periocular biometrics in the visible spectrum, IEEE Transactions on Information Forensics and Security 6(1) (2010) 96-106.
[6] Karahan, Şamil, Adil Karaöz, Ömer Faruk Özdemir, Ahmet Gökhan Gü, and Umut Uludag, On identification from periocular region utilizing sift and surf, in: 22nd European Signal Processing Conference (EUSIPCO), 2014, pp. 1392-1396.
[7] Mahalingam, Gayathri, and Karl Ricanek, LBP-based periocular recognition on challenging face datasets, EURASIP Journal on Image and Video processing 2013 (2013): 1-13.
[8] Cao, Zhicheng, and Natalia A. Schmid, Fusion of operators for heterogeneous periocular recognition at varying ranges, Pattern Recognition Letters 82 (2016) 170-180.
[9] Xu, Juefei, Miriam Cha, Joseph L. Heyman, Shreyas Venugopalan, Ramzi Abiantun, and Marios Savvides, Robust local binary pattern feature sets for periocular biometric identification, in: 4th IEEE International Conference on Biometrics: Theory, Applications and Systems (BTAS), 2010, pp. 1-8.
[10] Tiong, Leslie Ching Ow, Andrew Beng Jin Teoh, and Yunli Lee, Periocular Recognition in the Wild with Orthogonal Combination of Local Binary Coded Pattern in Dual-stream Convolutional Neural Network, in: International Conference on Biometrics (ICB), 2019, pp. 1-6.
[11] Long, Jonathan, Evan Shelhamer, and Trevor Darrell, Fully convolutional networks for semantic segmentation, in: IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2015, pp. 3431-3440.
[12] Proença, Hugo, and João C. Neves, Deep-prwis: Periocular recognition without the iris and sclera using deep learning frameworks, IEEE Transactions on Information Forensics and Security 13(4)
(2017) 888-896.

[13] Zhao, Zijing, and Ajay Kumar, Improving periocular recognition by explicit attention to critical regions in deep neural network, IEEE Transactions on Information Forensics and Security 13(12) (2018) 2937-2952.

[14] Oloyede, Muhtahir O., and Gerhard P. Hancke, Unimodal and multimodal biometric sensing systems: a review, IEEE Access 4 (2016) 7532-7555.

[15] Zanlorensi, Luiz A., Hugo Proença, and David Menotti, Unconstrained Periocular Recognition: Using Generative Deep Learning Frameworks for Attribute Normalization (arXiv:2002.03985).

[16] Goodfellow, Ian, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio, Generative adversarial nets, in: Advances in neural information processing systems, 2014, pp. 2672-2680.

[17] Hwang, Hyeonsang, and Eui Chul Lee, Near-Infrared Image-Based Periocular Biometric Method Using Convolutional Neural Network, IEEE Access 8 (2020) 158612-158621.

[18] Behera, Sushree S., Sapna S. Mishra, Bappaditya Mandal, and Niladri B. Puhan, Variance-guided attention-based twin deep network for cross-spectral periocular recognition, Image and Vision Computing 104 (2020) 104016.

[19] Jung, Yoon Gyo, Jaewoo Park, Leslie Ching Ow Tiong, and Andrew Beng Jin Teoh, Periocular recognition in the wild with learned label smoothing regularization, in: 25th International Conference on Digital Image Processing (ICDIP 2020), vol. 11519, 2020, pp. 115190T.

[20] Jung, Yoon Gyo, Cheng Yaw Low, Jaewoo Park, and Andrew Beng Jin Teoh, Periocular Recognition in the Wild Generalized Label Smoothing Regularization, IEEE Signal Processing Letters 27 (2020) 1455-1459.

[21] Hinton, Geoffrey, Oriol Vinyals, and Jeff Dean, Distilling the knowledge in a neural network (arXiv arXiv:1503.02531).

[22] Zagoruyko, Sergey, and Nikos Komodakis, Paying more attention to attention: Improving the performance of convolutional neural networks via attention transfer (arXiv:1612.03928).

[23] Yim, Junho, Donggyu Joo, Jihoon Bae, and Junmo Kim, A gift from knowledge distillation: Fast optimization, network minimization and transfer learning, in: IEEE Conference on Computer Vision and Pattern Recognition, 2017, pp. 4133-4141.

[24] Tung, Frederick, and Greg Mori, Similarity-preserving knowledge distillation, in: IEEE International Conference on Computer Vision, 2019, pp. 1365-1374.

[25] Park, Wonpyo, Dongju Kim, Yan Lu, and Minsu Cho, Relational knowledge distillation, in: IEEE Conference on Computer Vision and Pattern Recognition, 2019, pp. 3967-3976.

[26] Passalis, Nikolaos, and Anastasios Tefas, Learning deep representations with probabilistic knowledge transfer, in: European Conference on Computer Vision (ECCV), 2018, pp. 268-284.

[27] Zhang, Linfeng, Jiebo Song, Anni Gao, Jingwei Chen, Chenglong Bao, and Kaisheng Ma, Be your own teacher: Improve the performance of convolutional neural networks via self-distillation, in: IEEE International Conference on Computer Vision, 2019, pp. 3713-3722.

[28] Xu, Ting-Bing, and Cheng-Lin Liu, Data-distortion guided self-distillation for deep neural networks, in: AAAI Conference on Artificial Intelligence, vol. 33, 2019, pp. 5565-5572.
[29] Yun, Sukmin, Jongjin Park, Kimin Lee, and Jinwoo Shin, Regularizing class-wise predictions via self-knowledge distillation, in: IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2020, pp. 13876-13885.

[30] Aytar, Yusuf, Carl Vondrick, and Antonio Torralba, Soundnet: Learning sound representations from unlabeled video, in: Advances in neural information processing systems, 2016, pp. 892-900.

[31] Karlekar, Jayashree, Jiashi Feng, Zi Sian Wong, and Sugiri Pranata, Deep face recognition model compression via knowledge transfer and distillation (arXiv:1906.00619).

[32] Gupta, Saurabh, Judy Hoffman, and Jitendra Malik, Cross modal distillation for supervision transfer, in: IEEE conference on computer vision and pattern recognition, 2016, pp. 2827-2836.

[33] Wang, Mengjiao, Rujie Liu, Narishige Abe, Hidetsugu Uchida, Tomoaki Matsunami, and Shigefumi Yamada, Discover the effective strategy for face recognition model compression by improved knowledge distillation, in: IEEE International Conference on Image Processing (ICIP), 2018, pp. 2416-2420.

[34] Yuan, Li, Francis EH Tay, Guilin Li, Tao Wang, and Jiashi Feng, Revisiting Knowledge Distillation via Label Smoothing Regularization, in: IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2020, pp. 3903-3911.

[35] Ruder, Sebastian, An overview of multi-task learning in deep neural networks (arXiv:1706.05098).

[36] Maurer, Andreas, Massimiliano Pontil, and Bernardino Romera-Paredes, The benefit of multitask representation learning, The Journal of Machine Learning Research 17(1) (2016) 2853-2884.

[37] Wu, Sen, Hongyang R. Zhang, and Christopher Ré, Understanding and Improving Information Transfer in Multi-Task Learning (arXiv:2005.00944).

[38] Ioffe, Sergey, and Christian Szegedy, Batch normalization: Accelerating deep network training by reducing internal covariate shift (arXiv:1502.03167).

[39] Davies, David L., and Donald W. Bouldin, A cluster separation measure, IEEE transactions on pattern analysis and machine intelligence, 2 (1979) 224-227.

[40] Kumar, Neeraj, Alexander C. Berg, Peter N. Belhumeur, and Shree K. Nayar, Attribute and simile classifiers for face verification, in: IEEE 12th international conference on computer vision, 2019, pp. 365-372.

[41] Ng, Hong-Wei, and Stefan Winkler, A data-driven approach to cleaning large face datasets, in: IEEE international conference on image processing (ICIP), 2014, pp. 343-347.

[42] Internet Movie Database. in: IMDB. http://www.imdb.com

[43] Martinez A, Benavente R (1998) The AR face database, CVC Technical Report #24.

[44] Wolf, Lior, Tal Hassner, and Itay Maoz, Face recognition in unconstrained videos with matched background similarity, in: 2011 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2011, pp. 529-534.

[45] Parkhi, Omkar M., Andrea Vedaldi, and Andrew Zisserman, Deep face recognition, in: British Machine Vision Association (BMVC), 2015.

[46] Paszke, Adam, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen et al., Pytorch: An imperative style, high-performance deep learning library, in:
Advances in neural information processing systems, 2019, pp. 8026-8037.

[47] He, Kaiming, Xiangyu Zhang, Shaoqing Ren, and Jian Sun, Deep residual learning for image recognition, in: IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016, pp. 770-778.

[48] Hu, Jie, Li Shen, and Gang Sun, Squeeze-and-excitation networks, in: IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 7132-7141.

[49] He, Kaiming, Xiangyu Zhang, Shaoqing Ren, and Jian Sun, Delving deep into rectifiers: Surpassing human-level performance on imagenet classification, in: IEEE International Conference on Computer Vision (ICCV), 2015, pp. 1026-1034.

[50] Zhang, Ying, Tao Xiang, Timothy M. Hospedales, and Huchuan Lu, Deep mutual learning, in: IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 4320-4328.