

Article

HDC-Net: A Heterogeneous Dual-Stream Network with Supervised Contrastive Learning for Contactless Palmprint and Palm Vein Fusion Recognition

Zhiting Zhuang ¹, Fen Dai ^{1,*}, Yanyun Li ¹, Ze Xiong ¹, Fusheng Niu ¹ and Xiangqun Zou ^{2,3}

¹ College of Artificial Intelligence and Low-Altitude Technology, South China Agricultural University, Guangzhou 510642, China; 20243169053@stu.scau.edu.cn (Z.Z.); liyanyun2002@stu.scau.edu.cn (Y.L.); 20253172040@stu.scau.edu.cn (Z.X.); 20253169030@stu.scau.edu.cn (F.N.)

² College of Artificial Intelligence, Zhujiang College, South China Agricultural University, Guangzhou 510900, China; nkwavelet@163.com

³ Guangdong Engineering Technology Research Center for Artificial Intelligence Empowered Media Technologies and Applications, Guangzhou 510900, China

* Correspondence: daifen@scau.edu.cn

Abstract

With the advancement of contactless biometric technologies, improving recognition accuracy and robustness in unconstrained environments remains a significant challenge. To address insufficient feature representation caused by modality discrepancies, as well as non-compact feature distributions, we propose HDC-Net, a contactless palmprint and palm vein fusion recognition model based on a heterogeneous dual-stream network and supervised contrastive learning. Specifically, a heterogeneous dual-stream feature extraction architecture is designed to learn modality-specific representations from palmprint and palm vein images. Supervised contrastive learning is introduced to enhance intra-class compactness and improve inter-class separability. Furthermore, a cross-modal interaction fusion module is developed to facilitate complementary feature learning across modalities. Experimental results on Tongji, CASIA-MS, and the self-built SCAU-PM dataset demonstrate that the proposed method achieves equal error rates (EERs) of 0.05%, 0.21%, and 0.02%, respectively. These results indicate that the proposed method achieves reliable recognition performance across different datasets and provides a feasible approach for contactless palmprint and palm vein fusion recognition.

Keywords: palmprint and palm vein; multimodal fusion; heterogeneous dual-stream network; supervised contrastive learning; contactless biometrics



Academic Editor: Jules-Raymond Tapamo

Received: 18 May 2026

Revised: 28 June 2026

Accepted: 6 July 2026

Published: 8 July 2026

Copyright: © 2026 by the authors.

Licensee MDPI, Basel, Switzerland.

This article is an open access article distributed under the terms and

conditions of the [Creative Commons Attribution \(CC BY\)](https://creativecommons.org/licenses/by/4.0/) license.

1. Introduction

With the increasing demand for contactless identity authentication, biometric recognition has become increasingly important in practical applications. Compared with single-modal biometrics such as fingerprint and face recognition, palmprint and palm vein fusion recognition have attracted increasing attention due to their contactless acquisition and the complementarity of textural and structural features [1,2]. In complex, uncontrolled environments, multimodal fusion can partially compensate for the limitations of single modal information, thereby improving robustness [3]. Existing studies have systematically reviewed multimodal hand recognition methods [4].

Early palmprint recognition relied mainly on handcrafted features for identity representation. Two-dimensional Gabor phase encoding was employed for palmprint feature

extraction [1]. Later studies gradually shifted toward deep learning-based approaches. Under deep learning frameworks, parameter-sharing networks were proposed [5]. Spatial transformer networks were introduced for geometric alignment [6]. CompNet was proposed to extract texture features [7]. For palm vein recognition, local structure-based descriptors were used for feature extraction [8]. The effectiveness of deep neural networks has been validated [9]. In addition, lightweight models have been proposed to improve computational efficiency [10].

Early multimodal fusion methods mainly relied on handcrafted features and simple fusion strategies. Score-level fusion frameworks were introduced [11]. Subsequently, weighted score-level fusion methods were applied to palmprint and palm vein fusion recognition [12]. However, their robustness remained relatively limited. With the advancement of deep learning, multimodal fusion has gradually shifted toward deep feature learning. Fusion methods based on DCNN for palmprint and palm vein recognition were proposed [13]. Feature-level fusion methods have gradually attracted increasing attention [14]. Recently, feature-level fusion methods have been explored from the perspective of inter-modal correlation [3]. In recent years, hybrid fusion strategies [15], adaptive weighted fusion methods [16], image enhancement-based fusion recognition methods [17], and knowledge distillation-based single-stream fusion learning methods [18] have been proposed, further improving the collaborative representation capability of multimodal features. In addition, cross-modal matching and feature representation methods have also received attention, including feature representation approaches based on deep hashing and fuzzy commitment mechanisms [19], cross-modal matching methods based on collaborative hashing learning [20], and cross-modal feature construction methods based on geometric intersecting features of palmprint and palm vein [21]. Existing studies have achieved encouraging progress. Further research on enhancing the complementarity and representation capability of multimodal features, as well as optimizing feature space representation, remains worthwhile.

First, there is a mismatch between the feature extraction architecture and modality characteristics. Most existing fusion methods adopt symmetric backbone networks for feature extraction. However, palmprint patterns are dominated by fine-grained textural details, whereas palm veins capture vascular structural characteristics [2]. The two modalities differ in imaging mechanisms and feature representation. In this case, using identical network depth and channel configurations may lead to insufficient texture representation or redundant structural features, thereby degrading discriminative ability. Existing studies have mainly focused on fusion strategy design [15]. Cross-modal studies demonstrate that modality-specific feature extraction enhances feature representation capability [22]. Therefore, designing backbone architectures adapted to modality differences remains an important research direction. From the perspective of feature learning, different modalities contain distinct types of information and spatial distribution characteristics, resulting in different requirements for feature representation capability. Palmprint images are mainly characterized by fine-grained texture details and therefore require more effective preservation of texture information, whereas palm vein images are dominated by vascular topological structures and require greater emphasis on spatial structural representation. Therefore, adopting feature extraction architectures adapted to the characteristics of different modalities may help improve the complementarity and representation capability of multimodal features.

Second, a single optimization objective may lead to insufficient compactness in the feature embedding space. Existing training methods primarily rely on ArcFace loss [23] or Softmax loss, which enhance separability by enlarging inter-class margins but do not explicitly constrain distances between samples. In uncontrolled environments, variations in illumination and pose can increase intra-class variance, leading to more dispersed sample

distributions and unstable class boundaries, thereby degrading the model's generalization performance. Previous studies have improved robustness through feature decoupling [24], but generally do not explicitly introduce constraints to enhance intra-class compactness and improve inter-class separability. In contrast, supervised contrastive learning improves the feature space structure by enhancing intra-class compactness and improving inter-class separability [25], and further optimizes the feature space representation [26]. However, its application in palmprint and palm vein fusion recognition remains limited.

To address the above problems, this paper proposes HDC-Net, a contactless palmprint and palm vein fusion recognition model based on a heterogeneous dual-stream network and supervised contrastive learning. Inspired by MobileNetV3 [27] and MobileFaceNet [28], we construct a heterogeneous dual-stream backbone architecture. To address modality discrepancies, the palmprint branch integrates the SE module [29] to enhance texture representation, defined as the Texture Enhancement Branch (TE-Branch). The palm vein branch adopts the CoordAtt attention mechanism [30] to enhance spatial feature representation, defined as the Structure Perception Branch (SP-Branch). At the optimization level, supervised contrastive learning is introduced to enhance intra-class compactness and inter-class separability [26]. At the fusion level, inspired by the dual-attention mechanism of CBAM [31], we design a Cross-Palm Attention (CPA) module, enabling cross-modal feature fusion and collaborative representation via intermediate-layer spatial interaction.

The main contributions of this study are summarized as follows.

1. We propose HDC-Net, a contactless palmprint and palm vein fusion recognition model based on a heterogeneous dual-stream network and supervised contrastive learning. Specifically, a modality-specific feature extraction framework is constructed by designing a Texture Enhancement Branch (TE-Branch) for palmprint representation and a Structure Perception Branch (SP-Branch) for palm vein representation. This design adopts differentiated network structures according to the characteristics of different modalities, enhancing feature complementarity while preserving discriminative information from each modality, thereby improving the representation capability of the fused features.
2. A supervised contrastive learning loss is introduced to leverage label information, enhancing intra-class compactness and improving inter-class separability, thereby improving the discriminative ability of the model.
3. A Cross-Palm Attention (CPA) module is proposed to enable cross-modal feature interaction and fusion via coordinated channel and spatial attention.

Experimental results demonstrate the effectiveness and robustness of the proposed method across multiple datasets, further validating its applicability to contactless palmprint and palm vein fusion recognition.

2. Materials and Methods

2.1. Datasets

To evaluate the performance and robustness of HDC-Net in contactless multimodal fusion recognition, experiments were conducted on two public datasets and one self-constructed dataset. The detailed statistics are summarized in Table 1. The Tongji contactless palmprint and palm vein dataset [13,32] (hereinafter referred to as the Tongji dataset) contains visible-light palmprint images and near-infrared palm vein images from 300 subjects, comprising 600 classes. Images from the first acquisition session (Session 1) are used in this study, resulting in a total of 6000 multimodal sample pairs. The CASIA multispectral palmprint dataset [33,34] (hereinafter referred to as the CASIA-MS dataset) contains 200 classes. Based on wavelength differences, images captured at 460 nm are

selected as the palmprint modality, while those at 850 nm are used as the palm vein modality, resulting in 1200 multimodal sample pairs. In addition, a contactless palmprint and palm vein multimodal dataset (hereinafter referred to as SCAU-PM) is constructed, comprising 200 classes, to evaluate the model performance under complex conditions. In this dataset, palmprint and palm vein images are acquired under visible-light and near-infrared conditions, respectively. The dataset includes variations in illumination, pose, and background, with a total of 1537 sample pairs. The SCAU-PM dataset was collected at South China Agricultural University and contains images from 100 subjects, resulting in 200 classes from both left and right hands. Palmprint images were acquired using smartphone cameras under visible-light conditions, while palm vein images were captured using an 850 nm near-infrared imaging device. All images were collected in a contactless manner. To better simulate practical application scenarios, no strict constraints were imposed on hand pose, acquisition distance, or background conditions. Consequently, the dataset contains variations in background, viewing angle, and hand posture. The palmprint and palm vein images originate from the same subject and the same hand and are organized as paired multimodal samples, ensuring identity consistency across modalities.

Table 1. Detailed statistics of the three multimodal datasets used in the experiments.

Dataset	Number of Classes	Number of Sample Pairs	Total Number of Images
Tongji	600	6000	12,000
CASIA-MS	200	1200	2400
SCAU-PM	200	1537	3074

During the data preprocessing stage, the officially released region-of-interest (ROI) images provided by the Tongji dataset are used. For the CASIA-MS and SCAU-PM datasets, ROI extraction is performed on the original palmprint and palm vein images, followed by automatic cropping via keypoint localization and geometric normalization to obtain standardized palm regions. Finally, all ROI images are fed into the multimodal fusion network. As illustrated in Figure 1, representative multimodal ROI samples from the three datasets are presented.

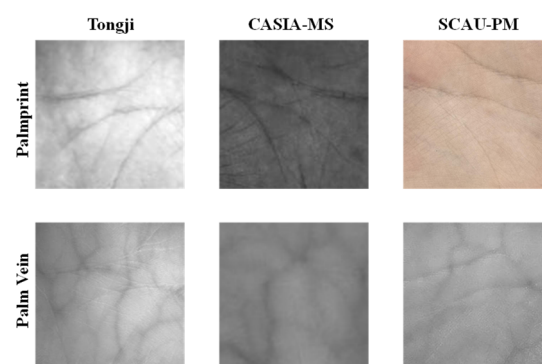


Figure 1. Examples of multimodal ROI samples from the three datasets.

2.2. Overall Framework

To address the efficient extraction and robust fusion of multimodal features in contactless palmprint and palm vein recognition, this study proposes HDC-Net, a contactless palmprint and palm vein recognition model based on a heterogeneous dual-stream network and supervised contrastive learning. As illustrated in Figure 2, the overall architecture of HDC-Net consists of three core components.

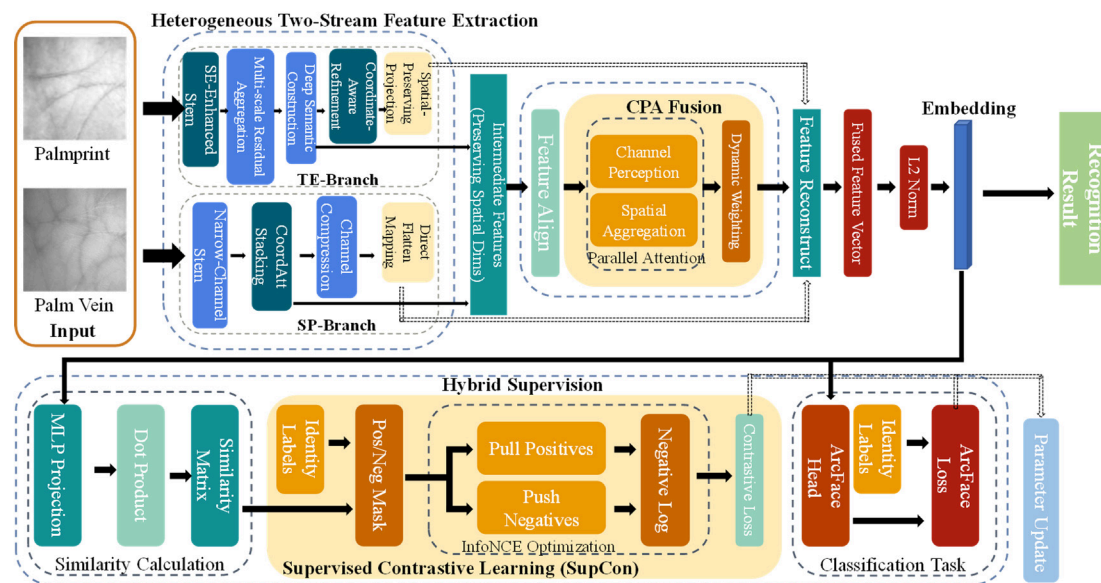


Figure 2. Overall architecture of HDC-Net.

1. Heterogeneous dual-stream feature extraction module. Considering the differences in texture and structural characteristics between palmprint and palm vein modalities, a Texture Enhancement Branch (TE-Branch) and a Structure Perception Branch (SP-Branch) are constructed for modality-specific feature extraction.
2. Supervised contrastive learning module. A hybrid optimization objective combining ArcFace classification loss and InfoNCE contrastive loss is employed to learn discriminative feature representations with intra-class compactness and inter-class separability in the embedding space.
3. Cross-Palm Attention (CPA) intermediate fusion module. Intermediate-level spatial features are extracted from feature maps, and cross-modal feature enhancement and fusion are achieved through parallel attention mechanisms and adaptive weighting.

2.3. Heterogeneous Dual-Stream Feature Extraction

Traditional dual-stream networks typically adopt fully symmetric backbone architectures such as dual-branch ResNet. However, such designs often neglect the modality gap between palmprint and palm vein images in terms of imaging mechanisms and structural distributions. Specifically, palmprint images mainly contain rich epidermal fine-grained texture information, whereas palm vein images exhibit relatively sparse subcutaneous vascular structures with topological characteristics. To address this issue, HDC-Net adopts a heterogeneous dual-stream architecture where two asymmetric feature extraction branches are designed to adapt to the characteristics of each modality. The detailed backbone structures and parameter settings are provided in Table 2. Given preprocessed palmprint and palm vein images, the model learns discriminative feature representations from both modalities and finally outputs an L2-normalized fused feature embedding.

Table 2. Parameter configuration of the heterogeneous dual-stream backbone network of HDC-Net.

Stage	TE-Branch (Palmprint)	SP-Branch (Palm Vein)	Output Size (C × H × W)
Input	1 × 128 × 128	1 × 128 × 128	PP: 1 × 128 × 128 PV: 1 × 128 × 128
Stage 1 (Stem)	Conv(3 × 3, s = 2, C = 16), SE Block	Conv(3 × 3, s = 2, C = 8)	PP: 16 × 64 × 64 PV: 8 × 64 × 64

Table 2. Cont.

Stage	TE-Branch (Palmprint)	SP-Branch (Palm Vein)	Output Size (C × H × W)
Stage 2 (Shallow)	1 × DW-ResBlock_CA(C = 32, s = 2); 3 × DW-ResBlock_CA(C = 32, s = 1)	1 × DW-ResBlock_CA(C = 16, s = 2)	PP: 32 × 32 × 32 PV: 16 × 32 × 32
Stage 3 (Mid-level)	1 × DW-ResBlock_CA(C = 64, s = 1); 5 × DW-ResBlock_CA(C = 64, s = 1)	4 × DW-ResBlock_CA(C = 16, s = 1)	PP: 64 × 32 × 32 PV: 16 × 32 × 32
Stage 4 (Deep)	1 × DW-ResBlock_CA(C = 128, s = 2); CoordAtt; Conv(1 × 1, C = 64)	1 × DW-ResBlock_CA(C = 8, s = 2)	PP: 64 × 16 × 16 PV: 8 × 16 × 16
Stage 5 (Embedding)	AdaptiveAvgPool(4 × 4), Flatten, FC(1024 → 512), FC(512 → 128)	Flatten(No pooling), FC(2048 → 128)	128

Note: PP denotes palmprint, and PV denotes palm vein; s denotes the stride of convolution; C denotes the number of output channels. DW-ResBlock_CA refers to the depthwise residual block with coordinate attention, and FC denotes the fully connected layer.

2.3.1. Texture Enhancement Branch (TE-Branch)

This branch aims to extract multi-scale texture representations using a deep network.

- (1) SE Enhancement Initialization. At the initial stage, a convolutional layer with a stride of 2 is applied to expand the channel dimension to 16, followed by the integration of a Squeeze-and-Excitation (SE) module. The SE module adaptively recalibrates channel-wise responses via global average pooling and a two-layer fully connected structure with a reduction ratio of 16, thereby suppressing illumination variations and background noise in shallow features.
- (2) Multi-scale Residual Aggregation and Deep Semantic Modeling. The features are processed by a stacked residual block architecture based on depthwise separable convolutions, consisting of 3 shallow residual blocks and 5 deep residual blocks, with the channel dimensions progressively increased to 64. Each residual block integrates a Coordinate Attention (CoordAtt) module, which captures directional contextual information via one-dimensional global pooling along horizontal and vertical axes, thereby enhancing the representation of principal palmprint ridge structures. Intermediate-layer features are preserved for subsequent cross-modal interaction.
- (3) Spatial Preservation Projection. After further spatial enhancement by CoordAtt, 4 × 4 adaptive pooling is used instead of global average pooling to preserve local spatial structure. The resulting feature maps are flattened and projected through a fully connected layer into a 128-dimensional palmprint feature embedding vector.

2.3.2. Structure Perception Branch (SP-Branch)

Due to the relatively sparse vascular structures in palm vein images, excessively deep or wide network architectures are prone to overfitting. Therefore, this branch adopts a structure-aware feature extraction strategy.

- (1) Narrow-channel initialization and Coordinate Attention stacking. The network is designed with a compact channel width of 8–16 channels to reduce model complexity. Feature extraction is conducted using a Coordinate Attention (CA) based module composed of four consecutive residual blocks. By capturing the tubular topological structure of subcutaneous vessels, this module produces intermediate vein features of size 16 × 32 × 32, which are subsequently fed into the CPA module for cross-modal interaction.
- (2) Structure-preserving channel compression. As shown in Figure 2, after obtaining intermediate features, stride convolutions with Coordinate Attention are introduced to jointly compress spatial and channel dimensions. The number of channels is reduced

- from 16 to 8, and the spatial resolution is reduced to $8 \times 16 \times 16$. This operation reduces computational redundancy while preserving local connectivity and global topological structure of veins, thereby enhancing feature compactness and robustness.
- (3) Compact feature embedding. After compression, the $8 \times 16 \times 16$ feature tensor is flattened and projected via a fully connected layer into a 128-dimensional vein embedding. This design avoids global average pooling to preserve vascular structural information.

2.4. Supervised Contrastive Learning

In uncontrolled non-contact multimodal acquisition scenarios, variations in hand pose (e.g., translation and rotation) and changes in ambient illumination may introduce feature variations among samples of the same subject in the feature space. Existing multimodal fusion models commonly rely on classification-based loss functions such as ArcFace for optimization. However, as illustrated in the left panel of Figure 3a, ArcFace primarily improves recognition performance by increasing inter-class feature margins, while providing limited constraints on intra-class compactness. Consequently, in complex scenarios, similarity distributions between intra-class and inter-class hard samples may overlap, thereby narrowing the decision boundary and degrading model robustness under uncontrolled variations. To further improve the structure of the feature space, we introduce a supervised contrastive learning strategy. As shown in the right panel of Figure 3b, this mechanism pulls intra-class samples closer and pushes inter-class samples apart, thereby improving intra-class compactness and inter-class separability and further enlarging the decision boundary to some extent.

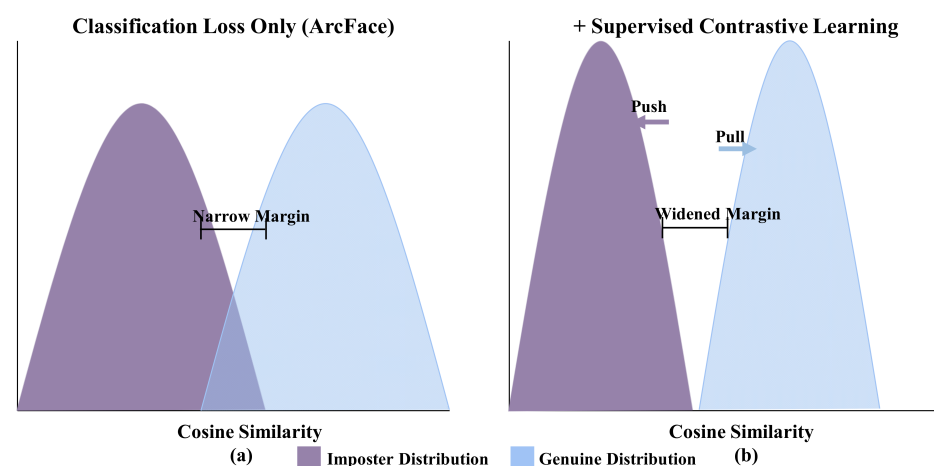


Figure 3. Illustration of the effect of supervised contrastive learning on feature space distribution. (a) Baseline model using ArcFace only. (b) Feature distribution after introducing supervised contrastive learning.

2.4.1. Identity-Label-Based Positive and Negative Sample Construction Strategy

Unlike self-supervised contrastive learning methods such as SimCLR that rely on image augmentation techniques including random cropping and flipping to construct positive pairs, this module leverages identity labels in the dataset to establish positive and negative sample relationships. Specifically, during the training data loading phase, a batch of samples with a fixed batch size is constructed. For each anchor palmprint and palm vein image pair, another sample belonging to the same identity but captured under different acquisition conditions such as pose or illumination variations is randomly selected as a positive sample, while all samples from different identities within the batch are treated as negative samples for contrastive learning.

Through this identity-label-based supervised sampling strategy, the model learns to minimize the feature distance between samples of the same identity while maximizing the distance between samples of different identities in the feature space, thereby improving feature discriminability and robustness.

2.4.2. Feature Space Optimization via InfoNCE Loss

At the network output, the anchor features and their corresponding positive and negative samples are first L2-normalized. Then, a dot-product similarity matrix is computed within a batch. Based on identity labels, a positive and negative mask matrix is constructed to distinguish positive and negative samples while excluding self-similarity.

The supervised InfoNCE loss L_{SupCon} is defined as follows:

$$L_{SupCon} = \sum_{i \in I} \frac{-1}{|P(i)|} \sum_{p \in P(i)} \log \frac{\exp(z_i \cdot z_p / \tau)}{\sum_{a \in A(i)} \exp(z_i \cdot z_a / \tau)} \quad (1)$$

where I denotes the index set of all samples within a batch, $P(i)$ denotes the set of positive samples corresponding to sample i , and $A(i)$ denotes the set of all samples excluding the sample itself. $z_i \cdot z_p$ represents the dot-product similarity between normalized features (i.e., cosine similarity), and τ is the temperature hyperparameter. In the experiments of this study, τ is set to 0.1. A lower temperature coefficient encourages the model to focus more on hard negative samples during training, i.e., negative samples that are close in the feature space. As shown in the right part of Figure 2, the loss function increases the similarity between positive pairs while reducing the similarity between negative pairs, making features of the same identity more compact in the embedding space and pushing different identities further apart, thereby effectively enlarging the decision boundary in regions with hard samples.

2.4.3. Hybrid Objective Joint Optimization

To further enhance the discriminative capability among different classes, HDC-Net retains the classification branch. In this study, the ArcFace loss L_{Arc} is adopted as the classification loss. The formulation is given as follows:

$$L_{Arc} = -\frac{1}{B} \sum_{i=1}^B \log \frac{\exp(s \cdot \cos(\theta_{y_i} + m))}{\exp(s \cdot \cos(\theta_{y_i} + m)) + \sum_{j \neq y_i} \exp(s \cdot \cos \theta_j)} \quad (2)$$

where the margin parameter m is set to 0.5 to enhance inter-class separability, the feature scaling factor s is set to 64, and B denotes the batch size. Finally, HDC-Net employs a joint optimization strategy, and the overall hybrid objective loss function is defined as:

$$L_{total} = L_{Arc} + \lambda L_{SupCon} \quad (3)$$

where λ denotes the weighting factor of the contrastive loss (set to 0.05 in this paper), which balances the contributions of the classification loss and the contrastive loss. The loss term L_{Arc} enhances global class discrimination and facilitates the formation of clear decision boundaries, while L_{SupCon} promotes a more compact feature distribution for samples of the same class. These two components are complementary in their optimization objectives, enabling the model to maintain strong discriminative capability while further improving intra-class compactness and inter-class separability in the feature space, thereby enhancing robustness in complex contactless scenarios.

2.5. Cross-Palm Attention (CPA) Intermediate Fusion Module

While the hybrid supervised loss is applied to the heterogeneous dual-stream network, modality discrepancies between palmprint and palm vein still exist, which may affect the fusion performance. To address this issue, the CPA module is introduced at the intermediate layer of the dual-stream network to adaptively refine spatial features. Specifically, the CPA module processes intermediate features from both streams through coordinated channel and spatial attention, and performs feature fusion via dynamic weighting. As illustrated in Figure 4, the CPA module operates on the intermediate features of palmprint and palm vein separately, and obtains fused features through concatenation and convolution operations, thereby enhancing cross-modal complementary representation.

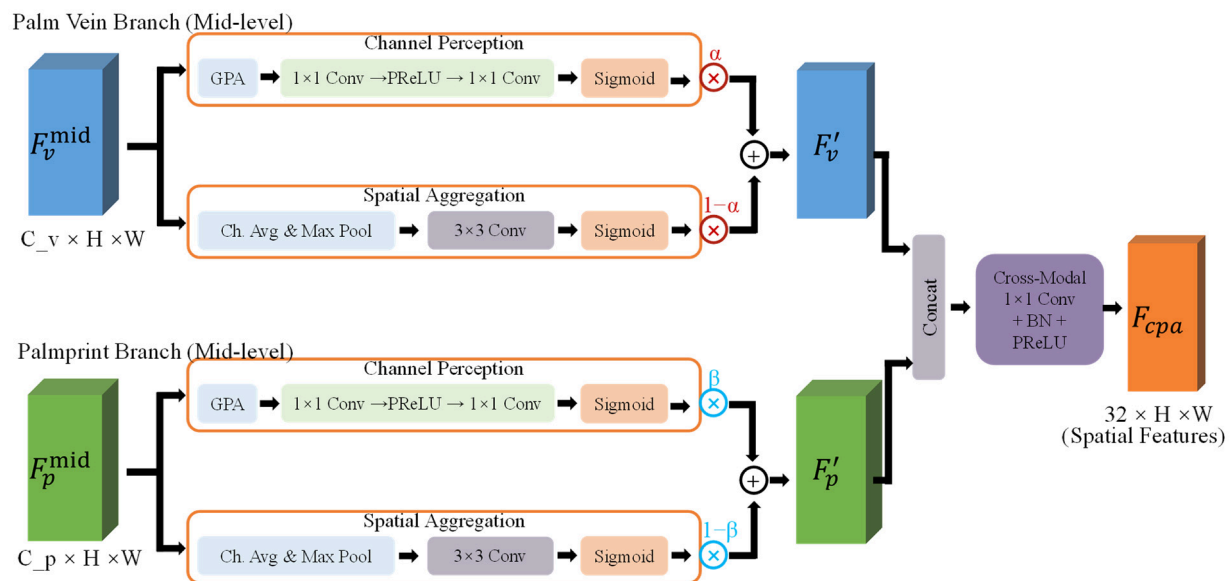


Figure 4. Internal structure of the Cross-Palm Attention (CPA) intermediate fusion module.

As illustrated in Figure 4, F_{cpa} corresponds to the intermediate fused features at the final output. The CPA module takes the intermediate features from the aforementioned dual-stream network (see the solid lines in Figure 1) as input and performs the following operations:

- (1) **Parallel attention refinement.** To facilitate subsequent processing, the bimodal features are first resized via bilinear interpolation to ensure consistent spatial dimensions. Subsequently, each modality is fed into parallel Channel Perception and Spatial Aggregation branches. In the Channel Perception branch, global average pooling (GAP) is applied to compress spatial information, followed by a 1×1 Conv $\rightarrow$ PReLU $\rightarrow 1 \times 1$ Conv structure to generate channel weights, which are activated by a Sigmoid function to obtain channel attention for feature reweighting. In the Spatial Aggregation branch, average pooling and max pooling are performed along the channel dimension, and the resulting features are concatenated and passed through a 3×3 convolution to generate spatial attention maps for feature refinement.
- (2) **Internal Dynamic Weight Regulation.** To balance the contributions of the channel branch and the spatial branch to the feature representation, learnable scalar parameters are introduced. As illustrated in Figure 4, for palm vein and palmprint features, the computation process is defined as follows:

$$F'_v = \alpha \cdot M_v^{ch}(F_v^{mid}) + (1 - \alpha) \cdot M_v^{sp}(F_v^{mid}) \quad (4)$$

$$F'_p = \beta \cdot M_p^{ch}(F_p^{mid}) + (1 - \beta) \cdot M_p^{sp}(F_p^{mid}) \quad (5)$$

where F'_v and F'_p denote feature tensors, F_v^{mid} and F_p^{mid} represent the intermediate features extracted from the aforementioned dual-stream network, α and β are learnable scalar parameters, and $M^{ch}(\cdot)$ and $M^{sp}(\cdot)$ correspond to the outputs of the channel branch and the spatial branch, respectively. The introduction of dynamic weighting enables the network to adaptively adjust the contributions of the channel and spatial branches according to variations in the input features.

- (3) Feature Concatenation and Global Skip Reconstruction. The processed feature tensors are concatenated along the channel dimension. Subsequently, a 1×1 convolution (followed by a BN layer and PReLU activation) is applied for channel mixing and dimensionality reduction, yielding the intermediate fusion feature.

To interface with the downstream metric learning module, the intermediate fusion feature is further mapped to a 128-dimensional intermediate-level fusion feature via adaptive pooling and a fully connected layer. Two learnable parameters (activated by a Sigmoid function) are then used to perform weighted fusion between the intermediate-level fusion feature and the deep features (as shown by the dashed line in Figure 1) from the dual-stream branches. The averaging operation is adopted to maintain balanced contributions from both modalities without introducing additional parameters, thereby improving the stability of the fusion process. The computation process is defined as follows:

$$E_{fused} = \sigma(\omega_1) \cdot E_{cpa} + \sigma(\omega_2) \cdot \frac{E_{vein} + E_{print}}{2} \quad (6)$$

where E_{fused} denotes the fusion feature, E_{cpa} represents the intermediate-level fusion feature obtained by linear mapping of the intermediate fusion feature F_{cpa} , and E_{vein} and E_{print} are the 128-dimensional deep features described in Section 2.3, which are introduced via global skip connections to preserve the original information. $\sigma(\omega_1)$ and $\sigma(\omega_2)$ are used to balance the contribution ratios between the intermediate-level fusion feature and the deep features. The final fusion feature is fed into the hybrid supervision module for joint optimization.

2.6. Experimental Settings and Evaluation Metrics

All models are implemented using PyTorch 1.12.0 and trained on an NVIDIA GeForce RTX 3080 GPU (NVIDIA Corporation, Santa Clara, CA, USA). To evaluate the generalization capability on unseen identities, all datasets are split into mutually exclusive training and testing sets with a 4:1 ratio at the class level, following an open-set protocol. During training, the AdamW optimizer is adopted with 300 epochs. The initial learning rate is set to 5×10^{-4} and adjusted according to specific experiments, while the batch size is set to 8 and may also be tuned depending on experimental settings. Data augmentation strategies including random translation, scaling, and rotation are applied to improve model robustness. ArcFace loss is used as the supervised objective with fixed parameters ($m = 0.50$, $s = 64$). In HDC-Net, supervised contrastive loss is further introduced ($\tau = 0.1$, $\lambda = 0.05$) to construct a joint optimization objective. For fair comparison, all competing methods adopt the same loss configuration, while in ablation studies, each component is gradually introduced to evaluate its individual contribution. The hyperparameter settings used for training HDC-Net are summarized in Table 3.

To comprehensively evaluate the performance of HDC-Net in contactless multimodal biometric recognition, multiple metrics are adopted following the international standard ISO/IEC 19795-1:2021 [35].

Table 3. Hyperparameter settings used for training HDC-Net.

Hyperparameter	Value
Optimizer	AdamW
Initial Learning Rate	5×10^{-4}
Weight Decay	5×10^{-4}
Batch Size	8
Training Epochs	300
ArcFace Margin (m)	0.50
ArcFace Scale (s)	64
SupCon Temperature (τ)	0.10
Loss Weight (λ)	0.05

- (1) False Accept Rate (FAR) and False Reject Rate (FRR). FAR represents the probability that samples from different individuals are incorrectly accepted as a match, while FRR represents the probability that samples from the same identity are incorrectly rejected as non-matching pairs. The formulations are defined as follows:

$$\text{FAR}(T) = P(S_{\text{imp}} > T) \quad (7)$$

$$\text{FRR}(T) = P(S_{\text{gen}} \leq T) \quad (8)$$

where the intra-class score is denoted by S_{gen} , the inter-class score is denoted by S_{imp} , and the decision threshold is denoted by T .

- (2) Equal Error Rate (EER). EER is defined as the error rate at the threshold T^* where FAR equals FRR.
- (3) Strict Operating Point Performance. The performance is evaluated in terms of FRR at $\text{FAR} = 10^{-4}$, which measures the model's behavior under high-security requirements.
- (4) Detection Error Trade-off (DET) Curve. The DET curve is plotted on a logarithmic scale to illustrate the relationship between FAR and FRR, emphasizing performance differences in the low-error region.
- (5) Model Complexity Metrics. Model complexity is evaluated from three perspectives: parameter scale, storage cost, and computational complexity, quantified by the number of parameters, model size, and floating-point operations (FLOPs), respectively.

3. Results

3.1. Overall Recognition Performance of HDC-Net

To evaluate the overall performance of the proposed contactless multimodal fusion recognition model and analyze the improvements brought by multimodal fusion, this section conducts a comprehensive evaluation on the palm vein single-modal branch, the palmprint single-modal branch, and the multimodal fusion network. The evaluation metrics include Equal Error Rate (EER), False Reject Rate (FRR) under low False Accept Rate conditions ($\text{FRR@FAR} = 10^{-4}$), and Detection Error Trade-off (DET) curves. As shown in Table 4, the proposed single-modal feature extraction networks already demonstrate strong discriminative capability. On the Tongji dataset, the TE-Branch and SP-Branch achieve EERs of 0.16% and 0.27%, respectively. Under strict security conditions ($\text{FAR} = 10^{-4}$), the FRRs are 1.44% and 1.07%, indicating that the single-modal models maintain a certain level of robustness in complex scenarios, while leaving room for improvement. Building upon this, the proposed HDC-Net, which integrates heterogeneous dual-stream feature extraction, CPA, and supervised contrastive learning, achieves improved performance across all evaluation metrics. On the Tongji, CASIA-MS, and SCAU-PM datasets, the EERs are 0.05%, 0.21%, and 0.02%, respectively. Under the condition of $\text{FAR} = 10^{-4}$, the

corresponding FRR values are also reduced; for example, on the Tongji dataset, the FRR is 0.52%.

Table 4. Performance evaluation of single-modal branches and multimodal fusion model (%).

Method	Tongji		CASIA-MS		SCAU-PM	
	EER	FRR@FAR = 10^{-4}	EER	FRR@FAR = 10^{-4}	EER	FRR@FAR = 10^{-4}
SP-Branch Only	0.27	1.07	0.99	8.82	0.14	0.82
TE-Branch Only	0.16	1.44	0.76	19.00	0.18	1.88
HDC-Net	0.05	0.52	0.21	5.88	0.02	0.57

The DET curve results shown in Figure 5 demonstrate that HDC-Net consistently outperforms the single-modal models across different threshold ranges. The proposed method maintains stable performance in the low-error region, indicating strong feature discriminability and robustness. In summary, the proposed model achieves performance improvements over single-modal baselines and exhibits high recognition accuracy and stability in contactless biometric recognition tasks.

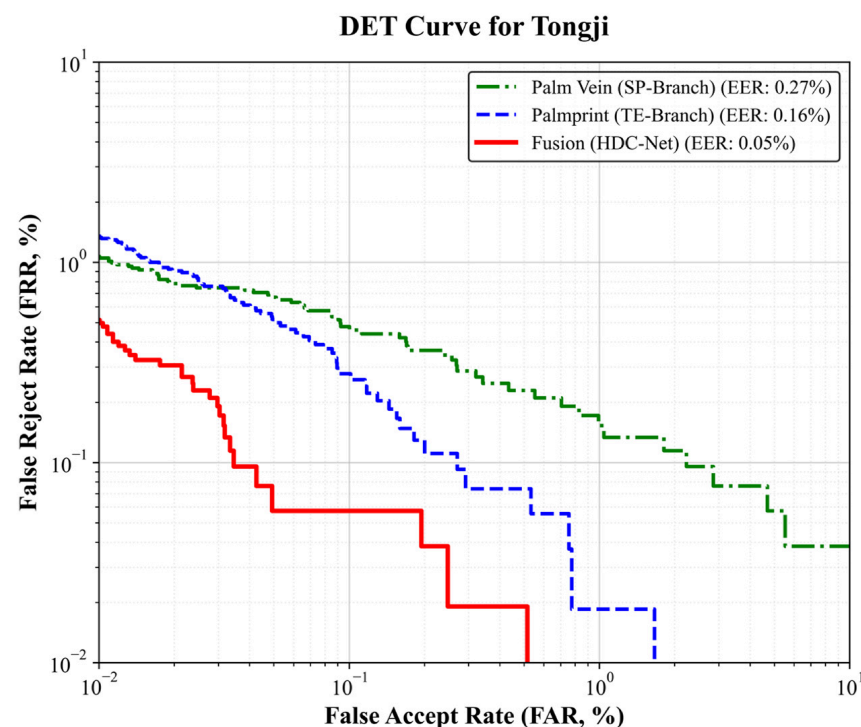


Figure 5. DET curve comparison of HDC-Net on the Tongji dataset.

To further evaluate the stability of HDC-Net under different data partitions, a 5-fold cross-validation experiment was conducted on the Tongji dataset. Specifically, all identity classes were randomly and evenly divided into five mutually exclusive subsets. For each fold, four subsets were used for training, and the remaining subset was used for testing. All experiments were conducted using the same training configurations and evaluation protocol as those used in the experiments described above. The results are summarized in Table 5. The EER and FRR@FAR = 10^{-4} values obtained from the five-fold cross-validation remain relatively stable, and the corresponding mean and standard deviation reflect the performance variation under different identity partitions. These results provide further support for the stability of HDC-Net under the adopted open-set evaluation protocol.

Table 5. Results of the 5-fold cross-validation of HDC-Net on the Tongji dataset (%).

Fold ID	EER	FRR@FAR = 10^{-4}
Fold 1	0.04	0.38
Fold 2	0.06	0.50
Fold 3	0.05	0.27
Fold 4	0.05	0.42
Fold 5	0.06	0.48
Mean $\pm$ Std	0.052 $\pm$ 0.008	0.410 $\pm$ 0.092

3.2. Ablation Study

To analyze the independent contributions of each key component in HDC-Net, ablation experiments are conducted on the Tongji dataset. The baseline model (Baseline, No.1) adopts a homogeneous dual-stream structure with simple feature-level concatenation, where both branches employ the same texture feature extraction architecture. On this basis, the heterogeneous dual-stream backbone, the CPA module, and the supervised contrastive learning (SupCon) strategy are introduced sequentially. According to the results in Table 6:

Table 6. Ablation study of the proposed components on the Tongji dataset (%).

No.	Heterogeneous Backbone	CPA	SupCon	EER	FRR@FAR = 10^{-4}
1	-	-	-	0.23	2.20
2	✓	-	-	0.14	0.84
3	✓	✓	-	0.11	2.14
4	✓	✓	✓	0.05	0.52

Note: “✓” indicates that the corresponding module is used, while “-” indicates that it is not used.

- (1) After introducing the heterogeneous dual-stream backbone (No.2), the EER decreases from 0.23% to 0.14%, and the FRR decreases from 2.20% to 0.84% (FAR = 10^{-4}), indicating that the heterogeneous feature extraction strategy effectively improves feature representation capability.
- (2) After introducing the CPA module (No.3), the EER is further reduced to 0.11%, demonstrating that cross-modal feature interaction enhances feature complementarity. It is worth noting that, although the introduction of CPA reduces the EER from 0.14% to 0.11%, the FRR@FAR = 10^{-4} increases from 0.84% to 2.14% on the Tongji dataset. This suggests that the improvement in overall discrimination performance does not necessarily translate into better performance under extremely strict operating points. A possible explanation is that CPA mainly enhances the discriminative capability of fused representations, while a small number of genuine pairs may still remain close to the decision boundary when only the classification objective is employed. Consequently, these samples are more likely to be rejected under FAR = 10^{-4} .
- (3) With the addition of supervised contrastive learning (SupCon) (No.4, i.e., the complete HDC-Net), the EER is further reduced to 0.05%, and the FRR reaches 0.52% at FAR = 10^{-4} , indicating that supervised contrastive learning enhances intra-class compactness and improves inter-class separability.

In summary, the proposed modules exhibit effective synergistic effects, jointly contributing to the overall performance improvement of the model.

The cosine similarity of features on the test set is analyzed from both macroscopic distribution and microscopic dynamic perspectives, as shown in Figure 6. Figure 6a,b presents the similarity distributions of all samples before and after introducing supervised contrastive learning, respectively. As shown in Figure 6a, although the classification loss alone achieves preliminary inter-class separation and intra-class compactness, the Genuine

distribution exhibits a left tail, and the Imposter distribution shows a right tail, indicating that the feature distribution remains insufficiently compact. It is worth noting that the left tail of the Genuine distribution in Figure 6a indicates that a small number of genuine pairs still remain close to the decision boundary when only the classification objective is employed. Under the strict operating point of $\text{FAR} = 10^{-4}$, these samples are more likely to be rejected, which provides a possible explanation for the increased FRR observed after introducing the CPA module without supervised contrastive learning. After introducing supervised contrastive learning (Figure 6b), the Genuine distribution shifts toward 1.0, while the Imposter distribution shifts toward 0.0, demonstrating enhanced intra-class compactness and improved inter-class separability.

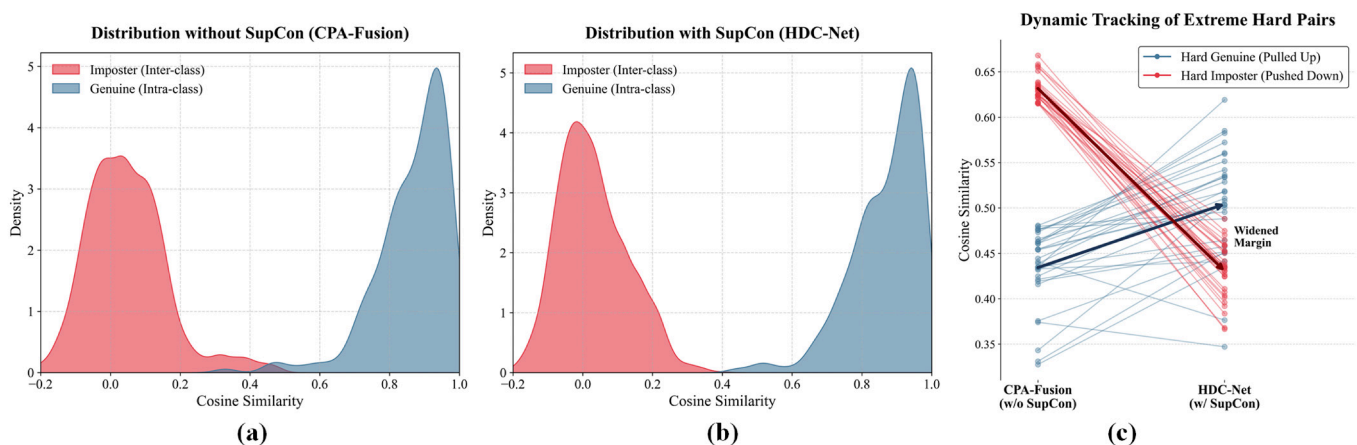


Figure 6. Visualization of the effect of supervised contrastive learning on feature space distribution. (a) Cosine similarity distribution before introducing SupCon. (b) Cosine similarity distribution after introducing SupCon. (c) Dynamic evolution of hard genuine and imposter sample similarities, where the arrows indicate that the genuine samples are pulled upward and the imposter samples are pushed downward.

Considering that misrecognition is mainly dominated by a small number of hard samples, 30 pairs of extremely hard genuine samples and 30 pairs of extremely hard imposter samples are further selected for dynamic analysis, as shown in Figure 6c. Under the w/o SupCon condition, some imposter samples exhibit higher similarity than genuine samples, indicating an overlap in the decision boundary. After introducing SupCon, the blue genuine samples are overall pulled upward, while the red imposter samples are pushed downward, forming an “X-shaped” evolution trend. These results demonstrate that SupCon effectively optimizes the distribution of hard samples and enlarges the discriminative margin, thereby improving model robustness.

To further quantitatively analyze the feature space variations shown in Figure 6, we compute the intra-class similarity, inter-class similarity, and embedding-space separation on the full test set. As shown in Table 7, after introducing supervised contrastive learning, the inter-class similarity decreases from 0.0475 to 0.0294, while the intra-class similarity changes from 0.8655 to 0.8679. Accordingly, the embedding-space separation increases from 0.8180 to 0.8385. These trends are consistent with the visual results shown in Figure 6a,b.

To further assess the influence of the supervised contrastive learning (SupCon) module on model performance from a statistical perspective, a Wilcoxon signed-rank test was conducted on the paired cosine similarity scores generated by the models before and after introducing SupCon on the test sets. The results are summarized in Table 8. On the Tongji, CASIA-MS, and SCAU-PM datasets, the obtained p -values were $p < 0.001$, $p = 0.0044$, and $p < 0.001$, respectively. These results reflect changes in the similarity score distributions

before and after introducing SupCon and provide a statistical basis for the associated performance differences.

Table 7. Quantitative comparison of feature space embedding with and without supervised contrastive learning.

Model	Intra-Class Similarity	Inter-Class Similarity	Embedding-Space Separation
Without SupCon	0.8655	0.0475	0.8180
With SupCon	0.8679	0.0294	0.8385

Note: Embedding-space separation is computed as intra-class similarity minus inter-class similarity.

Table 8. Wilcoxon signed-rank test results before and after introducing the SupCon module.

Dataset	Without SupCon EER (%)	With SupCon EER (%)	<i>p</i> -Value
Tongji	0.11	0.05	<0.001
CASIA-MS	0.38	0.21	0.0044
SCAU-PM	0.09	0.02	<0.001

3.3. Comparison with Existing Methods

To evaluate the effectiveness of each core component and the overall architecture of HDC-Net, comparative experiments are conducted on the Tongji, CASIA-MS, and SCAU-PM datasets. The analysis focuses on three aspects: heterogeneous dual-stream feature extraction, cross-modal fusion strategies, and overall performance with computational complexity.

3.3.1. Effectiveness Analysis of Homogeneous Networks and Heterogeneous Dual-Stream Networks

To evaluate the impact of feature extraction structures on multimodal fusion recognition performance, ResNet18, ShuffleNet, and MobileNet series are selected as homogeneous baselines for comparison. To ensure fairness, all multimodal models adopt channel concatenation for fusion and do not incorporate the contrastive learning loss. The results are shown in Table 9. In the single-modal evaluation, both the TE-Branch and SP-Branch demonstrate good recognition capability, validating the effectiveness of modality-specific feature extraction structures. In the multimodal evaluation, the fusion models generally outperform the single-modal models; however, homogeneous architectures (e.g., Dual-ResNet18 and Dual-MobileNetV3) show limited performance improvement, with EERs of 1.56% and 5.37%, and FRR@FAR = 10^{-4} of 15.14% and 55.68%, respectively. In contrast, the heterogeneous dual-stream network (TE-Branch + SP-Branch) achieves an EER of 0.14% and an FRR@FAR = 10^{-4} of 0.84%, indicating that modality-differentiated feature extraction is beneficial for improving feature discriminability.

Table 9. Performance results (%) of homogeneous networks and heterogeneous dual-stream networks under single-modal and multimodal settings.

Modality	Method	Tongji		CASIA-MS		SCAU-PM	
		EER	FRR@FAR = 10^{-4}	EER	FRR@FAR = 10^{-4}	EER	FRR@FAR = 10^{-4}
Palm Vein	ResNet18	3.12	32.62	3.96	54.07	0.83	10.87
	ShuffleNet	2.79	28.09	3.59	53.17	0.92	20.69
	MobileNetV1	3.67	40.76	3.81	63.12	1.59	19.95
	MobileNetV2	5.95	62.01	4.63	79.64	1.47	10.47
	MobileNetV3	6.26	56.07	7.47	81.67	2.80	57.24
	SP-Branch Only	0.27	1.07	0.99	8.82	0.14	0.82

Table 9. Cont.

Modality	Method	Tongji		CASIA-MS		SCAU-PM	
		EER	FRR@FAR = 10^{-4}	EER	FRR@FAR = 10^{-4}	EER	FRR@FAR = 10^{-4}
Palmprint	ResNet18	2.14	17.13	2.11	38.01	0.40	6.38
	ShuffleNet	2.32	27.15	4.44	71.95	0.58	10.96
	MobileNetV1	2.17	26.72	4.93	68.33	0.72	12.84
	MobileNetV2	2.83	44.26	3.74	68.10	0.69	26.17
	MobileNetV3	4.83	59.44	5.34	86.88	1.06	35.90
	TE-Branch Only	0.16	1.44	0.76	19.00	0.18	1.88
Multi-modal	Dual-ResNet18	1.56	15.14	1.01	19.91	0.17	0.64
	Dual-ShuffleNet	2.69	32.45	2.39	55.66	0.23	2.13
	Dual-MobileNetV1	3.27	40.76	4.32	66.75	0.48	8.67
	Dual-MobileNetV2	4.23	58.09	6.40	68.78	0.79	23.63
	Dual-MobileNetV3	5.37	55.68	4.67	68.33	1.02	11.20
	TE-Branch + SP-Branch	0.14	0.84	0.72	13.35	0.12	0.57

To further analyze the differences between homogeneous and heterogeneous architectures in the feature space, 128-dimensional fused features of different models are extracted on the Tongji test set, and Gram matrices are constructed based on cosine similarity for visualization, as shown in Figure 7. Samples indexed from 1 to 6 correspond to the same category across all sub-figures. Ideally, the feature space should exhibit high intra-class similarity and low inter-class similarity. As shown in Figure 7a,b, the homogeneous models still exhibit relatively high inter-class similarity, indicating feature overlap between different categories. For Dual-ResNet18 and Dual-MobileNetV3, the intra-class and inter-class similarities are approximately 0.99 and 0.94, respectively, with a small margin between them. In contrast, Figure 7c shows that the proposed heterogeneous dual-stream network achieves lower inter-class similarity (approximately 0.08) while maintaining higher intra-class consistency (approximately 0.88), resulting in a more discriminative feature distribution. According to the similarity gap between intra-class and inter-class regions, the heterogeneous model achieves a value of approximately 0.80, whereas the homogeneous models achieve only about 0.05, demonstrating the feature discriminability of the proposed method.

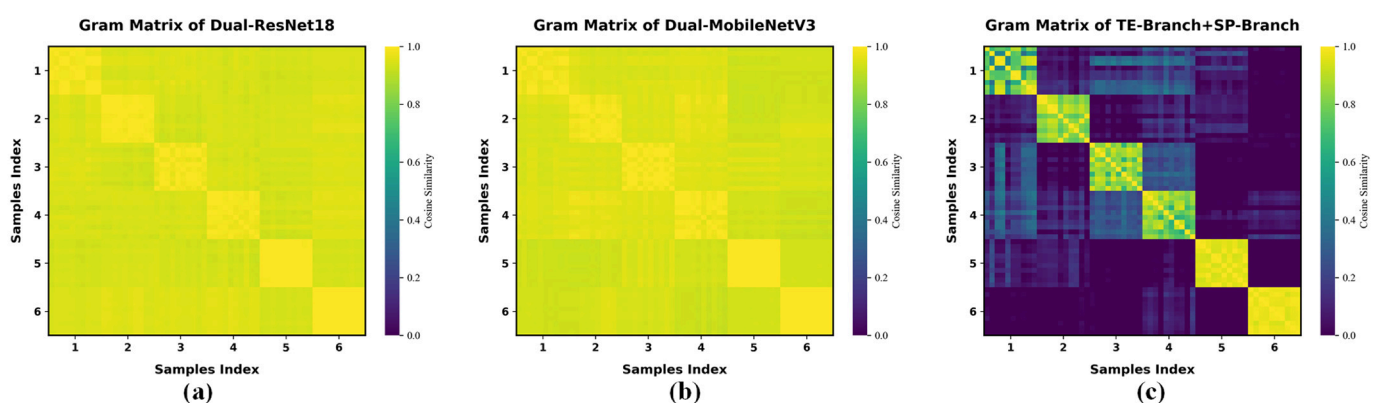


Figure 7. Visualization of cosine similarity-based Gram matrices of fused features on the Tongji dataset under different models. (a) Dual-ResNet18. (b) Dual-MobileNetV3. (c) The proposed heterogeneous dual-stream network (TE-Branch + SP-Branch).

To further quantitatively analyze the similarity distributions of homogeneous and heterogeneous architectures in the feature space, we computed the intra-class similarity, inter-class similarity, and their corresponding embedding-space separation for each model based on the results presented in Figure 7, as shown in Table 10.

Table 10. Quantitative analysis of cosine similarity distributions in Figure 7.

Model	Intra-Class Similarity	Inter-Class Similarity	Embedding-Space Separation
Dual-ResNet18	0.991	0.944	0.047
Dual-MobileNetV3	0.990	0.941	0.049
TE-Branch + SP-Branch	0.876	0.080	0.796

3.3.2. Comparison of Cross-Modal Fusion Strategies

Under a fixed heterogeneous backbone setting, to analyze the impact of different fusion strategies on recognition performance, comparative experiments are conducted while keeping other training settings consistent. The proposed CPA module is compared with Data-Fusion, Feature-Fusion, and Score-Fusion methods.

Specifically, Data-Fusion performs channel-wise stacking at the input level, Score-Fusion aggregates single-modality scores based on the Min-Rule, and Feature-Fusion achieves feature interaction through channel concatenation.

As shown in Table 11, the CPA module maintains low EER across the three datasets. Furthermore, when combined with supervised contrastive learning, HDC-Net exhibits good recognition performance across all evaluation metrics.

Table 11. Performance comparison of different cross-modal fusion strategies under a heterogeneous backbone setting (%).

Fusion Strategy	Tongji		CASIA-MS		SCAU-PM	
	EER	FRR@FAR = 10^{-4}	EER	FRR@FAR = 10^{-4}	EER	FRR@FAR = 10^{-4}
Data-Fusion	0.12	0.84	0.46	8.82	0.14	1.55
Score-Fusion	0.14	0.90	0.42	6.56	0.12	0.49
Feature-Fusion	0.14	0.84	0.72	13.35	0.12	0.57
CPA	0.11	2.14	0.38	9.28	0.09	0.57
HDC-Net	0.05	0.52	0.21	5.88	0.02	0.57

3.3.3. Performance Comparison and Computational Complexity Analysis

To evaluate the performance of HDC-Net, three palmprint and palm vein fusion methods from the past three years are selected for comparison, including MIES-Net [16], DH-Fuzzy [19], and CohashNet [20].

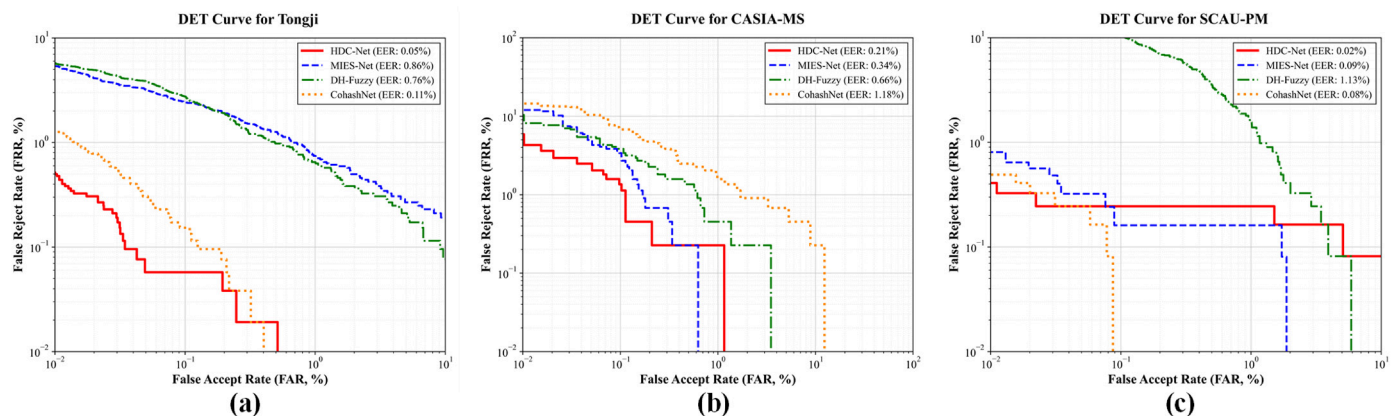
Considering the differences in experimental settings and implementation details across different studies, all comparison methods are re-implemented and evaluated under a unified protocol in this work. It should be noted that the original studies adopted different evaluation settings and dataset partition strategies, and discrepancies may therefore exist between the originally reported results and the re-implemented results. The re-implemented results are primarily used for relative performance comparison under the same evaluation protocol.

As shown in Table 12, HDC-Net achieves consistent recognition performance across all three datasets. On the Tongji, CASIA-MS, and SCAU-PM datasets, the proposed method achieves EERs of 0.05%, 0.21%, and 0.02%, respectively. Meanwhile, HDC-Net maintains stable performance under the condition of FAR = 10^{-4} .

As shown in Figure 8, the DET curves of HDC-Net across different datasets are lower than or comparable to those of the compared methods, indicating its good discriminative ability under various threshold conditions. On the SCAU-PM dataset, the curve shows some fluctuations in the low-error region. However, HDC-Net still achieves the lowest EER, demonstrating the effectiveness of the proposed method.

Table 12. Performance comparison between HDC-Net and comparison methods (%).

Method	Tongji		CASIA-MS		SCAU-PM	
	EER	FRR@FAR = 10^{-4}	EER	FRR@FAR = 10^{-4}	EER	FRR@FAR = 10^{-4}
MIES-Net [16]	0.86	5.39	0.34	16.06	0.09	0.89
DH-Fuzzy [19]	0.76	5.72	0.66	10.41	1.13	17.17
CohashNet [20]	0.11	1.30	1.18	15.61	0.08	0.57
HDC-Net	0.05	0.52	0.21	5.88	0.02	0.57

**Figure 8.** Comparison of DET curves between HDC-Net and the compared methods. (a) DET Curve for Tongji. (b) DET Curve for CASIA-MS. (c) DET Curve for SCAU-PM.

In contactless biometric applications, computational complexity is as important as recognition accuracy. In this study, the computational complexity of different models is evaluated under a unified setting following the ISO/IEC 19795-1:2021 standard, with a fixed input size of 128×128 single-channel images.

As shown in Table 13, HDC-Net contains 0.98 M parameters, a model size of 3.74 MB, and 0.1052 G FLOPs. In comparison, the baseline methods Dual-ResNet18, MIES-Net, DH-Fuzzy, and CohashNet exhibit higher computational complexity. Although Dual-MobileNetV3 achieves lower FLOPs (0.0395 G), its cross-modal recognition capability is limited. Furthermore, SP-Branch Only (0.0039 G) and TE-Branch Only (0.0490 G) demonstrate that the heterogeneous dual-stream backbone retains effective feature representation capability under low computational complexity. Overall, HDC-Net achieves a good balance between recognition performance and computational efficiency.

Table 13. Comparison of Computational Complexity of Different Models.

Method	Params (M)	Model Size (MB)	FLOPs (G)
Dual-ResNet18	22.50	85.92	1.1397
Dual-MobileNetV3	3.33	12.80	0.0395
MIES-Net	56.47	215.51	2.9157
DH-Fuzzy	15.02	57.96	1.8924
CohashNet	9.39	45.93	0.2725
SP-Branch Only	0.27	1.02	0.0039
TE-Branch Only	0.67	2.59	0.0490
HDC-Net	0.98	3.74	0.1052

4. Discussion

This study conducts fusion recognition experiments based on HDC-Net, which are built upon two hypotheses: first, modality-specific feature extraction contributes to improv-

ing feature representation capability; second, introducing supervised contrastive learning can optimize the structure of the feature space. The experimental results (as shown in Tables 3 and 5) are consistent with these hypotheses.

Existing multimodal fusion methods typically adopt symmetric backbone networks for feature extraction, where modality-specific characteristics may not be sufficiently exploited. Previous studies have demonstrated that modality-specific feature extraction can improve feature representation capability [22]. Most existing methods rely on optimization objectives such as ArcFace [23] or Softmax loss, whereas supervised contrastive learning further enhances the structural optimization of the feature space.

From the perspective of feature representation, the heterogeneous dual-stream network effectively improves modality-specific representation capability (as shown in Table 5). From the perspective of feature space distribution, after introducing supervised contrastive learning, intra-class samples become more compact and inter-class samples become more separable. As shown in Figure 6, the distributions of genuine and impostor samples exhibit clearer and more stable separation in the feature space; meanwhile, the similarity of hard genuine pairs increases while that of hard impostor pairs decreases, resulting in an enlarged discriminative margin. These observations are consistent with the reduction in EER and the stable performance under low FAR settings. Furthermore, the CPA module further enables cross-modal feature interaction and fusion through channel and spatial attention mechanisms. It should be further noted that an interesting phenomenon is observed on the Tongji dataset after introducing the CPA module. Although the EER is reduced, the $FRR@FAR = 10^{-4}$ increases to some extent. This phenomenon is mainly observed on the Tongji dataset and is not consistently reflected on the CASIA-MS and SCAU-PM datasets. A possible explanation is that CPA enhances cross-modal feature interaction and improves the discriminative capability of the fused representation. However, without additional feature-space constraints, certain genuine pairs may still exhibit relatively large intra-class variations. Under the strict operating point of $FAR = 10^{-4}$, these samples are more likely to be rejected. After introducing supervised contrastive learning, feature space separation is further improved, thereby alleviating this phenomenon.

In contactless biometric recognition, system robustness and computational complexity are both important considerations. While improving recognition performance, HDC-Net maintains relatively low computational complexity (0.1052 GFLOPs) and model size (3.74 MB), achieving a good balance between performance and computational efficiency. Nevertheless, the model still has certain limitations, including sensitivity to batch size, limited stability under varying conditions due to a fixed input resolution, and the need for further validation in cross-device and cross-domain scenarios. In future work, model compression techniques can be explored to further reduce computational complexity, and cross-domain generalization methods can be introduced to improve the model's adaptability under different data distributions.

5. Conclusions

We propose a contactless palmprint and palm vein fusion recognition model, HDC-Net, based on a heterogeneous dual-stream network and supervised contrastive learning. The proposed method improves cross-modal feature representation and discriminative ability by constructing a heterogeneous dual-stream feature extraction architecture and introducing supervised contrastive learning. In addition, the CPA module is incorporated to enhance fusion performance. Experimental results demonstrate that the proposed method achieves stable and good performance across multiple datasets, validating its effectiveness and robustness and further indicating its applicability and potential for contactless multimodal biometric fusion recognition.

Author Contributions: Conceptualization, F.D. and Z.Z.; methodology, Z.Z. and F.D.; software, Z.Z. and Y.L.; validation, Z.Z., Y.L. and F.D.; formal analysis, Z.Z. and F.D.; investigation, Z.Z. and Y.L.; resources, F.D. and X.Z.; data curation, Z.Z. and Z.X.; writing—original draft preparation, Z.Z.; writing—review and editing, F.D. and X.Z.; visualization, Z.Z. and F.N.; supervision, F.D.; project administration, F.D.; funding acquisition, F.D. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the National Natural Science Foundation of China (grant number 32371984).

Data Availability Statement: The Tongji palmprint image database used in this study is publicly available at: <https://cslinzhang.github.io/ContactlessPalm/> (accessed on 12 May 2025). The CASIA-MS-PalmprintV1 dataset used in this study is publicly available at: <http://biometrics.idealtest.org/> (accessed on 12 May 2025). The data presented in this study are available on request from the corresponding author due to restrictions on data sharing.

Acknowledgments: Portions of the research in this paper use the CASIA-MS-PalmprintV1 collected by the Chinese Academy of Sciences' Institute of Automation (CASIA). During the preparation of this manuscript, the authors used ChatGPT to improve language. The authors have reviewed and edited the output and take full responsibility for the content of this publication.

Conflicts of Interest: The authors declare no conflicts of interest.

Abbreviations

The following abbreviations are used in this manuscript:

TE-Branch	Texture Enhancement Branch
SP-Branch	Structure Perception Branch
CPA	Cross-Palm Attention
SupCon	Supervised Contrastive Learning
FAR	False Accept Rate
FRR	False Reject Rate
EER	Equal Error Rate
DET	Detection Error Trade-off
FLOPs	floating-point operations

References

1. Zhang, D.; Kong, W.K.; You, J.; Wong, M. Online palmprint identification. *IEEE Trans. Pattern Anal. Mach. Intell.* **2003**, *25*, 1041–1050. [\[CrossRef\]](#)
2. Zhou, Y.B.; Kumar, A. Human Identification Using Palm-Vein Images. *IEEE Trans. Inf. Forensics Secur.* **2011**, *6*, 1259–1274. [\[CrossRef\]](#)
3. Wu, W.; Zhang, Y.; Li, Y.; Li, C. Fusion recognition of palmprint and palm vein based on modal correlation. *Math. Biosci. Eng.* **2024**, *21*, 3129–3145. [\[CrossRef\]](#) [\[PubMed\]](#)
4. Li, S.Y.; Fei, L.K.; Zhang, B.; Ning, X.; Wu, L.F. Hand-based multimodal biometric fusion: A review. *Inf. Fusion* **2024**, *109*, 20. [\[CrossRef\]](#)
5. Zhong, D.; Yang, Y.; Du, X. Palmprint Recognition Using Siamese Network. In *Biometric Recognition. CCBR 2018*; Springer: Cham, Switzerland, 2018; pp. 48–55.
6. Xu, N.Y.; Zhu, Q.; Xu, X.Y.; Zhang, D.Q. An effective recognition approach for contactless palmprint. *Vis. Comput.* **2021**, *37*, 695–705. [\[CrossRef\]](#)
7. Liang, X.; Yang, J.Y.; Lu, G.M.; Zhang, D. CompNet: Competitive Neural Network for Palmprint Recognition Using Learnable Gabor Kernels. *IEEE Signal Process. Lett.* **2021**, *28*, 1739–1743. [\[CrossRef\]](#)
8. Mirmohamadsadeghi, L.; Drygajlo, A. Palm vein recognition with local texture patterns. *IET Biom.* **2014**, *3*, 198–206. [\[CrossRef\]](#)
9. Fanjiang, Y.Y.; Lee, C.C.; Du, Y.; Horng, S. Palm Vein Recognition Based on Convolutional Neural Network. *Informatica* **2021**, *32*, 687–708. [\[CrossRef\]](#)
10. Horng, S.-J.; Vu, D.-T.; Nguyen, T.-V.; Zhou, W.; Lin, C.-T. Recognizing palm vein in smartphones using RGB images. *IEEE Trans. Ind. Inform.* **2021**, *18*, 5992–6002. [\[CrossRef\]](#)
11. Kittler, J.; Hatef, M. On combining classifiers. *IEEE Trans. Pattern Anal. Mach. Intell.* **1998**, *20*, 226–239. [\[CrossRef\]](#)

12. Zhang, D.; Guo, Z.; Lu, G.; Zhang, L.; Liu, Y.; Zuo, W. Online joint palmprint and palmvein verification. *Expert Syst. Appl.* **2011**, *38*, 2621–2631. [[CrossRef](#)]
13. Zhang, L.; Cheng, Z.; Shen, Y.; Wang, D. Palmprint and Palmvein Recognition Based on DCNN and A New Large-Scale Contactless Palmvein Dataset. *Symmetry* **2018**, *10*, 78. [[CrossRef](#)]
14. Zhang, K.; Wang, H.B.; Yu, C.; Du, M.; Tao, L. Class Constraint-based Discriminative Features Learning Algorithm for Palm Print and Palm Vein Fusion Recognition. In *7th International Conference on Signal and Image Processing (ICSIP)*; IEEE: New York, NY, USA, 2022; pp. 275–280.
15. Fan, D.D.; Liang, X.; Jia, W.; Chen, J.A.; Zhang, D.V. A Novel Hybrid Fusion Combining Palmprint and Palm Vein for Large-Scale Palm-Based Recognition. *IEEE Trans. Syst. Man Cybern. Syst.* **2024**, *54*, 4471–4484. [[CrossRef](#)]
16. Pan, Z.; Xu, J.; Wang, J.; Jia, W. Palmprint and palm vein recognition method based on modal information evaluation strategy. *CAAI Trans. Intell. Syst.* **2024**, *19*, 1136–1148. (In Chinese) [[CrossRef](#)]
17. Wang, C.H.; Chen, W.R.; Yen, J.J.; Yang, X.S.; Siang, Y.S. Enhanced biometric authentication through integrated palm print and palm vein images. *Vis. Comput.* **2025**, *42*, 13. [[CrossRef](#)]
18. Yu, Y.; Jin, P.; Wang, J.; Pan, Z. Rethinking Early-Fusion Strategy for Palmprint and Palm Vein Fusion Recognition. In *IEEE International Joint Conference on Biometrics (IJCB)*; IEEE: New York, NY, USA, 2025; pp. 1–8.
19. Wu, T.; Leng, L.; Khan, M.K. A multi-spectral palmprint fuzzy commitment based on deep hashing code with discriminative bit selection. *Artif. Intell. Rev.* **2023**, *56*, 6169–6186. [[CrossRef](#)]
20. Dong, X.; Khan, M.K.; Leng, L.; Teoh, A.B.J. Co-Learning to Hash Palm Biometrics for Flexible IoT Deployment. *IEEE Internet Things J.* **2022**, *9*, 23786–23794. [[CrossRef](#)]
21. Lee, J.; Oh, B.S.; Toh, K.A. Extraction of intersecting palm-vein and palmprint features for cancellable identity verification. *CAAI Trans. Intell. Technol.* **2024**, *9*, 69–86. [[CrossRef](#)]
22. Lu, Y.; Wu, Y.; Liu, B.; Zhang, T.; Li, B.; Chu, Q.; Yu, N. Cross-modality Person re-identification with Shared-Specific Feature Transfer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, Seattle, WA, USA, 14–19 June 2020; pp. 13379–13389.
23. Deng, J.; Guo, J.; Yang, J.; Xue, N.; Kotsia, I.; Zafeiriou, S. ArcFace: Additive Angular Margin Loss for Deep Face Recognition. *IEEE Trans. Pattern Anal. Mach. Intell.* **2022**, *44*, 5962–5979. [[CrossRef](#)] [[PubMed](#)]
24. Pan, Z.Y.; Xu, J.M.; Jiang, S.T.; Wang, J. SSFD-Net: Shared-Specific Feature Disentanglement Network for Multimodal Biometric Recognition with Missing Modality. *Digit. Signal Prog.* **2025**, *159*, 11. [[CrossRef](#)]
25. Hadsell, R.; Chopra, S.; Lecun, Y. Dimensionality Reduction by Learning an Invariant Mapping. In *IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR)*; IEEE: New York, NY, USA, 2006; pp. 1735–1742.
26. Khosla, P.; Teterwak, P.; Wang, C.; Sarna, A.; Tian, Y.; Isola, P.; Maschinot, A.; Liu, C.; Krishnan, D. Supervised Contrastive Learning. In *Advances in Neural Information Processing Systems (NeurIPS)* 33; Neural Information Processing Systems Foundation, Inc.: San Diego, CA, USA, 2020; pp. 18661–18673.
27. Howard, A.; Sandler, M.; Chu, G.; Chen, L.C.; Chen, B.; Tan, M.; Wang, W.; Zhu, Y.; Pang, R.; Vasudevan, V.; et al. Searching for MobileNetV3. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, Seoul, Republic of Korea, 27 October–2 November 2019; pp. 1314–1324.
28. Chen, S.; Liu, Y.; Gao, X.; Han, Z. MobileFaceNets: Efficient CNNs for Accurate Real-time Face Verification on Mobile Devices. In *Biometric Recognition. CCBR 2018*; Springer: Cham, Switzerland, 2018; pp. 428–438.
29. Hu, J.; Shen, L.; Albanie, S.; Sun, G.; Wu, E. Squeeze-and-Excitation Networks. *IEEE Trans. Pattern Anal. Mach. Intell.* **2020**, *42*, 2011–2023. [[CrossRef](#)] [[PubMed](#)]
30. Hou, Q.; Zhou, D.; Feng, J. Coordinate Attention for Efficient Mobile Network Design. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, Virtual, 21–24 June 2021; pp. 13713–13722.
31. Woo, S.; Park, J.; Lee, J.Y.; Kweon, I.S. CBAM: Convolutional Block Attention Module. In *Proceedings of the European Conference on Computer Vision (ECCV)*, Munich, Germany, 8–14 September 2018; pp. 3–19.
32. Zhang, L.; Li, L.; Yang, A.; Shen, Y.; Yang, M. Towards contactless palmprint recognition: A novel device, a new benchmark, and a collaborative representation based identification approach. *Pattern Recognit.* **2017**, *69*, 199–212. [[CrossRef](#)]
33. CASIA-MS-PalmprintV1. Available online: <http://biometrics.idealtest.org/> (accessed on 12 May 2025).
34. Hao, Y.; Sun, Z.; Tan, T.; Ren, C. Multispectral palm image fusion for accurate contact-free palmprint recognition. In *15th IEEE International Conference on Image Processing*; IEEE: New York, NY, USA, 2008; pp. 281–284.
35. ISO/IEC 19795-1:2021; Information Technology—Biometric Performance Testing and Reporting—Part 1: Principles and Framework. ISO: Geneva, Switzerland; IEC: Geneva, Switzerland, 2021.

Disclaimer/Publisher’s Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.