A Quality-Guided Mixture of Score-Fusion Experts Framework for Human Recognition

Jie Zhu, Yiyang Su, Minchul Kim, Anil Jain, and Xiaoming Liu
Department of Computer Science and Engineering,
Michigan State University, East Lansing, MI 48824
{zhujie4, suyiyan1, kimminc2, jain, liuxm}@msu.edu

Abstract

*Whole-body biometric recognition is a challenging multimodal task that integrates various biometric modalities, including face, gait, and body. This integration is essential for overcoming the limitations of unimodal systems. Traditionally, whole-body recognition involves deploying different models to process multiple modalities, achieving the final outcome by score-fusion (e.g., weighted averaging of similarity matrices from each model). However, these conventional methods may overlook the variations in score distributions of individual modalities, making it challenging to improve final performance. In this work, we present **Q**uality-guided **M**ixture of score-fusion **E**xperts (QME), a novel framework designed for improving whole-body biometric recognition performance through a learnable score-fusion strategy using a Mixture of Experts (MoE). We introduce a novel pseudo-quality loss for quality estimation with a modality-specific Quality Estimator (QE), and a score triplet loss to improve the metric performance. Extensive experiments on multiple whole-body biometric datasets demonstrate the effectiveness of our proposed approach, achieving state-of-the-art results across various metrics compared to baseline methods. Our method is effective for multimodal and multi-model, addressing key challenges such as model misalignment in the similarity score domain and variability in data quality. Code is available at the Project Link.*

1. Introduction

Whole-body biometrics integrates diverse recognition tasks such as Face Recognition (FR) [10, 24], Gait Recognition (GR) [63, 66], and Person Re-identification (ReID) [15, 35] to overcome unimodal limitations. Whole-body biometrics benefits from the combined strengths of multiple modalities. This multimodal synergy ensures robust performance in non-ideal conditions (low-light, occlusion, and missing traits), making it indispensable for security-critical domains like surveillance and law enforcement.

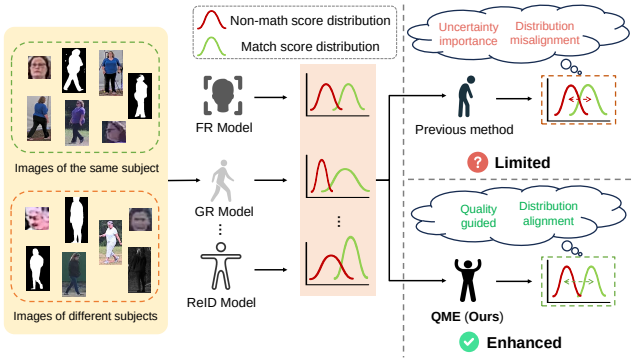

Figure 1. Illustration of score distribution alignment in multimodal human recognition. Different models and modalities (*e.g.*, face, gait, and body) produce distinct similarity score distributions. Conventional score-fusion methods struggle with optimal alignment and assigning importance weights to each modality, potentially degrading performance.

Effective fusion is pivotal to whole-body recognition. Current approaches include decision-level fusion, feature-level fusion, and score-level fusion [51]. In decision-level fusion, each modality first makes an identity decision based on its extracted features. The individual decisions are then combined based on either decision scores or ranks. Feature-level fusion combines extracted features from different modalities to obtain a single representation [5, 27]. However, this approach is often hindered by inconsistencies across modalities in biometrics, as different traits may not necessarily complement each other effectively. Most importantly, feature-level fusion requires suitable paired multimodal datasets. Many available datasets such as WebFace42M [67] for face recognition do not contain whole-body data, while other datasets like PRCC [61], LTCC [43], and CCPG [32] widely used in person ReID and gait recognition, are limited by dataset size, the masking of faces, or

insufficient number of subjects for generalizable training.

Compared to feature-level fusion, score-level fusion integrates the similarity scores or feature (embedding) distances generated by individual models. Score-level fusion offers computational efficiency and modular flexibility compared to feature-level fusion, enabling seamless integration of heterogeneous modalities while preserving individual models' performance. However, conventional score-fusion techniques are limited by their inability to fully utilize the different distributions of match (genuine) and non-match (impostor) scores produced by each model, as shown in Fig. 1. Additionally, finding the optimal weight for each model in the fusion process is challenging, even using grid search [33], leading to suboptimal performance.

To address these challenges, we propose a Quality Estimator (QE) and pseudo-quality loss that leverages pretrained models to generate pseudo-quality labels, eliminating laborious manual annotation. We develop a Mixture of Score-Fusion Experts method, where each expert learns a distinct fusion strategy (*e.g.*, one prioritizes face-gait synergy, and another handles occlusion scenarios). Experts' contributions are dynamically weighted by QE predictions, ensuring robustness to sensor noise and missing modalities. To improve metric learning performance, we present a score triplet loss that enforces margin separation between match/non-match scores while suppressing non-match magnitudes, directly aligning with metrics like 1:1 verification and 1:N open-set search. This approach improves score-level alignment between modalities without the need for retraining biometric backbones nor requiring tremendous training data. Our main contributions are:

- We propose a Quality Estimator (QE) that employs pseudo quality loss—derived from pretrained models and ranking performance—to assess biometric modality quality without the need for human-labeled data.
- We introduce **QME**, a multimodal biometric recognition framework that integrates a learnable, modality-specific score-fusion method. QME dynamically combines diverse fusion strategies, adapting to sensor noise, occlusions, and missing modalities.
- We introduce a novel score triplet loss for metric learning by enforcing a match/non-match score margin, directly improving key metrics like verification accuracy and open-set search effectiveness.
- Experiments on multiple whole-body biometric datasets validate our approach's superior robustness over leading score-fusion methods and models.

2. Related Work

2.1. Score-fusion

Score-level fusion integrates similarity scores from multiple modalities to optimize recognition decisions [51]. Traditional score-fusion methods include Z-score and min-max normalization. [19, 38, 41, 42, 58] introduce likelihood ratio based score fusion. Ross *et al.* propose mean, max, or min score-fusion, where the final score is determined by averaging, the highest, or the lowest score [23, 45, 64]. Recent literature categorizes score fusion into two paradigms: fixed-rule methods, employing predefined heuristics (*e.g.*, predefined weights), and trained-rule methods, utilizing learned parameters optimized through training (*e.g.*, SVM) [6, 40, 55]. Score-fusion methods offer several advantages: 1) they are robust to missing modality inputs, and 2) they simplify alignment, as the domain gap between modalities is smaller than feature-space alignment. However, challenges remain in determining the optimal alignment and weighting for each model and identifying the most effective fusion strategy. We aim to explore a better way of assessing the contribution of each modality and develop a more generalizable score-fusion method.

2.2. Biometric Quality Assessment

Unlike generic image quality assessment [46], biometric quality assessment is the process of evaluating the quality of biometric data (*e.g.*, facial images), which directly influences recognition performance [13, 39, 57]. This assessment typically follows initial authentication to filter out spoofed or synthetic samples [16, 17, 65]. While some studies target fingerprints and irises [3, 11, 28], others apply learning-based methods for facial image quality [2, 4, 21, 24, 25, 37, 50, 56]. However, many such methods rely on specialized training procedures incompatible with pretrained models. In this work, we introduce a method to train a general QE by distilling knowledge from the pretrained model, providing a versatile approach to biometric quality assessment.

2.3. Whole-Body Biometric Recognition

As illustrated in Fig. 2, whole-body biometric systems integrate detectors, encoders, and fusion modules to unify multi-modal traits (*e.g.*, face, gait) for robust identification [9]. Key to the design is effectively leveraging complementary strengths while mitigating individual weaknesses: facial recognition excels with high-resolution frontal images but degrades under non-ideal conditions (*e.g.*, large standoff, off-angle views), while gait and ReID models contend with clothing/posture variations [34, 36]. Recent advances [7, 18, 22, 44, 53, 60] highlight multi-attribute fusion but largely overlook the heterogeneity inherent in whole-body modalities, focusing mainly on homogeneous sensor data. Efforts to incorporate facial features into ReID [14, 27, 30, 31, 34] often prioritize modular additions over optimizing fusion efficacy. Fusion methods for comprehensive whole-body biometric recognition remain challenging, and require in-depth exploration.

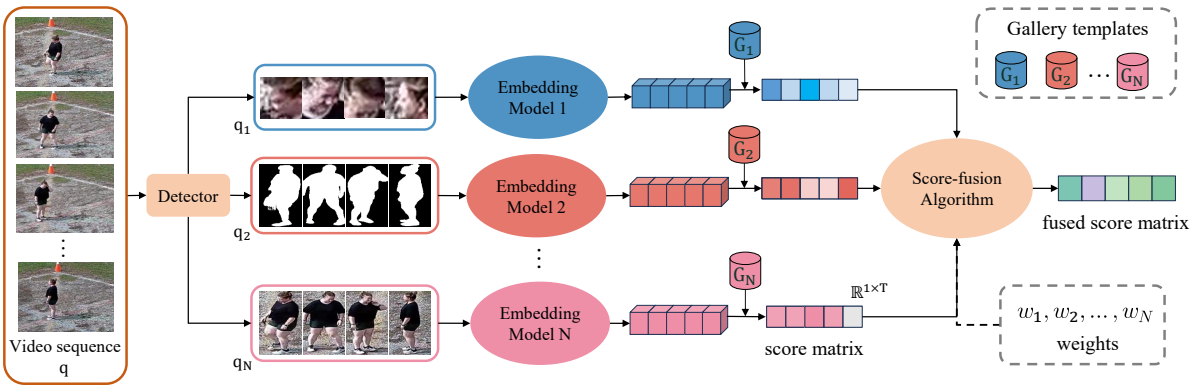

Figure 2. General framework for whole-body biometric recognition. An input video sequence q is processed by a detector to extract different modality queries, which are fed into multiple embedding models. Each model generates similarity scores by comparing the extracted features with T gallery templates. Our work focuses on score-fusion algorithms that produce the final decision based on input score matrices and modality weights.

3. Methodology

In this section, we introduce the proposed **QME** method, which leverages quality assessment and learnable score-fusion with MoE across multiple modalities. Our approach is specifically designed to tackle challenges related to model misalignment in score-level distributions and varying data quality in whole-body biometric recognition.

Overview. In biometric evaluation, a query (or probe) refers to a sample sequence needing identification/verification against a gallery of enrolled subjects in the system. Each gallery subject may have multiple videos/images to extract gallery templates. Given a model M_n in the embedding model set $\{M_1, M_2, \ldots, M_N\}$ with a query and gallery templates where N is the number of models, we compute the query feature $q_n \in \mathbb{R}^{L\times d_n}$ and gallery template features $G_n \in \mathbb{R}^{T\times d_n}$, where L represents the sequence length of the query (number of images) and T is the total number of gallery templates (videos/images) across all gallery subjects, and d_n is the feature dimension of M_n. We further compute the average of q_n to obtain the query-level feature vector in $\mathbb{R}^{d_n}$, and then compute its similarity with G_n to get the query score matrix $\mathbf{S}_n \in \mathbb{R}^{1\times T}$, representing the similarity score of the query with each gallery template. Our training process involves two stages: (1) training QE, and (2) freezing QE while training the learnable score-fusion model.

3.1. Quality Estimator (QE)

The goal of the QE is to predict the input quality of a given modality. We hypothesize that if the input quality for a particular modality is poor, the system should shift focus to other modalities to enhance overall performance. As illustrated in Fig. 3(a), to train a QE for M_n, we collect the intermediate features $\mathcal{I}_n \in \mathbb{R}^{L\times U\times P_n\times d_n}$ from M_n, where U is the number of blocks, P_n is the patch size of M_n. $\mathcal{I}_n$ captures various levels of semantic information from the model. We follow [25] to extract intermediate features from the backbone and compute the mean and the standard deviation, reducing $\mathcal{I}_n$ to a representation in $\mathbb{R}^{L\times 2d_n}$. This representation is then fed into an encoder to predict query-level quality weight $w_n \in \mathbb{R}$ produced by sigmoid function.

Pseudo Quality Loss. The challenge of training QE is the lack of human-labeled qualities. Empirically, we do not have the quality label of the query images. However, we can know the ranking result by sorting the similarities between the query feature and training gallery features. A higher ranking result indicates the input images are close to their gallery center. We assume that if the ranking result of the input is better, the quality of the input will be higher. Hence, we propose a pseudo quality loss $\mathcal{L}_{rank}$ using the ranking result of the input for the pretrained model M_n:

$$\mathcal{L}_{rank} = \sum_{i\in L} \text{MSELoss}\left(w_i, \text{ReLU}(\frac{\delta - r_i}{\delta - 1})\right). \quad (1)$$

Here r_i is the ranking result of the query feature q_i, w_i is the predicted quality weight, and δ is a hyperparameter to adjust the sensitivity of the ranking threshold. To obtain r_i, we compute the similarity matrix between q_i and G_n. Lower δ will push the predicted r_i to 0 if the ranking result is out of δ. Conversely, higher δ will cause the QE to predict a value closer to 1 as it has a higher tolerance for the ranking result. Our proposed QE offers several benefits: (1) It can generalize across all pretrained models (not only FR models) by learning from these models and identifying characteristics of challenging samples, and (2) it can be trained on any dataset, whether in-domain or out-of-domain. While pretrained models may exhibit biases toward their training data, which can hinder generalization, challenging samples may originate from either in-domain or out-of-domain data.

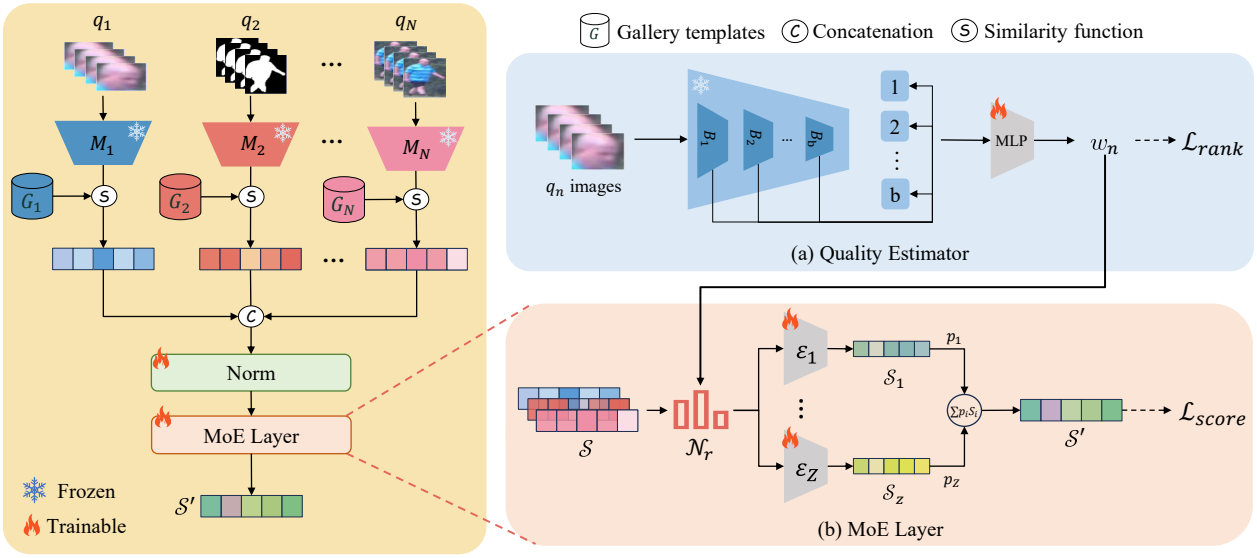

Figure 3. The architecture of the proposed QME framework. It includes a $Norm$ layer and an MoE layer to process concatenated score matrix $\mathcal{S}$ from the model set $M_1, M_2, \ldots, M_N$. The MoE layer contains experts $\varepsilon_1, \varepsilon_2, \ldots, \varepsilon_Z$ to individually encode the fused score matrices. A quality estimator (QE) uses the intermediate feature $\mathcal{I}_n$ from the backbone block $B_1, B_2, \ldots, B_b$ to generate weights w_n, which control $p_1, p_2, \ldots, p_Z$ for a weighted sum, producing the final fused score matrix $\mathcal{S}'$.

3.2. Mixture of Score-fusion Experts

The concept of MoE [12, 48] comes from the NLP community, where they use MoE layers to replace feed-forward network (FFN) layers in the transformer blocks. With the sparsity of experts and the router network, each expert can focus on handling different tokens. In addition, some special loss functions are designed to control the behavior of the router [8, 29, 48, 49, 54, 68].

Inspired by this, we design a MoE layer (shown in Fig. 3(b)) with multiple score-fusion experts, controlled by $\mathcal{N}_r$ that learns to perform score-fusion based on quality weights. Unlike in traditional MoE setups, where a router network predicts assignment probabilities from inputs, the similarity score in our case is a high-level semantic feature, lacks fine-grained cues about query quality. Instead, we use the proposed QE to predict the quality weight of the query to imply the reliability of the input modality, guiding the selection process. For an expert ϵ_z from expert set $\{\epsilon_1, ..., \epsilon_Z\}$ where Z is the number of experts, it receives a concatenated score matrix $\mathcal{S} \in \mathbb{R}^{T \times N}$ from all modalities and predict a fused score matrix $\mathcal{S}_z \in \mathbb{R}^{1 \times T}$. Given w_n as the modality-specific quality weight and ε_n controlled by $p_n = w_n$, we aim for expert ε_n to prioritize the selected modality when w_n is high. Conversely, when w_n is low, other experts contribute more to the final score matrix and shift the focus to other modalities. This approach ensures that higher-quality modalities have a greater influence on the output, while lower-quality ones contribute less, optimizing overall performance.

3.3. Quality-Guided Mixture of Score-fusion Experts (QME)

Based on Sec. 3.1 and 3.2, we further introduce QME. As illustrated in Fig. 3 (left), for a query feature set $\mathbf{Q} = \{q_1, q_2, \ldots, q_N\}$ processed by the model set $\{M_1, M_2, \ldots, M_N\}$, we generate the concatenated input score matrix $\mathcal{S} = \{\mathbf{S}_1, \mathbf{S}_2, \ldots, \mathbf{S}_N\} \in \mathbb{R}^{T \times N}$. For models that use Euclidean distance as a metric, we convert distances into similarity scores:

$$\frac{1}{1 + Euc(q, g)}, \tag{2}$$

where $Euc(q, g)$ represents Euclidean distance between the query feature q and the gallery feature g. This transformation remaps Euclidean distances to align with the range of Cosine Similarity, where larger values indicate higher similarity. We then normalize $\mathcal{S}$ using a $BatchNorm$ layer. After normalization, $\mathcal{S}$ is fed into the MoE layer, which contains a router network $\mathcal{N}_r$ and multiple score-fusion experts $\{\varepsilon_1, \varepsilon_2, \ldots, \varepsilon_Z\}$. Each expert is specialized to handle specific input conditions (*i.e.*, similarity values), with the router selecting the most suitable expert based on quality assessment. $\mathcal{N}_r$ takes w_n as the input and generates the weight of assigning input to all experts $\{p_1, p_2, \ldots, p_Z\}$ where p_Z is the weight of contribution of expert ε_Z. The final fused score matrix $\mathcal{S}'$ is computed as a weighted sum of the outputs from all experts:

$$\mathcal{S}' = \sum_{z \in Z} p_z \mathcal{S}_z, \tag{3}$$

where $\mathcal{S}_z$ is the output score matrix from ε_z. By using quality weight to modulate $\mathcal{S}'$, each expert learns how the contributions of different modalities' scores to $\mathcal{S}'$ should be adjusted in response to changes in their quality levels.

Score Triplet Loss. The triplet loss [47] optimizes relative distances between samples:

$$\mathcal{L}_{tri} = \text{ReLU}(d(a,p) - d(a,n) + m), \tag{4}$$

where $d(a,p)$ is the distance between anchor a and positive sample p, $d(a,n)$ is the distance between anchor a and negative sample n, and m enforces a margin. The triplet loss focuses on maintaining a boundary between positive and negative pairs, but it does not effectively constrain the value of non-match scores. The verification and open-set search rely on a threshold τ. For example, TAR@τ%FAR measures the acceptance rate of the match samples such that only $\tau\%$ of non-match scores can be accepted as matches. To optimize these metrics, we introduce the score triplet loss:

$$\mathcal{L}_{score} = \text{ReLU}(\mathcal{S}'_{nm}) + \text{ReLU}(m - \mathcal{S}'_{mat}), \tag{5}$$

where $\mathcal{S}'_{nm}$ is the non-match scores of $\mathcal{S}'$, $\mathcal{S}'_{mat}$ is the match score of $\mathcal{S}'$. Unlike the original triplet loss, this formulation provides more constraints:

- Directly suppresses non-match scores ($\text{ReLU}(\mathcal{S}'_{nm})$): encouraging they remain below decision thresholds.
- Enforces a margin on match scores ($\text{ReLU}(m - \mathcal{S}'_{mat})$): guaranteeing they exceed non-matches by m.

By jointly optimizing score magnitudes and relative margins, the loss aligns training objectives with evaluation metrics (*e.g.*, TAR@FAR), reducing false acceptances while maintaining discriminative power.

4. Experiments

To rigorously validate our method's robustness, we intentionally leverage a diverse set of embedding models spanning multiple modalities, including face recognition model [24, 26], gait recognition and person ReID models [15, 35, 59, 62, 63]. This cross-modal diversity systematically avoids overfitting to any single modality's biases, demonstrating that our framework generalizes across heterogeneous feature spaces. We stress-test our method's ability to harmonize divergent embeddings—a critical requirement for real-world deployment, where the distribution of the test set is unpredictable.

Baseline Setup. We benchmark our method against traditional and contemporary fusion strategies spanning three categories: (1) *Statistical Fusion*: Min/Max score fusion [23], Z-score normalization and min-max normalization [52]; (2) *Representation Harmonization*: Rank-based histogram equalization (RHE) [19]; and (3) *Model-driven learnable score-fusion*: Farsight [34], SVM-based (Support Vector Machine) score fusion (BSSF) [55], Weighted-sum with learnable coefficients [40] and AsymA-O1's asymmetric aggregation [20]. We also compare with SapiensID [27], a SoTA multimodal model for human recognition. This comprehensive comparison validates our method's superiority in balancing discriminative feature preservation.

Dataset	Type	#Subjects (Train/Test/Non-mated)	#Query	#Gallery
CCVID	Video	75 / 151 / 31	834	1074
MEVID	Video	104 / 54 / 11	316	1438
LTCC	Image	77 / 75 / 15	493	7050
BRIAR	Video	775 / 1103 / (566, 522)	10371	12264

Table 1. Statistics of the evaluation set of human recognition benchmarks. BRIAR has two gallery protocols (*i.e.*, 2 non-mated lists) for open-set search. The number of query and gallery indicate the number of images/sequences for image/video datasets.

Evaluation Metrics. We adopt standard person ReID metrics like Cumulative Matching Curve (CMC) at rank-1 and mean Average Precision (mAP) [15, 35]. To holistically assess whole-body biometric systems, we extend evaluation to verification (TAR@FAR: True Acceptance Rate at a False Acceptance Rate) and open-set search (FNIR@FPIR: False Non-Identity Rate at a specified False Positive Identification Rate).

- TAR@FAR reflects real-world security needs: measuring reliable genuine acceptance rates while rejecting impostors within controlled error tolerance.
- FNIR@FPIR handles open-set scenarios (common in surveillance), rejecting unseen identities robustly without compromising known match detection.

Together, these metrics ensure that the proposed methods achieve a balanced trade-off among accuracy (CMC/mAP), security (TAR@FAR), and generalizability (FNIR@FPIR), reflecting real-world deployment requirements through a comprehensive and practical performance evaluation.

Datasets. We evaluate our method on diverse datasets spanning static images, video sequences, multi-view captures, and cross-modal biometric data (shown in Tab. 1) to rigorously assess generalization across varying resolutions, viewpoints, and temporal dynamics. This multi-faceted benchmarking ensures robustness to real-world challenges such as occlusion, motion blur, and sensor heterogeneity, validating practical applicability in unconstrained environments. More details are provided in the Supplementary.

Evaluation Protocol. For CCVID, MEVID, and LTCC, we evaluate under general conditions, as the focus of score-fusion is not only on the Clothes-Changing (CC) scenario. For BRIAR, we follow Farsight [35] and conduct two test settings: Face-Included Treatment, where facial images are clearly visible, and Face-Restricted Treatment, where facial images are in side view or captured from long distances.

4.1. Implementation Details

In our experiments, we set N as either 2 or 3, incorporating multiple modalities as inputs for a comprehensive evaluation. We adopt the methodology of CAFace [25] to precompute gallery features for all training subjects across modalities. Specifically, pre-trained biometric backbones process all video sequences or images in the training dataset before training and use average pooling to generate modality-specific center features as gallery features. For open-set evaluation, we follow [53] to construct 10 random subsets of gallery subjects which contain around 20% of the subjects in the test set as the non-mated lists (numbers of non-mated subjects in Tab. 1), and report the median and standard deviation values. During training, we randomly sample $L = 8$ frames from each tracklet video and aggregate their features, either through averaging or using specific aggregation methods from the models, to produce query-level features. We set the number of experts to $Z = 2$, with $p_1 = w_n$, and $p_2 = 1 - p_1$. δ is set to 3 for CCVID, MEVID, and LTCC, and 20 for BRIAR. $\varepsilon_1, \varepsilon_2, \ldots, \varepsilon_z$ represents 3-layer MLPs. The parameter m in Eq. 5 is set to 3. We use Adam optimizer with a learning rate of $5e^{-5}$ and a weight decay of $1e^{-2}$. We apply a Cosine annealing warm-up strategy to adjust the learning rate. For learnable baseline methods, we train them on the same training set. More details are provided in the Supplementary.

4.2. Experimental Results

Tab. 2, 3, and 4 show the performance of our method on CCVID, MEVID, LTCC, and BRIAR compared with other score-fusion methods. For Z-score and Min-max normalization methods, we average the scores after the normalization. To ensure a fair comparison with GEFF [1], we replace the FR model in GEFF with AdaFace and apply Gallery Enrichment (GE) to our method, as GE adds selected query samples into the gallery. GEFF requires a hyperparameter α to combine the ReID and FR score matrices and cannot extend to three modalities.

In CCVID, the FR model performs particularly well, as most body images are front-view and contain well-captured faces. As a result, the improvement through multimodal fusion is understandably limited. In MEVID, LTCC, and BRIAR (Face-Restricted Treatment), the performance of the FR model is not comparable to that of the ReID models. This is mainly due to (1) the presence of multiple views and varying distances in captured images, which often results in low-quality images, and (2) label noise and detection errors. The performance of score fusion surpasses that of individual models and modalities, suggesting that each model contributes complementary information. Our method effectively harnesses additional useful information in complex scenarios, leading to an even greater performance boost in MEVID and LTCC than in CCVID. While other score-fusion approaches do not consistently perform well across all metrics or need to manually select hyperparameters, our method achieves higher performance across the board, with notable improvements in both closed-set and open-set evaluations, especially in MEVID and BRIAR. Additionally, our approach is generalizable, adapting effectively to various modality combinations, model combinations, and similarity metrics, irrespective of whether the backbones are fine-tuned on the target dataset or not. More experimental results can be found in the Supplementary.

Method	*Comb.*	Rank1↑	mAP↑	TAR↑	FNIR↓
*AdaFace** [24]	♦	94.0	87.9	75.7	13.0 ± 3.5
CAL [15]	♠	81.4	74.7	66.3	52.8 ± 13.3
*BigGait** [63]	♣	76.7	61.0	49.7	71.1 ± 6.1
SapiensID [27]	●	92.6	77.8	-	-
GEFF† [1]	♦ ♠	89.4	87.5	84.0	13.3 ± 1.3
Ours	♦ ♠	**93.3**	**89.5**	**86.9**	**11.4 ± 1.5**
Min-Fusion [23]		87.1	79.2	62.4	48.5 ± 8.7
Max-Fusion [23]		89.9	89.3	73.4	23.0 ± 10.1
Z-score [52]		92.2	90.6	73.9	15.1 ± 1.5
Min-max [52]		91.8	90.9	73.9	15.4 ± 2.5
RHE [19]		91.7	90.2	73.1	16.6 ± 2.5
Weigthed-sum [40]	♦ ♠ ♣	91.7	90.6	73.6	15.4 ± 1.8
Asym-AO1 [20]		92.3	90.0	74.0	15.9 ± 1.7
BSSF [55]		91.8	91.1	73.9	14.1 ± 1.3
Farsight [34]		92.0	91.2	73.9	13.9 ± 1.1
Ours (AdaFace-QE)		92.6	**91.6**	75.0	13.3 ± 1.2
Ours (CAL-QE)		**94.1**	90.8	**76.2**	**12.3 ± 1.4**

(a) **Performance on CCVID Dataset.**

Method	*Comb.*	Rank1↑	mAP↑	TAR↑	FNIR↓
*AdaFace** [24]	♦	25.0	8.1	5.4	98.8 ± 1.2
CAL [15]	♠	52.5	27.1	34.7	67.8 ± 7.3
AGRL [59]	■	51.9	25.5	30.7	69.4 ± 8.9
GEFF† [1]	♦ ♠	32.9	18.8	19.9	78.7 ± 8.1
Ours	♦ ♠	**33.5**	**19.9**	**26.2**	**72.5 ± 10.3**
Min-Fusion [23]		46.8	21.2	28.0	70.4 ± 8.0
Max-Fusion [23]		33.2	14.9	8.3	97.4 ± 1.6
Z-score [52]		54.1	27.4	30.7	66.5 ± 7.0
Min-max [52]		52.8	24.7	25.0	71.3 ± 6.1
RHE [19]		52.8	24.8	25.3	71.2 ± 6.2
Weigthed-sum [40]	♦ ♠ ■	54.1	27.3	30.3	66.3 ± 7.0
Asym-AO1 [20]		52.5	22.9	23.6	71.7 ± 5.8
BSSF [55]		53.5	27.4	30.5	65.9 ± 7.2
Farsight [35]		53.8	25.4	26.6	69.8 ± 6.4
Ours (AdaFace-QE)		**55.7**	**28.2**	**32.9**	64.6 ± 8.2
Ours (CAL-QE)		55.4	27.9	32.5	**64.3 ± 8.7**

(b) **Performance on MEVID Dataset.**

Table 2. Our performance on CCVID and MEVID datasets in the general setting. [Keys: **Best** and second best performance; *Comb.*: model combination; *: zero-shot performance; †: reproduced using AdaFace [24] as the face module; ♦: AdaFace for face modality; ♣; BigGait for gait modality; ♠: CAL of body modality; ■: AGRL for body modality; ●: SapiensID for face and body modality; TAR: TAR@1%FAR; FNIR: FNIR@1%FPIR.]

Method	*Comb.*	Rank1↑	mAP↑	TAR↑	FNIR↓
*AdaFace** [24]	♦	18.5	5.9	2.4	99.8 ± 0.2
CAL [15]	♠	74.4	40.6	36.7	59.7 ± 7.3
AIM [62]	■	74.8	40.9	37.0	66.2 ± 9.2
SapiensID [27]	●	72.0	34.6	-	-
Ours	♠ ■	**75.3**	**42.5**	**38.1**	**58.6 ± 9.6**
Min-Fusion [23]		38.1	13.5	12.4	81.9 ± 6.0
Max-Fusion [23]		62.5	33.3	16.8	94.8 ± 4.7
Z-score [52]		73.0	37.5	30.4	68.7 ± 9.2
Min-max [52]		73.2	38.1	31.9	75.1 ± 9.2
RHE [19]	♦ ♠ ■	70.4	34.2	21.5	78.0 ± 10.0
Weigthed-sum [40]		73.2	37.8	31.3	72.4 ± 8.6
Asym-AO1 [20]		71.2	32.9	19.1	76.3 ± 8.9
BSSF [55]		73.5	39.1	34.2	68.9 ± 8.5
Farsight [34]		73.2	37.8	31.3	72.4 ± 8.6
Ours		**73.8**	**39.6**	**35.0**	**64.3 ± 8.0**

Table 3. Our performance on LTCC. [Keys: **Best** and second best performance; *Comb.*: model combination; *: zero-shot performance; ♦: AdaFace for face modality; ♠: CAL of body modality; ■: AIM for body modality; ●: SapiensID for face and body modality; TAR: TAR@1%FAR; FNIR: FNIR@1%FPIR.]

Method	*Comb.*	*Face Incl. Trt.*			*Face Restr. Trt.*		
		TAR↑	R20↑	FNIR↓	TAR↑	R20↑	FNIR↓
KPRPE [26]	♦	66.5	80.5	54.8	31.5	44.5	81.3
BigGait [63]	♣	66.3	93.1	72.7	61.0	90.4	76.3
CLIP3DReID [35]	♠	55.8	83.5	80.1	47.9	79.3	83.4
Min-Fusion [23]		70.9	86.5	55.6	39.1	58.0	77.1
Max-Fusion [23]		68.7	93.0	72.5	61.6	90.6	76.1
Z-score [52]		78.5	92.3	43.8	51.1	83.9	72.2
Min-max [52]		82.4	**96.0**	46.9	61.4	**91.5**	68.5
RHE [19]	♦ ♣ ♠	82.8	95.7	44.2	64.9	90.8	67.1
Weigthed-sum [40]		84.0	95.4	43.2	62.6	90.2	68.1
Asym-AO1 [20]		83.4	95.1	42.4	58.5	90.0	66.9
Farsight [34]		82.4	95.8	46.1	65.7	91.0	68.2
Ours		**84.5**	**96.0**	**41.2**	**67.9**	90.6	**64.1**

Table 4. Our performance on BRIAR Evaluation Protocol 5.0.0. [Keys: **Best** and second best performance; *Comb.*: model combination; *Face Incl. Trt.*: Face-Included Treatment; *Face Restr. Trt.*: Face-Restricted Treatment; ♦: AdaFace for face modality; ♣: BigGait for gait modality; ♠: CLIP3DReID of body modality; TAR: TAR@0.1%FAR; R20: Rank20; FNIR: FNIR@1%FPIR.]

$\mathcal{L}_{\text{score}}$	QE	Z	Rank1↑	mAP↑	TAR↑	FNIR↓
✗	✗	1	49.4	21.6	23.3	84.0
✓	✗	1	53.8	24.5	25.3	70.4
✗	✗	2	54.1	25.5	30.8	65.4
✓	✗	2	55.1	27.0	31.3	66.5
✓	✓	2	**55.7**	**28.2**	**32.9**	**64.6**

Table 5. Ablation study results on MEVID. In the absence of the QE setting (*i.e.*, QE ✗), we average the outputs from experts. [Keys: TAR= TAR@1%FAR; FNIR= FNIR@1%FPIR.]

4.3. Analysis

Our experiments reveal two critical insights:

1. While existing methods perform well on high-quality facial datasets, they falter under challenging in-the-wild conditions characterized by non-frontal angles and variable capture quality.
2. Our framework demonstrates superior robustness in these complex scenarios, achieving markedly larger performance gains compared to controlled environments.

This divergence stems from fundamental dataset characteristics: constrained benchmarks predominantly contain optimal facial captures where conventional face recognition excels, whereas unconstrained datasets reflect real-world imperfections that degrade reliability. The limitations of prior approaches arise from their dependence on high-quality facial predictions, which introduce noise when inputs diverge from ideal conditions. Conversely, our method dynamically adapts to input quality variations, synthesizing multi-modal cues to maintain accuracy without additional hardware or data requirements. This capability underscores its practical viability in deployment scenarios where sensor fidelity and environmental conditions are unpredictable.

Single Model Could Be Better than Fusion. While fusion methods generally outperform individual models, exceptions exist (*e.g.*, LTCC), where 3-modality fusion underperforms due to weak face modality. However, fusion with CAL and AIM shows better results, serving as a direction for further mitigating such effects in future work. More results are in the Supplementary.

Comparison with SoTA Human Recognition Model. We benchmark against SapiensID [27] on the CCVID and LTCC datasets. While SapiensID demonstrates competitive or superior performance relative to certain score-fusion methods, our method consistently achieves optimal results. This performance advantage substantiates the critical importance of score-fusion algorithm and our proposed QME.

4.4. Ablation Studies

Effects of $\mathcal{L}_{\text{score}}$, QE, and Z. Tab. 5 illustrates the effects of $\mathcal{L}_{\text{score}}$, QE, and the number of score-fusion experts Z. Compared to $\mathcal{L}_{\text{tri}}$, $\mathcal{L}_{\text{score}}$ yields significant performance improvements across all metrics, regardless of z, underscoring the importance of extra boundary for non-match scores. We further observe that increasing the number of experts Z gradually improves performance, indicating that combining multiple experts enriches the model's decision-making process by capturing diverse perspectives in complex multi-modal settings. Moreover, incorporating QE guidance further boosts performance by enabling quality-aware weighting, allowing each expert to focus on the most relevant features for a given input. This reflective weighting strategy allows the experts to learn more effectively by prioritizing high-quality information, ultimately enhancing the overall robustness and accuracy of the model.

Expert	*Face Incl. Trt.*			*Face Restr. Trt.*		
	TAR↑	R20↑	FNIR↓	TAR↓	R20↑	FNIR↓
ε_1	83.6	95.5	41.7	62.0	**90.6**	66.7
ε_2	81.8	95.5	46.6	65.0	**90.6**	68.4
Ours ($\varepsilon_1 + \varepsilon_2$)	**84.5**	**95.7**	**41.2**	**67.9**	**90.6**	**64.1**

Table 6. Effects of the mixture of score-fusion experts on BRIAR. ε_1 has a better performance in *Face Incl. Trt.*, while ε_2 experts in *Face Restr. Trt.*. [Keys: *Face Incl. Trt.*= Face Included Treatment; *Face Restr. Trt.*= Face Restricted Treatment; TAR=TAR@0.1%FAR; R20=Rank20; FNIR=FNIR@1%FPIR.]

Effects of Mixture of Score-fusion Experts. Tab. 6 analyzes the effects of the mixture of score-fusion experts compared to single-expert performance. We conduct the ablation study on BRIAR as Face Included Treatment and Face Restricted Treatment settings are closely related to face quality weights. ε_1 achieves better results in TAR@0.1%FAR for Face Included Treatment and in FNIR@1%FPIR across all settings, while ε_2 performs better in TAR@0.1%FAR for Face Restricted Treatment. This is because the FR model excels in identifying true positive pairs, resulting in lower FNIR@1%FPIR. Guided by p_1, ε_1 learns to prioritize the FR model, while ε_2 focuses on ReID and GR models. Fusing both experts' scores improves overall performance, demonstrating that using multiple experts enhances final performance and allows each expert to capture distinct information.

Effects of QE for Other Modalities. We validate the proposed QE by evaluating the performance of QME using the QE trained from CAL as input to $\mathcal{N}_r$ in Tab. 2 (denoted as *CAL-QE*). When using QE from CAL, the performance is comparable to that of QE from AdaFace, with both significantly outperforming baseline methods. These results demonstrate the flexibility and robustness of QME.

4.5. Visualization

Score Distribution. Fig. 4 visualizes the distribution of non-match scores, match scores, and the threshold FAR@1% for both Z-score and our method on CCVID. To ensure a balanced comparison between the two distributions, we randomly sample an equal number of non-match and match scores. Compared to the Z-score score-fusion, our approach boosts match scores while keeping non-match scores within the same range. This adjustment validates the effects of score triplet loss to improve the model's ability to distinguish between matches and non-matches.

Quality Weights. Fig. 5 visualizes the distribution of predicted quality weights for facial images in the CCVID and MEVID test sets. Note that these weights represent video-level quality weights, obtained by averaging the quality weights of each frame in the video sequence. CCVID has

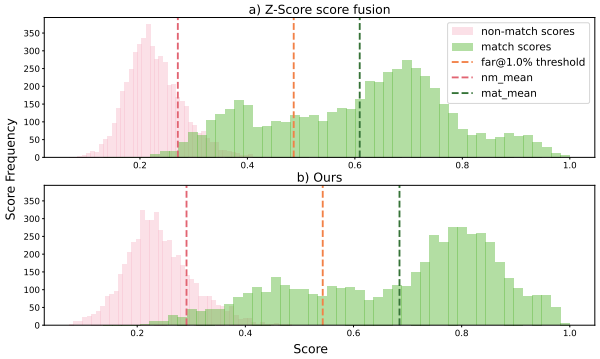

Figure 4. Score distributions of the CCVID test set. [Keys: nm_mean = mean value of non-match scores; mat_mean = mean value of match scores.]

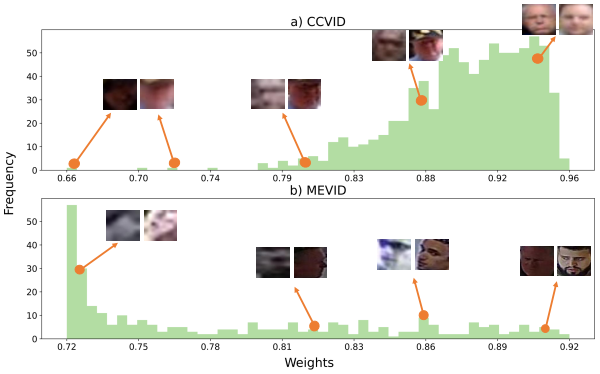

Figure 5. The distribution of AdaFace quality weights for the CCVID and MEVID datasets, illustrated with examples showcasing a range of quality weights.

a higher proportion of high-quality weights, as most images are captured from a front view. In contrast, MEVID shows more variability in quality weights due to detection noise and varying clarity. The visualization indicates that our method effectively estimates image quality. The use of ranking-based pseudo-labels encourages the model to focus on relative quality, making it more robust to outliers. This guides the score-fusion experts to prioritize the most reliable modality based on quality. Visualization of CAL quality weight can be found in the Supplementary.

5. Conclusion

We propose **QME**, a framework for whole-body biometric recognition that dynamically fuses modality-specific experts through a novel quality-aware weighting. To enhance discriminative power, we introduce a score triplet loss that explicitly enforces a margin between match and non-match scores. Experiments across diverse benchmarks demonstrate the superior performance of our method, serving as a general framework for multi-modal score fusion, which can be applied to any system with heterogeneous models.

Acknowledgments. This research is based upon work supported in part by the Office of the Director of National Intelligence (ODNI), Intelligence Advanced Research Projects Activity (IARPA), via 2022-21102100004. The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies, either expressed or implied, of ODNI, IARPA, or the U.S. Government. The U.S. Government is authorized to reproduce and distribute reprints for governmental purposes notwithstanding any copyright annotation therein.

References

[1] Daniel Arkushin, Bar Cohen, Shmuel Peleg, and Ohad Fried. Geff: improving any clothes-changing person ReID model using gallery enrichment with face features. In *WACV*, 2024. 6

[2] Lacey Best-Rowden and Anil K Jain. Learning face image quality from human assessments. *IEEE Transactions on Information forensics and security*, 13(12), 2018. 2

[3] Samarth Bharadwaj, Mayank Vatsa, and Richa Singh. Biometric quality: a review of fingerprint, iris, and face. *EURASIP journal on Image and Video Processing*, 2014, 2014. 2

[4] Jie Chang, Zhonghao Lan, Changmao Cheng, and Yichen Wei. Data uncertainty learning in face recognition. In *CVPR*, 2020. 2

[5] Junwen Chen, Jie Zhu, and Yu Kong. Atm: Action temporality modeling for video question answering. In *ACM MM*, 2023. 1

[6] Mohamed Cheniti, Zahid Akhtar, Chandranath Adak, and Kamran Siddique. An approach for full reinforcement-based biometric score fusion. *IEEE Access*, 2024. 2

[7] David Cornett, Joel Brogan, Nell Barber, Deniz Aykac, Seth Baird, Nicholas Burchfield, Carl Dukes, Andrew Duncan, Regina Ferrell, Jim Goddard, et al. Expanding accurate person recognition to new altitudes and ranges: The briar dataset. In *WACV*, 2023. 2

[8] Yongxing Dai, Xiaotong Li, Jun Liu, Zekun Tong, and Ling-Yu Duan. Generalizable person re-identification with relevance-aware mixture of experts. In *CVPR*, 2021. 4

[9] Maria De Marsico, Michele Nappi, and Daniel Riccio. Cabala—collaborative architectures based on biometric adaptable layers and activities. *PR*, 45(6):2348–2362, 2012. 2

[10] Jiankang Deng, Jia Guo, Niannan Xue, and Stefanos Zafeiriou. Arcface: Additive angular margin loss for deep face recognition. In *CVPR*, 2019. 1

[11] Mohamad El-Abed, Christophe Charrier, and Christophe Rosenberger. Quality assessment of image-based biometric information. *EURASIP Journal on Image and video Processing*, 2015, 2015. 2

[12] William Fedus, Barret Zoph, and Noam Shazeer. Switch transformers: Scaling to trillion parameter models with simple and efficient sparsity. *Journal of Machine Learning Research*, 23(120), 2022. 4

[13] Patrick Grother and Elham Tabassi. Performance of biometric quality measures. *TPAMI*, 29(4), 2007. 2

[14] Artur Grudzien, Marcin Kowalski, and Norbert Palka. Face re-identification in thermal infrared spectrum based on ThermalFaceNet neural network. In *MIKON*, 2018. 2

[15] Xinqian Gu, Hong Chang, Bingpeng Ma, Shutao Bai, Shiguang Shan, and Xilin Chen. Clothes-changing person re-identification with rgb modality only. In *CVPR*, 2022. 1, 5, 6, 7

[16] Xiao Guo, Yaojie Liu, Anil Jain, and Xiaoming Liu. Multi-domain learning for updating face anti-spoofing models. In *ECCV*, 2022. 2

[17] Xiao Guo, Xiufeng Song, Yue Zhang, Xiaohong Liu, and Xiaoming Liu. Rethinking vision-language model in face forensics: Multi-modal interpretable forged face detector. In *CVPR*, 2025. 2

[18] Yuxiang Guo, Cheng Peng, Chun Pong Lau, and Rama Chellappa. Multi-modal human authentication using silhouettes, gait and rgb. In *FG*, 2023. 2

[19] Mingxing He, Shi-Jinn Horng, Pingzhi Fan, Ray-Shine Run, Rong-Jian Chen, Jui-Lin Lai, Muhammad Khurram Khan, and Kevin Octavius Sentosa. Performance evaluation of score level fusion in multimodal biometric systems. *PR*, 43 (5), 2010. 2, 5, 6, 7

[20] Abderrahmane Herbadji, Zahid Akhtar, Kamran Siddique, Noubeil Guermat, Lahcene Ziet, Mohamed Cheniti, and Khan Muhammad. Combining multiple biometric traits using asymmetric aggregation operators for improved person recognition. *Symmetry*, 12(3):444, 2020. 5, 6, 7

[21] Javier Hernandez-Ortega, Javier Galbally, Julian Fierrez, Rudolf Haraksim, and Laurent Beslay. Faceqnet: Quality assessment for face recognition based on deep learning. In *ICB*, 2019. 2

[22] Siyuan Huang, Ram Prabhakar Kathirvel, Chun Pong Lau, and Rama Chellappa. Whole-body detection, recognition and identification at altitude and range. *arXiv preprint arXiv:2311.05725*, 2023. 2

[23] Anil Jain, Karthik Nandakumar, and Arun Ross. Score normalization in multimodal biometric systems. *PR*, 38(12), 2005. 2, 5, 6, 7

[24] Minchul Kim, Anil K Jain, and Xiaoming Liu. Adaface: Quality adaptive margin for face recognition. In *CVPR*, 2022. 1, 2, 5, 6, 7

[25] Minchul Kim, Feng Liu, Anil K Jain, and Xiaoming Liu. Cluster and aggregate: Face recognition with large probe set. In *NeurIPS*, 2022. 2, 3, 6

[26] Minchul Kim, Yiyang Su, Feng Liu, Anil Jain, and Xiaoming Liu. KeyPoint Relative Position Encoding for Face Recognition. In *CVPR*, 2024. 5, 7

[27] Minchul Kim, Dingqiang Ye, Yiyang Su, Feng Liu, and Xiaoming Liu. Sapiensid: Foundation for human recognition. In *CVPR*, 2025. 1, 2, 5, 6, 7

[28] Emine Krichen, Sonia Garcia-Salicetti, and Bernadette Dorizzi. A new probabilistic iris quality measure for comprehensive noise detection. In *BTAS*, 2007. 2

[29] Dmitry Lepikhin, HyoukJoong Lee, Yuanzhong Xu, Dehao Chen, Orhan Firat, Yanping Huang, Maxim Krikun, Noam

Shazeer, and Zhifeng Chen. Gshard: Scaling giant models with conditional computation and automatic sharding. *arXiv preprint arXiv:2006.16668*, 2020. 4

[30] Pei Li, Joel Brogan, and Patrick J Flynn. Toward facial re-identification: Experiments with data from an operational surveillance camera plant. In *BTAS*, 2016. 2

[31] Pei Li, Maria Loreto Prieto, Patrick J Flynn, and Domingo Mery. Learning face similarity for re-identification from real surveillance video: A deep metric solution. In *IJCB*, 2017. 2

[32] Weijia Li, Saihui Hou, Chunjie Zhang, Chunshui Cao, Xu Liu, Yongzhen Huang, and Yao Zhao. An in-depth exploration of person re-identification and gait recognition in cloth-changing conditions. In *CVPR*, 2023. 1

[33] Petro Liashchynskyi and Pavlo Liashchynskyi. Grid search, random search, genetic algorithm: a big comparison for nas. *arXiv preprint arXiv:1912.06059*, 2019. 2

[34] Feng Liu, Ryan Ashbaugh, Nicholas Chimitt, Najmul Hassan, Ali Hassani, Ajay Jaiswal, Minchul Kim, Zhiyuan Mao, Christopher Perry, Zhiyuan Ren, et al. Farsight: A physics-driven whole-body biometric system at large distance and altitude. In *WACV*, 2024. 2, 5, 6, 7

[35] Feng Liu, Minchul Kim, Zhiyuan Ren, and Xiaoming Liu. Distilling CLIP with Dual Guidance for Learning Discriminative Human Body Shape Representation. In *CVPR*, 2024. 1, 5, 6, 7

[36] Feng Liu, Nicholas Chimitt, Lanqing Guo, Jitesh Jain, Aditya Kane, Minchul Kim, Wes Robbins, Yiyang Su, Dingqiang Ye, Xingguang Zhang, et al. Person recognition at altitude and range: Fusion of face, body shape and gait. *arXiv preprint arXiv:2505.04616*, 2025. 2

[37] Qiang Meng, Shichao Zhao, Zhida Huang, and Feng Zhou. Magface: A universal representation for face recognition and quality assessment. In *CVPR*, 2021. 2

[38] Karthik Nandakumar, Yi Chen, Sarat C Dass, and Anil Jain. Likelihood ratio-based biometric score fusion. *TPAMI*, 30 (2), 2007. 2

[39] Necmiye Ozay, Yan Tong, Frederick W Wheeler, and Xiaoming Liu. Improving face recognition with a quality-based probabilistic framework. In *CVPRW*, 2009. 2

[40] Tae Jin Park, Manoj Kumar, and Shrikanth Narayanan. Multi-scale speaker diarization with neural affinity score fusion. In *ICASSP*, 2021. 2, 5, 6, 7

[41] Norman Poh and Josef Kittler. A unified framework for biometric expert fusion incorporating quality measures. *TPAMI*, 34(1), 2011. 2

[42] Norman Poh, Josef Kittler, and Thirimachos Bourlai. Improving biometric device interoperability by likelihood ratio-based quality dependent score normalization. In *BTAS*, 2007. 2

[43] Xuelin Qian, Wenxuan Wang, Li Zhang, Fangrui Zhu, Yanwei Fu, Tao Xiang, Yu-Gang Jiang, and Xiangyang Xue. Long-term cloth-changing person re-identification. In *ACCV*, 2020. 1

[44] Kaijie Ren and Lei Zhang. Implicit Discriminative Knowledge Learning for Visible-Infrared Person Re-Identification. In *CVPR*, 2024. 2

[45] Arun Ross and Anil Jain. Information fusion in biometrics. *PR letters*, 24(13), 2003. 2

[46] Avinab Saha, Sandeep Mishra, and Alan C Bovik. Re-iqa: Unsupervised learning for image quality assessment in the wild. In *CVPR*, pages 5846–5855, 2023. 2

[47] Florian Schroff, Dmitry Kalenichenko, and James Philbin. Facenet: A unified embedding for face recognition and clustering. In *CVPR*, 2015. 5

[48] Noam Shazeer, Azalia Mirhoseini, Krzysztof Maziarz, Andy Davis, Quoc Le, Geoffrey Hinton, and Jeff Dean. Outrageously large neural networks: The sparsely-gated mixture-of-experts layer. *arXiv preprint arXiv:1701.06538*, 2017. 4

[49] Noam Shazeer, Youlong Cheng, Niki Parmar, Dustin Tran, Ashish Vaswani, Penporn Koanantakool, Peter Hawkins, HyoukJoong Lee, Mingsheng Hong, Cliff Young, et al. Mesh-tensorflow: Deep learning for supercomputers. In *NeurIPS*, 2018. 4

[50] Yichun Shi and Anil K Jain. Probabilistic face embeddings. In *ICCV*, 2019. 2

[51] Maneet Singh, Richa Singh, and Arun Ross. A comprehensive overview of biometric fusion. *Information Fusion*, 52, 2019. 1, 2

[52] Robert Snelick, Mike Indovina, James Yen, and Alan Mink. Multimodal biometrics: issues in design and testing. In *ICMI*, 2003. 5, 6, 7

[53] Yiyang Su, Minchul Kim, Feng Liu, Anil Jain, and Xiaoming Liu. Open-set biometrics: Beyond good closed-set models. In *ECCV*, 2024. 2, 6

[54] Yiyang Su, Yunping Shi, Feng Liu, and Xiaoming Liu. Hamobe: Hierarchical and adaptive mixture of biometric experts for video-based person reid. In *ICCV*, 2025. 4

[55] Jackson Horlick Teng, Thian Song Ong, Tee Connie, Kalaiarasi Sonai Muthu Anbananthen, and Pa Pa Min. Optimized score level fusion for multi-instance finger vein recognition. *Algorithms*, 2022. 2, 5, 6, 7

[56] Philipp Terhorst, Jan Niklas Kolf, Naser Damer, Florian Kirchbuchner, and Arjan Kuijper. Ser-fiq: Unsupervised estimation of face image quality based on stochastic embedding robustness. In *CVPR*, 2020. 2

[57] Yan Tong, Frederick W Wheeler, and Xiaoming Liu. Improving biometric identification through quality-based face and fingerprint biometric fusion. In *CVPRW*, 2010. 2

[58] Mayank Vatsa, Richa Singh, and Afzel Noore. Integrating image quality in 2ν-svm biometric match score fusion. *International Journal of Neural Systems*, 17(05), 2007. 2

[59] Yiming Wu, Omar El Farouk Bourahla, Xi Li, Fei Wu, Qi Tian, and Xue Zhou. Adaptive graph representation learning for video person re-identification. *IEEE Transactions on Image Processing*, 29, 2020. 5, 6

[60] Bin Yang, Jun Chen, and Mang Ye. Shallow-Deep Collaborative Learning for Unsupervised Visible-Infrared Person Re-Identification. In *CVPR*, 2024. 2

[61] Qize Yang, Ancong Wu, and Wei-Shi Zheng. Person re-identification by contour sketch under moderate clothing change. *TPAMI*, 43(6), 2019. 1

[62] Zhengwei Yang, Meng Lin, Xian Zhong, Yu Wu, and Zheng Wang. Good is bad: Causality inspired cloth-debiasing for cloth-changing person re-identification. In *CVPR*, 2023. 5, 7

[63] Dingqiang Ye, Chao Fan, Jingzhe Ma, Xiaoming Liu, and Shiqi Yu. BigGait: Learning Gait Representation You Want by Large Vision Models. In *CVPR*, 2024. 1, 5, 6, 7

[64] Mustafa Berkay Yılmaz and Berrin Yanıkoğlu. Score level fusion of classifiers in off-line signature verification. *Information Fusion*, 32, 2016. 2

[65] Yue Zhang, Ben Colman, Xiao Guo, Ali Shahriyari, and Gaurav Bharaj. Common sense reasoning for deepfake detection. In *ECCV*, 2024. 2

[66] Ziyuan Zhang, Luan Tran, Xi Yin, Yousef Atoum, Xiaoming Liu, Jian Wan, and Nanxin Wang. Gait recognition via disentangled representation learning. In *CVPR*, 2019. 1

[67] Zheng Zhu, Guan Huang, Jiankang Deng, Yun Ye, Junjie Huang, Xinze Chen, Jiagang Zhu, Tian Yang, Jiwen Lu, Dalong Du, et al. Webface260m: A benchmark unveiling the power of million-scale deep face recognition. In *CVPR*, 2021. 1

[68] Simiao Zuo, Xiaodong Liu, Jian Jiao, Young Jin Kim, Hany Hassan, Ruofei Zhang, Tuo Zhao, and Jianfeng Gao. Taming sparsely activated transformer with stochastic experts. In *ICLR*, 2022. 4