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

Anonymous ICCV submission

Paper ID

Abstract

*Whole-body biometric recognition is a challenging multi-modal 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 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 multi-modal and multi-model, addressing key challenges such as model misalignment in the similarity score domain and variability in data quality. Code will be publicly released upon publication.*

1. Introduction

Whole-body biometrics integrates diverse recognition tasks such as face recognition (FR) [9, 22], person re-identification (ReID) [15, 32], and gait recognition (GR) [57, 59] 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. Effective fusion is pivotal to whole-body recognition. Current approaches include decision-level fusion, feature-level fusion, and score-level fusion [46]. 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. This fusion scheme does not incorporate any correlation among the modalities. Feature-level fusion combines extracted features from different modalities to obtain a single representation. However, this approach is often hindered by inconsistencies across modalities, as different biometric traits may not necessarily complement each other effectively. Most importantly, this kind of method requires suitable paired multi-modal datasets. Many available datasets such as WebFace42M [60] for face recognition do not contain whole-body data, while other datasets like PRCC [55], LTCC [38], and CCPG [29] 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.

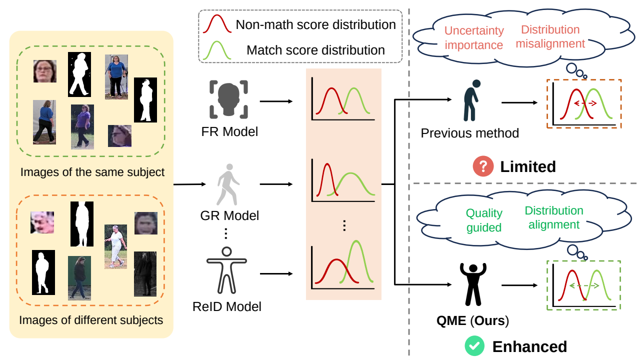

Figure 1. Illustration of score distribution alignment in multi-modal human recognition. Different models and modalities (*e.g.*, face, gait, and body) produce distinct similarity score distributions. Conventional score-fusion methods struggle with optimal alignment and assigning importance weights of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

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 model optimizations. 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 [30], 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 via ranking performance, eliminating laborious manual annotation. We develop a Mixture of Score-Fusion Experts method that each expert learns distinct fusion strategies (*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 score triplet loss that enforces margin separation between match/non-match scores while suppressing non-match magnitudes, directly aligning with metrics like verification and open-search. This approach improves score-level alignment between modalities without the need for retraining biometric backbones and tremendous training data. Our contributions are summarized as follows:

- We propose a Quality Estimator (QE) that employs pseudo quality loss—derived from pretrained models and ranking performance—to assess modality quality without the need for human-labeled data.
- We introduce **QME**, a multi-modal 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 develop a novel score triplet loss for metric learning that enforces a clear margin between match and non-match scores, directly optimizing key performance metrics such as verification accuracy and open-search effectiveness.
- Extensive experiments on multiple whole-body biometric datasets validate the superior performance and robustness of our approach compared to state-of-the-art score-fusion methods.

2. Related Work

2.1. Score-fusion

Score-level fusion integrates similarity scores from multiple modalities to optimize recognition decisions [46]. Traditional score-fusion methods include Z-score and min-max normalization. [34, 36, 37, 51] introduce likelihood ratio based score fusion. Ross *et al.* propose mean, max, or min score-fusion, where the final score is determined by averaging, highest, or lowest score [21, 40, 58]. The Reduction of High-scores Effect (RHE) normalization developed by [17] builds upon the min-max normalization approach by incorporating genuine pair scores. 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) [5, 35, 49]. 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 becomes smaller compared to 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

Biometric quality assessment is the process of evaluating the quality of biometric data (facial images and fingerprints), which directly impacts the performance and accuracy of biometric recognition systems [12]. [3, 10, 25] focus on fingerprint and iris, while [2, 4, 19, 22, 23, 33, 44, 50] focus on quality assessment using learning-based methods in face recognition. However, many of these approaches require specialized training paradigms that are 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 feature detectors, encoders, and fusion modules to unify multi-modal traits (*e.g.*, face, gait) for robust identification. Key to their 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.*, long-distance, off-angle views), while gait and ReID models contend with clothing/posture variations. [31]. Recent advances [6, 16, 20, 39, 48, 54] emphasize multi-attribute fusion, yet predominantly target homogeneous sensor data, neglecting the heterogeneous nature of whole-body modalities. Efforts to incorporate facial features into ReID [13, 27, 28, 31] often prioritize modular additions over-optimizing fusion efficacy. The challenge of fusion methods for comprehensive whole-body biometric recognition remains an open problem requiring in-depth exploration.

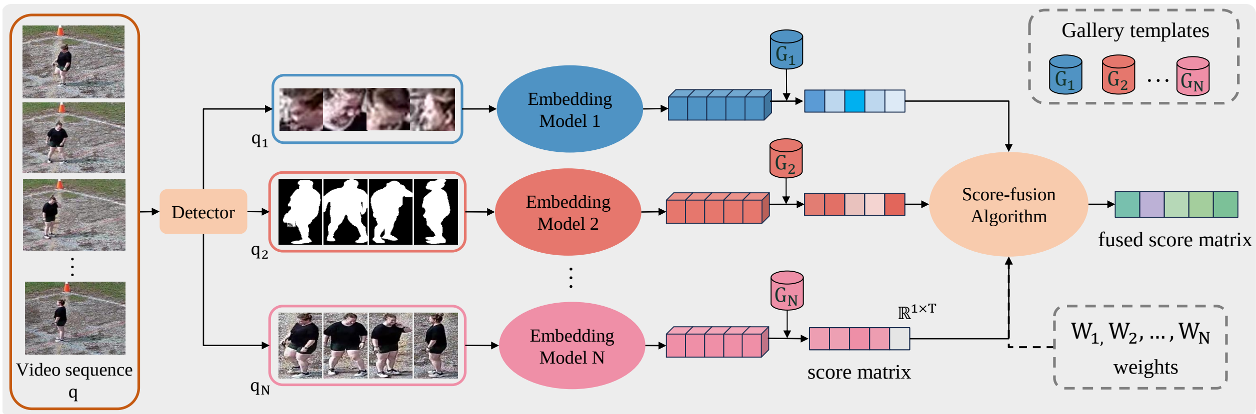

Figure 2. General framework for whole-body biometric recognition. Input video sequence q is processed by a detector to extract different modality queries, which are fed into multiple embedding models. Each model generates similarity scores by comparing the extracted features with gallery templates (T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

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, there are typically multiple queries (or probes) and a fixed set of gallery subjects. A query refers to a sample sequence that needs to be identified or verified, while the gallery consists of previously enrolled or known subjects in the system. Each gallery subject has multiple video sequences (or 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 features $q_n \in \mathbb{R}^{L\times d_n}$ and gallery template features $G_n \in \mathbb{R}^{T\times d_n}$ of all gallery subjects, where L represents the sequence length of the query (number of images) and T is the number of gallery templates (*i.e.*, number of videos/images), and d_n is the feature dimension of M_n. We further compute the average of q_n to obtain a feature vector in $\mathbb{R}^{1\times d_n}$, then compute the similarity between 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-stage training: (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), given a query feature set $\mathbf{Q}_n = \{q_n^1, q_n^2, \ldots, q_n^B\} \in \mathbb{R}^{B\times d_n}$ where B is the training batch size, we collect the intermediate features $\mathcal{I}_n \in \mathbb{R}^{B\times L\times U\times P_n\times d_n}$ from the model 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 [23] to extract intermediate features from the backbone and compute the mean and the standard deviation, reducing $\mathcal{I}_n$ to a representation in $\mathbb{R}^{B\times L\times 2d_n}$. This representation is then fed into an encoder to predict query-level quality weight $W_n \in \mathbb{R}^{B\times 1}$ produced by sigmoid function.

Pseudo Quality Loss. The challenge of training QE is the lack of human-labeled training set quality. 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 believe 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:

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

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. In order to get 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 toleration about the ranking result. Our proposed QE offers several

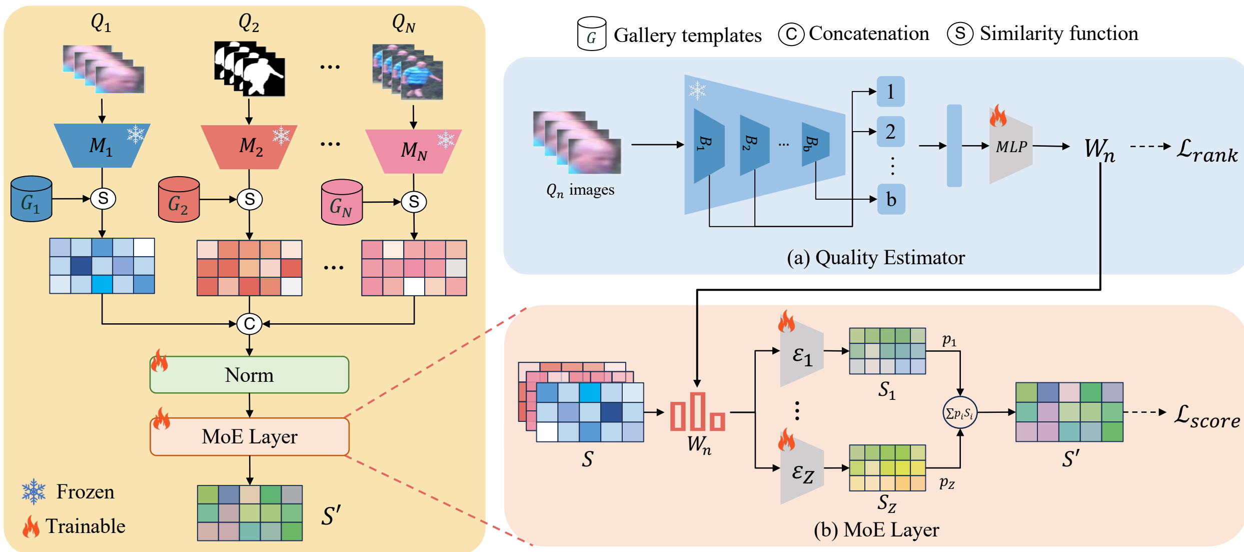

Figure 3. The architecture of the proposed QME framework. It includes a *Norm* layer and an *MoE* layer to process concatenated score matrices $\mathbf{S}$ from the model set $M_1, M_2, \ldots, M_N$. The *MoE* layer contains experts $\varepsilon_1, \varepsilon_2, \ldots, \varepsilon_Z$ to individually 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 $\mathbf{S}'$.

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.

3.2. Mixture of Score-fusion Experts

The concept of MoE [11, 42] 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 [7, 26, 42, 43, 61].

Inspired by this, we design an 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 the traditional MoE setup, 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 a expert ϵ_z from expert set $\{\epsilon_1, ..., \epsilon_Z\}$ where Z is the number of experts, they receive score matrix $\mathbf{S} \in \mathbb{R}^{T \times N}$ from all modalities and predict a fused score matrix $\mathbf{S}'_\mathbf{z} \in \mathbb{R}^{T \times 1}$. 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, another expert, ε_j (controlled by $1 - p_n$), shifts 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. Further details are provided in Sec. 4.4.

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

Based on Sec. 3.1 and 3.2, we further introduce QME. As illustrated on the left side of Fig. 3, for a query feature set $\mathbf{Q}_n = \{q_n^1, q_n^2, \ldots, q_n^B\} \in \mathbb{R}^{B \times d_n}$ processed by model set $\{M_1, M_2, \ldots, M_N\}$, we generate the input score matrix $\mathcal{S} = \{\mathbf{S}_1, \mathbf{S}_2, \ldots, \mathbf{S}_N\} \in \mathbb{R}^{B \times T \times N}$, respectively. N is the number of models. For models employing different distance metrics, such as cosine similarity and Euclidean distance, we convert Euclidean distances into similarity scores using:

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

where $Euc(q, g)$ represents Euclidean distance between feature q and 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 [41] optimizes relative distances between samples:

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

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 that only $\tau\%$ of non-match scores can be accepted as matches. To optimize these metrics, we introduce the score triplet loss $\mathcal{L}_{score}$:

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

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 ($\mathrm{ReLU}(\mathcal{S}'_{nm})$): ensuring they remain below decision thresholds.
- Enforces a margin on match scores ($\mathrm{ReLU}(m - \mathcal{S}'_{mat})$): 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 [22, 24], gait recognition and person ReID models [15, 32, 53, 56, 57] 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.

Dataset	Type	#Subjects (Train/Test)	#Query	#Gallery
CCVID	Video	75 / 151	834	1074
MEVID	Video	104 / 54	316	1438
LTCC	Image	77 / 75	493	7050
BRIAR	Video	775 / 424	10371	12264

Table 1. Statistics of the evaluation set of human recognition benchmarks. For the LTCC, the numbers indicate the number of images, while others are the number of sequences.

Baseline Setup. We benchmark our method against traditional and contemporary fusion strategies spanning three categories: (1) *Statistical Fusion*: Min/Max score fusion [21], Z-score normalization and min-max normalization [47]; (2) *Representation Harmonization*: Rank-based histogram equalization (RHE) [17]; and (3) *Model-driven learnable score-fusion*: Farsight [31], SVM-based (Support Vector Machine) score fusion (BSSF) [49], Weighted-sum with learnable coefficients [35] and AsymA-O1's asymmetric aggregation [18]. This comprehensive comparison validates our method's superiority in balancing discriminative feature preservation.

Evaluation Metrics. We adopt standard person ReID metrics like Cumulative Matching Curve (CMC) at rank-1 and mean Average Precision (mAP) [8, 14, 15, 32, 38, 45, 52, 53, 56, 57, 59]. 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 False Positive Identification Rate).

- TAR@FAR directly aligns with real-world security needs, measuring how reliably the system accepts genuine matches while rejecting imposters under controlled error tolerance.
- FNIR@FPIR addresses open-set scenarios (common in surveillance), where queries may belong to unknown individuals, ensuring robust rejection of "unknowns" without compromising true match detection.

These metrics collectively ensure methods balance accuracy (CMC/mAP), security (TAR@FAR), and generalizability (FNIR@FPIR), reflecting real-world deployment requirements with comprehensive 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 [32] 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 $N = 2, 3$, incorporating multiple modalities (face, gait, and body) as inputs for a comprehensive evaluation. We adopt the methodology of CAFace [23] to precompute gallery features for all training subjects across multiple biometric modalities. Specifically, pre-trained biometric backbones process each video sequence or image in the training dataset before training begins, and use average pooling to generate modality-specific gallery features. For open-set evaluation, we follow Su *et al.*'s work [48] to construct 10 random subsets of gallery subjects (covering 20% of the subjects in the test set) 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 $p_1 = W_f$, and $p_2 = 1 - p_1$. δ in Eq. 1 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. More details are provided in 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. Note that Z-score and min-max are normalization methods; after normalization, we average the scores for a more balanced comparison. 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. That is because GE adds selected query samples into the gallery, so the test set has changed. Note that GEFF requires a hyperparameter α to combine the score matrices from the ReID model and the FR model, which cannot be extended to the 3-modality setting.

In CCVID, the FR model performs particularly well, as most body images are front-view and contain well-captured faces. 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. However, 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 (+1.6% on Rank1, +0.8% on mAP, +2.2% on TAR@1%FAR and +1.3% on FNIR@1%FPIR on MEVID). 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 no-

Method	*Comb.*	Rank1↑	mAP↑	TAR↑	FNIR↓
*AdaFace** [22]	♦	94.0	87.9	75.7	13.0 ± 3.5
CAL [15]	♠	81.4	74.7	66.3	52.8 ± 13.3
*BigGait** [57]	♣	76.7	61.0	49.7	71.1 ± 6.1
GEFF† [1]	♦ ♠	89.4	87.5	84.0	13.3 ± 1.3
Ours		**93.3**	**89.5**	**86.9**	$\mathbf{11.4 \pm 1.5}$
Min-Fusion [21]		87.1	79.2	62.4	48.5 ± 8.7
Max-Fusion [21]		89.9	89.3	73.4	23.0 ± 10.1
Z-score [47]		92.2	90.6	73.9	15.1 ± 1.5
Min-max [47]		91.8	90.9	73.9	15.4 ± 2.5
RHE [17]		91.7	90.2	73.1	16.6 ± 2.5
Weigthed-sum [35]	♦ ♠ ♣	91.7	90.6	73.6	15.4 ± 1.8
Asym-AO1 [18]		92.3	90.0	74.0	15.9 ± 1.7
BSSF [49]		91.8	91.1	73.9	14.1 ± 1.3
Farsight [31]		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**	$\mathbf{12.3 \pm 1.4}$

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

Method	*Comb.*	Rank1↑	mAP↑	TAR↑	FNIR↓
*AdaFace** [22]	♦	25.0	8.1	5.4	98.8 ± 1.2
CAL [15]	♠	52.5	27.1	34.7	67.8 ± 7.3
AGRL [53]	■	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**	$\mathbf{72.5 \pm 10.3}$
Min-Fusion [21]		46.8	21.2	28.0	70.4 ± 8.0
Max-Fusion [21]		33.2	14.9	8.3	97.4 ± 1.6
Z-score [47]		54.1	27.4	30.7	66.5 ± 7.0
Min-max [47]		52.8	24.7	25.0	71.3 ± 6.1
RHE [17]		52.8	24.8	25.3	71.2 ± 6.2
Weigthed-sum [35]	♦ ♠ ■	54.1	27.3	30.3	66.3 ± 7.0
Asym-AO1 [18]		52.5	22.9	23.6	71.7 ± 5.8
BSSF [49]		53.5	27.4	30.5	65.9 ± 7.2
Farsight [32]		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	$\mathbf{64.3 \pm 8.7}$

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

Table 2. Our performance on CCVID and MEVID datasets in the general setting. **Bold**: best performance. Underline: second best performance. *Comb.*: model combination. *: zero-shot performance. †: reproduced using AdaFace [22] as the face module. ♦: AdaFace for face modality. ♣: BigGait for gait modality. ♠: CAL of body modality. ■: AGRL for body modality. [Keys: TAR=TAR@0.1%FAR. FNIR= FNIR@1%FPIR.].

Method	*Comb.*	Rank1↑	mAP↑	TAR↑	FNIR↓
*AdaFace** [22]	♦	18.5	5.9	2.4	99.8 ± 0.2
CAL [15]	♠	74.4	40.6	36.7	59.7 ± 7.3
AIM [56]	■	74.8	40.9	37.0	66.2 ± 9.2
Min-Fusion [21]		38.1	13.5	12.4	81.9 ± 6.0
Max-Fusion [21]		62.5	33.3	16.8	94.8 ± 4.7
Z-score [47]		73.0	37.5	30.4	68.7 ± 9.2
Min-max [47]		73.2	38.1	31.9	75.1 ± 9.2
RHE [17]	♦ ♠ ■	70.4	34.2	21.5	78.0 ± 10.0
Weigthed-sum [35]		73.2	37.8	31.3	72.4 ± 8.6
Asym-AO1 [18]		71.2	32.9	19.1	76.3 ± 8.9
BSSF [49]		73.5	39.1	34.2	68.9 ± 8.5
Farsight [31]		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. **Bold**: best performance. Underline: second best performance. *Comb.*: model combination. *: zero-shot performance. ♦: AdaFace for face modality. ♠: CAL of body modality. ■: AIM for body modality. [Keys: TAR=TAR@0.1%FAR. FNIR= FNIR@1%FPIR.]

Method	*Comb.*	*Face Incl. Trt.*			*Face Restr. Trt.*		
		TAR↑	R20↑	FNIR↓	TAR↑	R20↑	FNIR↓
KPRPE [24]	♦	66.5	80.5	54.8	31.5	44.5	81.3
BigGait [57]	♣	66.3	93.1	72.7	61.0	90.4	76.3
CLIP3DReID [32]	♠	55.8	83.5	80.1	47.9	79.3	83.4
Min-Fusion [21]		70.9	86.5	55.6	39.1	58.0	77.1
Max-Fusion [21]		68.7	93.0	72.5	61.6	90.6	76.1
Z-score [47]		78.5	92.3	43.8	51.1	83.9	72.2
Min-max [47]		82.4	**96.0**	46.9	61.4	**91.5**	68.5
RHE [17]	♦ ♣ ♠	82.8	95.7	44.2	64.9	90.8	67.1
Weigthed-sum [35]		84.0	95.4	43.2	62.6	90.2	68.1
Asym-AO1 [18]		83.4	95.1	42.4	58.5	90.0	66.9
Farsight [31]		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. **Bold**: best performance. Underline: 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. [Keys: TAR=TAR@0.1%FAR. R20= Rank20. FNIR= FNIR@1%FPIR.]

table 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.

4.3. Analysis

Our experiments reveal two critical insights: First, while existing methods enhance performance on constrained datasets with high-quality facial imagery, they falter under challenging in-the-wild conditions characterized by non-frontal angles and variable capture quality. Second, 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 feature 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.

$\mathcal{L}_{\text{score}}$	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.]

4.4. 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 the $\mathcal{L}_{\text{tri}}$, $\mathcal{L}_{\text{score}}$ 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 leads to incremental performance improvements. This trend suggests that the fusion of multiple experts enriches the model's decision-making process by capturing diverse perspectives, making it better equipped to handle complex, multi-modal data scenarios. Lastly, the inclusion of QE guidance results in even further performance enhancements. QE allows for quality-based weighting, which enables 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. We analyze the effects of the mixture of score-fusion experts compared to single-expert performance, as shown in Tab. 6. 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 generalizability of the proposed QE with the performance of **QME** using the QE of CAL as the 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 outperforming baseline methods. Visualization of CAL quality weight can be found in the Supplementary.

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 increases match scores while keeping non-match scores within the same range. This adjustment validates the effects of score triplet loss. This improved 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 on 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 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. This guides the score-fusion experts to prioritize the most reliable modality based on quality.

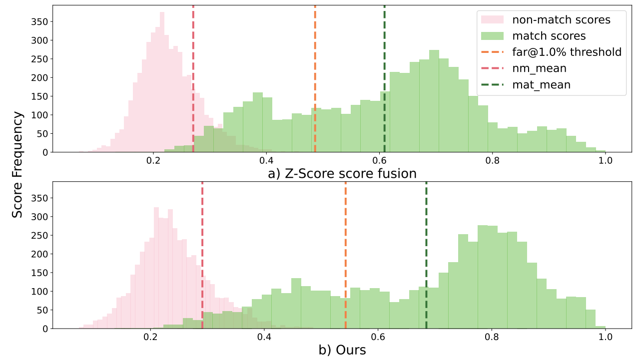

Figure 4. Score distributions of the CCVID test set. [Keys: nm_mean=mean value of non-match scores; mat_mean= mean value of match scores.]

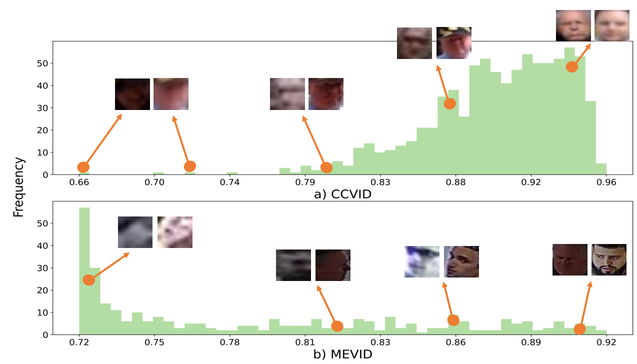

Figure 5. The distribution of AdaFace quality weights for the CCVID and MEVID datasets, illustrated with examples showcasing a range of quality weights.

5. Conclusion

We propose **QME** (Quality-guided Mixture of Experts), a framework for robust whole-body biometric recognition that dynamically fuses modality-specific experts through quality-aware weighting. The proposed score triplet loss enforces the margin between match and non-match scores. Experiments across diverse benchmarks demonstrate the superior performance of our method. QME serves as a general framework for multi-modal score fusion—applicable to any system combining heterogeneous models.

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 *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 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 *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2020. 2

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

[6] 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 *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 2023. 2

[7] Yongxing Dai, Xiaotong Li, Jun Liu, Zekun Tong, and Ling-Yu Duan. Generalizable person re-identification with relevance-aware mixture of experts. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2021. 4

[8] Daniel Davila, Dawei Du, Bryon Lewis, Christopher Funk, Joseph Van Pelt, Roderic Collins, Kellie Corona, Matt Brown, Scott McCloskey, Anthony Hoogs, et al. Mevid: Multi-view extended videos with identities for video person re-identification. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 2023. 5

[9] Jiankang Deng, Jia Guo, Niannan Xue, and Stefanos Zafeiriou. Arcface: Additive angular margin loss for deep face recognition. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2019. 1

[10] 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

[11] 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

[12] Patrick Grother and Elham Tabassi. Performance of biometric quality measures. *IEEE transactions on pattern analysis and machine intelligence*, 29(4), 2007. 2

[13] Artur Grudzien, Marcin Kowalski, and Norbert Palka. Face re-identification in thermal infrared spectrum based on ThermalFaceNet neural network. In *2018 22nd International Microwave and Radar Conference (MIKON)*. IEEE, 2018. 2

[14] Xinqian Gu, Hong Chang, Bingpeng Ma, Hongkai Zhang, and Xilin Chen. Appearance-preserving 3d convolution for video-based person re-identification. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part II 16*. Springer, 2020. 5

[15] Xinqian Gu, Hong Chang, Bingpeng Ma, Shutao Bai, Shiguang Shan, and Xilin Chen. Clothes-changing person re-identification with rgb modality only. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2022. 1, 5, 6, 7

[16] Yuxiang Guo, Cheng Peng, Chun Pong Lau, and Rama Chellappa. Multi-modal human authentication using silhouettes, gait and rgb. In *2023 IEEE 17th International Conference on Automatic Face and Gesture Recognition (FG)*. IEEE, 2023. 2

[17] 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. *Pattern Recognition*, 43(5), 2010. 2, 5, 6, 7

[18] 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

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

[20] 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

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

[22] Minchul Kim, Anil K Jain, and Xiaoming Liu. Adaface: Quality adaptive margin for face recognition. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2022. 1, 2, 5, 6, 7

[23] Minchul Kim, Feng Liu, Anil K Jain, and Xiaoming Liu. Cluster and aggregate: Face recognition with large probe set. *Advances in Neural Information Processing Systems*, 35, 2022. 2, 3, 6

[24] Minchul Kim, Yiyang Su, Feng Liu, Anil Jain, and Xiaoming Liu. KeyPoint Relative Position Encoding for Face Recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024. 5, 7

[25] Emine Krichen, Sonia Garcia-Salicetti, and Bernadette Dorizzi. A new probabilistic iris quality measure for comprehensive noise detection. In *2007 First IEEE International Conference on Biometrics: Theory, Applications, and Systems*. IEEE, 2007. 2

[26] 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

[27] Pei Li, Joel Brogan, and Patrick J Flynn. Toward facial re-identification: Experiments with data from an operational surveillance camera plant. In *2016 IEEE 8th International Conference on Biometrics Theory, Applications and Systems (BTAS)*. IEEE, 2016. 2

[28] 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 *2017 IEEE International Joint Conference on Biometrics (IJCB)*. IEEE, 2017. 2

[29] 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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023. 1

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

[31] 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 *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 2024. 2, 5, 6, 7

[32] Feng Liu, Minchul Kim, Zhiyuan Ren, and Xiaoming Liu. Distilling CLIP with Dual Guidance for Learning Discriminative Human Body Shape Representation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024. 1, 5, 6, 7

[33] Qiang Meng, Shichao Zhao, Zhida Huang, and Feng Zhou. Magface: A universal representation for face recognition and quality assessment. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2021. 2

[34] Karthik Nandakumar, Yi Chen, Sarat C Dass, and Anil Jain. Likelihood ratio-based biometric score fusion. *IEEE transactions on pattern analysis and machine intelligence*, 30(2), 2007. 2

[35] Tae Jin Park, Manoj Kumar, and Shrikanth Narayanan. Multi-scale speaker diarization with neural affinity score fusion. In *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2021. 2, 5, 6, 7

[36] Norman Poh and Josef Kittler. A unified framework for biometric expert fusion incorporating quality measures. *IEEE transactions on pattern analysis and machine intelligence*, 34(1), 2011. 2

[37] Norman Poh, Josef Kittler, and Thirimachos Bourlai. Improving biometric device interoperability by likelihood ratio-based quality dependent score normalization. In *2007 First IEEE International Conference on Biometrics: Theory, Applications, and Systems*. IEEE, 2007. 2

[38] 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 *Proceedings of the Asian Conference on Computer Vision*, 2020. 1, 5

[39] Kaijie Ren and Lei Zhang. Implicit Discriminative Knowledge Learning for Visible-Infrared Person Re-Identification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024. 2

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

[41] Florian Schroff, Dmitry Kalenichenko, and James Philbin. Facenet: A unified embedding for face recognition and clustering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015. 5

[42] 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

[43] 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. *Advances in neural information processing systems*, 31, 2018. 4

[44] Yichun Shi and Anil K Jain. Probabilistic face embeddings. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2019. 2

[45] Xiujun Shu, Xiao Wang, Xianghao Zang, Shiliang Zhang, Yuanqi Chen, Ge Li, and Qi Tian. Large-scale spatio-temporal person re-identification: Algorithms and benchmark. *IEEE Transactions on Circuits and Systems for Video Technology*, 32(7), 2021. 5

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

[47] Robert Snelick, Mike Indovina, James Yen, and Alan Mink. Multimodal biometrics: issues in design and testing. In *Proceedings of the 5th international conference on Multimodal interfaces*, 2003. 5, 6, 7

[48] Yiyang Su, Minchul Kim, Feng Liu, Anil Jain, and Xiaoming Liu. Open-set biometrics: Beyond good closed-set models. In *European Conference on Computer Vision*. Springer, 2025. 2, 6

[49] 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

[50] 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 *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2020. 2

[51] 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

[52] Fangbin Wan, Yang Wu, Xuelin Qian, Yixiong Chen, and Yanwei Fu. When person re-identification meets changing clothes. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition workshops*, 2020. 5

[53] 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

[54] Bin Yang, Jun Chen, and Mang Ye. Shallow-Deep Collaborative Learning for Unsupervised Visible-Infrared Person Re-Identification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024. 2

[55] Qize Yang, Ancong Wu, and Wei-Shi Zheng. Person re-identification by contour sketch under moderate clothing change. *IEEE transactions on pattern analysis and machine intelligence*, 43(6), 2019. 1

[56] 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 *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2023. 5, 7

[57] Dingqiang Ye, Chao Fan, Jingzhe Ma, Xiaoming Liu, and Shiqi Yu. BigGait: Learning Gait Representation You Want by Large Vision Models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024. 1, 5, 6, 7

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

[59] Ziyuan Zhang, Luan Tran, Xi Yin, Yousef Atoum, Xiaoming Liu, Jian Wan, and Nanxin Wang. Gait recognition via disentangled representation learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2019. 1, 5

[60] 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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021. 1

[61] 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. *arXiv preprint arXiv:2110.04260*, 2021. 4