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

Supplementary Material

6. Dataset Description

CCVID Dataset. CCVID [17] is a video-based clothes-changing ReID benchmark containing 75 identities for training and 151 for evaluation. There are 834 query sequences and 1074 gallery sequences in the test set.

MEVID Dataset. MEVID [9] is a multi-view video-based person ReID dataset published in 2023 with 104 training subjects and 54 testing subjects. It contains 316 query tracklets and 1438 gallery tracklets for evaluation.

LTCC Dataset. LTCC [45] is an image-based person ReID dataset comprising 77 training subjects and 75 testing subjects. For evaluation, it includes 493 query images and 7,050 gallery images.

BRIAR Dataset. Biometric Recognition and Identification at Altitude and Range (BRIAR) is a large whole-body multimodal biometric dataset published by IARPA [7]. It contains videos captured from various altitudes, distances, and angles. For training, we use BRS1-4 (775 subjects) as the training dataset. We use the latest Protocol EVP5.0.0 BLENDED, which contains 1103 subjects with 10371 queries for performance evaluation.

7. Additional Implementation Details

For the CCVID, MEVID, and LTCC datasets, we use RetinaFace [12] with a default detection threshold of 0.9 as the face detector to generate facial images. For BRIAR, we employ an internal face-body joint detector to obtain both facial and body images.

Gallery Features of Training Set. For CCVID and MEVID, we compute the center features as the gallery features based on the subject ID, camera ID, and clothing ID. Therefore, each subject may have multiple gallery features. For BRIAR, we perform subject-level center features and derive center features from images or videos categorized as "Control," as these are high-quality indoor captures. If "Control" images or videos are unavailable for a person, we randomly sample 100 frames from 50 videos instead.

Similarity Distances of Each Model. We follow the original papers to measure the distances between features. ArcFace [11], KPRPE [28], AdaFace [26], CAL [17] AGRL [61], and CLIP3DReID [37] use cosine-similarity to measure the distance, while BigGait [65] uses Euclidean distance. We use Eq. 2 to transform Euclidean distance into similarity scores.

Baseline Implementation. We implement Asym-AO1 [22], BSSF [57], and Weighted-sum [42] based on the paper. We train them on the training dataset with the default hyperparameters. For Weighted-sum [42], we use grid search to determine the best modality weights combination on the training dataset and evaluate the test set.

8. Additional Experimental Results

Our method supports both multimodal and multi-model score-level fusion, as demonstrated in the main paper. Additionally, it applies to unimodal score-level fusion. In the unimodal setting, we replace QE with the average sum of the score-fusion experts, as the input originates from a single modality.

Tab. 7 shows the results of our performance on unimodal combinations on CCVID. Our method outperforms the baseline methods in CCVID, and a similar results to the baseline method in LTCC. We hypothesize that the similar performance is due to CAL and AIM sharing similar model structures and being trained on the same dataset, leading to comparable decision outcomes. However, the FNIR@1%FPIR differs, with our method performing slightly better than the baseline.

Computation Cost of QME. The runtime cost of QME with 1k probes and 10k gallery is 0.03s with preselected quality weights, and 0.04s for real-time quality weight predictions on an RTX 4070 GPU.

9. Additional Ablation Experiments

Effects of Hyperparameters. We show the effect of different batch sizes B and sequence lengths L in Tab. 8. The performance remains consistent with respect to B and L, likely due to two key factors: (1) the range of scores is relatively narrow, which reduces sensitivity to variations in batch size and sequence length; and (2) our model efficiently captures the underlying patterns in the score, achieving stable performance even with smaller sequences and batch sizes. This stability highlights the model's robustness across a range of hyperparameter settings, suggesting that it effectively leverages available data without requiring extensive selection of B or L.

Effects of Ranking Threshold δ. Fig. 6 shows the predicted quality weight W_f of the selected facial images with QEs trained with different δ. A larger δ results in a larger W_f as we discuss in Sec. 3.1. With δ, we can control the sensitivity of quality weights.

Method	*Comb.*	Rank1↑	mAP↑	TAR↑	FNIR↓
*AdaFace** [26]	♦	94.0	87.9	75.7	13.0 ± 3.5
*ArcFace** [11]	♣	93.2	85.3	69.1	22.5 ± 6.8
Min-Fusion [25]		93.3	86.3	71.1	16.0 ± 6.6
Max-Fusion [25]		**94.1**	87.4	73.7	16.4 ± 4.7
Mean-Fusion [25]		**94.1**	87.1	71.8	15.6 ± 5.9
Z-score [54]	♦ ♣	**94.1**	87.1	71.9	15.5 ± 5.9
Min-max [54]		**94.1**	87.1	71.8	15.7 ± 5.9
RHE [21]		**94.1**	87.1	71.7	15.7 ± 5.9
Ours		93.9	**87.6**	**75.4**	$\mathbf{12.8 \pm 3.1}$

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

Method	*Comb.*	Rank1↑	mAP↑	TAR↑	FNIR↓
CAL [17]	♠	74.4	40.6	36.7	59.7 ± 7.3
AIM [64]	■	74.8	40.9	37.0	66.2 ± 7.5
Min-Fusion [25]		74.4	41.9	37.7	59.5 ± 11.4
Max-Fusion [25]		73.6	41.5	37.5	60.7 ± 6.8
Mean-Fusion [25]		**75.3**	**42.5**	**38.1**	58.7 ± 9.9
Z-score [54]	♠ ■	**75.3**	**42.5**	**38.1**	58.9 ± 9.9
Min-max [54]		**75.3**	**42.5**	**38.1**	58.8 ± 10.0
RHE [21]		75.1	**42.5**	**38.1**	59.0 ± 9.9
Ours		**75.3**	**42.5**	**38.1**	$\mathbf{58.6 \pm 9.6}$

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

Table 7. Performance on CCVID and LTCC. [Keys: **Best** and second best performance; *Comb.*: model combination; *: zero-shot performance; ♦: AdaFace for face modality; ♣: ArcFace for face modality; ♠: CAL of body modality; ■: AIM for body modality; TAR: TAR@1%FAR; FNIR: FNIR@1%FPIR.]

B	L	Rank1	mAP	TAR@1%FAR	FNIR@1%FPIR
8	1	92.6	91.6	74.7	13.3 ± 1.1
8	8	92.6	91.6	75.0	13.3 ± 1.2
8	16	92.3	91.6	75.0	13.3 ± 1.2
16	1	92.6	91.6	74.2	13.4 ± 1.0
16	8	92.6	91.6	74.5	13.4 ± 1.0
16	16	92.6	91.6	74.6	13.4 ± 1.0

Table 8. Effects of hyperparameters B and L on CCVID.

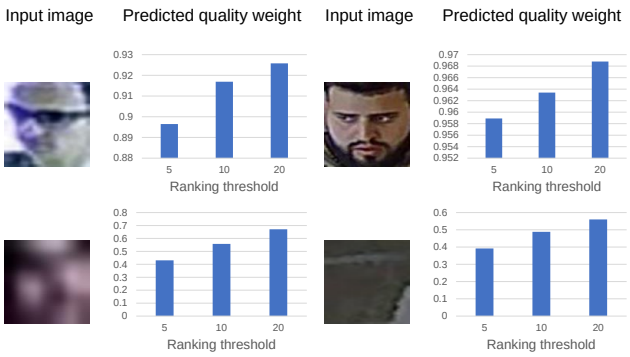

Figure 6. Effects of QEs with different δ in MEVID.

Effects of Quality Assessment Methods. We analyze the effects of using the norm of AdaFace [26] as quality weights for the input of $\mathcal{N}_r$. Since the face feature norm ranges from $[0, +\infty]$, we apply a transformation: $1 - 1/N_f$ where N_f is the face feature norm. Our result using QE is slightly better than using N_f. Moreover, compared to N_f, our QE offers greater flexibility by adjusting the ranking threshold (details in Sec. 9), and is applicable to other modalities. Further exploration of alternative quality assessment methods is encouraged in future work.

Method	Rank1↑	mAP↑	TAR↑	FNIR↓
Ours	92.6	91.6	75.0	13.3 ± 1.2
Ours (Norm)	93.9	89.0	75.6	14.0 ± 2.9

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

Method	Rank1↑	mAP↑	TAR↑	FNIR↓
Ours	55.7	28.2	32.9	64.6 ± 8.2
Ours (Norm)	55.4	27.9	32.1	63.8 ± 8.2

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

Table 9. Our performance on CCVID and MEVID datasets in the general setting. *Ours (Norm)*: the norm of AdaFace features as the quality weights. [Keys: TAR=TAR@1%FAR; FNIR=FNIR@1%FPIR.]

δ	m	CCVID		LTCC	
		Rank1	mAP	Rank1	mAP
3	3	94.1 ± 0.08	90.4 ± 0.45	74.2 ± 0.43	40.0 ± 0.6
20	3	94.2 ± 0.10	91.2 ± 0.37	75.1 ± 0.20	41.8 ± 0.60
3	1	94.0 ± 0.05	90.9 ± 0.42	75.0 ± 0.34	42.0 ± 0.48

Table 10. Effects of δ, m on CCVID and LTCC.

Effects of QE for other modalities. We visualize the quality weight distribution of CCVID and MEVID in Fig. 7, and the performance of QME using the QE of CAL as the input to $\mathcal{N}_r$ in Tab. 2 in the main paper (denote as *CAL-QE*).

Statistical Test. We provide statistical tests with mean and standard deviation in the last three rows of Tab. 10 using seeds 42, 333, and 2025. Our method demonstrates consistent performance across all trials.

Effects of δ, m. We provide additional ablation studies on ranking threshold δ and margin m in Tab. 10, rows 2 and 3. Our method is robust to hyperparameter settings. Z is set to 2 since the quality weight is a scalar in [0,1].

10. Facial Data Cleaning for MEVID

We illustrate the extra facial data cleaning process that applies in MEVID. We observe that there are many imposter subjects or false positive samples in each subject due to detection errors, as shown in imposters in Fig. 8. We use AdaFace [26] to compute feature vectors for all images in the training set. After that, we calculate the score distribution among all images belonging to the same subject.

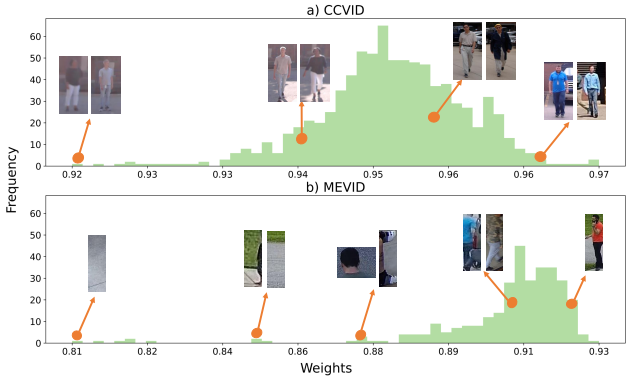

Figure 7. The distribution of CAL quality weights for the CCVID and MEVID datasets, illustrated with examples showcasing a range of quality weights.

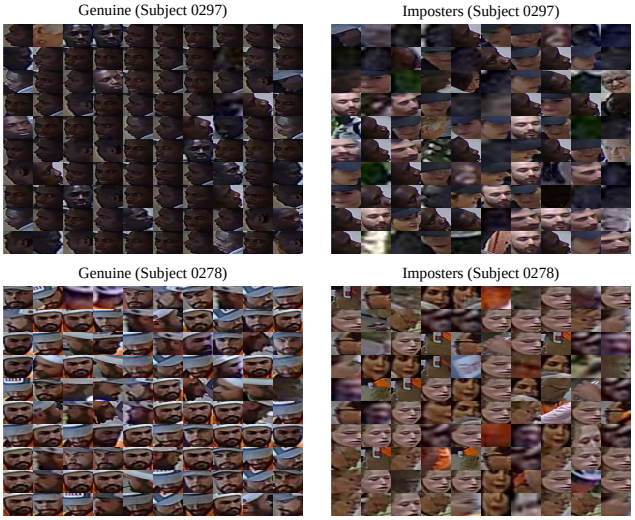

Figure 8. Visualization of imposters' facial images in MEVID.

Assuming we have a chunk size of 100 images for a subject, we can get a self-similarity score matrix with dimension $\mathbb{R}^{100\times100}$. We filter out the scores that are not in $[\mu - \alpha * \sigma, \mu + \alpha * \sigma]$, where μ and σ are the mean and standard deviation of self-similarity score matrix, α is a hyperparameter to control the sensibility of the threshold. We set up $\alpha = 0.7$ in our case. We set up the chunk size as 3000 and apply data cleaning for both the training and testing sets. Fig. 8 visualizes the filtered facial images (imposters) of selected subjects in the MEVID. Note that there are still some false positive samples remaining in the test set, and all baseline methods use the same dataset for fair comparison.

11. Limitation

While our method greatly enhances whole-body biometric performance, its impact on other domains remains unexplored. Future work could extend its application to broader tasks. Moreover, some other router functions or improving the number of score-fusion experts for MoE can be further explored to understand the effects of expert learning. Future research could investigate its application to broader tasks, extending its effectiveness across diverse domains.

12. Potential Societal Impacts

Our paper leverages multiple public biometric datasets for research purposes, with a focus on the similarity score domain, which is less directly tied to sensitive biometric data. As biometric recognition tasks grow increasingly complex, integrating multiple models has become a key trend to enhance system performance. It is essential to ensure that the use of biometric datasets and recognition systems adheres to ethical standards and complies with privacy regulations.