International Journal of Computer Information Systems and Industrial Management Applications
ISSN 2150-7988 Volume 18(11s) 2026 pp. 138-153
https://doi.org/10.70917/ijcisim-2026-3749
Article

Uncertainty-Aware Cross-Modal Feature Interaction Fusion for Face and Full-Hand Biometric Identification

Ahmed Alsafi[1,2], Yaghoub Farjami[1]

[1]Department of Computer and Information Technology, University of Qom, Qom, Iran
[2]Almakeen Institution for Science and Knowledge, Najaf, Iraq, ahmed.alsafi90@gmail.com
ORCID: https://orcid.org/0009-0004-3337-0283

Abstract: —Combining more than one biometric trait usually gives more reliable recognition than a single trait, but how the features are merged matters. Most systems simply concatenate the feature vectors, discarding information about how the two traits relate. We address this with UA-CFIF (Uncertainty-Aware Cross-Modal Feature Interaction Fusion), which fuses face and full hand images while explicitly modelling the interaction between them. Its Cross-Modal Interaction Module builds four views of the embeddings — the face, the hand, their difference, and their product — and a gated attention layer decides how much each should count. A dropout-based step then gives a sense of how sure the model is about each prediction. Where earlier MULBv1 studies cropped the hand to the palm, we keep the whole hand, fingers included. Across 528 test pairs from 176 people, UA-CFIF reaches 99.62% Rank-1 accuracy, 1.0000 AUC, and 0.15% EER, and unlike plain concatenation, it gives a clearer signal when it might be wrong.

Keywords: —Multimodal Biometrics, Deep Learning, Feature Fusion, Attention Mechanism, Uncertainty Estimation, Convolutional Neural Networks.

1. INTRODUCTION

Biometrics is a critical part of today's authentication systems, since biological and behavioral traits offer reliable and convenient identification [1]. Among the most popular biometric traits are face, iris, fingerprint, palmprint, and hand geometry [2], all useful but sensitive to acquisition conditions [1], [3]. This has led to the rise of multimodal biometric systems, which combine complementary information from several traits to give more accurate and robust recognition, with better resistance to spoofing attacks, than unimodal systems [3]-[4]. In multimodal biometrics, fusion can be carried out at the sensor, feature, score, or decision levels [5], [6]. Among these, feature-level fusion can preserve more discriminative information by merging modality-specific features before classification [5], [7], [8]. Even so, most existing feature-level approaches rely on either concatenation or correlation transformations of feature vectors [7], [9].

While useful, these approaches consider modality embeddings separately and almost completely ignore the relationships between different biometric traits. Yet biometric embeddings can carry several natural relationships — feature agreement, disagreement, and complementarity — that have proven useful in other multimodal contexts [10] and that simple concatenation cannot capture. To address this drawback, this work proposes Uncertainty-Aware Cross-Modal Feature Interaction Fusion (UA-CFIF), a novel framework for feature-level fusion that takes cross-modal interactions into account without concatenating feature vectors. To do this, we introduce the Cross-Modal Interaction Module (CMIM), which combines face and full-hand embeddings through four complementary interaction types, motivated by the success of interaction-based representation learning in multimodal contexts [10]. A gated attention mechanism then learns the importance of each interaction dynamically [11], and Bayesian dropout [12] estimates how uncertain the model is about each prediction, making biometric recognition more trustworthy [13]-[14]. In addition, this study is the first to use full hand images on the MULBv1 database, rather than the cropped palm sections used in

prior works [9], [15]. Preserving the finger geometry and boundary information provides additional discriminative features that are lost when only the palm is used.

The key contributions of this paper are:

A novel Cross-Modal Interaction Module (CMIM) that models the complementary relationships between face and hand representations, in contrast to concatenation-based methods.

An adaptive gated attention mechanism that learns the importance of each interaction component.

A Bayesian dropout scheme that produces an uncertainty estimate for every prediction.

The first evaluation, to our knowledge, of full-hand images on the MULBv1 dataset, with Grad-CAM visualization used to interpret which regions contribute to identification.

2. RELATED WORK

A. Face Recognition

Hand-based biometric recognition, including palmprint identification, has been studied for over two decades and has progressed from hand-crafted texture descriptors to deep-learning approaches. Taigman et al. [16] showed that deep CNNs can approach human-level performance on the LFW benchmark. FaceNet, introduced by Schroff et al. [17], learns a compact embedding using triplet loss and made embedding-based recognition the dominant approach. Deng et al. [18] proposed ArcFace, which uses an additive angular margin loss to push samples of the same identity closer together while spreading different identities apart; it has since become a widely used baseline. Wang and Deng [19] provide a survey of deep face recognition that covers architecture choices, loss functions, and evaluation benchmarks. More recently, interest has shifted to transformer-based designs. TransFace [20] reported that vision transformers can match or exceed CNN baselines on standard face recognition benchmarks, which indicates that attention-based architectures are competitive with CNNs for this task.

B. Hand-Based Biometric Recognition

Hand-based biometric recognition, including palmprint identification, has been studied for over two decades and has progressed from hand-crafted texture descriptors to deep-learning approaches. Zhang et al. [21] proposed an online palmprint identification system that uses 2D Gabor-based texture features on low-resolution images, establishing an early benchmark on the PolyU dataset. Liu and Kumar [2] extended this line to contactless settings, showing that deep residual networks can support robust identification without a fixed acquisition device. In the deep learning era, Shao and Zheng [22] developed a deep CNN with transfer learning for palmprint recognition, demonstrating strong cross-dataset generalization. A survey by Zhong et al. [23] of palmprint recognition methods over the past decade shows that most CNN-based systems focus on the palm surface, leaving the geometric and texture information in the finger regions underutilized — a gap that this work addresses by retaining the full hand image.

C. Multimodal Biometric Fusion

Ross and Jain [5], [6] laid the groundwork for multimodal fusion strategies at the sensor, feature, score, and decision levels. Haghighat et al. [7] proposed Discriminant Correlation Analysis (DCA) for feature-level fusion, which simultaneously maximizes the pairwise correlation between modality features and the separation between identity classes. The squeeze-and-excitation (SE) mechanism of Hu et al. [11] learns to weight feature channels adaptively, and this principle informed our gated attention design. For uncertainty estimation, the survey by Gawlikowski et al. [12] categorizes uncertainty estimation methods in deep networks and finds that dropout-based approaches offer a practical balance between accuracy and calibration. Shi and Jain [24] proposed Probabilistic Face Embeddings (PFEs), which represent each face as a Gaussian distribution whose variance captures the uncertainty in the learned features; this improves recognition on challenging and ambiguous samples. Unlike PFEs, which model uncertainty at the feature level, UA-CFIF estimates uncertainty at the prediction level using dropout-based entropy. Recent work on multimodal hand biometrics has also explored correlation-based fusion. Yang et al. [25] proposed a fusion network that enhances correlations between hand modalities and reports accuracy gains on multimodal hand recognition; our approach differs in that it operates between heterogeneous modalities (face and hand) and learns the importance of each interaction type through gated attention.

D. Prior Work on MULBv1

The MULBv1 dataset was introduced by Kadhim and Abdulameer [26], who also presented a face recognition case study reaching 97.41% accuracy. Later work by the same group combined face, palm, and iris features and achieved 99.88% accuracy [15]. Separately, a face-iris system using score-level fusion reached 100% Rank-1 accuracy on the same dataset [27]. A feature-level fusion system combining face and cropped palmprint using CNN features with an SVM classifier reached 99.28% [9]. All of these systems either cropped the palm region or did not use any hand image, and none of them provided uncertainty estimation or adaptive attention-based fusion. To our knowledge, UA-CFIF is the first system on MULBv1 to bring full hand images, uncertainty estimation, and adaptive attention-based fusion together in a single framework.

3. PROPOSED METHODOLOGY

A. General Phases of the Multimodal Biometric Framework

UA-CFIF consists of five phases that take a face-hand pair from raw images to a prediction with an associated uncertainty score: image preprocessing, deep feature extraction, cross-modal interaction construction (CMIM), attention-based weighting, and classification with uncertainty estimation. Figure 1 shows the overall workflow. Each phase is described in the following sections, with special focus on the Cross-Modal Interaction Module and the gated attention mechanism that form the core of UA-CFIF.

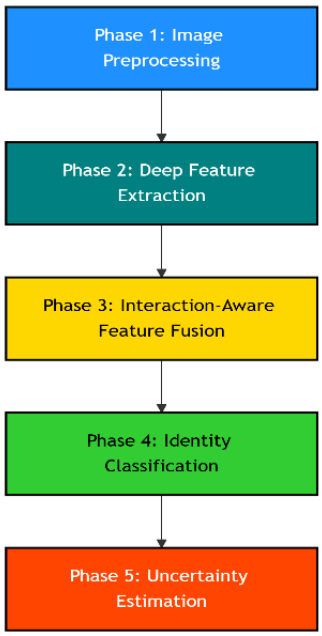

Fig 1. General phases of the proposed multimodal biometric framework.

B. The Proposed Uncertainty-Aware Cross-Modal Feature Interaction Fusion (UA-CFIF)

UA-CFIF is designed to improve multimodal biometric identification by explicitly modeling the relationships between face and hand features prior to fusion. Unlike conventional feature-level fusion methods that simply concatenate modality-specific feature vectors, UA-CFIF constructs a relationship-aware representation that captures complementary information between the two modalities. The overall architecture is illustrated in Figure 2. As outlined in Section III-A, the framework consists of five stages: image preprocessing, deep feature extraction, cross-modal interaction construction, attention-based weighting, and classification with uncertainty estimation.

Face and hand images are first preprocessed and encoded independently using two parallel CNN feature extractors (one for each modality) that produce 256-dimensional feature embeddings. Instead of concatenating them, the Cross-Modal Interaction Module (CMIM) builds an interaction representation by combining the face embedding, the hand embedding, their absolute difference, and their Hadamard product. These four representations capture complementary relationships that conventional concatenation cannot preserve. They are then refined by a gated

attention mechanism, which adaptively learns the importance of each view for every sample and produces a fused representation for identity classification. Finally, a dropout-based uncertainty estimate inspired by Bayesian approximation [13] provides an uncertainty value alongside each prediction, which can be used downstream to flag low-confidence cases for review.

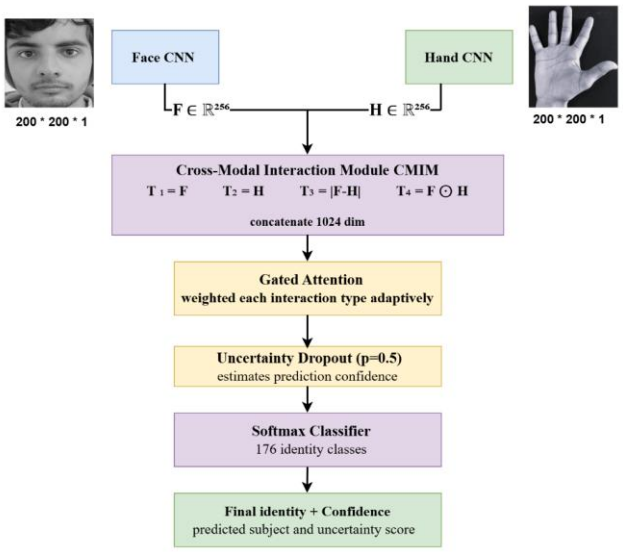

Fig 2. Overview of the proposed UA-CFIF framework.

C. Cross-Modal Interaction Module (CMIM)

The CMIM computes four relationships between F and H. The design is inspired by natural language inference research [10], where combining sentence embedding differences and products was shown to improve multimodal inference. We adopt the same scheme here, interpreting the Hadamard product as capturing shared identity evidence and the absolute difference as highlighting modality-specific patterns. We apply this four-way interaction scheme to the face and hand embeddings:

$T_1 = F$ — the original face embedding.

$T_2 = H$ — the original hand embedding.

$T_3 = |F - H|$ — the absolute difference, highlighting dimensions where the two modalities diverge.

$T_4 = F \odot H$ — the Hadamard product, capturing the element-wise co-activation of face and hand features.

These four vectors are combined into a 1024-dimensional interaction representation Z:

$$Z = [T_1 ; T_2 ; T_3 ; T_4] \in \mathbb{R}^{1024} \qquad (1)$$

D. Gated Attention Mechanism

A gated attention layer learns to up-weight informative dimensions of Z and down-weight less relevant ones, inspired by the channel-wise recalibration principle of squeeze-and-excitation networks [11], though applied here to dimension-wise weighting of the interaction representation rather than CNN feature maps. A fully connected layer with sigmoid activation maps $Z \in \mathbb{R}^{1024}$ to a weight vector $A \in [0, 1]^{1024}$:

$$A = \sigma(W_a \cdot Z + b_a) \in [0, 1]^{1024} \qquad (2)$$

where $W_a \in \mathbb{R}^{1024 \times 1024}$ is a learned weight matrix and $b_a \in \mathbb{R}^{1024}$ is a bias vector. The gated output G is then:

$$G = Z \odot A \in \mathbb{R}^{1024} \qquad (3)$$

Because A depends on Z directly, the system learns a sample-specific weighting strategy — assigning different importance to each interaction view depending on the particular face-hand pair being identified.

E. Uncertainty Estimation

Following Gal and Ghahramani [13], dropout with rate p = 0.5 is applied to G during training, which can be interpreted as approximate Bayesian inference [13], [28]. The resulting softmax output p = [p_1, ..., p_C] is then used to estimate prediction uncertainty via normalized Shannon entropy:

$$H(x) = -(1 / \log C) \cdot \Sigma_c \, p_c \cdot \log(p_c + \varepsilon) \qquad (4)$$

where C = 176 is the number of identity classes, p_c is the softmax probability for class c, and $\varepsilon = 10^{-12}$ ensures numerical stability. Values near 0 indicate high confidence, and values near 1 indicate maximum uncertainty.

F. Classification Head

The dropout output passes through a Dense(176) layer with softmax activation. The model is trained with categorical cross-entropy loss; full training settings are detailed in Section IV-B. Table I summarizes the complete model architecture.

Table 1. UA-CFIF Model Architecture Summary

Layer	**Output Shape**	**Parameters**	**Notes**
Face_Input / Hand_Input	(None, 200, 200, 1)	0	Grayscale, normalized; hand input uses full hand without cropping
Conv2D(32) ×2	(None, 200, 200, 32)	320 × 2	3×3, ReLU, same padding
MaxPool ×2	(None, 100, 100, 32)	0	2×2
Conv2D(64) ×2	(None, 100, 100, 64)	18,496 × 2	3×3, ReLU, same padding
MaxPool ×2	(None, 50, 50, 64)	0	2×2
Flatten ×2	(None, 160,000)	0	50×50×64
Dense(256) ×2	(None, 256)	40,960,256 × 2	ReLU embedding
$T_1 = F$	(None, 256)	0	Face embedding (direct)
$T_2 = H$	(None, 256)	0	Hand embedding (direct)
$T_3 = \lvert F - H \rvert$	(None, 256)	0	Absolute difference
$T_4 = F \odot H$	(None, 256)	0	Hadamard product
CMIM Concat $[T_1;T_2;T_3;T_4]$	(None, 1024)	0	4 × 256
Attention Gate Dense(1024)	(None, 1024)	1,049,600	Sigmoid
Gated Fusion	(None, 1024)	0	$Z \odot A$
Uncertainty Dropout (p=0.5)	(None, 1024)	0	Bayesian approximation
Classifier Dense(176)	(None, 176)	180,400	Softmax

Total Parameters	—	83,188,144	~317 MB

Note: Parameters are reported per branch where applicable; ×2 denotes identical parallel branches for face and hand inputs.

4. EXPERIMENTAL SETUP

A. Dataset

All experiments were run on MULBv1 [26], prepared by Kadhim and Abdulameer and publicly available online. The dataset contains face, hand, and iris images of 176 individuals, collected under controlled indoor conditions with an iPhone 14 Pro Max camera [26]. In this study, only the face and hand modalities are used. The dataset is pre-divided by its authors into training, validation, and test splits. Because face and hand images are stored separately, face-hand pairs were formed for each individual by pairing images up to the minimum count available across the two modalities. This yielded 2,465 pairs for training, 353 for validation, and 528 for testing (exactly 3 per individual across 176 individuals). Unlike prior work on MULBv1 [9], [15], [27], palm cropping is not applied, and the full hand image is used as captured. Figure 3 shows sample face-hand pairs from five individuals.

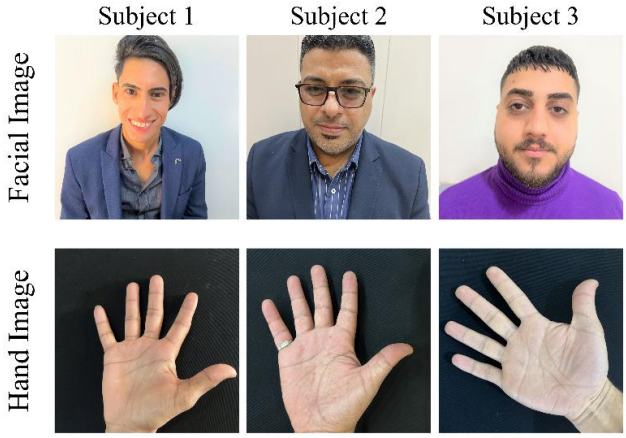

Fig 3. Selected face and full-hand images for five different individuals from the MULBv1 database. The full-hand images contain all finger regions, with no palmprint cropping.

B. Training Protocol

The model was implemented in TensorFlow 2.20 and trained on an NVIDIA Tesla T4 GPU via Google Colab. We used the Adam optimizer [29] with a learning rate of 0.001 and a batch size of 16. Training ran for a maximum of 30 epochs and stopped early at epoch 16 through EarlyStopping (patience = 10 epochs). The best model weights, saved at epoch 6 with a validation accuracy of 98.87%, were used for all evaluations. No data augmentation was applied. The validation set contains 353 pairs (approximately 2 per subject on average), which is smaller than ideal for robust early stopping decisions; however, the test set results are fully independent and unaffected by this limitation.

The Uncertainty_Dropout layer uses a dropout rate of 0.5, following the value recommended in the original dropout formulation [30]. The model is trained with categorical cross-entropy loss, and all weights are initialized using the Glorot uniform scheme [31].

C. Concatenation Baseline

To validate the contribution of the CMIM, the attention mechanism, and the uncertainty module, we trained a Concatenation Baseline using an identical training setup. This baseline uses the same CNN backbone as UA-CFIF, with the same architecture, data, optimizer, and training protocol, but replaces the CMIM, attention, and uncertainty dropout with a simple concatenation of F and H into a 512-dimensional vector, followed directly by the classifier. The baseline includes no uncertainty dropout; its prediction uncertainty is instead estimated from the softmax output entropy for a fair comparison. With this design, any difference in performance reflects the fusion strategy, including the additional representational capacity of the interaction module.

D. Evaluation Metrics

The following metrics were used:

- Rank-k Accuracy: in the closed-set identification setting, the proportion of test samples for which the correct identity appears within the top-k predicted classes (k = 1, 5, 10, 20).
- Macro-averaged Precision, Recall, and F1-Score across all 176 identity classes, with equal weight given to each identity.
- AUC: area under the ROC curve in a one-vs-rest micro-averaged scheme, where each of the 528 test samples is evaluated against 175 impostor comparisons.
- EER: the operating point at which the false acceptance rate equals the false rejection rate, derived from the ROC curve of the one-vs-rest evaluation.
- McNemar's test [32]: to assess whether the difference in error rates between UA-CFIF and the Concatenation baseline is statistically significant.
- Uncertainty comparison: mean prediction uncertainty for correctly classified vs. misclassified samples in both models, where uncertainty is estimated via dropout-based entropy in UA-CFIF and via softmax entropy alone in the Concatenation baseline.

E. Ablation Study

To understand the contribution of each component, six configurations were evaluated in increasing order of architectural complexity, from unimodal baselines to the full system. These configurations were evaluated by selectively enabling or disabling components of the trained UA-CFIF model, rather than being independently retrained from scratch; the ablation results therefore reflect relative trends rather than fully independent comparisons. The independently trained Concatenation baseline (Section IV-C) provides a controlled comparison between UA-CFIF and simple concatenation under identical training conditions.

It should be noted that the 'Concatenation' entry in the ablation table shares the same training run as the other variants and is therefore distinct from the independently retrained Concatenation Baseline in Section IV-C (Table IV).

- Face Only: the face CNN branch evaluated independently (hand input masked); direct classification from the face embedding alone.
- Hand Only: the hand CNN branch evaluated independently (face input masked); direct classification from the hand embedding alone.
- Concatenation: simple concatenation of face and hand embeddings with no CMIM or attention.
- CFIF (Equal Weights): all four CMIM interactions concatenated with uniform weighting and no attention gating.
- CFIF + Attention: CMIM with gated attention, without the Uncertainty_Dropout layer.

UA-CFIF (Full): the complete proposed system.

5. RESULTS AND DISCUSSION

A. Training Convergence

Figure 4 shows the training and validation accuracy and loss curves. Training stopped at epoch 16 via EarlyStopping. The model converges quickly: validation accuracy reaches 66.01% at epoch 1, jumps to 96.60% at epoch 2, reaches 98.58% at epoch 3, and achieves its peak value of 98.87% at epoch 6. From epoch 4 onward, training accuracy approaches 100%, while validation accuracy fluctuates between 96% and 99%. This fluctuation is typical of a small validation set. The moderate gap between training and validation accuracy is consistent with dropout regularization, and no divergence is observed.

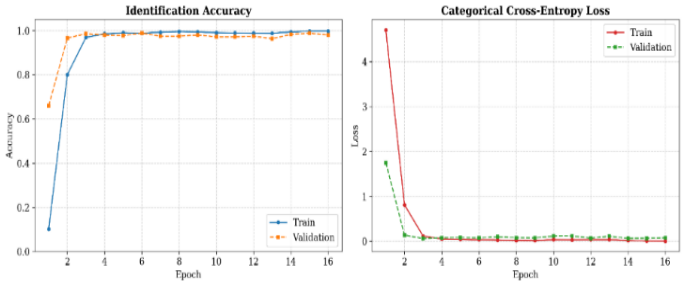

Fig 4. Training and validation accuracy (left) and loss (right) curves of the proposed UA-CFIF over 16 epochs. The curves show stable convergence without overfitting.

B. Identification Performance

Table II shows that UA-CFIF achieves 99.62% Rank-1 accuracy (only 2 errors) and a macro F1-score of 99.61%, which indicates consistent performance across all 176 identities. Rank-5 and Rank-10 remain at 99.62%, while Rank-20 rises to 99.81% (527/528). At Rank-20, one of the two errors recovers, which means the correct identity appears within the top-20 for that sample but not at Rank-1. The AUC of 1.0000 in a micro-averaged one-vs-rest scheme, where 528 genuine comparisons are evaluated against 92,400 impostor comparisons (528 × 175), shows near-perfect separation between the genuine and impostor score distributions. There is no data leakage between the training, validation, and test sets, and the non-zero EER of 0.15% confirms that the model is highly discriminant while ruling out overfitting as the cause of the high AUC.

Table 2. UA-CFIF Performance on the MULBv1 Test Set (528 pairs, 176 subjects, 3 per subject)

Metric	**Value**	**Notes**
Rank-1 Accuracy	99.62%	526 / 528 correct
Rank-5 Accuracy	99.62%	526 / 528 in top-5
Rank-10 Accuracy	99.62%	526 / 528 in top-10
Rank-20 Accuracy	99.81%	527 / 528 in top-20
Macro Precision	99.72%	Average over 176 classes
Macro Recall	99.62%	Average over 176 classes
Macro F1-Score	99.61%	Harmonic mean
AUC (micro-average)	1.0000	528 genuine vs. 92,400 impostor comparisons
EER	0.15%	FAR = FRR = 0.15%

C. ROC and CMC Curves

Figure 5 shows the ROC curve (left) and the CMC curve (right). The ROC curve rises almost vertically from the origin and reaches the top-left corner, which confirms near-perfect separation between genuine and impostor scores with AUC = 1.0000. The CMC curve shows that the Rank-1 recognition rate of 99.62% remains constant through Rank-19 and rises to 99.81% at Rank-20. For one of the two misclassified samples, the correct identity appears within the top-20 candidates, while the other error persists beyond Rank-20. The corresponding EER of 0.15% (Table II) indicates that, at the EER operating point, UA-CFIF has equally low false acceptance and false rejection rates, which supports its use as a biometric verifier in high-security settings.

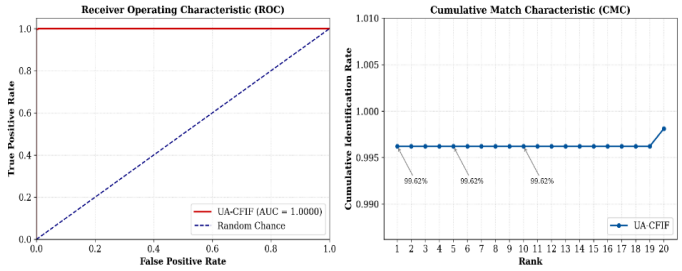

Fig 5. ROC and CMC curves for UA-CFIF.

D. *Genuine vs. Impostor Score Distribution*

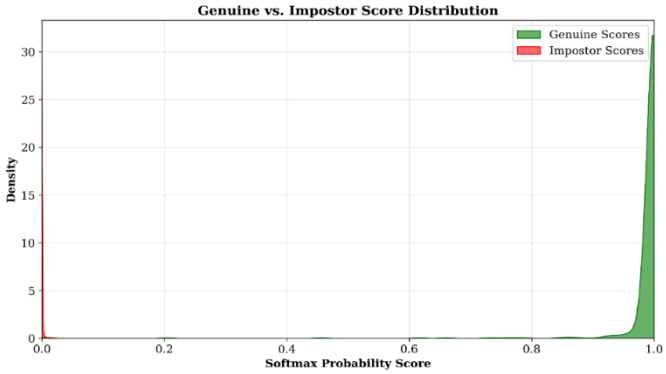

Figure 6 shows kernel density estimates of genuine scores (the softmax probability assigned to the correct class) and impostor scores (the softmax probabilities assigned to all 175 incorrect classes per test sample, for 92,400 impostor scores in total). The two distributions are almost completely separated: most genuine scores cluster near 1.0, while impostor scores concentrate near 0.0, with only minimal overlap. This separation allows threshold-based verification with minimal tuning. The EER of 0.15% (Table II) confirms that the observed AUC of 1.0000 reflects genuine score separation rather than data leakage or class imbalance artifacts.

Kernel density distributions of genuine and impostor softmax probabilities. Genuine scores concentrate around 1.0 and impostor scores around 0.0, showing almost perfect separation.

E. *Ablation Study*

Ablation results are listed in Table III. The face-only baseline attains 99.24% Rank-1 accuracy, while the hand-only configuration reaches 98.67%. This confirms the stronger discriminative power of the face modality on its own. In the ablation sequence, where all configurations share the same training run as UA-CFIF, simple concatenation of F and H achieves 99.81% Rank-1. Adding CMIM with equal weights leaves Rank-1 accuracy unchanged at 99.81%, which shows that the interaction terms do not by themselves alter the decision boundary. Introducing gated attention changes the accuracy to 99.43% under the shared-weight evaluation, a variation attributed to the shared-weight setting rather than a true degradation. The full UA-CFIF model, which adds uncertainty-aware dropout, achieves 99.62% accuracy.

Within the ablation sequence, concatenation achieves 99.81% Rank-1 accuracy compared to 99.62% for UA-CFIF — a difference of one sample that is not statistically significant (McNemar $p = 0.625$). UA-CFIF is designed for better-calibrated uncertainty alongside competitive accuracy, not for marginal accuracy gains. Under independently controlled training conditions (Table IV), UA-CFIF outperforms the Concatenation baseline in both Rank-1 accuracy (99.62% vs. 99.24%) and uncertainty calibration: misclassified samples receive substantially higher uncertainty scores in UA-CFIF than in the baseline (0.5188 vs. 0.2835).

Table 3. Ablation Study Results

Configuration	**CMIM**	**Attn**	**Unc. Est.**	**Rank-1 (%)**	**F1**
Face Only	No	No	No	99.24	0.9921

Hand Only	No	No	No	98.67	0.9864
Concatenation	No	No	No	99.81	0.9980
CFIF (Equal Weights)	Yes	No	No	99.81	0.9979
CFIF + Attention	Yes	Yes	No	99.43	0.9941
UA-CFIF (Proposed)	Yes	Yes	Yes	99.62	0.9961

F. Comparison with Concatenation Baseline

Table IV compares UA-CFIF directly with the Concatenation baseline trained under identical conditions. UA-CFIF achieves higher Rank-1 accuracy (99.62% vs. 99.24%) and a higher macro F1-score (99.61% vs. 99.22%), with only 2 misclassifications compared to 4 in the Concatenation baseline. The baseline marginally outperforms UA-CFIF at Rank-5 (99.81% vs. 99.62%): the baseline recovers 3 of its 4 Rank-1 errors within the top-5 candidates, whereas UA-CFIF's 2 errors do not appear within the top-5 — suggesting that the two misclassified samples in UA-CFIF are harder cases where the correct identity ranks lower in the score list.

Table 4. UA-CFIF vs. Concatenation Baseline (same backbone, same training conditions)

Metric	**UA-CFIF**	**Concatenation Baseline**
Rank-1 Accuracy	99.62%	99.24%
Rank-5 Accuracy	99.62%	99.81%
Macro F1-Score	99.61%	99.22%
Misclassifications	2 / 528	4 / 528
McNemar p-value	0.625 (between UA-CFIF and Concatenation Baseline)	
Mean unc. (correct)	0.0070	0.0076
Mean unc. (errors)	0.5188	0.2835
Error percentile rank (uncertainty)	0.4%	1.0%

McNemar's test gives $p = 0.625$, so the raw accuracy difference is not statistically significant on a 528-sample test set. The more telling difference lies in uncertainty calibration. In UA-CFIF, misclassified samples have a mean uncertainty score of 0.5188, placing them in the top 0.4% of the uncertainty distribution. In the Concatenation baseline, misclassified samples have a mean uncertainty of only 0.2835, placing them in the top 1.0% — a much weaker separation from the correctly classified samples. This means that UA-CFIF provides a more reliable confidence signal: its errors are more clearly flagged as uncertain, suggesting the system could be more safely deployed in practice. Figure 7 visualizes this difference.

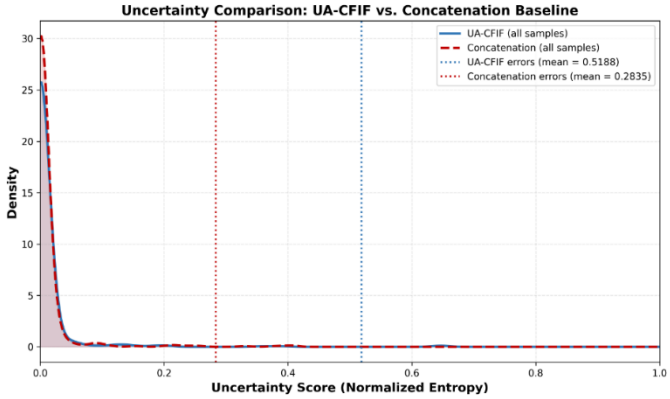

Fig 7. Comparison of the uncertainty values of UA-CFIF with those of the baseline concatenation model, which exhibits greater uncertainty for misclassified samples in UA-CFIF.

G. *Attention Weight Analysis*

Figure 8 illustrates the mean attention weights of the four interaction views in CMIM. If all four were equally important, their weights would be close to 0.5. In fact, T_3 (absolute difference) and T_2 (hand features) receive the highest attention weights, while T_1 (face features) has an intermediate value and T_4 (Hadamard product) receives the lowest weight (0.4976), approaching the uniform baseline of 0.5. Even though face features are more discriminative as a standalone modality (99.24% vs. 98.67% for hand-only, Table III), the greater weight assigned to T_2 suggests that hand features contribute more additional information to the fused representation — an observation consistent with the complementary nature of the two modalities.

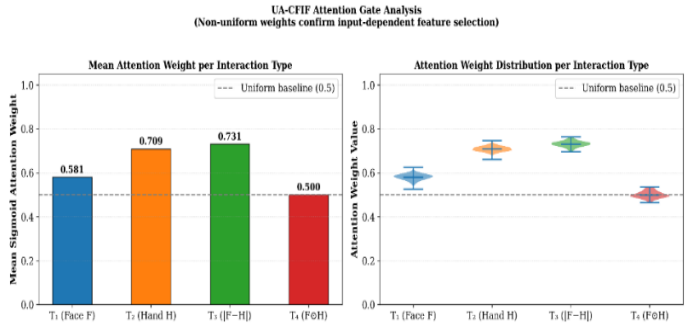

Fig 8. Mean gated attention weights for CMIM interaction types (T_1-T_4). The dashed line represents the uniform baseline value (0.5).

H. *Uncertainty and Error Analysis*

Figure 9 shows the uncertainty score distributions for correctly classified (n = 526) and misclassified (n = 2) samples. The correctly classified samples have a mean uncertainty of 0.0070, indicating that the model is highly confident for the vast majority of predictions. The two errors have uncertainty scores of 0.39 and 0.65, placing them within the top 0.6% and top 0.2% of the distribution respectively. If a threshold of $H(x) > 0.3$ were used to flag predictions for human review, both errors would be caught while only a small fraction of correctly classified predictions would also be flagged. This suggests that the uncertainty score could serve as a practical signal for flagging predictions during deployment.

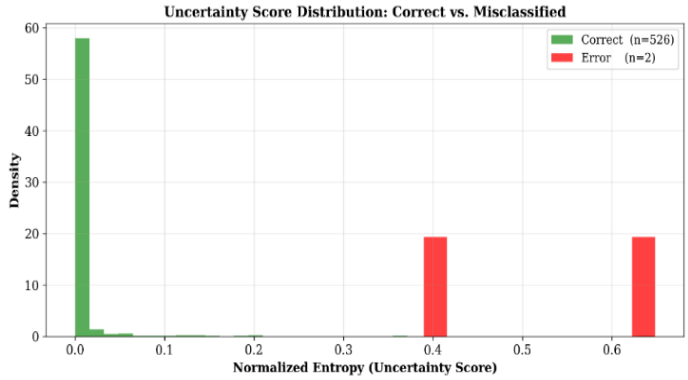

Fig 9. Uncertainty score distributions of the correctly classified (n = 526) and misclassified (n = 2) samples.

I. Interpretability: Grad-CAM

Grad-CAM [33] visualization heatmaps of the last convolutional layer of each CNN branch are presented in Figure 10. The attention map for faces concentrates on discriminative areas, including the eyes, nose bridge, and cheekbones. For the hand modality, the activated areas consist of finger edges and boundary contours that are typically excluded by palm-cropping approaches. This demonstrates the contribution of the finger region to identity-related information, supporting the motivation outlined in Section II-B: prior palmprint-based systems that crop the hand to the palm region may discard identity-relevant information located in the fingers.

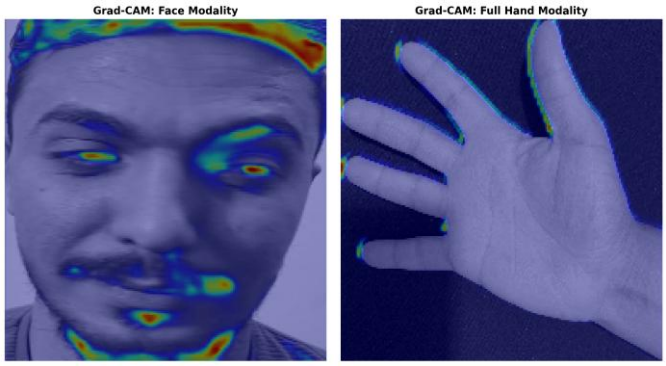

Fig 10. Grad-CAM heat maps of the last convolutional layer of each CNN branch.

J. Per-Subject Performance Analysis

Figure 11 illustrates per-subject Rank-1 accuracy across all 176 subjects. 174 of the 176 subjects (98.9%) achieve perfect accuracy (3/3 samples correct), while the remaining 2 subjects achieve 67% (2/3 samples correct). No subject has 0% accuracy, indicating that the model does not fail completely on any individual.

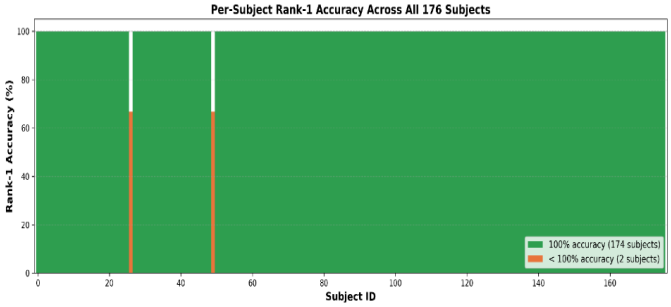

Fig 11. Accuracy of Rank-1 per each individual among all 176 subjects. 174 subjects get perfect 100% accuracy; there are just 2 individuals that result in wrong classification.

K. Computational Efficiency

Table V reports the computational cost and inference time for UA-CFIF. All GPU measurements use a tf.function-compiled prediction step with 10 warm-up runs and 100 timed runs per batch size. The CPU measurement uses the same protocol on the Colab CPU and provides a proxy for a resource-constrained deployment environment.

Table 5. Computational Efficiency of UA-CFIF

Metric	**Value**	**Notes**
GFLOPs per image pair	0.950	Computed from layer dimensions
Total Parameters	83.19M	float32
Parameter Memory	317.34 MB	83.19M × 4 bytes
Peak Inference Memory	0.75 MB	Per sample (batch=1)
GPU Latency (batch=1, Tesla T4)	1.56 ms	Single-query mode
GPU Latency (batch=32, Tesla T4)	0.33 ms	Optimal throughput
GPU Throughput (batch=32)	2,993 pairs/s	NVIDIA Tesla T4
CPU Latency (simulated edge)	122.26 ± 50.87 ms	1 pair, Colab CPU
GPU vs CPU Speedup	78.4x	batch=1 comparison

UA-CFIF requires 0.950 GFLOPs per image pair and consumes only 0.75 MB of peak inference memory per sample, despite its 83.19M parameter count. A high parameter count with a low FLOP count is not a contradiction: roughly 98% of the parameters sit in the two Dense(256) embedding layers, which cost far fewer FLOPs per parameter than convolutional layers. Most of the model's memory is in weights that are cheap to evaluate. The throughput numbers in Table V show the deployment picture: under 2 ms per query on a single GPU, and nearly two orders of magnitude slower on CPU (78.4×). GPU-based server deployment is the natural target. Adapting the model for edge devices through quantization or distillation is left as future work.

L. Comparison with Prior Work

Table VI compares UA-CFIF with previous works on MULBv1. This comparison is not fully controlled, because the compared approaches use different modality combinations. In particular, iris-based approaches rely on a highly discriminative modality but require specialized near-infrared acquisition equipment. UA-CFIF therefore does not claim superiority over iris-based approaches in raw discriminative power.

Table 6. Comparison with Prior Work on MULBv1

System	**Modalities**	**Hand Region**	**Acc. (%)**	**EER (%)**	**Uncertainty**	**Interpretable**
Kadhim et al. [26]	Face only	—	97.41	—	No	No
Kadhim et al. [9]	Face + Palm	Cropped	99.28	—	No	No
Kadhim et al. [15]	Face+Palm+Iris	Cropped	99.88	1.86	No	No
Kadhim et al. [27]	Face + Iris	—	100.00	0.26	No	No

UA-CFIF (Proposed)	Face + Full Hand	Full	99.62	0.15	Yes	Yes

Among the systems that report EER, UA-CFIF achieves the lowest value (0.15%), with a comparable Rank-1 accuracy of 99.62%. It is also the only MULBv1 system that combines complete hand images, uncertainty estimation, and attention-based interpretability. The face-iris approach [27] reaches 100% accuracy, but iris is generally regarded as a more discriminative modality than hand geometry, and its acquisition requires specialized near-infrared sensors that UA-CFIF does not need. This comparison therefore reflects a trade-off between modality discriminability and acquisition cost rather than a direct measure of algorithmic superiority. The main contribution of UA-CFIF is not accuracy alone. It is the combination of competitive accuracy with uncertainty estimation and interpretability.

6. CONCLUSION

This paper proposed a multimodal biometric system, UA-CFIF, that combines face and full-hand images on the MULBv1 dataset. Three main findings emerge from this work: First, CMIM represents four cross-modal interactions, where the absolute difference (T_3) receives the highest attention weight (0.7310), meaning that inter-modal discrepancy carries substantial identity-related information. Second, dropout-based uncertainty estimation provides a confidence signal that statistically distinguishes misclassified samples: errors are mostly concentrated in the top 0.4% of the uncertainty distribution (mean error uncertainty 0.5188), compared to the top 1.0% for the independently trained Concatenation baseline (mean error uncertainty 0.2835). Finally, this paper is the first to evaluate full-hand images as an input modality for MULBv1, and Grad-CAM confirms that finger-edge regions are informative for identity representation.

Overall, the proposed system achieves 99.62% Rank-1 accuracy, a 99.61% macro F1-score, an AUC of 1.0000, and an EER of 0.15% on 528 test pairs from 176 individuals, at a cost of 0.950 GFLOPs per pair, outperforming the independently trained Concatenation baseline in both accuracy and uncertainty calibration. The evaluation is carried out under a closed-set setting on a single dataset, and the results should not be assumed to generalize to open-set recognition or real-world deployment scenarios, such as cross-session or cross-environment variation.

We acknowledge that the ablation configurations in Table III share weights with the trained UA-CFIF model rather than being retrained independently, so these results should be read as relative trends rather than fully controlled comparisons. The one independently trained comparison available, UA-CFIF versus the Concatenation baseline under identical training conditions (Table IV), confirms the core finding. UA-CFIF achieves higher accuracy and substantially better-calibrated uncertainty on its errors (mean uncertainty over misclassified samples 0.5188 versus 0.2835; a well-calibrated model should be more uncertain on the samples it misclassifies), and the accuracy difference is supported by a non-significant McNemar test ($p = 0.625$) that rules out the alternative explanation that the baseline is simply better. Retraining each ablation configuration independently is one of the directions we plan to pursue next.

Future work will explore a lightweight backbone for edge deployment, open-set recognition, multi-session robustness, and improved hand feature extraction.

Future Work

Despite the promising results achieved by UA-CFIF, several directions remain to be explored to further improve its applicability and robustness.

- the generalization of the proposed framework should be evaluated on larger and more diverse multimodal databases, ideally containing cross-session and cross-environment variations. Open-set identification, where the system encounters unseen identities, is another critical scenario that has not been addressed in the current closed-set evaluation.
- to facilitate deployment on resource-constrained devices, future work will investigate lightweight backbone architectures, such as MobileNet or EfficientNet, along with model compression techniques including quantization and knowledge distillation. These efforts aim to reduce the inference latency and memory footprint while preserving competitive accuracy.
- the hand feature extraction could be enhanced by incorporating vision transformers (e.g., ViT or Swin Transformer) or by introducing finger segmentation and keypoint detection to explicitly exploit the geometric information present in finger edges, as highlighted by Grad-CAM visualizations.

- uncertainty estimation can be further improved using Monte Carlo Dropout at test time, deep ensembles, or Bayesian neural networks. The uncertainty score could also be integrated into a rejection mechanism, where low-confidence predictions are flagged for manual inspection in high-security applications.
- the gated attention mechanism within the Cross-Modal Interaction Module could be extended to a bidirectional or multi-head design, enabling more dynamic and fine-grained modeling of cross-modal dependencies.
- to provide a more rigorous evaluation of individual components, future ablation studies should retrain each configuration independently rather than sharing weights with the full model, as done in the current work. This would yield more reliable insights into the contribution of each architectural module.

Declaration on Generative AI in the Writing Process

During the preparation of this manuscript, the authors used Claude (Anthropic) to assist with language editing, restructuring, and clarity improvements throughout the text. All scientific content, including the research design, experiments, data analysis, results, and conclusions, was conceived, conducted, and verified entirely by the authors. The authors reviewed and edited all AI-assisted output and take full responsibility for the accuracy, originality, and integrity of this publication.

References

1. S. Dargan and M. Kumar, "A comprehensive survey on the biometric recognition systems based on physiological and behavioral modalities," Expert Systems with Applications, vol. 143, article 113114, 2020, doi: 10.1016/j.eswa.2019.113114.
2. C. Liu and A. Kumar, "Contactless palmprint identification using deeply learned residual features," IEEE Transactions on Biometrics, Behavior, and Identity Science, vol. 2, no. 2, pp. 172-181, 2020, doi: 10.1109/TBIOM.2020.2967073.
3. S. Pahuja and N. Goel, "Multimodal biometric authentication: A review," AI Communications, vol. 37, no. 4, pp. 525-547, 2024, doi: 10.3233/AIC-220247.
4. A. Alshardan, A. Kumar, M. Alghamdi, M. Maashi, S. Alahmari, A. A. K. Alharbi, W. Almukadi, and Y. Alzahrani, "Multimodal biometric identification: Leveraging convolutional neural network (CNN) architectures and fusion techniques with fingerprint and finger vein data," PeerJ Computer Science, vol. 10, article e2440, 2024, doi: 10.7717/peerj-cs.2440.
5. A. Ross and A. K. Jain, "Information fusion in biometrics," Pattern Recognit. Lett., vol. 24, no. 13, pp. 2115-2125, 2003, doi: 10.1016/S0167-8655(03)00079-5.
6. A. Ross and A. K. Jain, "Multibiometric systems," Commun. ACM, vol. 47, no. 1, pp. 34-40, Jan. 2004, doi: 10.1145/962081.962102.
7. M. Haghighat, M. Abdel-Mottaleb, and W. Alhalabi, "Discriminant correlation analysis: Real-time feature level fusion for multimodal biometric recognition," IEEE Trans. Inf. Forensics Security, vol. 11, no. 9, pp. 1984-1996, Sep. 2016, doi: 10.1109/TIFS.2016.2569061.
8. R. G. Praveen and J. Alam, "Recursive joint cross-modal attention for multimodal fusion in dimensional emotion recognition," in Proc. IEEE/CVF Conf. Computer Vision and Pattern Recognition Workshops (CVPRW), 2024, pp. 4803-4813.
9. O. N. Kadhim and M. H. Abdulameer, "Hybrid multimodal biometric identification system: Integrating face and palmprint traits through feature-level fusion," in Innovations of Intelligent Informatics, Networking, and Cybersecurity (3INC 2024), Cham: Springer, 2025, pp. 165-178, doi: 10.1007/978-3-031-81065-7_13.
10. A. Conneau, D. Kiela, H. Schwenk, L. Barrault, and A. Bordes, "Supervised learning of universal sentence representations from natural language inference data," in Proc. EMNLP, Copenhagen, Denmark, 2017, pp. 670-680, doi: 10.18653/v1/D17-1070.
11. J. Hu, L. Shen, and G. Sun, "Squeeze-and-excitation networks," in Proc. IEEE CVPR, Salt Lake City, UT, USA, 2018, pp. 7132-7141, doi: 10.1109/CVPR.2018.00745.
12. J. Gawlikowski et al., "A survey of uncertainty in deep neural networks," Artificial Intelligence Review, vol. 56, no. 1, pp. 1513-1589, 2023, doi: 10.1007/s10462-023-10562-9.
13. Y. Gal and Z. Ghahramani, "Dropout as a Bayesian approximation: Representing model uncertainty in deep learning," in Proc. ICML, New York, NY, USA, 2016, pp. 1050-1059.
14. C. Guo, G. Pleiss, Y. Sun, and K. Q. Weinberger, "On calibration of modern neural networks," in Proc. 34th Int. Conf. Machine Learning (ICML), PMLR, vol. 70, 2017, pp. 1321-1330.
15. O. N. Kadhim and M. H. Abdulameer, "Biometric identification advances: Unimodal to multimodal fusion of face, palm, and iris features," Adv. Electr. Comput. Eng., vol. 24, no. 1, pp. 91-100, 2024, doi: 10.4316/AECE.2024.01010.
16. Y. Taigman, M. Yang, M. Ranzato, and L. Wolf, "DeepFace: Closing the gap to human-level performance in face verification," in Proc. IEEE CVPR, Columbus, OH, USA, 2014, pp. 1701-1708, doi: 10.1109/CVPR.2014.220.
17. F. Schroff, D. Kalenichenko, and J. Philbin, "FaceNet: A unified embedding for face recognition and clustering," in Proc. IEEE CVPR, Boston, MA, USA, 2015, pp. 815-823, doi: 10.1109/CVPR.2015.7298682.

18. J. Deng, J. Guo, N. Xue, and S. Zafeiriou, "ArcFace: Additive angular margin loss for deep face recognition," in Proc. IEEE CVPR, Long Beach, CA, USA, 2019, pp. 4690-4699, doi: 10.1109/CVPR.2019.00482.
19. M. Wang and W. Deng, "Deep face recognition: A survey," Neurocomputing, vol. 429, pp. 215-244, Mar. 2021, doi: 10.1016/j.neucom.2020.10.081.
20. J. Dan, Y. Liu, H. Xie, J. Deng, H. Xie, X. Xie, and B. Sun, "TransFace: Calibrating transformer training for face recognition from a data-centric perspective," in Proc. IEEE/CVF Int. Conf. Computer Vision (ICCV), Paris, France, 2023, pp. 20642-20653.
21. D. Zhang, W. K. Kong, J. You, and M. Wong, "Online palmprint identification," IEEE Trans. Pattern Anal. Mach. Intell., vol. 25, no. 9, pp. 1041-1050, Sep. 2003, doi: 10.1109/TPAMI.2003.1227981.
22. H. Shao and B. Zheng, "Deep convolutional neural network-based palmprint recognition," IEEE Access, vol. 7, pp. 174090-174099, 2019, doi: 10.1109/ACCESS.2019.2956313.
23. D. Zhong, X. Du, and K. Zhong, "Decade progress of palmprint recognition: A brief survey," Neurocomputing, vol. 328, pp. 16-28, 2019, doi: 10.1016/j.neucom.2018.03.081.
24. Y. Shi and A. K. Jain, "Probabilistic face embeddings," in Proc. IEEE/CVF ICCV, Seoul, South Korea, 2019, pp. 6902-6911, doi: 10.1109/ICCV.2019.00700.
25. W. Yang, J. Huang, J. Luo, and W. Kang, "A hand features based fusion recognition network with enhancing multi-modal correlation," Computer Modeling in Engineering and Sciences, vol. 140, no. 1, pp. 537-555, 2024, doi: 10.32604/cmes.2024.049174.
26. O. N. Kadhim and M. H. Abdulameer, "A multimodal biometric database and case study for face recognition based deep learning," Bull. Electr. Eng. Inform., vol. 13, no. 1, 2024, doi: 10.11591/eei.v13i1.6605.
27. O. N. Kadhim, M. H. Abdulameer, and Y. M. H. Al-Mayali, "A multimodal biometric system for iris and face traits based on hybrid approaches and score level fusion," BIO Web Conf., vol. 97, article 00016, 2024, doi: 10.1051/bioconf/20249700016.
28. A. Kendall and Y. Gal, "What uncertainties do we need in Bayesian deep learning for computer vision?" in Proc. NeurIPS, Long Beach, CA, USA, 2017, pp. 5574-5584.
29. D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," in Proc. ICLR, San Diego, CA, USA, May 2015.
30. N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov, "Dropout: A simple way to prevent neural networks from overfitting," J. Mach. Learn. Res., vol. 15, no. 56, pp. 1929-1958, 2014.
31. X. Glorot and Y. Bengio, "Understanding the difficulty of training deep feedforward neural networks," in Proc. 13th Int. Conf. Artificial Intelligence and Statistics (AISTATS), vol. 9, 2010, pp. 249-256.
32. Q. McNemar, "Note on the sampling error of the difference between correlated proportions or percentages," Psychometrika, vol. 12, no. 2, pp. 153-157, 1947, doi: 10.1007/BF02295996.
33. R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, "Grad-CAM: Visual explanations from deep networks via gradient-based localization," in Proc. IEEE Int. Conf. Computer Vision (ICCV), 2017, pp. 618-626, doi: 10.1109/ICCV.2017.74.