ISSN: 3065-9817

Research Article

Journal of Clinical, Medical, and Diagnostic Research

Towards Energy-Effective Multimodal Biometric Recognition Via Information Bottleneck Fusion Spiking Neural Networks

Yan Shen[1], Xiaoxu Yang[1*], Xu Liu[2], Jiashan Wan[2] and Na Xia[2]

[1]*State Grid Electric Power Research Institute, 19 Chengxin Avenue, Jiangsu, Nanjing, China*

[2]*School of Computer Science and Information Engineering, Hefei University of Technology, Hefei, Anhui, China*

[*]Corresponding Author
Xiaoxu Yang, State Grid Electric Power Research Institute, 19 Chengxin Avenue, Jiangsu, Nanjing, China.

Submitted: 2025, May 09; **Accepted:** 2025, Jun 12; **Published:** 2025, July 15

Citation: Shen, Y., Yang, X., Liu, X., Wan, J., Xia, N. (2025). Towards Energy-Effective Multimodal Biometric Recognition Via Information Bottleneck Fusion Spiking Neural Networks. *J of Cli Med Dia Research, 3*(2), 01-14.

Abstract

With the development of multimodal biometric recognition technology, addressing substantial differences in data type, scale, resolution, and quality among biometric modalities has become one of the key challenges in ensuring enhanced security and accuracy. However, most existing techniques fail to address modality imbalance caused by disparities across modalities, leading to over-reliance on a single modality, degraded performance, and increased security vulnerabilities. Additionally, deploying traditional neural networks with full-precision floating-point representations on embedded devices is expensive and resource-intensive, further exacerbating security risks during end-to-end transmission. A new spiking neural networks multimodal biometric recognition model which incorporates two novel multimodal fusion methods is proposed. First, a spiking multimodal information bottleneck fusion approach is introduced to preserve cross-modal relevant information while utilizing sparse spiking mechanisms for efficient computation in low-power environments. Secondly, a dynamic adaptive dropout strategy is proposed to discard modality-specific spiking features during training, mitigating imbalance and enhancing multimodal representation learning. Through extensive experiments conducted on datasets such as CASIA, Iris-Fingerprint, and NUPT-FPV, the proposed model shows state-of-the-art performance, effectively addressing modality imbalance while improving security, robustness, and energy efficiency.

1. Introduction

Biometric recognition involves identifying individuals through their unique physiological and behavioral traits, allowing intelligent systems to autonomously conduct identity verification, status evaluation, and attribute prediction [1,2]. Over recent decades, single-modality biometric techniques—such as those based on face, ear, palmprint, vein structure, handwriting, gait, or voice—have been widely explored and implemented across societal and everyday applications. Nevertheless, these conventional single-modality approaches often encounter limitations, including reduced accuracy and inconsistent outcomes, stemming from incomplete data, suboptimal image quality, and similarities between classes. To overcome these challenges and ensure robust authentication, integrating data from multiple biometric sources to improve recognition effectiveness has gained increasing importance [3]. The application of Multimodal Biometric Recognition (MBR) has extended across multiple domains, including biometric interchange stations, blockchain, and federated learning, delivering marked performance improvements and practical results [4,5]. When compared to traditional single-modality biometric approaches, this technology presents the following distinctive advantages:

- Superior Recognition Accuracy [6]. MBR achieves enhanced authentication accuracy through the effective integration of multiple modality information.
- Robust Anti-forgery Defense [7]. The incorporation of multiple biometric traits significantly strengthens the system's security against fraudulent attempts, surpassing single-modality systems.

Although multimodal biometric authentication is gaining broader adoption, discrepancies among biometric modalities in MBR generate a multitude of extraneous features during the fusion process. This exacerbates a significant modality imbalance issue that undermines MBR performance. Modality imbalance describes a situation in training where specific modalities converge more rapidly, overshadowing the overall multimodal learning dynamics. As a result, the modalities with slower convergence rates are not sufficiently trained, leading to their potential being underutilized during both training and inference.

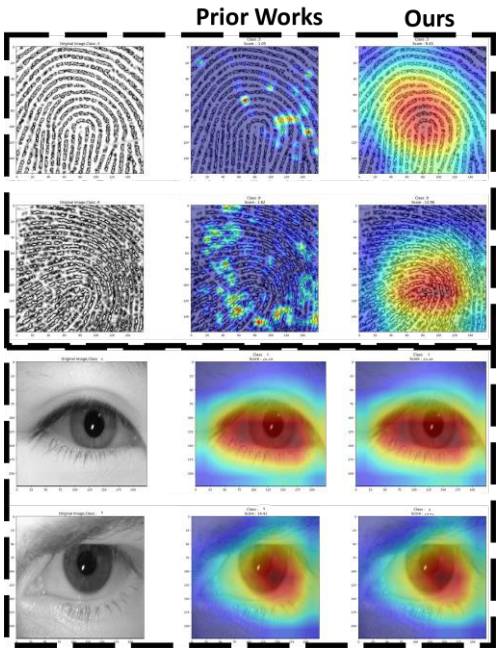

Figure 1: Illustration of the differences in prior works and our feature activation patterns. We observed that existing methods inadequately capture critical regions across input modalities, causing pronounced modal imbalance. Our proposed method effectively captures the significant activation regions across different modalities, thereby alleviating the modal imbalance problem

This phenomenon holds particular significance in multimodal biometric recognition systems, where insufficient modality learning can result in decreased modal sensitivity, subsequently compromising recognition accuracy and elevating vulnerability to modality-specific attacks, when training a multimodal model using four modalities (e.g., face, iris, fingerprint, and palmprint), the model may rely solely on facial and iris features during training. This dependency allows attackers to bypass authentication by falsifying fingerprints and palmprints, significantly increasing the risk of system fraud. We believe that the primary cause of modality imbalance in MBR is the presence of excessive redundancy across different modalities, which prevents the model from effectively extracting the most discriminative features across multiple modalities. As shown in Figure 1, when we use Grad-CAM visualization techniques to generate heatmaps visualizing the contribution of multimodal authentication modalities to each input biometric information, we observe that compared to unimodal systems, multimodal feature systems exhibit serious modal imbalance and information redundancy [8]. For example, the fingerprint feature heatmap fails to capture key discriminative information, indicating that the model heavily relies on iris features during authentication rather than fingerprint features. In addition, due to the large parameter size and slow computation of advanced MBR models, they are often deployed on cloud servers for inference, which further exacerbates security issues during data transmission and storage, such as data breaches, unauthorized access, or interception attacks during cloud-based authentication [9].

To address these challenges, we propose a multimodal data fusion method based on spiking neural networks and Information Bottleneck theory. First, we design a Spiking Multimodal Representation (SMR) aimed at converting multimodal input data into sparse binary spiking features and optimizing their representation during training. Specifically, we utilize spiking neural networks to extract spiking features from each modality and perform representation learning. Additionally, we design a Multimodal Information Bottleneck Fusion Module (MIBF), inspired by Information Bottleneck theory, which imposes variational constraints on fused features during training. This module is applied to the multimodal information fusion process to guide and reduce redundancy in multimodal fusion through Information Bottleneck theory, thereby mitigating the modality imbalance issue. Specifically, MIBF compresses the extracted modality features through a dual-objective optimization: minimizing input-fusion information overlap while maximizing fusion-output information sharing, leading to efficient multimodal integration. The goal is to retain information most relevant to prediction while eliminating redundant information unrelated to prediction across modalities. Finally, we propose a Multimodal Adaptive Dropout (MAD) method to dynamically discard features from different modalities during training. This method detects the convergence status of modalities during training and applies random dropout to features of converged and dominant modalities to reduce feature redundancy, enabling the model to thoroughly train modalities with slower convergence rates. We conducted comprehensive experiments on multiple multimodal biometric datasets (e.g., CASIA, Iris-Fingerprint, and NUPT-FPV) and compared our approach with previous works [10-12]. The experimental results demonstrate that our method effectively

mitigates modality imbalance and information redundancy, significantly reduces energy consumption, and yields more accurate and robust outcomes. Overall, the main contributions of this work are as follows:

- We designed a multimodal information bottleneck fusion method based on spiking neural networks, aimed at reducing redundant information among multiple modalities during the fusion process, thereby mitigating modality imbalance and achieving more energy-efficient and accurate multimodal biometric recognition.
- We propose an adaptive dropout method that dynamically discards features from each modality during training, with the goal of preventing rapid convergence of dominant modalities and balancing learning performance across different modalities.
- Extensive experiments demonstrate that our proposed method consistently outperforms baseline models across multiple datasets. Visualization results further confirm that our approach effectively alleviates modality imbalance in multimodal biometric fusion.

2. Related Works

2.1. Multimodal Biometric Recognition

With advancements in multimodal learning, biometric multimodal recognition has gained widespread attention and made significant progress. It primarily integrates multiple biometric features for identity authentication and feature analysis, with fusion methods classified into four types: pixel-level, feature-level, score-level, and decision-level [13-16]. As deep learning evolves, feature-level fusion has emerged as a research focus due to its strengths in feature extraction and integration.

Typical feature-level fusion methods include: Guo et al.'s NLNet, which enhances performance and efficiency through a lightweight design; Gona et al.'s transfer learning approach, combining multi-kernel bilateral filtering, deep convolutional residual networks, and GoogleNet classification to outperform traditional methods; Zhong's hand-based multimodal method, integrating deep hashing and biometric graph matching for improved accuracy and efficiency; and Abdullahi et al. sequential multimodal network, enhancing robustness via spatiotemporal feature integration [19,17-19]. Additionally, multimodal fusion plays a key role in medical applications. For instance, Lu et al. developed a hierarchical attention-based framework for Alzheimer's progression prediction, Jeong et al. applied adversarial learning to brain tumor grading with enhanced accuracy via medical imaging and biometric data, and Li et al. proposed a self-supervised progressive fusion network for reliable diagnosis and prognosis [20-22].

Despite its potential, current research focuses on designing modality-specific network structures and applying fusion mechanisms like attention or optimization-based feature selection. However, the impact of modality imbalance on multimodal biometric authentication remains underexplored, warranting further investigation.

2.2. Multimodal Information Bottleneck

The Information Bottleneck (IB) method transforms an initial state into a compact latent representation, aiming to retain the most critical information about the input state X and target state Y. This representation reduces redundancy through compression while preserving key details. The balance between compression and retention is expressed as the following optimization problem,

$$\min_{p(B|X)} I(X;B) - \beta I(B;Y), \#(1)$$

where βcontrols the trade-off. The IB framework maximizes information retention during compression and has been widely applied in multimodal learning and deep learning.

For instance, Xiao et al. proposed a hierarchical perception method based on information theory to extract multimodal information [23]. Jiang et al. introduced the Correlated Information Bottleneck method to improve the robustness of pre-trained multimodal models in Visual Question Answering [24]. Mai et al. explored IB applications across different stages of multimodal fusion [25]. The IB method also enhances model robustness. Kuang et al. used it to compress redundant information, improving resistance to adversarial attacks [26]. Nagrani et al. developed an attention bottleneck fusion method, achieving efficient multimodal data processing by filtering information [27].

Despite its potential, current research has limitations. Most works are context-specific and lack applicability to biometric multimodal feature fusion. Furthermore, the complex relationships between features from different modalities and fused features are often overlooked. Designing an IB method specifically for multimodal feature fusion is a critical direction for future research.

2.3. Imbalanced Multimodal Fusion

Modality imbalance presents a critical challenge in multimodal learning [28]. Conventional multimodal deep neural networks (DNNs) frequently struggle to surpass the performance of the best-performing unimodal models due to inherent disparities between modalities. Wang et al. demonstrated significant differences in overfitting and generalization patterns across modalities, while Peng et al. revealed that rapidly converging modalities tend to suppress the learning progression of other modalities during gradient updates [29,30]. Notably, they discovered that multimodal models may excessively rely on modality biases present in the data, substantially deviating from the desired objective of modality equilibrium [31-33]. To address this challenge, researchers have proposed various innovative approaches [31-33]. Wang et al. optimized gradient fusion by introducing additional classifiers for each modality. However, these methods inevitably increase model complexity and computational overhead. Furthermore, Xiao et al. introduced DropPathway, which employs random audio pathway dropping to balance learning rates, while Peng et al. developed the OGM-GE method, which dynamically adjusts learning rates for dominant modalities [30,34]. Despite these technical advances, existing approaches have not fundamentally activated the intrinsic

potential of slower-learning modalities, with improvements largely confined to passive regulation.

The core innovation of this paper lies in proposing an adaptive dynamic dropout mechanism tailored to different modality features. Unlike existing passive regulation methods, our approach achieves dynamic optimization during training through adaptively adjusting dropout strategies.

2.4. Spiking Neural Network

Spiking Neural Networks (SNNs) achieve discrete spike-based information transmission by replacing the traditional ReLU activation function in ANNs with spiking neurons. Recently, SNNs have demonstrated comparable performance to ANNs in single-modal image processing and classification tasks. For instance, Fang et al. proposed a deep residual spiking neural network that achieved excellent classification performance on the ImageNet dataset [35]. Furthermore, Liu et al. extended spiking networks to multi-modal audio-visual domains, achieving superior performance compared to ANN-based models [36]. In this paper, we adopt the Leaky Integrate-and-Fire (LIF) neuron as our default neuron model. In practical applications, we employ its discretized iterative form [37]:

$$U[t] = \left(1 - \frac{1}{\tau}\right) U[t-1] + C[t],$$
$$S[t] = \Psi(U[t] - V_t) = \begin{cases} 1, if\ U[t] \geq V_t \\ 0, otherwise \end{cases}, \quad \#(2)$$
$$U[t+1] = U[t] - V_t S[t],$$

The LIF neuron operates through three sequential processes: first, during the charging phase, the neuron accumulates membrane potential U[t] by integrating the input postsynaptic current $C[t]$ and membrane time constant τ; then, in the firing phase, when the membrane potential exceeds the threshold V_t as determined by the Heaviside function Ψ(·), a spike is generated; finally, in the reset phase, the membrane potential undergoes reset and decay, simulating the refractory period of biological neurons. This discrete event-driven information processing mechanism not only demonstrates biological plausibility but also significantly reduces computational overhead through sparse representation of discrete spikes. In contrast to prior works, we propose to utilize SNNs as encoders for multimodal biometric information encoding, enabling efficient spike-based sparse representations.

3. Method

As illustrated in Figure 2, our pipeline consists of three main components: (1) **Multi-modal Spiking Representation,** (2) **Multi-modal Adaptive Dropout (MAD)**, and (3) **Multi-modal Information Bottleneck Fusion (MIBF)**. First, Multi-modal Spiking Representation employs m independent spiking encoders to extract feature representations from different modalities. These spiking encoders capture the unique characteristics of each modality and transform them into discrete spike information, providing rich sparse spiking feature representations for subsequent processing while reducing computational overhead during modal fusion. Second, MAD applies adaptive dropout operations to the spiking features of individual modalities to mitigate the convergence process of the primary modality and balance the learning effects across other modalities. Finally, guided by information bottleneck theory, MIBF aims to effectively integrate spiking features from multiple modalities, focusing on preserving features most relevant to the final recognition task.

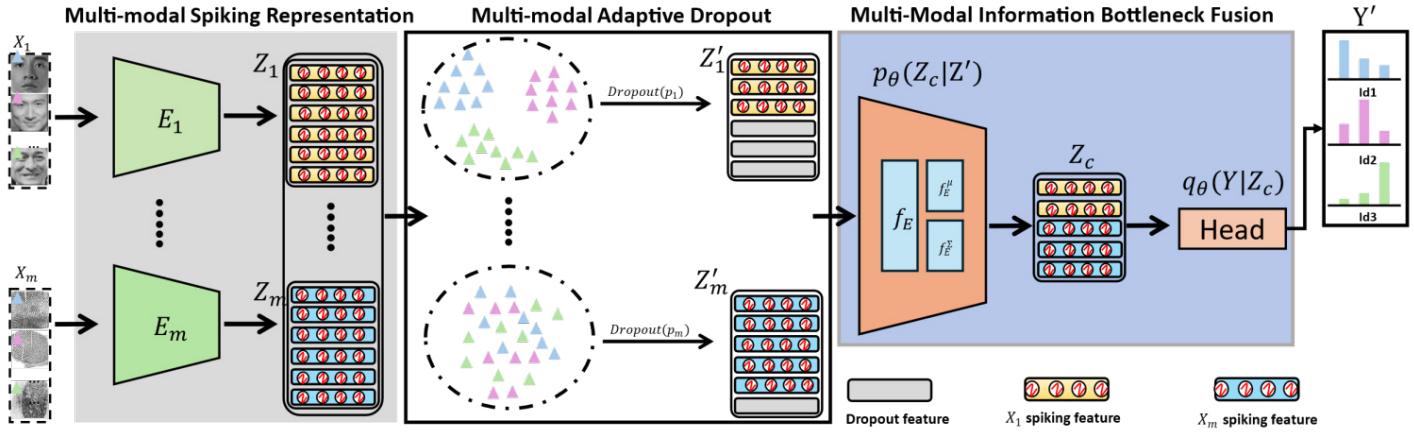

Figure 2: The pipeline of our proposed method comprises three key components: (1) Multimodal Spiking Representation, (2) Multimodal Adaptive Dropout, and (3) Multimodal Information Bottleneck Fusion

3.1. Multi-Modal Spiking Representation

Given the dataset $X = \{X_1, X_2, \ldots, X_m\}$, where m denotes the number of different biometric modalities, $\{X_1, X_2, \ldots, X_m\}$ represents each biometric multi-modal input, e.g., fingerprints, iris, and facial biometric inputs. The label Y represents the identity recognition or biometric authentication result determined by this set of inputs.

In the multi-modal spiking representation stage, we follow the

mainstream approach in existing research by initializing m modal spiking encoders, denoted as $E_1(\cdot)$, $E_2(\cdot)$,…, $E_m(\cdot)$. Here, m represents the number of modalities. Each spiking encoder $E_i(\cdot)$ is an independent encoder, *e.g.,* SEW-ResNet18 backbone [35,38]. Specifically, the process of transforming input X_i into representation vectors by the spiking encoders can be defined as,

$$Z_1, \dots, Z_m = E_1(X_1), \dots, E_m(X_m), \#(3)$$

where $Z_i \in R^{H\times W\times C}$ represents the spike-encoded features of the X_i input modality. This feature representation process aims to project input features into high-dimensional space, capturing key spatial and channel-wise dimensional information, including significant features, textures, structures, and semantic information of the modality, providing rich information for subsequent multi-modal fusion. Furthermore, through the built-in spiking neurons in the encoder, we transform these continuous features into spiking features. As shown in Figure 3, there is a fundamental difference between the distribution of spiking features and floating-point features. Specifically, compared to the floating-point features in previous works, spiking features exhibit discrete distributions, and these sparse discrete features can accelerate feature fusion computation and reduce memory overhead [37].

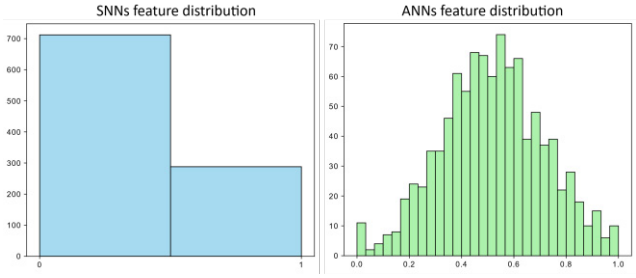

Figure 3: Comparative analysis of feature distributions: Binary-valued SNN features (**Left**) and continuous-valued ANN features (**Right**). We can observe that, compared to the continuous features of ANNs, the discrete features of SNNs reduce energy consumption in forward propagation by converting Multiply-Accumulate operations into Accumulate operations

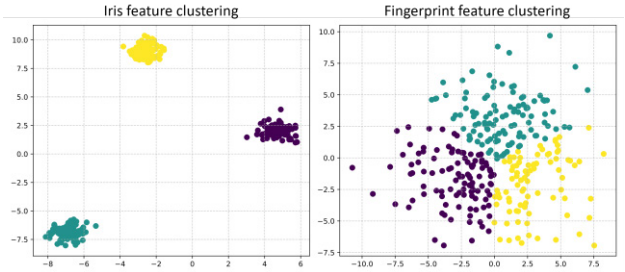

Figure 4: Illustration of t-SNE visualization results of different biometric modal features (**Left**: Iris; **Right**: Fingerprint) [39]. We can observe that when modal imbalance occurs, with iris modality being dominant, the iris modal features demonstrate good intra-class compactness and inter-class separability. In contrast, the fingerprint modality does not exhibit a clear clustering structure

3.2. Multi-modal Adaptive Dropout

In prior research, dropout has been extensively employed as a regularization technique, particularly in multi-modal systems to address overfitting and modal imbalance [34,40,41]. Traditional dropout methods typically utilize a uniform dropout rate, randomly discarding features during training to mitigate overfitting. However, this approach exhibits significant limitations: it cannot adapt to learning dynamics and convergence speeds across different modalities, and struggles to provide personalized optimization for each modality, thereby constraining learning efficiency and performance.

In-depth investigations reveal that during the later stages of deep network training, extracted high-level features become sufficiently rich for widespread pattern recognition tasks, such as cluster analysis. By removing the final fully connected layer (classification head) of neural networks, researchers discovered the ability to distinguish natural groups within datasets without explicit labels, as illustrated in Figure 4. Based on these findings, we can assess the convergence quality of each modality using the Convergence Quality Index [42,43]. We hypothesize that a shorter intra-class distance indicates more thorough feature training for a particular modality, while a longer intra-class distance suggests insufficient learning. We define the intra-class distance using the silhouette coefficient. Specifically, the input feature $z_i \in R^{B\times H\times W\times C}$ represents the high-dimensional feature of the i-th modality $i \in m$, and the label $Y \in R^{B\times N}$ denotes the category information for each sample.

For the features z_i of the i-th modality, we first flatten them

to $z_i' \in R^{B\times D}$ (where $D = H \times W \times C$) to facilitate distance calculations between samples. For each sample j = 1,2,…,B, its intra-class distance a_j is defined as the average Euclidean distance between this sample and other samples within its class C_k,

$$a(j) = \frac{1}{|C_k|-1}\sum_{x\in C_k, x\neq z_j'} |z_j' - x|_2 , \#(4)$$

where C_k is the class cluster to which sample j belongs (determined by Y or clustering results). z_j' represents the feature vector of sample j. ($|\cdot|$) denotes the Euclidean distance. $|C_k|$ is the number of samples in cluster C_k, which is calculated by grouping samples based on Y, specifically, it is mathematically defined as follows,

$$C_k = j \in |1,2,\dots,B \mid Yj, k = 1, \#(5)$$

the overall intra-class distance W_m for this modality is then defined as the average of intra-class distances across all samples,

$$W_m = \frac{1}{B}\sum_{j=1}^{B} a_j , \#(6)$$

where W_m reflects the compactness of features within the m-th modality, with smaller values indicating better convergence and larger values suggesting insufficient convergence.

$$p_m = \min\left(1, \max\left(0, \alpha \cdot \frac{W_m}{\max_i W_i}\right)\right), \#(7)$$

for modality m, the features are randomly dropped with a dropout rate of p_m, specifically defined in the following form,

$$z_m' = \text{dropout}(z_m, p_m) \#(8)$$

$$z_m' = z_m \odot M_m, \quad M_m \sim \text{Bernoulli}(1-p_m), \#(9)$$

where $\odot$ represents element-wise multiplication. $M_m \in \{0,1\}^{B\times D}$ is a mask matrix with the same dimensions as z_m, with each element independently sampled from the "Bernoulli" (1-p_m), retaining features with probability 1-p_m and dropping features with probability p_m. During the training phase, dropped feature values are set to 0; during the testing phase, the features are typically scaled (e.g., multiplied by 1-p_m to maintain consistent expectations.

3.3. Multi-Modal Information Bottleneck Fusion

After adaptive dropout, all intermediate features are represented as $Z = \{z_1', \dots, z_m'\}$. We hypothesize that these are the most discriminative features for each modality. Our next objective is to perform effective biometric multi-modal feature fusion, for which we draw inspiration from the Information Bottleneck (IB) theory. According to the IB, our goal is to find an optimal fusion strategy that transforms Z into a compressed representation Z_c, simultaneously maximizing the preservation of information relevant to Y.

Consequently, we define the following IB fusion objective function,

$$\mathcal{L}_{\text{MB}} = I(Z_c; Y) - I(Z_c; Z), \#(10)$$

the first term $I(Z_c;Y)$ represents the mutual information between the fused representation Z_c and the final prediction Y. Our objective is to maximize this value, which implies maximizing the information content of Z_c with respect to Y, thereby enabling Z_c to predict Y as accurately as possible. Through this approach, we aim to retain information most relevant to the prediction task while reducing redundancy across modalities.

However, directly solving this function is challenging due to the complex computation involved in high-dimensional joint distributions. According to the Variational Information Bottleneck (VIB) method (Alemi et al., 2016), we can reformulate the problem into maximizing an approximate variational lower bound [44]. First, given a probability distribution as,

$$p(Z_c, Z, Y) = p(Z_c|Z,Y)p(Y|Z)p(Z) = p(Z_c|Z)p(Y|Z)p(Z), \#(11)$$

we hypothesize that the generation of Z_c depends solely on Z_m' and is independent of Y, corresponding to the Markov chain $Y \leftrightarrow Z \leftrightarrow Z_c$.

Regarding the first term $I(Z_c,Y)$, based on variational inference theory, it can be transformed into the following form,

$$I(Z_c, Y) = \int dzdydz_c p(y \mid z) p(z_c \mid z) p(z) \log \frac{p(y \mid z_c)}{p(y)}. \#(12)$$

Due to the difficulty of directly calculating $p(y \mid z_c)$, we introduce a variational decoder $q_\theta (y \mid z_c)$ to approximate the estimation. Utilizing the non-negativity of Kullback-Leibler (KL) divergence, we can derive the lower bound of mutual information. For the mutual information $I(Z_c,Y)$, its lower bound can be expressed as,

$$I(Z_c, Y) \geq \int dzdydz_c p(y \mid z) p(z_c \mid z) p(z) \log q_\theta (y \mid z_c) + H(Y), \#(13)$$

where $H(Y)$ is the entropy of label Y, which can be ignored during the optimization process.
For the second term $I(Z_c;Z)$, we similarly use the variational inference method to derive its upper bound:

$$I(Z_c;Z) \le \int dz_c dz\, p(z) p(z_c \mid z'_m) \log \frac{p(z_c \mid z)}{r(z_c)}, \#(14)$$

where $p(z_c \mid z_m')$ is the posterior distribution of z_c, and $r(z_c)$ is the variational approximation of $p(z_c)$.
In summary, our multi-modal information bottleneck fusion objective function can be represented as:

$$J_{MIBF} = \frac{1}{M}\sum_{n=1}^{m} E\, z \sim p(z_c \mid z'_m) - \log q_\theta(y|z_c) + KL[p(z_c \mid z'_m), r(z_c)]. \#(15)$$

Through the above derivation process, we transform the complex mutual information computation into an optimizable objective function, converting the information bottleneck fusion problem into a variational optimization problem. There are many methods to optimize this problem. Where, we choose the method frequently used in previous works, optimizing through Variational Inference (e.g., Variational Autoencoders) [44]. Specifically, in practice, we replace $p(z_c|z_m')$ with a learned module (MLP) denoted as $p_\theta(z_c|z_m')$. The specific implementation method is as follows: we use MLPs to encode the feature set $\{z_1', \oplus, z_m'\}$ into two latent variables—mean $\mu(Z)$ and variance $\sigma(Z)$. Utilizing the reparameterization trick, we map these latent variables into a fused feature Z_c. Finally, another MLP maps Z_c to the final predicted output Y'. Specifically, it can be formally expressed as the following formula,

$$Z_l = f_{\mathrm{E}}(z'_1 \oplus z'_2 \oplus \cdots \oplus z'_m), \#(16)$$

$$\mu(Z_l), \sigma(Z_l) = f_{\mathrm{E}}^{\mu}(Z_l), f_{\mathrm{E}}^{\sigma}(Z_l). \#(17)$$

We generate Z_c from $\mu(Z_l)$ and $\sigma(Z_l)$, with the final formula as follows,

$$Z_c = \mu(Z_l) + \sigma(Z_l) \odot \epsilon, \quad \epsilon \sim \mathcal{N}(0, I). \#(18)$$

3.4. Training Objective

The final fused feature Z_c is utilized to generate the ultimate output $Y' \in R^{B\times N}$ for producing authentication results. During training, Y' represents the logits for each classification category, and we minimize the prediction error between Y' and Y using cross-entropy loss. Additionally, the regularization term of the information bottleneck is defined as minimizing the KL divergence between $p_\theta(Z_c \mid Z')$ and $p(Z_c)$. Specifically, the training objective is optimized through the following formula,

$$\mathcal{L}_{total} = \mathrm{CrossEntropy}(Y, Y') + \beta \cdot D_{\mathrm{KL}}\big(p_\theta(Z_c|Z')|p(Z_c)\big), \#(19)$$

where, $p_\theta(Z_c|Z'_m)$ parameterizes the distribution of the latent variable through $\mu(Z_l)$ and $\sigma(Z_l)$, while $p(Z)$ is defined as a standard normal distribution. The KL divergence $D_{\mathrm{KL}}\big(q_\theta(Z|X)|p(Z)\big)$ can be explicitly expressed as,

$$D_{\mathrm{KL}}\big(q_\theta(Z|X)|p(Z)\big) = \frac{1}{2}\sum_{j=1}^{D_z}\big(\mu(Z_l)_j^2 + \sigma(Z_l)_j^2 - 1 - \log\big(\sigma(Z_l)_j^2\big)\big), \#(20)$$

where D_z denotes the dimensionality of the latent variable Z_c.

In this way, we have completed the theoretical framework and method design of applying information bottleneck theory to multimodal fusion. As shown in Figure 2, during training, we followed previous work, which essentially involves training a classifier to classify multiple data features, and during verification, we perform identity authentication based on the confidence output by the model [9,45].

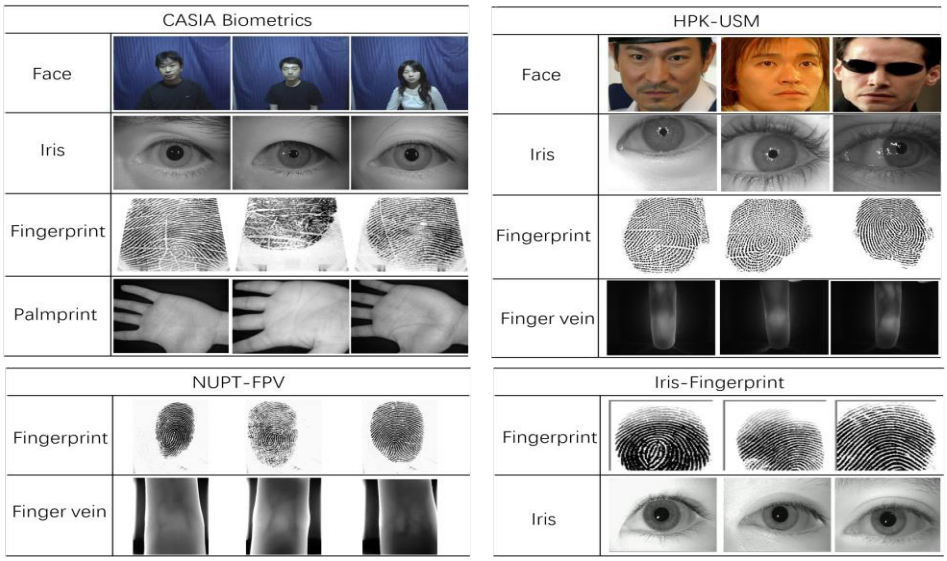

Figure 5: Illustration of the four datasets utilized in this experiment—CASIA Biometrics, HPK-USM, NUPT-FPV, and Iris-Fingerprint—which collectively encompass a wide range of biometric traits, including faces, irises, fingerprints, and palm prints

4. Experiment

4.1. Datasets

We evaluate our approach on four authoritative multimodal biometric datasets: CASIA, NUPT-FPV, Iris-Fingerprint, and the HPK-USM composite dataset [10,46]. Specifically, the CASIA dataset, released by the Chinese Academy of Sciences, encompasses multiple biometric modalities, from which we utilize fingerprint, face, iris, palmprint, and finger vein modalities. The HPK-USM dataset combines face, fingerprint, and iris data from Hong Kong Polytechnic University with finger vein data from the FV-USM dataset [46-50]. NUPT-FPV provides fingerprint and finger vein modalities, while the Iris-Fingerprint dataset from Kaggle, containing iris and fingerprint modalities, offers complementary experimental validation [51].

The selection and combination of these four datasets ensure experimental diversity and provide a comprehensive foundation for multimodal biometric analysis. All sample images are illustrated in Figure 5. Following previous work, we optimize these datasets through systematic preprocessing methods [9]. Specifically, for modalities with fewer images, we augment the data through random transformations including rotation, translation, cropping, and noise addition to achieve balanced image counts across modalities. The preprocessing stage includes resizing all images to a uniform dimension of 224×224, which aligns with the feature extraction network's specifications. Furthermore, we standardize category counts across modalities by matching them to the modality with the minimum number of categories. This optimization strategy effectively addresses both inter-modality image quantity discrepancies and category count variations. The final dataset statistics and processing results are presented in Table 1.

Name	Modality	No. of Classes	Train pairs	Train pairs	Test pairs
CASIA	Face	312	12480	7488	4992
	Palmprint				
	Fingerprint				
	Iris				
HPK-USM	Face	123	3690	2214	1476
	Fingerprint				
	Finger vein				
	Iris				
NUPT-FPV	Fingerprint	840	8400	5040	3260
	Finger vein				
Iris-Fingerprint	Fingerprint	45	540	270	180
	Iris				

Table 1: Biometric Dataset Statistics

4.2. Experimental Settings

We adapted SEW-ResNet18 for multimodal feature extraction, incorporating modifications guided by biometric dataset requirements and existing research [35,52,53]. The network architecture was streamlined by removing the final fully connected layer and adjusting to single-channel input for grayscale images. Training was conducted on an NVIDIA RTX 4090 GPU (PyTorch 2.0.0, Python 3.8, CUDA 11.8) using MSE loss and SGD optimization (weight decay: 1×10^{-3}). The training protocol included sequential learning rates (1e-3, 1e-4, 1e-5), a batch size of 64, and 20 epochs, with p_m capped at 0.3 for stability; the network structure comprises the p_θ fully connected layer with μ and σ projection layers (feature dimensions of 512 ×512 and 512 ×256 respectively), along with the q_θ classification layer. We utilize the classic LIF neuron as the basic neuron model and employ surrogate gradients for SNN backpropagation. In evaluating model performance, we employ the Correct Identification Rate (CIR) as the measurement metric, which is an essential method for assessing classification model performance. The CIR calculates classification accuracy through the following formula:

$$CIR = \frac{n_c}{n_t} \times 100\%, \#(22)$$

where n_c and n_t denote correct and total samples. To assess the stability and reliability of our method more objectively and accurately, we follow previous research by conducting each experiment five times [46]. We then calculate the mean (MEAN) and standard error (STD) as the final performance evaluation metrics. In terms of energy efficiency evaluation, following established practices in SNN research, we employ the number of synaptic operations (SOPs) as our theoretical energy consumption metric [54]. This approach provides a reliable estimation of computational efficiency by calculating the accumulated spike-driven synaptic operations throughout the network's inference process.

4.3. Comparison Study

To comprehensively validate our method's effectiveness, we selected four representative multimodal biometric recognition methods for comparison. For multimodal recognition, we chose FPV-Net by Ren et al. and NLNet by Guo et al.; for modality imbalance solutions, we selected PMR and OGM-GE [9,30,38]. As shown in Table 2, experimental results demonstrate that our method is competitive across multiple datasets, particularly excelling on NUPT-FPV and Iris-Fingerprint datasets [9,30,38,46].

Method	CASIA	HPK-USM	NUPT-FPV	Iris-Fingerprint	OPs(G)↓	Energy(mJ) ↓
FPV-Net [46]	91.05 ± 0.006%	92.05 ± 0.006%	88.05 ± 0.006%	91.03 ± 0.006%	85.28	216.40
NLNet [9]	92.05 ± 0.021%	91.21 ± 0.011%	85.05 ± 0.020%	91.05 ± 0.002%	67.33	126.30
PMR [38]	90.21 ± 0.006%	91.31 ± 0.031%	88.01 ± 0.018%	92.31 ± 0.004%	98.33	266.60
OMG-GE [30]	90.21 ± 0.018%	92.19 ± 0.042%	89.09 ± 0.017%	93.19 ± 0.012%	98.33	266.60
Ours	**95.31 ± 0.019%**	**95.05 ± 0.021%**	**90.71 ± 0.006%**	**94.21 ± 0.021%**	**4.83**	**1.32**

Table 2: Comparison of the Performance of Different Methods Across Four Datasets

According to the results shown in Table 2, our method achieves significant performance improvements across multiple datasets including CASIA Biometrics, NUPT-FPV, and Iris-Fingerprint. Additionally, it demonstrates lower energy consumption and computational operations, making it more suitable for embedded device deployment [9]. Specifically, on the CASIA Biometrics dataset, our method outperforms the previous best method NLNet by 3.26%, with similar advantages observed across other datasets. More importantly, our method exhibits the smallest standard deviation across multiple datasets, further confirming its robustness and reliability. Furthermore, our method requires only about 1/100 of the energy consumption compared to previous ANN-based methods, which greatly addresses the power consumption issues of resource-constrained devices, enabling efficient edge computing deployment.

4.4. Multimodal Verification Study

We conducted a comprehensive validation of our method's robustness through multimodal verification analysis, building upon established evaluation protocols [55]. The methodology transformed traditional many-to-one classification into one-to-one feature space metric learning. Feature vectors, extracted from the network's final fully connected layer, were compared using Euclidean distance to classify sample pairs as either "genuine matches" (same class) or "impostor matches" (different classes).

We conducted experimental evaluations on three datasets: CASIA, NUPT-FPV, and Iris-Fingerprint. For CASIA, samples 1-20 were used for training, 21-25 for optimization, and 26-40 for testing; for the other two datasets, samples 1-3 from each category were used for training, 4-5 for optimization, and 6-10 for testing. The experimental results, as show in Figure 6, demonstrated through ROC curves, showed significant performance advantages: on CASIA and NUPT-FPV datasets, our method exhibited superior recognition accuracy across multiple FAR intervals; on the Iris-Fingerprint dataset, despite narrower performance gaps, it maintained optimal EER in the low to medium FAR range.

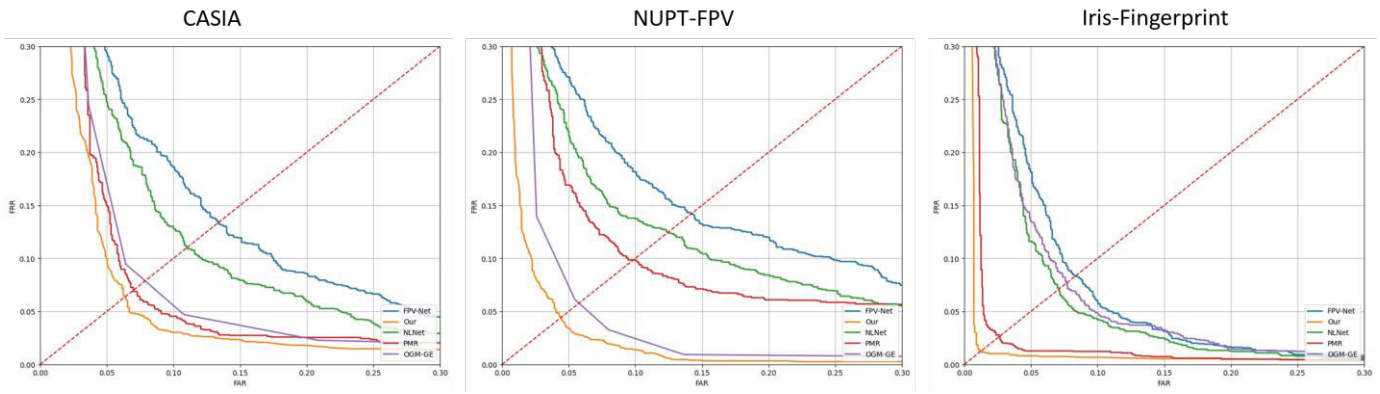

Figure 6: The ROC Curves of Different Methods on the CASIA, NUPT-FPV and Iris-Fingerprint

4.5. Modality Replacement Evaluations

In the study of multimodal biometric recognition, modality imbalance may result in an overreliance on one or a few prevailing modalities, which can compromise the stability and dependability of system outcomes. To thoroughly examine the model's behavior when certain modalities are entirely unavailable and to gauge its reliance on different modalities, we developed and executed **Protocol 1**. In multimodal biometric recognition, modality imbalance may lead to system over-reliance on specific modalities, affecting system stability and reliability. To evaluate the model's dependency on different modalities, we designed **Protocol 1**. This protocol simulates real-world modality spoofing and attacks by replacing specific modalities with Gaussian noise. We first train the model under complete modality conditions to establish baseline performance metrics P_mb, followed by modality replacement experiments. To ensure result reliability, each replacement condition is independently tested five times. Using CIR as the performance metric, we calculate the performance variation rate using:

$$\Delta P = \left| \frac{P_{mb} - P_{mr}}{P_{mb}} \right| \times 100\% \#(23)$$

where P_{mb} and P_{mr} represent the performance metrics before and after replacement, respectively. Through this variation rate, we assess each modality's impact on overall system performance.

Model	**CASIA (ΔCIA%) Missing Face/Iris/ Fingerprint/Palmprint**	**Iris-Fingerprint (ΔCIA%)**
Missing Iris/ Fingerprint		
FPV-Net [46]	9.03/21.06/30.92/4.04	30.05/0.039
NLNet [9]	17.93/12.03/9.93/4.8	18.86/15.84
PMR [38]	15.89/14.06/11.92/9.94	14.321/12.30
OMG-GE [30]	13.10/10.069/13.929/12.06	11.06/10.10
Ours	**10.81/10.64/11.28/10.93**	**11.29/11.48**

Table 3: Quantitative Performance Assessment (Δ*CIA*%) on CASIA And Iris-Fingerprint Databases

In Table 3, we systematically evaluated multimodal biometric recognition models under modality loss conditions. An ideal multimodal recognition model should exhibit balanced dependence across modalities, with consistent performance degradation when any modality is replaced by Gaussian noise. Analysis revealed significant disparities in multimodal balance among existing methods. The specialized architectures of NLNet and FPV-Net for multimodal biometric recognition revealed substantial constraints in managing modality imbalance, manifested through inconsistent performance degradation patterns following modality loss. In contrast, our proposed method and other modality balancing strategies demonstrated more stable performance characteristics, maintaining higher consistency under modality perturbation. These findings highlight a key advantage of our approach: the ability to extract critical information from each modality without over-relying on any single modality, thereby enhancing model robustness.

4.6. Ablation Study

To comprehensively dissect the functionality and value of core methodological components, we designed a meticulous ablation study. In the experiments, we systematically dismantled the Multi-modal Information Bottleneck Fusion (MIBF) and Multi-modal Adaptive Dropout (MAD) modules, replacing them with alternative approaches. This strategy aimed to precisely evaluate the actual impact of key components on overall performance and conduct an in-depth analysis of each module's independent contribution to multimodal learning.

Through comparative experiments, as shown in Table 4, we observed that applying different substitution strategies such as concatenation, attention, and addition to MIBF yielded differentiated performance on CASIA and Iris-Fingerprint datasets, with overall performance slightly inferior to the original approach. This finding highlights the unique advantages of MIBF in feature integration. Interestingly, removing the MAD module appeared to bring about minor performance improvements, which are likely attributable to the model's subsequent over-reliance on features from a dominant modality. This systematic module deconstruction not only revealed the critical roles of individual components but also provided valuable research perspectives for understanding the intrinsic mechanisms of multimodal feature fusion. By precisely parsing and reconstructing the model architecture, we gained deeper insights into the collaborative mechanisms and unique values of modules in multimodal learning.

Model	CASIA	Iris-Fingerprint
w/o MIBF+Concat	92.11%	92.11%
w/o MIBF+Atten	90.31%	92.87%
w/o MIBF+Add	91.75%	91.75%
w/o MAD	93.10%	92.95%

Table 4: Ablation Study Results

4.7. Qualitative Results

We deeply explored the intrinsic mechanisms and performance of our proposed method through two complementary visualization techniques: Grad-CAM saliency mapping and T-SNE feature space visualization [8,9]. These techniques not only reveal the interpretability of our model but also validate the effectiveness of our proposed approach.

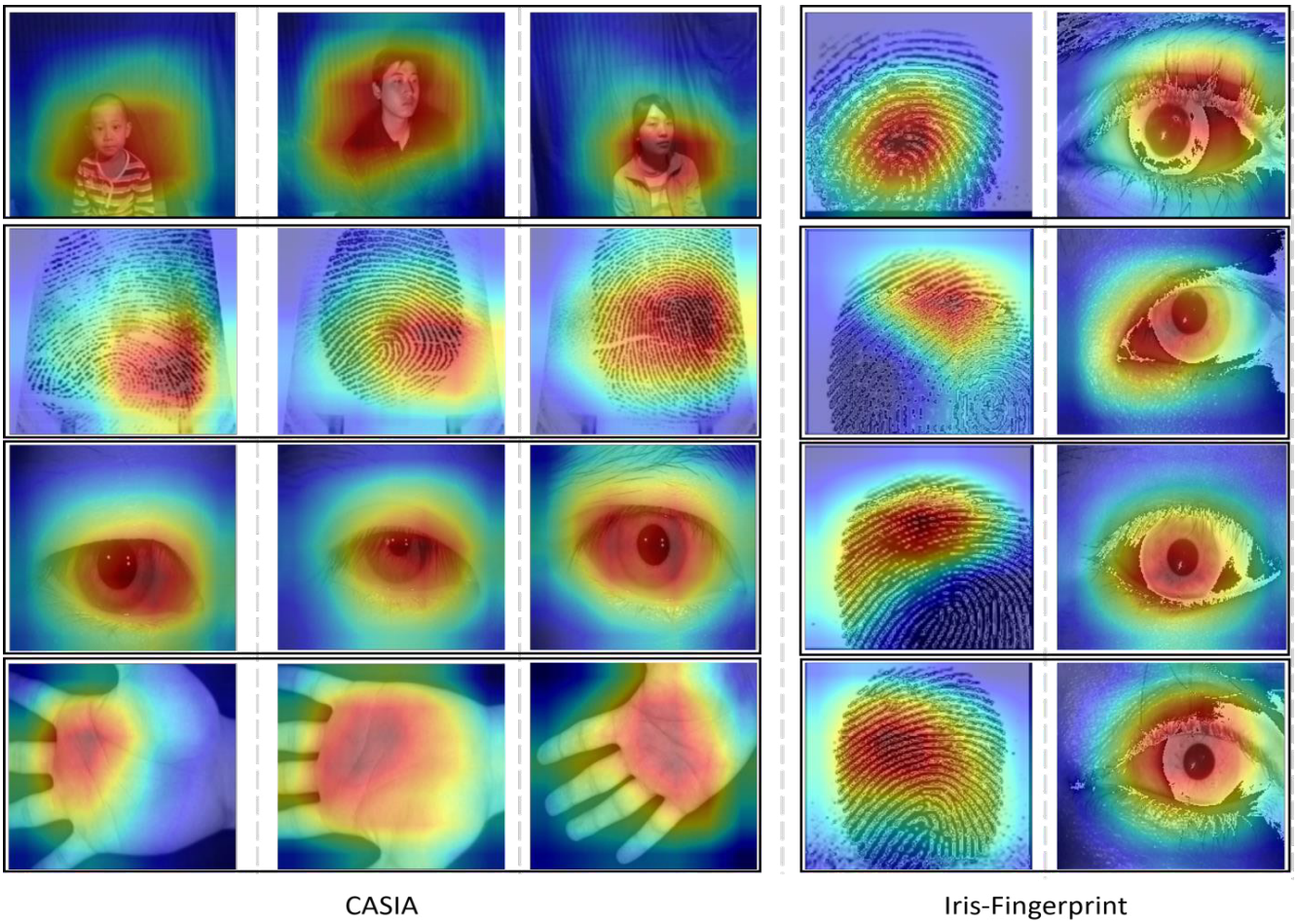

Figure 7: T-SNE Results. We visualized the features of each modality using t-SNE and compared the results with FPV-Net. Our findings reveal that, in contrast to FPV-Net, our approach successfully learned consistent and highly discriminative features across all modalities

Grad-CAM Results. Grad-CAM is a widely applied visualization technique that can generate category activation maps, As show in Figure 7, demonstrating the critical regions that deep neural networks focus on during classification decisions [8]. We performed Grad-CAM visualization on input data for each modality, aiming to reveal the significant regions of attention across different modalities. Specifically, we observed that our method can effectively capture key regions in different modalities, such as faces, fingerprints, irises, and palm prints, successfully capturing the most discriminative feature regions for each modality. This fine-grained visualization technique not only enhances the model's interpretability but also validates the effectiveness of our proposed multi-modal feature extraction strategy.

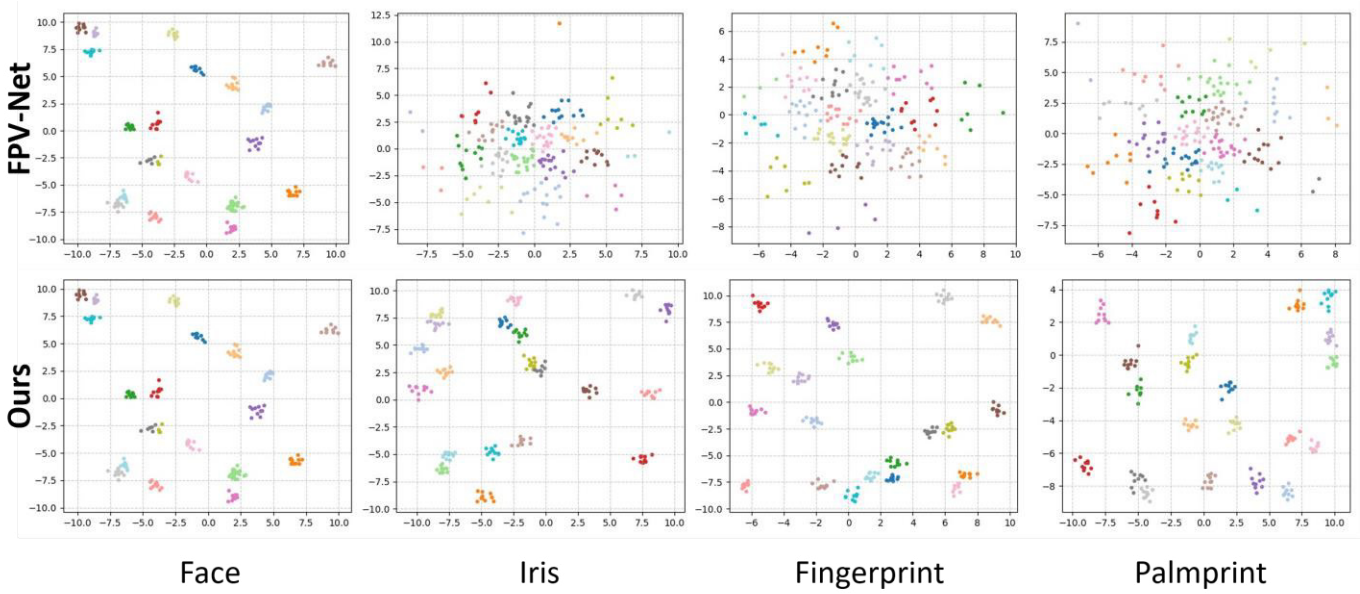

Figure 8: T-SNE Results. We visualized the features of each modality using t-SNE and compared the results with FPV-Net. Our findings reveal that, in contrast to FPV-Net, our approach successfully learned consistent and highly discriminative features across all modalities

T-SNE Results. To further understand the feature representations learned by our model, As show in Figure 8, we employed the T-SNE technique to perform low-dimensional visualization of the feature spaces across different modalities [39]. T-SNE is a non-linear dimensionality reduction method that can effectively map high-dimensional features to two- or three-dimensional spaces while maximally preserving the local structure between data points. Through T-SNE visualization, we obtained the following key observations: First, for each modality, data points from the same category exhibit a tight clustering characteristic in the feature space. This aggregation indicates that our multi-modal learning approach successfully learned discriminative feature representations. Second, data points from different categories are clearly separated in the feature space, forming distinctly distinguishable clusters. Finally, we observed consistent results across all modalities, demonstrating that our method can effectively learn discriminative features for each modality, successfully mitigating the modal imbalance problem.

5. Conclusion

In this paper, we present a spiking neural network-based framework for multimodal biometric recognition that tackles modality imbalance using information bottleneck fusion and dynamic dropout mechanisms. Our primary contributions are twofold: First, we introduce a novel multimodal fusion methodology that integrates spiking neural networks with information bottleneck theory, achieving optimal preservation of relevant information while minimizing redundancy. Second, we develop an adaptive dynamic dropout strategy that facilitates flexible feature adaptation throughout the training process. Extensive experiments on multiple public datasets demonstrate the superior performance of our approach. Furthermore, our method exhibits low energy consumption, making it particularly well-suited for embedded device deployment and applications.

Looking forward, the multimodal data fusion framework proposed in this research demonstrates broad application prospects. Our method is not limited to biometric recognition but may potentially extend to multiple domains such as medical imaging, cross-modal learning, and multimedia analysis. Future research can further refine the information bottleneck and adaptive dropout strategies, exploring their applicability to a wider range of multimodal data processing challenges.

Acknowledgments

This research received partial funding from the State Grid Corporation of China Headquarters Technology Project (Grant No. 5500-202440171A-1-1-ZN).

References

1. Hou, B., Zhang, H., & Yan, R. (2022). Finger-vein biometric recognition: A review. *IEEE Transactions on Instrumentation and Measurement, 71*, 1-26.
2. Jia, W., Xia, W., Zhang, B., Zhao, Y., Fei, L., Kang, W., ... & Guo, G. (2021). A survey on dorsal hand vein biometrics. *Pattern Recognition, 120,* 108122.
3. Sultana, M., Paul, P. P., & Gavrilova, M. L. (2017). Social behavioral information fusion in multimodal biometrics. *IEEE transactions on systems, man, and cybernetics: systems, 48*(12), 2176-2187.
4. Sharma, S., Saini, A., & Chaudhury, S. (2024). Multimodal biometric user authentication using improved decentralized fuzzy vault scheme based on Blockchain network. *Journal of Information Security and Applications, 82*, 103740.
5. Coelho, K. K., Tristão, E. T., Nogueira, M., Vieira, A. B., & Nacif, J. A. (2023). Multimodal biometric authentication method by federated learning. *Biomedical Signal Processing and Control, 85,* 105022.
6. Kumar, A., & Zhou, Y. (2011). Human identification using finger images. *IEEE Transactions on image processing, 21*(4), 2228-2244.
7. Kosmala, J., & Saeed, K. (2012). Human identification by vascular patterns. In *Biometrics and Kansei Engineering* (pp. 67-87). New York, NY: Springer New York.
8. Selvaraju, R. R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., & Batra, D. (2017). Grad-cam: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE international conference on computer vision* (pp. 618-626).
9. Guo, Z., Ma, H., & Liu, J. (2024). NLNet: A narrow-channel lightweight network for finger multimodal recognition. *Digital Signal Processing, 150,* 104517.
10. Muhammad, J., Wang, Y., Wang, C., Zhang, K., & Sun, Z. (2021). Casia-face-africa: A large-scale african face image database. *IEEE Transactions on Information Forensics and Security, 16,* 3634-3646.
11. Kamlaskar, C., & Abhyankar, A. (2021). Iris-fingerprint multimodal biometric system based on optimal feature level fusion model. *AIMS Electronics and Electrical Engineering, 5*(4), 229-250.
12. Kamil, B. Z., Salman, D. A., & Kareem, S. A. DETECT INTEREST REGION OF FINGER VEIN BASED ON K-MEANS CLUSTERING.
13. Yuan, C., Jiao, S., Sun, X., & Wu, Q. J. (2021). MFFFLD: A multimodal-feature-fusion-based fingerprint liveness detection. *IEEE Transactions on Cognitive and Developmental Systems, 14*(2), 648-661.
14. Aizi, K., & Ouslim, M. (2022). Score level fusion in multi-biometric identification based on zones of interest. *Journal of King Saud University-Computer and Information Sciences, 34*(1), 1498-1509.
15. Pradhan, A., He, J., & Jiang, N. (2021). Score, rank, and decision-level fusion strategies of multicode electromyogram-based verification and identification biometrics. *IEEE Journal of Biomedical and Health Informatics, 26*(3), 1068-1079.
16. Purohit, H., & Ajmera, P. K. (2021). Optimal feature level fusion for secured human authentication in multimodal biometric system. *Machine Vision and Applications, 32*(1), 24.
17. Gona, A., Subramoniam, M., & Swarnalatha, R. (2023). Transfer learning convolutional neural network with modified Lion optimization for multimodal biometric system. *Computers and Electrical Engineering, 108*, 108664.
18. Zhong, D., Shao, H., & Du, X. (2019). A hand-based multi-biometrics via deep hashing network and biometric graph matching. *IEEE Transactions on Information Forensics and Security, 14*(12), 3140-3150.
19. Abdullahi, S. B., Bature, Z. A., Chophuk, P., & Muhammad, A. (2023). Sequence-wise multimodal biometric fingerprint and finger-vein recognition network (STMFPFV-Net). *Intelligent Systems with Applications, 19,* 200256.
20. Lu, P., Hu, L., Mitelpunkt, A., Bhatnagar, S., Lu, L., & Liang, H. (2024). A hierarchical attention-based multimodal fusion framework for predicting the progression of Alzheimer's disease. *Biomedical Signal Processing and Control, 88*, 105669.
21. Jeong, S. W., Cho, H. H., Lee, S., & Park, H. (2022). Robust multimodal fusion network using adversarial learning for brain tumor grading. *Computer Methods and Programs in Biomedicine, 226*, 107165.
22. Li, L., Pan, H., Liang, Y., Shao, M., Xie, S., Lu, S., & Liao, S. (2024). PMFN-SSL: Self-supervised learning-based progressive multimodal fusion network for cancer diagnosis and prognosis. *Knowledge-Based Systems, 289*, 111502.
23. Xiao, X., Liu, G., Gupta, G., Cao, D., Li, S., Li, Y., ... & Bogdan, P. (2024). Neuro-inspired information-theoretic hierarchical perception for multimodal learning. *arXiv preprint arXiv:2404.09403.*
24. Jiang, J., Liu, Z., & Zheng, N. (2024). Correlation information bottleneck: Towards adapting pretrained multimodal models for robust visual question answering. *International Journal of Computer Vision, 132*(1), 185-207.
25. Mai, S., Zeng, Y., & Hu, H. (2022). Multimodal information bottleneck: Learning minimal sufficient unimodal and multimodal representations. *IEEE Transactions on Multimedia, 25,* 4121-4134.
26. Kuang, H., Liu, H., Wu, Y., Satoh, S. I., & Ji, R. (2023). Improving adversarial robustness via information bottleneck distillation. *Advances in Neural Information Processing Systems, 36,* 10796-10813.
27. Nagrani, A., Yang, S., Arnab, A., Jansen, A., Schmid, C., & Sun, C. (2021). Attention bottlenecks for multimodal fusion. *Advances in neural information processing systems, 34,* 14200-14213.
28. Du, C., Li, T., Liu, Y., Wen, Z., Hua, T., Wang, Y., & Zhao, H. (2021). Improving multi-modal learning with uni-modal teachers. *arXiv preprint arXiv:2106.11059*.
29. Wang, W., Tran, D., & Feiszli, M. (2020). What makes training multi-modal classification networks hard?. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 12695-12705).

30. Peng, X., Wei, Y., Deng, A., Wang, D., & Hu, D. (2022). Balanced multimodal learning via on-the-fly gradient modulation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 8238-8247).
31. Winterbottom, T., Xiao, S., McLean, A., & Moubayed, N. A. (2020). On modality bias in the tvqa dataset. *arXiv preprint arXiv:2012.10210.*
32. Goyal, Y., Khot, T., Summers-Stay, D., Batra, D., & Parikh, D. (2017). Making the v in vqa matter: Elevating the role of image understanding in visual question answering. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 6904-6913).
33. Jabri, A., Joulin, A., & Van Der Maaten, L. (2016, September). Revisiting visual question answering baselines. In *European conference on computer vision* (pp. 727-739). Cham: Springer International Publishing.
34. Xiao, F., Lee, Y. J., Grauman, K., Malik, J., & Feichtenhofer, C. (2020). Audiovisual slowfast networks for video recognition. *arXiv preprint arXiv:2001.08740.*
35. Fang, W., Yu, Z., Chen, Y., Huang, T., Masquelier, T., & Tian, Y. (2021). Deep residual learning in spiking neural networks. *Advances in Neural Information Processing Systems, 34,* 21056-21069.
36. Liu, X., Xia, N., Zhou, J., Li, Z., & Guo, D. Towards energy-efficient audio-visual classification via multimodal interactive spiking neural network. *ACM Transactions on Multimedia Computing, Communications and Applications.*
37. Wu, Y., Deng, L., Li, G., Zhu, J., Xie, Y., & Shi, L. (2019, July). Direct training for spiking neural networks: Faster, larger, better. In *Proceedings of the AAAI conference on artificial intelligence* (Vol. 33, No. 01, pp. 1311-1318).
38. Fan, Y., Xu, W., Wang, H., Wang, J., & Guo, S. (2023). Pmr: Prototypical modal rebalance for multimodal learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (pp. 20029-20038)
39. Van der Maaten, L., & Hinton, G. (2008). Visualizing data using t-SNE. *Journal of machine learning research, 9*(11).
40. Ke, Z., Wen, Z., Xie, W., Wang, Y., & Shen, L. (2020, April). Group-wise dynamic dropout based on latent semantic variations. In *Proceedings of the AAAI Conference on Artificial Intelligence* (Vol. 34, No. 07, pp. 11229-11236).
41. Poernomo, A., & Kang, D. K. (2018). Biased dropout and crossmap dropout: learning towards effective dropout regularization in convolutional neural network. *Neural networks, 104,* 60-67.
42. Du, Y., Zhou, D., Xie, Y., Lei, Y., & Shi, J. (2023). Prototype-guided feature learning for unsupervised domain adaptation. *Pattern Recognition, 135*, 109154.
43. Chen, G., Li, X., Yang, Y., & Wang, W. (2024). Neural clustering based visual representation learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (pp. 5714-5725).
44. Alemi, A. A., Fischer, I., Dillon, J. V., & Murphy, K. (2016). Deep variational information bottleneck. *arXiv preprint arXiv:1612.00410.*
45. Kingma, D. P., & Welling, M. (2013, December). *Auto-encoding variational bayes*
46. Ren, H., Sun, L., Guo, J., & Han, C. (2022). A dataset and benchmark for multimodal biometric recognition based on fingerprint and finger vein. *IEEE Transactions on Information Forensics and Security, 17,* 2030-2043.
47. Wang, T. Y., & Kumar, A. (2016, February). Recognizing human faces under disguise and makeup. In *2016 IEEE International Conference on Identity, Security and Behavior Analysis (ISBA)* (pp. 1-7). IEEE.
48. Lin, C., & Kumar, A. (2018). Matching contactless and contact-based conventional fingerprint images for biometrics identification. *IEEE Transactions on Image Processing, 27*(4), 2008-2021.
49. Nalla, P. R., & Kumar, A. (2016). Toward more accurate iris recognition using cross-spectral matching. *IEEE transactions on Image processing, 26*(1), 208-221.
50. Wang, K., & Kumar, A. (2019). Cross-spectral iris recognition using CNN and supervised discrete hashing. *Pattern Recognition, 86,* 85-98.
51. Asaari, M. S. M., Suandi, S. A., & Rosdi, B. A. (2014). Fusion of band limited phase only correlation and width centroid contour distance for finger based biometrics. *Expert Systems with Applications, 41*(7), 3367-3382.
52. Xue, G., Li, S., Hou, P., Gao, S., & Tan, R. (2023). Research on lightweight Yolo coal gangue detection algorithm based on resnet18 backbone feature network. *Internet of Things, 22,* 100762.
53. Guo, F., Wang, Y., & Qian, Y. (2023). Real-time dense traffic detection using lightweight backbone and improved path aggregation feature pyramid network. *Journal of Industrial Information Integration, 31*, 100427.
54. Zhou, Z., Zhu, Y., He, C., Wang, Y., Yan, S., Tian, Y., & Yuan, L. (2022). Spikformer: When spiking neural network meets transformer. *arXiv preprint arXiv:2209.15425.*
55. Lin, X., Wang, S., Cai, R., Liu, Y., Fu, Y., Tang, W., ... & Kot, A. (2024). Suppress and rebalance: Towards generalized multi-modal face anti-spoofing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (pp. 211-221).

Copyright: *©2025 Xu Liu, et al. This is an open-access article distributed under the terms of the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.*