scientific reports

Check for updates

OPEN

Multimodal feature enhancement via dynamically-aware heterogeneous network for face anti-spoofing

Yutao Yan, Liang Shi✉, Yijun Zhang & Xiaosong Chang

Presentation attacks pose a significant threat to face recognition systems, making face anti-spoofing (FAS) a critical security measure. However, many existing approaches suffer from inadequate exploitation of physical modality cues and rely on overly complex architectures, which hinder their deployment in practical applications. This paper introduces DAH-FAS, a Dynamically-Aware Heterogeneous Face Anti-Spoofing Network designed to mitigate the limitations of existing face anti-spoofing methods. To reinforce the RGB branch's capacity for detailed feature extraction, we have designed a Variance-Adaptive Multi-Scale Residual Block (VA-MSRB). To improve the model's perception of bio-thermal diffusion patterns, the BioThermal Enhancer (BTE) is integrated into the GhostNet backbone of the IR branch. On this basis, a Bidirectional Group Cross-Modal Attention (BGC-MA) mechanism is constructed between the IR and depth branches during the feature extraction stage, enabling cross-modal geometric feature alignment and enhancing the complementarity among features. We evaluate our method on the CASIA-SURF, CASIA-SURF CeFA, and WMCA datasets, and results demonstrate that the proposed approach achieves significant advantages in differentiating real from fake faces.

As face recognition technology progresses rapidly, its applications have become increasingly widespread in areas such as security surveillance, identity verification, and mobile payment. However, an increasing number of presentation attacks have been launched against face recognition systems. These span from basic 2D print/replay attacks to advanced 3D spoofing methods like high-fidelity silicone masks and lifelike 3D head models. By mimicking facial texture, motion patterns, and 3D deformations, attackers continuously challenge the defensive boundaries of conventional detection algorithms. Such spoofing attacks pose serious threats to user privacy and financial security. To ensure dependable face recognition and counter potential attacks, face anti-spoofing (FAS) technologies serve as an essential safeguard to reinforce system integrity and trust.

In recent years, with the rising focus from researchers, various face anti-spoofing techniques have emerged, which are typically grouped into two overarching types: traditional machine learning approaches based on handcrafted features and modern methods driven by deep learning approaches.

Traditional machine learning approaches focus on designing features that capture inherent properties and texture information in images or videos, such as Local Binary Patterns (LBP)[1] and Histogram of Oriented Gradient (HOG)[2,3], which are often used in conjunction with traditional machine learning approaches such as Support Vector Machines (SVMs) for extracting and classifying features. Additionally, motion-based methods typically require users to perform a series of predefined actions–such as blinking, lip movement, or head rotation–to cooperate with the verification process. For instance, Pan et al.[4,5] proposed using the entire blinking process as an indicator for liveness detection, while Kollreider et al.[6] introduced a face anti-spoofing method by analyzing mouth movement. Although these traditional machine learning approaches achieved certain levels of success, their limited feature representation capacity has become increasingly evident when confronting more sophisticated spoofing attacks.

Compared with the limitations of handcrafted features, deep learning approaches have demonstrated superior capabilities in adaptively capturing cross-modal spoofing cues through data-driven feature learning mechanisms. For instance, ResNet-101[7], a deep residual network, alleviates the gradient vanishing issue via residual links and enables high-dimensional feature representations. However, its ability to detect subtle spoofing cues remains insufficient. To overcome this limitation, CDCN[6] introduces the Central Difference Convolution

School of Computer, Jiangsu University of Science and Technology, Jiangsu 212000, China. ✉email: jsjxy_sl@just.edu.cn

(CDC) to enhance the detection of subtle spoof cues. Nevertheless, the single-path feature extraction structure of CDCN restricts its ability to fully exploit cross-layer information. Although CDCN++[6] improves performance by incorporating a Multiscale Attention Fusion Module (MAFM) and Neural Architecture Search (NAS), it still lacks sufficient sensitivity to subtle artifacts in complex materials, such as silicone masks, which limits its spoof detection capability.

Some researchers have introduced spatiotemporal information as auxiliary supervision for better classification of live/spoof face. However, these methods often lead to increased computational complexity. For example, Xu et al.[8] leveraged the strengths of Long Short-Term Memory (LSTM) networks and Convolutional Neural Networks (CNNs) by extracting frame-level features using CNNs and modeling their temporal dynamics through LSTM for binary classification of live and spoof faces. Khan et al.[9] built a lightweight face anti-spoofing system utilizing the MobileNetV3 architecture, which leverages temporal and spatial features extracted from video frames to enhance its capability in detecting presentation attacks.

George et al.[10] addressed the limitations of traditional approaches under complex spoofing scenarios by combining pixel-level supervision with attention mechanisms, which contributes to more accurate liveness verification. To improve intra-modal representation, Zhang et al.[11] designed a novel multimodal multi-scale fusion strategy that applies channel attention to boost discriminative features and reduce noise across different modalities.

However, these deep learning approaches still overly make use of data-driven representations and fail to effectively integrate physical priors such as biometric cues. This leads to blind spots in detecting highly realistic spoofing attacks. Moreover, their complex architectures pose challenges for achieving real-time performance, particularly when resources are limited.

This work presents Dynamically-Aware Heterogeneous Face Anti-Spoofing Network (DAH-FAS) to overcome the limitations discussed above. The key contributions can be summarized as follows:

- Variance-Adaptive Multi-Scale Residual Block (VA-MSRB) is introduced in the RGB branch. It utilizes a tri-branch heterogeneous structure combined with variance-guided fusion to overcome the limitations of fixed receptive fields in traditional convolutions and mitigate the loss of cross-scale information typically caused by single-path convolutional operations.
- BioThermal Enhancer (BTE) is embedded into the IR branch to capture the subtle thermodynamic differences between silicone masks and real human skin, thereby improving the model's capability to detect thermal camouflage.
- A Bidirectional Group Cross-Modal Attention (BGC-MA) mechanism is constructed between the depth and IR branches to compensate for information degradation resulting from geometric misalignment between the two modalities. This mechanism enables alignment of geometric features across modalities, thereby enhancing the effectiveness of multimodal feature fusion.

Related work

Inverted residual

The inverted residual structure was first introduced by Sandler et al.[12] in MobileNetV2, with the core idea of incorporating an efficient residual connection into lightweight networks to lower both computation overhead and model size, without compromising performance.

As shown in Fig. 1, The typical design of an inverted residual block begins with a 1×1 convolution to increase dimensionality, continues with a depthwise separable convolution (DSConv), and concludes with a 1×

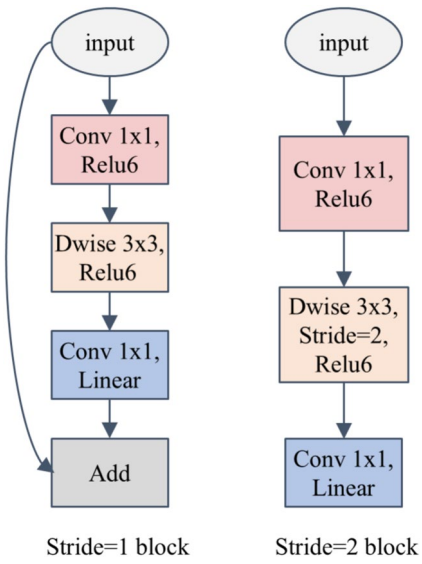

Figure 1. Structure of the inverted residual block.

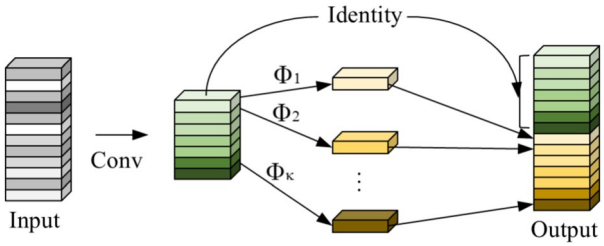

Figure 2. Structure of the Ghost module.

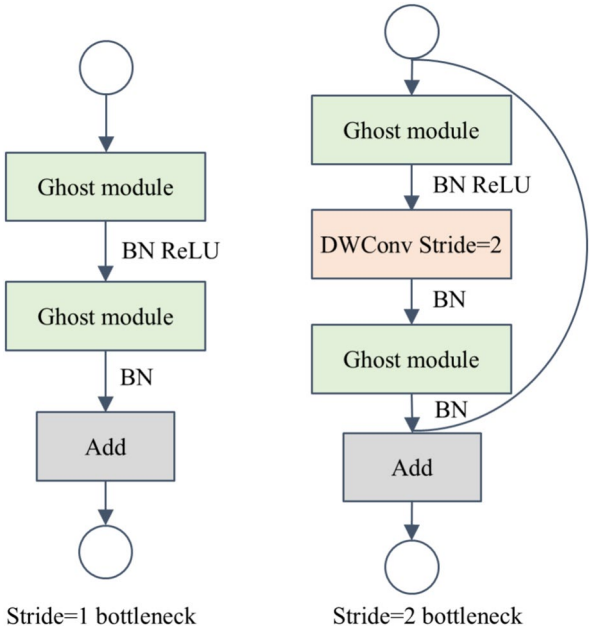

Figure 3. Structure of the Ghost bottlenecks.

1 convolution to reduce dimensions. For the case of stride = 1, a residual connection is added. This structure constrains the non-linear activation function ReLU6 within the channel transformation stages, which helps maintain the stability of feature representations.

This "inverted" design allows for more effective information retention in low-dimensional space, while enabling nonlinear transformations in high-dimensional space to enhance representational capacity.

Ghost module and Ghost bottlenecks

The foundational unit of GhostNet, known as the Ghost module, was originally designed by Kai Han et al.[13]. Its main idea is that there exists significant redundancy within the feature maps of convolutional neural networks. To address this, the Ghost module first generates a small set of intrinsic features using standard convolution, and then produces additional "ghost features" through inexpensive linear operations, significantly reducing computational cost. The GhostNet backbone primarily consists of Ghost bottlenecks units, whose foundational component is the Ghost module, as detailed in Fig. 2.

The Ghost bottlenecks design shares structural characteristics with the inverted residual structure from MobileNetV2. However, it employs the Ghost module instead of standard convolutions to further enhance lightweight properties. Specifically, Channel expansion is carried out by the first Ghost module, while the second one performs channel reduction to ensure compatibility with the input and enable a residual shortcut. When the stride equals 1, The feature maps, whose spatial dimensions are preserved, provide enhanced representation capability. When the stride equals 2, the feature maps undergo downsampling to compress spatial information. Moreover, the Squeeze-and-Excitation (SE) mechanism, which enhances channel-wise attention, is often integrated with the Ghost bottlenecks module to further boost its effectiveness. The Ghost bottlenecks structure, as detailed in Fig. 3.

Squeeze-and-excitation

Squeeze-and-Excitation (SE)[14] is a representative channel attention mechanism. Its core idea is to model inter-channel dependencies and dynamically recalibrate the importance of each channel feature. It initially applies

global average pooling to compress spatial dimensions and obtain global contextual information. Then, channel-wise relationships are modeled and nonlinear interactions are introduced through two fully connected layers that are combined with nonlinear activation functions. Finally, a Sigmoid activation function outputs the importance weight for each channel. Owing to its simple structure and significant performance improvement, the SE module has been widely integrated into various lightweight networks, such as MobileNetV3 and EfficientNet. The SE attention mechanism can be formulated as:

$$F_{\mathrm{SE}}(X) = X \cdot \sigma\Big(W_2 \cdot \delta\big(W_1 \cdot \mathrm{GAP}(X)\big)\Big) \tag{1}$$

Where X denotes the input feature map, GAP represents global average pooling, δ is the ReLU activation function, σ is the Sigmoid activation function, and W_1 and W_2 are the weights of the two fully connected layers.

CDC and face anti-spoofing

In addition to lightweight architectural components, several task-level studies have explored different perspectives to enhance the generalization, efficiency, and multimodal robustness of face anti-spoofing systems. Early representative work, Yu et al.[6] introduced the concept of Central Difference Convolution(CDC), which models both local intensity and gradient variations to enhance sensitivity to spoof-related texture inconsistencies. This idea inspired subsequent studies emphasizing pixel-level physical cues, including the Pixel-Inconsistency Data Augmentation (PIDA)[15] strategy, which extended this line by explicitly modeling cross-pixel dependency disruptions for fine-grained forgery localization, providing valuable insight into detecting 2D print or replay attacks.

Beyond visual cues, Kong et al. conducted a comprehensive investigation into both digital and physical face spoofing, highlighting the importance of multimodal defense strategies[16]. Meanwhile, they further explored acoustic-based face anti-spoofing by reconstructing 3D facial geometry from inaudible sound waves, demonstrating the potential of combining audio and visual modalities for robust spoofing detection[17]. In addition, Mu et al.[18] proposed a textually guided domain generalization framework, leveraging semantic supervision to align spoof representations across domains and further improve generalization.

Recent model-level innovations have also focused on efficiency and generalization. MoE-FFD[19] proposed a mixture-of-experts architecture combining lightweight adapters and dynamic expert routing to enhance generalization under cross-domain settings. S-Adapter[20] generalized Vision Transformers to FAS by introducing statistical token adapters and style regularization, effectively embedding texture statistics and mitigating domain shift. Yu et al.[21] further re-examined the role of Vision Transformers and Masked Autoencoders in multimodal FAS, emphasizing modality alignment and texture-aware reconstruction for robust fusion. Furthermore, M^{3}FAS[22] designed an accurate and robust multimodal mobile FAS system that fuses RGB and acoustic signals, achieving real-time performance and strong cross-environment robustness on smartphones.

In contrast to these approaches, our proposed DAH-FAS focuses on dynamically-aware heterogeneous feature extraction across RGB, infrared, and depth modalities. By integrating lightweight backbones (MobileNetV2, GhostNet, and ResNet-18) with modality-specific enhancement modules (VA-MSRB, BTE, and BGC-MA), DAH-FAS achieves a balance between generalization capability and multimodal complementarity.

Methods

This section elaborates on the core design of the presented Dynamically-Aware Heterogeneous Face Anti-Spoofing Network. As illustrated in Fig. 4, a heterogeneous multimodal feature extraction framework is established, where MobileNetV2[12], GhostNet[13], and ResNet-18 serve as the backbone networks for the RGB, IR, and depth modalities, respectively. Depth modalities typically encodes richer and more complex geometric structures and is prone to sensor noise or missing-value artifacts. Therefore, the depth branch adopts ResNet-18 to extract

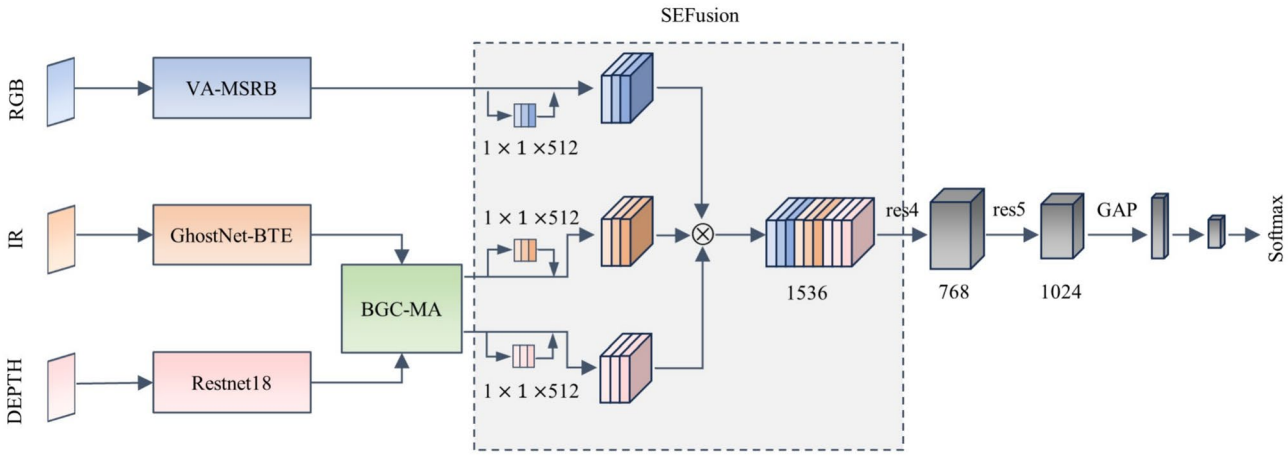

Figure 4. Structure of the DAH-FAS.

more stable and discriminative structural features, ensuring reliable multimodal geometric representation and benefiting the subsequent cross-modal fusion process. This design ensures computational efficiency while enabling modality-specific feature optimization.

RGB branch: variance-adaptive multi-scale residual block (VA-MSRB)

Based on the inverted residual structure of MobileNetV2, we design a tri-branch heterogeneous convolutional module to extract multi-scale deformable features and perform dynamic fusion, as illustrated in Fig. 5. The three branches are defined as follows:

The baseline branch adopts a 3×3 depthwise separable convolution[23] to preserve features within the original receptive field.

$$F_{\mathrm{dsc}} = \mathrm{DSConv}_{3\times3}(F_{\mathrm{rgb}}) \in \mathbb{R}^{B\times C\times H\times W} \tag{2}$$

Where $F_{\mathrm{rgb}} \in \mathbb{R}^{B\times C\times H\times W}$ denotes the input RGB feature map, where B, C, H, and W represent the batch size, number of channels, height, and width of the feature map, respectively. $\mathrm{DSConv}_{3\times3}$ refers to a 3×3 depthwise separable convolution operation.

The deformable branch adopts Deformable ConvNets v2 (DCNv2)[24], which enhances spatial adaptability by dynamically learning both sampling offsets and modulation scalars.

$$F_{\mathrm{dcn}} = \sum_{k=1}^{k} w_k \, F_{\mathrm{rgb}}(p + p_k + \Delta p_k) \tag{3}$$

Where p denotes the reference spatial location on the feature map, p_k is the predefined relative offset of the k-th convolutional kernel element, and Δp_k is a learnable offset vector in the horizontal and vertical directions. w_k controls the contribution of the sampled feature at the sampled location.

The detail branch adopts a two-stage convolutional structure: a channel compression stage for dimensionality reduction, followed by spatial feature extraction. This design enables effective capture of local texture details. The detail branch output is computed as:

$$F_{\mathrm{detail}} = \mathrm{Conv}_{3\times3}\Big(\mathrm{ReLU}\Big(\mathrm{Conv}_{1\times1}(F_{\mathrm{rgb}})\Big)\Big) \tag{4}$$

The input feature variance $\sigma_{\mathrm{var}} = \mathrm{Var}(F_{\mathrm{rgb}})$ is used to generate dynamic fusion weights for the three branches. Specifically, the variance is first globally averaged, followed by a 1×1 convolution and a Softmax layer to obtain a normalized weight vector:

$$w = \mathrm{Softmax}\Big(\mathrm{Conv}_{1\times1}\big(\mathrm{GAP}(\sigma_{\mathrm{var}})\big)\Big), \quad w = [w_1, w_2, w_3] \tag{5}$$

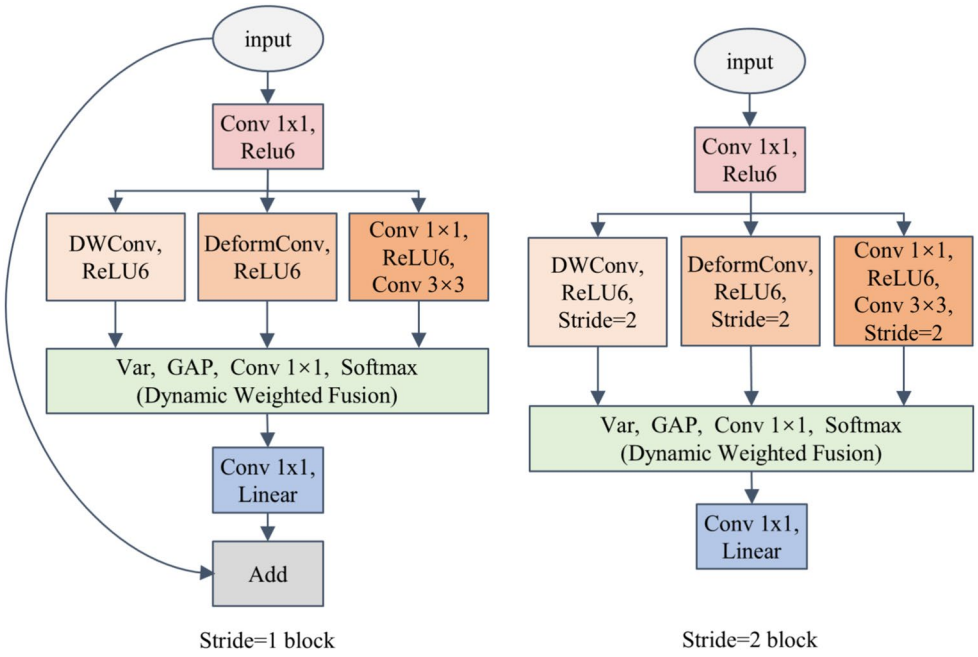

Figure 5. Structure of the VA-MSRB.

where $w = [w_1, w_2, w_3]$ represents the fusion weight vector for the three branches. The magnitude of each weight is determined by the channel-wise variance, which reflects the activation intensity of the corresponding features. The final output feature is computed as:

$$F_{\text{rgb}} = w_1 F_{\text{dsc}} + w_2 F_{\text{dcn}} + w_3 F_{\text{detail}} \tag{6}$$

During the fusion process, regions with high variance–such as edges and texture-rich areas–tend to assign greater weights to the deformable convolution branch, enabling more precise modeling of geometric deformations and improving robustness against complex spoofing attacks. In contrast, for low-variance regions, such as smooth and flat areas, the model emphasizes the detail extraction branch to suppress noise and maintain the stability of feature representation.

IR branch: biothermal enhancer (BTE)

To enhance the model's perception of bio-thermal diffusion patterns, a BioThermal Enhancer (BTE) module is embedded within the GhostNet backbone, as illustrated in Fig. 6. The module extracts thermal gradient cues by applying Sobel filtering[25] to the channel-averaged feature map. The computational procedure is detailed below.

First, let the input feature map be $F_{\text{ir}} \in \mathbb{R}^{C \times H \times W}$, where C denotes the number of channels, and H and W represent the height and width of the feature map, respectively. The channel-wise averaging thermal map $T \in \mathbb{R}^{H \times W}$ is then computed as:

$$T = \frac{1}{C} \sum_{c=1}^{C} F_{\text{ir}}^{(c)} \tag{7}$$

Then, horizontal and vertical gradients are computed by convolving T with Sobel kernels W_x and W_y :

$$G_x = W_x \times T, \quad G_y = W_y \times T \tag{8}$$

The final thermal gradient magnitude is obtained by:

$$G = \sqrt{G_x^2 + G_y^2 + \epsilon} \tag{9}$$

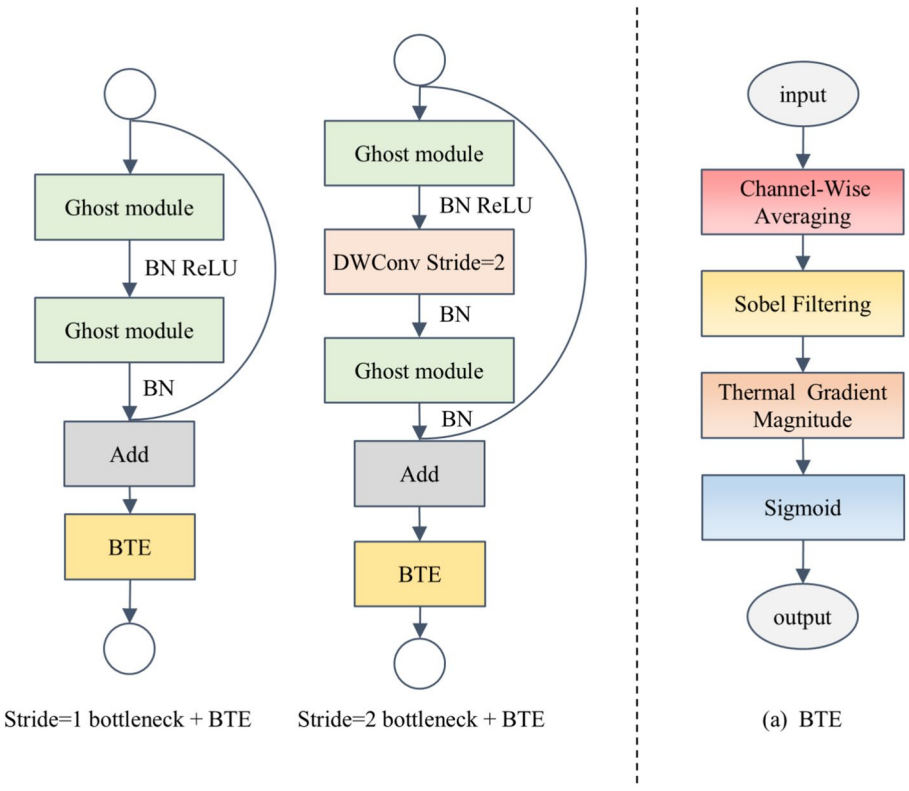

Figure 6. Structure of the BTE. The BTE module consists of four main components: Channel-Wise Averaging, Sobel Filtering, Thermal Gradient Magnitude, and Sigmoid.

Where ϵ is a small constant added to avoid numerical instability during computation. To further emphasize the thermal gradient information, a learnable scaling factor α is introduced. The scaled gradient map is then normalized to the range [0, 1] using the Sigmoid activation function to generate the thermal attention map:

$$A_T = \sigma(\alpha G) \tag{10}$$

Finally, the generated thermal attention A_T is applied to the GhostNet-extracted IR feature map F_{ir} via element-wise multiplication to obtain the enhanced representation F'_{ir}:

$$F'_{\text{ir}} = F_{\text{ir}} \odot A_T \tag{11}$$

Here, F_{ir} represents the original feature output from GhostNet, and $\odot$ denotes element-wise multiplication, which enhances the response in biologically active thermal regions. Focusing on regions with abrupt thermal gradients enables better capture of thermal features, which in turn increases model robustness in face anti-spoofing.

IR and depth branches: bidirectional group cross-modal attention (BGC-MA)

A Bidirectional Group Cross-Modal Attention (BGC-MA) module is constructed between the IR and depth branches. The BGC-MA module aims to enhance the complementary relationship between IR and depth modalities through bidirectional cross-modal feature interaction. It involves channel-wise interaction followed by spatial interaction to capture geometric correspondences between the modalities, as illustrated in Fig. 7. Since BGC-MA relies on accurate geometric cues for reliable IR-Depth alignment, a powerful depth backbone is required; hence, ResNet18 was adopted to provide robust structural representations and enhance the stability of multimodal fusion. The BGC-MA comprises two components: channel-wise interaction and spatial interaction.

The channel-wise interaction employs global average pooling and grouped convolution to generate attention weights, thereby strengthening the channel-level correlations between IR and depth features. In contrast, the spatial interaction integrates the original and enhanced features and utilizes depthwise separable convolution to compute global spatial attention weights, further improving the effectiveness of cross-modal fusion.

First, global average pooling is employed over both the IR feature $F_{\text{ir}} \in \mathbb{R}^{B\times C\times H\times W}$ and the depth feature $F_{\text{depth}} \in \mathbb{R}^{B\times C\times H\times W}$, obtaining global context representations. Then, grouped 1×1 convolution is

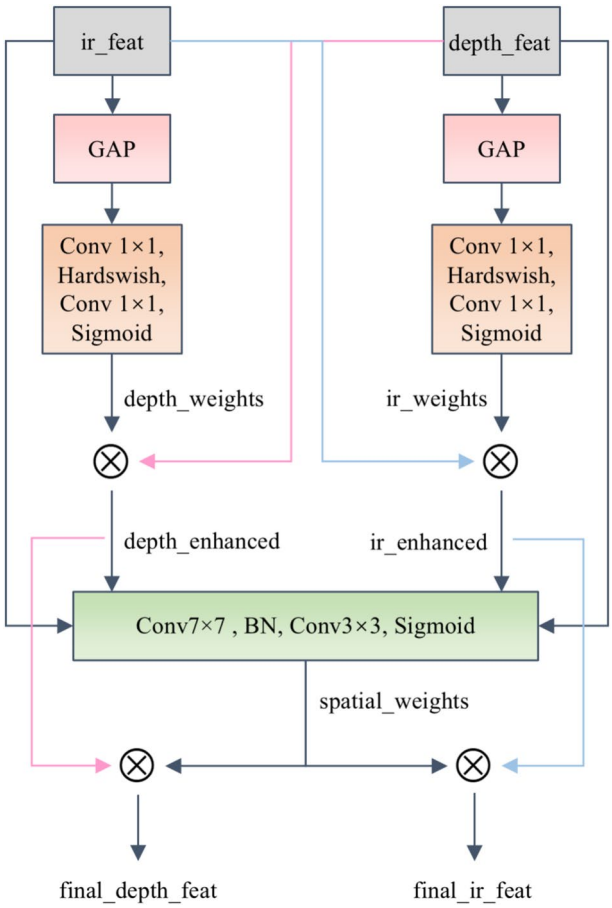

Figure 7. Structure of the BGC-MA.

performed for channel compression, reducing parameter complexity while enhancing the correlation between IR and depth features.The channel interaction is formulated as follows:

$$F_{\text{depth- enh}} = F_{\text{depth}} \odot \sigma\Big(W_2 \cdot \text{Hardswish}\big(W_1 \cdot \text{GAP}(F_{\text{ir}})\big)\Big) \tag{12}$$

$$F_{\text{ir- enh}} = F_{\text{ir}} \odot \sigma\Big(V_2 \cdot \text{Hardswish}\big(V_1 \cdot \text{GAP}(F_{\text{depth}})\big)\Big) \tag{13}$$

Here, W_1, W_2, V_1, and V_2 are the parameters of 1×1 grouped convolution layers, employed for channel compression. The function σ denotes the Sigmoid activation, which is applied to generate attention weights. The operator $\odot$ indicates element-wise multiplication, used to reweight the feature maps. Through this process, the cross-channel interaction between the IR and depth features is enhanced, thereby improving the effectiveness of multimodal fusion.

To capture spatial dependencies between the IR and depth features, we integrate both original and enhanced features through spatial interaction. Specifically, we concatenate the global average pooled features from the original IR and depth branches, along with the enhanced IR and depth outputs as follows:

$$F_{\text{spatial\_input}} = \text{Concat}\big(\text{GAP}(F_{\text{depth}}),\ \text{GAP}(F_{\text{ir}}),\ \text{GAP}(F_{\text{depth\_enh}}),\ \text{GAP}(F_{\text{ir\_enh}})\big) \tag{14}$$

Here, GAP denotes global average pooling. Then, the concatenated global features are fed into two depthwise separable convolutions with kernel size 7×7 to produce the spatial attention map W_{spatial} :

$$W_{\text{spatial}} = \sigma\Big(U_2 \cdot \text{ReLU}\big(U_1 \cdot F_{\text{spatial\_input}}\big)\Big) \tag{15}$$

Where U_1 and U_2 denote two depthwise separable convolutional layers with 7×7 kernels, responsible for extracting spatial context and generating the attention map. W_{spatial} are subsequently utilized to refine the enhanced IR and depth features through element-wise multiplication, yielding the final refined features:

$$F_{\text{depth- final}} = F_{\text{depth- enh}} \odot W_{\text{spatial}} \tag{16}$$

$$F_{\text{ir- final}} = F_{\text{ir- enh}} \odot W_{\text{spatial}} \tag{17}$$

This spatial interaction process effectively performs joint modeling of the spatial information from both IR and depth features, thereby further enhancing the fusion of the two modalities.

Ultimately, the enhanced IR and Depth features, refined through channel-wise and spatial interaction modules, are passed to subsequent fully connected layers for final classification.

Experiment

Datasets

To assess the performance and generalization ability of our approach, we select three widely recognized multimodal face anti-spoofing datasets, namely CASIA-SURF, CASIA-SURF CeFA, and WMCA. The ablation studies are specifically carried out on the CASIA-SURF dataset.

CASIA-SURF

CASIA-SURF[11], constructed by the Institute of Automation at the Chinese Academy of Sciences, serves as a widely used benchmark for multimodal face presentation attack detection. It includes RGB, IR, and depth modalities, comprising 21,000 video clips from 1,000 subjects. Among them, 3,000 are real face videos and 18,000 are spoofed face videos, which involve six different types of attack.

CASIA-SURF CeFA

CASIA-SURF CeFA[26] includes data from 1,607 participants representing three ethnicities: African, East Asian, and Central Asian. It features RGB, depth, and IR modalities, yielding 18,000 samples in total–comprising 4,500 genuine and 13,500 spoof instances. This dataset encompasses a variety of 2D and 3D presentation attacks, such as print-based, replay-based, 3D-printed mask, and silicone mask attacks. By adopting synchronized acquisition and facial region detection, the dataset ensures high quality and consistency, offering a valuable resource for studying face anti-spoofing algorithms under diverse ethnicity, modality, and attack conditions.

WMCA

WMCA[27] is a comprehensive multimodal face anti-spoofing dataset consisting of 1,941 short video recordings from 72 subjects. The data are captured simultaneously from four modalities: RGB, Depth, Infrared (IR), and Thermal, providing rich cross-modal information. The dataset covers seven attack types involving approximately 80 distinct attack tools, including both visible and invisible spoofing types. It adopts the grandtest protocol for evaluating "visible" attacks and the leave-one-out (LOO) protocol for assessing "invisible" attacks, making it one of the most diverse and challenging benchmarks for multimodal face anti-spoofing research.

Experiment preparation

The input images are resized to 112×112. During training, random flipping, rotation, and cropping are applied for data augmentation. All experiments are conducted on an NVIDIA GeForce GTX 4060 GPU, with model

Baseline	MobileNetV2(RGB)	GhostNet(IR)	APCER	BPCER	ACER	TPR@FPR=10E-2	TPR@FPR=10E-3	TPR@FPR=10E-4
+			3.80	1.00	2.40	96.70	81.80	56.80
+	+		4.11	1.08	2.59	96.46	88.19	77.26
+	+	+	3.47	0.55	2.01	97.76	88.94	79.93

Table 1. Ablation results of backbone replacement on the CASIA-SURF dataset (Unit: %).

VA-MSRB	BTE	BGC-MA	APCER	BPCER	ACER	TPR@FPR=10E-2	TPR@FPR=10E-3	TPR@FPR=10E-4
			3.47	0.55	2.01	97.76	88.94	79.93
+			3.31	0.36	1.84	98.46	89.78	72.26
	+		3.19	0.53	1.86	97.33	86.61	79.60
		+	3.32	0.76	2.04	98.61	90.88	75.79
+	+		1.82	1.47	1.65	98.85	90.54	82.35
	+	+	2.02	1.35	1.68	98.66	91.85	80.17
+		+	1.13	1.04	1.08	99.32	94.00	81.39
+	+	+	1.31	0.72	1.01	99.19	95.35	87.32

Table 2. Ablation results of different modules on the CASIA-SURF dataset (Unit: %).

construction, training, and dataset evaluation performed using the PyTorch framework and Python 3.8. The network is optimized using the Adam optimizer, with a cosine annealing learning rate schedule. The initial learning rate is set to 10E-6 , the batch size is 128, and one cosine cycle spans 10 epochs.

In the experimental evaluation, a comprehensive set of metrics is employed to assess the model's performance from different perspectives. For intra-dataset evaluation, we adopt three commonly used indicators: Attack Presentation Classification Error Rate (APCER), Bona Fide Presentation Classification Error Rate (BPCER), and the Average Classification Error Rate (ACER). To further assess the model's recognition capability under varying security levels, we report the True Positive Rate at fixed False Positive Rates of TPR@FPR=10E-2, 10E-3, and 10E-4. For cross-dataset testing, which evaluates generalization to unseen data distributions, we utilize the Half Total Error Rate (HTER) and the Area Under the Curve (AUC). Additionally, we report FLOPs and parameters to evaluate the model's computational efficiency and complexity.

Results and analysis

Ablation analysis

Backbone network selection for modality branches

The impact of different backbone feature extractors within each modality branch is evaluated by adopting SE Fusion[11] as the baseline multimodal fusion framework. Specifically, we replace the backbone networks of the RGB and IR branches to construct three comparative experiments, as shown in Table 1.

In the baseline configuration, all three modality branches adopt ResNet-18 as the backbone network, achieving an ACER of 2.40% and TPR@FPR=10E-4 of 56.80%. After replacing the RGB branch with MobileNetV2, the TPR@FPR=10E-4 improves significantly to 77.26%, while the ACER slightly increases to 2.59%, indicating enhanced spoof detection capability under strict false-positive constraints, albeit with a marginal increase in overall misclassification rate.

When the IR branch is further replaced with GhostNet, the model demonstrates improvements in both metrics, with the ACER reduced to 2.01% and TPR@FPR=10E-4 increased to 79.93%. Therefore, in multimodal face anti-spoofing tasks, adopting structurally heterogeneous backbones for different modalities proves to be an effective and practical design strategy.

Evaluation of the proposed modules and their combinations

After completing the ablation study on backbone replacement, we adopt the best-performing configuration as the new baseline model to further investigate the effectiveness of the proposed key components. We incrementally introduce the VA-MSRB, the BTE, and the BGC-MA to evaluate both their individual contributions and combined impact on face anti-spoofing performance.

The results are presented in Table 2. When introducing each module individually, all three modules contribute to performance improvement to varying degrees. For VA-MSRB alone, the ACER decreased from 2.01% to 1.84% and TPR@FPR=10E-2 increased to 98.46%, demonstrating that VA-MSRB enhances local texture modeling and improves discrimination against attack samples. Separately, when only the BTE module was introduced, the ACER decreased from 2.01% to 1.86%, suggesting that temperature gradient information effectively enhances the identification of bona fide cues. In comparison, the standalone BGC-MA module achieves better performance at TPR@FPR=10E-2 and 10E-3 levels, but demonstrates slightly lower robustness at TPR@FPR=10E-4, with a marginal increase in ACER. This indicates that while BGC-MA facilitates cross-modal alignment, it may require cooperation with other modules to achieve optimal performance under strict low-FPR constraints.

Model	APCER	BPCER	ACER	TPR@FPR=10E-2	TPR@FPR=10E-3	TPR@FPR=10E-4
Halfway fusion[11]	5.60	3.80	4.70	89.1	33.6	17.8
SE fusion[11]	3.80	1.00	2.40	96.7	81.8	56.8
PipeNet[28]	2.08	2.45	2.26	95.90	82.10	56.7
MA-Net[29]	2.40	1.70	2.00	96.00	82.60	58.1
MFF-CNN[30]	2.84	4.12	3.48	–	–	–
Conv-MLP[31]	1.50	1.80	1.60	–	–	–
MF²ShrT[32]	1.60	1.20	1.40	–	–	–
MFViT and MRF[33]	1.50	1.70	1.60	–	–	–
DACA-CNN[34]	2.77	3.13	2.95		–	–
ECA-ICD[35]	5.57	0.65	3.11	–	–	–
Ours	1.31	0.72	1.01	99.19	95.35	87.32

Table 3. Comparison between the proposed method and state-of-the-art methods on CASIA-SURF (Unit: %).

Protocol	APCER	BPCER	ACER
4@1	1.13	1.11	1.12
4@2	2.89	0.51	1.70
4@3	1.63	0.55	1.09
Ours	1.88 ± 0.9	0.72 ± 0.33	1.30 ± 0.34

Table 4. Experimental results of the proposed method on CASIA-SURF CeFA under different protocols (Unit: %).

In the dual-module combination experiments, the overall performance is further improved. Notably, the combination of VA-MSRB and BTE achieves a TPR@FPR=10E-4 of 82.35%, which is significantly higher than that of each module used individually. The combination of VA-MSRB and BGC-MA yields the best ACER performance at 1.08%, while maintaining competitive performance under medium-to-high security evaluation scenarios. Finally, when all three modules are jointly applied, the model achieves the best performance across all metrics, with the ACER reduced to 1.01% and the TPR@FPR=10E-4 increased to 87.32%. These results indicate that the three modules are functionally complementary, and their enhancements in spatial, multi-scale, and cross-modal feature modeling significantly improve the robustness and discriminative capability of the fusion model under different security thresholds.

Performance comparison

To further evaluate the performance of our proposed multimodal anti-spoofing, comparative analyses are carried out against existing approaches on the CASIA-SURF, CASIA-SURF CeFA, and WMCA.

CASIA-SURF

As shown in Table 3, our method achieves the best overall performance among all evaluated approaches, with the ACER reduced to 1.01%, significantly lower than those of DACA-CNN 2.95%, MA-Net 2.00% and MF²ShrT 1.40%, among others. Under stringent security conditions, the proposed method also demonstrates excellent performance, achieving TPR@FPR=10E-2, 10E-3, 10E-4 and of 99.19%, 95.35%, and 87.32%, respectively, ranking among the best-performing models that report these metrics. These results validate the effectiveness of our module designs and fusion strategy in improving both the accuracy and robustness of multimodal face anti-spoofing.

CASIA-SURF CeFA

To further evaluate the generalization ability and stability of the proposed method in cross-ethnicity scenarios, we conducted systematic experiments on three sub-protocols (Protocol 4@1,Protocol 4@2, and Protocol 4@3) of the CASIA-SURF CeFA dataset. As shown in Table 4, the proposed method achieves ACERs of 1.12%, 1.70%, and 1.09%, respectively, demonstrating strong robustness and consistent performance.

In addition, we carried out comparative experiments with several face anti-spoofing methods under the same protocol settings on the CASIA-SURF CeFA, and the results are summarized in Table 5. As shown, the proposed method balances APCER and BPCER, achieving 1.88 ± 0.9% and 0.72 ± 0.33%, respectively, and an ACER of 1.30 ± 0.34%. This demonstrates its competitive performance among all listed methods. Compared to MA-Net, which yields a lower BPCER but suffers from a significantly higher APCER with BPCER = 1.20 ± 1.60%, our method demonstrates a more balanced ability to distinguish between genuine and attack samples. In addition, unlike methods such as FaceBagNet and PSMM-Net, whose ACER standard deviations exceed ±1.5%, the proposed approach consistently maintains lower variance across all metrics, reflecting superior training stability and generalization capability. Overall, these results confirm that the proposed method ensures high

Model	APCER	BPCER	ACER
MA-Net[29]	20.90 ± 6.80	1.20 ± 1.60	11.10 ± 4.40
PSMM-Net[26]	7.80 ± 2.90	5.50 ± 3.00	6.70 ± 2.20
MFViT and MRF[33]	10.67 ± 8.33	2.37 ± 3.24	6.50 ± 2.59
FaceBagNet[36]	5.59 ± 0.20	3.28 ± 2.66	4.59 ± 1.54
PipeNet[28]	3.25 ± 1.98	1.16 ± 1.12	2.21 ± 1.26
Conv-MLP[31]	1.33 ± 0.30	1.42 ± 0.26	1.37 ± 0.27
DACA-CNN[34]	1.63 ± 1.79	2.32 ± 2.13	1.98 ± 0.83
ECA-ICD[35]	4.04 ± 3.25	1.53 ± 0.68	2.74 ± 1.57
DRWT-RDIA[37]	1.76 ± 1.12	0.87 ± 0.42	1.31 ± 0.76
Ours	1.88 ± 0.9	0.72 ± 0.33	1.30 ± 0.34

Table 5. Comparison between the proposed method and state-of-the-art methods on CASIA-SURF CeFA (Unit: %).

Model	HETR↓	AUC↑
FeatherNet[38]	39.22	62.52
FlexModal-FAS[39]	39.22	65.76
FaceBagNet[36]	28.06	78.73
PipeNet[28]	12.90	93.11
ViT-S/16[40]	10.30	95.49
Conv-MLP[31]	10.17	96.09
Ours	9.56	97.36

Table 6. Performance comparison under cross-dataset testing in terms of HTER and AUC (Unit: %).

accuracy while maintaining robust cross-ethnicity recognition performance, thereby demonstrating superior robustness in multimodal face anti-spoofing tasks.

After verifying the effectiveness of our approach using 112×112 input images on CASIA-SURF and CASIA-SURF CeFA, we further evaluated its robustness and generalization under more challenging and practical conditions. Specifically, we conducted experiments using higher-resolution input images resized to 224×224, along with a cross-dataset evaluation where the model was trained on CASIA-SURF CeFA and tested on CASIA-SURF.

The cross-dataset results are summarized in Table 6. Under this challenging evaluation setting that tests model generalization across different data distributions, our method demonstrates highly competitive performance. It achieves a Half Total Error Rate of 9.56% and an Area Under the Curve of 97.36%, positioning it among the top-performing approaches in the comparison.As shown, our method performs comparably to or even slightly better than other competitive models such as ViT-S/16 (HTER=10.30%, AUC=95.49%) and Conv-MLP (HTER=10.17%, AUC=96.09%). These results collectively validate the competitive generalization capability of our method, underscoring its potential for reliable deployment in practical scenarios involving domain shifts.

WMCA

As shown in Table 7, which details the ACER for each "invisible" attack type, our method achieves a highly competitive mean ACER of 5.40%, ranking among top-performing models such as DaR-ViT at 4.79% and DRWT-RDIA at 5.49%. Notably, our approach excels in two challenging attack types, Papermask with 0.2% and Replay with 0.1%, demonstrating its effectiveness against diverse spoofing techniques.

However, under the LOO protocolor assessing "invisible" attack, the performance on the Glasses attack reaches 26.2%, as this attack involves only partial occlusion of the eye region, while the bonafide set includes subjects wearing real glasses, creating strong confounding patterns. In such cases, the global variance–guided fusion in VA-MSRB becomes dominated by genuine facial regions, reducing sensitivity to the small spoofed area. Additionally, BGC-MA receives misleading IR–Depth cues, as the rigid 3D structure of the glasses remains geometrically aligned with surrounding real facial regions, producing an appearance of consistent cross-modal correspondence. In contrast, full-face attacks like Papermask exhibit global IR–Depth inconsistency, enabling more reliable detection.

Although our method does not achieve the best performance on the Glasses attack, it remains highly stable across the remaining attack types. When excluding this attack type, DAH-FAS achieves an average ACER of 1.93% with a standard deviation of ±2.46%, which are substantially lower than that of methods such as DaR-ViT at 4.78±4.00% and DRWT-RDIA at 3.73±4.88%, indicating more consistent and robust performance on the other attack types. Notably, in practical deployment where training sets typically include known attack samples, the model can learn discriminative features, thus mitigating this limitation.

Model	Fakehead	Glasses	Papermask	Rigidmask	Flexiblemask	Replay	Print	Mean ± std
ResNet[7]	2.5	48.3	18.2	15.4	33.2	15.8	3.5	19.56 ± 16.09
MA-Net[29]	2.1	36.7	0.9	9.8	25.3	3.2	0.3	11.18 ± 13.22
FaceBagNet[36]	1.2	13.7	2.3	2.6	31.6	8.5	4.5	9.20 ± 9.99
CMFL[41]	2.5	33.5	1.8	1.7	12.4	1.0	0.7	7.60 ± 11.20
ViTFAS[42]	2.7	15.9	2.3	9.5	2.6	12.4	–	7.56 ± 5.36
Conv-MLP[31]	0.2	32.5	0.9	2.3	12.6	0.8	0.1	7.05 ± 11.16
Dual-Stream[43]	0.4	50.0	0.4	1.4	18.1	1.4	0.7	10.34 ± 18.64
DRWT-RDIA[37]	1.1	18.4	0.5	4.1	12.9	0.6	0.8	5.49 ± 7.23
DaR-ViT[44]	3.89	4.91	1.79	9.31	10.15	3.0	0.54	4.79 ± 3.65
Ours	0.8	26.2	0.2	3.8	6.1	0.1	0.6	5.40 ± 9.44

Table 7. Comparison of different methods under LOO protocol in WMCA(Unit: %).

Model	FLOPs(G)	Parameters(M)
ResNet[7]	7.85	42.5
MCCNN[27]	10.88	37.7
ViTFAS[42]	16.85	85.8
MLP-Mixer[45]	3.30	64.0
MF^2ShrT[32]	–	37.9
DRWT-RDIA[37]	4.22	87.72
Ours	1.78	29.6

Table 8. Comparison results of different model in terms of efficiency.

Efficiency analysis

In addition to the performance comparison, we further evaluate the computational efficiency of different methods. As reported in Table 8, our model requires only 1.78 G FLOPs and 29.6 M parameters, achieving the lowest computational cost among all compared approaches. Even when compared with the MLP-Mixer, which has 3.30 G FLOPs and 64.0 M parameters, DAH-FAS still demonstrates a more favorable trade-off with substantially lower FLOPs and fewer parameters. These results show that the overall model maintains a moderate parameter size and low computational cost, despite integrating modality-specific enhancement modules and cross-modal alignment.

Conclusion

This paper proposes a Dynamically-Aware Heterogeneous Face Anti-Spoofing Network (DAH-FAS), aiming to enhance the performance of existing multimodal liveness detection methods in terms of physical attribute modeling and modality collaboration. To address the representational discrepancies among different modalities, the VA-MSRB module is introduced in RGB branch to strengthen texture feature representation, BTE module is embedded in the IR branch to enhance the perception of bio-thermal cues, and the BGC-MA mechanism is constructed between the IR and depth branches to achieve geometric alignment and efficient information exchange.

Extensive experiments on three challenging datasets, CASIA-SURF, CASIA-SURF CeFA, and WMCA, demonstrate that the proposed method achieves state-of-the-art detection performance under multiple protocols and security levels. For example, it achieves an ACER of only 1.01% and a TPR@FPR=10E-4 of 87.32% on CASIA-SURF. Moreover, it exhibits excellent generalization and cross-ethnicity adaptability across the three sub-protocols of CASIA-SURF CeFA. Furthermore, the model demonstrates strong generalization capability in demanding evaluations, with an HTER of 9.56% in cross-dataset tests and a mean ACER of 5.40% on the WMCA LOO protocol, highlighting its robustness against unseen domains and attack types. These results thoroughly validate the effectiveness and robustness of the proposed modular architecture and fusion strategy. However, certain limitations remain in handling partial-occlusion attacks with confounding patterns, such as the Glasses attack, where genuine accessories in bonafide samples create ambiguous cross-modal cues that challenge both the variance-guided fusion in VA-MSRB and the geometric alignment in BGC-MA. Future work will address these challenges by exploring region-aware feature extraction and context-sensitive fusion mechanisms, while also focusing on model optimization for lightweight deployment and enhanced adaptability in real-world application scenarios.

Data availability

The data presented in this study are openly available in [CASIA-SURF] at [https://sites.google.com/view/face-anti-spoofing-challenge/dataset-download/casia-surfcvpr2019], [CASIA-SURF CeFA] at [https://sites.google.co

m/view/face-anti-spoofing-challenge/dataset-download/casia-surf-cefacvpr2020], and [WMCA] at [https://www.idiap.ch/en/scientific-research/data/wmca], reference number [11, 26, 27].

Received: 4 July 2025; Accepted: 19 December 2025
Published online: 29 December 2025

References

1. Patel, K., Han, H. & Jain, A. K. Secure face unlock: Spoof detection on smartphones. *IEEE Trans. Inf. Forensics Secur.* **11**, 2268–2283 (2016).
2. Maatta, J., Hadid, A. & Pietikainen, M. Face spoofing detection from single images using texture and local shape analysis. *IET Biometrics* **1**, 3–10 (2012).
3. Komulainen, J., Hadid, A. & Pietikäinen, M. Context based face anti-spoofing. In *2013 IEEE Sixth International Conference on Biometrics: Theory, Applications and Systems (BTAS)* 1–8 (2013).
4. Pan, G., Sun, L., Wu, Z. & Lao, S. Eyeblink-based anti-spoofing in face recognition from a generic webcamera. In *2007 IEEE 11th International Conference on Computer Vision* 1–8 (2007).
5. Sun, L., Pan, G., Wu, Z. & Lao, S. Blinking-based live face detection using conditional random fields. In *Advances in Biometrics: International Conference, ICB 2007, Seoul, Korea, August 27-29, 2007. Proceedings* 252–260 (2007).
6. Yu, Z. et al. Searching central difference convolutional networks for face anti-spoofing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* 5295–5305 (2020).
7. He, K., Zhang, X., Ren, S. & Sun, J. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* 770–778 (2016).
8. Xu, Z., Li, S. & Deng, W. Learning temporal features using lstm-cnn architecture for face anti-spoofing. *2015 3rd IAPR asian Conference on Pattern Recognition (ACPR)* 141–145 (2015).
9. Khan, S., Siddique, T. H. M., Ibrahim, M. S., Siddiqui, A. J. & Huang, K. Spatio-temporal deep learning for improved face presentation attack detection. *Knowledge-Based Syst.* **12**, 113059 (2025).
10. George, A. & Marcel, S. Deep pixel-wise binary supervision for face presentation attack detection. In *2019 International Conference on Biometrics (ICB)* 1–8 (2019).
11. Zhang, S. et al. A dataset and benchmark for large-scale multi-modal face anti-spoofing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* 919–928 (2019).
12. Sandler, M., Howard, A., Zhu, M., Zhmoginov, A. & Chen, L.-C. Mobilenetv2: Inverted residuals and linear bottlenecks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* 4510–4520 (2018).
13. Han, K. et al. Ghostnet: More features from cheap operations. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* 1580–1589 (2020).
14. Hu, J., Shen, L. & Sun, G. Squeeze-and-excitation networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* 7132–7141 (2018).
15. Kong, C. et al. Pixel-inconsistency modeling for image manipulation localization. *IEEE Trans. Pattern Anal. Mach. Intell.* **47**, 4455–4472 (2025).
16. Kong, C., Wang, S. & Li, H. Digital and physical face attacks: Reviewing and one step further. *ArXiv* https://doi.org/10.48550/arXiv.2209.14692 (2022).
17. Kong, C., Zheng, K., Wang, S., Rocha, A. & Li, H. Beyond the pixel world: A novel acoustic-based face anti-spoofing system for smartphones. *IEEE Trans. Inf. Forensics Secur.* **17**, 3238–3253 (2022).
18. Mu, L. et al. Teg-dg: textually guided domain generalization for face anti-spoofing. *arXiv* https://doi.org/10.48550/arXiv.2311.18420 (2023).
19. Kong, C. et al. Moe-ffd: Mixture of experts for generalized and parameter-efficient face forgery detection. *IEEE Trans. Depend. Secur. Comput.* https://doi.org/10.1109/TDSC.2025.3604443 (2025).
20. Cai, R. et al. S-adapter: Generalizing vision transformer for face anti-spoofing with statistical tokens. *IEEE Trans. Inf. Forensics Secur.* **19**, 8385–8397 (2024).
21. Yu, Z. et al. Rethinking vision transformer and masked autoencoder in multimodal face anti-spoofing. *Int. J. Comput. Vision* **132**, 5217–5238 (2024).
22. Kong, C. et al. M^3FAS: An accurate and robust multimodal mobile face anti-spoofing system. *IEEE Trans. Depend. Secur. Comput.* **21**, 5650–5666 (2024).
23. Chollet, F. Xception: Deep learning with depthwise separable convolutions. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* 1251–1258 (2017).
24. Zhu, X., Hu, H., Lin, S. & Dai, J. Deformable convnets v2: More deformable, better results. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* 9308–9316 (2019).
25. Aoun, A., Masadeh, M. & Tahar, S. On the design of approximate sobel filter. In *2022 International Conference on Microelectronics (ICM)* 102–106 (2022).
26. Liu, A. et al. Casia-surf cefa: A benchmark for multi-modal cross-ethnicity face anti-spoofing. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision* 1179–1187 (2021).
27. George, A. et al. Biometric face presentation attack detection with multi-channel convolutional neural network. *IEEE Trans. Inf. Forensics Secur.* **15**, 42–55 (2020).
28. Yang, Q. et al. Pipenet: Selective modal pipeline of fusion network for multi-modal face anti-spoofing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops* 644–645 (2020).
29. Liu, A. et al. Face anti-spoofing via adversarial cross-modality translation. *IEEE Trans. Inf. Forensics Secur.* **16**, 2759–2772 (2021).
30. Zou, W., Zhang, D. & Lee, D.-J. A new multi-feature fusion based convolutional neural network for facial expression recognition. *Appl. Intell.* **52**, 2918–2929 (2022).
31. Wang, W., Wen, F., Zheng, H., Ying, R. & Liu, P. Conv-mlp: A convolution and mlp mixed model for multimodal face anti-spoofing. *IEEE Trans. Inf. Forensics Secur.* **17**, 2284–2297 (2022).
32. Antil, A. & Dhiman, C. Mf2shrt: multimodal feature fusion using shared layered transformer for face anti-spoofing. *ACM Trans. Multimed. Comput.Commun. Appl.* **20**, 1–21 (2024).
33. Li, Z. et al. A multimodal face antispoofing method based on multifeature vision transformer and multirank fusion. *Concurrency Comput. Pract. Exp.* **35**, e7824 (2023).
34. Li, N., Weng, Z., Liu, F., Li, Z. & Wang, W. Dual-path adaptive channel attention network based on feature constraints for face anti-spoofing. *IEEE Access* **13**, 22855–22867 (2025).
35. Li, Y., Sun, W., Li, Z. & Guo, X. Face anti-spoofing based on adaptive channel enhancement and intra-class constraint. *J. Imaging* **11**, 116 (2025).
36. Shen, T., Huang, Y. & Tong, Z. Facebagnet: Bag-of-local-features model for multi-modal face anti-spoofing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops* 0–0 (2019).
37. Sun, R., Wang, F., Yu, X., Gao, X. & Zhang, X. Robust multimodal face anti-spoofing via frequency-domain feature refinement and aggregation. *Pattern Recognition Lett.* **197**, 31–36 (2025).

38. Zhang, P. et al. Feathernets: Convolutional neural networks as light as feather for face anti-spoofing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops* (2019).
39. Yu, Z. et al. Flexible-modal face anti-spoofing: A benchmark. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops* 6346–6351 (2023).
40. Dosovitskiy, A. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv* https://doi.org/10.48550/arXiv.2010.11929 (2020).
41. George, A. & Marcel, S. Cross modal focal loss for rgbd face anti-spoofing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* 7882–7891 (2021).
42. George, A. & Marcel, S. On the effectiveness of vision transformers for zero-shot face anti-spoofing. In *2021 IEEE International Joint Conference on Biometrics (IJCB)* 1–8 (2021).
43. Deng, P., Ge, C., Qiao, X., Wei, H. & Sun, Y. Attention-aware dual-stream network for multimodal face anti-spoofing. *IEEE Trans. Inf. Forensics Secur.* **18**, 4258–4271 (2023).
44. Tong, G., Shao, H. & Yan, X. Enhanced depth-guided rgb image feature fusion for robust face anti-spoofing. *Arab. J. Sci. Eng.* https://doi.org/10.1007/s13369-025-10632-w (2025).
45. Tolstikhin, I. O. et al. Mlp-mixer: An all-mlp architecture for vision. *Adv. Neural Inf. Processing Syst.* **34**, 24261–24272 (2021).

Author contributions
Conceptualization, Y.Y. and L.S.; methodology, Y.Y.; software, Y.Y.; validation, Y.Y. and L.S.; formal analysis, Y.Z. and X.C.; investigation, Y.Y.; data curation, Y.Z. and X.C.; writing–original draft preparation, Y.Y.; writing–review and editing, L.S.; visualization, Y.Y.; supervision, L.S. All authors have read and agreed to the published version of the manuscript.

Funding
This research did not receive any specific funding.

Declarations

Competing interests
The authors declare no competing interests.

Additional information
Correspondence and requests for materials should be addressed to L.S.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.