FA-Net: Dual-Branch YOLOv5 with Triplet-Aware Dynamic Filtering and Bilateral Coordinate Fusion for RGB-D Hand Gestures

LuYang Wang[1], Jing Qi[1,*], You Zhou[1,*]
[1]School of Cyber Security and Computer,Hebei University, Qiyi East Road, Baoding 071002, Hebei China

***Abstract*—Hand posture recognition using RGB-D cameras offers intuitive interaction but remains challenged by RGB instability (e.g., illumination changes, skin-like backgrounds) and depth noise. Thus, to address these challenges, we propose a Fusion-Attention Network (FA-Net), a multi-stage dynamic attention fusion network built on a dual-branch YOLOv5 backbone. Firstly, our proposed Triplet-Aware Dynamic Filtering (TADF) module generates data-driven convolutional kernels via channel, width and spatial interactions to precisely align and enhance RGB and depth features. Secondly, our developed a Bilateral Coordinate Fusion (BCF) module boosts cross-modal feature exchange and integrates spatial coordinates into channel attention to suppress background clutter and emphasize key gesture regions. FA-Net adaptively adjusts modality weights and performs scale-aware fusion, fully exploiting RGB-depth complementarity. On CUG, NTU and a self-built robot dataset, FA-Net outperforms state-of-the-art fusion methods, achieving notable gains in accuracy and robustness. Deployment on a wheeled robot with an Intel RealSense camera confirms high-precision gesture recognition and reliable motion command generation, demonstrating FA-Net's practicality for real-world human-robot interaction.**

***Index Terms*—RGB-D hand gesture recognition, multimodal fusion,human-computer interaction.**

I. Introduction

Natural human–computer interaction (HCI) modalities are increasingly studied, especially for contactless visual interfaces. Gesture recognition, being intuitive and efficient, has become a key technology for applications such as brain–machine interfaces, augmented reality, and service robotics [1]. Traditional HCI uses devices like keyboards and mice, which differ from natural communication and impose physical constraints, limiting real-time flexibility in dynamic scenarios.

Existing methods are divided into dynamic and static approaches by temporal characteristics. Dynamic gesture recognition models spatiotemporal video sequences to capture motion trajectories and velocities, but requires high-frame-rate inputs and temporal networks, leading to high computational cost and redundant-frame interference. Static recognition analyzes hand shape in single frames, reducing data redundancy and complexity, making it suitable for fixed-point commands and symbolic gestures [2].

This work is supported in part by Scientific Research Foundation of Hebei University for Distinguished Young Scholars (Grant No.521100221081), and in part by Demonstration lessons for graduate students of Hebei Province (Grant No. KCJSX2025009).

* Corresponding authors: qijingalice@163.com, zhouyou@hbu.edu.cn

Early static methods using RGB images alone suffer from instability under illumination changes and background texture interference, which hinders hand contour and joint-detail extraction. Depth-only approaches provide strong geometric cues but are prone to noise and missing surface details. With the proliferation of RGB-D sensors (e.g., Kinect, Intel RealSense), researchers have explored fusing color and depth data to improve recognition performance [3]. YOLOv5, in particular, is popular as a backbone for real-time deployment on mobile platforms due to its balance of accuracy and inference speed.

Multimodal fusion has since become a major research focus. Chen. [4] proposed a bidirectional cross-modality propagation and gated aggregation mechanism for RGB-D semantic segmentation, suppressing depth noise and misalignment; however, such schemes often estimate fusion weights from globally pooled descriptors or coarse channel statistics, which can discard fine-grained spatial structure and fail to correct local modality misalignments. Ren. [5] introduced a multi-scale weighted histogram of contour directions (MSWHCD) for first-person static gestures, capturing discriminative geometric details to improve classification robustness, but descriptor-based approaches of this kind typically rely on handcrafted or histogram representations that are not end-to-end learnable and are brittle under severe occlusion and scale variation. EMSANet [6] integrated multi-task attention branches for parallel semantic and instance segmentation, balancing real-time efficiency with precision on mobile platforms, yet it and similar architectures often treat fusion as a late aggregation step and lack tight cross-modal alignment at multiple feature scales, reducing robustness when one modality is locally unreliable.

Further advances include BCF-DPM [7], which extracts bag-of-contour fragments from depth projections to enhance resilience against occlusion and deformation; while effective, such pipelines depend heavily on depth preprocessing and tend to discard complementary RGB texture cues. A 3D-CNN-based system for fingertip detection and gesture recognition under complex backgrounds [8] shows strong spatiotemporal modeling but at high computational cost, limiting real-time deployment on resource-constrained platforms. RFNet [9] embeds attention-based feature complementarity for small-object detection in autonomous driving, yet many attention schemes (e.g., those relying on channel pooling) erode positional pre-

cision when applied naively.

FA-Net adopts a multi-stage dynamic attention fusion strategy to address limitations of existing RGB-D fusion methods. First, the proposed Triplet-Aware Dynamic Filtering (TADF) generates data-driven dynamic convolutional kernels via cross-dimensional interactions (channel–height, channel–width, and spatial), preserving fine-grained local structures and correcting local misalignments often smoothed by global pooling. Second, Bilateral Information Exchange (BIE) enables bidirectional cross-layer feature propagation, allowing RGB and depth features to mutually refine before weight estimation, thus avoiding late static fusion and producing more context-aware importance weights. Third, Coordinate Attention (CA) embeds positional information into channel attention, suppressing background interference and improving localization under occlusion and clutter. Applied across multiple stages with learnable dynamic weighting, these components enable scale-aware fusion, adaptive modality weighting, and reduced information loss while maintaining deployment efficiency. The main contributions of this article are shown as follows:

- We design the TADF module, which generates dynamic convolutional kernels through cross-modal interactions along channel-height, channel-width, and spatial dimensions to precisely align and enhance RGB and depth features, overcoming information loss and limited spatial interaction in traditional channel-attention methods.
- We design the BCF module, comprising the BIE for explicit bidirectional channel interactions to strengthen cross-modal information flow and CA to fuse spatial coordinates into channel attention, suppress background clutter, and dynamically weight features for improved discriminability and robustness.
- Extensive experiments on CUG, NTU, and a self-built RGB-D dataset demonstrate that FA-Net significantly outperforms state-of-the-art fusion methods in accuracy and robustness, and real-world deployment on a wheeled manipulator with an Intel RealSense camera confirms high-precision gesture recognition and reliable motion command generation.

II. Methodology

In this section, we describe FA-Net and its TADF, BIE, and CA modules. The fusion pipeline begins with multi-scale feature extraction and gating, applies TADF independently to RGB and depth for fine-grained enhancement, then uses BCF to weight and fuse modality-specific features.

Fig. 1 depicts FA-Net's integration of fusion modules into two parallel YOLOv5 branches, enabling multi-stage RGB-D feature processing. Fusion modules are placed immediately after each CSP Bottleneck and its C3 block, enhancing cross-modal aggregation. The C3 module, a core YOLOv5 component, excels at feature extraction and compression. By appending our fusion modules after each C3 block, FA-Net increases adaptability to diverse RGB and depth cues, improving RGB-D learning accuracy and robustness.

In this article, RGB and depth channels are identical during processing, so only RGB channels are shown in the following formulas.

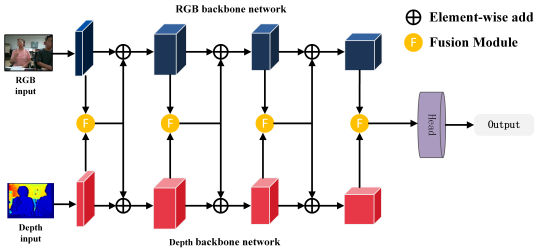

Fig. 1. An overview of the framework

A. TADF Module

Effective cross-modal fusion is crucial for RGB-D gesture recognition. Conventional channel-attention mechanisms suffer from over-reliance on channel-wise compression, information loss, and lack spatial interactions for modeling geometric relations across modalities. To address this, we propose TADF for adaptive dynamic fusion via cross-dimension interactions. Fig. 2 shows TADF as the core fusion hub with three parallel attention paths modeling channel, spatial, and depth interdependencies, enhancing the discriminative power of fused features.

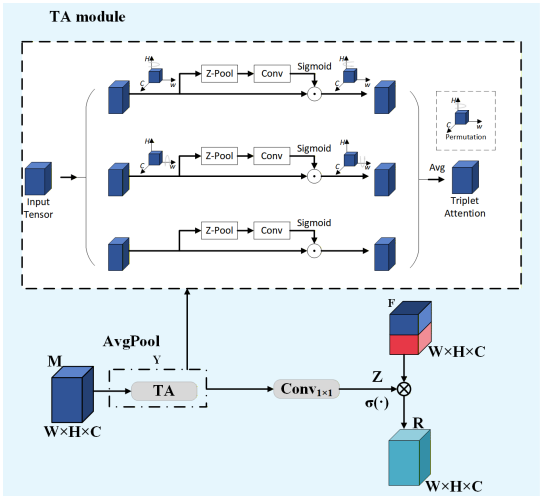

Fig. 2. Details of the TADF module

Given input tensor $\chi \in R^{C\times H\times W}$, global adaptive average pooling produces a descriptor Y fed into a Triple Attention (TA) submodule with three parallel branches:

X_h is rotated 90° counter-clockwise around height, yielding *(W×H×C)*. Z-Pool aggregates channel max and average into *(2×H×C)*, then a $k\times k$ convolution + BatchNorm produces *(1×H×C)* attention weights. Sigmoid modulates the rotated tensor, which is rotated back to *(C×H×W)*.

Similarly, X_w is rotated 90° around width to *(H×C×W)*. Z-Pool compresses to *(2×C×W)*; a $k\times k$ conv + BatchNorm

produces *(1×C×W)* attention map. After Sigmoid, weights are applied and the tensor rotated back.

For spatial attention, Z-Pool χ along channels into *(2×H×W)*; a *k×k* conv + BatchNorm yields *(1×H×W)* map; after Sigmoid, weights multiply χ.

The three branches output tensors of shape *(C×H×W)*, fused by element-wise averaging. Triplet Attention highlights salient regions along channel-height, channel-width, and spatial dimensions. Formally:

$$y = 1/3\left(\overline{\hat{\chi}_1\sigma\left(\psi_1(\hat{\chi}_1^*)\right)} + \overline{\hat{\chi}_2\sigma\left(\psi_2(\hat{\chi}_2^*)\right)} + \chi\sigma\left(\psi_3(\hat{\chi}_3)\right)\right) \tag{1}$$

Here, $\sigma(\cdot)$ denotes Sigmoid; ψ_1, ψ_2, ψ_3 are 2D convolutions in each branch. Simplified:

$$y = 1/3\left(\overline{\hat{\chi}_1\omega_1} + \overline{\hat{\chi}_2\omega_2} + \chi\omega_3\right) = 1/3\left(\overline{y_1} + \overline{y_2} + y_3\right) \tag{2}$$

$\omega_1, \omega_2, \omega_3$ denote attention weights; $\overline{y_1}$, $\overline{y_2}$, y_3 are outputs of channel-height, channel-width, and spatial branches. Rotations align dimensions to *(C×H×W)*.

Static features are extracted via TADF; a 1×1 conv generates weight map Z, followed by Sigmoid:

$$Z = \sigma(Conv_{1\times1}(TA(y))) \tag{3}$$

Finally, Z multiplies fused input features F element-wise:

$$O = Z \odot F \tag{4}$$

Overall TADF formulation:

$$R = \sigma(Conv_{1\times1}(TA(AvgPool(M)))) \odot F \tag{5}$$

TADF achieves cross-modal alignment, adaptive weighting, and discriminative enhancement, reducing RGB-depth discrepancies, ensuring fusion consistency, and emphasizing critical features in complex scenes.

B. BCF Module

For a specific task, not all features carry equal semantic importance. Enabling the model to adaptively adjust feature weights based on input content enhances representational capacity and discriminative power. Motivated by this idea, we propose a weighted feature enhancement module, the BCF module, as illustrated in Fig. 3.

The BCF employs a dual-branch design integrating complementary attention along distinct dimensions. It comprises a BIE submodule, which aligns RGB and depth features for efficient fusion, and a CA submodule, which boosts spatial sensitivity, suppresses background noise, and focuses on critical gesture regions.

Below, we first detail the BIE and CA modules, then outline the BCF workflow.

1) BIE Submodule: Traditional RGB-D fusion methods process RGB and depth features separately, combining them with global-pooling-derived weights, which discard fine-grained details and reduce precision. We propose a BIE module that explicitly exchanges feature maps between RGB and depth before weight estimation: RGB cues inform the depth branch, and depth cues inform the RGB branch. This bidirectional interaction strengthens cross-modal fusion, exploits shared features, and yields more accurate RGB-D importance weights, as shown in Fig. 4.

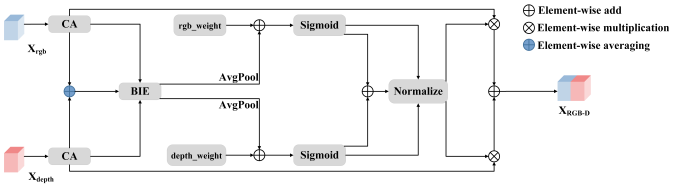

Fig. 3. Details of the BCF module

In BIE notation, positive polarity H_p^l denotes the RGB-branch feature map at layer l, and negative polarity H_n^l denotes the depth-branch feature map. The cross-layer interaction representation (CIR), H_{int}^l, fuses both modalities at the same layer with cross-layer context for refinement. For an event at time step t, H_p^l and H_n^l represent Layer-1 features, while $H_{int,t}^l$ injects hierarchical context during propagation.

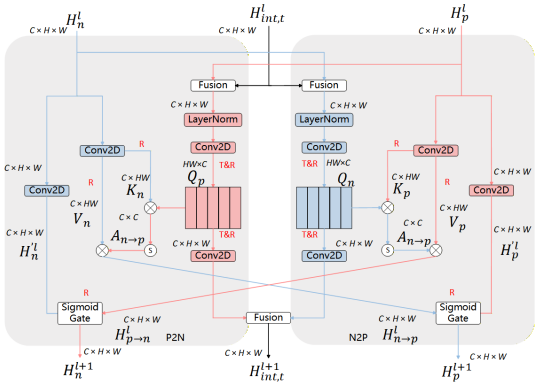

Fig. 4. Details of the BIE module

We detail propagation from the positive to the negative event ($H_p^l \rightarrow H_n^{l+1}$); the reverse follows analogously.

The CIR of the current layer, $H_{int,t}^l$, is normalized and transformed via LayerNorm and a 1×1 convolution. The updated CIR is fused with H_n^l through another 1×1 convolution to produce the query $Q_n^l \in R^{M\times HW}$, where M denotes the number of global structures.

The positive-event representation H_p^l is processed by convolutional layers and reshaped to produce the key $K_p^l \in R^{M\times HW}$ and value $V_p^l \in R^{M\times HW}$. We compute similarity between Q_n^l and K_p^l, apply Softmax along the horizontal dimension, and obtain the attention matrix $A_{p\rightarrow n}^l \in R^{M\times M}$, reflecting relations between global semantic units. Following Scaled Dot-Product Attention, a scaling factor $\sqrt{HW}$ is used.

We generate forward interaction features and fuse them. The attention weights perform a weighted sum over V_p^l, yielding:

$$H_{p\rightarrow n}^{'l} = Conv_{1\times1}\left(Softmax\left(\frac{Q_n^l(K_p^l)^\top}{\sqrt{HW}}\right)V_p^l\right) \tag{6}$$

A residual branch of convolutional layers captures the local structure of the original negative event, yielding $H_n^{'l}$. The negative-event representation at $l+1$ is updated via gated fusion:

$$Z_n^l = \sigma\Big(W_{n1}H_n^{'l} + W_{n2}H_{p\to n}^l\Big) \tag{7}$$

$$H_n^{l+1} = Z_n^l \odot H_n^l + \big(1 - Z_n^l\big) \odot H_{p\to n}^l \tag{8}$$

Here, $Z_n^l \in R^{C\times H\times W}$ denotes gating weights. Analogously, negative-event information propagates back to the positive event.

To enhance cross-layer modeling, the CIR updates by concatenating Q_p^l and Q_n^l along the channel dimension, followed by a convolution:

$$H_{int,t}^{l+1} \;=\; Conv\left(\langle Q_p^l, Q_n^l\rangle\right) + H_{int,t}^l \tag{9}$$

Here, $\langle\cdot\rangle$ denotes channel-wise concatenation.

In summary, the BIE module assigns varying importance to features, enhancing those more relevant to the task. This adaptive weighting avoids information dilution common in methods treating all features equally.

2) CA Submodule: Current gesture recognition systems face two challenges in complex environments: 1) spatial ambiguity in cluttered backgrounds, and 2) structural fragmentation due to articulated motion. Existing attention modules, relying on global pooling, fail under occlusion where fine-grained features are overwhelmed by background noise.

To address these issues, we propose the Coordinate Attention (CA) mechanism inspired by Hou et al. [10], illustrated in Fig. 5. By decomposing spatial dimensions, CA enables collaborative refinement across spatial and channel dimensions, integrating positional information to enhance spatial awareness and robustness to background interference.

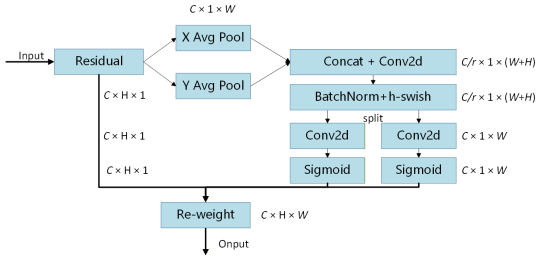

Fig. 5. Details of the CA module

Given an input tensor $\chi \in R^{C\times H\times W}$, average pooling is applied along horizontal and vertical directions to obtain $X_{avg}^x \in R^{C\times H\times 1}$ and $X_{avg}^y \in R^{C\times 1\times W}$. Taking X_{avg}^x as an example, it captures long-range dependencies along the horizontal direction while preserving vertical positional information, facilitating accurate localization: $X_{avg}^x = AvgPool_x(X), X_{avg}^y = AvgPool_y(X)$.

To capture global context while preserving positional information, X_{avg}^x and X_{avg}^y serve as key components. We further introduce Coordinate Attention Generation, which is lightweight, exploits positional information to highlight regions of interest, and models inter-channel dependencies.

Specifically, the two features are concatenated along the spatial dimension to obtain $X_{avg}^{xy} \in R^{C\times 1\times(W\times H)}$, then passed through a 1×1 convolution (Conv-BatchNorm-H-Swish) to generate $F \in R^{C/r\times 1\times(W\times H)}$, where r is a channel reduction ratio. The feature F encodes directional information along both spatial axes.

$$F = H - Swish(BatchNorm(Conv(X_{avg}^{xy}))) \tag{10}$$

Next, F is split into $F_x \in R^{C/r\times 1\times H}$ and $F_y \in R^{C/r\times 1\times W}$. Each is processed by a 1×1 convolution followed by Sigmoid, producing attention weights $W_x \in R^{C\times H\times 1}$ and $W_y \in R^{C\times 1\times W}$: $F_x, F_y = Split(F)$:

$$W_x = sigmoid(Conv(F_x)) \tag{11}$$

The input tensor is modulated as $Y = X \odot W_X \odot W_y$. Unlike SENet [11], which focuses on channel relationships, CA also encodes spatial cues via row and column attention, enabling precise localization and improving recognition accuracy.

In summary, the CA module captures inter-channel dependencies and positional cues by encoding spatial orientation with cross-channel interactions, enabling precise attention to target regions.

3) Workflow of BCF module: First, CA is applied to RGB and depth features to enhance spatial-channel attention: $F_{rgb\_ca} = CA(X_{rgb})$.

Here, $F_{rgb\_ca} \in R^{B\times C\times H\times W}$ retains the same shape as its input [B,C,H,W]. The two modality-specific CA outputs are then averaged to obtain an intermediate feature:

$$X_{int} = 1/2(F_{rgb\_ca} + F_{depth\_ca}) \in R^{B\times C\times H\times W} \tag{12}$$

This intermediate representation provides a preliminary fused global context for cross-modal interaction. The three features F_{rgb_ca}, F_{depth_ca}, and X_{int} are then fed into the BIE module:

$$(F_{rgb\_bie}, F_{depth\_bie}) = BIE(F_{rgb\_ca}, F_{depth\_ca}, X_{int}) \tag{13}$$

Here, F_{rgb_bie} and F_{depth_bie} represent modality-interactive features of the RGB and depth branches.

Next, global average pooling is applied to F_{rgb_bie} and F_{depth_bie} to obtain scalar vectors. These are added to learnable weighting parameters α_{rgb} and α_{depth}, followed by Sigmoid activation, and normalized to produce fusion weights:

$$S_{rgb} = \sigma(AvgPool(F_{rgb\_bie}) + \alpha_{rgb}) \tag{14}$$

Here, α_{rgb} and α_{depth} are learnable dynamic parameters, with outputs in (0,1). After Sigmoid, tensors $S_{rgb}, S_{depth} \in R^{B\times 1\times 1\times 1}$ are obtained.

Following normalization, we derive the final attention weights:

$$\omega_{rgb} = \frac{S_{rgb}}{S_{rgb} + S_{depth}} \tag{15}$$

Ensuring $\omega_{rgb} + \omega_{depth} = 1$, the weights are properly scaled. This Sigmoid-plus-normalization strategy enables flexible and learnable bias adjustment. The final fusion weights are $\omega_{rgb}, \omega_{depth} \in R^{\times 1\times 1\times 1}$.

Finally, the normalized weights are element-wise multiplied with F_{rgb_ca} and F_{depth_ca}, then summed to obtain the fused feature:

$$X_{fused} = \omega_{rgb} \odot F_{rgb\_ca} + \omega_{depth} \odot F_{depth\_ca} \tag{16}$$

Here, the final output is $X_{fused} \in R^{B\times C\times H\times W}$.

Accordingly, the overall fusion equation is:

$$X_{fused} = \frac{\sigma\big(AvgPool(F_{rgb\_bie}) + \alpha_{rgb}\big)}{S_{rgb} + S_{depth}} \odot F_{rgb\_ca} + \frac{\sigma\big(AvgPool(F_{depth\_bie}) + \alpha_{depth}\big)}{S_{rgb} + S_{depth}} \odot F_{depth\_ca} \tag{17}$$

The BCF module fuses coordinate attention with cross-modal interactions to guide modality-specific perception, assign adaptive weights, and enhance features robustly. By jointly modeling RGB and depth with dynamic weighting, it improves cross-modal consistency, reduces redundancy and imbalance, and enhances discriminability of critical regions across diverse environments.

C. FUSION Module

The core goal of fusion in deep learning is to extract and integrate multi-source features into a more discriminative global representation. As shown in Fig. 6, our fusion module begins with multi-scale feature extraction and a gating mechanism that dynamically adjusts responses at different resolutions, boosting hand-pose perception, suppressing noise, and highlighting key details.

Next, TADF refines RGB and depth features for improved intra-modality discrimination. Finally, BCF models importance disparities between RGB and depth branches and fuses them via adaptive weights, ensuring focus on task-relevant regions and improving robustness and performance.

Specifically, the aggregation module operates as follows.

To capture spatial information at multiple levels, RGB and depth features are processed using convolutions with three receptive fields, namely 1×1, 3×3, and 5×5, denoted as $Conv_{1\times1}$, $Conv_{3\times3}$, and $Conv_{5\times5}$.

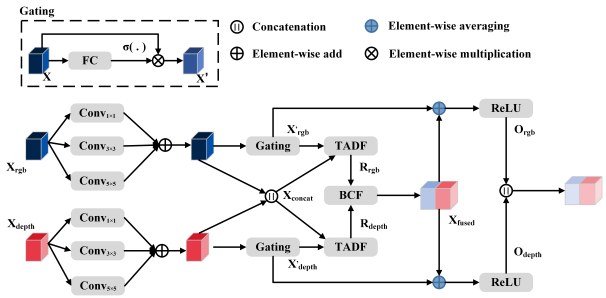

Fig. 6. Details of the fusion module

The resulting multi-scale features are summed to obtain the fused representation:

$$X_{rgb} = Conv_{1\times1}(X_{rgb}) + Conv_{3\times3}(X_{rgb}) + Conv_{5\times5}(X_{rgb}) \tag{18}$$

Next, a gating mechanism performs saliency-aware adjustment. A convolutional layer extracts a gating signal for each feature, normalized to [0,1] using a Sigmoid to produce attention weights.

These signals are applied via element-wise multiplication, enhancing salient information while suppressing redundancy, thus improving discriminability and robustness.

$$G_{rgb} = \sigma\left(Conv_{gate}(X_{rgb})\right), \; X'_{rgb} = G_{rgb} \odot X_{rgb} \tag{19}$$

Here, $Conv_{gate}$ denotes the convolution generating gating signals.

The enhanced RGB and depth features are then fed into TADF. This module performs adaptive enhancement and constructs a joint representation (X_{concat}) by concatenation, improving cross-modal consistency.

Through cross-dimensional attention, TADF dynamically adjusts weight distribution between RGB and depth, enabling the model to highlight key regions under complex backgrounds or interference, improving fusion stability.

$$R_{rgb} = TADF(X'_{rgb}, X_{concat}) \tag{20}$$

Here, X_{concat} denotes concatenation along the channel dimension. After TADF, features are passed into BCF for weighted fusion.

The BCF module performs dynamic weighting, enabling adaptive focus across modalities based on relative importance: $X_{concat} = BCF(R_{rgb}, R_{depth})$.

This enhances the flexibility and task adaptiveness of fusion, improving the final representation.

To balance modality-specific information and fused features, BCF introduces a residual connection after fusion. The original and fused features are combined by element-wise addition and averaging to produce the final output:

$$O = ReLU\left(\frac{X_{rgb} + X_{fused}}{2} + \frac{X_{depth} + X_{fused}}{2}\right) \tag{21}$$

This design preserves original RGB and depth information while incorporating enhanced cross-modal representations.

In summary, the proposed framework leverages three components: multi-scale extraction with gating, TADF for channel-spatial interaction modeling, and BCF for adaptive weighted fusion.

Together, these modules exploit complementary RGB-depth cues to enhance feature expressiveness and integration. The fusion strategy improves semantic understanding and adapts modality attention in complex or noisy scenes, boosting gesture-recognition robustness and accuracy.

III. EXPERIMENTS

A. Datasets

This section covers public datasets and a self-constructed dataset. To ensure randomness and representativeness, all datasets were split into training, validation, and testing sets using a randomized strategy.

1) Public Datasets: For objectivity and reproducibility, FA-Net was evaluated on two public RGB-D hand datasets alongside our own. The CUG dataset [12] contains unconstrained samples from 27 participants, with one to seven subjects (up to eight hands) per frame. We re-annotated it for classification, yielding 1,055 training, 351 validation, and 351 test images. The NTU dataset [13], [14] comprises 10 gesture classes, each performed in 10 variations by 10 subjects (1,000 samples). After adding detailed labels, 800 images were used for training, and the remaining 200 were split evenly into 100-image validation and test sets.

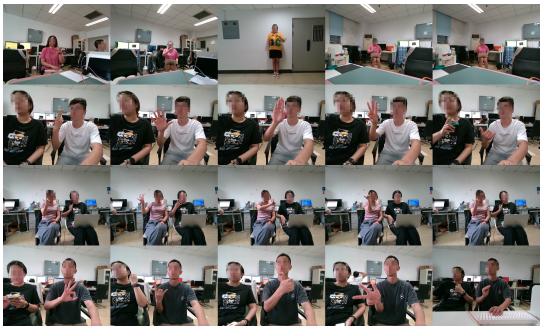

Fig. 7. Seventeen gestures from our self-built dataset, together with examples of interfering gestures.

2) Self-Built Datasets: For human–robot interaction in reconnaissance and search-and-rescue scenarios, we built an experimental platform using a wheeled mobile manipulator and collected a new RGB-D hand gesture dataset. A gesture-to-robot-command mapping table covering 17 distinct gestures was designed. To enhance diversity, samples were captured at distances from 0.4m to 4m, with each gesture including at least 500 RGB-D image pairs at 640×480 resolution.

The dataset was partitioned into 4,250 training, 2,550 validation, and 1,700 test samples. All gestures were performed by four participants in a lab, and identical gestures executed with either hand were assigned to the same class. Fig. 7 shows example images.

B. Implementation Details

Training was conducted on an AMD EPYC 7302 16-core processor and NVIDIA RTX A4000 GPU using PyTorch with batch size 16. Models were trained for 250 epochs on CUG, 60 on NTU, and 100 on the self-constructed dataset. Optimization used SGD with an initial learning rate of 0.01 (decayed to 0.002 via One-Cycle schedule), weight decay 0.0005, and early stopping to prevent overfitting. Data augmentation (random translation, scale jitter, horizontal flipping) was synchronously applied to both RGB and depth modalities to improve diversity and robustness in complex scenarios.

C. Comparative Experiments

To ensure fair and consistent comparison, we benchmarked FA–Net against state-of-the-art RGB–D hand detection and fusion methods. All competing modules were integrated into a uniform dual-branch YOLOv5 backbone to maintain architectural consistency. During training, both training and validation sets were used for model fitting, while the test set was reserved for final evaluation, ensuring generalization and avoiding overfitting or dataset-specific tuning.

As shown in Tables I, FA-Net demonstrates superior performance across multiple metrics. On the CUG dataset, it achieves an AP_{50} of 92.2% and an $AP_{50:95}$ of 73.5%, outperforming existing approaches and validating the effectiveness of our method. Similarly, on the NTU and our self-built datasets, FA-Net attains AP scores of 93.3%/63.3% and 93.1%/74.2%, respectively. Overall, these results highlight FA-Net's advantage in RGB-D fusion and hand-pose recognition, offering a more precise and robust solution for real-world applications.

D. Ablation Experiments

This study evaluates the synergistic effects of three attention mechanisms, TADF, BIE, and CA, within the fusion module. Built on a parallel-branch YOLOv5 framework, we start with a baseline model that directly fuses RGB and depth features, then incrementally introduce TADF, BIE, and CA to assess their individual and combined contributions.

The results show that each module yields clear improvements. Specifically, TADF, by modeling channel-spatial interactions, enhances feature weighting; BIE reinforces complementary fusion between RGB and depth streams, boosting cross-modal integration; and CA strengthens sensitivity to spatial cues, suppressing background noise and sharpening responses to critical gesture regions. Quantitatively, both AP_{50} and $AP_{50:95}$ show gains over the baseline, indicating consistently high detection accuracy across IoU thresholds.

The comparative results in Table II further confirm the positive impact of each module, particularly the combined use of TADF and BIE, demonstrating the effectiveness of the proposed FA-Net fusion module in enhancing RGB-D hand-gesture recognition.

TABLE I
COMPARISON OF RECOGNITION ACCURACY (%) ON THREE DATASETS

Methods	**Input**	AP_{50}			$AP_{50:95}$		
		CUG	**NTU**	**Self-Built**	**CUG**	**NTU**	**Self-Built**
Baseline	RGB-D	86.4	89.2	86.4	68.2	59.8	68.2
CPNet [15]	RGB-D	87.1	77.2	90.2	68.8	50.2	71.6
RD3DNet [16]	RGB-D	86.9	83.3	87.6	66.6	49.7	69.5
HIDANet [17]	RGB-D	86.2	86.9	88.7	64.4	61.9	69.3
DWANet [18]	RGB-D	89.4	76.1	89.2	71.2	47.1	71.5
DVSONet [19]	RGB-D	84.9	78.9	91.4	66.4	52.5	72.3
RDVSNet [20]	RGB-D	88.1	92.8	91.4	69.4	61.5	72.2
Ours	RGB-D	**92.2**	**93.3**	**93.1**	**73.5**	**63.3**	**74.2**

E. Application In A Laboratory Robot

To assess FA-Net's practical utility, we deployed it on a wheeled mobile manipulator featuring an Ackermann chassis and a three-degree-of-freedom robotic arm for on-site testing. Using an Intel RealSense D455, the robot captures gesture

TABLE II
ABLATION EXPERIMENTS FOR EACH MODULE IN THE FUSION MODULE ON THE CUG DATASET

Methods	AP_{50}	$AP_{50:95}$
Baseline	86.4	68.2
Baseline + TADF	90.2	71.6
Baseline + BIE	87.6	69.5
Baseline + CA	88.7	69.3
Baseline + TADF + BIE	90.5	71.9
Baseline + TADF + CA	91.4	72.3
Baseline + BIE + CA	91.4	72.2
Baseline + TADF + BIE + CA	**92.2**	**73.5**

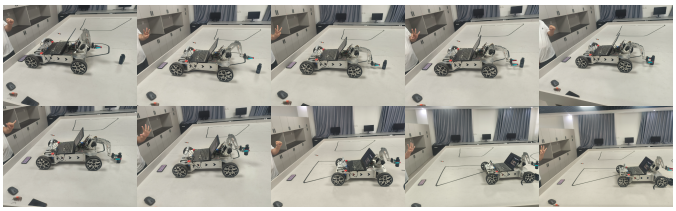

Fig. 8. FA-Net enabling real-time hand-gesture control of a robotic vehicle in a hazardous scenario.

inputs and executes commands via FA-Net. In a hazardous scenario, it successfully controlled the robot to place a simulated hand grenade into an explosive containment device in real time, as shown in Fig. 8. Field tests verified FA-Net's high accuracy in translating gestures into precise motion commands, highlighting its reliability and effectiveness in real-world robotic control.

IV. CONCLUSIONS

The proposed FA-Net architecture addresses key challenges in RGB-D static gesture recognition, namely feature alignment, cross-modal interaction, and spatial localization, through integrating three core modules: TADF, BIE, and CA. Comparative experiments on the CUG, NTU, and self-constructed datasets show that FA-Net consistently outperforms mainstream fusion methods, achieving an average AP_{50} improvement of about 5 percentage points. Ablation studies further confirm that multi-stage, multi-module synergy contributes to optimal performance.

More importantly, FA-Net has been deployed on a wheeled mobile manipulator platform, where it uses a RealSense camera to decode gesture commands in real time and translate them into precise motion directives, validating its effectiveness and robustness in complex environments. Future work may explore lightweight model design and cross-task transfer learning to expand FA-Net's applicability across broader human–robot interaction scenarios.

REFERENCES

[1] A. Carfì and F. Mastrogiovanni, "Gesture-based human–machine interaction: Taxonomy, problem definition, and analysis," *IEEE Transactions on Cybernetics*, vol. 53, no. 1, pp. 497–513, 2021.

[2] Y. Li, X. Wang, W. Liu, and B. Feng, "Deep attention network for joint hand gesture localization and recognition using static rgb-d images," *Information Sciences*, vol. 441, pp. 66–78, 2018.

[3] F. Mueller, D. Mehta, O. Sotnychenko, S. Sridhar, D. Casas, and C. Theobalt, "Real-time hand tracking under occlusion from an egocentric rgb-d sensor," in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 1154–1163.

[4] X. Chen, K.-Y. Lin, J. Wang, W. Wu, C. Qian, H. Li, and G. Zeng, "Bi-directional cross-modality feature propagation with separation-and-aggregation gate for rgb-d semantic segmentation," in *European conference on computer vision*. Springer, 2020, pp. 561–577.

[5] R. Yiyi, X. Xie, G. Li, and Z. Wang, "Hand gesture recognition with multi-scale weighted histogram of contour direction (mswhcd) normalization for wearable applications," *IEEE Transactions on Circuits and Systems for Video Technology*, 2016.

[6] D. Seichter, S. B. Fischedick, M. Köhler, and H.-M. Groß, "Efficient multi-task rgb-d scene analysis for indoor environments," in *2022 International joint conference on neural networks (IJCNN)*. IEEE, 2022, pp. 1–10.

[7] B. Feng, F. He, X. Wang, Y. Wu, H. Wang, S. Yi, and W. Liu, "Depth-projection-map-based bag of contour fragments for robust hand gesture recognition," *IEEE Transactions on Human-Machine Systems*, vol. 47, no. 4, pp. 511–523, 2016.

[8] D.-S. Tran, N.-H. Ho, H.-J. Yang, E.-T. Baek, S.-H. Kim, and G. Lee, "Real-time hand gesture spotting and recognition using rgb-d camera and 3d convolutional neural network," *Applied Sciences*, vol. 10, no. 2, p. 722, 2020.

[9] L. Sun, K. Yang, X. Hu, W. Hu, and K. Wang, "Real-time fusion network for rgb-d semantic segmentation incorporating unexpected obstacle detection for road-driving images," *IEEE robotics and automation letters*, vol. 5, no. 4, pp. 5558–5565, 2020.

[10] Q. Hou, D. Zhou, and J. Feng, "Coordinate attention for efficient mobile network design," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2021, pp. 13 713–13 722.

[11] J. Hu, L. Shen, and G. Sun, "Squeeze-and-excitation networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 7132–7141.

[12] C. Xu, J. Zhou, W. Cai, Y. Jiang, Y. Li, and Y. Liu, "Robust 3d hand detection from a single rgb-d image in unconstrained environments," *Sensors*, vol. 20, no. 21, p. 6360, 2020.

[13] Z. Ren, J. Yuan, and Z. Zhang, "Robust hand gesture recognition based on finger-earth mover's distance with a commodity depth camera," in *Proceedings of the 19th ACM international conference on Multimedia*, 2011, pp. 1093–1096.

[14] Z. Ren, J. Yuan, J. Meng, and Z. Zhang, "Robust part-based hand gesture recognition using kinect sensor," *IEEE transactions on multimedia*, vol. 15, no. 5, pp. 1110–1120, 2013.

[15] X. Hu, F. Sun, J. Sun, F. Wang, and H. Li, "Cross-modal fusion and progressive decoding network for rgb-d salient object detection," *International Journal of Computer Vision*, vol. 132, no. 8, pp. 3067–3085, 2024.

[16] Q. Chen, Z. Zhang, Y. Lu, K. Fu, and Q. Zhao, "3-d convolutional neural networks for rgb-d salient object detection and beyond," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 35, no. 3, pp. 4309–4323, 2022.

[17] Z. Wu, G. Allibert, F. Meriaudeau, C. Ma, and C. Demonceaux, "Hidanet: Rgb-d salient object detection via hierarchical depth awareness," *IEEE Transactions on Image Processing*, vol. 32, pp. 2160–2173, 2023.

[18] Z. Tu, X. Qian, and W. Zhou, "Distilled wave-aware network: Efficient rgb-d co-salient object detection via wave learning and three-stage distillation," *Information Sciences*, vol. 700, p. 121855, 2025.

[19] J. Li, W. Ji, S. Wang, W. Li *et al.*, "Dvsod: Rgb-d video salient object detection," *Advances in Neural Information Processing Systems*, vol. 36, pp. 8774–8787, 2023.

[20] A. Mou, Y. Lu, J. He, D. Min, K. Fu, and Q. Zhao, "Salient object detection in rgb-d videos," *IEEE Transactions on Image Processing*, 2024.