Mission-Level UAV Intent Recognition Using Image-Trajectory Fusion

TANG Li[1,2*], *WANG Zijie*[2,3], *TU Pengyue*[3]

1. School of Automobile and Transportation, Xihua University, Chengdu 610039, P. R. China; 2. Intelligent Air-Ground Fusion Vehicle and Control Engineering Research Center of the Ministry of Education, Xihua University, Chengdu 610039, P. R. China; 3. School of Aerospace and Intelligent Equipment, Xihua University, Chengdu 610039, P. R. China

(Received 22 May 2026; revised 23 July 2026; accepted 30 July 2026)

Abstract: Low-altitude traffic management requires timely inference of both unmanned aerial vehicle (UAV) platform type and likely mission intent. Declared flight plans can provide intent information for cooperative operations, but non-cooperative, abnormal or partially observed UAVs require inference from observable evidence. The existing surveillance methods mainly detect or track UAVs, classify platform type, or predict trajectories. They therefore remain limited when different missions share similar short-term motion and appearance alone lacks temporal context. This study introduces MCFNet, a multimodal framework that combines an external-view RGB image with an 8 s trajectory history for mission-level UAV intent recognition. Its bidirectional cross-attention fusion links local visual regions with trajectory time steps before classification, allowing cues such as payload state or spray effects to be interpreted with the relevant maneuver segments. Modality dropout reduces single-modality shortcuts, and a consistency loss penalizes invalid type-intent pairs. On 7 290 simulated samples covering three UAV types and 15 intents, MCFNet achieves 89.63%±0.99% intent accuracy, 89.41%±1.00% type-intent joint accuracy and 91.33%±1.05% macro-F_1, outperforming the strongest late-fusion baseline by 3.04% and achieving a slightly higher mean intent accuracy than the multimodal bottleneck Transformer (MBT) token-level fusion baseline. The results provide an observation-based route to earlier warning and differentiated decision support for low-altitude traffic management.

Key words: UAV intent recognition; image-trajectory fusion; low-altitude airspace; trajectory disambiguation; domain-aware constraints

CLC number: V279 **Document code:** A **Article ID:** 1005-1120(2026)04-0529-17

0 Introduction

Low-altitude airspace is becoming a dense and heterogeneous traffic environment. Unmanned aerial vehicles (UAVs) now support inspection, logistics, agriculture, emergency response and security-sensitive operations. Recent work on urban air mobility (UAM) traffic-flow recovery, low-altitude air-transport management and trajectory conformance alarms shows that future operations will need continuous monitoring, conformance assessment and decision support, not only vehicle detection[1-3]. Recent reviews of low-altitude surveillance further emphasize that future low-altitude operations will involve cooperative and non-cooperative targets, heterogeneous sensing conditions and increasing autonomy, requiring integrated surveillance and fusion-based monitoring capabilities[4]. In this setting, mission-level UAV intent is a decision-relevant traffic state. It helps determine whether an observed vehicle should be monitored, prioritized, rerouted, isolated or treated as a potential violation. Cooperative systems can use declared flight plans or assigned procedures. Non-cooperative, abnormal or partially observed operations instead require intent to be in-

*Corresponding author, E-mail address: tangli@xhu.edu.cn.

How to cite this article: TANG Li, WANG Zijie, TU Pengyue. Mission-level UAV intent recognition using image-trajectory fusion[J]. Transactions of Nanjing University of Aeronautics and Astronautics, 2026, 43(4): 529-545.

http://dx.doi.org/10.16356/j.1005-1120.2026.04.004

ferred from observable evidence.

The existing UAV surveillance methods provide an essential but incomplete perception layer. Recent reviews and RF/WiFi- or vision-based anti-UAV studies show rapid progress in detection, classification and tracking across radar, radio-frequency, acoustic and image modalities[5-7]. These methods mainly estimate presence, location, trajectory or platform type. They therefore remain exposed to trajectory ambiguity. Different missions can produce similar short-term motion patterns, and a single image cannot represent temporal maneuver evolution. Mission-level intent recognition consequently requires joint reasoning about what the UAV looks like and how it moves.

For low-altitude traffic management, UAV type and mission intent answer different but coupled questions. Type recognition indicates what a UAV is, including its platform capability and feasible mission set. Intent recognition indicates what the UAV is likely to do, including whether its current behavior is routine, safety-critical or potentially anomalous. Inferring both is therefore necessary for operationally meaningful surveillance, because a behavior that is plausible for one UAV type may be invalid or suspicious for another.

UAV intent recognition is formulated here as a multimodal traffic-surveillance problem. Images provide airframe, payload and operating-state cues, whereas trajectories provide recent motion dynamics. A simple combination of these signals is insufficient. Conventional late fusion compresses each modality into a global descriptor before merging, discarding the spatial and temporal indices needed for cross-modal interpretation. For trajectory-ambiguous missions, a model should associate local visual cues with relevant motion segments before classification, rather than combine already pooled features at the decision stage.

To address this problem, MCFNet is proposed as a multimodal cross-fusion network for mission-level UAV intent recognition. MCFNet uses an ImageNet-pretrained ResNet-18 branch to encode visual evidence and a Transformer branch to encode an 8 s trajectory history. Its core component, bidirectional cross-attention fusion (BCAF), aligns image spatial tokens and trajectory temporal tokens in both directions. This design enables motion-conditioned visual selection and appearance-conditioned trajectory interpretation. Modality-level random dropout reduces single-modality shortcut learning, and a domain-aware consistency loss penalizes physically implausible type-intent combinations.

MCFNet is evaluated on a simulation dataset built with Unreal Engine 4 (UE4) and AirSim. The dataset contains 7 290 image-trajectory samples covering three UAV types, 15 mission-level intent categories and 15 photorealistic UAV models. Several intent groups deliberately share overlapping trajectory envelopes, making the benchmark suitable for testing whether multimodal fusion can resolve operational ambiguity. Comparisons with unimodal, late-fusion variants and multimodal bottleneck Transformer (MBT)[8] baselines show that MCFNet improves over conventional baselines and remains competitive with a strong token-level fusion baseline, especially on trajectory-confusable groups.

The main contributions are as follows.

(1) Mission-level UAV intent recognition is formulated as an observation-based perception task for low-altitude traffic management, especially when declared intent is unavailable or unreliable.

(2) MCFNet is introduced to align image spatial tokens and trajectory temporal tokens through bidirectional cross-attention, enabling fine-grained cross-modal reasoning.

(3) Modality-level random dropout is combined with a domain-aware consistency loss to improve robustness and encode mission-platform feasibility.

(4) A controlled image-trajectory benchmark is constructed with 7 290 samples, three UAV types and 15 intent categories. MCFNet achieves $89.63\%\pm0.99\%$ intent accuracy, $89.41\%\pm1.00\%$ joint accuracy and $91.33\%\pm1.05\%$ macro-F_1 across three random seeds.

The remainder of this paper is organized as follows. Section 1 reviews related work on low-altitude surveillance, intent recognition and multimodal fusion. Section 2 describes the MCFNet architec-

ture and training objective. Section 3 presents the dataset, experimental results, ablation and robustness analyses, and a detailed analysis of confusable intent groups. Section 4 concludes the paper and discusses future work.

1 Related Work

1.1 Low-altitude surveillance

Recent low-altitude airspace studies have shifted from conceptual UAM feasibility to traffic density, congestion and service-level control. Aarts et al.[9] quantified constrained urban-airspace capacity under separation and speed assumptions. Safadi et al.[10] constructed macroscopic fundamental diagrams for low-altitude air city transport, and Cummings et al.[11] linked advanced air mobility (AAM) density, conflict rate, speed loss and throughput through 4 D fundamental diagrams. Structured-airspace work, including the sky-highway model of Fu et al.[12], further indicates that scalable operations require continuous traffic-state interpretation rather than isolated vehicle monitoring. In parallel, unmanned traffic management (UTM) research has formalized automated services for identification, surveillance, monitoring, deconfliction, communication and decision making[13]. Dai et al.[14] made the intent problem explicit by treating conformance monitoring as a test of whether unmanned aircraft adhere to declared operational intent. That assumption is inadequate for non-cooperative or abnormal UAVs, where intent must be inferred from observations.

UAV sensing research supplies the detection layer but not the semantic layer needed for intent-aware management. Rahman et al.[15] surveyed RF, visual, acoustic, radar and hybrid UAV detection and classification, while Leng et al.[16] identified scale, orientation, spatial and class imbalance as persistent obstacles in aerial object detection. Alshaer et al.[17] combined deep visual detection with Kalman filtering for tracking. These studies improve presence, location and platform-type estimation. Their outputs, however, remain geometric or categorical and do not infer the mission being executed.

1.2 Intent recognition

Recent transportation-intent models increasingly treat intent as a context-conditioned latent variable rather than a post hoc label. Jiang et al.[18] proposed an intention-aware interactive Transformerfor dense vehicle trajectory prediction. Huang et al.[19] modelled lane-change intention through driver-behavior and traffic-context graphs, and Ling et al.[20] used spatial-temporal attention graph convolution for fast pedestrian crossing-intention prediction. Xu et al.[21] further showed that pedestrian crossing intent depended on pedestrian-vehicle interaction. These studies support intent-aware perception, but their assumptions are tied to road topology, ground-agent interaction and planar motion.

UAV-specific intent research is closer to the present task but remains dominated by kinematic or control-state evidence. Perrusquía et al.[22] introduced a control-physics-informed framework that classified drone intentions, including mapping, point-to-point, package-delivery and perimeter flights, and predicts future occupied airspace. Trajectory-based behavior analysis has also been studied in air-traffic contexts. Wu et al.[23] used ADS-B trajectory data to define trajectory redundancy and deviation metrics for detecting abnormal flight behaviors. Zhang et al.[24] and Shi et al.[25] showed that flight-state recognition improved UAV trajectory prediction, while Cao et al.[26] embedded maneuver-intention probabilities into an attention-based Bi-GRU predictor. These methods show that behavioral abstraction can improve motion forecasting. However, they primarily infer flight mode or future trajectory from dynamic states. They do not resolve mission ambiguity when visually distinct operations produce near-identical trajectories, such as spraying versus field survey or delivery flight versus zone intrusion.

1.3 Multimodal fusion

Recent multimodal learning has shifted from simple feature aggregation to structured interaction. Zhang et al.[27] reviewed vehicle-information fusion and emphasized the value of motion and control signals as complements to visual sensing. Xu et al.[28]

surveyed transformer-based multimodal learning, and Gu et al.[29] showed that camera-LiDAR Transformerfusion can improve semantic segmentation under adverse driving conditions. These studies motivate attention-based interaction, but their target domains are vehicle perception, vision-language alignment or camera-LiDAR scene understanding. The UAV image-trajectory problem has a different structure. Visual data contain local payload or operation cues, whereas trajectories contain temporally distributed motion evidence.

Late fusion remains a necessary baseline because it is modular and competitive when modalities provide complementary global descriptors. For mission-level UAV intent recognition, however, global-vector concatenation or addition removes the spatial and temporal indices needed to relate a small visual cue, such as spray particles or cargo state, to a particular maneuver segment. Recurrent alternatives such as LSTM-concat preserve temporal order during trajectory encoding but still expose only a compressed trajectory descriptor to the fusion layer. The research gap is therefore not UAV detection, type recognition or trajectory prediction alone. It is fine-grained, domain-constrained image-trajectory fusion for mission-level intent recognition when different tasks intentionally share similar kinematic signatures.

2 The Proposed Method

MCFNet formulates UAV intent recognition as a multimodal surveillance task. In low-altitude traffic monitoring, an external observer can receive two complementary streams: A visual snapshot of platform appearance and payload state, and a short trajectory history of recent maneuvering behavior. The model associates these streams before classification, rather than treating them as independent descriptors that meet only at the decision stage. As shown in Fig.1, MCFNet contains four components: An image branch for spatial visual encoding, a trajectory branch for temporal motion encoding, a BCAF module for token-level modality interaction, and two heads for UAV type and mission-level intent prediction. In Fig.1, MLP represents the multilayer perceptron.

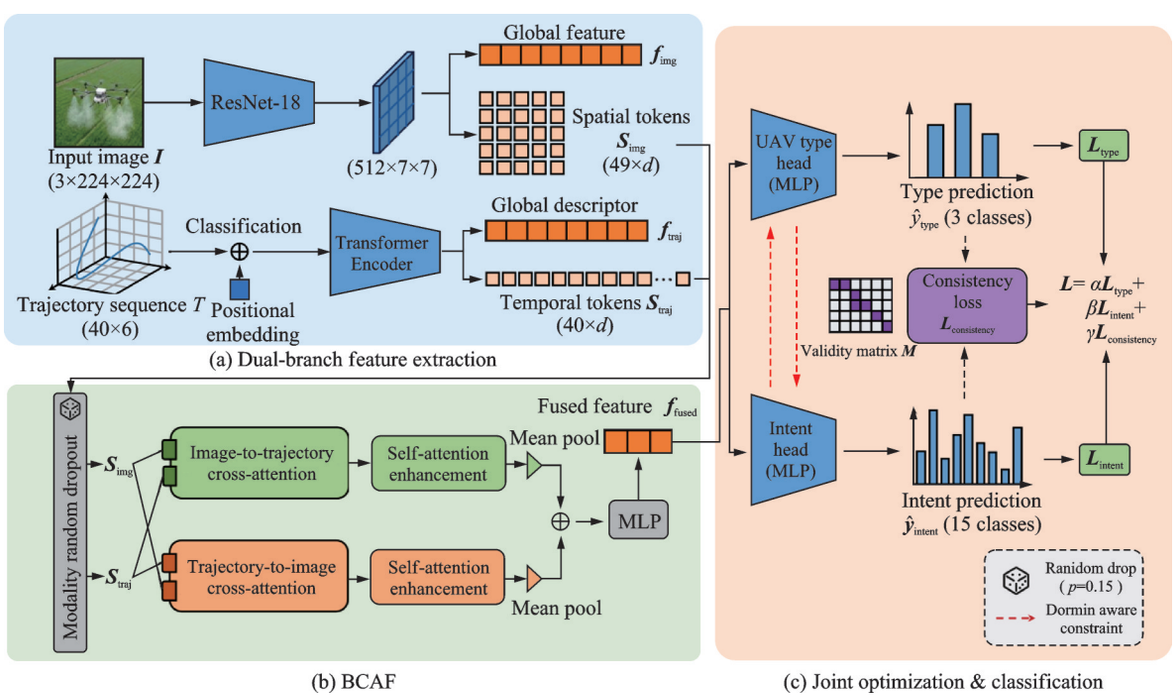

Fig.1 Overall architecture of MCFNet

Given an RGB image $I \in R^{3\times 224\times 224}$ and a trajectory sequence $T \in R^{40\times 6}$, the objective is to infer both UAV type and mission-level intent. The trajectory contains 40 frames sampled at 5 Hz, giving an 8 s observation window. Each frame contains three displacement components relative to

the first observed position and three velocity components, denoted by $[\Delta x, \Delta y, \Delta z, v_x, v_y, v_z]$. This representation captures accumulated spatial evolution and local motion dynamics. Both are needed to distinguish operational maneuvers that can share similar instantaneous speeds but differ in temporal pattern. The output space consists of three UAV types and 15 mission-level intent categories. The type labels are aerial-photography UAVs (T1), plant-protection UAVs (T2) and logistics UAVs (T3). Table 1 summarizes the intent taxonomy.

Table 1 Mission-level UAV intent taxonomy

Category	ID	Intent name	Description	Valid type
Common	I01	Takeoff	Vertical ascent from ground	T1, T2, T3
	I02	Landing approach	Controlled descent to landing	T1, T2, T3
	I03	Idle hover	Stationary hovering in place	T1, T2, T3
	I04	Emergency return	High-speed direct return to home	T1, T2, T3
Aerial photography	I05	Panoramic scan	In-place rotation with camera sweep	T1
	I06	Tracking shot	Following a ground target	T1
	I07	Orbit shot	Circular flight around a point	T1
Plant protection	I08	Spraying	Zigzag coverage with liquid dispersal	T2
	I09	Field survey	Grid-pattern inspection flight	T2
Logistics	I10	Delivery flight	Point-to-point cargo transport	T3
	I11	Pickup/Delivery hover	Low-altitude hover at pickup/drop-off	T3
	I12	Route following	Waypoint-based corridor flight	T3
Anomalous	I13	Zone intrusion	Direct flight toward restricted area	T1, T2, T3
	I14	Evasive maneuvering	Erratic heading and velocity changes	T1, T2, T3
	I15	Suspicious loitering	Bounded random walk in small area	T1, T2, T3

2.1 Dual-branch feature extraction

The image branch uses an ImageNet-pretrained ResNet-18 backbone[30]. Global average pooling and fully connected layers are removed, retaining the last residual feature map. For a 224×224 input, this map has size $512 \times 7 \times 7$. Global average pooling followed by linear projection produces a $d = 256$ visual descriptor. The same 7×7 map is reshaped into 49 spatial tokens, each projected from 512 to 256 dimensions. These tokens preserve coarse image layout, allowing later fusion layers to attend to local cues, such as spray particles below a plant-protection UAV or a cargo box on a delivery platform. Fig.2 shows visual variation.

The trajectory branch uses a Transformer encoder[31] to model temporal dependencies. The 40-frame trajectory is mapped from six dimensions to $d = 256$ by a linear embedding layer. A learnable classification token is prepended, and learnable positional embeddings encode temporal order. The resulting 41-token sequence is processed by two Transformer encoder layers with four attention heads, a feed-forward dimension of 512, GELU activation and dropout of 0.1. The output classification token provides a global motion descriptor, while the remaining 40 temporal tokens retain time-step-level context. Thus, global descriptors summarize each modality, whereas spatial and temporal tokens are retained for cross-modal interaction. Fig.3 shows why token-level information is needed for intent groups with overlapping kinematic signatures.

2.2 Bidirectional cross-attention fusion

BCAF is the central component of MCFNet. It addresses a limitation of conventional late fusion in UAV traffic surveillance. Global image and trajectory descriptors are concatenated only after their spatial and temporal structures have been compressed. A classifier can therefore observe that visual and kinematic information are both present, but it cannot explicitly learn cross-modal correspondences. For example, it cannot learn that a small visual region becomes important because a specific maneuver is

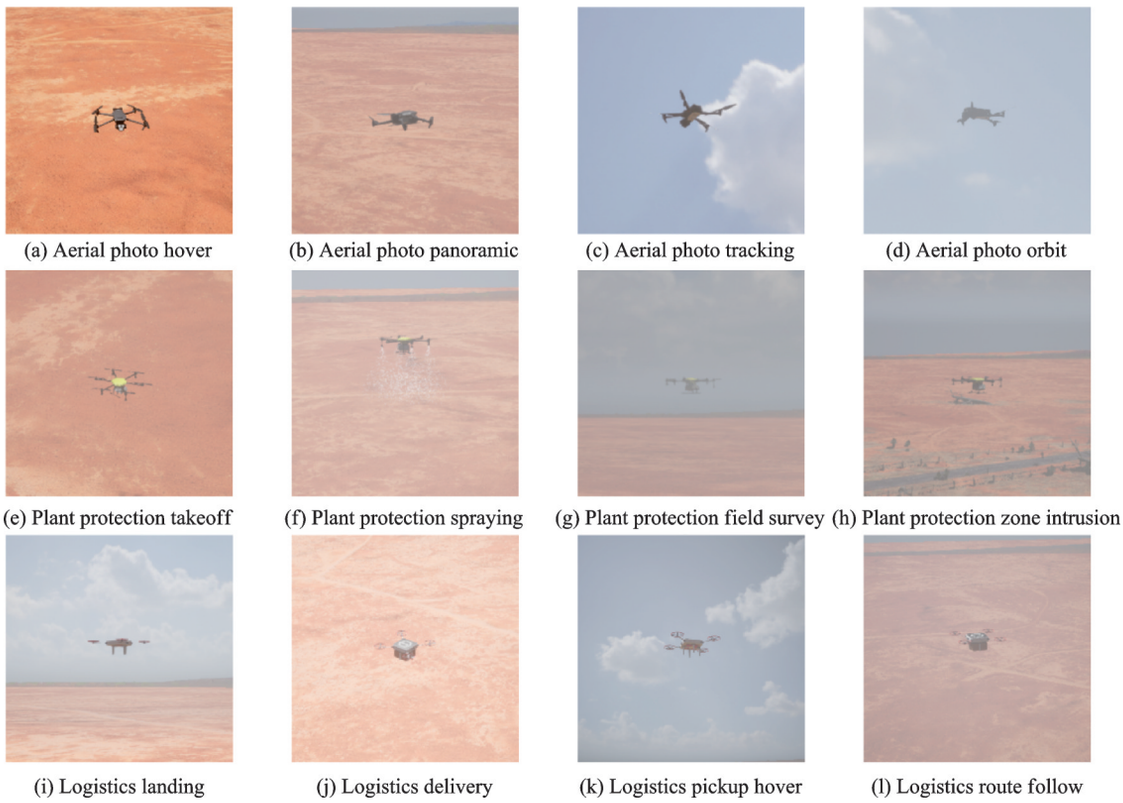

Fig.2 Example images from the dataset showing different UAV types and visual states

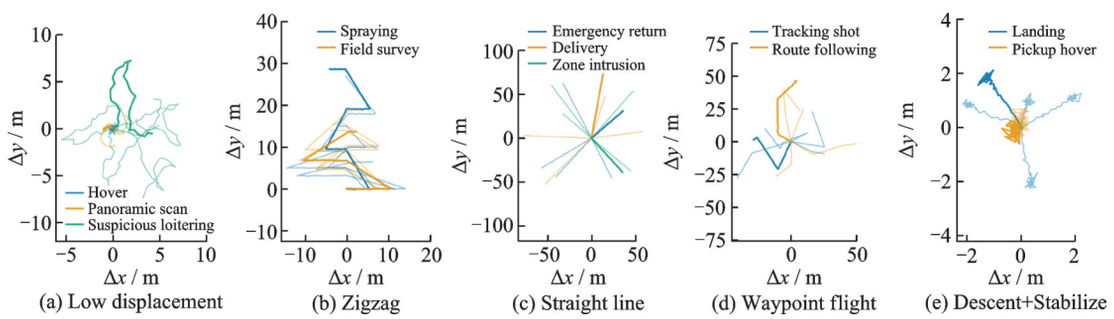

Fig.3 Trajectory examples from the five confusable groups

being performed, or that a trajectory segment should be interpreted differently because a payload or spray pattern is visible. This limitation is particularly relevant to trajectory-confusable intents, where fine-grained visual evidence must be associated with specific motion segments.

To address this limitation, BCAF is designed for the structurally heterogeneous image-trajectory setting considered in this study. Existing token-level cross-modal attention frameworks are commonly developed for vision-language alignment, such as ViLBERT[32] and LXMERT[33], or for audio-visual fusion, such as MBT[8]. These settings differ from the present task, where a two-dimensional image-token grid must be associated with a one-dimensional trajectory-token sequence. The image branch encodes relatively static airframe, payload and operating-state cues, whereas the trajectory branch encodes temporally distributed maneuver evidence. BCAF applies bidirectional cross-attention directly between image spatial tokens and trajectory temporal tokens before global pooling. This design preserves the spatial location of visual evidence and the temporal position of motion evidence during fusion. It differs from late fusion, which combines already pooled descriptors, and from bottleneck-token fusion, where cross-modal interaction is mediated by a compact set of latent tokens.

The module implements this direct interaction through two parallel cross-attention operations. In the image-to-trajectory direction, the 49 image tokens are used as queries, and the 40 trajectory tokens are used as keys and values. Each visual region is augmented with motion information relevant to that region. In the trajectory-to-image direction, the 40 trajectory tokens query the 49 image tokens, allowing each time step to incorporate visually grounded task cues.

Formally, let $\boldsymbol{S}_{\text{img}} \in \boldsymbol{R}^{49\times d}$ denote the image spatial tokens and $\boldsymbol{S}_{\text{traj}} \in \boldsymbol{R}^{40\times d}$ denote the trajectory temporal tokens. The cross-attention in each direction is computed as multi-head scaled dot-product attention

$$\text{CrossAttn}(\boldsymbol{Q},\boldsymbol{K},\boldsymbol{V}) = \text{Concat}(\text{head}_1,\text{head}_2,\cdots,\text{head}_h)\boldsymbol{W}^{\text{O}} \quad (1)$$

$$\text{head}_j = \text{softmax}\left(\frac{\boldsymbol{Q}\boldsymbol{W}_j^{\text{Q}}(\boldsymbol{K}\boldsymbol{W}_j^{\text{K}})^{\text{T}}}{\sqrt{d_h}}\right)\boldsymbol{V}\boldsymbol{W}_j^{\text{V}} \quad (2)$$

where $\boldsymbol{Q}$, $\boldsymbol{K}$ and $\boldsymbol{V}$ denote the query, the key and the value input matrices, respectively. The index j identifies one of the $h=4$ attention heads, and head_j denotes its output. The matrices $\boldsymbol{W}_j^{\text{Q}}$, $\boldsymbol{W}_j^{\text{K}}$ and $\boldsymbol{W}_j^{\text{V}}$ are the corresponding learnable query, key and value projection matrices, respectively. The feature dimension of each attention head is $d_h = d/h = 64$. The matrix $\boldsymbol{W}^{\text{O}}$ is the learnable output projection for the concatenated head outputs. The two directional cross- attention outputs are computed as

$$\boldsymbol{Z}_{\text{img}\leftarrow\text{traj}} = \text{LN}(\text{CrossAttn}(\boldsymbol{S}_{\text{img}},\boldsymbol{S}_{\text{traj}},\boldsymbol{S}_{\text{traj}})) \quad (3)$$

$$\boldsymbol{Z}_{\text{traj}\leftarrow\text{img}} = \text{LN}(\text{CrossAttn}(\boldsymbol{S}_{\text{traj}},\boldsymbol{S}_{\text{img}},\boldsymbol{S}_{\text{img}})) \quad (4)$$

where the operator LN denotes layer normalization along the feature dimension.

Each directional output is then refined by a two-layer feed-forward network (FFN) with GELU activation and residual normalization, shown as

$$\overline{\boldsymbol{Z}}_{\text{img}\leftarrow\text{traj}} = \text{LN}(\boldsymbol{Z}_{\text{img}\leftarrow\text{traj}} + \text{FFN}(\boldsymbol{Z}_{\text{img}\leftarrow\text{traj}})) \quad (5)$$

with an analogous operation for $\boldsymbol{Z}_{\text{traj}\leftarrow\text{img}}$. Finally, self-attention is applied within each cross-enhanced sequence to consolidate the newly injected information, shown as

$$\tilde{\boldsymbol{S}}_{\text{img}} = \text{LN}(\overline{\boldsymbol{Z}}_{\text{img}\leftarrow\text{traj}} + \text{MultiHeadSelfAttn}(\overline{\boldsymbol{Z}}_{\text{img}\leftarrow\text{traj}})) \quad (6)$$

The trajectory-enhanced sequence $\tilde{\boldsymbol{S}}_{\text{traj}}$ is obtained in the same way. The final fused feature is formed by mean pooling the two enhanced token sequences, concatenating the pooled vectors into a 512-D representation, and projecting it to 256 D with a two-layer MLP followed by layer normalization.

To reduce modality dominance during training, modality-level dropout is applied at the input of BCAF. With probability 0.15, all image spatial tokens are replaced by zeros. With probability 0.15, all trajectory tokens are replaced by zeros. Otherwise, both modalities are retained. The operation is disabled during inference. This strategy discourages reliance on the stronger unimodal cue and encourages both branches to learn representations that remain useful when one modality is weak, noisy or partially missing.

2.3 Joint optimization with consistency constraints

The 256D fused feature is fed into two task heads. The type head uses a 64D hidden layer and outputs logits over three UAV types. The intent head uses a 128D hidden layer and outputs logits over 15 intent classes. Both heads use ReLU activation and dropout of 0.2. Joint optimization is adopted because UAV type is not merely an auxiliary label. It also encodes operational feasibility, since several intents are valid only for specific platform categories. The training objective is

$$\mathcal{L} = \alpha\mathcal{L}_{\text{type}} + \beta\mathcal{L}_{\text{intent}} + \gamma\mathcal{L}_{\text{consistency}} \quad (7)$$

where the type and intent losses are standard cross-entropy terms. The weights are fixed to $\alpha = 2.0$, $\beta = 1.0$ and $\gamma = 0.5$ in all experiments. A differentiable consistency loss further penalizes physically invalid type-intent combinations. The coefficients α, β and γ weight the type, the intent and the consistency losses, respectively. They are fixed at $\alpha = 2.0$, $\beta = 1.0$ and $\gamma = 0.5$ throughout the experiments. Let $\boldsymbol{M}^{\text{inv}} \in \{0,1\}^{3\times 15}$ denote an invalidity mask, where $M_{t,i}^{\text{inv}} = 1$ indicates that UAV type t and intent i cannot co-occur. For example, an aerial- photography UAV (T1) cannot perform spraying (I08), and a logistics UAV (T3) cannot execute an orbit shot (I07). The consistency term is defined as the expected probability assigned to invalid combinations, shown as

$$\mathcal{L}_{\text{consistency}}=\frac{1}{B}\sum_{b=1}^{B}\sum_{t=1}^{3}\sum_{i=1}^{15}M_{t,i}^{\text{inv}}\,\sigma(\hat{\boldsymbol{y}}_{\text{type}}^{(b)})_t\,\sigma(\hat{\boldsymbol{y}}_{\text{intent}}^{(b)})_i \quad (8)$$

where σ denotes the softmax function and B the batch size. The vectors $\hat{\boldsymbol{y}}_{\text{type}}^{(b)}$ and $\hat{\boldsymbol{y}}_{\text{intent}}^{(b)}$ denote the type and the intent logits for the b-th sample in the batch, respectively. Unlike hard output masking, this soft penalty does not forcibly remove any class during inference. Instead, it provides gradient signals that encourage the two heads to produce mutually consistent predictions. The larger weight on type classification reflects the lower ambiguity of type labels and their value as a supervisory signal for learning type-dependent intent cues. Table 2 lists the valid type-intent combinations.

Table 2 Valid type-intent combination matrix

ID	I01	I02	I03	I04	I05	I06	I07	I08	I09	I10	I11	I12	I13	I14	I15
T1	✓	✓	✓	✓	✓	✓	✓						✓	✓	✓
T2	✓	✓	✓	✓				✓	✓				✓	✓	✓
T3	✓	✓	✓	✓						✓	✓	✓	✓	✓	✓

3 Experiments

The experiments address three questions central to low-altitude traffic surveillance: Whether multimodal sensing improves intent recognition over image-only or trajectory-only observation, whether token-level cross-attention is more effective than conventional late fusion, and which MCFNet components contribute to the final performance. This section describes the simulation dataset and training protocol, then reports the main comparison, ablation results and analyses of confusable intent groups.

3.1 Dataset and setup

Large-scale real-world UAV intent datasets with reliable ground-truth labels are difficult to collect because many safety-critical or anomalous behaviors are rare, regulated or unsafe to reproduce repeatedly. A simulation-based dataset was therefore constructed using UE4 with the AirSim plugin[34], providing photorealistic rendering and physics-based flight dynamics. The simulator makes it possible to control UAV type, payload state, camera viewpoint, weather and trajectory parameters while preserving exact type and intent annotations.

The dataset covers three operational UAV types: Aerial photography (T1), plant protection (T2) and logistics (T3). For each type, multiple airframe variants are modelled to reduce reliance on a single appearance template. T1 includes three quadrotor photography platforms. T2 includes three agricultural UAVs with hexarotor, quadrotor and octarotor configurations, each rendered under spray-on and spray-off states. T3 includes three delivery UAVs, each rendered under loaded and empty states. These variants are important because several intents are associated with local visual cues, such as spray particles or a cargo box. Table 3 lists the specifications of all 15 UAV blueprints.

Table 3 UAV blueprint specifications

Type	Model ID	Airframe	Wingspan/m	Visual state
T1	T1_M1	Quadrotor	0.4	Default
	T1_M2	Quadrotor	0.6	Default
	T1_M3	Quadrotor	1.0	Default
T2	T2_M1	Hexarotor	1.2	Spray / Idle
	T2_M2	Quadrotor	1.5	Spray / Idle
	T2_M3	Octarotor	2.0	Spray / Idle
T3	T3_M1	Quadrotor	0.5	Loaded / Empty
	T3_M2	Hexarotor	1.0	Loaded / Empty
	T3_M3	Hexarotor	1.5	Loaded / Empty

For each valid type-intent combination, multiple trajectories were generated by parameterized motion generators. Each simulated episode contains 80 frames at 5 Hz. MCFNet used the first 40 frames as the observation window for intent recognition. The remaining frames were retained only as part of the full simulated episode and were not used as a prediction target. Trajectory parameters, including speed, altitude, row width, turning radius and maneuver amplitude, were sampled from randomized opera-

tional ranges. To create a realistic intent-recognition challenge, several intent groups were deliberately assigned overlapping kinematic envelopes. For example, spraying (I08) and field survey (I09) share identical zigzag trajectory parameters and differ mainly in the visual presence of spray particles. Emergency return (I04), delivery flight (I10) and zone intrusion (I13) share similar high-speed straight-line motion and differ mainly in platform type and mission context. This design prevents trajectory-only recognition from trivially solving the task.

For each trajectory, an RGB image was captured from a simulated external observer camera. Camera distance was sampled from 10 m to 35 m, elevation angle from -60° to 30°, and azimuth from 0 to 360°. The camera was always oriented towards the UAV. Weather conditions, including cloud cover, rain intensity, fog density and dust level, were randomized to increase visual diversity. Images were captured at $1\,280 \times 960$ resolution, cropped around the UAV center with a context factor of 5.0 times the estimated target size, and resized to 224×224. Fig.4 illustrates the data collection pipeline. The final dataset contains 7 290 samples distributed across all valid type-intent combinations. It was split into 70% training (5 102 samples), 15% validation (1 094 samples) and 15% test (1 094 samples) sets using stratified sampling. The benchmark was designed to test controlled trajectory ambiguity and does not replace real-world deployment validation. Table 4 summarizes the per-intent distribution.

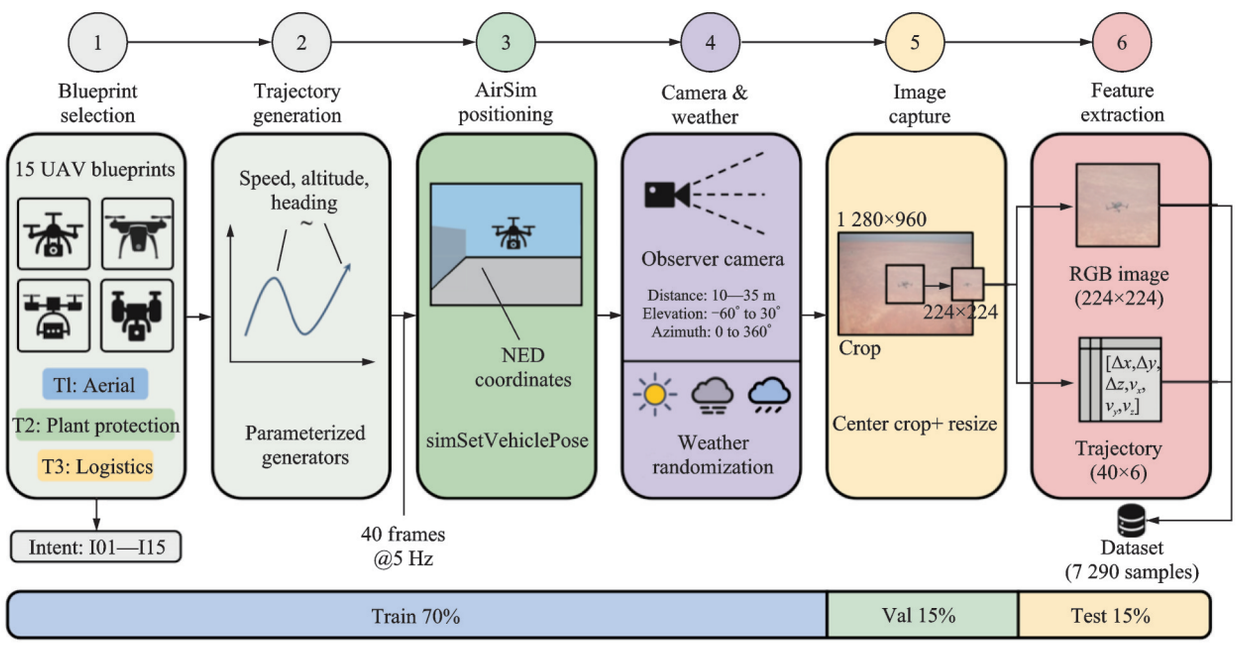

Fig.4 Data collection pipeline diagram

Trajectory features were normalized using z-score statistics computed from the training set and then applied to all splits. During training, trajectory augmentation included random temporal shifts of up to two frames, additive Gaussian noise with standard deviation 0.1, and 10% random frame dropout followed by linear interpolation[35]. Image augmentation included random resized cropping with a conservative scale range of 0.9 to 1.0, horizontal flipping and color jittering with brightness 0.3, contrast 0.3, saturation 0.2 and hue 0.1. Validation and test images used deterministic resizing and ImageNet normalization.

All models were trained with the AdamW optimizer and weight decay of 10^{-4}. The pretrained ResNet-18 backbone used a learning rate of 5×10^{-5}, while newly initialized layers used 1×10^{-4}. A cosine annealing schedule with a five-epoch warmup was adopted. Training ran for up to 50 epochs, with early stopping based on validation intent

Table 4 Dataset statistics by intent and UAV type

ID	Intent name	T1	T2	T3	Total
I01	Takeoff	240	240	240	720
I02	Landing approach	240	240	240	720
I03	Idle hover	240	240	240	720
I04	Emergency return	240	240	240	720
I05	Panoramic scan	240	0	0	240
I06	Tracking shot	240	0	0	240
I07	Orbit shot	240	0	0	240
I08	Spraying	0	450	0	450
I09	Field survey	0	240	0	240
I10	Delivery flight	0	0	300	300
I11	Pickup	0	0	240	240
I12	Route following	0	0	300	300
I13	Zone intrusion	240	240	240	720
I14	Evasive maneuvering	240	240	240	720
I15	Suspicious loitering	240	240	240	720
	Total	2 400	2 370	2 520	7 290

Table 5 Hyperparameter settings used for training all models

Parameter	Value
Image backbone	ResNet-18 (pretrained)
Feature dimension d	256
Transformer layers	2
Attention heads	4
FFN hidden dim	512
Transformer dropout	0.1
Fusion dropout	0.3
Head dropout	0.2
Modality dropout	0.15
Optimizer	AdamW
Learning rate (new)	1×10^{-4}
Learning rate (backbone)	5×10^{-5}
Weight decay	1×10^{-4}
Scheduler	Cosine annealing + warmup
Batch size	32
Max epoch	50
Early stopping patience	15
Gradient clipping	Max norm 1.0
Type loss α	2.0
Intent loss β	1.0
Consistency γ	0.5

accuracy and a patience of 15 epochs. Gradient clipping with a maximum norm of 1.0 and FP16 mixed-precision training were used. Table 5 lists the hyperparameters. Six baselines were included. Image-only and trajectory-only evaluated the discriminative value of each modality in isolation. Late fusion (Concat) and late fusion (Add) represented conventional multimodal fusion strategies that combine global image and trajectory descriptors after unimodal encoding. LSTM-concat replaced the Transformertrajectory encoder with a recurrent encoder while retaining late concatenation. MBT was also included as a representative token-level cross-modal fusion baseline, where a small set of shared bottleneck tokens mediates information exchange between the two modalities. All multimodal baselines used the same ResNet-18 visual backbone and comparable feature dimensions. Main comparison experiments were repeated with three random seeds (42, 1 337 and 3 407), and results are reported as mean±standard deviation. Ablation experiments used seed 42 unless stated otherwise.

3.2 Main results

Table 6 reports type accuracy, intent accuracy, joint accuracy and macro-F_1 on the test set. Joint accuracy is counted as correct only when both UAV type and intent are correctly predicted. It is therefore a stricter metric for traffic-management use cases, where platform capability and behavior must be interpreted consistently.

Table 6 Main comparison results on the UAV intent recognition dataset

Method	Modality	Type acc/%	Intent acc/%	Joint acc/%	Macro-F_1/%
Image-only	Image	99.85±0.16	32.60±0.77	32.57±0.78	31.06±0.87
Trajectory-only	Trajectory	50.06±0.35	79.07±0.61	39.37±1.09	69.65±0.73
Late fusion (Concat)	Both	99.82±0.20	86.59±1.09	86.41±1.27	86.31±1.17
Late fusion (Add)	Both	99.97±0.04	86.45±0.49	86.42±0.52	87.55±0.58
LSTM-concat	Both	99.88±0.04	86.14±0.42	86.01±0.47	86.89±0.55
MBT	Both	99.82±0.13	88.94±0.99	88.82±1.05	91.06±1.04
MCFNet	Both	99.88±0.04	89.63±0.99	89.41±1.00	91.33±1.05

The unimodal baselines reveal the complementary roles of the two information sources. Image-only achieves nearly saturated type accuracy because airframe appearance is directly visible. Its intent accuracy is only 32.60%, however, showing that a single image is insufficient for recognizing most operational behaviors. Trajectory-only reaches 79.07% intent accuracy, confirming that motion history is highly informative. Its type accuracy is weak because platform category cannot be reliably inferred from kinematics alone. The gap between unimodal and multimodal models shows that low-altitude intent recognition requires both operational motion and visual capability cues.

This unimodal result also addresses whether the joint type-intent formulation encourages a shortcut through UAV type or static visual cues. The low image-only intent accuracy indicates that platform appearance alone is insufficient for intent recognition. In a single-seed prediction analysis, MCFNet further achieves 86.24% intent accuracy and 85.21% macro-F_1 on the seven type-shared intents (I01—I04 and I13—I15), which are valid for all three UAV types and cannot be resolved by type identity. These results suggest that MCFNet learns motion-conditioned visual representations rather than relying only on type-based shortcuts.

Among the late-fusion variants, Late fusion (Concat) achieved the highest intent accuracy of 86.59%. MCFNet increased intent accuracy to 89.63%, corresponding to a gain of 3.04%, and improved joint accuracy from 86.41% to 89.41%. This gain indicates that the improvement does not come simply from using two modalities, because all multimodal baselines receive the same image and trajectory inputs. Instead, the advantage is associated with where the two modalities interact: MCFNet aligns spatial image tokens with temporal trajectory tokens before global pooling, whereas late fusion combines already compressed unimodal descriptors.

MCFNet was further compared with MBT, a token-level cross-modal fusion baseline. MBT achieved 88.94%±0.99% intent accuracy and 88.82%±1.05% joint accuracy, showing aggregate performance comparable to MCFNet. Under the three-seed evaluation, the two methods showed no clear separation in overall accuracy. Thus, the comparison with MBT should not be interpreted as a large aggregate accuracy advantage. Rather, MCFNet differs from MBT in providing a direct spatial-temporal interaction path, with attention weights that remain indexed by image regions and trajectory time steps. This distinction is qualitatively illustrated by the attention visualizations in Section 3.4.

The per-intent breakdown in Fig.5 further shows where the performance differences arise. All three multimodal models achieved perfect F_1 on intents with an unambiguous cue in at least one modality, including takeoff (I01), spraying (I08), field survey (I09) and evasive maneuvering (I14). MCFNet obtained higher F_1 than MBT on seven intent classes, tied on six classes and was slightly lower on two classes. Compared with Late fusion (Concat), the largest gains were observed for pickup hover (I11, 45.8%), zone intrusion (I13, +13.8%) and landing approach (I02, +11.2%). Emergency return (I04) remained the most difficult intent for all models, with F_1 below 40%.

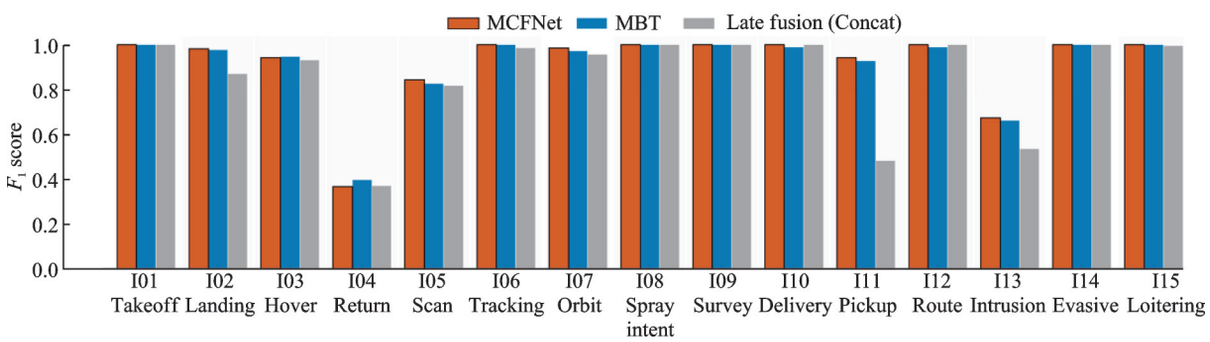

Fig.5 Per-intent F_1 score comparison between MCFNet, MBT, and Late fusion (Concat)

3.3 Ablation and robustness

Table 7 isolates the contribution of the main design choices. Because ablations are run with a single seed, the full-model value in this table is not identical to the three-seed mean reported in Table 6.

Table 7 Ablation study results

Variant	Intent acc/%	Δ/%
MCFNet (Full)	90.40	—
w/o consistency loss	89.21	−1.19
w/o multi-task (Intent only)	89.31	−1.09
LSTM encoder	86.56	−3.84
ResNet-50 backbone	88.48	−1.92

Removing the consistency loss reduces intent accuracy by 1.19 percentage points. This indicates that type-intent feasibility constraints provide useful regularization beyond standard cross-entropy supervision, but the modest decrease also shows that the validity matrix acts as a soft prior rather than a deterministic rule. Valid type-intent associations can still be learned largely from the image-trajectory data when the prior is removed. For modified UAVs, improvised payloads or unforeseen mission profiles where the predefined matrix may not hold, the consistency weight can be reduced, or the fixed matrix can be replaced by a learnable feasibility prior without changing the model architecture.

Training with the intent objective alone, by disabling both type supervision and the consistency term, reduces accuracy by 1.09%. This result supports the multi-task formulation. Type recognition supplies an auxiliary signal that helps the shared representation learn platform-dependent operational cues, especially for spraying, delivery and other type-specific intents.

Replacing the Transformertrajectory encoder with a bidirectional LSTM causes the largest degradation, reducing intent accuracy by 3.84%. This suggests that self-attention is better suited to capturing nonlocal temporal dependencies in UAV motion, such as repeated zigzag coverage, gradual descent during landing or bounded movement during loitering.

Using ResNet-50 instead of ResNet-18 does not improve performance. This indicates that the limiting factor is not visual backbone capacity alone. Increasing image-encoder depth cannot replace the contribution of the fusion architecture.

Table 8 reports the effect of observation-window length. A 20-frame window provides only 4 s of motion history and is insufficient for recognizing patterns such as zigzag coverage or orbital movement. Extending the window from 40 frames to 60 frames yields only a marginal gain of 0.29% in intent accuracy and 0.11% in joint accuracy. The 40-frame setting is therefore used as the default operating point because it provides a better accuracy-latency trade-off for early intent recognition.

Table 8 Observation-window sensitivity

Window	Duration/s	Intent acc/%	Joint acc/%
20	4	86.21	86.03
40	8	90.40	90.40
60	12	90.69	90.51

Trajectory-noise robustness was further evaluated by adding Gaussian perturbations to the normalized trajectory features at test time. As shown in Fig.6, MCFNet remains stable under mild-to-moderate trajectory perturbations. Intent accuracy decreases only from 90.40% to 89.76% at $\sigma=0.05$, 89.40% at $\sigma=0.10$ and 88.21% at $\sigma=0.15$. Macro-F_1 follows the same trend, decreasing from 91.57% to 89.00% at $\sigma=0.15$. Performance drops more sharply under stronger perturbations, reaching 81.17% intent accuracy at $\sigma=0.20$ and 70.57% at $\sigma=0.25$. This pattern indicates that MCFNet is robust within a mild-to-moderate trajectory-noise range, but remains sensitive to severe trajectory corruption.

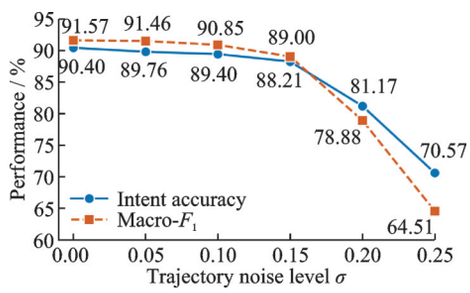

Fig.6 Trajectory-noise robustness of MCFNet

3.4 In-depth analysis

This analysis examines where the proposed direct spatial-temporal interaction is most useful and whether the learned attention can be interpreted in terms of image regions and trajectory time steps. After the aggregate comparison and ablation results, the remaining question is why specific intent groups benefit from multimodal fusion and why some errors persist. This subsection analyses normalized confusion matrices and cross-attention visualizations for representative confusable groups.

Fig.7 compares the normalized confusion matrices of MCFNet and Late fusion (Concat). The diagonal pattern is stronger for MCFNet, and the remaining errors are concentrated in operationally plausible confusions rather than random class swaps.

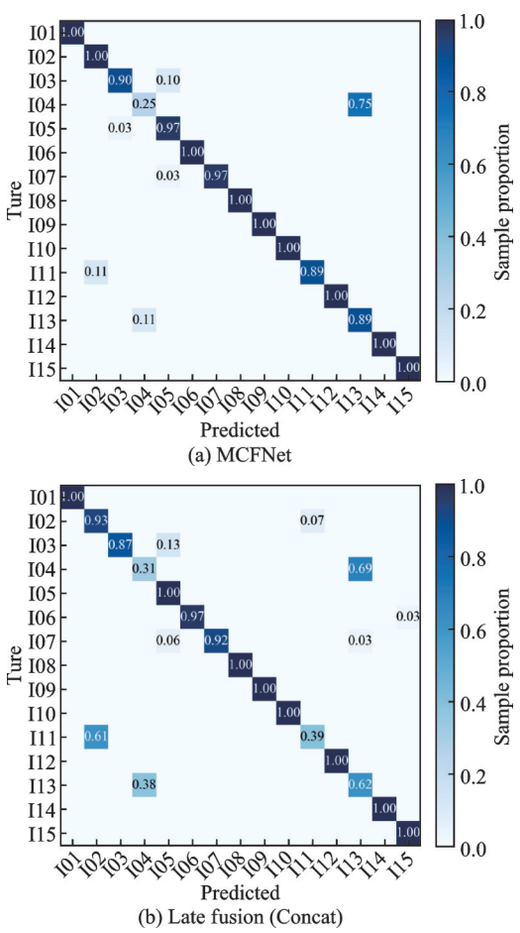

Fig.7 Confusion matrices of MCFNet and Late fusion (Concat)

As shown in Fig.8, the three confusable groups reveal a graded pattern rather than a uniform advantage. Multimodal fusion is sufficient when visual cues are highly salient, whereas BCAF is most useful when local visual or contextual evidence must be interpreted together with ambiguous motion.

For Group A, which includes idle hover (I03), panoramic scan (I05) and suspicious loitering (I15), all trajectories show limited displacement over the observation window. MCFNet improves the recall of idle hover from 87.04% to 89.81% and preserves perfect recall for suspicious loitering, indicating that the model can better separate part of this low-motion intent group. Panoramic scan remains a borderline case: Late fusion (Concat) reaches perfect recall, whereas MCFNet shows a slight residual confusion. This result suggests that the benefit of token-level fusion is selective rather than uniform, and is most apparent when local visual evidence and subtle motion differences must be interpreted jointly.

For Group B, spraying (I08) and field survey (I09) share the same zigzag motion pattern and differ mainly in the spray-particle cue beneath the UAV. All three multimodal methods reach 100% F_1 on this pair, indicating that the visual cue is sufficiently distinctive once image information is included. This result reinforces the need for image-trajectory fusion, because the two intents are deliberately trajectory-indistinguishable. It also shows that Group B is an easier multimodal case, where a salient visual cue can already be captured by global fusion.

For Group C, emergency return (I04), delivery flight (I10) and zone intrusion (I13) share high-speed, near-linear trajectories. This group better reflects the high-ambiguity setting in which fine-grained image-trajectory interaction is expected to be most useful. The clearest gain is observed for zone intrusion, where recall increases from 62.04% for Late fusion (Concat) to 88.89%, indicating that BCAF helps disambiguate direct-flight behaviors when trajectory evidence alone is insufficient. Emergency return remains the main failure case: Its recall under MCFNet is 25.00%, slightly lower than the 31.48% recall of Late fusion (Concat). A closer look at the confusion matrix shows that 75% of I04

samples are predicted as zone intrusion (I13). This error reflects a practical ambiguity rather than a random failure. Emergency return and zone intrusion both involve high-speed direct flight and differ mainly in destination semantics, which cannot be fully inferred from a single 8 s observation window without geofence, destination or mission-context information.

To examine whether BCAF learns meaningful cross-modal correspondences, attention weights are visualized for representative samples in Fig.8. Each row shows the original image, the trajectory-to-image attention heatmap overlaid on the image, and the image-to-trajectory attention weights over time.

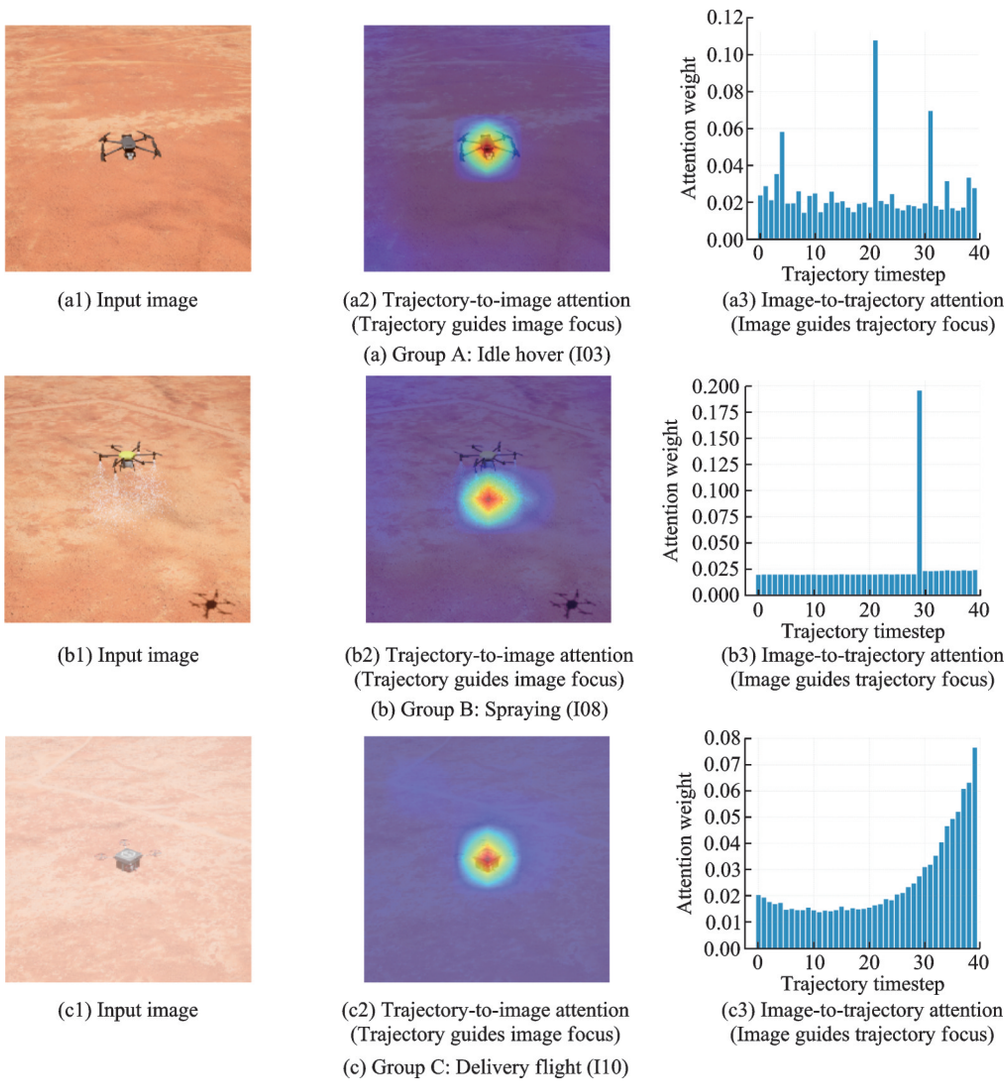

Fig.8 Cross-attention visualization examples from Groups A, B, and C

The attention maps are consistent with the intended fusion behavior. For the spraying sample, trajectory-to-image attention concentrates near the area where spray particles are rendered below the rotors. For the delivery sample, attention is concentrated around the payload region beneath the fuselage. For the idle-hover sample, attention is more broadly distributed across the UAV body, consistent with the absence of a single local visual cue. These patterns do not prove a causal explanation by themselves, but they provide qualitative evidence that the cross-attention module learns interpretable associations between motion context and visual regions. This direct spatial-temporal indexing also distinguishes BCAF from bottleneck-token fusion, where cross-modal interaction is routed through latent tokens that are less directly tied to image positions or trajectory time steps.

4 Conclusions and Future Work

This study investigated mission-level UAV intent recognition for low-altitude airspace surveillance. MCFNet was proposed to jointly infer UAV type and mission intent from an external-view RGB image and an 8 s trajectory history. By aligning image spatial tokens with trajectory temporal tokens, BCAF enables local visual evidence to be interpreted together with recent maneuver patterns. The multi-task type-intent formulation and differentiable consistency loss further incorporate mission-platform feasibility during training.

Experiments show that MCFNet achieves 89.63% mean intent accuracy, 89.41% joint type-intent accuracy and 91.33% macro-F_1 across three random seeds. It improves intent accuracy by 3.04% over the strongest late-fusion baseline and by 10.56% over the trajectory-only baseline. Its aggregate performance is also comparable to MBT, while retaining attention maps directly indexed by image regions and trajectory time steps. Ablations show that removing the consistency loss, disabling multi-task learning and replacing the Transformer trajectory encoder with an LSTM reduce intent accuracy by 1.19%, 1.09% and 3.84%, respectively. These results indicate that UAV intent recognition should be treated as context-aware behavior interpretation, not as isolated visual identification or trajectory classification. By linking what a UAV looks like with how it moves, MCFNet provides a step towards earlier warning and more differentiated decision support for future low-altitude traffic management.

Three limitations should be noted. First, the current evaluation is simulation-based. Although airframe variants, viewpoints and weather conditions were randomized to reduce template overfitting, performance may still decrease under real-world sensor noise, motion blur, occlusion, illumination changes and imperfect target localization. A representative real-world validation set is therefore needed to quantify the simulation-to-reality gap and support synthetic-to-real fine-tuning, feature alignment or confidence-filtered self-training. Second, the type-intent validity matrix was specified manually according to the predefined mission taxonomy. This prior may be too restrictive for modified UAVs, improvised payloads or repurposed mission profiles. Future work could replace the fixed binary matrix with a data-driven compatibility prior or a continuous type-intent compatibility score. Third, the task remains closed-set, with three UAV types and 15 predefined intents, and it does not yet incorporate broader operational context. A more deployable system should include unknown-intent detection, incremental updates for new UAV platforms and mission profiles, and contextual inputs such as geofences, planned corridors, landing zones and restricted areas.

References

[1] PANG B, LOW K H, DUONG V N. Chance-constrained UAM traffic flow optimization with fast disruption recovery under uncertain waypoint occupancy time[J]. Transportation Research Part C: Emerging Technologies, 2024, 161: 104547.

[2] WENG C, PAN T, CHEN C, et al. Urban low-altitude air transport management: Bridging dynamic traffic control and static network equilibrium[J]. Transportation Research Part C: Emerging Technologies, 2025, 178: 105237.

[3] RUSKIN K J, KAUFMANN K, WILLEMS B, et al. Multiple applications for a trajectory conformance monitoring alarm[J]. Transportation Research Interdisciplinary Perspectives, 2025, 33: 101605.

[4] TANG Xinmin, GU Junwei, LIU Bing, et al. Review on low-altitude surveillance technology and its development trend[J]. Journal of Nanjing University of Aeronautics and Astronautics, 2024, 56(6): 973-993. (in Chinese)

[5] SEIDALIYEVA U, ILIPBAYEVA L, TAISSARIYEVA K, et al. Advances and challenges in drone detection and classification techniques: A state-of-the-art review[J]. Sensors, 2023, 24(1): 125.

[6] BISIO I, GARIBOTTO C, HALEEM H, et al. RF/WiFi-based UAV surveillance systems: A systematic literature review[J]. Internet of Things, 2024, 26: 101201.

[7] YASMEEN A, DAESCU O. Recent research progress on ground-to-air vision-based anti-UAV detection and tracking methodologies: A review[J]. Drones, 2025, 9(1): 58-87.

[8] NAGRANI A, YANG S, ARNAB A, et al. Atten-

tion bottlenecks for multimodal fusion[J]. Advances in Neural Information Processing Systems, 2021, 34: 14200-14213.

[9] AARTS M J M, ELLERBROEK J, KNOOP V L. Capacity of a constrained urban airspace: Influencing factors, analytical modelling and simulations[J]. Transportation Research Part C: Emerging Technologies, 2023, 152: 104173.

[10] SAFADI Y, FU R, QUAN Q, et al. Macroscopic fundamental diagrams for low-altitude air city transport[J]. Transportation Research Part C: Emerging Technologies, 2023, 152: 104141.

[11] CUMMINGS C, MAHMASSANI H. Airspace congestion, flow relations, and 4-D fundamental diagrams for advanced urban air mobility[J]. Transportation Research Part C: Emerging Technologies, 2024, 159: 104467.

[12] FU R, SAFADI Y, QUAN Q, et al. Sky highway: An air traffic structure for low-altitude heterogeneous VTOL aircraft[J]. Transportation Research Part C: Emerging Technologies, 2026, 185: 105531.

[13] HAMISSI A, DHRAIEF A. A survey on the unmanned aircraft system traffic management[J]. ACM Computing Surveys, 2023, 56(3): 1-37.

[14] DAI W, QUEK Z H, PANG B, et al. Analysis of UTM tracking performance for conformance monitoring via hybrid SITL Monte Carlo methods[J]. Drones, 2023, 7(10): 597.

[15] RAHMAN M H, SEJAN M A S, AZIZ M A, et al. A comprehensive survey of unmanned aerial vehicles detection and classification using machine learning approach: Challenges, solutions, and future directions[J]. Remote Sensing, 2024, 16(5): 879.

[16] LENG J, YE Y, MO M, et al. Recent advances for aerial object detection: A survey[J]. ACM Computing Surveys, 2024, 56(12): 1-36.

[17] ALSHAER N, ABDELFATAH R, ISMAIL T, et al. Vision-based UAV detection and tracking using deep learning and Kalman filter[J]. Computational Intelligence, 2025, 41(1): e70026.

[18] JIANG T, LIU Y, DONG Q, et al. Intention-aware interactive Transformerfor real-time vehicle trajectory prediction in dense traffic[J]. Transportation Research Record, 2023, 2677(3): 946-960.

[19] HUANG T, FU R, SUN Q, et al. Driver lane change intention prediction based on topological graph constructed by driver behaviors and traffic context for human-machine co-driving system[J]. Transportation Research Part C: Emerging Technologies, 2024, 160: 104497.

[20] LING Y, MA Z, ZHANG Q, et al. PedAST-GCN: Fast pedestrian crossing intention prediction using spatial-temporal attention graph convolution networks[J]. IEEE Transactions on Intelligent Transportation Systems, 2024, 25(10): 13277-13290.

[21] XU L, YOU S, HE G, et al. Pedestrian-vehicle information modulation for pedestrian crossing intention prediction[J]. IEEE Transactions on Intelligent Vehicles, 2024, 10(3): 1919-1930.

[22] PERRUSQUÍA A, GUO W, FRASER B, et al. Uncovering drone intentions using control physics informed machine learning[J]. Communications Engineering, 2024, 3(1): 36-49.

[23] WU Yexin, ZHAO Yifei, WANG Hongyong. Deviation behavior analysis and detection based on flight trajectory data[J]. Transactions of Nanjing University of Aeronautics and Astronautics, 2026, 43(1): 127-144.

[24] ZHANG J, SHI Z, ZHANG A, et al. UAV trajectory prediction based on flight state recognition[J]. IEEE Transactions on Aerospace and Electronic Systems, 2023, 60(3): 2629-2641.

[25] SHI Z, ZHANG J, SHI G, et al. Design of a UAV trajectory prediction system based on multi-flight modes[J]. Drones, 2024, 8(6): 255-271.

[26] CAO Y, ZHANG J, SHI G, et al. Research on trajectory prediction algorithm based on unmanned aerial vehicles behavioral intentions[J]. Drones, 2025, 9(9): 640-661.

[27] ZHANG X, GONG Y, LU J, et al. Multi-modal fusion technology based on vehicle information: A survey[J]. IEEE Transactions on Intelligent Vehicles, 2023, 8(6): 3605-3619.

[28] XU P, ZHU X, CLIFTON D A. Multimodal learning with Transformers: A survey[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2023, 45(10): 12113-12132.

[29] GU J Y, BELLONE M, PIVOŇKA T, et al. CLFT: Camera-lidar fusion Transformerfor semantic segmentation in autonomous driving[J]. IEEE Transactions on Intelligent Vehicles, 2024. DOI: 10.1109/TIV.2024.3454971.

[30] HE K, ZHANG X, REN S, et al. Deep residual learning for image recognition[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern recognition. Las Vegas, USA: IEEE, 2016: 770-778.

[31] VASWANI A, SHAZEER N, PARMAR N, et al. Attention is all you need[C]//Proceedings of the 31st Conference on Neural Information Processing Systems (NIPS 2017). Long Beach, CA, USA: [s.n.], 2017.

[32] LU J, BATRA D, PARIKH D, et al. VilBERT: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks[C]//Proceedings of the

33rd Conference on Neural Information Processing Systems (NeurIPS 2019). Vancouver, Canada: [s. n.], 2019.

[33] TAN H, BANSAL M. LXMERT: Learning cross-modality encoder representations from Transformers[C]//Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP). Hong Kong, China: ACL, 2019.

[34] SHAH S, DEY D, LOVETT C, et al. AirSim: High-fidelity visual and physical simulation for autonomous vehicles[C]//Proceedings of the Field and service robotics: Results of the 11th International Conference. Cham: Springer International Publishing, 2017.

[35] IGLESIAS G, TALAVERA E, GONZÁLEZ-PRIETO Á, et al. Data augmentation techniques in time series domain: A survey and taxonomy[J]. Neural Computing and Applications, 2023, 35(14): 10123-10145.

Acknowledgements This work was supported by the National Natural Science Foundation of China (No.52102384), the Project for Enhancing R&D Capabilities in Green, Low-Carbon and New Energy Technologies (No.2025GH-JSHZ-02), and the Xihua University Education and Teaching Reform Project (Special Initiative on Industry-Education Integration) (No.xcjz2025001).

Author

The first/corresponding author Dr. **TANG Li** received her Ph.D. degree from Southwest Jiaotong University in 2015. She is currently an associate professor at Xihua University. Her research interests include low-altitude traffic management and control technologies, and transportation big data mining.

Author contributions Dr. **TANG Li** conceptualized the study, secured funding, and administered the project. Ms. **WANG Zijie** designed and performed the experiments, and wrote the manuscript. Dr. **TU Pengyue** provided critical feedback and suggested revisions. All authors commented on the manuscript draft and approved the submission.

Competing interests The authors declare no competing interests.

(Production Editor: SUN Jing)

基于图像-轨迹融合的任务级无人机意图识别

唐 立[1,2], 王梓洁[2,3], 涂彭越[3]

(1.西华大学汽车与交通学院,成都 610039,中国; 2.西华大学智能空地融合载具及管控教育部工程研究中心,成都 610039,中国; 3.西华大学航空航天与智能装备学院,成都 610039,中国)

摘要:低空交通管理需要及时判断无人机(Unmanned aerial vehicle, UAV)平台类型及其潜在任务意图。对于协作目标,申报飞行计划能够提供一定的意图信息;但对于非协作、异常或仅能被部分观测的无人机,意图需要从可观测线索中推断。现有无人机监视方法主要关注目标检测、跟踪、平台类型识别或轨迹预测,当不同任务具有相似短时运动特征,且单帧图像缺少时间上下文时,难以支持任务级意图判别。针对这一问题,本文提出一种图像-轨迹融合框架MCFNet,将外部视角RGB图像与8 s轨迹历史相结合,用于任务级无人机意图识别。该方法通过双向交叉注意力融合模块,在分类前建立局部视觉区域与轨迹时间步之间的关联,使载荷状态、喷洒效果等局部视觉线索能够结合相应机动片段进行解释。同时,引入模态随机dropout以降低单一模态捷径学习,并采用类型-意图一致性损失约束物理上不合理的类型-任务组合。在包含3类无人机、15类任务意图和7 290个仿真样本的数据集上,MCFNet取得了89.63%±0.99%的意图识别准确率、89.41%±1.00%的类型-意图联合准确率和91.33%±1.05%的宏平均F_1值。与最强的晚期融合基线相比,MCFNet的意图识别准确率提高了3.04%,并取得了略高于多模态瓶颈(Multimodal bottkneck transfomer, MBT)token级融合基线的平均意图识别准确率。结果表明,图像-轨迹融合能够为低空交通管理中的早期预警和差异化决策支持提供一种基于观测证据的技术路径。

关键词:无人机意图识别;图像-轨迹融合;低空空域;轨迹消歧;领域感知约束

研究亮点:

1.面向低空交通管理场景,将无人机任务级意图识别建模为基于外部观测证据的多模态感知问题。

2.通过双向交叉注意力在图像空间token与轨迹时间token之间建立直接关联。

3.引入模态随机dropout和类型-意图一致性损失,以降低单模态捷径学习并约束不合理的类型-任务组合。

4.所提方法相比基线取得更高的意图识别准确率,并在可解释的空间-时间注意力分析中体现出细粒度融合能力。