SCIENCE CHINA
Information Sciences

• RESEARCH PAPER •

November 2026, Vol. 69, Iss. 11, 210306:1–210306:17
https://doi.org/10.1007/s11432-026-5089-4

Special Topic: Communication, Sensing, and Target Recognition for Low-Altitude Economy

Robust radar-optical fusion for low-altitude target recognition under time-varying heterogeneous degradation

Cheng HU[1,2], Longxiang JIAO[1,2], Zhiyang CHEN[1,2*], Rui WANG[1,2], Lianjun WANG[3], Qi JIANG[1,2] & Qiuhui WANG[1,2]

[1]*Radar Technology Research Institute, School of Information and Electronics, Beijing Institute of Technology, Beijing 100081, China*
[2]*Advanced Technology Research Institute, Beijing Institute of Technology, Jinan 250300, China*
[3]*Research Department of Cyber-Electromagnetic Space Information Technology, Aerospace Information Research Institute, Chinese Academy of Sciences, Beijing 100094, China*

Received 15 March 2026/Revised 8 June 2026/Accepted 20 August 2026/Published online 26 October 2026

Abstract Reliable recognition of low-altitude targets, such as unmanned aerial vehicles (UAVs) and birds, is important for low-altitude airspace security. Radar-optical fusion provides complementary micro-motion and appearance cues for target recognition but is challenged by time-varying heterogeneous degradation: the two modalities undergo distinct degradations whose severities vary dynamically with weak cross-modal correlation. Specifically, fog and low illumination weaken texture and contour information in optical images, whereas clutter interference obscures micro-Doppler structures in radar spectrograms. Consequently, the reliability of the two modalities fluctuates independently over time, deteriorating multimodal fusion performance. To address this challenge, this paper proposes a radar-optical robust fusion network (RORF Net). The framework integrates three key components. First, a gradient-aware and semantically guided enhancement network extracts degradation-robust structural representations by modeling optical texture and contour patterns together with radar micro-Doppler ridges. Second, a confidence-driven dynamic gating module estimates sample-level modality reliability and adaptively regulates the contribution of each modality through bidirectional cross-modal interaction. Third, a cross-modal semantic alignment module maps radar and optical representations into a common semantic space through supervised contrastive learning. The network is optimized using a two-stage training strategy that first establishes cross-modal alignment and then enforces global semantic refinement. Experiments on our radar-optical multimodal dataset demonstrate that RORF Net attains an overall accuracy of 96.45%, surpassing existing state-of-the-art multimodal fusion algorithms by 8.71%. Ablation studies confirm its effectiveness under time-varying heterogeneous degradation.

Keywords low-altitude target recognition, radar-optical multimodal fusion, time-varying heterogeneous degradation, structure-aware feature extraction, dynamic gating fusion, contrastive learning

Citation Hu C, Jiao L X, Chen Z Y, et al. Robust radar-optical fusion for low-altitude target recognition under time-varying heterogeneous degradation. Sci China Inf Sci, 2026, 69(11): 210306, https://doi.org/10.1007/s11432-026-5089-4

1 Introduction

With the rapid expansion of the low-altitude economy, reliable recognition of low-altitude targets, such as unmanned aerial vehicles (UAVs) and birds, has become increasingly critical for aviation security and low-altitude airspace management, as these targets pose serious threats to airport operations and safety [1–3]. To address this challenge, various sensors have been explored. Acoustic sensors capture rotor noise but are limited by range and environmental interference [4]. 6G integrated sensing and communication (ISAC) base stations enable real-time UAV localization, but their sensing range is inherently bounded by the spatial extent of base-station deployment in non-cooperative scenarios [5]. WiFi-based methods monitor communication signals between UAVs and their controllers, yet they are constrained by prior protocol knowledge [6]. Optical sensors provide rich texture and contour information under favorable conditions, but their performance degrades under fog or low light environments [7]. Radar distinguishes low-altitude targets primarily through micro-Doppler signatures induced by periodic motions of rotors or wings [8,9], but the echoes are often contaminated by strong ground clutter.

Radar, operating in the microwave frequency band, can observe low-altitude targets over long ranges under diverse environmental conditions, making it a primary sensing modality for low-altitude target recognition [10, 11]. Nevertheless, micro-Doppler signatures of different targets may be quite similar, resulting in overlapping

* Corresponding author (email: zychen12@163.com)

© Science China Press 2026

feature distributions and reduced class separability [12–15]. Integrating radar with complementary modalities, particularly optical sensors that provide rich texture and edge information, offers an effective approach to improve recognition accuracy. Consequently, radar-optical fusion becomes a promising solution for accurate low-altitude target recognition in open-area scenarios.

Significant progress has been made in radar-optical information fusion for autonomous driving and remote sensing applications. In autonomous driving, methods such as CenterFusion [16] and CRN [17] establish geometric associations between radar point clouds and optical images in Bird's-Eye-View (BEV) space to enable multimodal feature interaction. LRFusionPR [18] and DPFT [19] further extend multimodal interaction into polar BEV space or multi-view representations, improving the capability of 3D object detection. In remote sensing, optical images are frequently integrated with synthetic aperture radar (SAR) scattering characteristics to enhance object detection, semantic segmentation and cloud removal [20]. In such fields, radar mainly provides geometric or spatial information to complement optical sensing, with fusion methods focusing on spatial association or pixel-level integration. Consequently, these methods may fail to fully capture the heterogeneous characteristics of radar micro-Doppler signatures and optical texture features in low-altitude target recognition.

In the domain of low-altitude target perception, recent studies have explored the integration of radar point clouds and optical imagery for UAV detection [21], where radar measurements are utilized as geometric cues for target localization. Other studies typically fuse range-Doppler matrices or target track information with optical image features using simple feature concatenation [22] or decision-level fusion [23]. These approaches rely on shallow radar representations and preliminary measurement information, largely neglecting the informative micro-Doppler signatures embedded in radar returns. Moreover, they rarely consider the impact of modality quality degradation.

In open environments, weather conditions and background clutter can vary significantly over time and across locations, subjecting multimodal data to time-varying heterogeneous degradation. On the one hand, different modalities are affected by heterogeneous disturbances: optical images are highly susceptible to fog and low light, which reduce image contrast and blur texture details [24–28], whereas radar signals are often corrupted by ground clutter that distorts micro-Doppler signatures [29]. On the other hand, the severity of degradation varies dynamically, and the quality variations of different modalities are generally uncorrelated. As a result, the reliability of radar and optical data may degrade independently at different observation moments. Such time-varying heterogeneous degradation poses substantial challenges for robust radar-optical recognition of low-altitude targets.

Several studies have begun to consider modality quality degradation in multimodal fusion. Existing approaches mainly focus on modality incompleteness, including missing point cloud, image occlusion or modality dropout [30, 31]. However, they pay less attention to signal-level degradation caused by clutter interference or image quality deterioration. MGFNet [32] introduces a dynamic gating mechanism that estimates multimodal data quality and adaptively adjusts fusion weights for optical and SAR data. Although this strategy provides useful insights, it benefits from the inherent spatial correspondence between optical and SAR images and therefore does not explicitly address cross-modal alignment. In contrast, radar time-frequency spectrograms and optical images exhibit large semantic discrepancies and lack explicit spatial correspondence, requiring dedicated alignment mechanisms. Consequently, existing approaches still fall short of addressing the joint challenges of heterogeneous modality representations and time-varying degradation in radar-optical fusion.

To address this problem, this paper proposes a radar-optical robust fusion network (RORF Net) for low-altitude target recognition under time-varying heterogeneous degradation. The network decomposes the recognition problem under time-varying heterogeneous degradation into three progressively structured sub-tasks: robust single-modality representation, stable cross-modal alignment, and selective multimodal fusion. First, we design a gradient-aware and semantically-enhanced network which explicitly enhances micro-Doppler ridge structures in radar spectrograms and contour information in optical images, and further aggregates multi-scale features under semantic guidance, thereby reinforcing the most discriminative characteristics for low-altitude target recognition. Second, a confidence-driven dynamic gating module is introduced to adaptively adjust fusion weights based on sample-level modality confidence via bidirectional cross-attention. Third, a cross-modal semantic alignment module employs two-stage contrastive learning to map radar and optical features into a common semantic subspace, enabling unified representation learning for heterogeneous modalities. The entire framework is trained via a two-stage strategy that first establishes cross-modal alignment and then enforces degradation-robust feature learning. Furthermore, we construct a radar-optical dataset covering diverse degradation scenarios. Experimental results demonstrate that the proposed method outperforms representative multimodal fusion methods and maintains robust recognition performance across different sensing conditions.

The remainder of this paper is organized as follows. Section 2 analyzes the characteristics of time-varying heterogeneous degradation in real-world outfield environments. Section 3 presents the proposed method and network architecture. Section 4 describes the dataset construction, experimental setup, and performance evaluation. Finally,

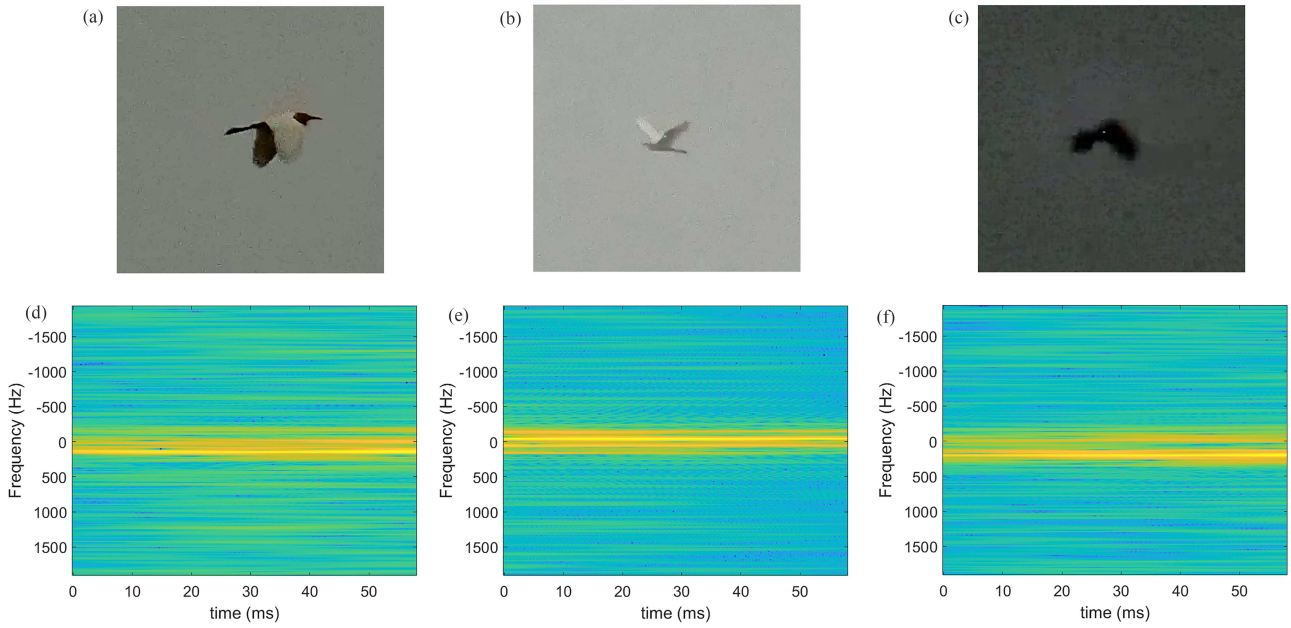

Figure 1 (Color online) Sample of time-varying heterogeneous degradation. (a) t_1 optical image; (b) t_2 optical image; (c) t_3 optical image; (d) t_1 radar spectrogram; (e) t_2 radar spectrogram; (f) t_3 radar spectrogram.

Section 5 concludes the paper and outlines future research directions.

2 Problem analysis

To illustrate the core challenge of time-varying heterogeneous degradation in radar-optical fusion, we analyze real-world samples of a bird observed at three distinct time instances, denoted as t_1, t_2 and t_3. Figure 1 presents both the radar time-frequency spectrogram and the corresponding optical image for each instance. The bird's micro-motion induces Doppler frequency spreading, resulting in time-varying micro-Doppler ridges in the radar spectrogram. The visibility of these ridges provides an indication of radar information quality, while energy concentrated near zero Doppler frequency corresponds to clutter interference.

At t_1, the optical image is clear, with the bird's contours and fine textures well preserved. The radar spectrogram exhibits minimal zero-frequency clutter, producing distinct micro-Doppler ridges with a high signal-to-clutter ratio. At t_2, corresponding to foggy conditions, the optical image deteriorates due to blur and a slight increase in brightness, causing the white bird to nearly merge with the misty background. Simultaneously, the radar spectrogram is affected by stronger zero-frequency clutter, partially obscuring the micro-Doppler ridges. At t_3, under dusk conditions, poor illumination leads to substantial loss of optical contour and fine-texture information. In contrast, the radar spectrogram shows a distinct degradation pattern. Although the signal-to-clutter ratio is similar to that at t_1, the bird's increased radial velocity shifts its micro-Doppler ridges away from low-frequency clutter, resulting in reduced interference compared to t_2.

These observations show that radar and optical degradation are not synchronized and evolve dynamically over time. The underlying causes are twofold. First, intrinsic differences in sensing mechanisms and signal formation processes lead to modality-specific degradation patterns. Second, environmental conditions change across time, further influencing the severity of degradation for each modality. Consequently, the reliability of each modality is sample-dependent, posing significant challenges to robust multimodal feature extraction and target recognition in open-area surveillance.

Based on this analysis, robust radar-optical recognition of low-altitude targets requires three coupled capabilities.

• Degradation-robust representation learning. Extract features that remain informative under degradation, such as optical edges and radar micro-Doppler ridges.

• Confidence-based dynamic fusion. Estimate the reliability of each modality for each sample and adaptively fuse their features to suppress the contribution of the lower-quality modality.

• Cross-modal feature alignment. Adjust the feature extraction space of each modality and align radar and optical features into a common representation.

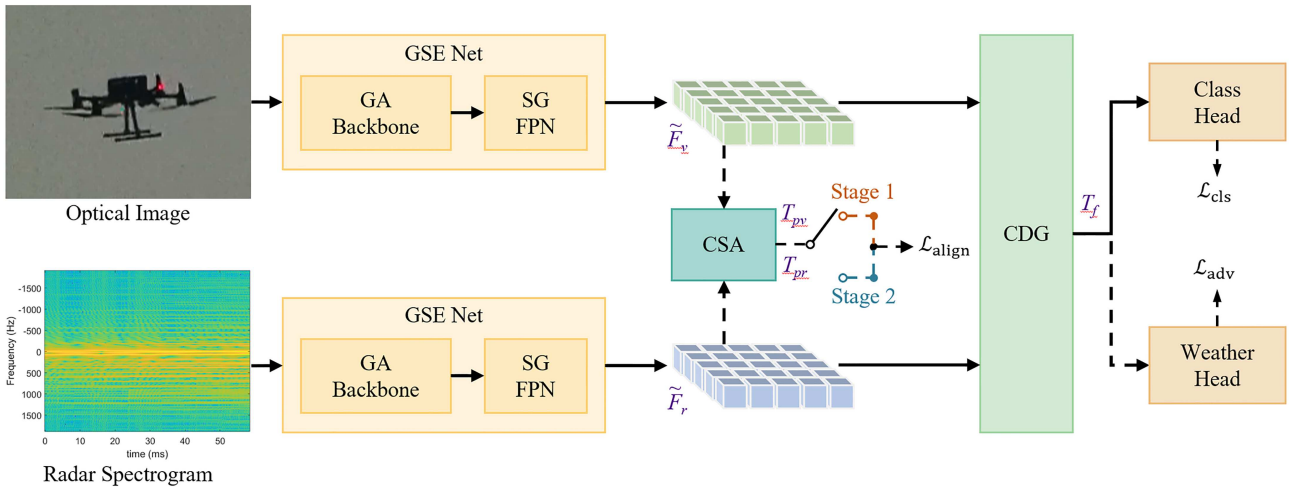

Figure 2 (Color online) Architecture of the proposed radar-optical robust fusion network (RORF Net). GSE Net extracts degradation-robust radar and optical features by combining gradient-aware local structural modeling with semantically guided multi-scale aggregation. The dynamic gating module establishes bidirectional cross-modal interaction and estimates sample-level modality confidence from the interaction features to adjust the fusion weights. During training, the cross-modal semantic alignment module projects radar and optical features into a common semantic space and enforces cross-modal semantic consistency through a two-stage supervised contrastive learning strategy, where the training objective focuses on cross-modal alignment in the early epochs and shifts to global semantic refinement in the later epochs. This branch operates only during training and is removed during inference.

3 Method

3.1 Overview

In open-area low-altitude surveillance, radar spectrograms and optical images are affected by different disturbances. Fog and low illumination weaken optical texture and contour information, whereas clutter interference obscures micro-Doppler ridges in radar spectrograms. Since the degradation severities of the two modalities are generally uncorrelated, the relative quality of paired radar-optical observations varies across samples. Robust recognition must therefore preserve target-dependent structural information within each modality, adapt their contributions during fusion, and reduce the semantic discrepancy between the multimodal representations.

Based on these requirements, we propose the radar-optical robust fusion network, as illustrated in Figure 2. Given a radar spectrogram and its corresponding optical image, the gradient-aware and semantically-enhanced network (GSE Net) first extracts degradation-robust features by modeling micro-Doppler ridge structures and optical texture and contour structures, followed by semantically guided multi-scale aggregation. Confidence-driven dynamic gating (CDG) module then establishes bidirectional cross-modal interaction and determines the contribution of each modality according to sample-level confidence. During training, the cross-modal semantic alignment (CSA) module acts on the modality-specific features in parallel with the fusion path and imposes two-stage supervised contrastive constraints on their class-level correspondence, while this module is removed during inference.

Formally, given a spatio-temporally registered pair consisting of a radar spectrogram X_r and optical image X_o, the network learns a mapping $\mathcal{F} : (X_r, X_o) \to y \in \mathbb{R}^C$, where C denotes the number of target categories. The overall pipeline can be summarized as

$$\begin{aligned} \tilde{F}_r &= \mathrm{GSE}(X_r), \\ \tilde{F}_o &= \mathrm{GSE}(X_o), \\ T_f &= \mathrm{CDG}(\tilde{F}_r, \tilde{F}_o), \\ T_{pr}, T_{po} &= \mathrm{CSA}(\tilde{F}_r, \tilde{F}_o), \\ y_{\mathrm{cls}} &= \mathrm{ClsHead}(T_f), \end{aligned} \tag{1}$$

where $\tilde{F}_r$ and $\tilde{F}_o$ are the features extracted by GSE Net, T_f is the fused feature produced by CDG, T_{pr} and T_{po} are the projected features generated by CSA, and y_{cls} denotes the classification result.

3.2 Gradient-aware and semantically-enhanced network

Fog, low illumination, and clutter interference can substantially change pixel intensity and spectral amplitude, whereas the spatial organization of target-dependent structures is comparatively stable. These structures appear

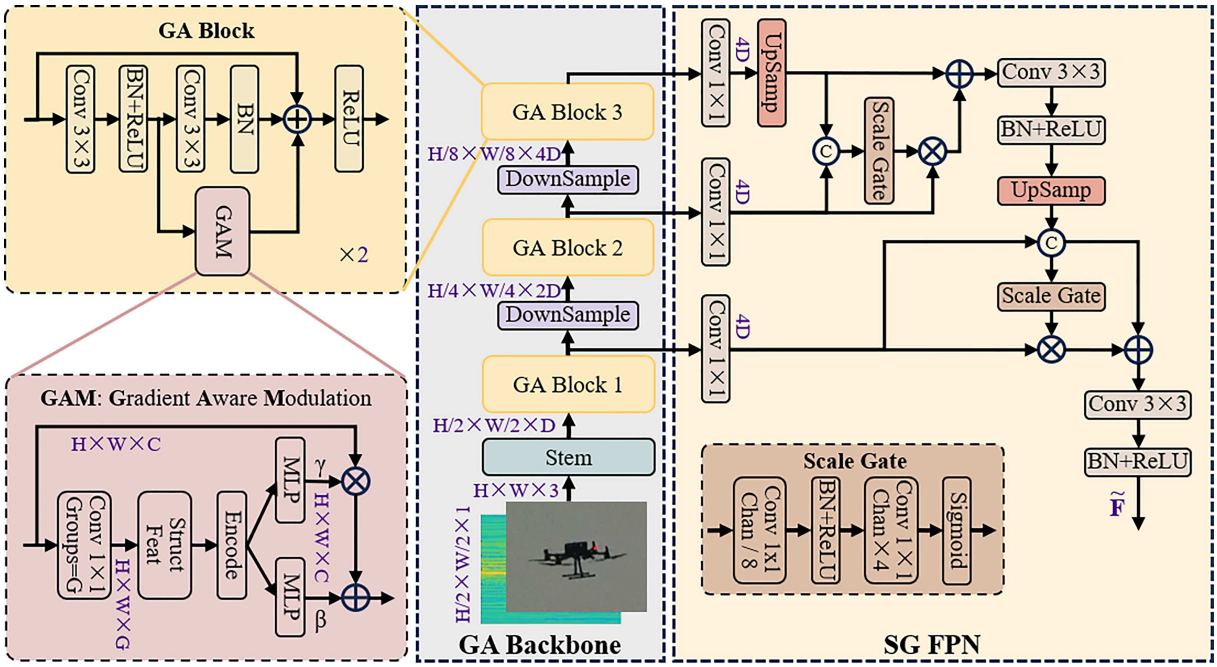

Figure 3 (Color online) Architecture of the gradient-aware and semantically-enhanced network (GSE Net). In the gradient-aware backbone, the GAM branch applies Sobel operators and structure tensor analysis to describe the orientation and coherence of local contour, texture, and micro-Doppler ridge structures. The resulting descriptors generate feature modulation parameters. The multi-scale outputs of the backbone are then processed by SG FPN, where high-level semantic features guide the selection and aggregation of informative low-level structural responses. By jointly modeling gradient structures and semantic context, GSE Net produces robust multi-scale representations that remain stable under time-varying heterogeneous degradation.

as texture and contour patterns in optical images and as the distribution, orientation, and continuity of micro-Doppler ridges in radar spectrograms. Gradient-aware and semantically-enhanced network captures and aggregates these cues through a gradient-aware backbone and a semantically-guided feature pyramid network, as illustrated in Figure 3.

3.2.1 *Gradient-aware backbone*

The backbone combines convolutional representation learning with explicit local structural modeling. Each block contains an identity branch, a convolution branch, and a gradient-aware modulation (GAM) branch. The first two branches preserve the input and extract general local responses, while the modulation branch encodes the orientation and coherence of optical contours, texture patterns, and micro-Doppler ridges.

The GAM branch first applies grouped convolution to obtain a compact feature map for structural analysis:

$$F_{\mathrm{grp}} = \mathrm{Conv}_{\mathrm{grp}=G}(F_{\mathrm{GAMin}}), \tag{2}$$

where F_{GAMin} denotes the input feature map, F_{grp} is the output of the grouped convolution, and G is the number of groups in the convolution. This operation reduces feature dimensionality while retaining the local response patterns required for the subsequent gradient computation.

Local variation in F_{grp} is then measured along the horizontal and vertical directions using Sobel operators:

$$I_x = \mathrm{Sobel}_x(F_{\mathrm{grp}}), \quad I_y = \mathrm{Sobel}_y(F_{\mathrm{grp}}), \tag{3}$$

where $\mathrm{Sobel}_x(\cdot)$ and $\mathrm{Sobel}_y(\cdot)$ represent the Sobel filtering operations along the horizontal and vertical directions, respectively. The gradient maps I_x and I_y do not serve as independent edge responses. Their joint statistics are needed to determine whether a local neighborhood contains an organized contour or ridge and to characterize its dominant orientation.

A structure tensor is therefore constructed at spatial location p:

$$J_\rho(p) = \mathrm{Gauss}_\rho * \begin{bmatrix} I_x^2(p) & I_x(p)I_y(p) \\ I_x(p)I_y(p) & I_y^2(p) \end{bmatrix}, \tag{4}$$

where $J_\rho(p)$ is the structure tensor at spatial location p, Gauss_ρ denotes a Gaussian smoothing kernel with scale ρ, and $*$ indicates convolution.

Eigen-decomposition of the structure tensor yields the principal direction θ and the coherence $c \in [0, 1]$. The principal direction indicates the dominant orientation of local feature variation, while the coherence measures the directional consistency of the neighborhood. Organized contours and ridges tend to produce relatively high coherence values, whereas flat regions and irregular responses produce lower values. These descriptors are encoded and passed through a multi-layer perceptron (MLP) to generate the modulation parameters γ and β.

$$\gamma, \beta = \text{MLP}([\sin 2\theta, \cos 2\theta, c]). \tag{5}$$

The parameters are then applied to the input feature

$$F_{GAMout} = \gamma \cdot F_{GAMin} + \beta. \tag{6}$$

The modulation is therefore determined by local orientation and coherence. Responses associated with organized contour, texture, or ridge structures can be strengthened, whereas flat regions and irregular local responses receive less emphasis. Together with the identity and convolution branches, the modulation branch allows the backbone to retain target-dependent structural patterns while preserving the flexibility of data-driven feature learning.

3.2.2 *Semantically-guided feature pyramid network*

Backbone features at different levels contain complementary information. Shallow features preserve fine contour, texture, and ridge details, but they also contain strong background responses. Deeper features provide clearer target semantics at a lower spatial resolution. The feature pyramid network therefore uses top-down semantic guidance to select and aggregate informative low-level structural features.

Given the K backbone outputs, 1×1 pointwise convolutions first generate the lateral features $L^{(1)}, L^{(2)}, \ldots, L^{(K)}$ and place features from different levels in a common channel space. The aggregation starts from the highest-level feature $L^{(K)}$. After being upsampled, $L^{(K)}$ is concatenated with the adjacent low-level feature $L^{(K-1)}$, and the combined feature is processed by a scale gate attention module to generate the spatial attention map $A^{(K-1)}$.

The attention map modulates $L^{(K-1)}$ through element-wise multiplication. Fine-grained structures that agree with the high-level semantic response are retained, while irrelevant low-level background responses are reduced. The modulated feature is then fused with the upsampled high-level feature to obtain $E^{(K-1)}$. This enhanced feature serves as the semantic input to the next lower level, and the same operation is repeated down the pyramid until the final multi-scale representation $\tilde{F}$ is obtained.

The feature pyramid network performs semantically-guided multi-scale feature aggregation. Rather than simply combining features across scales, high-level semantic information is leveraged to selectively enhance informative low-level details. Given multi-scale backbone outputs, 1×1 pointwise convolutions are first applied to generate lateral features $L^{(1)}, L^{(2)}, \ldots, L^{(K)}$, where K denotes the number of feature levels extracted by GSE Net. A top-down fusion process is then conducted as follows.

This progressive aggregation is suited to low-altitude target recognition because the useful structures do not occur at a fixed scale. Optical target size varies with target type, imaging distance, and camera focal length, while the spectral extension and local distribution of micro-Doppler ridges differ across target categories and motion states. By using high-level semantics to guide each fusion step, the feature pyramid network retains fine-grained structures that support target recognition without indiscriminately introducing all low-level responses.

3.3 Confidence-driven dynamic gating

The relative quality of radar and optical observations varies across samples, making fixed fusion weights sensitive to severely degraded inputs. Confidence-driven dynamic gating module addresses this problem by constructing bidirectional cross-modal interaction features, estimating sample-level confidence from these features, and using the resulting scores to control their fusion.

Given the multi-scale features $\tilde{F}_r$ and $\tilde{F}_o$ extracted by GSE Net, independent linear projections are first applied to obtain the query, key, and value embeddings of the radar modality, denoted as Q_r, K_r, and V_r, and those of the optical modality, denoted as Q_o, K_o, and V_o. Cross-attention is then computed in both directions:

$$F_{r\to o} = \text{Softmax}\left(\frac{Q_r \cdot K_o^{\mathrm{T}}}{\sqrt{d_k}}\right) \cdot V_o, \quad F_{o\to r} = \text{Softmax}\left(\frac{Q_o \cdot K_r^{\mathrm{T}}}{\sqrt{d_k}}\right) \cdot V_r, \tag{7}$$

where d_k denotes the token dimension and T indicates matrix transposition. $F_{r\to o}$ represents optical information selected under radar guidance, whereas $F_{o\to r}$ represents radar information selected under optical guidance. The two interaction directions describe how each modality supports the shared target semantics and provide the basis for relative confidence estimation.

Global average pooling (GAP) is applied to the cross-attention outputs to obtain the descriptors $\bar{\mu}_r$ and $\bar{\mu}_o$. The descriptors are concatenated and passed through a lightweight MLP followed by a Softmax to generate the confidence score α:

$$\begin{aligned} \bar{\mu}_r &= \mathrm{GAP}(F_{r\to o}), \\ \bar{\mu}_o &= \mathrm{GAP}(F_{o\to r}), \\ \alpha &= \mathrm{Softmax}(\mathrm{MLP}([\bar{\mu}_r, \bar{\mu}_o])). \end{aligned} \tag{8}$$

The confidence scores are then used to fuse the two interaction features:

$$\begin{aligned} T_{mm} &= \alpha(1)\cdot F_{r\to o} + \alpha(2)\cdot F_{o\to r}, \\ T_f &= \mathrm{Concat}(T_{mm}, Q_r). \end{aligned} \tag{9}$$

The weighted feature T_{mm} adjusts the contribution of the two interaction directions according to the estimated sample-level confidence, while the radar feature Q_r is further concatenated with T_{mm} to retain radar information that remains robust under diverse environmental conditions. The resulting feature T_f is used by the classification head for target recognition.

3.4 Cross-modal semantic alignment

Radar spectrograms and optical images represent target motion and appearance in fundamentally different forms, and time-varying heterogeneous degradation may further enlarge their feature discrepancy. Cross-modal semantic alignment module is therefore introduced during training to project the two modalities into a common semantic space and optimize their class-level correspondence through two sequential stages: cross-modal alignment followed by global semantic refinement.

Global average pooling is applied to the multi-scale features $\tilde{F}_r$ and $\tilde{F}_o$ to obtain global representations g_r and g_o, which are then mapped into a low-dimensional contrastive space through a shared projection head:

$$T_{pr} = \phi(g_r), \quad T_{po} \quad = \phi(g_o), \tag{10}$$

where T_{pr} and T_{po} denote the projected radar and optical embeddings, respectively.

Stage 1: Cross-modal alignment. At the beginning of training, the dominant problem is the separation between radar and optical feature distributions. The module therefore restricts the positive relation to same-class samples from different modalities. For an anchor radar embedding T^i_{pr}, only optical embeddings with the same class label are used as positive samples. The positive set is defined as

$$P_{\mathrm{cross}}(i) = \{T^j_{po} \mid C(i) = C(j)\}, \tag{11}$$

where $C(i)$ denotes the class label of sample i. This constraint directly reduces the cross-modal distance between samples that share the same target semantics, without allowing the larger number of intra-modal relations to dominate the early optimization.

Stage 2: Global semantic refinement. After cross-modal correspondence has been established, the positive set is expanded to include same-class embeddings across both modalities. For an anchor T^i_p, which can be either T^i_{pr} or T^i_{po}, the positive set is defined as

$$P_{\mathrm{full}}(i) = \{T^j_{pr}, T^j_{po} \mid C(i) = C(j)\} \setminus \{T^i_p\}. \tag{12}$$

The expanded relation jointly constrains cross-modal and intra-modal samples. It further reduces same-class variation and increases the separation between different target categories across the complete multimodal feature space.

The two stages use the same supervised contrastive loss and differ only in the definition of the positive set $P(i)$:

$$\mathcal{L}_{\mathrm{align}} = \sum_{i=1}^{2B} \frac{-1}{|P(i)|} \sum_{T^+_p \in P(i)} \ln \frac{\exp(T^i_p \cdot T^+_p/\tau)}{\sum_{k=1}^{2B} \mathbf{1}_{[k\neq i]} \exp(T^i_p \cdot T^k_p/\tau)}, \tag{13}$$

where B is the batch size, τ denotes the temperature parameter, and $P(i)$ is the positive set associated with the anchor T^i_p at the current training stage. The sequential positive-set design first resolves the cross-modal discrepancy that directly affects feature interaction and then organizes the full embedding space according to target categories.

3.5 Loss function

The overall training objective of RORF Net combines target classification, cross-modal semantic alignment, and environment-invariant representation learning:

$$\mathcal{L} = \mathcal{L}_{\text{cls}} + \lambda_{\text{align}}\mathcal{L}_{\text{align}} + \lambda_{\text{adv}}\mathcal{L}_{\text{adv}}, \tag{14}$$

where $\mathcal{L}_{\text{cls}}$ supervises target recognition, $\mathcal{L}_{\text{align}}$ is the supervised contrastive loss used by the semantic alignment module, and $\mathcal{L}_{\text{adv}}$ reduces the dependence of the learned representation on environmental conditions.

The classification loss is the standard cross entropy between the predicted class distribution y_{pr} and the ground-truth label y_{gt}:

$$\mathcal{L}_{\text{cls}} = -\sum_{i=1}^{C} y_{\text{gt}}(i) \ln \left(\frac{\exp(y_{\text{pr}}(i))}{\sum_{j=1}^{C} \exp(y_{\text{pr}}(j))} \right), \tag{15}$$

where C denotes the number of target categories.

The alignment loss $\mathcal{L}_{\text{align}}$ follows (13). Its positive set changes from $P_{\text{cross}}(i)$ to $P_{\text{full}}(i)$ according to the two-stage training strategy, so that cross-modal correspondence is established before the global class structure is refined.

For $\mathcal{L}_{\text{adv}}$, an auxiliary environment classifier predicts the sensing condition from the learned features using standard cross entropy. A gradient reversal layer reverses the gradient passed from this classifier to the feature extractor, discouraging the representation from encoding environment-specific information while preserving its discriminability for target recognition.

4 Experimental results and analysis

4.1 Experimental platform

This study is conducted using a radar-guided and optical-confirmed cooperative surveillance system deployed at the outfield test site in Dongying, Shandong Province, China. The radar and optical systems are spatially separated by approximately 400 m, as shown in Figure 4. The site provides a representative environment for low-altitude target recognition in open-area scenarios, featuring open terrain, ground-clutter-dominated radar backgrounds, and frequent bird activity.

The radar operates at a pulse repetition frequency of 3.8 kHz with a coherent processing interval of 133 ms (512 pulse accumulations). Short-time Fourier transform (STFT) is applied to the radar slow-time signals using 224 pulses per frame with a sliding step of one pulse, producing 224×224 radar time-frequency spectrograms. The normalized logarithmic amplitude spectrum is used as the radar feature representation. Upon receiving guidance information from the radar, the optical system captures target images at a frame rate of 30 Hz. The captured images are cropped into 448×448 patches centered on the detected target region.

To ensure accurate spatial alignment between the radar and optical systems, high-precision integrated positioning and orientation modules are employed. This module provides latitude, longitude, altitude and three-axis attitude (yaw, pitch, roll), which can align the absolute geodetic coordinates and axis orientations of both sensors. Clock synchronization between the two sensing systems is ensured via the precision time protocol (PTP). During data preprocessing, radar time-frequency spectrograms and optical images are temporally matched according to their nearest timestamps to construct paired multimodal samples.

Various low-altitude targets, including rotary-wing UAVs, birds, and bionic flapping-wing UAVs, were observed at distances ranging from 900 to 2000 m. The collected targets exhibit diverse motion behaviors, including hovering, straight-flight, and maneuvering, thereby covering representative target dynamics encountered in open-area low-altitude surveillance.

4.2 Dataset construction

4.2.1 *Data collection under diverse conditions*

The constructed dataset is designed to cover typical threat objects encountered in open-area low-altitude surveillance and comprises three primary categories with eight subcategories: rotary-wing UAVs (e.g., DJI Phantom 4, DJI M350), birds, and bionic flapping-wing UAVs. A total of 8940 paired radar-optical samples are collected under clear conditions. The detailed sample distribution is summarized in Table 1.

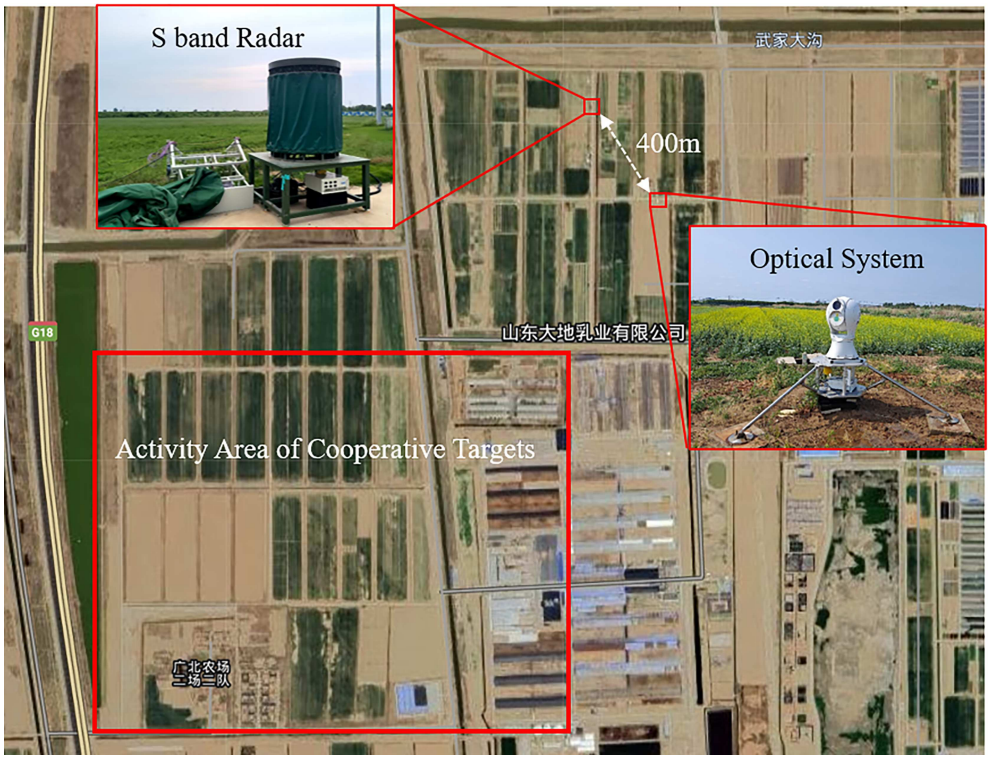

Figure 4 (Color online) Experimental equipment displacement.

Table 1 Sample distribution of low-altitude targets by category and condition.

Target	Sample number		
	Clear	Foggy	Dusk
DJI M350	537	72	112
DJI T100	648	28	54
DJI AVATA	216	39	47
DJI Mini	221	48	41
DJI Flip	706	45	61
Autel Lite+	742	48	50
Bionic UAV	2569	80	256
Birds	3301	291	272
Total	8940	651	893

Additionally, 651 foggy samples and 893 dusk samples were acquired to evaluate robustness under adverse environments. Since the quantity of real degraded samples remains relatively limited compared with the clear-condition dataset, these samples are reserved exclusively for testing.

4.2.2 *Degradation characterization and physics-guided dataset augmentation*

Due to the limited availability of real degraded samples, we employ a physics-guided simulation strategy to augment the dataset. Specifically, we first characterize real-world degradation patterns and then synthesize corresponding degraded samples. To analyze optical degradation under different environmental conditions, representative samples of various target categories are selected from Clear, Foggy, and Dusk scenes to examine pixel intensity distributions. For each sample, the target region is manually annotated, while the remaining pixels are treated as background. Figure 5 illustrates the corresponding intensity histograms.

Across all environmental conditions, target pixels exhibit a bimodal gray-scale distribution. The darker mode mainly corresponds to dark-colored targets and the non-illuminated regions of bright-colored targets, whereas the brighter mode is associated with surfaces receiving stronger illumination and reflection. Under clear conditions, the target distribution remains well separated from the background, resulting in pronounced contour gradients and well-preserved texture details. Under foggy conditions, atmospheric scattering reduces target-background contrast

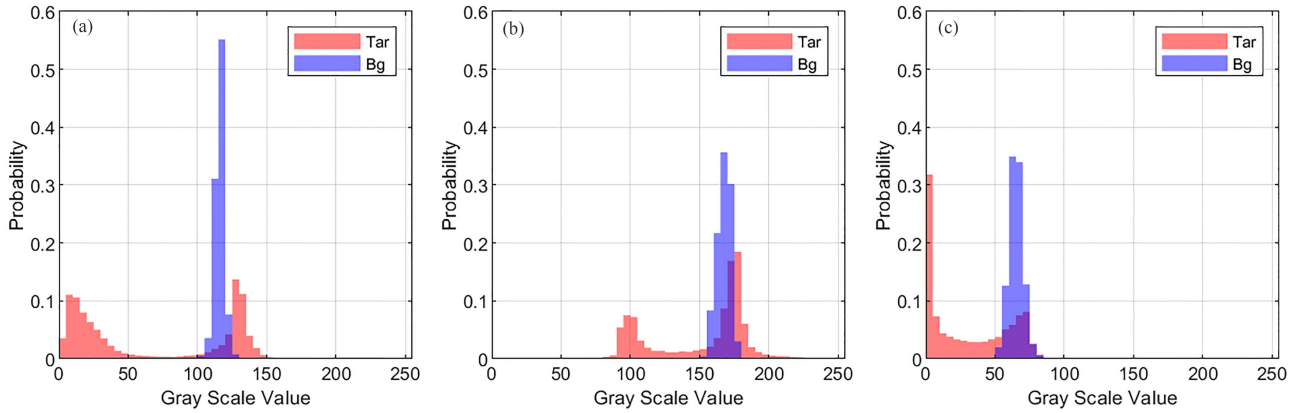

Figure 5 (Color online) Gray-scale distribution of target (red) and background (blue) pixels under different conditions. (a) Clear; (b) Foggy; (c) Dusk.

Table 2 Composition of the constructed Clear, Foggy, and Dusk datasets.

	Clear	Foggy	Dusk
Train	7147	7147	7147
Test	1793	2444	2686

and causes the target distribution to gradually overlap with the surrounding scene. Under dusk conditions, the overall luminance of both targets and background decreases substantially due to reduced illumination, causing the two modes to shift toward lower gray-scale values and making targets appear as dark silhouettes with severely degraded contour and texture information.

Guided by the above analysis, we expand all 8940 measured clear-condition samples into foggy and dusk conditions. The degradation parameters are calibrated to align with the statistical characteristics observed in the real degraded samples, while additional randomness is introduced to improve diversity and cover a broader range of degradation severities.

• Optical degradation simulation. For the foggy conditions, synthetic haze layers with randomly varying transparency are superimposed on optical images. Meanwhile, brightness is slightly increased and contrast is reduced to simulate the visual characteristics associated with fog densities. For the dusk conditions, motion blur or Gaussian blur is randomly applied with a probability of 40% each, where the blur severity is jointly determined by the simulated illumination level and target velocity obtained from radar measurements. Subsequently, brightness and contrast are reduced to reproduce the combined effects of low light and motion-induced artifacts typically observed in dusk scenarios.

• Radar degradation simulation. The target echo intensity varies with the radar cross section (RCS) and observation geometry, while the clutter intensity depends on background scattering characteristics. We therefore extract clutter segments with randomly varying intensity recorded at the same deployment site and independently inject them into the target echoes. For each optically degraded sample, a corresponding radar sample is generated with an independently sampled clutter injection, constructing a complete set of paired multimodal degraded data.

The three constructed datasets are denoted as Clear, Foggy, and Dusk datasets. The Clear dataset is directly formed using the 8940 measured clear-condition samples, while the Foggy and Dusk datasets each contain the same number of synthetically degraded paired samples. Each dataset condition is independently divided into training and testing subsets using an 8:2 ratio. Finally, the real samples collected under foggy and dusk conditions are added exclusively to the testing subsets. The final composition of the constructed datasets after data splitting is summarized in Table 2.

4.2.3 *Quantitative evaluation of degradation severity*

To quantitatively assess the degradation severity of the constructed Foggy and Dusk datasets, we evaluate both optical and radar modalities using complementary metrics, as summarized in Table 3.

For optical images, the structural similarity index measure (SSIM) is employed to evaluate image quality degradation. Specifically, SSIM is computed between each degraded optical image and its corresponding image in the clear condition, which serves as the reference. The Foggy dataset maintains relatively high SSIM values (0.90 ± 0.03),

Table 3 Statistics of degradation levels in the foggy and dusk conditions relative to the Clear dataset.

Indicator	Foggy	Dusk
Image SSIM	0.90 ± 0.03	0.73 ± 0.05
Radar SCNR (dB)	21.94 ± 12.18	20.62 ± 13.99

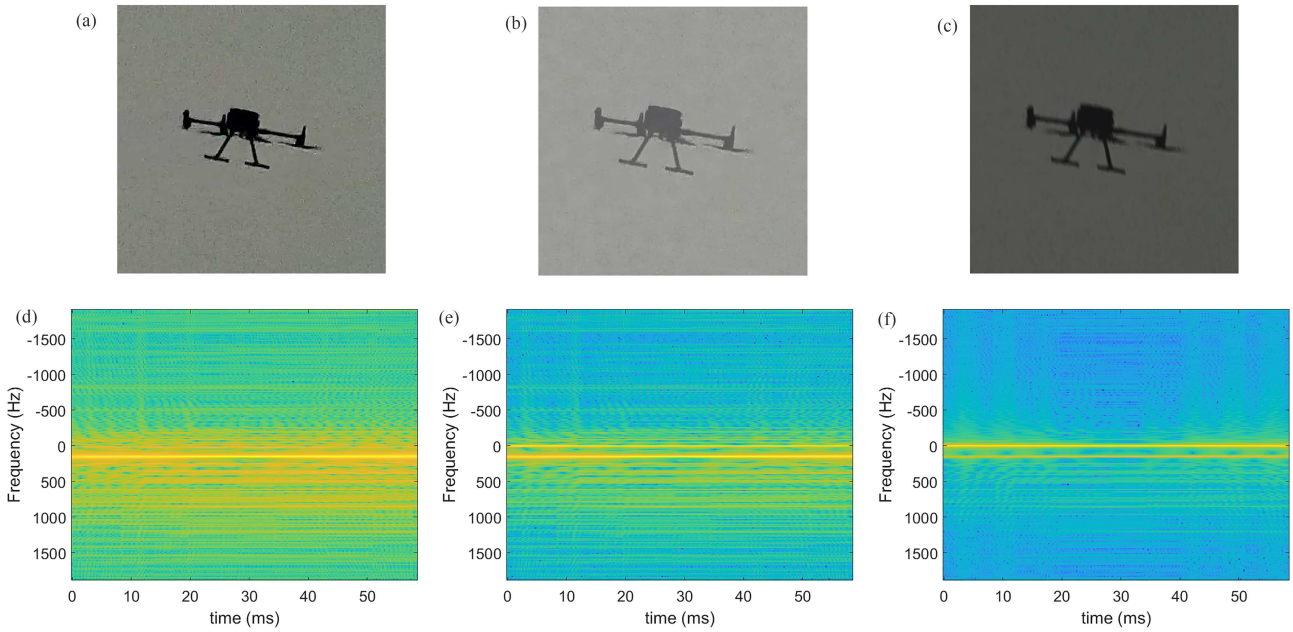

Figure 6 (Color online) Visual examples of modality degradation under different environmental conditions. (a) Clear image; (b) foggy image with SSIM = 0.90; (c) dusk image with SSIM = 0.73; (d) clear radar spectrogram; (e) radar spectrogram with 21 dB SCNR; (f) radar spectrogram with 9 dB SCNR.

indicating that although contrast and brightness are altered, the underlying structural information remains largely preserved. In contrast, the Dusk dataset exhibits substantially lower SSIM values (0.73 ± 0.05), reflecting severe structural distortion caused by the combined effects of blur and low light. For radar signals, the signal-to-clutter-plus-noise ratio (SCNR) is adopted to quantify the severity of clutter interference. SCNR is computed in the range-Doppler domain as the ratio between the target-echo power and the average clutter-plus-noise power in neighboring cells, where a protection window is used to exclude the target response from the background estimation. The Foggy and Dusk datasets exhibit mean SCNR values of 21.94 and 20.62 dB, respectively, with relatively large standard deviations (12.18 dB and 13.99 dB). These statistics indicate considerable variability in clutter and noise contamination across samples, confirming that the measured-clutter injection strategy produces diverse radar data qualities and effectively reflects the time-varying nature of practical radar interference.

Figure 6 presents representative visual examples of the constructed degradation scenarios with explicit SSIM and SCNR levels. For optical images, the Foggy dataset (SSIM = 0.90) mainly reduces scene visibility while still preserving a considerable amount of texture details. In contrast, the Dusk dataset (SSIM = 0.73) leads to a significant decrease in overall brightness and produces more severe blurring effects, resulting in noticeable loss of fine structural details. For radar signals, clutter interference directly affects the visibility of target micro-Doppler signatures in the time-frequency spectrogram. When SCNR decreases to 21 dB, the micro-Doppler patterns begin to be disturbed by clutter components. When it further increases to 9 dB, the micro-Doppler ridges become visually weak and partially obscured, making reliable feature extraction considerably more challenging.

4.3 Implementation details

4.3.1 *Network configuration*

For the backbone, the base channel dimension is set to $D = 64$, and the number of feature levels is set to $K = 3$. In the GAM branch, the grouped convolution uses four groups. The Gaussian kernel employed in the structure tensor computation has a kernel size of 3 and a variance of 1. In the feature pyramid network, the output dimension of the lateral convolution layers is set to 256.

Table 4 Recognition performance of different modalities on the clear condition (trained on clear condition). The best results are in bold.

Class	RORF Net radar only	RORF Net optical only	RORF Net fusion
DJI M350	97.22%	98.15%	99.07%
DJI T100	26.15%	100.00%	100.00%
DJI AVATA	22.73%	86.36%	72.73%
DJI Mini	48.89%	91.11%	97.78%
DJI Flip	94.37%	100.00%	99.30%
Autel Lite+	96.64%	100.00%	100.00%
Bionic UAV	78.99%	85.80%	99.81%
Birds	98.49%	99.70%	99.55%
OA	83.99%	95.15%	**98.94%**
AA	70.43%	95.14%	**96.03%**
Kappa	0.7856	0.9357	**0.9861**

For the confidence-driven dynamic gating module, the number of attention heads is set to 8, and the query, key, and value dimensions are all set to 64.

The network is optimized using the AdamW optimizer with an initial learning rate of 1×10^{-4}, which is gradually annealed to 1×10^{-6} using a cosine decay schedule over 20 epochs.

Training is performed in two stages. In Stage 1, the model is trained for 15 epochs with loss weights $\lambda_{\text{align}} = 0.8$ and $\lambda_{\text{adv}} = 0.0$. In Stage 2, training continues for 35 epochs with loss weights $\lambda_{\text{align}} = 0.5$ and $\lambda_{\text{adv}} = 0.1$.

4.3.2 *Evaluation metrics*

Three metrics are used to assess classification performance: overall accuracy (OA), average accuracy (AA) and the Kappa coefficient, defined as

$$\begin{aligned} \text{AA} &= \frac{1}{C}\sum_{i=1}^{C}\frac{TP_i}{S_i}, \\ \text{OA} &= \frac{\sum_{i=1}^{C} TP_i}{S_{all}}, \\ \text{Kappa} &= \frac{\text{OA} - p_e}{1 - p_e}, \quad p_e = \frac{\sum_{i=1}^{C} S_i \cdot b_i}{S_{all}^2}, \end{aligned} \tag{16}$$

where TP_i denotes the number of correctly classified samples for class i, S_i represents the total number of samples in class i, b_i is the number of samples predicted as class i, and S_{all} is the total number of samples in the test set. The Kappa coefficient is used to measure classification performance in imbalanced datasets, with values closer to 1 indicating better classification reliability and improved balance across classes.

4.4 Performance evaluation

4.4.1 *Modality contribution analysis*

We evaluate three input configurations to examine the contribution of each sensing modality: radar-only recognition, optical-only recognition, and full multimodal fusion. All models are trained on the Clear training set. Performance on the Clear test set reflects recognition under matched sensing conditions, whereas evaluation on the Foggy and Dusk test sets measures cross-condition generalization under previously unseen degradation.

Table 4 reports the results on the clear condition. The optical-only model achieves an OA of 95.15%, and an AA of 95.14%, significantly outperforming the radar-only configuration. Under clear visibility, optical images preserve detailed target shape, contour, and texture information and are therefore highly discriminative for distinguishing specific UAV types. Radar remains useful for classes whose visual appearance is similar but whose motion characteristics differ. Combining the two modalities increases the OA to 98.94% and the Kappa coefficient to 0.9861, indicating that optical appearance and radar micro-Doppler information provide complementary evidence under favorable sensing conditions. In particular, the motion dynamics encoded in radar spectrograms help resolve visually similar categories such as birds and bionic UAVs, leading to more robust recognition than either modality alone.

The difference between the two modalities becomes more pronounced on the Foggy test set, as shown in Table 5.

Evaluation results on the foggy condition are shown in Table 5. The OA of the optical-only model decreases from 95.15% to 25.67%, while the radar-only model retains an OA of 80.81%. This collapse is not caused simply by a reduction in image brightness. Atmospheric scattering and scene whitening reduce target-background contrast and

Table 5 Recognition performance of different modalities on the foggy condition (trained on clear condition). The best results are in bold.

Class	RORF Net radar only	RORF Net optical only	RORF Net fusion
DJI M350	90.56%	0.00%	88.33%
DJI T100	23.42%	56.33%	31.65%
DJI AVATA	26.51%	0.00%	36.14%
DJI Mini	51.61%	10.75%	59.14%
DJI Flip	95.72%	0.00%	95.72%
Autel Lite+	90.36%	0.00%	86.29%
Bionic UAV	71.21%	16.84%	80.81%
Birds	96.64%	45.17%	95.06%
OA	80.81%	25.67%	**83.01%**
AA	68.51%	16.22%	**71.65%**
Kappa	0.7510	0.0295	**0.7782**

Table 6 Recognition performance of different modalities on the dusk condition (trained on clear condition). The best results are in bold.

Class	RORF Net radar only	RORF Net optical only	RORF Net fusion
DJI M350	93.18%	23.98%	95.00%
DJI T100	24.46%	27.72%	54.35%
DJI AVATA	28.57%	72.53%	54.95%
DJI Mini	51.16%	59.30%	69.77%
DJI Flip	95.57%	14.29%	93.60%
Autel Lite+	90.45%	39.20%	80.40%
Bionic UAV	70.26%	75.71%	77.92%
Birds	95.50%	66.24%	90.89%
OA	79.01%	57.53%	**82.54%**
AA	68.84%	47.37%	**77.11%**
Kappa	0.7304	0.4641	**0.7720**

suppress the fine texture and contour cues on which the optical model trained under clear conditions largely relies. For bright-colored targets, the target intensity distribution can overlap strongly with the foggy background, making target boundaries difficult to distinguish. The resulting loss of visual structure, together with the distribution shift between clear training images and foggy test images, leads to severe class confusion.

The full RORF Net achieves an OA of 83.01% on the Foggy test set, exceeding both single-modality configurations. Although the optical observation is severely degraded, it can still contain partial contour or appearance information for some samples. The confidence-driven dynamic gating module reduces the contribution of unreliable optical features rather than discarding the modality completely, allowing the fused representation to retain useful complementary evidence while relying primarily on the more stable radar observation.

A similar but less extreme pattern appears under the dusk condition in Table 6. Low illumination and motion blur reduce the optical-only OA to 57.53%, whereas the radar-only model achieves 79.01%. The full model reaches 82.54%. Under this condition, fine texture information is weakened and local contour details become blurred. Nevertheless, coarse contours and partial shape cues remain visible in some samples, allowing the optical modality to retain a certain degree of discriminative information. As the degradation severity varies with illumination, target motion, and the resulting blur, leading to considerable sample-to-sample fluctuations in optical reliability. By retaining useful optical cues when they remain available and reducing their contribution when they become unreliable, the full RORF Net achieves an OA of 82.54%, outperforming both single-modality configurations.

Taken together, these results show that the performance of radar-optical fusion depends on sensing quality. Optical features dominate under clear visibility, whereas radar provides greater stability under fog and low illumination. A fusion model with fixed modality contributions cannot accommodate these changes, leading to degraded recognition performance. These findings motivate the design of adaptive fusion mechanisms capable of perceiving degradation severity and dynamically adjusting modality importance for robust low-altitude target recognition.

4.4.2 *Comparison with state-of-the-art methods*

To further evaluate the effectiveness of the proposed method, we compare RORF Net with several representative multimodal fusion approaches, including DPFFN [33], Panopticon [34], and LRFusionPR [18]. The selected baselines comprehensively utilize concatenation-based and attention-based methods to perform multimodal feature fusion,

Table 7 Performance comparison. The best results are in bold.

Class	DPFFN [33]			Panopticon [34]			LRFusionPR [18]			RORF Net		
	Clear	Foggy	Dusk	Clear	Foggy	Dusk	Clear	Foggy	Dusk	Clear	Foggy	Dusk
DJI M350	95.37%	73.33%	80.91%	94.44%	84.44%	79.09%	100.00%	95.56%	37.73%	100.00%	98.33%	95.91%
DJI T100	99.23%	79.75%	55.43%	100.00%	94.94%	73.37%	99.23%	94.30%	10.87%	100.00%	96.84%	86.96%
DJI AVATA	68.18%	83.13%	75.82%	59.09%	62.65%	30.77%	81.82%	71.08%	76.92%	81.82%	62.65%	93.41%
DJI Mini	71.11%	58.06%	68.60%	57.78%	59.14%	61.63%	80.00%	84.95%	75.58%	100.00%	77.42%	93.02%
DJI Flip	99.30%	97.86%	95.57%	92.25%	80.21%	74.88%	97.89%	96.26%	73.89%	100.00%	99.47%	94.58%
Autel Lite+	97.32%	81.22%	93.97%	95.97%	81.22%	30.15%	97.32%	92.39%	24.12%	97.32%	97.46%	87.94%
Bionic UAV	85.21%	87.37%	67.79%	95.14%	95.96%	87.01%	99.61%	94.44%	94.81%	99.81%	99.66%	94.55%
Birds	95.16%	92.86%	96.25%	98.03%	94.54%	94.32%	99.24%	93.49%	94.53%	99.55%	98.95%	96.78%
OA	91.86%	87.03%	82.24%	94.53%	89.57%	80.12%	98.22%	92.96%	76.25%	**99.11%**	**96.81%**	**94.34%**
AA	88.86%	81.70%	79.29%	86.59%	81.64%	66.40%	94.39%	90.31%	61.06%	**97.31%**	**91.35%**	**92.89%**
Kappa	0.8926	0.8304	0.7700	0.9280	0.8636	0.7426	0.9765	0.9079	0.6925	**0.9882**	**0.9583**	**0.9268**
Total OA		86.42%			87.19%			87.74%			**96.45%**	
Total AA		82.60%			77.01%			79.88%			**93.30%**	
Total Kappa		0.8230			0.8330			0.8415			**0.9537**	

which have demonstrated competitive performance in related multimodal perception tasks.

All of the methods are trained on the training set of clear and foggy conditions and evaluated on the whole test set, including Clear, Foggy and Dusk conditions. The quantitative results are summarized in Table 7.

Compared with representative multimodal fusion methods, the proposed RORF Net achieves the best overall performance across all metrics. Specifically, RORF Net attains a total OA of 96.45%, a total AA of 93.30%, and a total Kappa of 0.9537. In comparison, the second-highest total OA is obtained by LRFusionPR at 87.74%, giving an improvement of 8.71%. For total AA, RORF Net exceeds the second-best result, obtained by DPFFN, by 10.70%. Its total Kappa coefficient is 0.1122 higher than that of LRFusionPR, which obtains the second-highest value of 0.8415.

The condition-specific results provide a clearer view of this improvement. RORF Net achieves OA values of 99.11%, 96.81%, and 94.34% on the Clear, Foggy, and Dusk test sets, respectively. Its OA decreases by only 4.77 percentage points from Clear to Dusk. Over the same conditions, the decreases for DPFFN, Panopticon, and LRFusionPR are 9.62%, 14.41%, and 21.97%, respectively. The advantage of RORF Net is therefore not limited to higher performance under matched conditions; it also maintains a substantially smaller performance loss as the degradation pattern changes.

Such performance gains can be attributed to the joint design of structure-aware feature extraction and adaptive multimodal fusion. Unlike most existing methods that mainly improve the fusion strategy while relying on conventional backbone networks, the proposed framework introduces a structure-aware backbone that explicitly enhances structural patterns in both modalities. Through the modeling of local structural information, the network strengthens responses associated with coherent micro-Doppler ridges in radar spectrograms and contour structures in optical images. These structural cues remain more stable than raw spectral amplitude or pixel intensity under degradation, enabling the network to learn more discriminative and robust target representations. As a result, more discriminative and robust representations reflecting the intrinsic physical characteristics of low-altitude targets can be extracted.

Furthermore, the proposed confidence-driven dynamic gating module enables reliability-aware fusion by dynamically regulating the contributions of radar and optical features. In practical surveillance environments, the sensing quality of different modalities may fluctuate due to factors such as fog, low light, and clutter interference. Conventional fusion strategies typically combine multimodal features using fixed patterns, which may introduce unreliable information when one modality becomes severely degraded. In contrast, this module estimates modality reliability and adaptively adjusts their fusion weights, allowing the network to better exploit complementary information while suppressing degraded signals.

Overall, the experimental results demonstrate that the integration of structure-aware feature extraction and confidence-driven adaptive fusion significantly enhances recognition robustness, enabling RORF Net to achieve state-of-the-art performance under time-varying heterogeneous degradation.

In addition to recognition accuracy, we further evaluate the computational efficiency of the proposed method. The model complexity of different approaches is compared in terms of the number of parameters and floating-point operations (FLOPs), as summarized at the bottom of Table 8.

Table 8 Parameter counts and FLOPs on different network. The best results are in bold.

	DPFFN [33]	Panopticon [34]	LRFusionPR [18]	RORF Net
Params	15.34M	87.23M	24.70M	**14.63M**
FLOPs	**5.09G**	69.35G	314.42G	53.51G

Table 9 Ablation study. The best results are in bold.

Test set condition	Indicator	RORF Net w/o GAM	RORF Net w/o SG FPN	RORF Net w/o CDG	RORF Net w/o CSA	RORF Net w/o adversarial loss	RORF Net
Clear	OA	93.08%	73.95%	98.55%	98.55%	96.65%	**99.11%**
	AA	87.50%	47.71%	96.72%	97.02%	86.95%	**97.31%**
	Kappa	0.9099	0.6372	0.9809	0.9809	0.9557	**0.9882**
Foggy	OA	88.72%	73.43%	92.05%	91.85%	90.22%	**96.81%**
	AA	82.06%	46.61%	89.08%	88.12%	76.14%	**91.35%**
	Kappa	0.8543	0.6297	0.9102	0.9056	0.8805	**0.9583**
Dusk	OA	73.76%	68.54%	74.03%	69.45%	76.53%	**94.34%**
	AA	61.14%	43.63%	65.12%	53.07%	54.09%	**92.89%**
	Kappa	0.6647	0.5592	0.6808	0.5895	0.6912	**0.9268**

As shown in Table 8, RORF Net maintains a moderate parameter scale (14.63M), which is only slightly smaller than DPFFN (15.34M). This result indicates that the proposed architecture does not rely on excessively large model capacity to achieve its performance improvements. Its computational cost (53.51G FLOPs) is higher than DPFFN but substantially lower than Panopticon and LRFusionPR. The increased FLOPs mainly arise from the GAM module, which uses eigenvalue decomposition to compute structure tensors at every spatial location of both radar spectrograms and optical images. Although these operations are computationally intensive, they are essential for extracting robust, discriminative structural features and directly contribute to RORF Net's superior recognition performance.

Overall, these results indicate that the increased computational cost of RORF Net primarily stems from its structure-aware feature extraction, yet the network achieves substantial performance gains without requiring a large model size. Future work may investigate more efficient strategies for structure-aware feature learning to reduce computational overhead while preserving the benefits of incorporating structural inductive bias.

4.4.3 *Ablation study*

To further investigate the effectiveness of each component in the proposed framework, a series of ablation experiments are conducted. Starting from the RORF Net, the key modules of RORF Net are selectively removed to evaluate their individual contributions. The quantitative results are summarized in Table 9.

Removing the GAM branch leads to a clear classification performance drop across all test sets. Under foggy and dusk conditions, the OA decreases from 96.81% and 94.34% to 88.72% and 73.76%, respectively. This confirms that gradient-aware structural features extracted from radar spectrograms play an important role in improving robustness under degraded sensing conditions. By explicitly modeling local gradient structures, the GAM module enhances the representation of micro-Doppler ridge patterns and visual contours, which are essential for reliable target recognition.

Replacing SG FPN with simple multi-scale concatenation results in the most severe performance degradation, with OA dropping to around 70% in each test set. This observation highlights the importance of semantically guided multi-scale aggregation for capturing hierarchical target features and maintaining stable feature representations under varying degradation patterns.

Removing the CDG module leads to a dramatic reduction in generalization ability, particularly on the Dusk test set where OA drops from 94.34% to 74.03%. This demonstrates the significance of dynamic reliability estimation and adaptive modality weighting when sensing conditions are degraded.

Similarly, removing the CSA module causes the OA to decrease to 69.45% on the dusk condition, while removing the adversarial loss leads to an OA of 76.53% on the same condition. These results confirm that cross-modal semantic alignment and environment-invariant representation learning both contribute meaningfully to the overall robustness of the framework.

Overall, these ablation results confirm that each component contributes complementary benefits to the final performance of RORF Net. Their joint optimization enables robust multimodal recognition under time-varying heterogeneous degradation.

5 Conclusion

This paper investigates the problem of robust radar-optical fusion for low-altitude target recognition under time-varying heterogeneous degradation. The proposed RORF Net preserves local structural information in radar spectrograms and optical images, reduces the semantic gap between the two modalities during training, and adjusts their fusion contributions according to sample-level confidence. These operations jointly address the loss of discriminative information within individual modalities and the changing quality balance between paired observations. Experiments on the radar-optical dataset covering clear, foggy, and dusk conditions show that RORF Net achieves a total OA of 96.45%, a total AA of 93.30%, and a total Kappa coefficient of 0.9537. The total OA exceeds the best competing method by 8.71%, while the smaller performance decrease from clear to degraded conditions demonstrates improved cross-condition robustness. The ablation results further show that gradient-aware structural modeling, semantically guided multi-scale aggregation, cross-modal alignment, confidence-driven fusion, and environment-invariant learning make complementary contributions to this performance.

The current framework operates on spatio-temporally paired radar and optical observations and introduces additional computation for structure tensor analysis. Future work will investigate asynchronous temporal fusion that better reflects practical sensor operating modes and more efficient structure-aware architectures for real-time deployment.

Acknowledgements This work was supported in part by National Natural Science Foundation of China (Grant No. 62225104) and Key R&D Program of Shandong Province, China (Grant No. 2025CXGC010209).

References

1 State Council of the People's Republic of China. Report on the work of the government. 2025. https://english.www.gov.cn/news/202503/12/content_WS67d17f64c6d0868f4e8f0c10.html
2 Civil Aviation Administration of China. Guiding opinions on the application of surveillance technology in low altitude airspace in civil aviation. 2017. https://www.caac.gov.cn/English/News/202305/t20230515_218894.html
3 Chen W S, Niu H C, Wang X, et al. Review on multi-source detection technologies for birds and drones in airport clearance area (in Chinese). Acta Aeronautica et Astrona Sin, 2025, 46: 031251
4 Shi Z, Chang X, Yang C, et al. An acoustic-based surveillance system for amateur drones detection and localization. IEEE Trans Veh Technol, 2020, 69: 2731–2739
5 Liu Q, Liu R, Meng Q, et al. Theoretical model and key technologies for low-altitude aircraft positioning and sensing integrated with 6G communication networks (in Chinese). Sci Sin Inform, 2025, 55: 2401–2426
6 Bisio I, Garibotto C, Lavagetto F, et al. Blind detection: advanced techniques for WiFi-based drone surveillance. IEEE Trans Veh Technol, 2019, 68: 938–946
7 Zhang H, Xiao L, Cao X, et al. Multiple adverse weather conditions adaptation for object detection via causal intervention. IEEE Trans Pattern Anal Mach Intell, 2024, 46: 1742–1756
8 Chen V C, Fayin Li V C, Shen-Shyang Ho V C, et al. Micro-Doppler effect in radar: phenomenon, model, and simulation study. IEEE Trans Aerosp Electron Syst, 2006, 42: 2–21
9 Kang K B, Choi J H, Cho B L, et al. Analysis of micro-Doppler signatures of small UAVs based on Doppler spectrum. IEEE Trans Aerosp Electron Syst, 2021, 57: 3252–3267
10 Chen X, Chen W, Rao Y, et al. Progress and prospects of radar target detection and recognition technology for flying birds and unmanned aerial vehicles. J Radars, 2020, 9: 803–827
11 Semenyuk V, Kurmashev I, Lupidi A, et al. Advances in UAV detection: integrating multi-sensor systems and AI for enhanced accuracy and efficiency. Int J Crit Infrastructure Protection, 2025, 49: 100744
12 Chen X, Ma C, Zhao C, et al. UAV classification based on deep learning fusion of multidimensional UAV micro-Doppler image features. IEEE Geosci Remote Sensing Lett, 2024, 21: 1–5
13 Oh B S, Lin Z. Extraction of global and local micro-Doppler signature features from FMCW radar returns for UAV detection. IEEE Trans Aerosp Electron Syst, 2021, 57: 1351–1360
14 Ren J, Jiang X. Regularized 2-D complex-log spectral analysis and subspace reliability analysis of micro-Doppler signature for UAV detection. Pattern Recognition, 2017, 69: 225–237
15 Zhang W P, Li X, Liu Y X. Research on representation, estimation and recognition for radar targets with micro-motion (in Chinese). Sci Sin Inform, 2023, 53: 1469–1489
16 Nabati R, Qi H. Centerfusion: center-based radar and camera fusion for 3d object detection. In: Proceedings of 2021 IEEE Winter Conference on Applications of Computer Vision (WACV), Waikoloa, 2021. 1526–1535
17 Kim Y, Shin J, Kim S, et al. CRN: camera radar net for accurate, robust, efficient 3d perception. In: Proceedings of 2023 IEEE/CVF International Conference on Computer Vision (ICCV), Paris, 2023. 17569–17580
18 Qi Z, Cheng L, Zhou Z, et al. LRFusionPR: a polar BEV-based LiDAR-radar fusion network for place recognition. IEEE Robot Autom Lett, 2025, 10: 11784–11791

19 Fent F, Palffy A, Caesar H. DPFT: dual perspective fusion transformer for camera-radar-based object detection. IEEE Trans Intell Veh, 2025, 10: 4929–4941

20 Sun X, Tian Y, Lu W, et al. From single- to multi-modal remote sensing imagery interpretation: a survey and taxonomy. Sci China Inf Sci, 2023, 66: 140301

21 Wan Y, Ji J, Zhou F, et al. From static dense to dynamic sparse: vision-radar fusion-based UAV detection. IEEE Trans Inform Forensic Secur, 2025, 20: 4454–4468

22 Sakellariou N, Lalas A, Votis K, et al. Multi-sensor fusion for UAV classification based on feature maps of image and radar data. 2024. ArXiv:2410.16089

23 Li Z, Yang F, Wang C, et al. Target attribute recognition method based on image and track information fusion. J Ordnance Equipment Eng, 2024, 45: 232–237

24 He K, Sun J, Tang X. Single image haze removal using dark channel prior. In: Proceedings of 2009 IEEE Conference on Computer Vision and Pattern Recognition, 2009. 1956–1963

25 Lore K G, Akintayo A, Sarkar S. Llnet: a deep autoencoder approach to natural low-light image enhancement. 2016. ArXiv:1511.03995

26 Jiang Y, Gong X, Liu D, et al. EnlightenGAN: deep light enhancement without paired supervision. IEEE Trans Image Process, 2021, 30: 2340–2349

27 Guo C, Li C, Guo J, et al. Zero-reference deep curve estimation for low-light image enhancement. In: Proceedings of 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020. 1777–1786

28 Al Sobbahi R, Tekli J. Comparing deep learning models for low-light natural scene image enhancement and their impact on object detection and classification: overview, empirical evaluation, and challenges. Signal Process-Image Commun, 2022, 109: 116848

29 Hu C, Yan Y, Wang R, et al. High-resolution, multi-frequency and full-polarization radar database of small and group targets in clutter environment. Sci China Inf Sci, 2023, 66: 227301

30 Park K, Kim Y, Kim D, et al. Resilient sensor fusion under adverse sensor failures via multi-modal expert fusion. In: Proceedings of 2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Nashville, 2025. 6720–6729

31 Hu P, Qian Y, Zheng T, et al. t-READi: transformer-powered robust and efficient multimodal inference for autonomous driving. IEEE Trans Mobile Comput, 2025, 24: 135–149

32 Wei K, Dai J K, Hong D, et al. MGFNet: an MLP-dominated gated fusion network for semantic segmentation of high-resolution multi-modal remote sensing images. Int J Appl Earth Observation GeoInf, 2024, 135: 104241

33 Zhou Y, Liu S, Gao H W, et al. A dual-polarization feature fusion network for radar automatic target recognition based on HRRP sequences. IEEE Trans Aerosp Electron Syst, 2025, 61: 14642–14655

34 Waldmann L, Shah A, Wang Y, et al. Panopticon: advancing any-sensor foundation models for earth observation. In: Proceedings of 2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), Nashville, 2025. 2195–2205