ORIGINAL RESEARCH Open access

Visible-Infrared Multi-Modal Fusion Framework for LSS-UAV Detection in Complex Environments

Maria Gonzales[1*], Jose Santos[1], Anna Cruz[2], Carlos Reyes[3], Luis Torres[4]

Abstract

To resolve the technical difficulties associated with recognizing Low-Slow-Small Unmanned Aerial Vehicle (LSS-UAV) swarm formations—namely, weak signal signatures coupled with strong-feature environmental interference in complex scenes—a deep learning-driven visible-infrared multi-modal fusion recognition strategy is introduced. The approach leverages deep learning architectures to independently capture shape-based patterns in visible imagery and heat-based radiation patterns in infrared imagery. A tiered multi-modal fusion blueprint that unifies feature-level and decision-level integration is devised, embedding an Environment-Aware Dynamic Weighting (EADW) module alongside Dempster-Shafer evidence theory (D-S). This blueprint harnesses the mutual reinforcements of feature-level and decision-level merging. Consequently, both detection/recognition performance and the system's resilience against complex backgrounds are substantially improved for LSS-UAV swarms. Experimental findings indicate that the developed solution achieves 93.5% detection accuracy for LSS-UAV swarms in intricate urban landscapes, yielding an average gain of 18.7% relative to unimodal baselines. In comparison, the false alarm rate drops to 4.2%. The technique also demonstrates considerable adaptability to varying environmental conditions, maintaining high efficacy even under demanding conditions such as darkness and haze. An efficient and trustworthy technical pathway for spotting LSS-UAV swarm formations is thereby provided.

Keywords Deep learning, Low-slow-small UAVs, Multi-modal fusion, Attention mechanism, Target detection, D-S evidence theory

*Correspondence:
Maria Gonzales
maria.gonzales@up.edu.ph
[1] Department of AI for Banana and Pineapple Disease Monitoring, Faculty of Agriculture, University of the Philippines Los Baños, Laguna, Philippines
[2] Department of Crop Water Stress Index Prediction, Faculty of Agriculture, University of Santo Tomas, Manila, Philippines
[3] Department of Smart Fertilization Management, Faculty of Agriculture, Ateneo de Manila University, Quezon City, Philippines
[4] Department of Precision Harvest Scheduling, Faculty of Agriculture, University of the Philippines Mindanao, Davao City, Philippines

Introduction

Over recent years, LSS-UAVs have found broadened utility across civilian and military spheres owing to their minimal bulk (wingspan < 1 m), reduced cruise velocity (5–15 m/s), low ceiling (< 500 m), and capacity for swarm-based coordinated operations [1]. Nevertheless, the rapid expansion of LSS-UAV deployments has heightened low-altitude safety risks, confronting established low-altitude surveillance infrastructure with serious obstacles [2-4]. Legacy unimodal sensing approaches reveal considerable shortcomings in cluttered or ambiguous environments: radar-based methods often falter due to ground clutter and ornithological false targets [5]; visible-band detection is heavily contingent on ambient brightness, with reliability deteriorating markedly under nocturnal or poorly lit settings [6]; thermal infrared sensing suffers pronounced sensitivity to background temperature shifts, with metropolitan heat-island patterns inflating false positive counts [7]. As a direct result, multi-modal fusion sensing technologies endowed with environmental

This is an open access article distributed under the terms of the Creative Commons Attribution-NonCommercial-ShareAlike 4.0 International License (CC BY-NC-SA 4.0), which permits non-commercial use, distribution, reproduction, and adaptation in any medium or format, provided that appropriate credit is given to the original author(s) and source, a link to the license is provided, and any changes made are indicated. Adapted material must be distributed under the same license.

© 2023 The Author(s). | Published by Agro Intelligence Research Press

resilience have become both a focal research pursuit and an inexorable trend within the LSS-UAV detection landscape [8-12].

Contemporary investigations into LSS-UAV detection cluster around two principal trajectories: refining unimodal sensing technologies and fusing multi-modal information streams. Within the unimodal refinement track, investigators have advanced radar sensitivity via improved signal conditioning (e.g., wavelet-based noise reduction, pulse compression) [13] or have harnessed deep network architectures (e.g., YOLO family, Faster R-CNN) to boost optical object detection outcomes [14, 15].

In the realm of multimodal information integration, prevailing fusion strategies fall into three categories: data-level, feature-level, and decision-level. Frequently combined modalities include visible and infrared (RGBT), vision and acoustics, and vision and radar. Li *et al.* [16] applied generative adversarial network (GAN)-driven image blending methods to enrich input representations, refining the YOLOv5 detector with an attention mechanism to increase UAV detection rates markedly. Gao and Lin [17] explored the use of satellite remote sensing cartography to inject environmental context as ancillary input into radar-based detection, embedding background features into the decision process to strengthen both precision and robustness. Guo *et al.* [18] proposed a multi-scale differential attention fusion detection network that, by deriving scene illumination parameters and local target contrast, steered the differential attention module toward deep cross-modal and intra-modal feature fusion across visible and thermal bands, thereby boosting air-to-ground drone identification in dim conditions. Reference [19] developed a combined sensing system integrating optical devices, acoustic arrays, and radar, deploying the MUSIC algorithm for acoustic source triangulation, augmenting coverage with radar technology, and merging short-wave infrared with visible sensor streams for precise UAV coordinate estimation. Taken together, multimodal fusion paradigms have made notable strides in LSS-UAV detection. Even so, persistent hurdles remain, including pronounced inter-modal representation gaps, multi-scale target fluctuations, and dynamically shifting detection backdrops. What is more, numerous present-day fusion frameworks are short of both adaptive controllers for responding to transient environmental variations and principled uncertainty calculus for resolving contradictory decisions, factors that are indispensable for dependable real-world deployment [20].

To address the aforementioned gaps, this paper advances a visible-infrared multimodal fusion detection framework grounded in deep learning. Its principal novel contributions consist of: (1) Engineering specialized, noise-tolerant deep learning backbones (Visible: Enhanced ResNet-50 + DPAM; Infrared: Dual-path ConvNeXt-Tiny + UNet) dedicated to pulling out distinctive unimodal representations under entangled interference conditions; (2) Devising an original layered fusion construct that intertwines feature-level with decision-level integration. Within this construct, an Environment-Aware Dynamic Weighting (EADW) controller adaptively reconciles cross-modal feature streams contingent on instantaneous environmental conditions. At the same time, D-S evidence theory rigorously synthesizes feature-stage outputs with standalone unimodal decisions, enabling uncertainty governance and conflict arbitration. This formulation directly addresses the drawbacks of weak environmental adaptability, insufficient cross-modal information exchange, and vulnerability to contradictory decisions that afflict conventional fusion pipelines, yielding marked gains in both accuracy and stability for LSS-UAV swarm detection in complex scenes; (3) Furnishing exhaustive empirical corroboration and dissection on an openly available reference dataset, showcasing clear superiority in performance and ruggedness vis-à-vis leading contemporary methodologies.

Deep Learning-Based Multi-Modal Feature Extraction

Relevant deep learning theories and methods

Deep learning enables high-efficiency feature representation and decision-making from complex data by constructing multilayered nonlinear neural networks. Its central tenet is to progressively adjust network weights via error backpropagation, extracting abstract representations level by level to bypass the manual feature-engineering bottlenecks of classical machine learning [21]. Archetypal deep network families include Convolutional Neural Networks (CNNs), Recurrent Neural Networks (RNNs), and Transformers. CNNs are well-suited for image-domain processing, harnessing

© 2023 The Author(s). | Published by Agro Intelligence Research Press

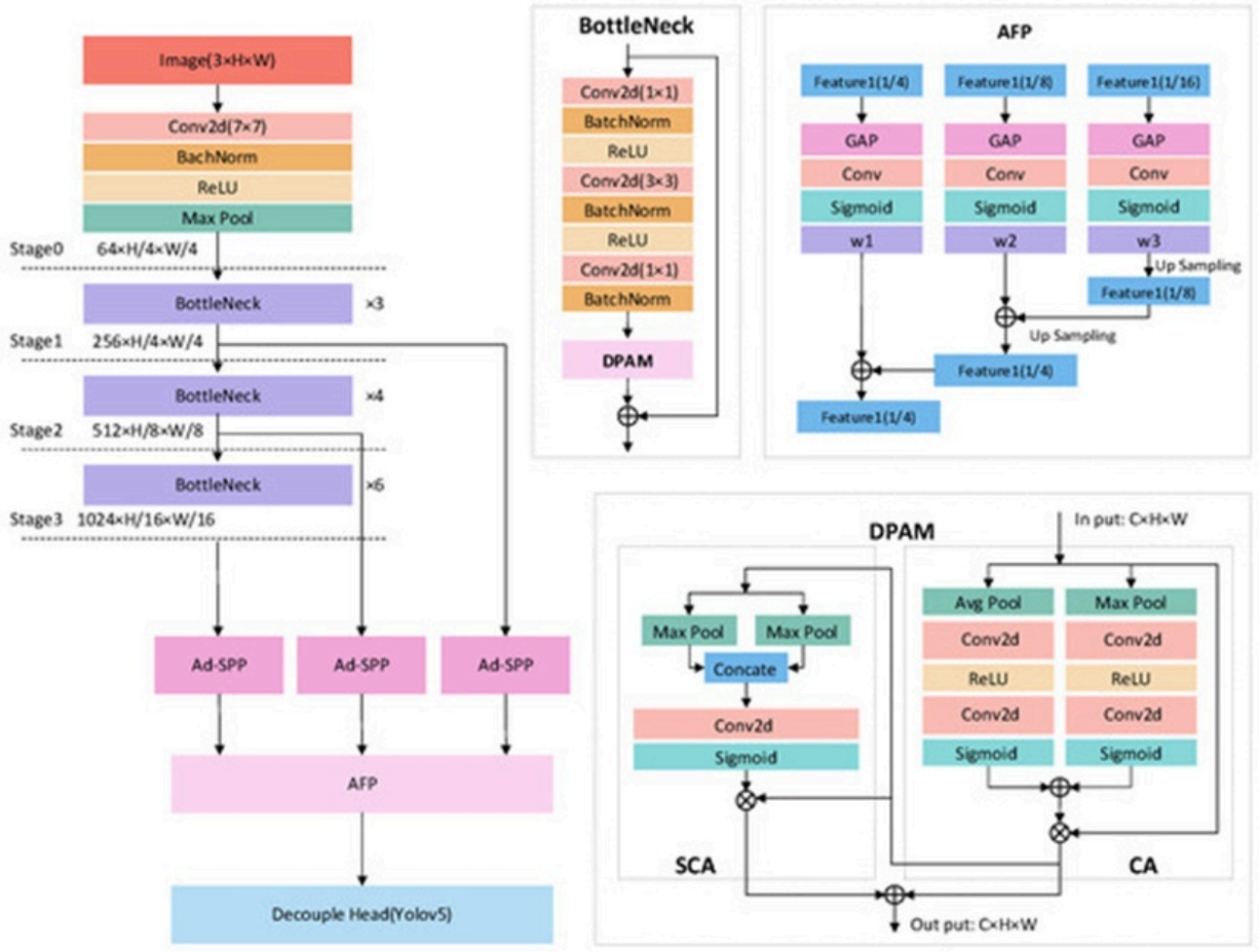
convolutional filters to isolate localized spatial motifs and leveraging pooling layers to reduce dimensionality while introducing translational invariance. RNNs accommodate sequential data by encapsulating temporal linkages through cyclical connectivity; Long Short-Term Memory (LSTM) and Gated Recurrent Unit (GRU) variants amplify long-range retention via gating mechanisms, making them apt for modeling continuous UAV flight paths. Transformers, built upon self-attention mechanisms, dismantle recursion-imposed constraints in sequence handling through holistic context modeling, affording efficient linkage of multi-scale features and presenting clear merits for cross-modal data registration [22].

Photoelectric target feature extraction based on deep learning

In cluttered environments, UAV swarm entities captured in visible-light frames often exhibit degraded textural definition, distracting scene artifacts, and overlapping, dense formations, posing challenges for conventional image analytics pipelines. Detection engines rooted in deep learning circumvent these difficulties by embedding attention mechanisms and erecting hierarchical feature abstraction networks to mine multi-granular textural attributes of LSS entities, enabling effective swarm localization under adverse conditions.

Deep learning network design

The architecture adopts a modified ResNet-50 backbone (with Stage 4 discarded to preserve higher spatial-resolution feature maps while reducing computational burden), combined with an Adaptive Feature Pyramid (AFP) component and a Dual-Path Attention Module (DPAM), as shown in **Figure 1**.



© 2023 The Author(s). | Published by Agro Intelligence Research Press

Figure 1. Schematic diagram of the visible light multi-scale texture feature extraction network structure.

The DPAM presented in this study reworks existing dual-path attention paradigms [23, 24] by computing channel- and spatial-oriented attention in a concurrent rather than successive manner. Serial designs risk prematurely filtering out informative signal components; DPAM circumvents this pitfall through cross-layer skip linkages and an adaptive gating function, thereby safeguarding and boosting mutually reinforcing cues. This characteristic proves especially useful for picking up the weak, multi-resolution signatures typical of LSS-UAV formations.

DPAM units are integrated into ResNet-50's first through third stages within the backbone, where cross-layer skip pathways help retain fine-scale structural detail. The feature map spatial dimensions emerging from these stages amount to 1/4, 1/8, and 1/16 of the original input extent, respectively. The aggregation stage deploys Adaptive Spatial Pyramid Pooling (Adaptive-SPP), under which the pooling kernel dimensions adapt dynamically:

$$\mathbf{size}_k = \left\lfloor s \cdot s_p^{(k)} \right\rfloor \quad (1)$$

Here, size$k$ corresponds to the kernel span chosen for the k-th pooling tier, s gives the spatial measure of the incoming feature map, and (k) constitutes a learnable proportionality parameter.

Target detection builds on a modified YOLOv5 base, splitting classification and coordinate regression into separate processing streams via a Decoupled Head and adopting the GIoU loss formulation to sharpen bounding box placement accuracy:

$$\Psi_{GIoU} = 1 - IoU + \frac{\|C - B \cup B^{gt}\|}{\|C\|} \quad (2)$$

Where C represents the minimal bounding region that encloses both the predicted box B and the annotation box Bgt, while IoU=$\|B \cap Bgt\|$/$\|B \cup Bgt\|$ expresses the Intersection over Union ratio.

To handle fluctuations in lighting and background distractors, the Dual-Path Attention Module (DPAM) is introduced. Setting it apart from standard sequential attention pipelines that cascade channel recalibration and spatial recalibration (and may inadvertently dampen valuable responses) [23], and taking cues from findings on the potency of parallel attention constructs [24], DPAM is structured around two co-active processing arms: Channel Attention (CA) and Spatial Attention (SA). Operating side by side, these two arms concurrently refine feature representations along channel and spatial axes, markedly strengthening discriminative power for tiny, low-visibility objects situated within cluttered observation conditions.

Channel Attention (CA): Channel-wise weighting coefficients are derived by pooling information through Global Average Pooling (GAP) jointly with Global Max Pooling (GMP):

$$\mathrm{W}_c = \sigma(\mathrm{MLP}(\mathrm{GAP}(F)) + \mathrm{MLP}(\mathrm{GMP}(F))) \quad (3)$$

With (·) standing for the Sigmoid mapping and F representing the feature tensor being processed.

Spatial Attention (SA): Taking the channel-weighted tensor $\mathbf{F}ca$ as input, spatial weighting maps Ws are produced by combining the outputs of GAP and GMP along the channel axis and then applying a convolutional pass:

© 2023 The Author(s). | Published by Agro Intelligence Research Press

$$\mathbf{W}_s = \sigma\left(\mathbf{W}^{7\times7} * [\mathrm{GAP}(\mathbf{F}_{ca}); \mathrm{GMP}(\mathbf{F}_{ca})]\right) \quad (4)$$

Where (·) again denotes the Sigmoid mapping, $\mathbf{F}ca$ is the channel-recalibrated feature representation, $\mathbf{W}7\times7$ refers to a convolution operator with a 7 × 7 kernel, and ∗ signals the convolution operation.

The accompanying diagram shows the upgraded ResNet-50 backbone with DPAM blocks in Stages 1–3, along with the Adaptive Feature Pyramid (AFP) and the task-decoupled detection head. Principal functional blocks are marked for easy interpretation.

Dense target decoupling strategy

To address overlapping detection candidates arising from textural crowding in clustered formations, a Non-Maximum Suppression (NMS) procedure is integrated. The suppression cutoff si is tuned on-the-fly in relation to the localized object concentration metric ϱij:

$$s_t = \begin{cases} s_t\left(1-\rho_{tf}\right), \text{if } IoU\left(b_t, b_f\right) > \tau \\ s_t \qquad \text{otherwise} \end{cases} \quad (5)$$

Where ϱij acts as a measure of local object packing density, and τ establishes the baseline permissible overlap threshold.

Infrared target feature extraction based on deep learning

Drone clusters captured in infrared imagery introduce complications, including subdued radiant heat output, strong background thermal disturbances, and inconsistent object dimensions, leaving conventional infrared target-detection routines ill-equipped for use across diverse operational environments. Borrowing principles from the dim-target sensing framework described by Chi *et al.* [25], this work proposes a two-stream infrared small-target detection architecture that adds a compact detection stage after feature encoding to achieve end-to-end thermal target localization and labeling.

Deep learning network design

The architecture is structured as a two-stream design, pairing multi-resolution thermal-signal encoding with cross-modal attention aggregation (merging features from different depths within the infrared image stack), as depicted in **Figure 2**.

© 2023 The Author(s). | Published by Agro Intelligence Research Press

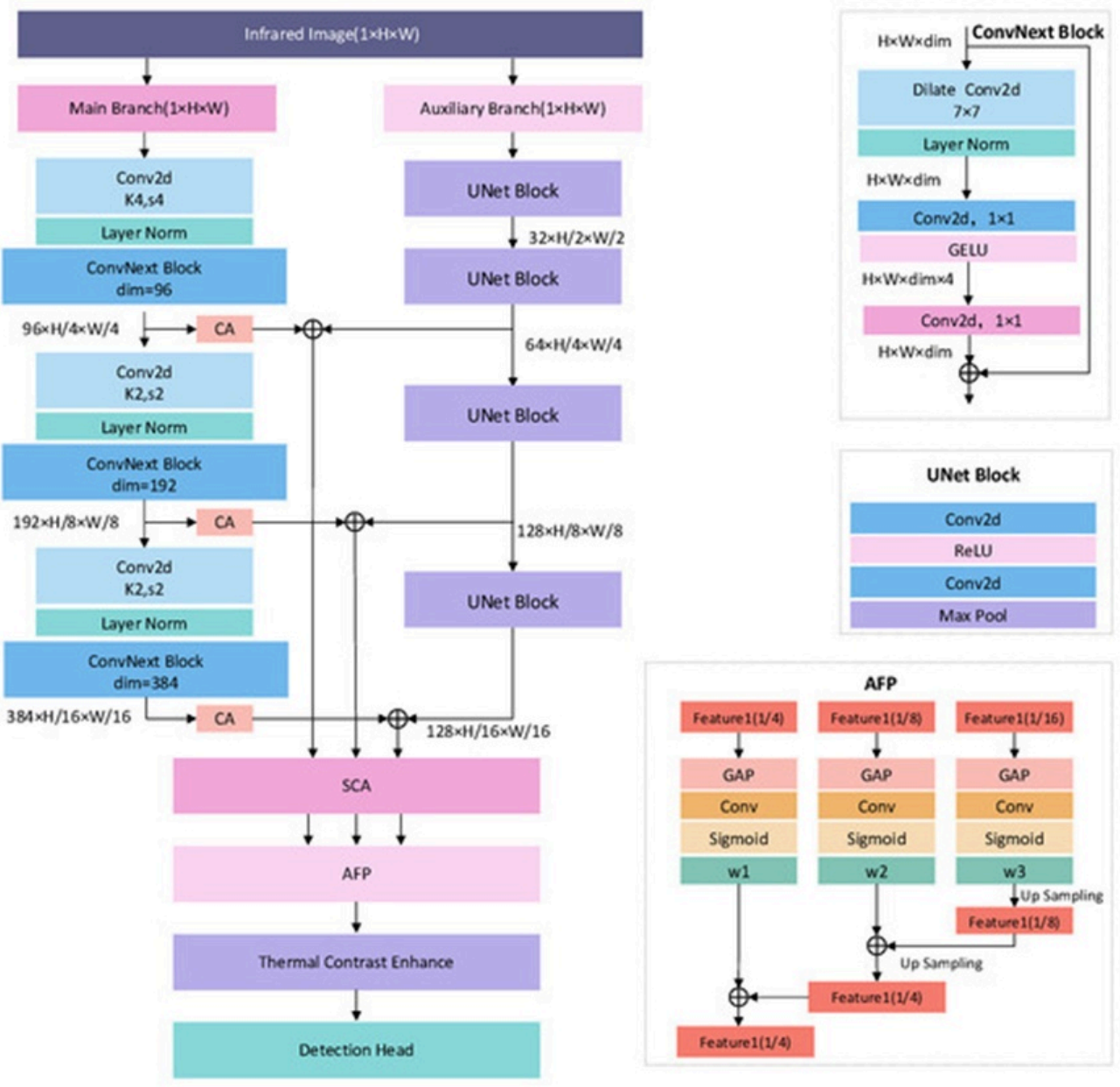


Figure 2. Schematic diagram of the infrared thermal radiation feature extraction network structure.

The diagram lays out the dual-pathway configuration, comprising a dominant pathway rooted in a reworked ConvNeXt-Tiny and a secondary pathway built from a lightweight UNet, followed by feature merging and a pared-down detection head.

The dominant pathway is based on a revised ConvNeXt-Tiny [26], augmented with dilated convolutional layers to broaden the receptive field. Its outputs cover spatial resolutions from 1/4 down to 1/16 of the input frame size. The secondary pathway runs a lightweight UNet-style module to extract thermal-gradient patterns while retaining fine-grained heat-distribution cues. Its output representation $F\text{thermal} \in \mathbb{R}C\times H\times W$ is spatially matched with the corresponding tiers of the dominant pathway. Signals coming from the dominant branch $F\text{main}$ and the secondary branch $F\text{thermal}$ are combined using Channel Attention (CA) along with Spatial Cross-Attention (SCA):

© 2023 The Author(s). | Published by Agro Intelligence Research Press

$$F_{\text{fused}} = \text{SCA}\left(F_{\text{main}}, F_{\text{thermal}}\right) + \text{CA}\left(F_{\text{main}} \otimes F_{\text{thermal}}\right) \quad (6)$$

Wherein SCA is expressed as:

$$\text{SCA}\left(Q, K, V\right) = \text{Softmax}\left(\frac{QW_Q(KW_K)^T}{\sqrt{d}}\right)VW_V \quad (7)$$

With WQ, WK, and WV constituting trainable projection weight arrays, and d acting as a dimensionality scaling constant.

Object decision-making is performed by a lightweight thermal detection head comprising three functional modules: bounding-box coordinate estimation (**B**final), object presence confidence (**O**final), and semantic class assignment (**C**final).

$$\begin{cases} \mathbf{B}_{\text{final}} = 0.1 \times \left(\mathbf{W}^{(2)} * \text{ReLU}\left(\mathbf{W}^{(1)} * \mathbf{X} + \mathbf{b}^{(1)}\right) + \mathbf{b}^{(2)}\right) \\ \mathbf{O}_{\text{final}} = \sigma\left(\mathbf{W}^{(obj)} * \mathbf{X} + \mathbf{b}^{(obj)}\right) \\ C_{\text{final}} = \sigma\left(\mathbf{W}^{(cls)} \cdot \text{Flatten}\left(\text{GAP}(\mathbf{X})\right) + b^{(cls)}\right) \end{cases} \quad (8)$$

Where $\mathbf{X} \in \mathbb{R}^{B \times C \times H \times W}$ denotes the incoming feature tensor, obj indicates the confidence estimation layer, cls indicates the category assignment layer, $\mathbf{W}(i)$ captures the convolution weight coefficients at layer i, b(i) captures the corresponding convolution bias offsets, $\sigma(\cdot)$ stands for the Sigmoid nonlinearity, ReLU(·) stands for the Rectified Linear Unit nonlinearity, and GAP(·) refers to the global average pooling operation. A module for boosting thermal contrast is placed immediately before the detection head:

$$F_{\text{enhanced}} = F_{\text{fused}} \otimes \left(1 + \text{Sigmoid}(\text{Conv}(F_{\text{fused}}))\right) \quad (9)$$

Multi-Scale detection and group target decoupling

To enhance feature representation across multiple scales, an Adaptive Feature Pyramid (AFP) is integrated, forming a dynamic scale-fusion conduit at the tail of the backbone. Its per-scale weighting factors ωs emerge from a gating operation:

$$\omega_s = \sigma(\text{MLP}(\text{GAP}(F_S))) \quad (10)$$

Within which FS stands for feature mappings drawn from distinct hierarchical depths, and (·) denotes the Sigmoid nonlinearity.

To cope with dense, overlapping swarming targets, the Density-Aware NMS rule **(Eq. 5)** is applied identically.

Multi-Modal Fusion Architecture Design

Conventional strategies for merging multiple modalities can be grouped into three tiers: data-level, feature-level, and decision-level fusion. Data-level schemes merge raw sensor streams immediately, without discarding any original

© 2023 The Author(s). | Published by Agro Intelligence Research Press

measurements. When inputs are strongly complementary and spatially synchronized, this tier delivers peak fusion quality; however, it is plagued by difficulties in spatiotemporal registration, heavy processing demands, and vulnerability to modality dropout. Feature-level fusion pursues mutual supplementation by abstracting high-level descriptors from each source, striking a compromise between preserving information content and keeping computation manageable. Yet it struggles to reconcile heterogeneous feature manifolds while risking the loss of useful signal during the merging process. Decision-level fusion pools verdicts formed independently within each modality. This paradigm yields fully decoupled modal processing, high tolerance against individual sensor faults, and supreme adaptability to surrounding conditions. Its drawbacks are the complete lack of cross-modal information sharing and the heightened likelihood of contradictory final outputs.

To address these shortcomings, the present work proposes a layered fusion blueprint that combines feature-level and decision-level merging **(Figure 3)**. In this scheme, a feature-level aggregation layer first sharpens feature salience through an Environment-Aware Dynamic Weighting (EADW) controller. Afterward, a decision-level aggregation layer applies D-S evidence theory to unify the unimodal decisions with the determination produced by the feature-fusion layer. The design thus directly addresses the dual pitfalls of absent inter-modal communication and clashing decisions, yielding stronger overall system resilience.

© 2023 The Author(s). | Published by Agro Intelligence Research Press

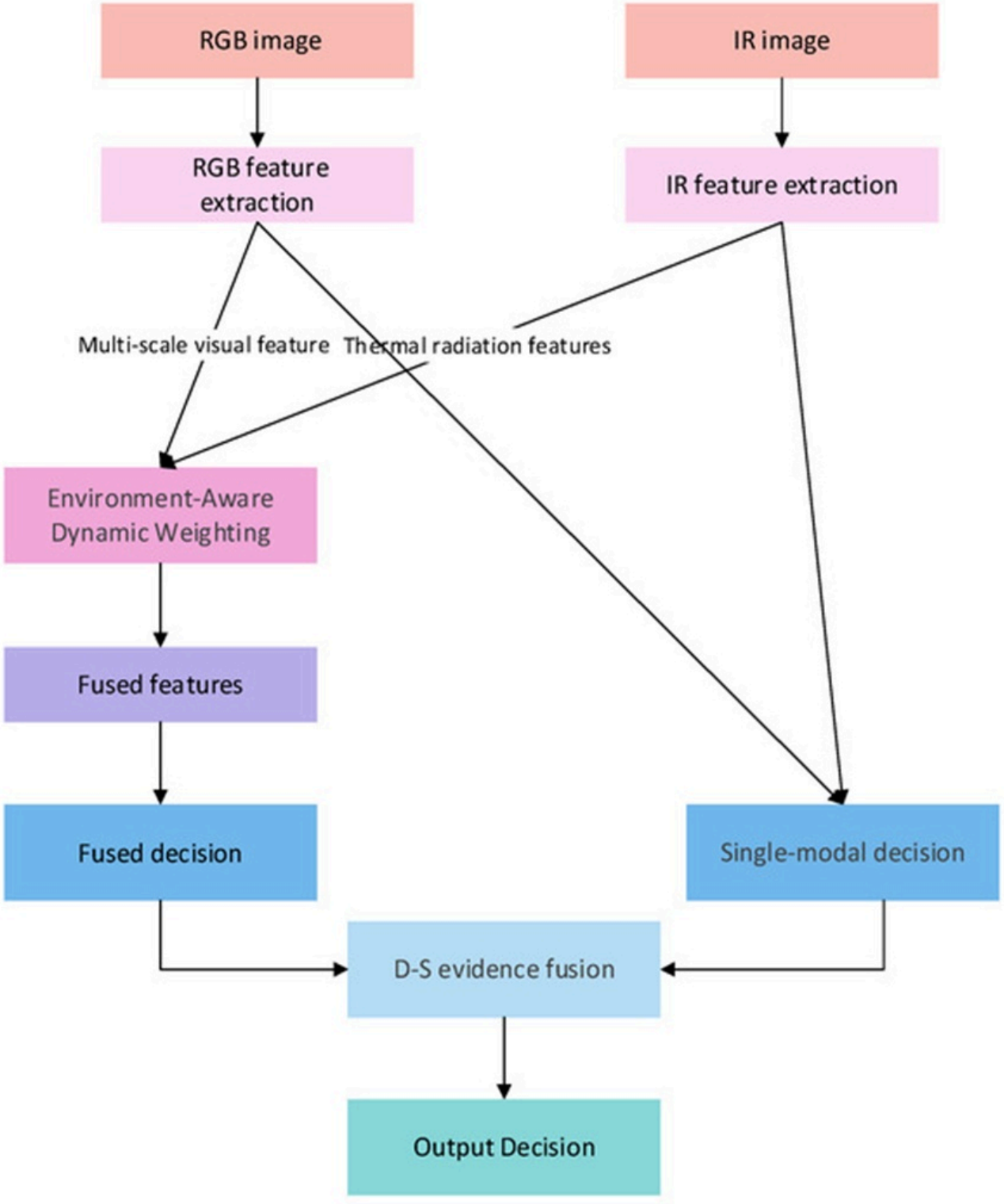


Figure 3. Hierarchical feature-level + decision-level fusion architecture.

The graphic outlines the complete processing chain: visible and infrared streams travel through their respective feature-extraction pipelines, advance into the EADW-governed feature-fusion layer, and finally reach the decision-fusion layer based on D-S evidence theory.

© 2023 The Author(s). | Published by Agro Intelligence Research Press

Feature fusion layer design

Recent literature [27] points out that fusion techniques responsive to dynamic scene conditions can markedly lift detection robustness in unpredictable environments. To overcome the collapse of conventional weighted averaging approaches when environments shift abruptly, an Environment-Aware Dynamic Weighting (EADW) fusion mechanism is introduced here. At its heart lies the idea of adaptively modulating each modality's contribution share during fusion as a function of real-time environmental state. The mechanism is built from three integral modules: perception of the environment, generation of dynamic weights, and feature alignment, followed by subsequent blending.

Dynamic environment awareness

A descriptor vector $E=[e1,e2,e3,e4,e5]T$ encoding environmental state is assembled, spanning illumination level, weather condition, and time-of-day factors. These quantities are inferred on the fly using onboard luminance sensors, publicly accessible weather API outputs (yielding visibility and humidity data), and system clock readings. The term $e1$ captures the illumination component, reflecting its effect upon the visible-band sensor [22].

$$e_1 = \frac{\log_{10}(L+1) - \log_{10}(L_{\min}+1)}{\log_{10}(L_{\max}+1) - \log_{10}(L_{\min}+1)} \quad (11)$$

Where L signifies the present illuminance reading (expressed in lux). Lmax represents the brightness reference under midday solar exposure (e.g., 100,000 lux), and Lmin represents the brightness reference under nighttime lunar illumination (e.g., 0.1 lux). The terms $e2,3,e4$ encapsulate weather-driven influences, quantifying atmospheric signal dampening under distinct meteorological settings (Clear, Rainy, Foggy). Their values are derived from sensor-recorded metrics:

$$\begin{bmatrix} e_2 \\ e_3 \\ e_4 \end{bmatrix} = \text{Softmax}\left(W_{weather} \cdot \begin{bmatrix} r/50 \\ v/10000 \\ h/100 \end{bmatrix} \right) \quad (12)$$

With r indicating rainfall rate (mm/h), v indicating visible range (m), h indicating relative humidity (%), and $Wweather$ being a learnable weighting array. The term $e5$ captures the temporal influence factor, describing how the hour within the diurnal cycle shapes background interference patterns for the infrared sensor.

$$e_5 = \frac{1}{2}\left[\sin\left(2\pi\frac{t}{24}\right) + 1\right] \quad (13)$$

With t standing for the local hour of day.

Dynamic weight generation

For every modality $m\in${Visible, Infrared}, a concatenated input vector $xm=[E,Pm]$ is built, where Pm denotes the average detection confidence over the current frame (supplied by the respective unimodal detection network). A two-layer fully connected neural architecture (MLP) computes a modality-specific scoring value hm:

© 2023 The Author(s). | Published by Agro Intelligence Research Press

$$\begin{cases} h_m = \mathrm{ReLU}\left(W_1 x_m + b_1\right) \\ e_m = W_2^T h_m + b_2 \end{cases} \quad (14)$$

In which $W1$, $W2$ are trainable weight matrices and $b1$, $b2$ are bias parameters. The adaptive fusion coefficient αm for modality m, following Softmax normalization, is defined as:

$$\alpha_m = \frac{\exp\left(e_m\right)}{\sum_{k\in\{V,I\}} \exp\left(e_k\right)} \quad (15)$$

The adaptive coefficient αm assigned to each modality m, derived via **Eq. 15**, is bounded between 0 and 1 after Softmax normalization, and the coefficients across all modalities add up precisely to 1.

Feature alignment and fusion

To prevent mismatches in dimensionality, each modality's representation Fm is first transformed linearly into a shared embedding space:

$$\tilde{F}_m = W_m F_m + b_m \quad (16)$$

Where $Wm \in \mathbb{R}^{df\times dm}$ constitutes a projection matrix, $\mathrm{bm} \in \mathbb{R}^{\mathrm{df}}$ constitutes a bias term, dm indicates the native feature size of modality m, and df indicates the target unified embedding size (kept at 512 throughout experimental runs). The combined output from the fusion stage is then given by:

$$F_{fused} = \sum_{m\in\{V,I\}} \alpha_m \cdot \tilde{F}_m \quad (17)$$

Decision fusion layer design

To exploit the global information synergy inherent in the feature-fusion verdict alongside the localized fidelity advantages of the standalone unimodal verdicts, while rigorously handling uncertainty and reconciling conflicting judgments, a complementary fusion strategy is proposed that merges the feature-fusion decision with the independent single-modality decisions. D-S evidence theory is invoked to perform the ultimate decision-level synthesis [28], as visualized in **Figure 4**.

© 2023 The Author(s). | Published by Agro Intelligence Research Press

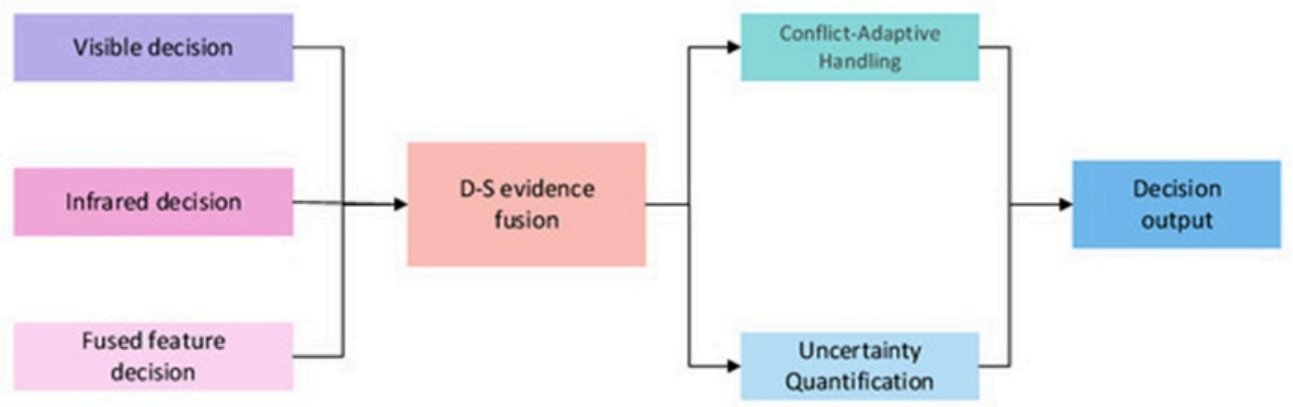


Figure 4. Schematic diagram of the decision fusion process.

The graphic elaborates the D-S evidence theory-centered fusion workflow: construction of BPA functions for the fusion-layer decision and each unimodal decision, aggregation of evidence bodies, conflict-adaptive treatment, and derivation of the final decision outcome.

Evidence space modeling

D-S evidence theory provides an uncertainty calculus whose effectiveness in arbitrating contradictory judgments has been demonstrated in recent investigations [27]. The frame of discernment is set as Θ={θ1,θ2}, where θ1 corresponds to "Target Present" and θ2 corresponds to "Target Absent." Its associated power set becomes 2Θ={Ø,θ1,θ2,{θ1,θ2}}. From the incoming decision streams, a three-way evidence space is established: E={Efused,EV,EI}, within which each constituent ES∈{fused,V,I} links to a dedicated Basic Probability Assignment (BPA) mapping ms:2Θ→[0, 1], obeying:

$$\begin{cases} \sum_{A \subseteq \Theta} m_s(A) = 1 \\ m_s(\emptyset) = 0 \end{cases} \quad (18)$$

Basic Probability Assignment (BPA) construction

Distinct BPA formulation protocols are adopted for the two input varieties: the combined feature fusion ruling $m$fused and the separate unimodal rulings mm.

The BPA belonging to the Fusion Decision ($m$fused) is built according to the following specification:

$$\begin{cases} m_{\text{fused}}(\{\theta_1\}) = \kappa_f \cdot \max\left(0, 2P_{\text{fused}}(\theta_1) - \lambda_f\right) \\ m_{\text{fused}}(\{\theta_2\}) = \kappa_f \cdot \max\left(0, 2P_{\text{fused}}(\theta_2) - \lambda_f\right) \\ m_{\text{fused}}(\Theta) = 1 - \sum_{l=1}^{2} m_{\text{fused}}(\{\theta_l\}) \end{cases} \quad (19)$$

Where κf serves as the fusion ruling's dominance factor (held at 0.9, surpassing those of the unimodal counterparts), while λf operates as a reliability floor on decisions (held at 0.6, discarding low-certainty outputs).

The BPA belonging to a Single-Modal Decision mm is built according to:

© 2023 The Author(s). | Published by Agro Intelligence Research Press

$$\begin{cases} m_m\left(\{\theta_1\}\right) = \kappa_m \cdot \sigma\left(\beta_m P_m\left(\theta_1\right) - \tau_m\right) \\ m_m\left(\Theta\right) = \gamma_m \cdot \left(1 - \max P_m\left(\theta_l\right)\right) \end{cases} \quad (20)$$

Within which (·) stands for the Sigmoid transformation that constrains probability magnitudes inside [0, 1], βm steers the steepness characteristic of the confidence curve (held at 5), τm supplies a bias offset to counterbalance sensor-specific drift (held at 0.5 for visible, 0.4 for infrared), and γm captures the modality's native uncertainty margin tied to sensor fidelity (held at 0.3 for visible, 0.2 for infrared).

D-S evidence combination rule

The tripartite evidence structure so formed is fused via Dempster's rule of composition:

$$m_{\text{final}}\left(A\right) = \frac{1}{1-K} \sum_{B_{\text{fused}} \cap B_V \cap B_I = A} \left(m_{\text{fused}}\left(B_{\text{fused}}\right) \cdot m_V\left(B_V\right) \cdot m_I\left(B_I\right)\right) \quad (21)$$

With $A \subseteq \Theta$ indicating the proposition under evaluation, and $K \in [0, 1]$ acting as the conflict measure (invoked when intersecting the full collection of evidence sets yields an empty outcome), gauging the intensity of discord spread across the evidential pieces. A larger numerical value of K reveals a more acute contradiction among the evidence sources, obtained through

$$K = \sum_{B_{\text{fused}} \cap B_V \cap B_I = \emptyset} \left(m_{\text{fused}}\left(B_{\text{fused}}\right) \cdot m_V\left(B_V\right) \cdot m_I\left(B_I\right)\right) \quad (22)$$

Broadly, $K < 0.3$ signals routine conditions, values in the band $0.3 \leq K < 0.7$ signal a medium level of contradiction, and $K \geq 0.7$ signals intense discord.

The aggregation procedure traverses every viable configuration in the evidence space (24 scenarios). A consistency screening step follows: mass is allocated solely when the intersection across all evidence combination sets coincides with the universal set (A), meaning each evidential body endorses the samel recognition conclusion (A). Simultaneously, the conflict measure (K) is determined for those configurations where disagreement surfaces among the evidential bodies. Through the normalizing factor 1/(1-K), conflict-assigned mass is redistributed across the consensus configurations.

Conflict-adaptive handling

Taking a lead from Murphy's methodology for gauging discord among evidential sources [29], a conflict-adaptive countermeasure routine activates once the system registers conflict at a medium tier or higher ($K \geq 0.3$). The procedure starts by appraising the dependability of each evidential origin to generate adaptive weighting terms αs:

$$\alpha_s = \frac{Q_s}{\max Q_k} \times \left(1 - \gamma_s\right) \times \xi_s \quad (23)$$

Within which Qs stands for a quality score assigned to that evidential piece (e.g., track record of correctness), γs equals that origin's built-in uncertainty allowance, and ξs works as a calibration constant (fixed empirically at 0.9). In scenarios where a unimodal ruling sharply clashes with the remaining sources, the fusion ruling's influence is scaled upward (×1.2).

© 2023 The Author(s). | Published by Agro Intelligence Research Press

Conversely, when the fusion ruling itself exhibits elevated uncertainty, its influence is scaled downward (×0.8). Under these rectified weights, the evidence bodies undergo aggregation via a weighted synthesis rule:

$$m_{\text{final}}(A) = \frac{1}{1 - K_{\alpha}} \sum_{\cap Bs = A} \left(a_{\text{used}} m_{\text{fused}}(B_s) \cdot \alpha_V m_V(B_s) \cdot \alpha_I m_I(B_s) \right) \quad (24)$$

Decision generation and uncertainty management

Let the belief mapping be specified as $\text{Bel}(\text{A}) = \sum_{\text{B} \subseteq \text{A}} \text{mfinal}(\text{B})$, giving the lower bound of credence committed to proposition A (the sum of probability masses apportioned among all subsets that imply A). Let the plausibility mapping be specified as $\text{Pl}(\text{A}) = \sum_{\text{B} \cap \text{A} \neq \emptyset} \text{mfinal}(\text{B})$, giving the upper bound of credence that proposition A might attract (the sum of probability masses apportioned among all propositions not ruling out A).

Building on these constructs, the terminal decision rule takes the form.

$$\text{Decision output} = \begin{cases} \theta_1 & \text{if } \text{Bel}(\{\theta_1\}) > 0.85 \\ \theta_2 & \text{if } \text{Pl}(\{\theta_1\}) < 0.35 \\ \text{Uncertain / Redetection} & \text{otherwise} \end{cases} \quad (25)$$

Experiments and Results Analysis

Experimental setup

Dataset: The experimental foundation draws upon the Anti-UAV multi-modal resource made available by Jiang *et al.* [30], which contains 318 temporally matched visible-infrared footage pairs. To shape a dataset both varied in coverage and representative while reducing frame repetition, we sampled 10 keyframes from each recording through a structured extraction scheme: (1) five frames picked at evenly spaced intervals to capture variation across time, plus (2) five frames gathered around instances of pronounced target displacement or appearance fluctuation, pinpointed via optical flow scrutiny. The resulting corpus, termed the Anti-UAV-RGBT dataset, consists of 6360 infrared-visible image dyads aligned with strict spatial and temporal fidelity (misalignment of less than 3 pixels at 1920 × 1080 resolution). The imagery traverses a spectrum of scene types—including crowded urban tower clusters, semi-urban outskirts, and waterfront stretches—and covers multiple observational circumstances (daylight, darkness, haze). Partitioning allocates 3200 samples to training, 1820 to testing, and 1340 to validation.

Hardware/Software Environment: Hardware: NVIDIA RTX 3090 GPU (24 GB VRAM), Intel Xeon Gold 6330 CPU @ 2.5 GHz, 45 GB RAM. Software: PyTorch 2.0.0, Python 3.8 (Ubuntu 20.04), CUDA 11.8.

Single-modal Model Training: Visible: Input dimensions: 1920 × 1080 visible imagery. Network: Enhanced ResNet-50 (equipped with FPN, DPAM, task-separated head). Optimizer: SGD (lr = 0.01, momentum = 0.9). Batch Size: 16. Epochs: 150. Augmentation: Stochastic cropping, flipping, chromatic perturbation. Infrared: Input dimensions: 640 × 512 infrared imagery. Network: Two-stream ConvNeXt-Tiny + lightweight UNet. Optimizer: Adam (lr = 5 × 10−4). Batch Size: 16. Epochs: 120. Augmentation: Luminance/contrast modulation.

Fusion Model Training: The parameter sets governing the feature-level fusion module (EADW) and the decision-level fusion module (D-S evidence theory) were jointly refined end-to-end concurrently with the unimodal detection backbones.

© 2023 The Author(s). | Published by Agro Intelligence Research Press

Optimization relied on a multi-component loss (categorical cross-entropy + localization GIoU). Optimizer: Adam (lr = 1 × 10−4). Batch Size: 8. Early halting trigger: validation loss variance remaining below 0.1% for a window of 20 epochs.

Evaluation Metrics: Commonly established object detection performance indicators were employed, namely: Accuracy, Precision, Recall, False Alarm Rate (FAR), and F1-Score.

Expressed precisely, Accuracy = (TP + TN)/(TP + TN + FP + FN), Precision = TP/(TP + FP), Recall = TP/(TP + FN), FAR = FP/(FP + TN), and F1-Score = 2 × Precision × Recall/(Precision + Recall), with TP, TN, FP, and FN denoting True Positives, True Negatives, False Positives, and False Negatives in respective order.

All reported figures are presented as the mean ± standard deviation calculated from three separate runs to uphold statistical dependability.

Overall performance comparison

Table 1 collates the full set of performance metrics measured on the test partition, contrasting the introduced fusion scheme against unimodal reference systems (Visible Light, Infrared) along with six prominent fusion approaches drawn from both classical blueprints and the latest state-of-the-art (SOTA) developments: Cross-modal Attention Fusion [31], Feature Weighted Average Fusion [32], Voting Decision Fusion [33], Dynamic Feature Fusion [20], as well as the more recently contributed TransFuser [23] and CMX [34].

Table 1. Overall performance comparison of different methods on the test set (%).

Approach	F1-score (%)	False alarm rate (FAR) (%)	Recall (%)	Precision (%)	Accuracy (%)
Visible-spectrum only	83.1	8.5	81.0	85.4	83.1
Infrared-spectrum only	79.1	10.3	81.5	76.8	78.9
Cross-modal attention mechanism [31]	90.0	5.4	90.8	89.3	90.1
Weighted feature averaging method [32]	89.1	5.9	89.5	88.7	89.2
Voting-based decision fusion [33]	87.5	7.1	89.0	86.0	87.5
Adaptive feature fusion approach [33]	88.7	6.3	89.9	87.5	88.7
TransFuser model [23]	91.4	4.9	92.0	90.9	91.5
CMX fusion framework [34]	91.7	4.7	92.3	91.2	91.8
Proposed approach	93.4	4.2	94.1	92.8	93.5

The introduced scheme delivers top-tier results across all central indicators. Its Accuracy lands at 93.5%, Precision at 92.8%, Recall at 94.1%, FAR sinks to a mere 4.2%, and the F1-Score reaches 93.4%. All values are reported as means; the standard deviation calculated over three separate executions remains below 0.5% across all metrics, signaling reliably stable behaviour. These figures provide compelling validation of the visible-infrared multimodal fusion paradigm and the layered fusion strategy proposed in this work.

When set against the strongest unimodal baseline (Visible Light detection), the proposed method raises Accuracy by 10.4% and trims FAR by 4.3%. This advantage principally flows from the synergy of multi-source data and the rigorous

© 2023 The Author(s). | Published by Agro Intelligence Research Press

handling of doubt and contradictory judgments provided by D-S evidence theory, which drastically reduces the probability of erroneous conclusions arising from the malfunction or perturbation of any single sensor.

Relative to established fusion baselines, Accuracy outstrips Feature Weighted Average Fusion by 4.3%, Voting Decision Fusion by 6.0%, Cross-modal Attention by 3.4%, and Dynamic Feature Fusion by 4.8%. It is worth emphasizing that the method likewise surpasses the recent SOTA entry, TransFuser [23], by 2.0% in Accuracy and 2% in F1-Score. The lead widens further when considering FAR suppression. These observations highlight the centrality of the Environment-Aware Dynamic Weighting (EADW) module in adaptively reshaping modal influence during feature amalgamation, alongside the proven edge of D-S evidence theory in accommodating ambiguity and conflict among heterogeneous data streams to achieve truly complementary decision fusion—outclassing simplistic averaging or plurality-vote heuristics.

Computational Cost Analysis: A supplementary breakdown of computational load and inference throughput is laid out in **Table 2**. The proposed method realizes an appealing performance–efficiency compromise. With a footprint of 46.2 M parameters and 81.9 G FLOPs, it proves leaner than the transformer-centric CMX [35] (55.3 M, 98.7 G) and TransFuser [34] (62.1 M, 110.5 G) counterparts, while achieving superior accuracy **(Table 1)**. Its inference pace of 28 FPS, although fractionally behind the Cross-Modal Attention baseline (31 FPS) owing to sequential stages inside the EADW and D-S blocks, continues to satisfy real-time constraints and exemplifies a well-judged trade-off between detection fidelity and resource consumption.

Table 2. Computational complexity and inference speed comparison.

Method	FPS	FLOPS (G)	Params (m)
Visible Branch	45	45.3	25.1
Infrared Branch	52	32.1	18.7
Cross-Modal Attention [31]	31	82.1	46.5
CMX [35]	26	98.7	55.3
TransFuser [34]	22	110.5	62.1
Proposed Method	**28**	**81.9**	**46.2**

Environmental adaptability analysis

To probe the method's ruggedness and versatility under environmental variation, **Table 3** itemizes detection accuracy figures across three canonical weather and lighting regimes present in the Anti-UAV-RGBT dataset.

Table 3. Accuracy comparison under different environmental conditions (%).

Environmental condition	Proposed method	CMX	Trans fuser	Cross-modal attn	Weighted avg	IR only	Visible only
Daytime	**94.8**	91.8	91.5	90.2	88.3	72.8	82.1
Nighttime	**92.3**	90.8	88.5	83.7	80.1	76.5	46.3
Haze	**89.7**	89.7	85.9	81.5	78.4	65.2	51.7
Average Performance	**92.3**	90.8	88.6	85.1	82.3	71.5	60

© 2023 The Author(s). | Published by Agro Intelligence Research Press

The proposed method maintains high detection accuracy across all examined environmental regimes (band: 89.7–94.8%), achieving a mean of 92.3% and outperforming all alternative approaches by a considerable margin. This attests to the method's pronounced capacity to adjust and endure across complex real-world conditions.

Nighttime Conditions: The visible-band imager undergoes severe functional collapse, leaving infrared as the dominant source of usable signal. Through the EADW mechanism, the system autonomously and decisively downscales the visible modality's fusion contribution while upscaling that of the infrared modality. Crucially, D-S evidence theory competently merges the feature-fusion layer's output (potentially handicapped by the visible channel's failure) with the standalone infrared determination. Accordingly, the proposed method yields nighttime results that are significantly stronger than those of both other fusion pipelines and unimodal setups.

Haze Conditions: Aided by the EADW module, which dynamically depresses the visible sensor's weight (as it suffers acutely from atmospheric extinction) while capitalizing on infrared input (rooted in thermal emission that penetrates haze more effectively), detection accuracy receives a pronounced boost.

Figures 5 and 6 provide a side-by-side visual illustration of detection behaviour for the visible-only and infrared-only baselines and the introduced method across dissimilar operating scenarios.

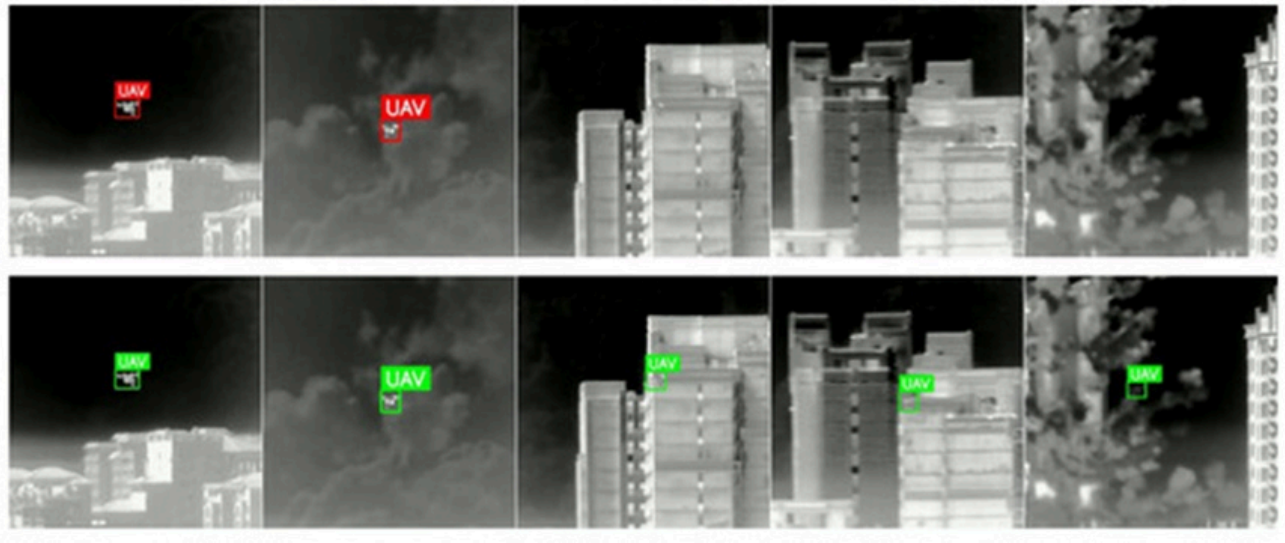


Figure 5. Comparison of infrared single-modal detection and multi-modal hierarchical fusion detection results.

As depicted, the upper row captures the infrared-only output, and the lower row captures the output from the introduced fusion approach. Amid urban thermal radiation clutter, the infrared-only pipeline fails to flag targets hidden behind buildings and vegetation, whereas the proposed approach successfully uncovers them.

© 2023 The Author(s). | Published by Agro Intelligence Research Press

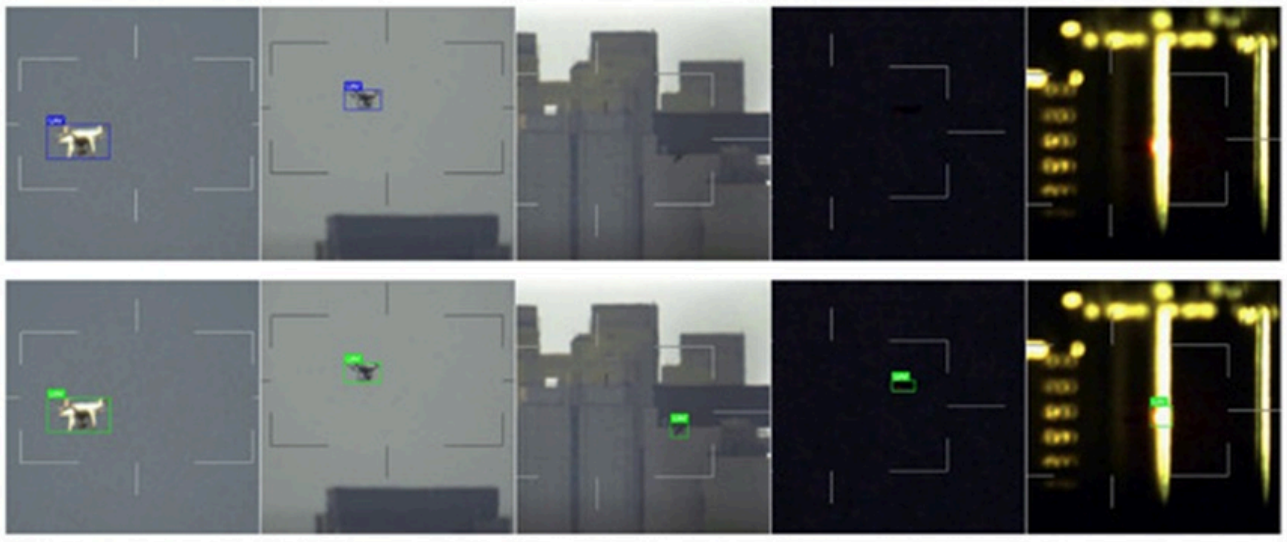


Figure 6. Comparison of visible light single-modal detection and multi-modal hierarchical fusion detection results.

As depicted, the upper row shows the visible-only output, and the lower row shows the output from the introduced fusion approach. In conditions with confusing background similarity, extremely faint lighting, and nocturnal luminous interference, the visible-only pipeline misses several targets, whereas our method correctly identifies all of them.

Ablation study

To isolate the contribution of each core module in the proposed fusion architecture, an ablation analysis was conducted. The setup combining feature-level naive averaging (FeatAvg) with decision-level plain voting (Vote) is treated as the reference point (Baseline). The original components—Environment-Aware Dynamic Weighting (EADW), D-S Evidence Theory (D-S), and Conflict-Adaptive (CA) handling—are integrated cumulatively. Findings are reported in **Table 4**.

Table 4. Ablation study results of key components (%).

Method	F1-Score	FAR	Recall	Precision	Accuracy
Baseline (FeatAvg + Vote)	87.5	7.1	89.0	86.0	87.5
EADW	90.3	5.5	91.0	89.6	90.3
EADW + D-S	92.0	4.8	92.8	91.3	92.1
EADW + D-S + CA (ours)	**93.4**	**4.2**	**94.1**	**92.8**	**93.5**

Effect of EADW

Installing the EADW mechanism on top of the baseline increases Accuracy by 2.8 percentage points and reduces FAR by 1.6 points. This powerfully confirms the worth of environment-responsive dynamic weighting during the feature-fusion stage. EADW adaptively remaps the contributions of distinct modalities based on real-time environmental readings (illumination, weather, time) and unimodal confidence estimates, substantially strengthening the fusion stage's environmental reactivity and its ability to distinguish targets.

Effect of D-S evidence theory

Replacing plain voting with D-S evidence theory after the EADW addition yields a further Accuracy bump of 1.8 points and a FAR cut of 0.7 points. This bears out the advantage conferred by D-S theory at the decision-fusion stage. By treating

© 2023 The Author(s). | Published by Agro Intelligence Research Press

uncertainty and contradictory signals from distinct evidential origins (the feature-fusion verdict and the unimodal verdicts) quantitatively and merging heterogeneous decisions through principled combination rules, D-S theory effectively circumvents the shortcomings of simplistic voting when information exchange is lacking, and decisions collide.

Gain from the Conflict-Adaptive (CA) mechanism

Appending the CA handling block to the D-S fusion chain produces the full method's peak figures (Ours), lifting Accuracy by an additional 1.4 points and cutting FAR by a further 0.6 points. This indicates that the CA block becomes indispensable when moderate or stronger discord ($K > 0.3$) surfaces. CA dynamically reassesses the credibility of evidential sources and recalculates their influence according to the conflict intensity, safeguarding the fusion system's reliability and decisional correctness even when substantial contradictions exist among evidential inputs.

Aggregate Impact: The ablation experiment clearly shows the evolution of performance from the baseline to the finalized method (ranging from 87.5% to 93.5%). Each sequentially added component yields a tangible performance uplift, unequivocally demonstrating the necessity and efficacy of the devised EADW, D-S, and CA mechanisms in forging a high-accuracy, highly resilient multi-modal fusion detection platform.

Conclusion

This paper addresses the problem of identifying Low-Slow-Small (LSS) UAV swarm formations by presenting a visible-infrared multimodal fusion detection methodology based on deep learning. A tiered, multimodal fusion framework that intertwines feature-level and decision-level amalgamation is engineered, meaningfully elevating detection and recognition capabilities and systemic robustness for LSS-UAV swarms operating in complex environments. The central contributions can be summarized as follows:

(1) Purpose-built deep feature extraction backbones (Enhanced ResNet-50 with DPAM for visible-band processing, and Dual-branch ConvNeXt-Tiny with UNet for infrared-band processing) are crafted to suit each modality. These architectures confront challenges including degraded textural definition, distracting scene clutter, and fluctuating object scales across visible and infrared imagery, concretely boosting single-modality feature discrimination under entangled disturbance conditions.

(2) An original EADW-DS hierarchical feature- and decision-level fusion construct is set forth. The EADW mechanism dynamically recalibrates fusion contributions by sensing instantaneous environmental state variables (illumination, weather, time) together with unimodal confidence readings, markedly elevating the environmental resilience of feature fusion. D-S evidence theory thereafter synthesizes the feature-fusion layer's determination with the autonomous unimodal determinations. Through formal uncertainty quantification and the embedding of a Conflict-Adaptive (CA) handling procedure, the design effectively resolves the twin issues of absent information interchange and contradictory verdicts, substantially reinforcing overall system robustness.

Evaluated on the Anti-UAV-RGBT dataset, the proposed method achieves a detection accuracy of 93.5%, an average gain of 18.7% over the best-performing unimodal method, while suppressing the false alarm rate to 4.2%. Performance uplift proves especially pronounced (>8%) under taxing conditions such as haze and nighttime. Ablation experiments confirmed the effectiveness of the key constituents—EADW, D-S, and CA.

Future investigations will be directed along the following lines:

(1) Integration of additional modalities: Incorporate further sensing channels—radar, RF (Radio Frequency), and acoustic arrays—to explore LSS target detection solutions spanning richer modality combinations and more entangled scene interference scenarios. The proposed tiered fusion architecture is intrinsically structured to accommodate more than two input modalities.

© 2023 The Author(s). | Published by Agro Intelligence Research Press

(2) Advancing the dynamic weighting mechanism: Advance the dynamic weighting logic inside EADW through the incorporation of reinforcement learning or online adaptation schemes, enabling more intelligent and finer-grained modulation of each modality's contribution in response to instantaneous scene dynamics (e.g., abrupt interference, target manoeuvres). Such advances could further augment the system's autonomy and extended adaptation in non-stationary environments.

(3) Robustness against adversarial perturbations: Probe the susceptibility of the proposed multi-modal system to adversarial attacks and devise corresponding defensive strategies [35], reinforcing the security and trustworthiness of the detection platform within potentially contested operational contexts.

(4) Theoretical underpinning: Undertake a deeper theoretical examination of the interdependencies among environmental descriptors, fusion weight distributions, and end-to-end performance, thereby supplying a more rigorous analytical foundation for the EADW mechanism.

Acknowledgements

None

Conflict of interest

None

Financial support

This research was funded by the National Natural Science Foundation of China (grant number 62471508).

Ethics statement

None

Received: 30 Mar 2022 Revised: 04 Jun 2022 Accepted: 21 Jun 2022
Published online: 25 May 2023

Rights and permissions

Open Access The author(s) retain copyright. This article is licensed under the Creative Commons Attribution-NonCommercial-ShareAlike 4.0 International License. It may be shared and adapted for non-commercial purposes with appropriate attribution, an indication of changes, and distribution of adaptations under the same license. Third-party material may be subject to separate terms identified in its credit line. View the license at https://creativecommons.org/licenses/by-nc-sa/4.0/.

References

1. Na Z, Cheng L, Sun H, Lin B. Survey of UAV detection and recognition research based on deep learning. Signal Process. 2024;40:609-24.

2. Adnan WH, Khamis MF. Drone use in military and civilian application: Risk to national security. J Media Inf Warf. 2022;15:60-70.

3. Qiu X, Luo B, Fu Z, Tan X, Xiong C. Review of anti-UAV technology development at home and abroad. Tact Missile Technol. 2024:63-73,98.

4. Liu L, Liu D, Wang X, Wang F, Li Y, He Y, et al. Development status and prospect of UAV swarm and anti-UAV swarm. Acta Aeronaut Astronaut Sin. 2022;43:4-20.

5. Cheng Y, Zou R, Chen J, Wu H, Hua X. LSS-Ku-1.0: A radar low-slow-small UAV detection dataset under ground clutter background. Signal Process. 2025;41:807-20.

6. Zhao J, Zhang J, Li D, Wang D. Vision-based anti-UAV detection and tracking. IEEE Trans Intell Transp Syst. 2022;23:25323-34.

7. Yang F, Wang M. Infrared small target detection method for low-altitude surveillance systems. Opt Tech. 2024;50:120-8.

8. Xu S, Chen X, Li H, Liu T, Chen Z, Gao H, et al. Airborne small target detection method based on multimodal and adaptive feature fusion. IEEE Trans Geosci Remote Sens. 2024;62:5637215.

9. Song H. Small target detection based on multimodal data fusion. In: Proceedings of the IEEE International Symposium on Computer Applications and Information Technology (ISCAIT); China; 2025. p. 537-40.

10. Ouyang J, Jin P, Wang Q. Multimodal feature-guided pretraining for RGB-T perception. IEEE J Sel Top Appl Earth Obs Remote Sens. 2024;17:16041-50.

11. Li M, Zhou M, Zhi R. Survey of UAV recognition research based on multi-modal fusion. Comput Eng Appl. 2025;61:1-14. http://kns.cnki.net/kcms/detail/11.2127.TP.20250523.1539.008.html (accessed on 10 April 2025).

12. Han Z, Yue M, Zhang C, Gao Q. Multi-modal fusion detection for UAV targets based on Siamese network. Infrared Technol. 2023;45:739-45.

13. Hu N, Tian X. OFDM-MIMO radar signal design and processing method for small UAVs. Signal Process. 2024;40:878-86.

14. Kaleem Z. Lightweight and computationally efficient YOLO for rogue UAV detection in complex backgrounds. IEEE Trans Aerosp Electron Syst. 2025;61:5362-6.

15. Ren S, He K, Girshick R, Sun J. Faster R-CNN: Towards real-time object detection with region proposal networks. IEEE Trans Pattern Anal Mach Intell. 2017;39:1137-49.

16. Li R, Peng Y, Yang Q. Fusion enhancement: UAV target detection based on multi-modal GAN. In: Proceedings of the IEEE Information Technology and Mechatronics Engineering Conference (ITOEC); Chongqing, China; 2023. p. 1953-7.

17. Gao M, Lin S. Deep learning LSS target detection algorithm based on radar signal and remote sensing map fusion. Signal Process. 2024;40:82-93.

18. Guo R, Sun B, Sun X, Bu D, Su S. Multi-modal fusion target detection method for UAVs under low-light conditions. Chin J Sci Instrum. 2025;46:338-50.

19. Hengy S, Laurenzis M, Schertzer S, Hommes A, Kloeppel F, Shoykhetbrod A, et al. Multimodal UAV detection: Study of various intrusion scenarios. In: SPIE Electro-Optical Remote Sensing XI; 2017. Bellingham (WA): SPIE; 2017. p. 203-12.

20. Feng D, Haase-Schütz C, Rosenbaum L, Hertlein H, Glaeser C, Timm F, et al. Deep multi-modal object detection and semantic segmentation for autonomous driving: Datasets, methods, and challenges. IEEE Trans Intell Transp Syst. 2021;22:1341-60.

© 2023 The Author(s). | Published by Agro Intelligence Research Press

21. LeCun Y, Bengio Y, Hinton G. Deep learning. Nature. 2015;521:436-44.

22. Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez AN, et al. Attention is all you need. arXiv. 2017;arXiv:1706.03762. https://arxiv.org/abs/1706.03762 (accessed on 8 December 2025).

23. Woo S, Park J, Lee JY, Kweon IS. CBAM: Convolutional Block Attention Module. In: Proceedings of the European Conference on Computer Vision (ECCV); Munich, Germany; 2018. p. 3-19.

24. Fu J, Liu J, Tian H, Li Y, Bao Y, Fang Z, et al. Dual Attention Network for Scene Segmentation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR); Long Beach, CA, USA; 2019. p. 3146-54.

25. Chi W, Liu J, Wang X, Feng R, Cui J. DBGNet: Dual-branch gate-aware network for infrared small target detection. IEEE Trans Geosci Remote Sens. 2023;61:5003714.

26. Liu Z, Mao H, Wu C-Y, Feichtenhofer C, Darrell T, Xie S. A ConvNet for the 2020s. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR); New Orleans, LA, USA; 2022. p. 11966-76.

27. Sun Y, Cao B, Zhu P, Hu Q. Drone-based RGB-infrared cross-modality vehicle detection via uncertainty-aware learning. IEEE Trans Circuits Syst Video Technol. 2022;32:6700-13.

28. Zhao G, Chen A, Lu G, Liu W. Data fusion algorithm based on fuzzy sets and D-S theory of evidence. Tsinghua Sci Technol. 2020;25:12-9.

29. Murphy CK. Combining belief functions when evidence conflicts. Decis Support Syst. 2000;29:1-9.

30. Jiang N, Wang K, Peng X, Yu X, Wang Q, Xing J, et al. Anti-UAV: A large-scale benchmark for vision-based UAV tracking. IEEE Trans Multimed. 2023;25:486-500.

31. Xu K, Wang B, Zhu Z, Jia Z, Fan C. A contrastive learning enhanced adaptive multimodal fusion network for hyperspectral and LiDAR data classification. IEEE Trans Geosci Remote Sens. 2025;63:4700319.

32. Liu Z, Shen Y, Lakshminararasimhan VB, Liang PP, Zadeh AB, Morency L-P. Efficient low-rank multimodal fusion with modality-specific factors. In: Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (ACL); Melbourne, Australia; 2018. p. 2247-56.

33. Perez-Rua J-M, Vielzeuf V, Pateux S, Baccouche M, Jurie F. MFAS: Multimodal fusion architecture search. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR); Long Beach, CA, USA; 2019. p. 6959-68.

34. Li H, Zhang S, Kong X. CMX: Cross-Modal Fusion for RGB-X Semantic Segmentation with Transformers. IEEE Trans Geosci Remote Sens. 2023;61:5003713.

35. Madry A, Makelov A, Schmidt L, Tsipras D, Vladu A. Towards Deep Learning Models Resistant to Adversarial Attacks. In: Proceedings of the International Conference on Learning Representations (ICLR); Vancouver, BC, Canada; 2018.

© 2023 The Author(s). | Published by Agro Intelligence Research Press