drones

MDPI

Article

STDG-UAV: A Lightweight Spatial–Temporal Domain-Generalized Network for Drone-to-Drone Detection and Onboard Collision-Risk Perception

Bingyan Cui [1] and Zhen Liu [2,*]

[1] Department of Civil and Environmental Engineering, Rutgers, The State University of New Jersey, Piscataway, NJ 08854, USA; bingyan.cui@rutgers.edu
[2] Institute of Space and Earth Information Science, Fok Ying Tung Remote Sensing Science Building, The Chinese University of Hong Kong, Hong Kong SAR, China
* Correspondence: zhenliu@cuhk.edu.hk

Highlights

What are the main findings?

- STDG-UAV achieved 95.4% mAP@0.5 and 64.7% AP_s, while improving temporal detection continuity to 96.2% with only two maximum consecutive missed frames.
- The model maintained an average mAP@0.5 of 87.6% in unseen real-flight domains and achieved real-time onboard inference at 28.7 FPS, with a TTC estimation MAE of 0.42 s.

What are the implications of the main findings?

- The proposed lightweight spatial–temporal and domain-generalized framework enables robust drone-to-drone perception across target scales, motion conditions, and previously unseen flight environments.
- Stable onboard detection and TTC-based risk estimation provide practical perception support for real-time warning and future autonomous sense-and-avoid systems.

Abstract

Reliable drone-to-drone perception is essential for safe autonomous operation in increasingly complex low-altitude airspace. This study proposes STDG-UAV, a lightweight spatial–temporal domain-generalized network for onboard drone-to-drone detection and collision-risk perception. The framework combines spatial tiny-target feature preservation with adaptive multi-scale fusion to enhance weak structural cues of distant UAVs. A motion-guided temporal memory module further exploits adjacent-frame information to improve detection continuity under fast motion, blur, and weak visibility. To enhance robustness to unseen environments, image-level domain randomization, feature-style perturbation, and consistency learning are introduced under a source-only training setting. Experiments show that STDG-UAV achieves 95.4% mAP@0.5, 69.8% mAP@0.5:0.95, and 64.7% AP_s, with only 5.8 M parameters and 14.9 GFLOPs. In continuous UAV videos, the model reaches a detection continuity rate of 96.2% and reduces the maximum consecutive missed frames to two. On unseen real-flight domains, it achieves an average mAP@0.5 of 87.6% with a generalization gap of 7.8 percentage points. Onboard experiments further demonstrate real-time inference at 28.7 FPS and accurate TTC estimation with an MAE of 0.42 s. These results demonstrate the potential of STDG-UAV for robust, real-time drone-to-drone perception and collision-risk assessment.

Academic Editors: Filipe Pereira, Joaquim Gabriel Mendes, Pedro Sousa, António Ramos Silva, Higinio González-Jorge, Pedro Arias-Perez

Received: 17 September 2026
Revised: 03 November 2026
Accepted: 04 November 2026
Published: 5 November 2026

Copyright: © 2026 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license.

Keywords: drone-to-drone detection; small-UAV detection; spatial–temporal learning; domain generalization; onboard perception; collision-risk perception

1. Introduction

Road infrastructure is continuously exposed to traffic loading and environmental stresses, which can contribute to pavement deterioration and potential particle release [1], highlighting the importance of regular infrastructure inspection and monitoring. In recent years, unmanned aerial vehicles (UAVs) have been increasingly used in infrastructure inspection, logistics, environmental monitoring, emergency response, and low-altitude intelligent transportation [2–5]. Their growing presence in shared low-altitude airspace highlights the need for reliable detection of nearby aircraft and timely collision-risk assessment to maintain safe separation. For non-cooperative UAVs without reliable inter-vehicle communication, predefined trajectories or information exchange alone cannot guarantee safe operation [6,7]. Therefore, autonomous perception based on onboard sensors has become a key component of sense-and-avoid systems [8,9]. Compared with radar, acoustic, and radio-frequency sensing, RGB vision offers lower hardware weight and cost while providing rich environmental information, making it suitable for integration into small and medium-sized UAV platforms [3,10,11]. Previous studies have investigated vision-based collision avoidance, time-to-collision (TTC) estimation, and experimental validation of sense-and-avoid systems for small UAVs [12–14]. However, in drone-to-drone scenarios, large inter-UAV separation often reduces the target to a small image region with limited identifiable details. Their appearance is also affected by rapid scale changes, complex backgrounds, and motion blur. Recent cross-scene UAV datasets further indicate that small target size and substantial domain shifts remain major challenges in air-to-air visual detection [15,16]. Therefore, achieving accurate, stable, and environment-adaptive UAV perception under limited onboard computational resources remains an important problem for autonomous low-altitude flight.

Recent studies on small-UAV detection have mainly focused on high-resolution feature preservation, multi-scale feature fusion, attention mechanisms, and lightweight network design [17–20]. For example, shallow detection branches, improved feature pyramids, and spatial attention mechanisms have been introduced to preserve small-target details during feature extraction [21–23]. In addition, Drone-DETR extends the RT-DETR architecture to improve feature interaction across spatial resolutions for small-object localization [24]. At the same time, depthwise separable convolution, lightweight backbones, and simplified detection heads have been used to reduce model parameters and computational cost for real-time onboard inference [25–28]. Although these methods have significantly improved small-object detection in UAV imagery, most still rely mainly on single-frame inputs, and performance gains are often achieved by adding detection layers, attention modules, or increasingly complex feature-fusion structures [25,29]. For distant UAVs, fine structural cues such as body edges, rotors, and local contours are mainly represented by weak high-frequency information, which can be easily suppressed during repeated downsampling [26,30]. In addition, UAV scale varies greatly with viewing distance, and fixed multi-scale fusion weights may not be suitable for both distant small targets and nearby large targets [31,32]. Therefore, preserving weak structural information and dynamically adjusting the contribution of features at different scales, without substantially increasing model complexity, remains a relevant research problem.

Real onboard perception is also inherently a continuous video process rather than a sequence of independent static images [33,34]. Fast relative motion, camera vibration, short-term occlusion, and motion blur can significantly degrade target appearance in

https://doi.org/10.3390/drones10110827

individual frames. As a result, even a detector with high average mAP may still produce consecutive missed detections, sudden confidence drops, or unstable bounding boxes [35]. Such transient errors may have limited influence on conventional image detection, but they can directly affect relative-motion estimation, TTC calculation, and collision-risk assessment in UAV sense-and-avoid applications. Recent video object detection methods have exploited cross-frame information through adjacent-frame feature aggregation, external memory, temporal attention, and motion-guided modeling [36–38]. In particular, studies on micro-UAV video detection have shown that spatiotemporal feature fusion can improve target representation in blurred or weakly visible frames [36]. However, some temporal methods require optical-flow estimation, dense multi-frame aggregation, or complex memory structures, resulting in high inference costs that are not well suited to resource-constrained onboard deployment [38,39]. A more lightweight temporal modeling strategy is therefore needed to exploit useful historical target information while avoiding the introduction of redundant background features into the current frame, thereby improving detection continuity and stability.

In addition to target-scale variation and temporal degradation, domain shift between training data and real flight environments is another important factor affecting the reliability of UAV vision systems [40–42]. Many detection methods use random splits of the same dataset for training and testing, so both subsets often share similar backgrounds, weather conditions, and imaging characteristics. In real flight, however, visual conditions may vary substantially in terms of sky appearance, building textures, vegetation, illumination, haze, and camera exposure [43,44]. Domain adaptation can reduce these differences by using target-domain samples, but such samples are often unavailable before deployment in unknown environments. Domain generalization is therefore more suitable for practical deployment because it aims to learn domain-invariant representations using only source-domain data. Recent studies have explored generalizable object detection through feature disentanglement, consistency constraints, object-aware enhancement, and single-domain generalization [45]. Despite this progress, domain generalization for drone-to-drone small-object detection remains relatively underexplored. In addition, many existing studies lack systematic validation on onboard edge-computing devices and under real dual-UAV flight conditions [46,47]. For sense-and-avoid applications, high accuracy on public datasets alone is insufficient to demonstrate practical effectiveness. Cross-scene generalization, temporal stability in continuous video, distance-dependent detection, real-time onboard inference, TTC estimation, and collision-risk output should also be evaluated.

Therefore, this study proposes STDG-UAV, a lightweight spatial–temporal domain-generalized network for drone-to-drone detection and onboard sense-and-avoid. First, a spatial tiny-target feature preservation module and an adaptive multi-scale fusion mechanism are developed. The spatial-frequency representation enhances edge and structural information of distant UAVs, while the contributions of P2-P5 features are dynamically adjusted according to target scale. Second, a motion-guided temporal memory module is introduced. It selectively uses historical information through lightweight feature alignment, motion-difference modeling, and gated fusion, thereby reducing consecutive missed detections and localization fluctuations under fast motion and blur. Third, image-level domain randomization, feature-style perturbation, and consistency learning are incorporated under a source-only training setting to improve generalization to unseen real-flight environments. Finally, a dual-UAV flight platform is established, and the proposed model is deployed on an onboard edge-computing device to evaluate real-time inference, distance-dependent detection, TTC estimation, and collision-risk perception. Through these designs, this study aims to provide an integrated visual perception framework that connects small-target spatial representation, continuous temporal perception, cross-

domain generalization, and real onboard UAV applications. The main methodological contribution lies in the coordinated integration and joint optimization of spatial-detail preservation, temporal memory, and source-only domain generalization for resource-constrained onboard UAV perception.

2. Methodology

We developed STDG-UAV, a lightweight spatial–temporal domain-generalized network for onboard sense-and-avoid. The framework first enhances distant UAV features through spatial tiny-target feature preservation and adaptive multi-scale fusion. A motion-guided temporal memory module is then introduced to aggregate useful information across adjacent frames and improve detection continuity under fast motion and blur. To further improve robustness to unseen flight environments, image-level domain randomization, feature-style perturbation, and consistency learning are incorporated during training. Finally, a lightweight detection head produces real-time UAV detections, which are further used for relative-motion analysis, time-to-collision estimation, and collision-risk perception.

2.1. Overall Framework of STDG-UAV

To improve the accuracy, temporal stability, and cross-scene generalization of small-UAV detection in complex low-altitude environments, this study proposes a lightweight spatial–temporal domain-generalized network, termed STDG-UAV. The overall framework is shown in Figure 1. The model takes consecutive onboard RGB frames as input and first extracts P2–P5 multi-scale features through a lightweight backbone. The backbone is based on Ultralytics YOLOv8.0, nano configuration, producing four feature levels (P2–P5) with downsampling strides of 4, 8, 16, and 32, respectively. The baseline retains the same backbone and basic detection head without the proposed STFP, AMSF, MGTM, and domain-generalization components. In the spatial branch, a tiny-target feature preservation module and an adaptive multi-scale fusion strategy are introduced to enhance weak texture and structural information of distant UAVs. A motion-guided temporal memory module is then used to align adjacent-frame features, model motion differences, and perform gated feature fusion, thereby improving detection continuity under fast motion, image blur, and short-term occlusion. During training, image-level domain randomization, feature-level style perturbation, and consistency learning are further employed to reduce the model's dependence on specific backgrounds and imaging conditions. The fused spatial–temporal features are processed by a lightweight decoupled detection head for classification, localization, and confidence prediction. Based on consecutive detection results, target-scale variation and relative-motion cues are further extracted to estimate the visual time to collision and determine the collision-risk level, providing perception support for real-time onboard warning and sense-and-avoid decision making.

https://doi.org/10.3390/drones10110827

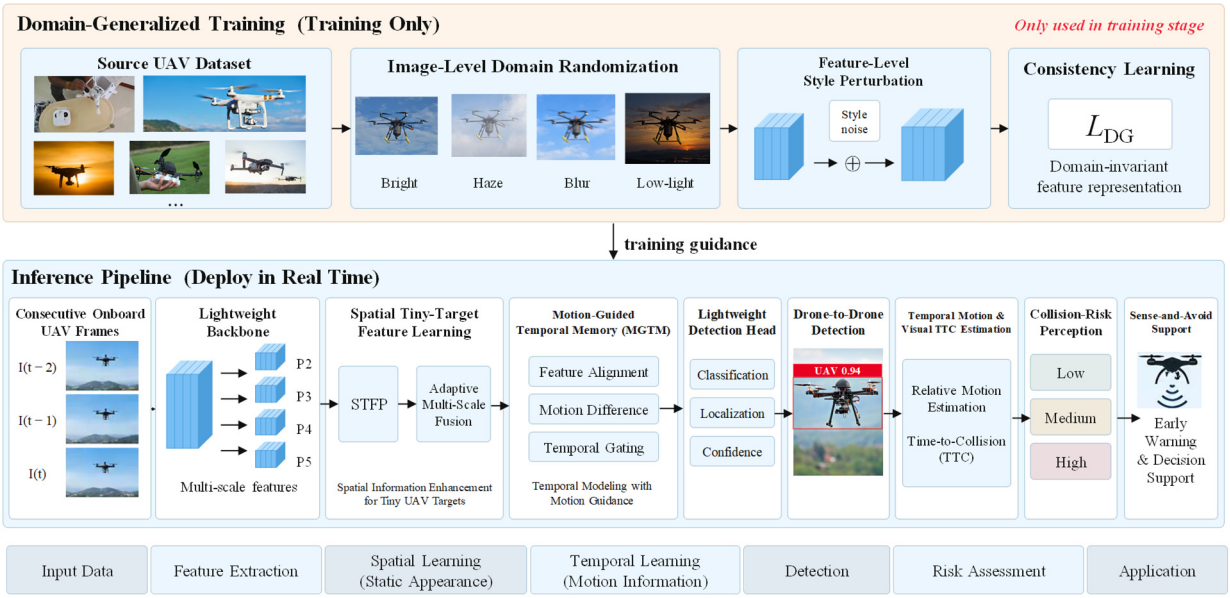


Figure 1. Overall architecture of STDG-UAV framework for drone-to-drone detection and onboard collision-risk perception.

2.2. Spatial Tiny-Target Feature Preservation and Adaptive Multi-Scale Fusion

To address the loss of fine details in distant UAV targets caused by their small size, weak texture, and repeated downsampling, a spatial tiny-target feature preservation (STFP) module and an adaptive multi-scale fusion strategy are developed, as shown in Figure 2. STFP takes shallow high-resolution features as input and consists of a spatial and a frequency branch. The spatial branch uses depthwise convolution to extract local structural information. In the frequency branch, the one-level Haar decomposition produces four subbands: LL (low-low), LH (low-high), HL (high-low), and HH (high-high). The three high-frequency subbands are used to enhance directional edge and structural responses, while the spatial branch employs a 3 × 3 depthwise convolution followed by BatchNorm and SiLU [48]. The high-frequency components mainly preserve fine-grained cues. These features are further enhanced by lightweight convolution and channel attention and then combined with the spatial features through gated fusion to obtain detail-preserved representations.

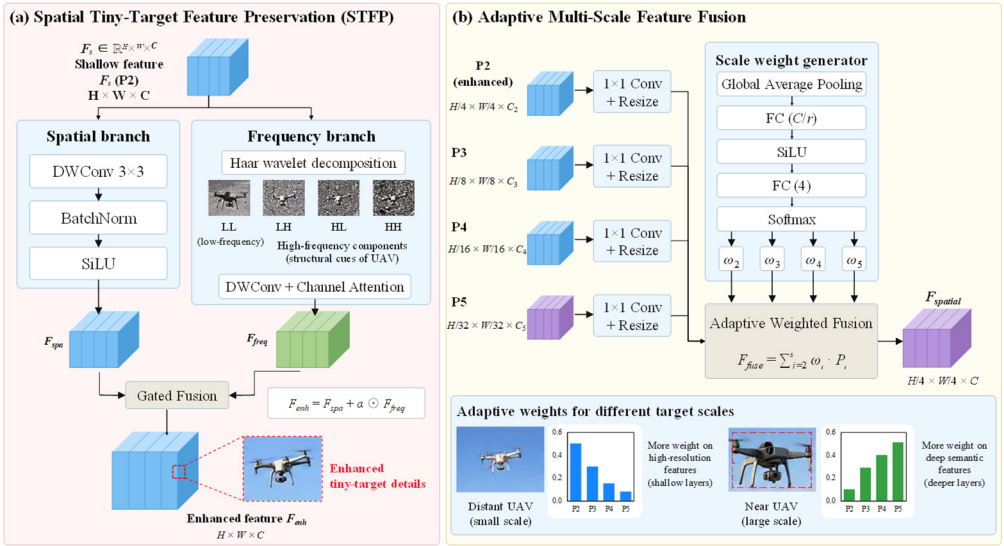


Figure 2. Architecture of the spatial tiny-target feature preservation and adaptive multi-scale fusion module.

On this basis, features from P2–P5 are further integrated. Each feature map is first aligned in channel dimension and spatial resolution using a 1 × 1 convolution and resizing operation. The P2–P5 features are projected to a common channel dimension and resized to the P2 spatial resolution (H/4 × W/4). The Softmax-normalized fusion weights satisfy $\sum_{i=2}^{5} \omega_i = 1$. Global average pooling and lightweight fully connected layers are then used to generate scale-specific weights, which are normalized by Softmax. The fused feature is expressed as

$$F_{fuse} = \sum_{i=2}^{5} \omega_i F_i, \tag{1}$$

where ω_i denotes the adaptive weight of the i-th scale. This mechanism assigns greater importance to high-resolution shallow features for distant small UAVs, while increasing the contribution of deeper semantic features for larger targets. It therefore improves the model's adaptability to large variations in UAV scale.

2.3. Motion-Guided Temporal Feature Aggregation

To fully exploit temporal motion information in UAV videos and reduce missed detections caused by fast motion, motion blur, and short-term weak visibility, a Motion-Guided Temporal Memory (MGTM) module is developed, as shown in Figure 3. The module takes the current spatial feature F_t and the temporal memory from the previous frame M_{t-1} as inputs, enabling cross-frame feature aggregation with low computational cost [49,50]. First, to reduce feature misalignment caused by target displacement between adjacent frames, the historical memory is processed by a 1 × 1 convolution, depthwise convolution, and nonlinear activation. This produces an aligned memory feature $\tilde{M}_{t-1}$.

The motion information is then explicitly extracted from the absolute difference between the current feature and the aligned historical memory:

$$D_t = \left| F_t - \tilde{M}_{t-1} \right|, \tag{2}$$

where D_t represents local feature variations between adjacent frames. Compared with directly combining neighboring-frame features, this difference operation better highlights moving target regions and reduces the repeated introduction of static background information. The motion-difference feature is further processed by depthwise 1 × 1 convolution, and a Sigmoid function to calculate a temporal gating weight:

$$G_t = \sigma\left(Conv_{1\times1}(DWConv(D_t))\right), \tag{3}$$

where $G_t \in [0,1]$ adaptively controls the contribution of historical information to the current frame. The temporal-enhanced feature is then obtained as

$$F_t^T = F_t + G_t \odot \tilde{M}_{t-1}, \tag{4}$$

where Θ denotes element-wise multiplication.

The fused feature F_t^T is sent to the detection head for target classification, localization, and confidence prediction. It is also used to update the current temporal memory M_t, which is passed to the next frame. At the beginning of each independent sequence, the temporal memory is initialized using the first-frame feature, $M_0 = F_0$. For subsequent frames, the memory is updated as $M_t = F_t^T$ and reset at the beginning of each new sequence. This prevents temporal information from being propagated across unrelated video sequences. In this way, temporal information is propagated recursively across consecutive frames. The proposed mechanism strengthens the continuous representation of targets under short-term blur or weak texture, reduces temporary missed detections and bounding-box fluctuations, and provides more reliable detection sequences for subsequent relative-motion estimation and collision-risk perception.

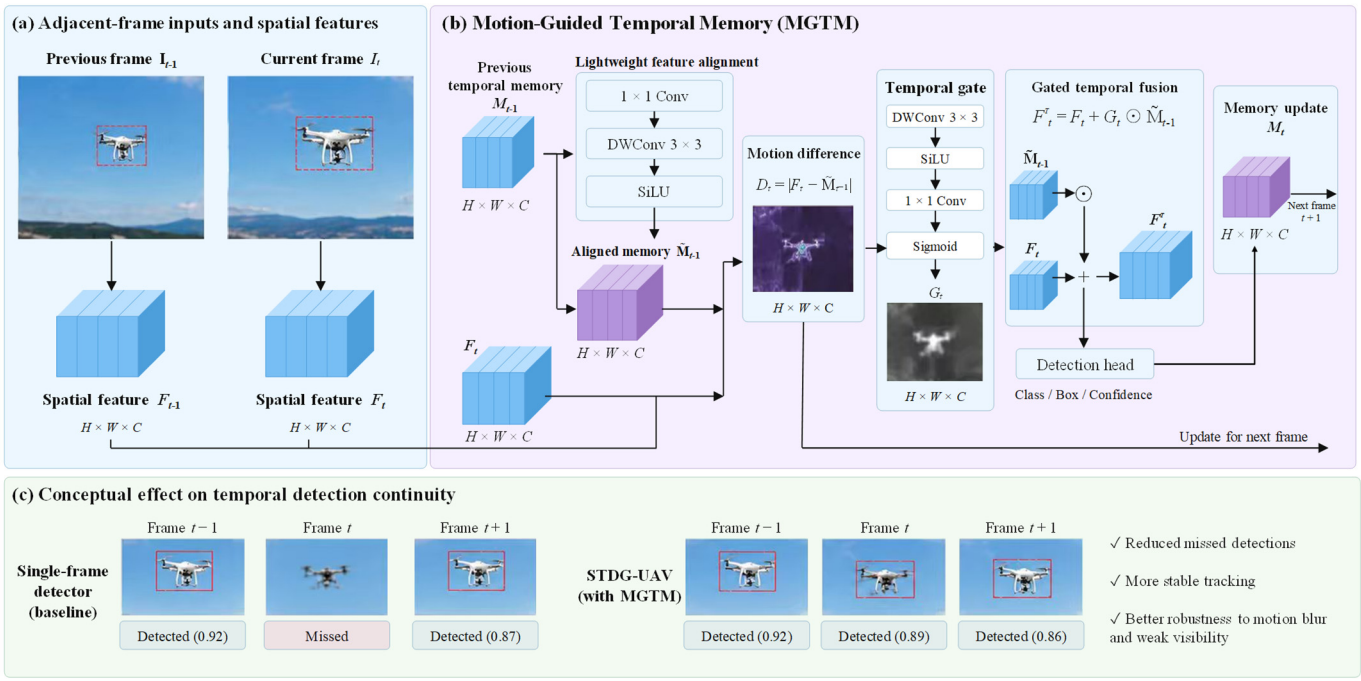


Figure 3. Architecture of the motion-guided temporal memory module for continuous UAV perception. (**a**) Adjacent-frame inputs and spatial feature extraction. (**b**) MGTM architecture, including historical feature alignment, motion-difference modeling, temporal gating, gated fusion, and memory updating. (**c**) Conceptual comparison of temporal detection continuity between the single-frame detector and STDG-UAV with MGTM.

2.4. Domain-Generalized Representation Learning

To reduce the model's dependence on specific backgrounds, illumination conditions, and imaging characteristics, and to improve its generalization to unseen real-flight scenarios, a domain-generalized representation learning strategy is developed, as shown in Figure 4. Only source-domain data are used during training, with no target-domain samples involved. Domain-invariant representations are learned progressively at the image, feature, and prediction levels.

First, image-level domain randomization is used to generate diverse training views. Given a source image x, randomized samples are produced through illumination variation, haze simulation, motion blur, and color shift:

$$x' = T(x;\theta), \tag{5}$$

where $T(\cdot)$ denotes a stochastic domain transformation. These operations change the image appearance and imaging style while preserving the target geometry and annotations, thereby simulating visual variations that may occur in real flight environments.

Feature-level style perturbation is then applied to intermediate features extracted by the shared encoder. Following the feature-statistics mixing principle of MixStyle, the channel-wise means and standard deviations of randomly paired source samples are interpolated to generate perturbed feature styles. For a feature map F, the channel-wise statistics $\mu(F)$ and $\sigma(F)$ are first calculated. Style statistics from different samples are randomly mixed to obtain a new mean $\tilde{\mu}$ and standard deviation $\tilde{\sigma}$. The perturbed feature is expressed as

$$\tilde{F} = \tilde{\sigma}\frac{F - \mu(F)}{\sigma(F)} + \tilde{\mu}, \tag{6}$$

This operation changes the feature style while preserving the spatial structure and semantic content of the UAV, thereby expanding the range of virtual domains observed during training.

Finally, the original and perturbed features are processed by a shared detection head. Feature- and prediction-level consistency constraints are imposed to reduce differences between the two branches. The domain-generalization loss is defined as

$$\begin{aligned} L_{DG} &= \lambda_f L_{feat} + \lambda_p L_{pred} \\ L_{feat} &= \frac{1}{CHW}\left\| N(F) - N(\tilde{F}) \right\|_2^2 \\ L_{pred} &= \frac{1}{N}\sum_{i=1}^{N}\left[\left\| p_i - \tilde{p}_i \right\|_2^2 + \left\| b_i - \tilde{b}_i \right\|_1 + (c_i - \tilde{c}_i)^2 \right] \end{aligned} \tag{7}$$

where F and $\tilde{F}$ represent the original and style-perturbed feature maps, respectively, and $N(\cdot)$ denotes channel-wise feature normalization. C, H, and W are the feature dimensions, while N is the number of matched prediction locations. The variables p_i, b_i, and c_i denote the class probabilities, normalized bounding-box coordinates, and object confidence, respectively. Their corresponding predictions from the perturbed branch are denoted by $\tilde{p}_i$, $\tilde{b}_i$, and $\tilde{c}_i$. Since the style perturbations preserve spatial geometry, predictions are matched at corresponding locations. The resulting domain-generalization loss is jointly optimized with the detection loss to improve prediction consistency under different source-domain appearances without using target-domain data.

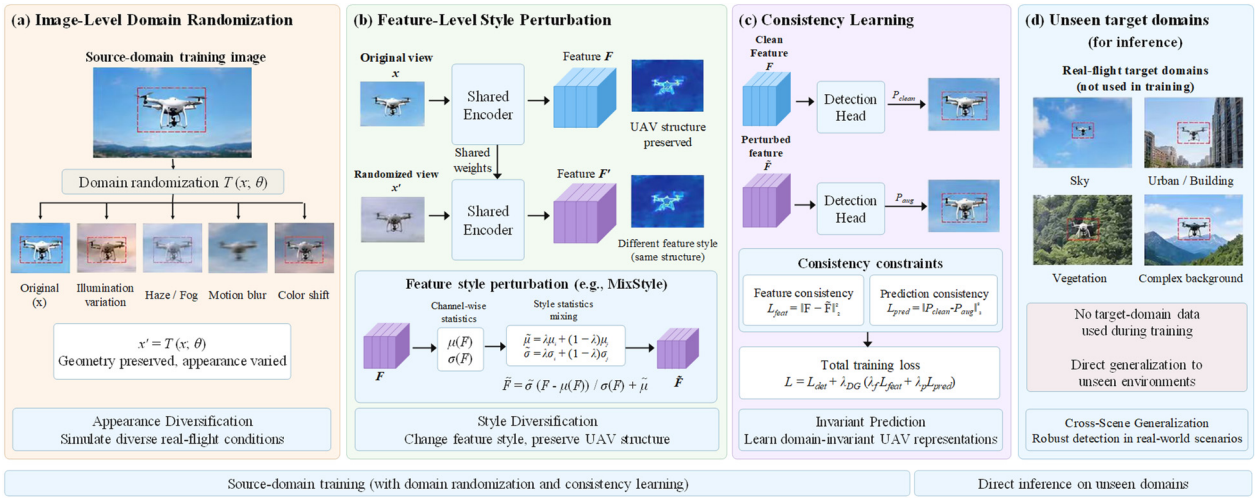


Figure 4. Domain-generalized representation learning strategy for cross-scene UAV detection. (**a**) Image-level domain randomization to simulate diverse visual conditions. (**b**) Feature-level style perturbation using feature-statistics mixing. (**c**) Feature- and prediction-level consistency learning between original and perturbed representations. (**d**) Inference on unseen real-flight target domains without using target-domain data during training.

2.5. Lightweight Detection Head and Joint Optimization

To balance detection accuracy with the requirements of real-time onboard inference, STDG-UAV adopts a lightweight decoupled detection head to process the fused spatial–temporal features, as summarized in Table 1. The detection headfirst applies a shared feature transformation for channel compression, followed by separate branches for classification, bounding-box regression, and confidence prediction. This design reduces feature interference among different prediction tasks [51,52]. Depthwise separable convolutions and other lightweight convolution operations are used in each branch to reduce the number of parameters and computational cost while maintaining effective responses to small UAV targets.

A joint optimization strategy is adopted during training to simultaneously consider object detection, domain generalization, and temporal consistency. The total loss is defined as

$$L_{total} = \lambda_{box} L_{box} + \lambda_{cls} L_{cls} + \lambda_{obj} L_{obj} + \lambda_{DG} L_{DG} + \lambda_{T} L_{T}, \quad (8)$$

where L_{box}, L_{cls}, and L_{obj} are used for bounding-box regression, class prediction, and object-confidence optimization, respectively. L_{DG} constrains feature and prediction consistency under different domain perturbations, while L_{T} improves temporal stability between adjacent frames. Through this joint objective, the model maintains a lightweight structure while achieving a balance among detection accuracy, cross-scene robustness, and stable perception in continuous UAV videos.

Table 1. Configuration and functional roles of the main components in STDG-UAV.

Component	Input	Main Operation	Output	Main Function
Lightweight backbone	RGB frame I_t	Lightweight CNN; P2-P5 feature extraction (strides 4/8/16/32)	P_2, P_3, P_4, P_5	Efficient multi-scale feature extraction
STFP	Shallow feature P_2	3 × 3 DWConv-BN-SiLU; Haar wavelet decomposition; channel attention and gated fusion	F_{enh}	Preserve fine structural cues of tiny UAVs
Adaptive multi-scale fusion	Enhanced P_2 and P_3-P_5	1 × 1 Conv + resize + GAP + FC-SiLU-FC + Softmax weighting	$F_{spatial}$	Adaptive fusion across UAV scales
MGTM	Current feature F_t and memory M_{t-1}	Lightweight feature refinement + motion difference + sigmoid gating + memory update	F_t^T, M_t	Improve temporal continuity and reduce missed detections under motion blur or weak visibility
Domain-generalized learning	Source images/features	Image randomization + style-statistics mixing + feature/prediction consistency	Domain-generalized model (training)	Improve unseen domain robustness
Lightweight detection head	Fused feature F_t^T	Shared stem + decoupled classification, box regression, and confidence branches	Class, box, confidence	Efficient onboard UAV detection
Joint optimization	Detection and DG/temporal losses	Weighted detection + DG + temporal consistency losses	Optimized STDG-UAV	Balance accuracy, continuity, and robustness

2.6. Vision-Based Collision-Risk Perception for Onboard Sense-And-Avoid

To further convert detection results into risk information for onboard collision avoidance, a vision-based collision-risk perception module is developed using consecutive-frame detections [53]. For a target UAV at time t, the center of its bounding box is denoted as $c_t = (x_t, y_t)$, and its scale is defined as

$$s_t = \sqrt{w_t h_t}, \quad (9)$$

where w_t and h_t are the bounding box's width and height, respectively. Changes in the target center and scale across consecutive frames are used to describe relative motion. When the target UAV approaches the observer UAV, its image scale generally increases. Therefore, the visual time to collision can be estimated from the target expansion rate as

$$\tau_t = \frac{s_t}{\dot{s}_t + \varepsilon}, \tag{10}$$

where $\dot{s}_t$ denotes the temporal rate of scale change and ε is a small constant. To reduce frame-level fluctuations, the target scale was smoothed using a K-frame moving average. The expansion rate was then calculated from the difference between consecutive smoothed scales divided by the actual frame interval. TTC was estimated only when the expansion rate exceeded a small positive threshold; otherwise, no finite TTC estimate was assigned. This avoids negative or unstable TTC values under stationary or receding motion. To reduce the influence of frame-level detection fluctuations, a sliding temporal window is used to smooth the target position, scale, and estimated time to collision. A collision-risk indicator is then constructed by combining the approaching rate, image-center offset, and τ_t:

$$\begin{aligned} R_t &= g_t(\alpha r_{\tau,t} + \beta r_{s,t} + \gamma r_{c,t}) \\ r_{\tau,t} &= \mathrm{clip}(1 - \frac{\tau_t}{\tau_0}, 0, 1) \\ r_{s,t} &= \mathrm{clip}(1 - \frac{\dot{\bar{s}}_t}{\bar{s}_t \rho_0}, 0, 1) \\ r_{c,t} &= 1 - d_t \end{aligned} \tag{11}$$

The normalized risk score was divided into three levels using two predefined thresholds. Low, medium, and high risk corresponded to scores below the lower threshold, between the two thresholds, and above the upper threshold, respectively. The fusion weights, normalization parameters, smoothing window, and warning thresholds were fixed before evaluation and remained unchanged across the test flights. The risk indicator was intended to support approaching-target warning rather than provide a complete collision-avoidance decision for arbitrary flight trajectories.

3. Results and Discussion

3.1. Experimental Setup and Evaluation Protocol

3.1.1. Datasets and Real-Flight Experimental Platform

As shown in Figure 5, the experimental data consisted of the publicly available Drone Dataset (UAV) [54] and independently collected real-flight UAV videos. The public dataset contained 1359 annotated images with 1486 UAV instances belonging to a single drone category. The original 1012 training images were divided into 911 training and 101 validation images, while the remaining 347 images were reserved for independent testing. Among the annotated UAV instances in the 347-image source-domain test set, N instances were classified as small objects according to the adopted AP_S criterion. For temporal learning, 911 five-frame pseudo-sequences (4555 frames) were constructed from the source training images using temporally consistent geometric transformations to train MGTM. The real-flight evaluation comprised 16 independent video sequences containing 11,150 frames. Among them, four sequences (2600 frames) were used for temporal robustness evaluation, while 12 sequences (8550 frames) covered four unseen domains, namely sky, building, vegetation, and mixed, with three sequences per domain. These videos included frontal approaching, lateral crossing, same-direction flight, and altitude variation. All unseen target-domain sequences were excluded from training, validation, and hyperparameter selection to prevent information leakage. To further evaluate temporal stability and cross-domain generalization, a dual-UAV real-flight experimental platform was established. The observer UAV was equipped with an RGB camera, a flight-control system, and an onboard edge-computing device. The target UAV followed predefined trajectories,

including frontal approaching, lateral crossing, same-direction flight, and altitude variation. The experiments covered sky, building, vegetation, and mixed backgrounds, and continuous video sequences were collected. All real-flight video sequences were reserved exclusively for evaluation. The temporal robustness and unseen-domain test sets were separated at the flight-sequence level, with no overlapping sequences. No real-flight evaluation data were used for training, validation, or hyperparameter tuning, ensuring a strictly source-only training protocol. The onboard perception system sequentially performed image acquisition, real-time STDG-UAV detection, relative-motion analysis, and collision-risk estimation to evaluate the practical performance in drone-to-drone scenarios.

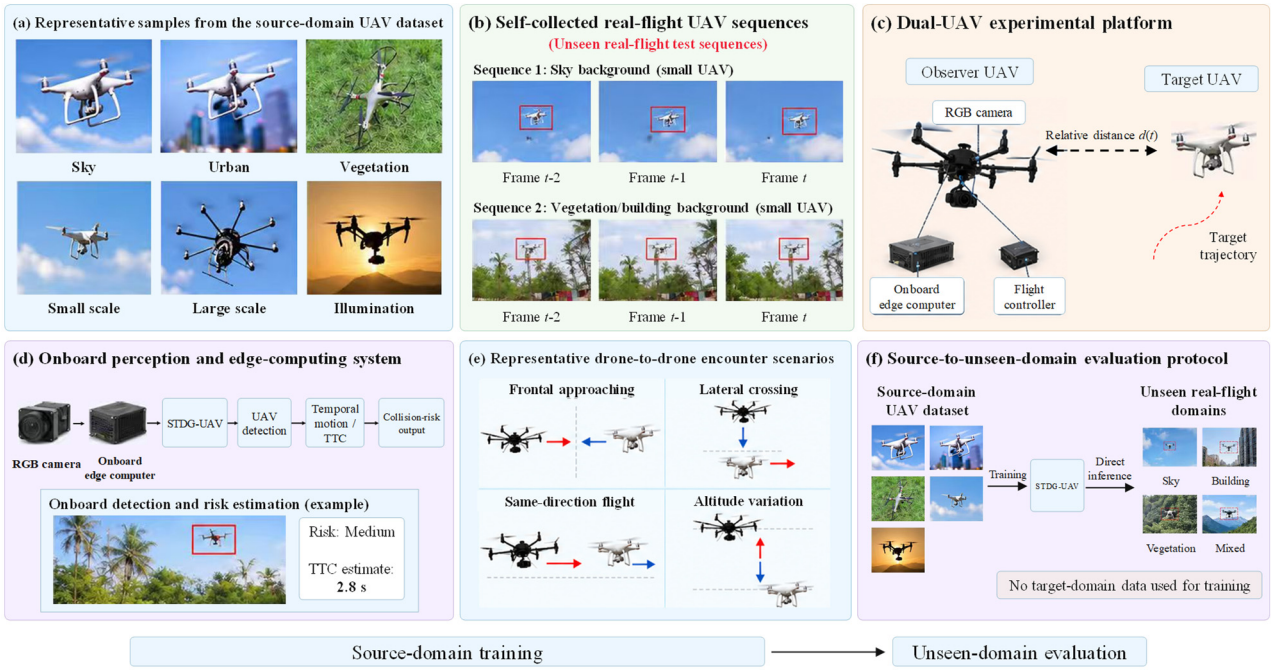


Figure 5. Datasets, dual-UAV experimental platform, and real-flight evaluation protocol.

3.1.2. Implementation Details and Evaluation Metrics

All comparison models were trained using the same source-domain data split and an input resolution of 640 × 640 pixels. We used PyTorch 2.3.1 with CUDA 12.1 on an NVIDIA RTX 4090 GPU (24 GB). The static detection stage was trained for 200 epochs with a batch size of 16 using AdamW, an initial learning rate of 0.001, cosine annealing with a five-epoch warm-up, and a weight decay of 0.0005. The MGTM component underwent 40 epochs of temporal refinement using five-frame sequences, a batch size of four sequences, and a learning rate of 0.0001. Standard augmentations included horizontal flipping (p = 0.5), random scaling (0.8–1.2), rotation (±5°), and brightness/contrast adjustments (±20%). Additional haze, motion-blur, and feature-style perturbations were applied only during STDG-UAV training. The loss weights λ_{box}, λ_{cls}, λ_{obj}, λ_{DG}, and λ_T were set to 5.0, 1.0, 1.0, 0.2, and 0.1, respectively, with λ_f = 1.0 and λ_p = 0.5. Three independent runs were conducted using random seeds of 42, 123, and 2026, with all other training settings kept consistent across runs. Model parameters and GFLOPs were calculated at 640 × 640 resolution, including the computational operations of MGTM and temporal-memory updating. For independent static-image evaluations in Tables 2 and 3, each image was treated as a separate single-frame sequence. Following Section 2.3, the temporal memory was initialized as $M_0 = F_0$ and reset for each image, preventing information exchange between unrelated samples.

Table 2. Quantitative comparison of STDG-UAV with representative object detectors.

Method	Precision (%)	Recall (%)	mAP@0.5 (%, Mean ± SD)	mAP@0.5:0.95 (%, Mean ± SD)	AP_s (%)	Params (M)	GFLOPs
YOLOv8n	89.2	84.6	91.8 ± 0.6	62.4 ± 0.8	54.1	3.2	8.7
YOLO11n	90.1	85.8	92.7 ± 0.5	64.1 ± 0.7	56.3	2.6	6.5
YOLOX-s	88.4	84.9	91.2 ± 0.7	61.8 ± 0.9	53.0	9.0	26.8
YOLOv8s	91.0	87.1	93.6 ± 0.4	66.9 ± 0.6	59.5	11.2	28.6
RT-DETR-R18	91.2	86.7	93.4 ± 0.5	66.3 ± 0.7	58.1	20.0	60.2
YOLOv5s-NGN	91.3	97.0	93.5 ± 0.5	66.4 ± 0.7	58.7	2.39	5.9
STDG-UAV	93.0	89.4	95.4 ± 0.3	69.8 ± 0.5	64.7	5.8	14.9

Table 3. Ablation study of the main components in STDG-UAV.

Setting	STFP	AMSF	MGTM	DG	mAP@0.5 (%)	mAP@0.5:0.95 (%)	AP_s (%)	Params (M)	GFLOPs
Baseline	--	--	--	--	92.6	64.2	56.7	3.1	7.9
+ STFP	✓	--	--	--	93.6	65.9	60.2	3.8	9.7
+ AMSF	✓	✓	--	--	94.3	67.3	62.3	4.6	11.6
+ MGTM	✓	✓	✓	--	94.9	69.1	64.1	5.8	14.9
+ Full STDG-UA	✓	✓	✓	✓	95.4	69.8	64.7	5.8	14.9
+ Full w/o STFP	--	✓	✓	✓	94.1	67.2	60.9	5.1	13.1
+ Full w/o AMSF	✓	--	✓	✓	94.4	67.8	62.0	5.8	14.9
+ Full w/o MGTM	✓	✓	--	✓	94.6	67.9	62.9	4.6	11.6
+ Full w/o memory refinement	✓	✓	✓	✓	95.0	68.9	63.8	5.5	14.0
+ Full w/o style perturbation	✓	✓	✓	✓	95.2	69.5	64.4	5.8	14.9
+ Full w/o consistency learning	✓	✓	✓	✓	95.1	69.4	64.3	5.8	14.9

For continuous UAV videos, temporal robustness was evaluated at a sampling rate of 30 FPS using frame timestamps. A detection was considered valid when its confidence was at least 0.25 and its intersection over union (IoU) with the ground-truth bounding box was at least 0.50. The detection continuity rate (DCR) was defined as the proportion of visible target frames with valid detections, while the maximum consecutive missed frames (MCMF) represented the longest continuous missed-detection interval. Normalized center jitter measured frame-to-frame variations in localization error relative to the image diagonal, and confidence variation was calculated as the mean absolute change in

detection confidence between consecutive matched frames. The generalization gap was defined as the difference between source-domain mAP@0.5 and average mAP@0.5 across the four unseen domains. For real-flight evaluation, distance-dependent detection and TTC estimation were assessed using synchronized flight measurements, with TTC errors quantified by MAE and RMSE. Small-object average precision (AP_s) was evaluated following the COCO criterion, where small objects were defined as ground-truth bounding boxes with an area below 32 × 32 pixels. AP_S was averaged over IoU thresholds from 0.50 to 0.95 in steps of 0.05.

3.2. Overall Detection Performance and Comparison with Representative Methods

Table 2 compares the overall performance of STDG-UAV with several representative object detectors. In addition to general-purpose detectors, YOLOv5s-NGN, a lightweight model specifically developed for air-to-air UAV detection, was included to provide a more task-relevant comparison under the same experimental protocol. The results show that STDG-UAV obtained a good balance between detection accuracy and complexity. Its precision and recall reached 93.0% and 89.4%, while mAP@0.5 and mAP@0.5:0.95 reached 95.4% and 69.8%. STDG-UAV achieved the highest precision and mAP values among the compared methods, although YOLOv5s-NGN showed a higher recall. Compared with YOLOv8s, STDG-UAV improved mAP@0.5 by 1.8%, with the number of parameters reducing from 11.2 M to 5.8 M. This indicates that our method improves detection performance while maintaining relatively low computational complexity. Compared with RT-DETR-R18, STDG-UAV achieved a 3.5% improvement in mAP@0.5:0.95, making it more suitable for resource-limited onboard deployment. Compared with YOLOv5s-NGN, STDG-UAV improved mAP@0.5, mAP@0.5:0.95, and AP_s by 1.9, 3.4, and 6.0 percentage points, respectively, although it requires more parameters and GFLOPs. This highlights its accuracy–complexity trade-off rather than simply its ability to minimize model size.

Notably, STDG-UAV achieves an AP_s of 64.7% for small targets, which is 5.2 and 8.4 percentage points higher than YOLOv8s and YOLO11n, respectively. This improvement is more pronounced than the gain in overall mAP. The result suggests that the proposed STFP mainly improves feature preservation and recognition for distant and small UAV targets, rather than simply increasing overall performance through greater model complexity.

Figure 6 shows qualitative detection results under challenging conditions. In particular, Figure 6a illustrates a representative distant-UAV detection example, where STDG-UAV provides more reliable target localization than the compared baseline detectors. In urban and vegetation backgrounds, baseline methods are more easily affected by complex textures, leading to false detections, missed detections, or localization errors. In contrast, STDG-UAV produces more concentrated responses around the target. The proposed model also shows more stable detection under motion blur and low-illumination conditions, indicating that spatial-detail enhancement and temporal information can provide complementary cues when target appearance is degraded. Overall, the quantitative and qualitative results demonstrate that STDG-UAV improves detection of small, weakly visible, and background-confused UAV targets while retaining a lightweight structure, providing a solid basis for continuous-video perception and real-time onboard applications.



Figure 6. UAV detection results under representative challenging conditions.

3.3. Ablation Study and Mechanism Analysis

3.3.1. Ablation Study of the Proposed Components

To evaluate the contribution of each core component to STDG-UAV, Table 3 presents both cumulative and controlled removal-based ablation results. The baseline model achieves 92.6% mAP@0.5, 64.2% mAP@0.5:0.95, and 56.7% AP_S. After introducing STFP, these metrics increase to 93.6%, 65.9%, and 60.2%, respectively. In particular, AP_S improves by 3.5 percentage points, indicating that spatial-detail and high-frequency structure preservation is especially effective for distant small targets. With the further addition of AMSF, mAP@0.5:0.95 and AP_S reached 67.3% and 62.3%, respectively, showing that dynamic fusion of P2-P5 features improves adaptation to UAV scale variation.

After MGTM is introduced, mAP@0.5:0.95 further increases to 69.1%, while AP_S reaches 64.1%. This suggests that cross-frame motion information can compensate for feature loss caused by blur or weak visibility in individual frames. With the domain-generalization strategy included, the complete model achieves 95.4% mAP@0.5, 69.8% mAP@0.5:0.95, and 64.7% AP_S. Notably, the domain-generalization module is activated only during training. Therefore, the model size and computational cost remain unchanged at 5.8 M parameters and 14.9 GFLOPs during inference.

To further isolate the contribution of each component, controlled removal-based ablations were conducted using the complete STDG-UAV as the reference. Removing STFP, AMSF, and MGTM reduced mAP@0.5:0.95 to 67.2%, 67.8%, and 67.9%, respectively, while AP_S decreased to 60.9%, 62.0%, and 62.9%. Removing memory refinement further reduced mAP@0.5:0.95 to 68.9%, indicating the contribution of historical feature refinement. In contrast, removing feature-style perturbation or consistency learning resulted in smaller source-domain performance reductions, with mAP@0.5:0.95 values of 69.5% and 69.4%, respectively. Their effects on unseen-domain robustness are further examined in Section 3.5. Overall, these results indicate that STFP and AMSF primarily enhance small-target

representation, MGTM improves temporal information aggregation, and DG contributes to cross-scene generalization.

3.3.2. Mechanism Analysis and Feature Visualization

Figure 7 further illustrates how the proposed modules work. Compared with the baseline network, the feature responses after STFP are more concentrated around the UAV target, while background responses are reduced. The Haar wavelet decomposition shows that the LH, HL, and HH high-frequency components retain useful information about body edges, rotors, and local contours. After fusion, the enhanced high-frequency response becomes more sensitive to small-target structures.

The adaptive weights for different scales further demonstrate the effect of AMSF. For distant small UAVs, the model assigns larger weights to P2 and P3 to preserve high-resolution spatial information. As the target size increases, more weight is gradually shifted toward deeper semantic features such as P4 and P5, indicating that the network can adapt dynamically to target-scale changes. In motion-blurred sequences, the motion-difference and temporal-gate responses generated by MGTM are mainly concentrated in the target region. Historical memory is then used to supplement degraded features in the current frame. The target-response curves also show that the single-frame baseline exhibits a clear response drop during the blurred period, whereas the model with MGTM maintains a more stable and continuous response. These visual results are consistent with the quantitative gains in Table 3 and show that STDG-UAV improves UAV perception from three complementary aspects: spatial detail, multi-scale information, and temporal motion.

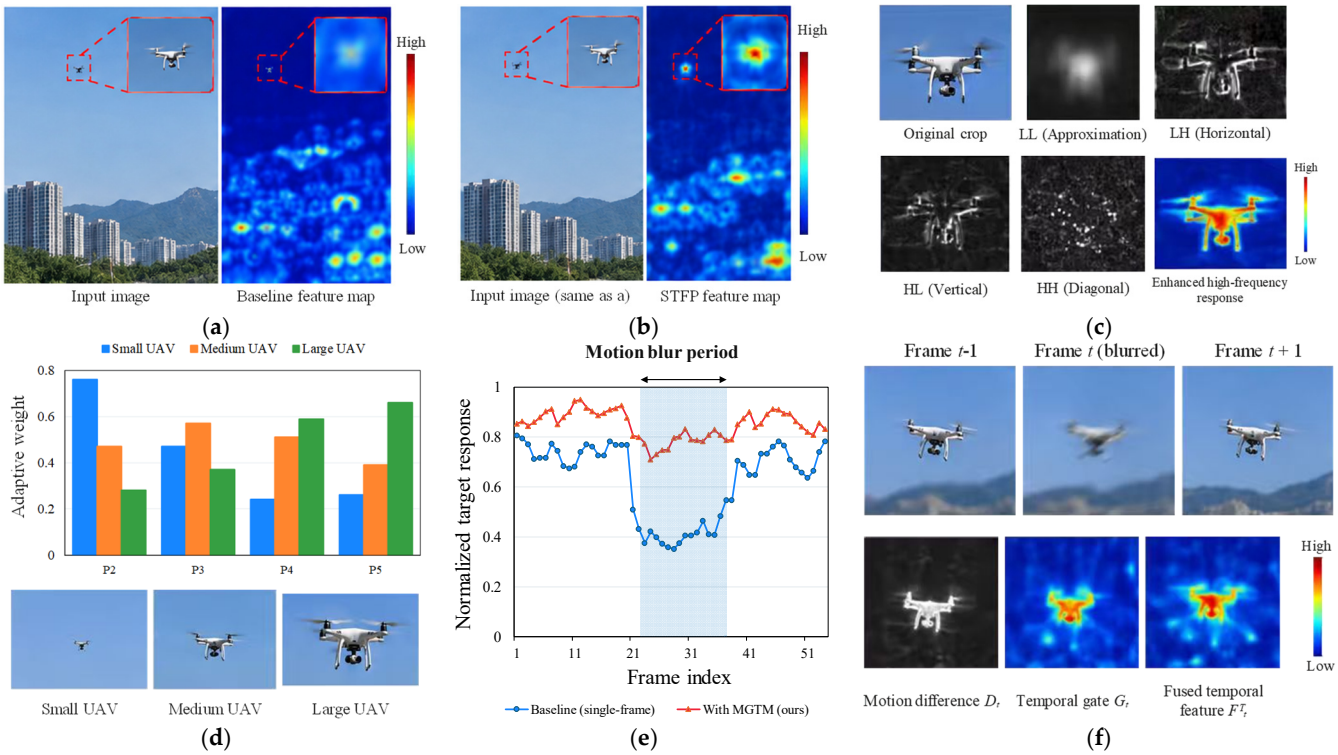


Figure 7. Mechanism visualization of spatial enhancement, adaptive multi-scale fusion, and temporal feature aggregation in STDG-UAV. (**a**) Baseline feature response for a distant UAV. (**b**) STFP-enhanced feature response. (**c**) High-frequency structural responses extracted by STFP. (**d**) Adaptive P2-P5 weights for different UAV scales. (**e**) Temporal target-response stability. (**f**) Motion-difference and temporal-gate responses under motion blur.

3.4. Temporal Robustness and Detection Continuity Analysis

To further evaluate the temporal stability of STDG-UAV in continuous flight videos, Table 4 compares the detection performance of conventional detectors and STARD-Net [37], a spatiotemporal detector designed for tiny airborne objects. The complete STDG-UAV achieves the highest detection continuity rate (DCR) of 96.2%, outperforming STARD-Net (94.1%) and the other comparison methods. Compared with STDG-UAV without MGTM, the DCR increases from 92.3% to 96.2%, while the number of maximum consecutive missed frames (MCMF) decreases from five to two. Compared with STARD-Net, STDG-UAV improves DCR by 2.1 percentage points and reduces the MCMF value from three to two. At a video evaluation rate of 30 FPS, five and two missed frames correspond to approximately 0.167 s and 0.067 s, respectively. These results indicate that temporal memory can reduce missed detections caused by fast motion, motion blur, and short-term weak visibility.

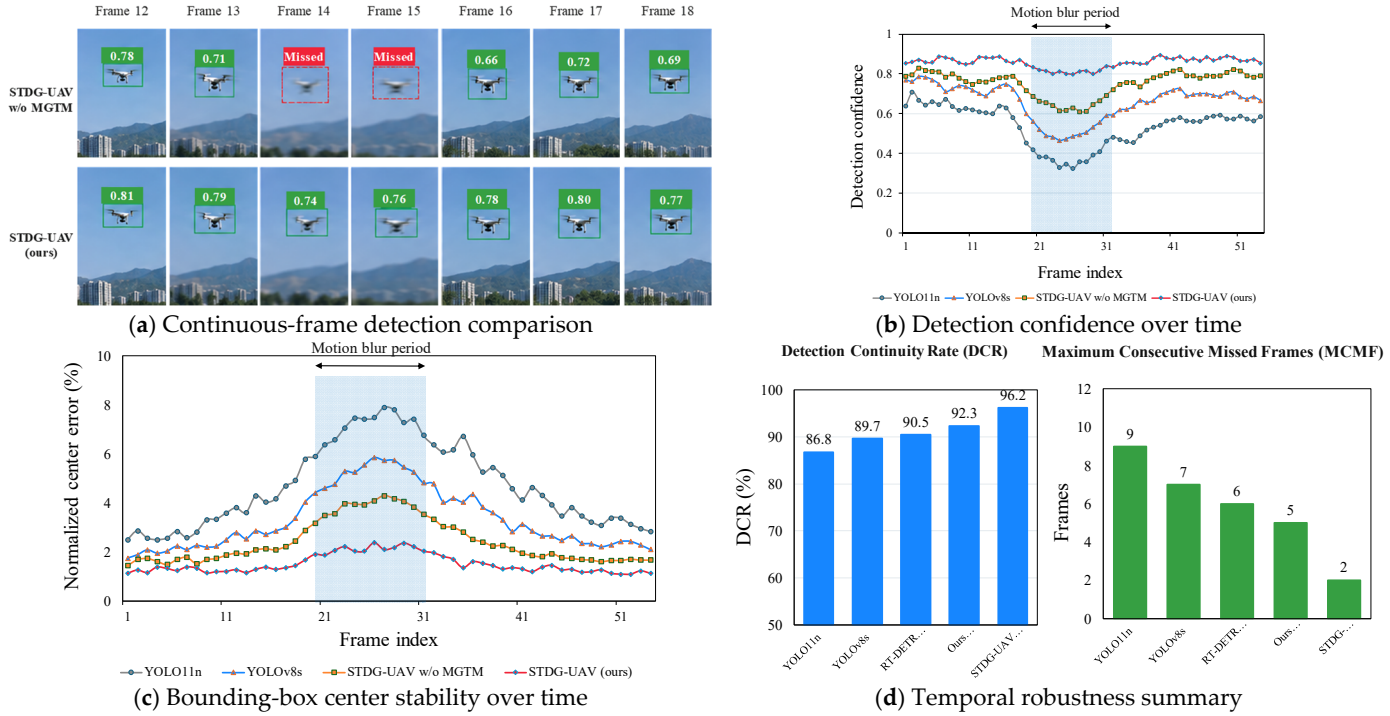
STDG-UAV also shows better bounding-box stability. Its normalized center jitter is only 2.0%, lower than the 3.1% obtained without MGTM and lower than those of the other comparison models. In addition, the confidence variation decreases from 0.081 to 0.049, lower than the 0.066 obtained by STARD-Net, indicating more stable predictions across consecutive frames. Although the real-flight videos contain stronger motion blur, scale variation, and background disturbance, STDG-UAV still achieves a video mAP@0.5 of 92.8%, exceeding STARD-Net by 1.1%. To further examine sequence-level robustness, the four independent test videos (S1-S4) were evaluated separately. The sequences contained 600, 650, 670, and 680 frames, respectively, with corresponding DCRs of 96.8%, 95.8%, 95.6%, and 96.6%. The mean sequence-level DCR was 96.2% ± 0.6% (SD), while the MCMF values were one, two, two, and two, respectively. The overall DCR was 96.2%, consistent with the results in Table 4. These sequence-level results indicate stable detection continuity across the evaluated videos.

Table 4. Temporal detection performance of different methods on challenging UAV video sequences.

Method	DCR (%) ↑	MCMF (Frames) ↓	Center Jitter (%) ↓	Confidence Variation ↓	Video mAP@0.5 (%) ↑
YOLO11n	86.8	9	4.8	0.124	86.9
YOLOv8s	89.7	7	4.1	0.106	88.5
RT-DETR-R18	90.5	6	3.8	0.098	89.1
STARD-Net [55]	94.1	3	2.7	0.066	91.7
STDG-UAV w/o MGTM	92.3	5	3.1	0.081	90.7
STDG-UAV	96.2	2	2.0	0.049	92.8

Figure 8 further illustrates these results. During the motion-blur interval, the model without MGTM produces consecutive missed detections, whereas the complete STDG-UAV continues to detect the target effectively. The confidence curves show that the baseline methods experience a larger confidence drop when the target appearance degrades, while STDG-UAV shows smaller fluctuations and recovers more quickly. The temporal variation in bounding-box center error also confirms that MGTM can reduce localization jitter. Overall, both the quantitative results and the frame-by-frame visualizations show that the main advantage of MGTM is not simply an improvement in single-frame accuracy, but greater consistency and reliability in continuous perception. Such stable

temporal output is particularly important for subsequent relative-motion estimation, TTC calculation, and collision-risk assessment.



(**a**) Continuous-frame detection comparison

(**b**) Detection confidence over time

(**c**) Bounding-box center stability over time

(**d**) Temporal robustness summary

Figure 8. Temporal robustness and detection continuity of representative methods in challenging UAV video sequences. Panel (**a**) shows a selected seven-frame segment (Frames 12–18), while panels (**b**–**d**) illustrate representative temporal responses and robustness metrics. Full evaluation results are reported in Table 4.

3.5. Cross-Domain Generalization and Robustness in Unseen Scenes

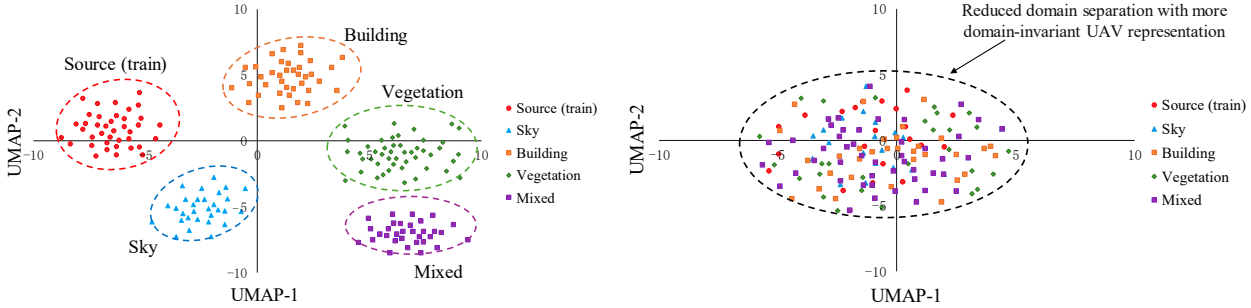
To evaluate the cross-domain generalization of STDG-UAV in unseen real-flight environments, Table 5 compares the detection performance of different models from the source domain to several unseen target domains. A MixStyle-based baseline (STDG-UAV w/o DG + MixStyle) was also included using the same detection architecture and source-domain training data, with feature-statistics mixing replacing the complete DG strategy. The unseen-domain evaluation comprised 12 independent flight sequences (8550 frames), with three sequences for each background: sky (2160 frames), building (2040 frames), vegetation (2250 frames), and mixed (2100 frames). The target-domain average mAP@0.5 was calculated as the unweighted mean of the four domain-specific results. This evaluation was conducted separately from the four challenging video sequences (2600 frames) used for temporal analysis in Section 3.4. All model parameters and evaluation thresholds were fixed based exclusively on source-domain validation data, without using unseen target-domain data for model selection. The complete STDG-UAV achieves an mAP@0.5 of 95.4% on the source domain, only 0.5 percentage points higher than the 94.9% obtained without DG. However, the average mAP@0.5 on the unseen target domains increases from 83.9% to 87.6%. The MixStyle-based baseline achieves an average target-domain mAP@0.5 of 85.2% and a generalization gap of 9.8 percentage points. Compared with MixStyle, the complete STDG-UAV improves the average target-domain mAP@0.5 by 2.4 percentage points and reduces the generalization gap by 2.0 percentage points, suggesting additional benefits from domain randomization and consistency learning. The improvements in the building, vegetation, and mixed scenes are 3.8, 4.0, and 4.3 percentage points, respectively, showing that the domain-generalization strategy is more effective under complex backgrounds and larger visual domain shifts. The sky domain is closer to the source-domain

distribution, so the improvement is relatively smaller. Moreover, STDG-UAV has the lowest generalization gap of 7.8 percentage points among all compared methods, indicating reduced performance degradation when transferring to unseen scenes. This indicates less performance degradation when transferring to unseen scenes.

Table 5. Cross-domain generalization performance on unseen real-flight scenes.

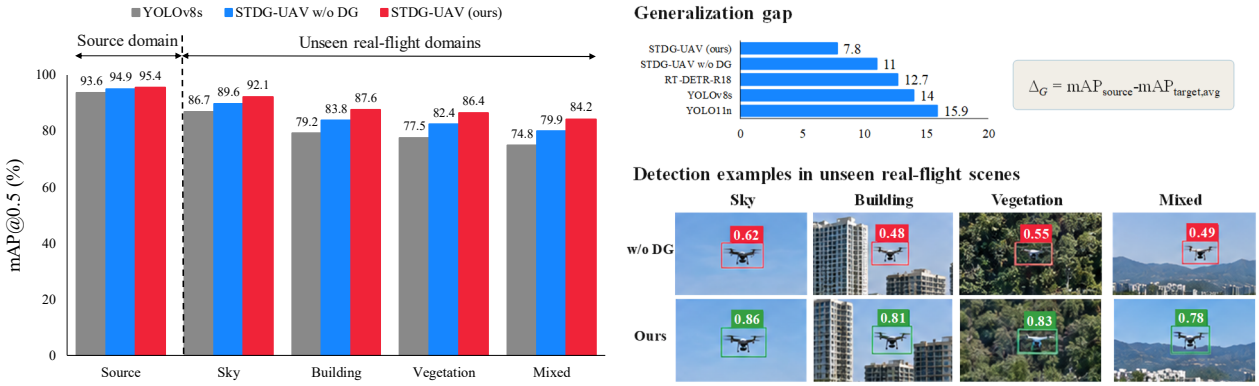
Method	Source mAP@0.5 (%)	Sky (%)	Building (%)	Vegetation (%)	Mixed (%)	Target Avg. (%)	Generalization Gap (%) ↓
YOLO11n	92.7	84.8	76.5	74.1	71.8	76.8	15.9
YOLOv8s	93.6	86.7	79.2	77.5	74.8	79.6	14.0
RT-DETR-R18	93.4	87.1	80.5	79.0	76.2	80.7	12.7
STDG-UAV w/o DG	94.9	89.6	83.8	82.4	79.9	83.9	11.0
STDG-UAV w/o DG + MixStyle	95.0	90.4	85.1	83.8	81.5	85.2	9.8
STDG-UAV	95.4	92.1	87.6	86.4	84.2	87.6	7.8

Figure 9 further shows the effect of domain generalization on feature distributions and detection results. The unseen-domain features used for UMAP visualization were extracted only after model training and selection, and were not used for parameter updating, threshold tuning, or model selection. Without DG, UAV features from different background conditions form clearly separated clusters in the UMAP space. After introducing image-level randomization, feature-style perturbation, and consistency learning, the feature distributions from different domains become closer, and the distance between the source and target domains is reduced. This suggests that the model becomes less dependent on background and imaging style. Detection examples from unseen scenes also show improved confidence and localization stability after DG is applied, especially under building, vegetation, and mixed backgrounds. Overall, the main role of DG is not to increase source-domain benchmark accuracy, but to improve robustness to unseen real-flight environments, leading to more stable cross-scene UAV perception.



(**a**) Feature distribution without domain generalization (**b**) Feature distribution with domain generalization

https://doi.org/10.3390/drones10110827



(**c**) Cross-domain detection performance

(**d**) Generalization gap and detection examples in unseen scenes

Figure 9. Cross-domain generalization of STDG-UAV across unseen real-flight environments.

3.6. Onboard Real-Flight Validation and Collision-Risk Analysis

3.6.1. Onboard Real-Time and Distance-Dependent Detection

To evaluate the practical performance of STDG-UAV in a real onboard environment, the model was deployed on an edge-computing platform and tested in dual-UAV flight experiments. The onboard platform consisted of an NVIDIA Jetson Orin NX (16 GB, 25 W power mode), a Sony IMX219 RGB camera (Arducam B0183, 2.8 mm lens, approximately 75° horizontal field of view), and a Pixhawk-based flight controller. Videos were captured at 1920 × 1080 pixels and 30 FPS, while 640 × 640 pixels for network inference. It was deployed using NVIDIA TensorRT with FP16 precision and a batch size of one. As shown in Figure 10a, the model achieved real-time inference at approximately 28.7 FPS with 1920 × 1080 video input, with a per-frame latency of about 34.8 ms, using a 1920 × 1080 camera stream resized to 640 × 640 pixels for network inference. An illustrative processing-time breakdown assigns 4.5 ms to image preprocessing, 25.0 ms to network inference, 3.5 ms to detection postprocessing, and 1.8 ms to TTC/risk computation, totaling 34.8 ms per frame (28.7 FPS). Under the same onboard hardware and inference settings, the estimated processing throughputs of YOLOv8s and RT-DETR-R18 are 56.2 and 37.3 FPS, respectively. Although STDG-UAV requires additional processing for spatial–temporal feature enhancement, its processing throughput remains close to the 30 FPS camera acquisition rate. These results indicate that STDG-UAV can meet the requirements of continuous onboard perception. Figure 10b further shows the detection rate at different relative distances. As the target distance increased from 10 m to 70 m, the detection rate of all models gradually decreased because the target occupied fewer pixels and contained less visual detail. Compared with YOLOv8s, RT-DETR-R18, and STDG-UAV without MGTM, the complete STDG-UAV maintained higher detection rates at medium and long distances. The advantage was more evident within the 40–70 m range, indicating that STFP, AMSF, and temporal feature fusion improve the detectability of distant small UAVs.

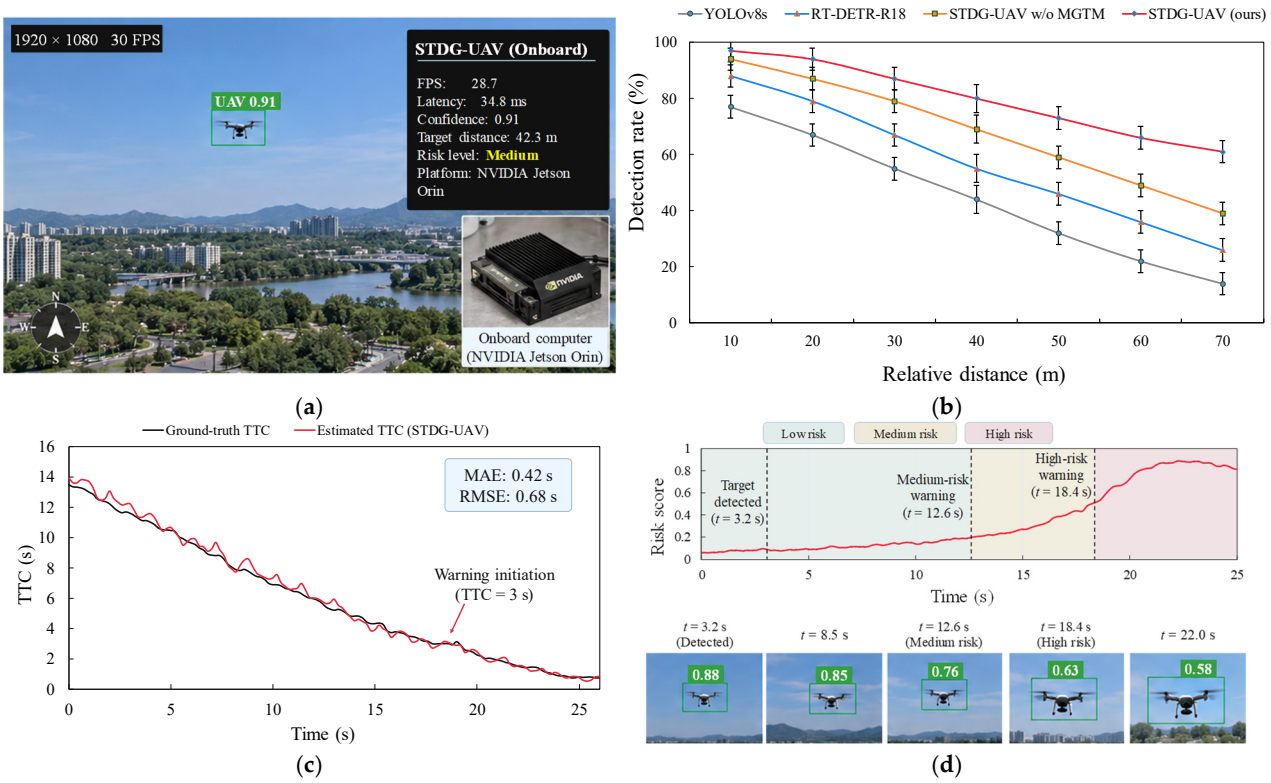


Figure 10. Onboard real-flight validation and collision-risk evaluation of STDG-UAV. (**a**) Onboard real-time UAV detection during flight. (**b**) Detection performance at different relative distances. (**c**) Comparison between estimated and ground-truth TTC. (**d**) Temporal evolution of collision-risk level during an approaching encounter.

3.6.2. Collision-Risk and TTC Evaluation

The TTC estimation and collision-risk perception performance was first examined in a representative frontal approaching scenario using consecutive visual observations. The reference relative distance was obtained from RTK-GNSS positioning records of the observer and target UAVs, with three-dimensional positions recorded at 10 Hz. The flight logs were synchronized with the 30 FPS camera stream using GNSS timestamps and a calibrated time offset. Linear interpolation was applied to obtain the corresponding UAV positions at each video frame. The relative distance was calculated from the synchronized positions, while the radial closing speed was estimated using a 0.5 s sliding linear fit. For approximately frontal approaches, the reference TTC was calculated as the ratio of relative distance to radial closing speed under a constant-velocity assumption. The resulting RTK-derived TTC values were compared with visual estimates at matched timestamps. As shown in Figure 10c, the visually estimated TTC follows the ground-truth TTC closely, with an MAE of 0.42 s and an RMSE of 0.68 s. As the two UAVs continued to approach each other, the TTC gradually decreased, and the system consistently captured the relative approaching trend. When the TTC approached the predefined warning threshold, the risk module began to output a higher warning level. Figure 10d shows the evolution of collision risk during the complete approaching process. The risk state gradually changed from low to medium and then to high, consistent with the increasing target size and approaching motion. These results show that STDG-UAV can not only perform real-time UAV detection but also convert stable temporal detections into interpretable TTC and collision-risk information. This provides perception-level support for warning, trajectory planning, and collision-avoidance decisions in upper-level sense-and-avoid systems.

To further evaluate the reliability of collision-risk perception, Table 6 summarizes the results of 12 flight trials under different encounter conditions. For frontal approaches, the

TTC MAE/RMSE values were 0.38/0.62 s and 0.46/0.74 s at lower and higher closing speeds, respectively, with warning lead times of 5.0 and 3.6 s. The higher-speed approaches produced one false warning and one missed warning. One false warning was also observed in the lateral-crossing and near-miss scenarios, respectively, while no false warnings occurred during receding flights. TTC errors were not reported for lateral or non-closing trajectories because the monocular expansion-based TTC estimate is primarily applicable to approaching motion. These results indicate the potential of STDG-UAV for early collision-risk warning while highlighting its limitations under complex relative-motion conditions.

Table 6. Multi-flight TTC estimation and collision-warning performance under different UAV encounter scenarios.

Flight Scenario	Flights (*n*)	TTC MAE/RMSE (s)	False Warning (*n*)	Missed Warning (*n*)	Warning Lead Time (s)
Frontal approach (lower closing speed)	3	0.38/0.62	0	0	5.0
Frontal approach (higher closing speed)	3	0.46/0.74	1	1	3.6
Lateral crossing	2	N/A	1	0	2.7
Near-miss encounter	2	N/A	1	--	N/A
Receding/non-closing flight	2	N/A	0	--	N/A

Several limitations should be acknowledged. First, monocular expansion-based TTC estimation is mainly applicable to approximately frontal approaches and may be affected by camera ego-motion, changes in UAV attitude, and lateral motion. Second, the evaluated UAV types and flight environments do not fully cover severe weather, prolonged occlusion, or more diverse aircraft configurations. Third, although the proposed framework provides real-time detection and collision-risk warnings, closed-loop autonomous collision avoidance has not been validated. Moreover, the reported onboard processing latency does not represent the complete camera-to-action response time. Future work will focus on uncertainty-aware motion estimation, broader real-flight validation, and closed-loop collision-avoidance experiments.

4. Conclusions

This study developed STDG-UAV, a lightweight spatial–temporal domain-generalized framework for drone-to-drone detection and onboard collision-risk perception. The proposed method integrates spatial tiny-target feature preservation, adaptive multi-scale fusion, motion-guided temporal memory, and source-only domain-generalized learning to address small target size, temporal instability, and cross-scene domain shifts in real flight environments. Experimental results show that STDG-UAV achieved a precision of 93.0%, recall of 89.4%, mAP@0.5 of 95.4%, and mAP@0.5:0.95 of 69.8%, while reaching an AP_S of 64.7% with only 5.8 M parameters and 14.9 GFLOPs. The temporal evaluation further demonstrates a detection continuity rate of 96.2%, with the maximum consecutive missed frames reduced to two and the normalized center jitter limited to 2.0%. Under unseen real-flight domains, STDG-UAV achieves an average mAP@0.5 of 87.6% and reduces the generalization gap to 7.8 percentage points. Onboard flight experiments further confirm real-time operation at 28.7 FPS with 34.8 ms latency. The estimated TTC closely

follows the ground truth, with an MAE of 0.42 s and an RMSE of 0.68 s. Overall, STDG-UAV provides an integrated perception framework that combines spatial, temporal, and domain-generalized learning for robust drone-to-drone detection and onboard collision-risk assessment. In the future, we will extend this method to more diverse UAV types and challenging environments, with particular emphasis on closed-loop autonomous avoidance validation.

Author Contributions: Conceptualization, B.C. and Z.L.; methodology, B.C. and Z.L.; software, B.C.; validation, B.C.; formal analysis, Z.L.; investigation, Z.L.; resources, B.C. and Z.L.; data curation, B.C. and Z.L.; writing—original draft preparation, B.C. and Z.L.; writing—review and editing, B.C. and Z.L.; visualization, B.C. and Z.L.; supervision, B.C. and Z.L.; project administration, B.C. and Z.L. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Data Availability Statement: The public source-domain UAV dataset used for model training and evaluation in this study is available from the Drone Dataset (UAV) at https://www.kaggle.com/datasets/dasmehdixtr/drone-dataset-uav (accessed on 16 September 2026). The self-collected real-flight videos, corresponding UAV annotations, experimental data partitions, synchronized flight-positioning logs, and evaluation scripts supporting the reported results are available from the corresponding author upon reasonable request.

Acknowledgments: During the preparation and revision of this manuscript, the authors used OpenAI ChatGPT (GPT-5.6 Sol) for English-language editing and refinement of academic expression. The authors have reviewed and edited the output and take full responsibility for the content of this publication.

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Liu, Z.; Nyoni, H.; Mina, O.; Shen, S. Microplastics release from recycled plastic modified asphalt: Detection, characterization, and environmental implications. *J. Clean. Prod.* **2026**, *577*, 149471. https://doi.org/10.1016/j.jclepro.2026.149471.
2. Tang, G.; Ni, J.; Zhao, Y.; Gu, Y.; Cao, W. A Survey of Object Detection for UAVs Based on Deep Learning. *Remote Sens.* **2024**, *16*, 149.
3. Al-lQubaydhi, N.; Alenezi, A.; Alanazi, T.; Senyor, A.; Alanezi, N.; Alotaibi, B.; Alotaibi, M.; Razaque, A.; Hariri, S. Deep learning for unmanned aerial vehicles detection: A review. *Comput. Sci. Rev.* **2024**, *51*, 100614. https://doi.org/10.1016/j.cosrev.2023.100614.
4. Liu, Z.; Kwan, M.P.; Jiang, W.; Liu, Y.; Cui, B. A Lightweight Multiscale Fusion Framework for Traffic Vehicle Detection From Satellite Remote Sensing, UAV, and CCTV Imagery. *IEEE Trans. Geosci. Remote Sens.* **2026**, *64*, 5632717. https://doi.org/10.1109/TGRS.2026.3713370.
5. Liu, Z.; Kwan, M.-P.; Hu, X.; Shen, S.; Gu, X.; Cui, B. Automated InSAR-LiDAR Fusion Monitoring and Decision Support for Urban Road Collapse Risk Assessment. *Autom. Constr.* 2026, *Accepted*.
6. Jia, W.; Lv, L.; Duan, R.; Sun, T.; Sun, W. A Reinforcement Learning-Based Adaptive Grey Wolf Optimizer for Simultaneous Arrival in Manned/Unmanned Aerial Vehicle Dynamic Cooperative Trajectory Planning. *Drones* **2025**, *9*, 723.
7. Freitas, A.; Rabbath, C.A.; Williams, C.; Lechevin, N.; Givigi, S. Multi-Unmanned Aerial Vehicles Cooperative Tactics: A Survey. *IEEE Access* **2025**, *13*, 119897–119921. https://doi.org/10.1109/ACCESS.2025.3583981.
8. Seidaliyeva, U.; Ilipbayeva, L.; Taissariyeva, K.; Smailov, N.; Matson, E.T. Advances and Challenges in Drone Detection and Classification Techniques: A State-of-the-Art Review. *Sensors* **2024**, *24*, 125.
9. Fiorio, M.; Galatolo, R.; Di Rito, G. Development and Experimental Validation of a Sense-and-Avoid System for a Mini-UAV. *Drones* **2025**, *9*, 96.
10. Cui, B.; Liu, Z.; Yang, Q. UAV-YOLO12: A Multi-Scale Road Segmentation Model for UAV Remote Sensing Imagery. *Drones* **2025**, *9*, 533.

11. Biswas, T.K.; Zannat, A.; Ishtiaq, W.; Hossain, M.A. A novel unified lightweight temporal-spatial transformer approach for intrusion detection in drone networks. *Sci. Rep.* **2026**, *16*, 14473. https://doi.org/10.1038/s41598-026-45063-6.
12. Liang, C.; Liu, L.; Liu, C. Multi-UAV autonomous collision avoidance based on PPO-GIC algorithm with CNN–LSTM fusion network. *Neural Netw.* **2023**, *162*, 21–33. https://doi.org/10.1016/j.neunet.2023.02.027.
13. Kim, M.; Ladosz, P.; Oh, H. Monocular vision-based time-to-collision estimation for small drones by domain adaptation of simulated images. *Expert Syst. Appl.* **2022**, *199*, 116973. https://doi.org/10.1016/j.eswa.2022.116973.
14. Park, J.; Choi, A.J. Vision-Based In-Flight Collision Avoidance Control Based on Background Subtraction Using Embedded System. *Sensors* **2023**, *23*, 6297.
15. Tang, W.; Li, Q.; Peng, Y.; Hao, H.; Kang, W.; Zhang, X.; Hou, L.; Lu, H. TriCross-D2D: A Cross-Scene, Cross-View, and Cross-Weather Dataset for Drone-to-Drone Detection. *Drones* **2026**, *10*, 459.
16. Cheng, Q.; Wang, Y.; He, W.; Bai, Y. Lightweight air-to-air unmanned aerial vehicle target detection model. *Sci. Rep.* **2024**, *14*, 2609. https://doi.org/10.1038/s41598-024-53181-2.
17. Zhao, Y.; Wang, L.; Lei, G.; Guo, C.; Ma, Q. Lightweight UAV Small Target Detection and Perception Based on Improved YOLOv8-E. *Drones* **2024**, *8*, 681.
18. Xu, L.; Zhao, Y.; Zhai, Y.; Huang, L.; Ruan, C. Small Object Detection in UAV Images Based on YOLOv8n. *Int. J. Comput. Intell. Syst.* **2024**, *17*, 223. https://doi.org/10.1007/s44196-024-00632-3.
19. Feng, F.; Hu, Y.; Li, W.; Yang, F. Improved YOLOv8 algorithms for small object detection in aerial imagery. *J. King Saud Univ. Comput. Inf. Sci.* **2024**, *36*, 102113. https://doi.org/10.1016/j.jksuci.2024.102113.
20. Liu, Z.; Wu, W.; Wang, D.; Cui, B.; Gu, X. Automatic extraction and 3D modeling of real road scenes using UAV imagery and deep learning semantic segmentation. *Int. J. Digit. Earth* **2024**, *17*, 2365970.
21. Luo, W.; Yuan, S. Enhanced YOLOv8 for small-object detection in multiscale UAV imagery: Innovations in detection accuracy and efficiency. *Digit. Signal Process.* **2025**, *158*, 104964. https://doi.org/10.1016/j.dsp.2024.104964.
22. Xiao, Y.; Di, N. SOD-YOLO: A lightweight small object detection framework. *Sci. Rep.* **2024**, *14*, 25624. https://doi.org/10.1038/s41598-024-77513-4.
23. Yang, G.; Jiang, Y.; Wang, S.; Chen, K. VinsFusion-Line: Binocular Vision Inertial Navigation Real-Time SLAM System Based on Line Features. *Cybern. Syst.* **2026**, *57*, 350–375. https://doi.org/10.1080/01969722.2025.2606002.
24. Kong, Y.; Shang, X.; Jia, S. Drone-DETR: Efficient Small Object Detection for Remote Sensing Image Using Enhanced RT-DETR Model. *Sensors* **2024**, *24*, 5496.
25. Luo, J.; Liu, Z.; Wang, Y.; Tang, A.; Zuo, H.; Han, P. Efficient Small Object Detection You Only Look Once: A Small Object Detection Algorithm for Aerial Images. *Sensors* **2024**, *24*, 7067.
26. Deng, H.; Zhang, S.; Wang, X.; Han, T.; Ye, Y. USD-YOLO: An Enhanced YOLO Algorithm for Small Object Detection in Unmanned Systems Perception. *Appl. Sci.* **2025**, *15*, 3795.
27. Cui, B.; Wang, H. Coupled prediction of viscoelastic properties of asphalt mixtures using a multi-task physics-constrained neural network. *Adv. Eng. Inform.* **2026**, *72*, 104437. https://doi.org/10.1016/j.aei.2026.104437.
28. Liu, Z.; Gu, X.; Kwan, M.-P.; Cui, B. Integrated ultrasonic testing and numerical simulation for damage detection in steel bridge deck pavements. *Eng. Struct.* **2026**, *355*, 122421. https://doi.org/10.1016/j.engstruct.2026.122421.
29. Ni, J.; Zhu, S.; Tang, G.; Ke, C.; Wang, T. A Small-Object Detection Model Based on Improved YOLOv8s for UAV Image Scenarios. *Remote Sens.* **2024**, *16*, 2465.
30. Han, W.; Li, J.; Wang, S.; Wang, Y.; Yan, J.; Fan, R.; Zhang, X.; Wang, L. A context-scale-aware detector and a new benchmark for remote sensing small weak object detection in unmanned aerial vehicle images. *Int. J. Appl. Earth Obs. Geoinf.* **2022**, *112*, 102966. https://doi.org/10.1016/j.jag.2022.102966.
31. Jiang, L.; Yuan, B.; Du, J.; Chen, B.; Xie, H.; Tian, J.; Yuan, Z. MFFSODNet: Multiscale Feature Fusion Small Object Detection Network for UAV Aerial Images. *IEEE Trans. Instrum. Meas.* **2024**, *73*, 5015214. https://doi.org/10.1109/TIM.2024.3381272.
32. Fang, H.; Ding, L.; Wang, L.; Chang, Y.; Yan, L.; Han, J. Infrared Small UAV Target Detection Based on Depthwise Separable Residual Dense Network and Multiscale Feature Fusion. *IEEE Trans. Instrum. Meas.* **2022**, *71*, 5019120. https://doi.org/10.1109/TIM.2022.3198490.
33. Chen, H.; Sun, Q.; Li, F.; Tang, Y. Computer vision tasks for intelligent aerospace perception: An overview. *Sci. China Technol. Sci.* **2024**, *67*, 2727–2748. https://doi.org/10.1007/s11431-024-2714-4.
34. Pan, Y.; Hu, X.; List, G.; Zhou, X. Review of volume-delay functions integrating traditional models and emerging AI technologies. *Commun. Transp. Res.* **2026**, *6*, 9640003. https://doi.org/10.26599/COMMTR.2026.9640003.

35. Liu, X.; Fang, J.; Liu, H.; Gong, Z.; Peng, Z.; Dong, J. OIL: A Closed-Loop Periodic Batch Adaptation Framework for Object Detection on Industrial Variable-Speed Conveyor Lines. *Sensors* **2026**, *26*, 5423.
36. Xu, H.; Ling, Z.; Yuan, X.; Wang, Y. A video object detector with Spatio-Temporal Attention Module for micro UAV detection. *Neurocomputing* **2024**, *597*, 127973. https://doi.org/10.1016/j.neucom.2024.127973.
37. He, F.; Li, Q.; Zhao, X.; Huang, K. Temporal-adaptive sparse feature aggregation for video object detection. *Pattern Recognit.* **2022**, *127*, 108587. https://doi.org/10.1016/j.patcog.2022.108587.
38. Wang, X.; Huang, Z.; Liao, B.; Huang, L.; Gong, Y.; Huang, C. Real-time and accurate object detection in compressed video by long short-term feature aggregation. *Comput. Vis. Image Underst.* **2021**, *206*, 103188. https://doi.org/10.1016/j.cviu.2021.103188.
39. Cores, D.; Brea, V.M.; Mucientes, M. Spatiotemporal tubelet feature aggregation and object linking for small object detection in videos. *Appl. Intell.* **2023**, *53*, 1205–1217. https://doi.org/10.1007/s10489-022-03529-w.
40. Cai, Z.; Zhang, Y.; Miao, Q. AEDA-Net: Attention-Embedded Domain Adaptation Network for UAV Fault Detection Under Heterogeneous Flight Status. *IEEE Trans. Instrum. Meas.* **2026**, *75*, 3518511. https://doi.org/10.1109/TIM.2026.3706227.
41. Xiao, D.; Yu, L.; Zhang, Y.; Fang, J.; Li, Y.; Xu, Q. Cross-Condition Aerodynamic Control Surface Fault Diagnosis for Fixed-Wing UAV With Dynamic Domain Adaptive Dropout Regularization and Hybrid Alignment. *IEEE Trans. Instrum. Meas.* **2026**, *75*, 3507020. https://doi.org/10.1109/TIM.2026.3666043.
42. Liu, X.; Li, Z.; Zhou, Y.; Peng, Y.; Luo, J. Camera–Radar Fusion with Modality Interaction and Radar Gaussian Expansion for 3D Object Detection. *Cyborg and Bionic Systems* **2024**, *5*, 0079. https://doi.org/10.34133/cbsystems.0079.
43. Zhang, H.; Lin, W.; Guo, C.; Cao, L.; Gou, H.; Li, J.; Li, J. Lower-Boundary-Aware Sky–Ground Segmentation for Robust Horizon Estimation in Low-Altitude UAV Remote Sensing. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2026**, 19, 1–22. https://doi.org/10.1109/JSTARS.2026.3744321.
44. Cho, Y.; Karmann, C.; Andersen, M. Clear vs. overcast: Effects of sky condition on daylit view perception in virtual reality. *Build. Environ.* **2025**, *285*, 113615. https://doi.org/10.1016/j.buildenv.2025.113615.
45. Wang, M.; Liu, Y.; Liu, H. An enhanced domain generalization method for object detection based on text guided feature disentanglement. *Digit. Signal Process.* **2025**, *156*, 104855. https://doi.org/10.1016/j.dsp.2024.104855.
46. Michailidis, E.T.; Maliatsos, K.; Skoutas, D.N.; Vouyioukas, D.; Skianis, C. Secure UAV-Aided Mobile Edge Computing for IoT: A Review. *IEEE Access* **2022**, *10*, 86353–86383. https://doi.org/10.1109/ACCESS.2022.3199408.
47. Cheriguene, Y.; Jaafar, W.; Yanikomeroglu, H. Federated Learning in UAV-Assisted MEC Systems: A Comprehensive Survey. *IEEE Open J. Commun. Soc.* **2025**, *6*, 7645–7676. https://doi.org/10.1109/OJCOMS.2025.3608657.
48. UH, D.; Kumar, J.P. Enhanced Transfer Learning-Based CNN for Abnormal Human Activity Detection in Video Surveillance Using Spatial-Temporal Features. *Cybern. Syst.* **2026**, *57*, 1996–2025. https://doi.org/10.1080/01969722.2025.2521708.
49. Wei, J.; Chen, H.; Yao, L.; Hou, X.; Zhang, R.; Shi, L.; Sun, J.; Hu, C.; Wei, X.; Jia, W. BioCompNet: A Deep Learning Workflow Enabling Automated Body Composition Analysis toward Precision Management of Cardiometabolic Disorders. *Cyborg Bionic Syst.* **2025**, *6*, 0381. https://doi.org/10.34133/cbsystems.0381.
50. Li, A.; Xu, Z.; Li, W.; Chen, Y.; Pan, Y. Urban Signalized Intersection Traffic State Prediction: A Spatial–Temporal Graph Model Integrating the Cell Transmission Model and Transformer. *Appl. Sci.* **2025**, *15*, 2377.
51. Cui, B.; Wang, H. Optimization of risk-based and cost-minimized maintenance strategy for corroded pipeline using reinforcement learning. *Reliab. Eng. Syst. Saf.* **2026**, *276*, 112929. https://doi.org/10.1016/j.ress.2026.112929.
52. Li, Z.; Zheng, W.; Yang, L.; Ma, L.; Zhou, Y.; Peng, Y. MonoAux: Fully Exploiting Auxiliary Information and Uncertainty for Monocular 3D Object Detection. *Cyborg Bionic Syst.* **2024**, *5*, 0097. https://doi.org/10.34133/cbsystems.0097.
53. Shajeena, J.; Govindasamy, B.; Gnanasundaram, M.; Robinson Joel, M. Mobile-Le Harmonic Fusion Network for Object Recognition and SiamMoT Based Multi-Object Tracking Using Video Surveillance. *Cybern. Syst.* **2026**, *57*, 866–896. https://doi.org/10.1080/01969722.2025.2590761.
54. Özel, M. Drone Dataset (UAV). 2019. Available online: https://github.com/dasmehdix/drone-dataset (accessed on 16 September 2026).
55. Rahman, M.H.; Madria, S. STARD-Net: SpatioTemporal Attention for Robust Detection of Tiny Airborne Objects from Moving Drones. *ACM Trans. Spat. Algorithms Syst.* **2026**, *12*, 3. https://doi.org/10.1145/3787467.

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.