drones

MDPI

Article

TMRGBT-D2D: A Temporal Misaligned RGB-Thermal Dataset for Drone-to-Drone Target Detection

Hexiang Hao [1], Yueping Peng [1,*], Zecong Ye [1,2], Baixuan Han [1], Wei Tang [1], Wenchao Kang [1], Xuekai Zhang [1], Qilong Li [1] and Wenchao Liu [1]

[1] School of Information Engineering, Engineering University of PAP, Xi'an 710086, China; hhx1214s@163.com (H.H.); yzc6666@yeah.net (Z.Y.); hbx911wj@163.com (B.H.); 13414667546@163.com (W.T.); 18259530623@163.com (W.K.); 18510004596@163.com (X.Z.); liqilong20020705@163.com (Q.L.); liuwch3@mail3.sysu.edu.cn (W.L.)

[2] Unit Command Department, Officers College of PAP, Chengdu 610213, China

* Correspondence: percy001@163.com

Abstract

In the field of drone-to-drone detection tasks, the issue of fusing temporal information with infrared and visible light data for detection has been rarely studied. This paper presents the first temporal misaligned rgb-thermal dataset for drone-to-drone target detection, named TMRGBT-D2D. The dataset covers various lighting conditions (i.e., high-light scenes captured during the day, medium-light and low-light scenes captured at night, with night scenes accounting for 38.8% of all data), different scenes (sky, forests, buildings, construction sites, playgrounds, roads, etc.), different seasons, and different locations, consisting of a total of 42,624 images organized into sequential frames extracted from 19 RGB-T video pairs. Each frame in the dataset has been meticulously annotated, with a total of 94,323 annotations. Except for drones that cannot be identified under extreme conditions, infrared and visible light annotations are one-to-one corresponding. This dataset presents various challenges, including small object detection (the average size of objects in visible light images is approximately 0.02% of the image area), motion blur caused by fast movement, and detection issues arising from imaging differences between different modalities. To our knowledge, this is the first temporal misaligned rgb-thermal dataset for drone-to-drone target detection, providing convenience for research into rgb-thermal image fusion and the development of drone target detection.

Keywords: drone-to-drone detection; temporal information; misaligned rgb-thermal dataset; deep learning

check for updates

Academic Editors: Diego González-Aguilera and David R. Green

Received: 11 August 2025
Revised: 26 September 2025
Accepted: 8 October 2025
Published: 10 October 2025

Citation: Hao, H.; Peng, Y.; Ye, Z.; Han, B.; Tang, W.; Kang, W.; Zhang, X.; Li, Q.; Liu, W. TMRGBT-D2D: A Temporal Misaligned RGB-Thermal Dataset for Drone-to-Drone Target Detection. *Drones* **2025**, *9*, 694. https://doi.org/10.3390/drones9100694

Copyright: © 2025 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 (https://creativecommons.org/licenses/by/4.0/).

1. Introduction

Object detection, a core technology in computer vision, facilitates real-time identification and localization of objects within images or video sequences. It has been widely employed in numerous domains, including autonomous driving, intelligent transportation, security surveillance, medical imaging, remote sensing, and defense-related applications, substantially advancing the level of automation and intelligence in these systems.

In recent years, drones have been widely used in many fields [1–4]. Due to their high stealth, low cost, and ease of acquisition, they are easily exploited by criminals for close-range reconnaissance and intelligence gathering, especially in sensitive areas. They can also be used to carry improvised explosive devices, posing a threat to public safety. Additionally, unauthorized civilian drones may infringe on personal privacy or interfere

with air traffic, thereby posing a threat to aviation safety. The misuse of drones has become one of the primary threats to aviation safety.

Therefore, the detection and early warning of unauthorized drones in restricted airspace is of significant practical importance. Current drone detection technologies primarily include radar-based detection [5], acoustic sensing [6], radio frequency (RF) signal analysis [7], and computer vision-based methods [8]. However, drones present distinct challenges: their small size and low radar cross section make them difficult to detect with radar systems, while their flight paths are often highly unpredictable. Acoustic detection is limited by rapid signal attenuation with distance and susceptibility to environmental noise. RF-based approaches often rely on pre-defined signal signatures, rendering them ineffective against autonomous drones that do not emit known communication signals. Given the stringent payload constraints on micro-drones, vision-based detection coupled with computer vision algorithms remains one of the most feasible approaches for air-to-air detection scenarios.

Optical detection is the most suitable solution for detecting drones on drone platforms. With the continuous development of drone technology and the improvement in onboard computer performance, it has become possible to equip drone platforms with high-performance computers and optical cameras, and use artificial intelligence algorithms to control drone platforms and accurately detect and identify drone targets. Optical detection offers the following advantages: first, drone platforms have limited payload capacity, and visual sensors are lightweight; second, most drone products already come equipped with visual sensors, making development and implementation convenient; third, visual sensors can provide the drone with rich information about its surrounding environment; and fourth, visual sensors have low power consumption, which benefits the drone platform's endurance, enabling it to track drone targets over an extended period of time.

To our knowledge, there is currently no bimodal drone detection benchmark for air-to-air scenarios, and existing drone target detection datasets typically only include unimodal data. To address this issue, we propose a temporal misaligned rgb-thermal dataset for drone-to-drone target detection, TMRGBT-D2D, to facilitate research on drone detection tasks. In our context, "temporal misaligned rgb thermal dataset" is defined as follows: The RGB and thermal image frames are captured in parallel and are temporally correlated (i.e., they form sequential pairs from the same moment in time), but they are not spatially aligned due to the parallax effect and differences in the fields of view between the two sensors. This means that while the frames are temporal pairs, a target drone may appear at different relative positions or with different perspectives in the two modalities, especially when at varying depths, making traditional image registration methods ineffective. The dataset contains high-quality, high-resolution visible and infrared video sequences, each of which has been annotated with high-quality labels. Notably, in TMRGBT-D2D, visible and infrared video sequences are paired to enable research on bimodal information fusion problems.

The main contributions of this paper are as follows: building a temporal misaligned rgb-thermal dataset for drone-to-drone target detection, TMRGBT-D2D, which consists of 19 RGB-T video pairs, where "RGB" and "T" represent visible light and thermal infrared, respectively, and conducting experimental analysis on 10 target detection algorithms in TMRGBT-D2D.

2. Related Works

2.1. RGBT Object Detection

For the fusion of infrared and visible light images, there are typically three strategies: pre-fusion (image fusion), mid-fusion (feature-level fusion), and post-fusion (decision-level fusion) [9]. Pre-fusion refers to first performing image fusion on images of different

modalities and then inputting them into the object detection network. Mid-stage fusion involves performing fusion during the feature extraction stage. Currently, many deep learning-based fusion object detection methods adopt mid fusion, as it enables neural networks to fuse information from different modalities at an intermediate stage, thereby enhancing feature representation capabilities. Post-stage fusion involves obtaining detection results from multimodal images separately and then integrating the detection results from different modalities through specific strategies, which can improve detection accuracy and robustness.

2.1.1. Pre-Fusion

SuperFusion [10] introduces a bidirectional deformation field registration module to correct geometric distortions in input images and incorporates a global spatial attention mechanism to ensure that the fusion results are suitable for high-level vision tasks. UMF-CMGR [11] employs a style transfer network to convert visible images into pseudo-infrared images, utilizes a multi-level refinement registration network to estimate displacement vector fields between infrared and pseudo-infrared image pairs—enabling registration in a unimodal space—and finally adopts a dual-path interactive fusion network for feature integration. SemLA [12] embeds semantic information at multiple network stages, aligns cross-modal semantic representations via a semantic alignment module, jointly optimizes registration and semantic features, and further improves spatial matching accuracy using a semantic structure representation module. Li et al. [13] propose a fusion framework applicable to loosely aligned source images. Their method designs a dynamic feature aggregation module to model correlations between unaligned features and reassembles features via a self-attention mechanism.

2.1.2. Mid-Fusion

Mid-fusion (feature-level fusion) is one of the core methods for processing multimodal images, with its core idea being how to solve the key problem of feature misalignment between modalities. Mainstream methods can be divided into supervised alignment methods, adaptive/weakly supervised alignment methods, and attention/Transformer-based alignment methods.

Supervised alignment methods: AR-CNN [14] explicitly learns cross-modal object position offsets through supervised learning on the re-paired KAIST [15] dataset. Similarly, multimodal RPN [16] performs dual bounding box regression using RPN and a detection head, and proposes a mini-batch sampling strategy that combines dual-modal IoU. However, such methods rely on precisely paired training labels, which are costly to annotate, and are typically built on computationally intensive two-stage detection frameworks, resulting in limited speed.

Adaptive/weakly supervised alignment methods: DCNet [17] targets robust goal detection (RGBT) by modeling modality-relatedness through spatial affine transformations, feature affine transformations, and dynamic convolutions, achieving weak alignment. ADCNet [18] focuses on unaligned visible-infrared detection, incorporating spatial difference calibration (achieved through adaptive affine transformations for spatial alignment) and domain difference calibration (refining object/background features across different modalities) to enhance the distinguishability of fused features. SACNet [19] uses asymmetric correlation modules to model correlations and performs feature alignment via deformable convolution sampling. CMA-Det [20] designs an alignment network to estimate deformation fields for feature correction and combines Transformer target search to enhance robustness.

Attention/Transformer-based alignment methods: ICAFusion [21] utilizes query-guided cross-attention mechanisms to effectively enhance the distinguishability of object features in misaligned scenarios. C2Former [22] leverages the Transformer's powerful relational modeling capabilities to address misalignment and fusion issues, and reduces feature map size through adaptive sampling with offset predictions to alleviate computational burden.

It can be observed that technological innovations in the field of feature-level fusion primarily focus on efficient, weakly supervised, or adaptive modal alignment strategies to overcome strict reliance on paired labels and potential efficiency bottlenecks.

2.1.3. Post-Fusion

Decision-level fusion is a commonly used multimodal integration technique. Its basic idea is to fuse the results of multiple-modal detectors and use non-maximum suppression (NMS) [23] to eliminate redundant and overlapping detection boxes. However, this simple threshold screening method has limitations; it often fails to effectively utilize useful information contained in modalities with lower detection scores. To overcome the shortcomings of traditional methods and enhance model robustness, recent research has shifted toward utilizing more advanced mathematical theories for probabilistic inference. For example, methods such as ProbEn3 [24] employ Bayesian rules and introduce the basic assumption of conditional independence between modalities. The advantage of this approach is that it can typically fuse information more effectively, even when some modalities are missing or not fully independent in practice, thereby improving the final detection performance.

Decision-level fusion approaches incur certain computational costs. This is because they typically require multiple images (or modal data) to be input into independent detection models for processing before fusion can be performed. Therefore, compared to feature-level fusion, which directly fuses information at the feature level, decision-level fusion generally requires more computational resources and longer inference times. In uncalibrated dual-modality drone small-object detection tasks, due to the small size of the targets and significant differences in views between modalities (e.g., RGB and infrared), the bounding box positions of the two modalities may exhibit significant deviations, making it difficult to form effective overlapping regions. This limits the detection performance of decision-level fusion methods that heavily rely on the overlap relationship between predicted boxes (e.g., fusion based on NMS).

2.2. Drone-to-Drone Detection

With the maturation of ground-to-air drone detection technology and the growing demand for practical applications, air-to-air target detection has gradually become a research hotspot. Saribas et al. [25] proposed a drone detection and tracking algorithm that combines the appearance-based detector YOLOv3 [26] with a kernel correlation filter-based tracker. They use the detector to mitigate the tracker's shortcomings; whenever the tracker loses the target, the detector serves as a localization initializer and self-correction mechanism. Ashraf et al. [27] proposed using a segmentation-based method to extract spatiotemporal attention information from UAVs, addressing the poor performance of feature aggregation methods due to the extremely small size and complex motion of UAV targets. The Intelligent Unmanned Systems Laboratory at West Lake University proposed a new dataset, Det-Fly [28], for air-to-air drone target detection. They evaluated eight representative deep learning algorithms based on this dataset and analyzed the impact of environmental background, target scale, viewpoint, and other challenging conditions on detection performance. Additionally, their team proposed a global-local drone detector [29] that can fuse motion and appearance features for drone detection under challenging conditions. First, a global detector is used to

search for drone targets, then it switches to a local detector operating in an adaptive search region to improve accuracy and efficiency. They also created a new dataset, ARD-MAV, to train and validate the effectiveness of the proposed detector. Lyu et al. [30] utilized the differences in motion trajectories between aerial objects and the background to locate potential objects. They established pixel correspondences between adjacent frames based on the local similarity of spatial feature vectors, enabling motion modeling to obtain local similarity, thereby determining the motion consistency between target pixels and their surrounding pixels. This approach led to the development of a simple and effective onboard micro-object detection method. Some researchers [31–33] achieved a balance between detection accuracy and efficiency by modifying the network structure of the YOLO series algorithms, using lightweight convolutional modules and attention mechanisms. Other studies not only modified the network structure but also altered the loss function to improve detection accuracy, such as Cheng et al. [34]'s YOLOv5s-ngn and Zuo et al. [35]'s UAV-STD.

2.3. Drone-to-Drone Datasets

In the field of deep learning, the quality of data directly determines the upper limit of model performance. The diversity, representativeness, and balance of data not only affect the recognition accuracy of the model but also significantly determine the algorithm's generalization ability in complex scenarios. Currently, in the drone air-to-air target detection scenario, data collection faces numerous challenges such as dynamic blurring and background interference. Many researchers have released datasets, which facilitates other researchers in conducting horizontal comparisons of algorithm performance on these datasets, thereby promoting the development of this field.

The FL dataset [36] includes 14 video sequences with a total of 38,948 frames. The dataset targets distant small-sized UAVs, which operate under complex lighting conditions and variable environments, often blending into the background and making them difficult to distinguish. The minimum and average sizes of the drone targets are 9 × 9 and 25.5 × 16.4, respectively, with image resolutions of 640 × 480 and 752 × 480. This dataset only includes grayscale images (single-channel), lacking color (RGB) information, which limits the algorithm's ability to utilize color features. It is suitable for studying how to detect fast-moving targets.

The NPS dataset [25] includes 50 videos, totaling 70,250 frames of color images, captured by a camera mounted on a drone flying at high speed, targeting three drone objects. The resolution is 1920 × 1080 or 1280 × 760, with object sizes ranging from 10 × 8 to 65 × 21, and an average size of 16.2 × 11.6. The challenges of this dataset lie in the irregular motion of the drone targets, their small size (often less than 0.1% of the image area), occlusions, and dynamic background interference (such as clouds and buildings).

The Det-Fly dataset [28] is an airborne visual multi-scale target image dataset obtained by mounting a camera on a DJI M210 drone (SZ DJI Technology Co., Ltd., Shenzhen, China) to capture multi-scene images of another DJI Mavic2 drone (SZ DJI Technology Co., Ltd., Shenzhen, China). The dataset comprises over 13,000 images, each with a resolution of 3840 × 2160. The dataset covers three drone attitudes—downward view, level view, and upward view—and includes various scenarios such as urban areas, skies, and complex backgrounds, as well as data under different lighting conditions (strong light, weak light). It also includes motion blur and partially occluded scenarios to simulate real-world detection conditions. However, it only includes the DJI Mavic2, limiting the model's generalization ability to other types of drones, and does not involve multi-drone interactions or high-density cluster scenarios.

The AOT (https://www.aicrowd.com/challenges/airborne-object-tracking-challenge, accessed on 11 August 2025) dataset is an onboard visual object detection and tracking

dataset for drone autonomous obstacle avoidance, with a total duration of 164 h (approximately 5.9 million images) and a resolution of 2448 × 2048. It is one of the largest and most densely annotated datasets in the field of aerial vision. The annotated object sizes range from 4 to 1000 pixels, and the dataset includes drone roll angles (up to 60°), pitch angle (altitude 24–1600 m), and dynamic turning scenarios, covering complex flight states. It encompasses weather and lighting conditions such as sunny, cloudy, overcast, backlit, and overexposed (5% with strong light interference). The image backgrounds include diverse terrains such as plains, mountains, and coastlines, with 80% of targets located above the horizon to minimize ground interference. However, the dataset only includes aircraft and helicopters, lacking other drone types (such as multirotors), which affects cross-category generalization capabilities. Additionally, small targets (<100 pixels) are annotated using approximate circles, which may introduce positioning errors.

The ARD-MAV dataset [29] uses DJI Mavic 2 Pro (SZ DJI Technology Co., Ltd., Shenzhen, China) and M300 (SZ DJI Technology Co., Ltd., Shenzhen, China) drones for low-altitude and medium-altitude filming, including 60 video sequences totaling 106,665 frames, with the smallest target being 6 × 3 pixels and the largest 136 × 75 pixels. The average object size is only 0.02% of the image size (1920 × 1080), significantly smaller than other mainstream datasets (e.g., 0.05% in NPS). The dataset features diverse shooting environments, including urban, natural, and aquatic settings, as well as real-world challenges such as lighting changes and weather conditions. This dataset is primarily designed for tasks involving small targets, dynamic backgrounds, and low computational resources, with a focus on dynamic interference scenarios such as camera movement (drones equipped with cameras may experience sudden or intense movement), target movement (blurring caused by rapid drone movement or hovering), and is suitable for evaluating algorithm performance in complex air-to-air environments, as well as providing benchmark support for real-world deployments such as real-time detection on aircraft.

The UAVfly dataset [34] was collected using a DJI AIR2s drone (SZ DJI Technology Co., Ltd., Shenzhen, China) for air-to-air capture, containing 10,281 images with a resolution of 1280 × 720 pixels. Samples with target sizes smaller than 10% of the image account for 30.2% of the dataset. In terms of environmental background distribution, the dataset comprehensively covers diverse geographical environments, with relatively balanced distribution across various environments. From the perspective of viewpoint distribution, the dataset features relatively balanced front, top, and bottom viewpoints, each accounting for approximately 33%. Additionally, data collection includes three time periods—morning, noon, and evening—with data volumes across these periods being nearly equal.

The MOT-FLY [37] dataset comprises a total of 11,186 images, all with a resolution of 1920 × 1080 pixels. In terms of target size, approximately 90% of the drone targets occupy less than 5% of the image area, 32.26% of instances have a width of less than 20 pixels, and each image contains 1–3 micro-drones, including the DJI Phantom 4 (SZ DJI Technology Co., Ltd., Shenzhen, China) and two custom-made drones, focusing on air-to-air multi-drone target tracking. The dataset covers diverse scenarios, backgrounds, drone motion states, and relative distances, encompassing four distinct environmental settings: sky (11.36%), plain (12.26%), village (35.78%), and urban (40.58%). MOT-FLY focuses on multi-UAV tracking applications in the civilian sector, with a significant proportion of images set in rural and urban backgrounds. Additionally, the dataset considers different viewpoints, including top-down (37.82%), front-view (32.76%), and bottom-view (29.43%), as well as varying lighting conditions (daytime, nighttime, and twilight).

The UAV2UAV-dataset [38] (hereinafter referred to as U2U) is an airborne visual single-target tracking dataset comprising 54 videos and 25,000 images. The U2U dataset targets fixed-wing drones, including standard horizontal flight, steep turns in a top-down view,

and acrobatic maneuvers such as vertical climbs and lateral rolls. The ground background includes relatively monotonous wastelands and grasslands, as well as urban buildings, sports fields, and other complex scenes.

To facilitate a comparative analysis of the characteristics of different datasets, Table 1 systematically summarizes them in five dimensions: data scale, target type, data type, modality, and characteristics. The analysis indicates that current aerial visual drone target datasets are primarily based on visible light or single-channel grayscale images, exhibiting significant scene limitations: These datasets generally lack samples under extremely low-light conditions (such as completely dark nighttime scenes), making it difficult to effectively evaluate algorithm performance in low-light environments. However, in practical countermeasure applications, it is necessary to achieve all-weather, round-the-clock drone intrusion detection capabilities, particularly reliable identification under nighttime darkness or extremely low-light conditions.

Table 1. Comparison of UAV datasets.

Name	Data Scale	Target Type	Data Type	Modality	Characteristics
FL	14 videos, 38,948 images	Multi-object	Video	Single-channel grayscale	Lacks color info, limits color feature use. Ideal for fast target detection/tracking
NPS	50 videos, 70,250 images	Multi-object	Video	Visible light	Irregular motion, small size, occlusion, dynamic background
Det-Fly	13,000 images	Single-object	Image	Visible light	Diverse perspectives. Limited to one drone type, reduces generalization
AOT	164 h video, 5.9 M images	Multi-object	Video	Single-channel grayscale	Largest aviation vision dataset. Contains depth information for depth estimation
ARD-MAV	60 videos, 106k+ images	Multi-object	Video	Visible light	Complex scenarios: background clutter, occlusion, small targets
UAVfly	10,281 images	Single-object	Image	Visible light	Diverse geography. One drone type only, large target size
MOT-FLY	16 videos, 11k+ images	Multi-object	Video	Visible light	Civil multi-UAV tracking focus. Significant rural/urban coverage
U2U	54 videos, 25k images	Single-object	Video	Visible light	Fixed-wing UAVs. Includes level flight, steep dives, maneuvers

3. The Proposed Dataset

A temporal infrared-visible light dataset refers to a dataset that includes temporal sequence information and two different modal data types (e.g., visible light + thermal imaging, video + audio, visible light + depth, etc.). Such datasets hold significant value in fields such as computer vision, autonomous driving, robot perception, and multimodal learning. For example, VT-VOD50 [39] is the first benchmark dataset for RGB-T video object detection. Its visible light and infrared thermal imaging images are synchronously captured using two non-overlapping cameras and manually aligned. In the specific field of drone target detection based on onboard vision, the introduction of infrared modality

images can effectively compensate for the shortcomings of visible light images in low-light or adverse weather conditions, while visible light images can provide a strong complement to infrared images in capturing details of small targets.

Currently, there are no temporal misaligned rgb-thermal dataset for drone-to-drone target detection. This data gap severely limits the application effectiveness of drone-to-drone target detection algorithms in nighttime and complex environments, necessitating the construction of a comprehensive dataset incorporating multispectral (e.g., thermal imaging) and low-light scenarios. To address this, this paper constructs a temporal misaligned rgb-thermal dataset for drone-to-drone target detection named TMRGBT-D2D, providing a crucial data foundation and conditions to fully leverage the advantages of temporal and dual-modal data.

3.1. Data Collection and Annotations

Data collection. The processed dataset contains 19 RGB-T video pairs, each consisting of one RGB video and one infrared video, totaling 42,624 frames. We used the professional DJI Mavic 3T (SZ DJI Technology Co., Ltd., Shenzhen, China) drone as the data collection platform. The DJI Mavic 3T, as an industry-grade drone integrating visible light and thermal imaging dual cameras, has technical characteristics highly compatible with RGB-T video target detection tasks, particularly demonstrating significant advantages in complex scene perception and multimodal data fusion. The integrated visual imaging system comprises two complementary RGB cameras. The wide-angle camera utilizes a 1/2-inch CMOS sensor (model not specified by manufacturer) with 48 effective megapixels. It features a fixed focal length of 24 mm (equivalent) and offers an 84° field of view (FOV). The telephoto camera is equipped with a 1/2-inch CMOS sensor providing 12 effective megapixels, a 162 mm equivalent focal length, and a narrower 15° FOV. This dual-camera configuration facilitates simultaneous multi-scale geospatial data acquisition. In terms of thermal camera, the key payload is an uncooled VOx microbolometer thermal camera with a resolution of 640 × 512 pixels. It operates in the 8–14 μm long-wave infrared (LWIR) spectrum and boasts a thermal sensitivity (NETD) of <50 mK at F1.0. With a 40 mm equivalent focal length and a 61° diagonal field of view. The DJI Mavic 3T can directly capture time-aligned RGBT video streams.

During capture, the visible light video resolution is set to 1920 × 1080 pixels, and the infrared video resolution is set to 640 × 512 pixels. Both the visible light and thermal infrared cameras have a frame rate of 30. The drone targets used for capture are DJI Avata 2 (SZ DJI Technology Co., Ltd., Shenzhen, China), DJI Mini 3 Pro (SZ DJI Technology Co., Ltd., Shenzhen, China), and DJI Mini 4 Pro, (SZ DJI Technology Co., Ltd., Shenzhen, China) as shown in Figure 1. The DJI Avata2 drone is predominantly black in color and relatively small in size, while the DJI Mini3 Pro and DJI Mini4 Pro are predominantly white in color with slight differences in appearance.



Figure 1. The drones used in this article.

We classified the collected video segments based on multiple dimensions, including lighting conditions (daytime, nighttime), platform status (stationary, moving), target perspective (different angles of the drone target), target movement status (stationary, moving), and background environment (sky, forest, buildings, construction sites, playgrounds, roads, etc.), to obtain video segments categorized by different scenarios. Concurrently, the classified segments were further divided into multiple sub-segments to facilitate subsequent high-quality annotation using multi-person annotation and cross-validation methods.

Dataset annotation. The temporal misaligned rgb-thermal dataset constructed in this paper was annotated using Darklabel (https://github.com/darkpgmr/DarkLabel, accessed on 11 August 2025) video annotation software. It is worth noting that, except for annotations that could not be identified under extreme conditions, RGBT annotations are one-to-one corresponding.

Firstly, the annotators conducted a standardized annotation process and training, clearly defining how the bounding boxes of the drone targets should be delineated (for example, to enclose all rotors and the main body as closely as possible), and providing a large number of positive and negative examples. All annotators received unified training to ensure that they fully understood and followed the same set of standards. At the same time, a multi-person annotation and cross-validation mechanism was adopted. To maximize the annotation quality, we used the "multiple independent annotations—cross-validation" process. Each video sequence would be independently annotated by at least two different annotators. After completion, a verification person would conduct cross-validation. The verifier would check each frame image's annotations for both modalities one by one to ensure they were in line with the norms and corresponded to each other. For conflicting annotations, the verifier would mark them out and have the entire annotation team discuss them, ultimately reaching a consensus.

For the quantitative assessment of annotation consistency, we defined cross-modal bounding box matching degree (IoU). To evaluate the alignment quality of annotations between the visible light and infrared modalities, we calculated the intersection over union (IoU) of all "one-to-one corresponding" annotation pairs. Despite the differences in perspective and imaging, the average IoU of these paired boxes reached above 0.75. This fully demonstrates the high consistency of our cross-modal annotations.

In the process of handling extreme cases, as mentioned in the paper, for frames where the target could not be clearly identified due to extreme conditions (such as severe motion blur, thermal cross-effect), we followed the principle of "better not to annotate than to annotate wrongly" and did not annotate them. The judgment criteria for such situations were clearly specified in the annotation guidelines to ensure the traceability and consistency of all annotation decisions.

3.2. Dataset Properties and Statistics

The number of annotations of the visible light images in this paper is higher than that of the thermal images. This is because the targets in thermal images are smaller and affected by thermal crossover phenomena. Our dataset covers various lighting conditions (i.e., high-light scenes captured during the day and medium-light/low-light scenes captured at night), with nighttime scenes accounting for 38.8% of all data. Thermal images can provide additional supplementary information under low-light and visually inaccessible conditions. The sequences were acquired in six different scenarios (sky, forest, buildings, construction sites, playgrounds, roads, etc.), across different seasons and locations. In summary, we provide the first temporal misaligned rgb-thermal dataset for drone-to-drone target detection, featuring diverse scenarios and high-quality annotations, which facilitates the development of RGBT image fusion and drone target detection.

Small-scale objects. Following the general scale classification rules [40], we further divide small-scale objects into three levels: extremely small $\in [1^2, 8^2)$, tiny $\in [8^2, 16^2)$, and small $\in [16^2, 32^2)$.

The total number of RGB images in this dataset is 21,312, with each image containing 1–3 drone targets, totaling 55,086 annotations. The distribution of target areas is shown in Figure 2. Among them, the following are noted:

- Extremely small targets (1–64 px^2): 1469.
- Tiny targets (64–256 px^2): 26,854.
- Small targets (256–1024 px^2): 23,167.
- Medium targets (1024–9216 px^2): 3594.
- Large targets ($\geq$9216 px^2): 2.

The average target area is 423.7 px^2, accounting for approximately 0.02% of the image area.

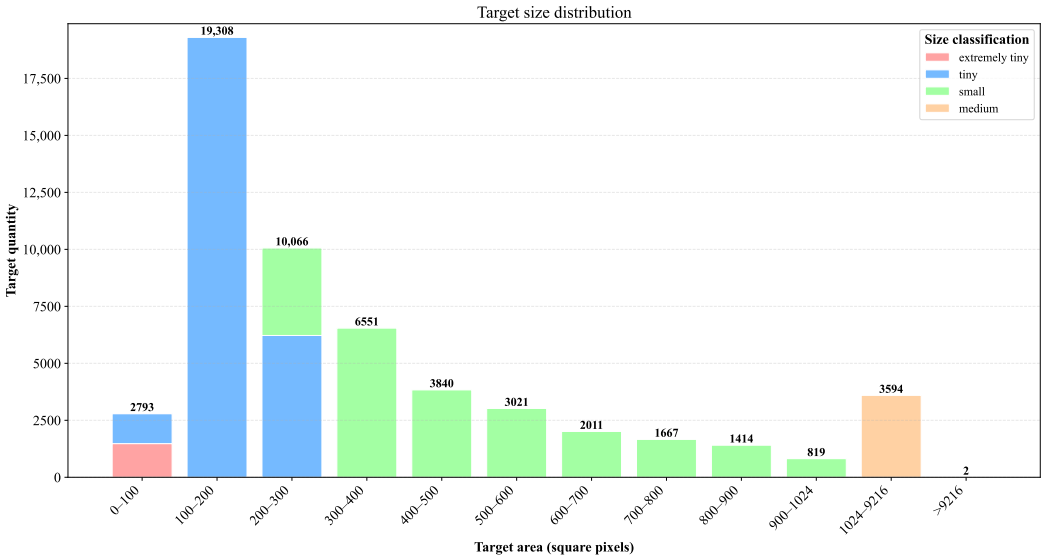


Figure 2. The distribution of target areas.

The total number of thermal images is 21,312, with a total of 39,237 annotations. Among these, the following are noted:

- Extremely small targets (1–64 px^2): 1234.
- Tiny targets (64–256 px^2): 31,538.
- Small targets (256–1024 px^2): 6059.
- Medium targets (1024–9216 px^2): 406.

The average target area for thermal images is 193.7 px^2.

It can be observed that tiny targets account for 48.7% and 80.4% of infrared and visible light images, respectively, representing the largest proportion of the dataset size distribution. Accurate detection requires consideration of comprehensive attributes such as appearance, context, density, and motion information.

Fast motion. How to accurately detect drones moving at high speeds is a key issue in drone detection tasks. When drones move at high speeds, their position on the imaging sensor shifts significantly within the single frame exposure time of the camera. This results in the target appearing not as a clear point but as a blurred strip or trajectory in a single image, as shown in Figure 3. This blurring significantly reduces the contrast and clarity of the image, causing loss of target details, decreased signal-to-noise ratio, and severe distortion of target features (edges, shape, texture), leading to reduced detector confidence, increased false negatives, or increased false positives. Both human eye recognition and

algorithmic detection become extremely challenging. Additionally, the faster the speed, the longer the exposure time, or the greater the distance (resulting in increased angular velocity of the target), the more severe the blurring effect becomes. High-speed movement not only causes blurring but may also cause the target to traverse background areas with different brightness and texture within a single frame. In some background areas, the target's contrast may be low, leading to missed detections. In this dataset, we define fast movement as situations where the drone target moves more than 10 pixels between adjacent frames. After screening, a total of 1830 frames met the criteria for fast movement, accounting for 8.59% of the total dataset.

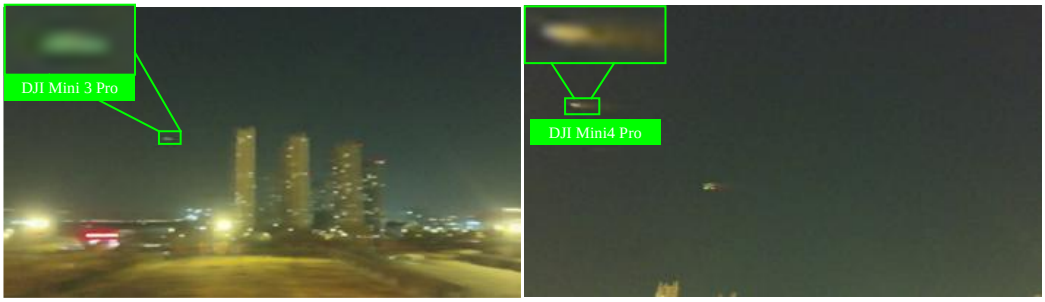
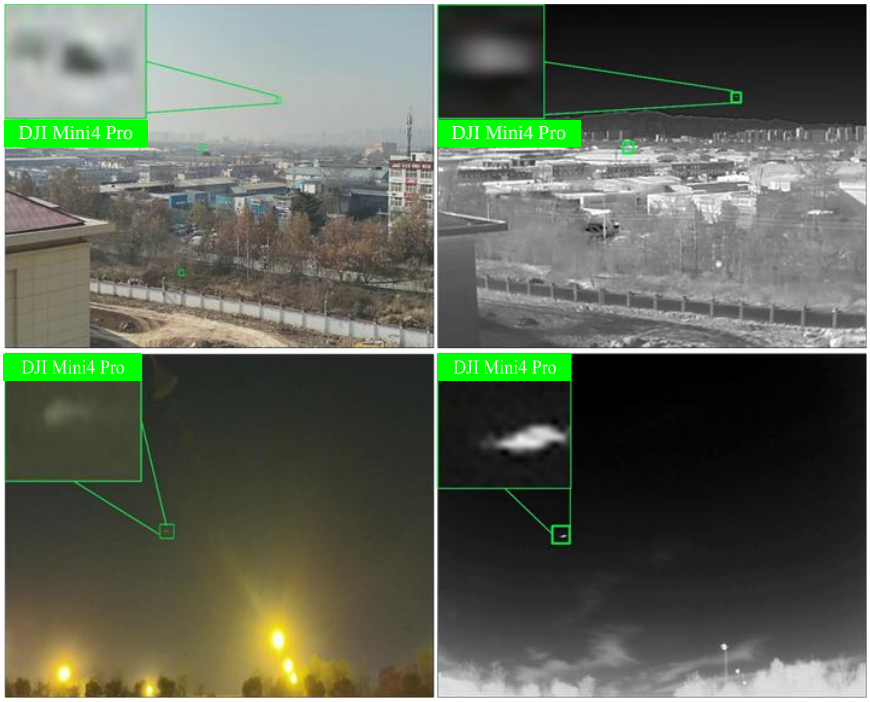


Figure 3. Motion blur caused by fast motion.

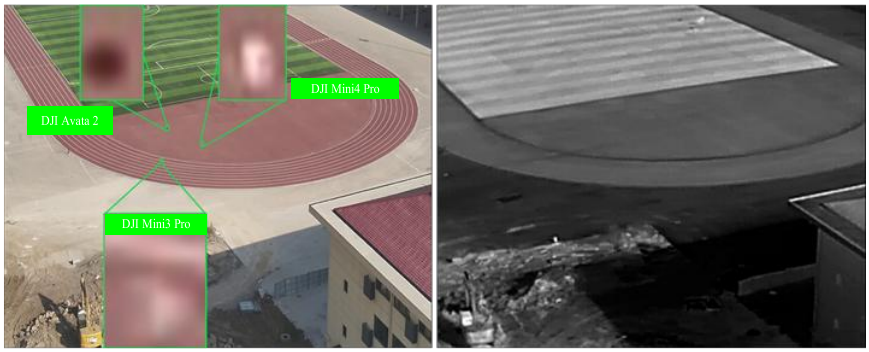
Significant differences in imaging performance of different modalities. As shown in Figure 4a, the impact on image quality varies between daytime and nighttime conditions. For the visible light modality, due to the influence of artificial lighting and other weak light sources, the quality of visible light images at night is generally less reliable or distinguishable than thermal images in simple backgrounds, especially in nighttime long-range scenarios. Unless the drone target has its own light source, it is difficult to visually discern the presence of a drone target with the naked eye. For the infrared modality, in high-temperature environments (daytime), thermal imaging produces low-quality target images. Conversely, in low-temperature environments (nighttime), thermal imaging produces high-quality images, with more pronounced temperature differences between the target and background, facilitating target identification. The takeoff time of the drone significantly affects infrared imaging performance. As shown in Figure 4b, shortly after takeoff on a playground (simple background), the drone target is not visible in infrared mode. This is because it takes time for the heat generated by components like motors to spread throughout the drone, resulting in an insignificant thermal effect at this stage, making it difficult to distinguish the thermal imaging target from the background. After the drone target has been flying for a certain period of time, its temperature increases, significantly improving the thermal imaging effect. The temperature difference between the target and the background becomes more pronounced, facilitating target detection.

Challenge in image alignment. We attempt to use projection transformation for alignment, but due to the different depth of field in this scene compared to typical remote sensing images, conventional alignment methods cannot achieve global alignment. Under the same depth of field conditions, affine transformation or projection transformation can be used for alignment. However, in the aerial visual drone target detection scenario, the depth of field of each drone target and background is inconsistent, making global alignment impossible. As shown in Figure 5, the Figure 5a,b illustrate the selection of four corresponding points in each modality image. Additionally, the target is annotated with a red box in the infrared image and a green box in the visible light image. Based on these four corresponding points, the homography matrix is calculated, and the infrared image is transformed using the homography matrix to obtain the projected infrared image, as shown in Figure 5c. Figure 5b is then overlaid with Figure 5c to obtain Figure 5d. After

transformation, it can be observed that the four corresponding points are all aligned, but the drone targets may not all be aligned: the yellow box indicates aligned drone targets, while the white box indicates drone targets that are not aligned between the two modalities.



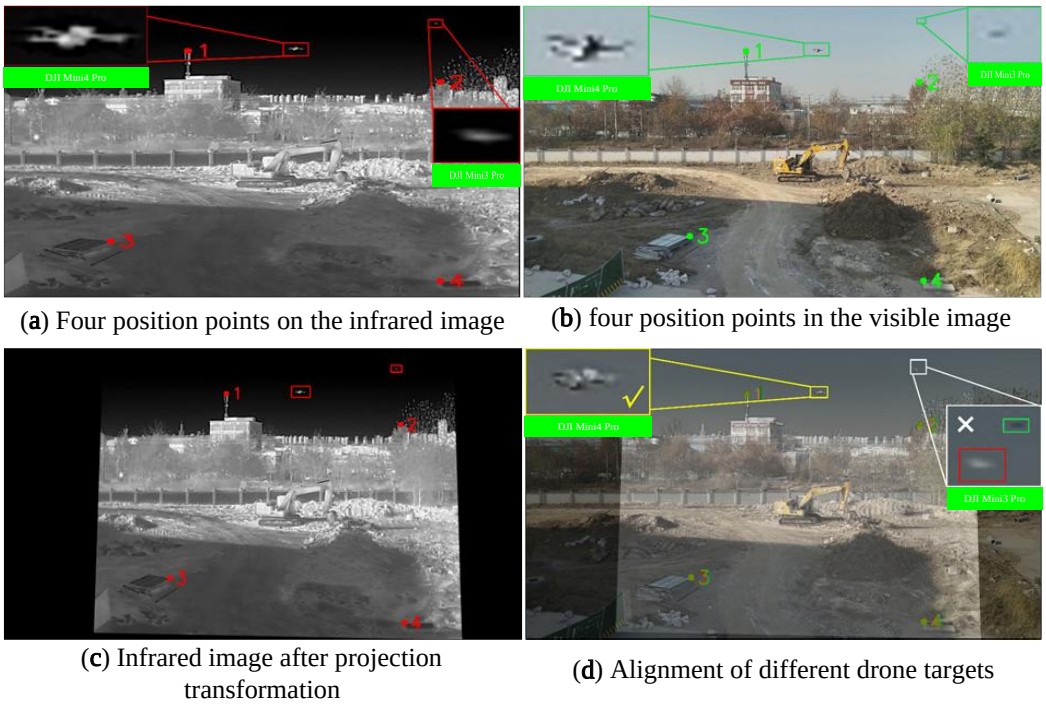
(**a**) Comparison of imaging effects between different modes during daytime and nighttime



(**b**) Effect of takeoff time on infrared imaging results

Figure 4. Differences in imaging performance of different modalities.

Given the above situation, this paper differs from the general multi-spectral dataset annotation method, which typically only annotates a single modality. This paper annotates targets in both modalities separately, and a target annotated in one modality may not necessarily be annotated in the other modality, as shown in Figure 5b. The green box indicates the annotated drone target. In the infrared modality, annotation is not performed when thermal cross-talk effects are significant and the target cannot be identified through temporal information. In the visible light modality, annotation is also not performed when brightness is insufficient to distinguish the presence of the drone target. This annotation method ensures label quality, aligning more closely with actual human visual analysis scenarios, and provides high-quality data support for subsequent algorithm research and performance evaluation.



(**a**) Four position points on the infrared image

(**b**) four position points in the visible image

(**c**) Infrared image after projection transformation

(**d**) Alignment of different drone targets

Figure 5. Challenge in image alignment.

4. Experiments

In this paper, we evaluate 10 classic deep learning-based object detection methods, including 6 general object detection methods: YOLOv5 (https://github.com/ultralytics/yolov5, accessed on 11 August 2025), YOLOv8 (http://github.com/Pertical/YOLOv8/blob/main/README.zh-CN.md, accessed on 11 August 2025), YOLOv9 [41], YOLO11 [42], RT-DETR [43], and Hyper-YOLO [44]. These methods were evaluated on the COCO dataset using mean Average Precision (mAP) as the evaluation metric, and they exhibited similar performance for small-scale objects. Additionally, we tested four RGBT detection methods (e.g., SuperYOLO [45], ICAFusion [21], ADCNet [18], and ProbEn3 [24]).

4.1. Experiments Setup

To reduce training time, our experiments were conducted on an Intel Xeon W-2245 CPU and an RTX3090 24 G GPU, using Ubuntu 20.04, PyTorch 1.11, and CUDA 11.3. In the training phase parameter settings, we set the batch size to 32, trained for 300 epochs, and used an image size of 640. Evaluation metrics include precision (P), recall (R), average precision at IoU = 0.5 (AP50), parameters (Param), and floating-point operations per second (FLOPs). In terms of dataset splitting, we simultaneously sampled infrared and visible light images, extracting one frame every five frames. From the sampled subset, we used 70% of the images for model training, 10% for validation, and the remaining 20% for testing.

4.2. Baseline Results

The results of the experiment are shown in Table 2. In this paper, we tested Transformer-based detectors, CNN-based detectors, and bimodal detection algorithms.

In Table 2, 'Visible' and 'Infrared' represent algorithms trained and tested using visible light data and infrared data, respectively, primarily targeting general-purpose detection algorithms. "Visible + Infrared" represents algorithms trained and tested using both visible light data and infrared data simultaneously. For general-purpose object detection algorithms, the testing also uses bimodal data as the test set. As shown in Table 2, YOLOv5 achieves an AP50 of 50.5% in the visible light modality, 88.2% in the infrared modality,

and 65.9% when trained with dual-modality data; YOLOv8 achieves an AP50 of 51.9% in the visible light modality and 89.1% in the infrared modality, with an AP50 of 66.7% when trained simultaneously using bimodal data; YOLOv9 achieves an AP50 of 50.7% in the visible light modality and 87.9% in the infrared modality, with an AP50 of 65.5% when trained simultaneously using bimodal data; YOLO11 performs similarly to YOLOv8, achieving AP50 values of 51.5%, 88.7%, and 66.6% when trained on visible light, infrared, and dual-modality data, respectively; Hyper-YOLO achieves an AP50 of 89.5% when using infrared modality data and an AP50 of 67.2% when using dual-modality data; RT-DETR achieves an AP50 of 90.3% in infrared modality (the best among general-purpose object detection algorithms), and an AP50 of 73.5% when using dual-modality data, but the number of parameters and computational complexity increase significantly. Among general-purpose object detection algorithms, YOLOv5 and YOLOv9 achieve AP50 of 65.9% and 65.5%, respectively, when trained with dual-modal data; YOLOv8 and YOLO11 achieve AP50 of 66.7% and 66.6%, respectively, when trained with dual-modal data. Compared to anchor-based algorithms like YOLOv5 and YOLOv9, anchor-free algorithms like YOLOv8 and YOLO11, and other anchor-free algorithms achieve higher accuracy. Small objects account for a small proportion of the image, and the anchor box mechanism tends to generate a large number of negative samples (background), causing the classifier to bias toward negative samples. Anchor-free methods (such as FCOS, CenterNet, YOLO11, YOLOv8, and RT-DETR) directly predict the center point and bounding box of the target without being constrained by fixed anchor boxes, resulting in more precise regression for small targets. In terms of accuracy, the infrared modality significantly outperforms the visible light modality. All models achieve AP50 (AP@IoU = 0.5) values in the IR modality that are far higher than those in the RGB modality (on average, approximately 30–40% higher). Under low-light conditions, the visible light modality contains limited information, while the infrared modality can obtain information such as target location and shape.

Table 2. Performance comparison of object detection models.

Model	Visible				Infrared				Visible+Infrared				Params (M)	FLOPs (G)
	P (%)	R (%)	AP50 (%)	mAP 50–95	P (%)	R (%)	AP50 (%)	mAP 50–95	P (%)	R (%)	AP50 (%)	mAP 50–95		
YOLOv5	80.3	43.0	50.5	30	95.5	80.8	88.2	60.9	89.8	57.4	65.9	41.5	2.2	5.8
YOLOv8	82.0	43.0	51.9	28.5	94.0	82.3	89.1	61.7	90.6	57.9	66.7	42.4	2.7	6.8
YOLOv9	83.0	42.8	50.7	27.2	95.7	80.6	87.9	60.1	88.5	57.8	65.5	40	1.7	6.4
YOLOv11	82.4	43.7	51.5	28.1	94.5	82.3	88.7	60.8	90.1	58.0	66.6	41.9	2.6	6.3
Hyper-YOLO	80.8	43.6	52.0	29.1	96.5	82.0	89.5	62.1	88.1	58.8	67.2	43.0	3.6	9.5
RT-DETR	85.0	52.3	60.1	30.2	97.6	86.0	90.3	68.9	92.3	65.8	73.5	48.7	42.8	130.5
SuperYOLO	–	–	–	–	–	–	–	–	95.9	72.0	76.3	40.3	7.7	14.0
ICAFusion	–	–	–	–	–	–	–	–	96.4	80.8	85.2	52.5	120.2	55.1
ADCNet	–	–	–	–	–	–	–	–	96.1	70.6	73.5	40.6	100.2	59.8
ProbEn3	–	–	–	–	–	–	–	–	–	–	82.6	45.3	180.4	70.9

SuperYOLO [45] is a method for modal fusion at the network input end and can be regarded as a pre-fusion detection algorithm, with an AP50 of 76.3%; ICAFusion [21] is a feature-level fusion detection algorithm based on Transformer design, with an AP50 of 85.2% (the best fusion performance), but with a high parameter count of 55.1 M; ADC-Net [18] is a feature-level fusion detection algorithm designed based on STN for misaligned images, with an AP50 of 73.2%; ProbEn3 [24] is a decision-level fusion detection algorithm, with an AP50 of 82.6%. RGBT detection methods can fuse dual-modal information, using infrared information to assist visible light image training. ICAFusion significantly outperforms other bimodal algorithms in terms of accuracy (96.4%), recall (80.8%), and AP50 (85.2%), validating the effectiveness of its fusion strategy. The high recall rate indicates that the model effectively combines complementary information from visible light and infrared (e.g., visible light details + infrared thermal features), enhancing detection robustness in

complex environments. The performance of all bimodal algorithms in fusion mode (Visible + Infrared) significantly outperforms general object detection algorithms, highlighting the necessity of multimodal fusion.

Figure 6 shows some undetected targets. Compared to existing drone detection datasets, TMRGBT-D2D is more challenging and difficult. The extremely small target size and significantly fewer visual cues impose severe constraints on feature representation learning, leading to high rates of missed detections and false alarms, as shown in Figure 6a. Additionally, TMRGBT-D2D includes scenarios where drone movement causes motion blur, particularly under low-light conditions, where motion blur is more pronounced, severely impacting detection performance, as shown in Figure 6b. However, rapid movement has minimal impact on infrared imaging, and how to fuse infrared and visible light features under such conditions remains an area for further research. In TMRGBT-D2D, due to the different shooting angles of drones, the background of the target drone in the forward and up perspectives is typically buildings, ground, forests, etc., as shown in Figure 6c. This results in a highly complex background for the target, and the extremely small target size further complicates the accurate detection of targets in complex backgrounds. Figure 6d illustrates the thermal cross-talk phenomenon present in the infrared images of TMRGBT-D2D. In the figure, the infrared radiation intensity of the drone target is extremely close to that of the branches in the background. The "color" or "gray value" of the region where the drone is situated is nearly identical to that of the surrounding background, rendering it visually indistinguishable from the background. As a result, the thermal features of the drone become less prominent. This phenomenon becomes more pronounced under factors such as extended distances, complex backgrounds, or atmospheric attenuation. It is one of the inherent physical limitations of infrared detection technology. Figure 6e shows that when using projection transformation for registration, due to inconsistent depth of field (DoF) among drone targets and background, the monotonic transformation can only align frames within a fixed DoF, unable to achieve global registration. Figure 6f shows that in TMRGBT-D2D, drones are typically small targets, and infrared image resolution is limited, resulting in limited target features (edges, shape, texture).



Figure 6. Examples of undetected targets.

4.3. The Impact of Temporal Information

In air-to-air drone detection tasks, single-frame imagery provides limited information, particularly under complex backgrounds and dynamic scenes, where the robustness of target detection remains to be improved.

As a key focus of this dataset, this section further explores and analyzes methods to leverage temporal information to enhance algorithm performance, emphasizing efficient utilization of video temporal cues to address the critical issue of insufficient motion feature

exploitation in video sequences. An efficient adjacent frame fusion mechanism (EAFF) [46] is introduced to verify whether temporal information can improve drone detection accuracy in air-to-air scenarios.

EAFF [46] aims to balance detection precision with computational efficiency while addressing challenges such as motion blur and small target feature loss. The core of the approach involves utilizing the motion information of target pixels between adjacent frames through a dual-module design to enhance target features: in the feature alignment and fusion module, inspired by the TAD [30], local similarity calculations establish pixel-level motion correspondences between adjacent and key frames, generating a quasi-optical flow motion field for feature alignment, and dynamically fusing target region features based on similarity-weighted metrics; in the background subtraction module, the motion field separates background regions, and dynamic background modeling suppresses static interference to highlight the true motion responses of targets. Unlike TAD, which relies solely on two-frame edge motion, this method employs a dual-cue fusion of motion and background, capable of capturing subtle appearance changes in hovering targets and improving large target localization accuracy through global feature alignment. Additionally, through lightweight local similarity computations and an end-to-end architecture, the parameter count and inference speed are significantly optimized compared to traditional video detection methods, making it more suitable for onboard platforms requiring low latency and high precision.

The core concept of this mechanism is to efficiently leverage motion information between adjacent frames and background cues to enhance key frame feature representations without incurring significant computational overhead. It primarily comprises two modules: Feature Alignment Fusion Module: this module establishes correspondences between features of neighboring frames and the key frame by computing local similarity, subsequently fusing the aligned neighboring frame features with those of the key frame; Background Subtraction Module: inspired by background subtraction techniques in motion object detection, this module aims to compute the difference between the foreground features of the key frame and the background features extracted from adjacent frames, thereby more effectively amplifying target features. The results of the experiment are shown in Table 3.

Table 3. Experiments on the ablation of the introduction of the EAFF.

Model	Visible				Infrared				Visible + Infrared				Params (M)	FLOPs (G)
	P (%)	R (%)	AP50 (%)	mAP 50–95	P (%)	R (%)	AP50 (%)	mAP 50–95	P (%)	R (%)	AP50 (%)	mAP 50–95		
YOLOv5	80.3	43.0	50.5	30	95.5	80.8	88.2	60.9	89.8	57.4	65.9	41.5	2.2	5.8
YOLOv5-EAFF	82.6	43.4	52.1	31	95.9	81.8	89.6	61.7	90.2	58.5	67.2	42.6	2.3	6.6
YOLOv8	82.0	43.0	51.9	28.5	94.0	82.3	89.1	61.7	90.6	57.9	66.7	42.4	2.7	6.8
YOLOv8-EAFF	83.4	43.6	52.7	29.1	96.1	84.4	90.3	63.1	91.2	58.9	68.9	43.6	2.7	7.6
YOLOv11	82.4	43.7	51.5	28.1	94.5	82.3	88.7	60.8	90.1	58.0	66.6	41.9	2.6	6.3
YOLOv11-EAFF	83.7	44.4	53.3	29.2	96.3	84.9	90.5	63.5	91.7	59.0	69.1	43.8	2.6	7.2

We compared YOLOv5 (YOLOv5n + EAFF), YOLOv8 (YOLOv8n + EAFF), and YOLO11 (YOLO11n + EAFF), all equipped with the EAFF mechanism, against the original single-frame algorithms on the TMRGBT-D2D dataset. Experimental results demonstrate that incorporating temporal information via the EAFF mechanism consistently enhances model performance. When trained solely on visible light modality data, YOLOv5n + EAFF, YOLOv8n + EAFF, and YOLO11n + EAFF achieved AP50 improvements of 1.6%, 0.6%, and 1.8%, respectively, and mAP50–95 increases of 1.0%, 0.6%, and 1.1%. Training exclusively on infrared modality data yielded AP50 gains of 1.4%, 1.2%, and 1.8%, and mAP50–95 improvements of 0.8%, 1.4%, and 2.7%, with YOLO11-EAFF exhibiting the most significant enhancements, +1.8% in AP50 and +2.7% in mAP50–95, indicating the strongest perfor-

mance leap among all modalities and suggesting a favorable integration of EAFF with the YOLOv11 architecture for infrared tasks. When trained on dual-modal infrared-visible data, YOLOv5n + EAFF, YOLOv8n + EAFF, and YOLO11n + EAFF showed AP50 increases of 1.3%, 2.2%, and 2.5%, and mAP50–95 improvements of 1.1%, 1.2%, and 1.9%, respectively, while maintaining comparable model parameters and computational complexity. These results demonstrate superior or comparable accuracy and speed relative to existing methods. The dual-modal fusion scenario maximizes the advantages of the EAFF mechanism, with all models exhibiting over 1 percentage point improvements in AP50 and mAP metrics. Notably, YOLO11-EAFF outperformed others in the fused-modal context, achieving the largest gains in P, AP50, and mAP, with AP50 increasing by 2.5% and mAP50–95 by 1.9%, culminating in the highest overall performance among all models (AP50: 69.1%, mAP50–95: 43.8%). These findings confirm that leveraging temporal information from adjacent frames within the sequence significantly enhances model performance on dual-modal datasets.

The introduction of this mechanism enables single-frame detectors such as YOLOv5 to utilize temporal information with minimal parameter increase and computational cost, laying a foundational basis for comparative experimental analysis.

5. Conclusions

This paper proposes the temporal misaligned rgb-thermal dataset for drone-to-drone target detection, named TMRGBT-D2D (A Temporal Misaligned RGB-Thermal Dataset for Drone-to-Drone Target Detection). Based on this dataset, we evaluated 10 representative deep learning algorithms, providing a crucial data foundation and conditions for fully leveraging the advantages of temporal and bimodal data to expand UAV target detection tasks based on onboard vision.

In future work, our goal is to further expand the dataset, improve annotation quality, and leverage temporal information to establish new algorithmic frameworks.

Author Contributions: Conceptualization, H.H., Z.Y. and Y.P.; methodology H.H., Y.P. and Z.Y.; software, H.H. and Z.Y.; validation, W.T., W.K. and Z.Y.; investigation, W.L.; resources, Z.Y. and Y.P.; data curation, Z.Y. and B.H.; writing—original draft preparation, H.H., Z.Y. and Y.P.; writing—review and editing, H.H., Z.Y., Y.P., X.Z., W.K. and Q.L.; visualization, Z.Y., X.Z., W.K. and Q.L.; supervision, Y.P.; project administration, Y.P.; funding acquisition, Y.P. All authors have read and agreed to the published version of the manuscript.

Funding: This work was supported by the Applied Research Advancement Project in Engineering University of PAP [grant no. WYY202304]; Research and Innovation Team Project in Engineering University of PAP [grant no. KYTD202306]; Graduate Student Funded Project Priorities [grant no. JYWJ2023B003].

Data Availability Statement: The data used in this analysis are publicly available, and access is provided in the text. The dataset associated with this study will be made publicly available at the following GitHub repository: https://github.com/HexiangH/TMRGBT-D2D, accessed on 11 August 2025.

Acknowledgments: The authors would like to thank all coordinators and supervisors involved and the anonymous reviewers for their detailed comments that helped to improve the quality of this article.

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

References

1. Pan, L.; Song, C.; Gan, X.; Xu, K.; Xie, Y. Military Image Captioning for Low-Altitude UAV or UGV Perspectives. *Drones* **2024**, *8*, 421. [CrossRef]
2. Ahirwar, S.; Swarnkar, R.; Bhukya, S.; Namwade, G. Application of Drone in Agriculture. *Int. J. Curr. Microbiol. Appl. Sci.* **2019**, *8*, 2500–2505. [CrossRef]
3. Raivi, A.M.; Huda, S.M.A.; Alam, M.M.; Moh, S. Drone Routing for Drone-Based Delivery Systems: A Review of Trajectory Planning, Charging, and Security. *Sensors* **2023**, *23*, 1463. [CrossRef] [PubMed]
4. Hassanalian, M.; Abdelkefi, A. Classifications, applications, and design challenges of drones: A review. *Prog. Aerosp. Sci.* **2017**, *91*, 99–131. [CrossRef]
5. Yuan, S.; Yang, Y.; Nguyen, T.H.; Nguyen, T.M.; Yang, J.; Liu, F.; Li, J.; Wang, H.; Xie, L. MMAUD: A Comprehensive Multi-Modal Anti-UAV Dataset for Modern Miniature Drone Threats. In Proceedings of the 2024 IEEE International Conference on Robotics and Automation (ICRA), Yokohama, Japan, 13–17 May 2024; IEEE: Piscataway, NJ, USA, 2024; pp. 2745–2751. [CrossRef]
6. Yang, B.; Matson, E.T.; Smith, A.H.; Dietz, J.E.; Gallagher, J.C. UAV Detection System with Multiple Acoustic Nodes Using Machine Learning Models. In Proceedings of the 2019 Third IEEE International Conference on Robotic Computing (IRC), Naples, Italy, 25–27 February 2019; pp. 493–498. [CrossRef]
7. Xiao, Y.; Zhang, X. Micro-UAV Detection and Identification Based on Radio Frequency Signature. In Proceedings of the 2019 6th International Conference on Systems and Informatics (ICSAI), Shanghai, China, 2–4 November 2019; pp. 1056–1062. [CrossRef]
8. Wang, B.; Li, Q.; Mao, Q.; Wang, J.; Chen, C.L.P.; Shangguan, A.; Zhang, H. A Survey on Vision-Based Anti Unmanned Aerial Vehicles Methods. *Drones* **2024**, *8*, 518. [CrossRef]
9. Wang, Q.; Tu, Z.; Li, C.; Tang, J. High performance RGB-Thermal Video Object Detection via hybrid fusion with progressive interaction and temporal-modal difference. *Inf. Fusion* **2025**, *114*, 102665. [CrossRef]
10. Tang, L.; Deng, Y.; Ma, Y.; Huang, J.; Ma, J. SuperFusion: A Versatile Image Registration and Fusion Network with Semantic Awareness. *IEEE/CAA J. Autom. Sin.* **2022**, *9*, 2121–2137. [CrossRef]
11. Wang, D.; Liu, J.; Fan, X.; Liu, R. Unsupervised Misaligned Infrared and Visible Image Fusion via Cross-Modality Image Generation and Registration. In Proceedings of the IJCAI, Vienna, Austria, 23–29 July 2022; pp. 3508–3515.
12. Xie, H.; Zhang, Y.; Qiu, J.; Zhai, X.; Liu, X.; Yang, Y.; Zhao, S.; Luo, Y.; Zhong, J. Semantics lead all: Towards unified image registration and fusion from a semantic perspective. *Inf. Fusion* **2023**, *98*, 101835. [CrossRef]
13. Li, H.; Liu, J.; Zhang, Y.; Liu, Y. A Deep Learning Framework for Infrared and Visible Image Fusion Without Strict Registration. *Int. J. Comput. Vis.* **2023**, *132*, 1625–1644. [CrossRef]
14. Zhang, L.; Zhu, X.; Chen, X.; Yang, X.; Lei, Z.; Liu, Z. Weakly Aligned Cross-Modal Learning for Multispectral Pedestrian Detection. In Proceedings of the 2019 IEEE/CVF International Conference on Computer Vision (ICCV), Seoul, Republic of Korea, 27 October–2 November 2019; pp. 5126–5136. [CrossRef]
15. Hwang, S.; Park, J.; Kim, N.; Choi, Y.; So Kweon, I. Multispectral Pedestrian Detection: Benchmark Dataset and Baseline. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Boston, MA, USA, 7–12 June 2015.
16. Wanchaitanawong, N.; Tanaka, M.; Shibata, T.; Okutomi, M. Multi-Modal Pedestrian Detection with Large Misalignment Based on Modal-Wise Regression and Multi-Modal IoU. In Proceedings of the 2021 17th International Conference on Machine Vision and Applications (MVA), Aichi, Japan, 25–27 July 2021; pp. 1–6. [CrossRef]
17. Tu, Z.; Li, Z.; Li, C.; Tang, J. Weakly Alignment-Free RGBT Salient Object Detection With Deep Correlation Network. *IEEE Trans. Image Process.* **2022**, *31*, 3752–3764. [CrossRef]
18. He, M.; Wu, Q.; Ngan, K.N.; Jiang, F.; Meng, F.; Xu, L. Misaligned RGB-Infrared Object Detection via Adaptive Dual-Discrepancy Calibration. *Remote Sens.* **2023**, *15*, 4887. [CrossRef]
19. Wang, K.; Lin, D.; Li, C.; Tu, Z.; Luo, B. Alignment-Free RGBT Salient Object Detection: Semantics-guided Asymmetric Correlation Network and A Unified Benchmark. *IEEE Trans. Multimed.* **2024**, *26*, 10692–10707. [CrossRef]
20. Song, K.; Xue, X.; Wen, H.; Ji, Y.; Yan, Y.; Meng, Q. Misaligned Visible-Thermal Object Detection: A Drone-Based Benchmark and Baseline. *IEEE Trans. Intell. Veh.* **2024**, *9*, 7449–7460. [CrossRef]
21. Shen, J.; Chen, Y.; Liu, Y.; Zuo, X.; Fan, H.; Yang, W. ICAFusion: Iterative cross-attention guided feature fusion for multispectral object detection. *Pattern Recognit.* **2024**, *145*, 109913. [CrossRef]
22. Yuan, M.; Wei, X. C²Former: Calibrated and Complementary Transformer for RGB-Infrared Object Detection. *IEEE Trans. Geosci. Remote Sens.* **2024**, *62*, 5403712. [CrossRef]
23. Neubeck, A.; Van Gool, L. Efficient Non-Maximum Suppression. In Proceedings of the 18th International Conference on Pattern Recognition (ICPR'06), Hong Kong, China, 20–24 August 2006; Volume 3, pp. 850–855. [CrossRef]
24. Chen, Y.T.; Shi, J.; Ye, Z.; Mertz, C.; Ramanan, D.; Kong, S. Multimodal object detection via probabilistic ensembling. In Proceedings of the Computer Vision—ECCV 2022: 17th European Conference, Tel Aviv, Israel, 23–27 October 2022; Proceedings, Part IX; Springer: Cham, Switzerland, 2022; pp. 139–158.

25. Li, J.; Ye, D.H.; Chung, T.; Kolsch, M.; Wachs, J.; Bouman, C. Multi-target detection and tracking from a single camera in Unmanned Aerial Vehicles (UAVs). In Proceedings of the 2016 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), Daejeon, Republic of Korea, 9–14 October 2016; pp. 4992–4997. [CrossRef]
26. Redmon, J.; Farhadi, A. YOLOv3: An Incremental Improvement. *arXiv* **2018**, arXiv:1804.02767. [CrossRef]
27. Ashraf, M.W.; Sultani, W.; Shah, M. Dogfight: Detecting Drones from Drones Videos. In Proceedings of the 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Nashville, TN, USA, 20–25 June 2021; pp. 7063–7072. [CrossRef]
28. Zheng, Y.; Chen, Z.; Lv, D.; Li, Z.; Lan, Z.; Zhao, S. Air-to-Air Visual Detection of Micro-UAVs: An Experimental Evaluation of Deep Learning. *IEEE Robot. Autom. Lett.* **2021**, *6*, 1020–1027. [CrossRef]
29. Guo, H.; Zheng, Y.; Zhang, Y.; Gao, Z.; Zhao, S. Global-Local MAV Detection Under Challenging Conditions Based on Appearance and Motion. *IEEE Trans. Intell. Transp. Syst.* **2024**, *25*, 12005–12017. [CrossRef]
30. Lyu, Y.; Liu, Z.; Li, H.; Guo, D.; Fu, Y. A Real-time and Lightweight Method for Tiny Airborne Object Detection. In Proceedings of the 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), Vancouver, BC, Canada, 17–24 June 2023; pp. 3016–3025. [CrossRef]
31. Zhou, X.; Yang, G.; Chen, Y.; Li, L.; Chen, B.M. VDTNet: A High-Performance Visual Network for Detecting and Tracking of Intruding Drones. *IEEE Trans. Intell. Transp. Syst.* **2024**, *25*, 9828–9839. [CrossRef]
32. Hao, H.; Peng, Y.; Ye, Z.; Han, B.; Zhang, X.; Tang, W.; Kang, W.; Li, Q. A High Performance Air-to-Air Unmanned Aerial Vehicle Target Detection Model. *Drones* **2025**, *9*, 154. [CrossRef]
33. Huang, M.; Mi, W.; Wang, Y. EDGS-YOLOv8: An Improved YOLOv8 Lightweight UAV Detection Model. *Drones* **2024**, *8*, 337. [CrossRef]
34. Cheng, Q.; Wang, Y.; He, W.; Bai, Y. Lightweight air-to-air unmanned aerial vehicle target detection model. *Sci. Rep.* **2024**, *14*, 2609. [CrossRef]
35. Zuo, G.; Zhou, K.; Wang, Q. UAV-to-UAV Small Target Detection Method Based on Deep Learning in Complex Scenes. *IEEE Sens. J.* **2025**, *25*, 3806–3820. [CrossRef]
36. Rozantsev, A.; Lepetit, V.; Fua, P. Detecting Flying Objects Using a Single Moving Camera. *IEEE Trans. Pattern Anal. Mach. Intell.* **2017**, *39*, 879–892. [CrossRef]
37. Chu, Z.; Song, T.; Jin, R.; Jiang, T. An Experimental Evaluation Based on New Air-to-Air Multi-UAV Tracking Dataset. In Proceedings of the 2023 IEEE International Conference on Unmanned Systems (ICUS), Hefei, China, 13–15 October 2023; pp. 671–676. [CrossRef]
38. Wang, Y.; Huang, Z.; Laganière, R.; Zhang, H.; Ding, L. A UAV to UAV tracking benchmark. *Knowl.-Based Syst.* **2023**, *261*, 110197. [CrossRef]
39. Tu, Z.; Wang, Q.; Wang, H.; Wang, K.; Li, C. Erasure-based Interaction Network for RGBT Video Object Detection and A Unified Benchmark. *arXiv* **2023**, arXiv:2308.01630. [CrossRef]
40. Lin, T.Y.; Maire, M.; Belongie, S.; Hays, J.; Perona, P.; Ramanan, D.; Dollár, P.; Zitnick, C.L. Microsoft COCO: Common Objects in Context. In Proceedings of the Computer Vision—ECCV 2014, Zurich, Switzerland, 6–12 September 2014; Fleet, D., Pajdla, T., Schiele, B., Tuytelaars, T., Eds.; Springer: Cham, Switzerland, 2014; pp. 740–755.
41. Wang, C.Y.; Yeh, I.H.; Mark Liao, H.Y. Yolov9: Learning what you want to learn using programmable gradient information. In Proceedings of the European Conference on Computer Vision, Milan, Italy, 29 September–4 October 2024; Springer: Cham, Switzerland, 2025; pp. 1–21.
42. Khanam, R.; Hussain, M. YOLOv11: An Overview of the Key Architectural Enhancements. *arXiv* **2024**, arXiv:2410.17725. [CrossRef]
43. Zhao, Y.; Lv, W.; Xu, S.; Wei, J.; Wang, G.; Dang, Q.; Liu, Y.; Chen, J. DETRs Beat YOLOs on Real-time Object Detection. In Proceedings of the 2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Seattle WA, USA, 17–21 June 2024; pp. 16965–16974. [CrossRef]
44. Feng, Y.; Huang, J.; Du, S.; Ying, S.; Yong, J.H.; Li, Y.; Ding, G.; Ji, R.; Gao, Y. Hyper-YOLO: When Visual Object Detection Meets Hypergraph Computation. *IEEE Trans. Pattern Anal. Mach. Intell.* **2025**, *47*, 2388–2401. [CrossRef] [PubMed]
45. Zhang, J.; Lei, J.; Xie, W.; Fang, Z.; Li, Y.; Du, Q. SuperYOLO: Super Resolution Assisted Object Detection in Multimodal Remote Sensing Imagery. *IEEE Trans. Geosci. Remote Sens.* **2023**, *61*, 5605415. [CrossRef]
46. Ye, Z.; Peng, Y.; Liu, W.; Yin, W.; Hao, H.; Han, B.; Zhu, Y.; Xiao, D. An Efficient Adjacent Frame Fusion Mechanism for Airborne Visual Object Detection. *Drones* **2024**, *8*, 144. [CrossRef]

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.