remote sensing

MDPI

Article

Seeing Through Sparse Foliage: Quality–Occlusion-Guided RGB–Thermal Fusion for Drone-Based Person Detection

Ziming Gui [1,†], Shaobo Liu [2,†], Dong Yang [2], Tongyuan Zou [2], Haoran Zhu [1] and Wen Yang [1,*]

[1] School of Electronic Information, Wuhan University, Wuhan 430072, China; guiziming@whu.edu.cn (Z.G.); zhuhaoran@whu.edu.cn (H.Z.)
[2] Space Star Technology Co., Ltd. (SSTC), China Academy of Space Technology, Beijing 100095, China; liu_shaobo@163.com (S.L.); topyangdong@163.com (D.Y.); zouty_cast@126.com (T.Z.)
* Correspondence: yangwen@whu.edu.cn
† These authors contributed equally to this work.

Highlights

What are the main findings?

- Physics-Driven Occlusion Taxonomy: We formally define "soft" and "hard" occlusion for drone-based detection, leveraging cross-spectral physical consistency to filter feature contamination caused by sparse foliage.
- Decoupled Quality–Occlusion Arbitration: A novel fusion mechanism is proposed that decouples global modality reliability (QAF) from localized spatial occlusion (OQF), enabling dynamic, pixel-wise multi-modal feature integration.

What are the implication of the main finding?

- Robust Search and Rescue: The UOT framework significantly enhances detection reliability in complex aerial environments, particularly recovering critical semantic cues for occluded tiny pedestrians in search and rescue missions.
- Theoretical Scalability: By utilizing cross-modal physical consistency as a noise filter, our approach offers a scalable solution for multi-spectral fusion that addresses modality-specific sensor degradation.

Abstract

Drone-based RGBT person detection facilitates critical applications such as search and rescue, owing to its high maneuverability and inherent capability to mitigate visual occlusion. However, despite the complementary nature of RGBT systems, existing detectors often overlook the specific impact of occlusion during the fusion process, leading to feature contamination and subsequent detection failures. In this work, we address this limitation by formally defining two categories of occlusion: "soft occlusion," where targets remain partially visible in at least one modality, and "hard occlusion," which involves complete obstruction. To tackle these challenges, we propose Unveiling Occluded Targets (UOT), a novel multi-modal fusion framework that implements a Quality–Occlusion Arbitration (QOA) mechanism. By leveraging both quality-related and occlusion-related cues, UOT dynamically arbitrates the fusion process to maximize information recovery from the clearer modality. Extensive experiments on the RGBTDronePerson and VTUAV-det datasets demonstrate significant improvements, achieving an mAP_{50}^{all} of 53.42% and an mAP_{50}^{tiny} of 54.70% in densely occluded scenes. Qualitative analysis further confirms UOT's robustness in reliably identifying targets obstructed by sparse foliage.

Keywords: drone image; RGB-T fusion; occlusion problem; tiny object detection

Check for updates

Academic Editor: Susana Lagüela López

Received: 27 December 2025
Revised: 14 February 2026
Accepted: 27 February 2026
Published: 4 March 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.

1. Introduction

Drone-based person detection has become a pivotal technology for critical applications such as search and rescue, disaster response, and maritime surveillance [1–3], primarily due to the high maneuverability and wide coverage of Unmanned Aerial Vehicles (UAVs). However, relying solely on single-modality sensors often limits operational reliability under adverse conditions. To ensure 24/7 reliability under varying illumination and weather conditions, integrating Visible (RGB) and Thermal (T) sensors—known as RGBT detection—has emerged as the standard paradigm. Theoretically, RGBT systems offer a robust solution: when one modality fails (e.g., RGB in darkness or heavy smoke), the other (e.g., Thermal) compensates by capturing the heat signatures of targets, thereby ensuring uninterrupted target perceptibility across diverse environments.

Despite this theoretical promise, the reality of drone-based sensing is far more complex than this idealized complementarity. Unlike ground-level surveillance with fixed, horizontal viewpoints, drones often capture images from oblique angles and high altitudes amidst intricate environments like dense forests or complex urban canyons. This unique perspective introduces a pervasive and formidable challenge: complex occlusion. As illustrated in Figure 1, we conducted a systematic investigation of the RGBTDronePerson dataset to quantify the prevalence of this issue. Our analysis reveals that occluded instances account for 42% of the total targets, underscoring that occlusion is not an isolated edge case but a dominant phenomenon in aerial surveillance. While existing RGBT detectors exhibit impressive performance when handling clear or fully visible targets, their accuracy degrades significantly when targets are obscured by natural obstacles such as dynamic foliage, interlaced branches, or artificial structures. In these scenarios, the spatial relationship between the target and the occluder varies erratically, making robust detection a persistent bottleneck in aerial RGBT perception.

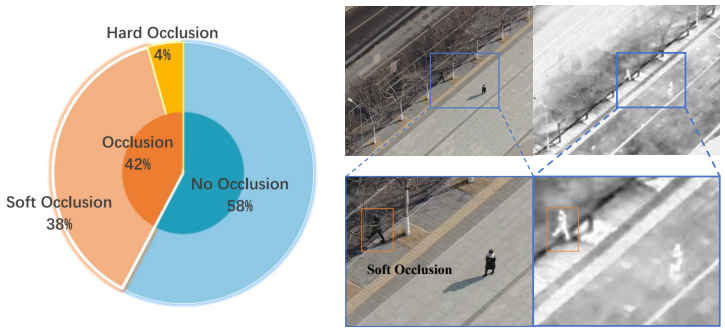


Figure 1. Figure (**a**) illustrates the distribution of various occlusion scenarios in the RGBTDronePerson dataset, while Figure (**b**) presents a typical example of soft occlusion as viewed from a drone's perspective.

Most RGBT detection methods focus on feature alignment or fusion strategy, operating under the implicit assumption that both modalities contain valid and high-quality information. Consequently, they lack an explicit "quality control" or "arbitration" mechanism. When a target suffers from what we term "soft occlusion"—where it remains partially visible in one modality (e.g., thermal signature visible through leaves) but is completely obscured or corrupted by noise in the other (e.g., RGB blocked by dense foliage)—blind fusion strategies inadvertently integrate the noise or irrelevant background features from the occluded modality into the final representation. This phenomenon, which we identify as "feature contamination," dilutes the discriminative power of the clear modality, leading to frequent missed detections or environmental hallucinations.

In this work, we argue that effective RGBT fusion in drone scenarios requires more than just feature aggregation; it demands occlusion-aware arbitration. As summarized in Figure 2, we formally categorize occlusion into two distinct types to guide this process: (1) Soft Occlusion, where at least one modality retains partial discriminative cues, and (2) Hard Occlusion, where the target is physically blocked across both spectra. We posit that while hard occlusion is an information-loss problem, soft occlusion is a fusion-weighting problem that can be resolved by intelligently suppressing the contaminated modality. To this end, we propose Unveiling Occluded Targets (UOT), a novel quality–occlusion arbitrated fusion framework. Unlike existing black-box methods, UOT introduces a Pre-Discrimination Module (PDM) that explicitly estimates two critical priors: a Quality-Aware Factor (QAF) for texture richness and an Occlusion–Quality Factor (OQF) for obstruction confidence. These factors act as a "gatekeeper" for our Occlusion Fusion Module (OFM), which dynamically re-weights features at a pixel-level granularity. This ensures that the network selectively "trusts" the clear modality while actively "ignoring" the occluded regions, thereby maximizing information recovery without introducing noise.

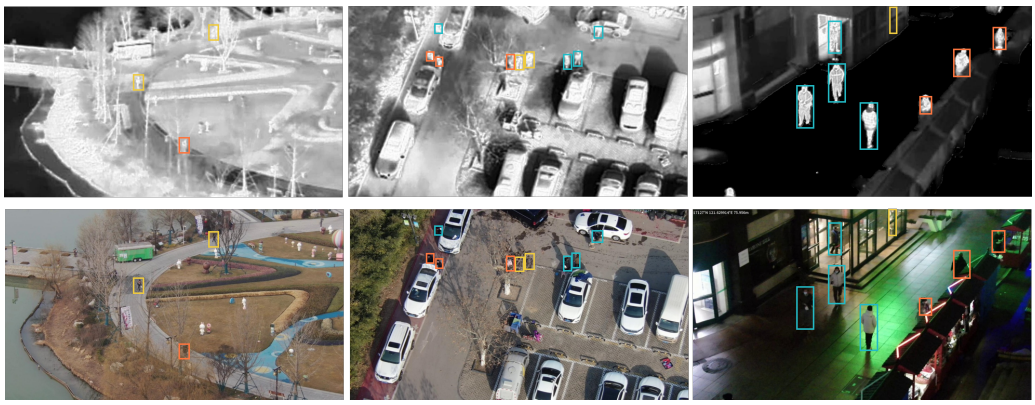


Figure 2. Definition and examples of different occlusion types. The categories are color-coded for clarity: cyan frames indicate non-occluded targets, yellow frames indicate soft-occluded targets (partially visible in at least one modality), and orange frames indicate hard-occluded targets (completely obstructed in both modalities).

In summary, this paper has the following contributions:

- For the first time, we identify the unique occlusion challenge in RGBT imagery from a drone's perspective, introducing the concepts of hard occlusion and soft occlusion to effectively address different occlusion scenarios.
- To systematically address these identified limitations, we propose the Pre-Discrimination Module (PDM) and the Occlusion Fusion Module (OFM) as a robust solution for adaptive multi-modal arbitration.
- We conducted extensive experiments on the RGBT datasets VTUAV-det and RGBT-DronePerson from a drone's perspective. The results demonstrate that our proposed method effectively mitigates the impact of soft occlusion on fusion.

2. Related Work

2.1. Tiny Object Detection

Tiny object detection (TOD) remains a fundamental yet formidable challenge in computer vision, characterized by targets with extremely limited pixel coverage and low signal-to-noise ratios. As underscored in a recent comprehensive survey [4], aerial person detection (APD) in search and rescue (SaR) scenarios is particularly demanding due to varied postures, sparsity, and the diminutive size of targets from an air-to-ground perspective. Conventional methodologies have evolved from basic data augmentation, such

as Stitcher [5] and copy–paste strategies [6], to sophisticated multi-scale feature learning. Representative architectures like Feature Pyramid Networks (FPNs) [7], along with its variants ASFF [8] and Recursive FPN [9], attempt to preserve shallow high-resolution cues. However, simply aggregating multi-scale features often fails to recover the discriminative structural information of tiny objects.

To address this, recent research has shifted toward more intricate feature interaction and contextual modeling. FRLI-Net [10] introduces a feature reconstruction and learning interaction paradigm specifically for remote sensing, emphasizing the synergy between feature reconstruction and inter-layer interaction. In parallel, the emergence of end-to-end Transformers has led to UAV-DETR [11], which optimizes computational efficiency for aerial platforms. Despite these methodological advancements, most single-modality detectors encounter performance bottlenecks in extreme lighting or "soft occlusion" conditions. This necessitates an arbitration-based fusion strategy that can leverage multi-modal complementarity to maintain detection stability when visual information is severely degraded.

2.2. Multi-Modal Fusion

RGBT dual-modal detection leverages the complementary strengths of visible (RGB) and thermal (T) sensors to overcome the inherent limitations of single-sensor systems under adverse conditions, such as total darkness or dense fog. The evolution of this field has transitioned from rudimentary linear fusion strategies, including element-wise summation and concatenation, to more sophisticated adaptive mechanisms that dynamically calibrate modal contributions at different feature levels. The primary goal of these methodologies is to effectively exploit the high-resolution texture information from RGB images alongside the robust thermal signatures from infrared sensors, thereby achieving a synergistic representation that surpasses the capabilities of any single modality.

Recent research has introduced various sophisticated architectures to optimize this information integration. QFDet [2] represents a pivotal work that ensures adequate supervision for tiny objects while prioritizing high-quality samples through a query-based fusion strategy. Building upon this, TCAINet [12] utilizes an adaptive decoder to refine cross-modal features. Notably, VTSaRNet [13] incorporates a Union Transformer Network to model long-range dependencies across RGB and thermal modalities, specifically addressing the challenge of sparse person detection in complex SaR scenes. In the specific context of drone-based surveillance, He et al. [14] introduced a progressive target-aware network designed for RGBT person detection, which meticulously refines features to suppress background clutter in volatile aerial scenes. The application of these fusion strategies has also extended to specialized environmental monitoring; for instance, Zhang et al. [3] developed a specialized RGB–Thermal fusion model for forest fire detection, demonstrating enhanced reliability in high-contrast and complex terrains. Additionally, Tang et al. [15] proposed a bi-directional fusion mechanism combined with a deformation-sensitive loss to address geometric variations caused by UAV maneuvers.

Despite these significant advancements, most current fusion strategies operate under the implicit assumption that both modalities contain valid information, thereby lacking an explicit occlusion-aware guidance mechanism. When a target is heavily obscured in one modality—such as an RGB target hidden by dense foliage while remaining visible in thermal—existing models often inadvertently integrate noise from the degraded channel. This leads to what we identify as "feature contamination," where the discriminative power of the clear modality is diluted by the corrupted features of the occluded one. Furthermore, while architectures like APDNet [16] prioritize efficiency through lightweight distillation networks for edge deployment, they may encounter robustness bottlenecks in extreme foliage occlusion where high-fidelity feature recovery is paramount. These limitations

demand a more granular, factor-driven approach to cross-modal arbitration to maintain detection robustness under the extreme perspective fluctuations and erratic viewing angles characteristic of UAV maneuvers.

2.3. Occlusion Detection

Occlusion remains one of the most significant impediments to robust object detection, particularly in complex aerial environments. Generally, occlusion scenarios can be categorized into two primary types: foreground-to-foreground occlusion, characterized by the overlap between multiple targets, and foreground-to-background occlusion, where targets are blocked by environmental elements such as foliage or structures.

To mitigate these challenges, various paradigms have been developed. In foreground-to-foreground scenarios, research focuses on refining Non-Maximum Suppression (NMS) [17], with variants like Soft-NMS [18] and IoU-aware adjustments [19] reducing missed detections in crowded areas. Regarding foreground-to-background occlusion, the mainstream approach involves occlusion localization combined with feature compensation. Some methodologies utilize motion vector features [20] or generative adversarial networks (GANs) [21] to compensate for obscured regions. Notably, Zhang et al. [22] proposed a multi-scale occlusion-aware network for UAV-based vehicle detection using feature alignment. OGMN [23] further identifies occlusion coordinates to guide the subsequent feature extraction process.

Notwithstanding these advancements, existing single-modality methods often rely on synthetically generating "phantom" features for occluded areas, which can introduce environmental noise. Moreover, they struggle when occlusion is pervasive across the visual spectrum. To address these limitations, our UOT framework departs from generative strategies by leveraging the intrinsic complementarity of multi-modal data. By utilizing the thermal modality to "see through" visual obstructions, our approach resolves the ambiguity of "soft occlusions" without the need for noisy feature synthesis, providing a more reliable foundation for drone-based detection in cluttered environments.

3. Methodology

3.1. Overview

The structure of our proposed UOT is shown in Figure 3. The input consists of paired images that undergo a detailed processing procedure within the Pre-Discrimination Model (PDM). This model is designed to extract two essential factors that significantly influence the image fusion process: the Quality-Aware Factor (QAF) and the Occlusion–Quality Factor (OQF). These two factors play a pivotal role in directing the Occlusion Fusion Module (OFM), which operates within a sophisticated framework based on joint optimization techniques. This framework effectively merges the strengths of both images from a bidirectional perspective, reintegrating the dominant representation of unoccluded regions with the unimodal advantages of occluded areas to yield a more coherent and comprehensive representation. Once the image fusion is completed, the resulting representation is forwarded to the detection head. Here, we follow the design of AODet [24], the Cross-level Query Selection (CQS) method is employed to sift through the fused data, ensuring that the final detection outcomes are not only robust but also precise. This multi-layered approach contributes to significantly improved performance in small object detecting and identifying objects within the processed images. The specific technical implementations and synergies between the PDM and OFM are elaborated upon in Sections 3.2 and 3.3, respectively.

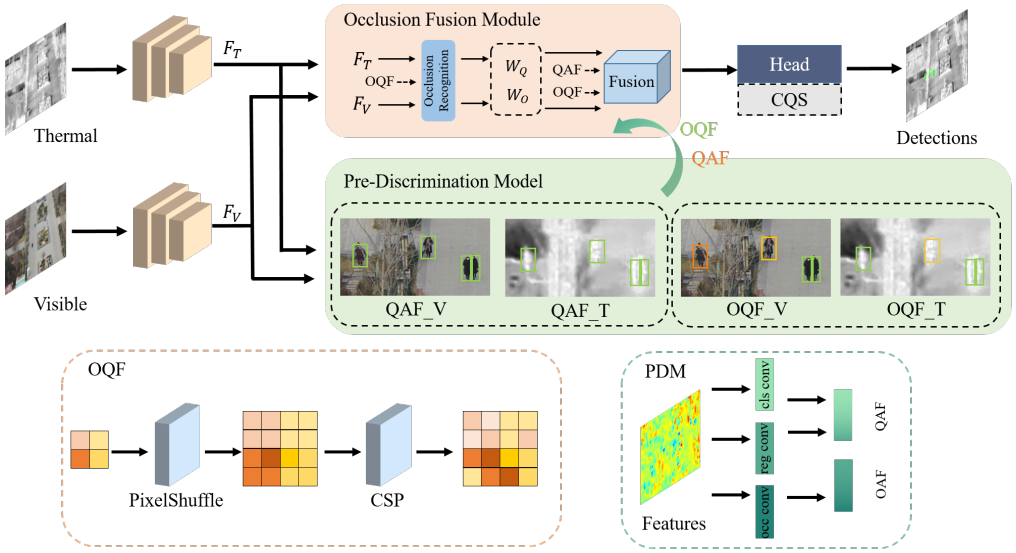


Figure 3. Overview of UOT. The input comprises paired images, which are processed by the Pre-Discrimination Model (PDM) to extract two critical factors: the Quality-Aware Factor (QAF) and the Occlusion–Quality Factor (OQF). These factors serve as pivotal components for guiding the Occlusion Fusion Module (OFM) in executing a joint optimization-based fusion framework. Specifically, the QAF encapsulates quality-aware guidance, whereas the OQF provides targeted occlusion-aware guidance. The fused representation is subsequently forwarded to the detection head, where the Cross-level Query Selection (CQS) is employed to achieve robust and accurate final detection outcomes.

3.2. Pre-Discrimination Module

The efficacy of multi-modal fusion hinges on the accurate identification of modal reliability prior to feature aggregation. While state-of-the-art methods, such as QFDet [2], have introduced quality-aware mechanisms to prioritize high-quality samples, they often operate under the assumption that target quality is predominantly dictated by global texture or illumination. Consequently, these approaches struggle to decouple target-related features from environmental noise during "soft occlusion" events. In such scenarios, a modality may exhibit rich texture richness (high quality) yet be entirely composed of occluding features (e.g., dense foliage), leading to inevitable feature contamination during fusion.

To bridge this gap, we propose the Pre-discrimination Module (PDM), which implements a dual-factor evaluation protocol to refine fusion weights at a granular level. Unlike existing strategies, PDM does not solely rely on a single quality metric; instead, it synergistically computes a Quality-Aware Factor (QAF) and a novel Occlusion–Quality Factor (OQF). Crucially, QAF and OQF characterize distinct dimensions of feature degradation—sensing reliability versus physical visibility—and are not inherently redundant. QAF represents the global modality-level reliability, primarily derived from high-level feature responses and classification confidence. It assesses whether a modality is intrinsically 'informative' (e.g., free from motion blur or sensor noise) but lacks the spatial granularity to perceive localized physical obstructions. In contrast, OQF is driven by a specialized spatial perception mechanism that evaluates extrinsic visibility at each coordinate. By decoupling these two factors, PDM prevents the model from mistakenly assigning high trust to "high-quality" background noise that lacks target-related information.

Specifically, as illustrated in Figure 4, for the input feature tensors F_v and F_t, the PDM first employs a specialized spatial perception mechanism to localize regions of interest (RoIs). Deviating from conventional spatial attention, our mechanism is designed to capture sub-pixel occlusion cues while maintaining semantic integrity. To achieve this, the dual-modal features are processed through a PixelShuffle layer. Unlike standard pooling or interpolation that discards spatial variance, PixelShuffle reorganizes channel-wise depth in-

formation into spatial coordinates to implement a sub-pixel feature reconstruction strategy. This is particularly vital for drone-captured imagery, where small targets often occupy minimal pixel areas; by magnifying the feature granularity, the model can effectively distinguish between the micro-textures of foliage and the actual silhouettes of human targets.

Subsequently, a Cross Stage Partial (CSP) architecture is utilized to aggregate multi-scale semantic signatures. The identification of RoIs within this architecture follows a Saliency-Guided Criteria: a region is flagged as an RoI if it exhibits high semantic consistency with "person-shaped" clusters in at least one modality, while the other modality shows an attenuated or conflicting signal. This hierarchical verification allows the CSP branch to filter out non-informative environmental noise while maintaining computational efficiency.

Through this specialized architecture, the PDM executes a two-stage identification criterion: (1) a coarse pre-discrimination to confirm the potential presence of person-shaped clusters; and (2) the generation of high-fidelity reliability masks. Within these localized regions, the QAF is derived from the local feature entropy, reflecting the richness of discriminative cues. Simultaneously, the OQF is computed via a CSP-based perceptual branch to estimate the spatial confidence of target occlusion at each coordinate. This dual-factor representation enables the framework to distinguish between "high-quality targets" and "high-quality occluders," ensuring that the subsequent Occlusion Fusion Module (OFM) can selectively trust the modality less affected by environmental obstructions.

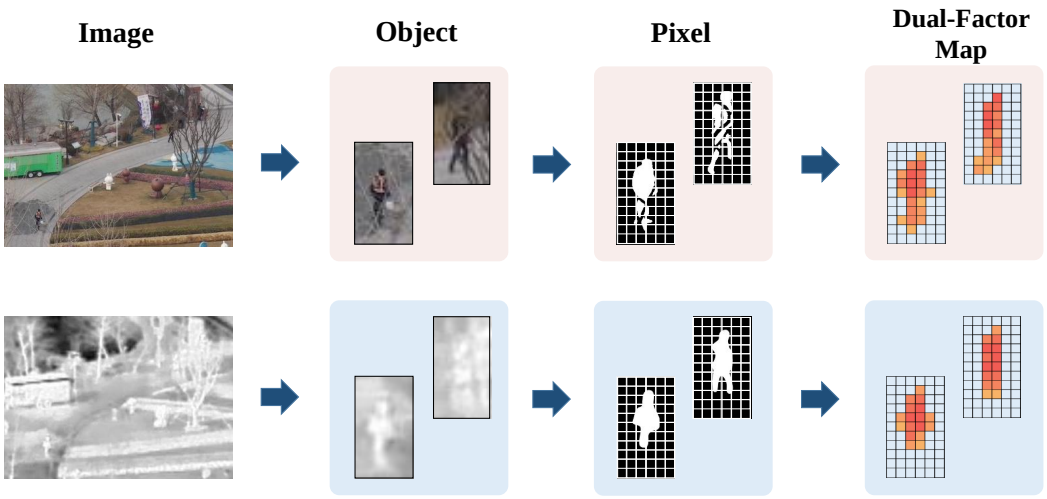


Figure 4. Our proposed process for generating the Dual-factor score map.

3.3. Occlusion Ground Truth Generation

Using the Pre-Discrimination Model (PDM), we are able to obtain precise prediction scores for both modalities independently. From the perspective of target detection, it is observed that occluded targets generally receive lower quality scores, particularly in the absence of effective anti-occlusion mechanisms. Conversely, non-occluded targets are likely to achieve higher quality scores, reflecting their clearer visibility. To enhance the supervision of the neural network, accurate ground truth data concerning occlusion is of paramount importance. However, a significant limitation arises from the fact that current UAV datasets often lack explicit annotations for occlusion.

To mitigate the potential 'confirmation bias' during initial training, we do not train the PDM from scratch blindly. Instead, we introduce a cross-modal consensus prior during a dedicated warm-up phase. For the first N epochs, the initial occlusion masks are generated based on the normalized pixel-wise discrepancy across spectra. This strategy utilizes the inherent physical consistency between RGB and thermal sensors as a noise filter:

background noise typically manifests independently in each modality, whereas true targets exhibit correlated responses. This ensures that the PDM is initialized with robust physical priors before it begins to refine complex occlusion patterns.

Unlike the OGMN [23], which relies heavily on pre-existing ground truth data to generate occlusion labels, our method capitalizes on the quality results produced by the PDM. This approach not only enables the capture of occlusion occurring between various targets but also facilitates a comprehensive understanding of occlusion dynamics between foreground and background elements. The occlusion loss is computed using binary cross-entropy loss, as follows:

$$Loss_{occ} = -\frac{1}{N}\sum_{i=1}^{N}[y_i \cdot \log(\hat{y}_i) + (1 - y_i) \cdot \log(1 - \hat{y}_i)], \tag{1}$$

where y_i represents the generated ground truth of occlusion, and $\hat{y}_i$ represents the occlusion prediction result from the PDM.

In RGBT person detection, 'soft' occluders like foliage often produce high-magnitude but irrelevant feature responses. A standard regression loss would force the model to fit these irrelevant high-frequency textures. Our Occlusion–Quality Loss, however, introduces a cross-modal consensus weighting, ensuring that gradients are back-propagated only when features exhibit cross-spectral physical consistency. This prevents the optimization process from being dominated by high-quality yet target-irrelevant background features.

3.4. Occlusion Fusion Module

To effectively address the challenges posed by occlusion, the Occlusion Fusion Module (OFM) utilizes the Occlusion–Quality Factor (OQF) and Quality-Aware Factor (QAF) generated by the PDM to dynamically arbitrate the fusion process. A higher OQF value indicates significant occlusion in a specific modality, signaling the module to assign larger weights to the complementary, clearer stream. Conversely, when occlusion is minimal, the module shifts its focus toward enhancing feature representation based on the QAF, which reflects the intrinsic signal richness.

Specifically, to ensure numerical stability and scale consistency across modalities, we apply min–max normalization to these factors. The process is formulated as:

$$\begin{aligned} QAF_i &= \frac{Q_i - \min(Q_V, Q_T)}{\max(Q_V, Q_T) - \min(Q_V, Q_T) + \epsilon}, \\ OQF_i &= \frac{O_i - \min(O_V, O_T)}{\max(O_V, O_T) - \min(O_V, O_T) + \epsilon}, \end{aligned} \tag{2}$$

where $i \in \{V, T\}$ denotes the Visible and Thermal modalities, respectively. Q_i and O_i represent the raw QAF and OQF values for modality i. The $\max(\cdot)$ and $\min(\cdot)$ operations are computed element-wise across the spatial dimensions of the feature maps. The parameter ϵ is a stability constant, fixed at 1×10^{-5} in our implementation, to preclude numerical singularity (division-by-zero).

The normalized factors act as a dynamic gating mechanism. The enhancement of visible and thermal features is achieved through an adaptive re-weighting process that prioritizes reliable information:

$$\begin{aligned} \hat{F}_V &= (1 + W_q \cdot QAF_V + W_o \cdot OQF_V) \odot F_V, \\ \hat{F}_T &= (1 + W_q \cdot QAF_T + W_o \cdot OQF_T) \odot F_T, \end{aligned} \tag{3}$$

where $\odot$ denotes the Hadamard product (element-wise multiplication), and $F_i \in \mathbb{R}^{C \times H \times W}$ represents the input feature tensor. W_q and W_o are learnable weight parameters (imple-

mented as 1×1 convolutional layers) that automatically balance the trade-off between quality-driven enhancement and occlusion-driven mitigation.

The rationale for adopting min–max normalization instead of an exponential function lies in the necessity of preserving the linear characteristics of feature information. An exponential function tends to induce a "winner-takes-all" effect, which can prematurely suppress modalities with slightly weaker response intensities—potentially resulting in the loss of residual complementary feature cues in "soft occlusion" scenarios (where target visibility is partially reduced rather than completely blocked). In contrast, min–max normalization performs a linear transformation that strictly retains the relative variance and scaling range between the two modalities. This ensures the arbitration process remains sensitive to subtle differences in target visibility, thereby enabling a more balanced and inclusive cross-modal fusion that fully leverages the complementary nature of RGB and thermal modalities.

Finally, the recalibrated features $\hat{F}_V$ and $\hat{F}_T$ are concatenated along the channel dimension and fused through a 3×3 convolutional layer to generate the unified multi-modal representation. This bidirectional guidance ensures that the system can effectively decouple non-occluded and occluded regions, selectively trusting the modality that is less affected by environmental obstructions.

4. Experiment

4.1. Implementation Details

The paper adopts two datasets, VTUAV-det [2,25] and RGBTDronePerson [2] to comprehensively assess the distinct uniqueness of multimodal occlusion.

VTUAV-det: The VTUAV [25] dataset serves as a comprehensive benchmark specifically designed for the dual tracking of persons using visible and thermal imaging on unmanned aerial vehicles (UAVs). It encompasses a rich variety of scenes and diverse weather conditions to ensure robustness in real-world applications. VTUAV-det [2] carefully selected sequences filled with individuals and adopted a sampling strategy of one frame every 50 frames. This approach enabled it to compile a specialized drone-based person detection dataset. The division between training and testing subsets reflects the original structure of the VTUAV dataset, and this newly created dataset is focused only on the "person" category, including a total of 11,392 pairs of training images and 5378 pairs of test images. In the training set, a total of 74,010 individual instances are annotated, providing a rich resource for developing and evaluating tracking algorithms.

RGBTDronePerson: The RGBTDronePerson dataset contains image pairs captured by a DJI Mavic 2 Enterprise Advanced drone, which features a dual-camera system. This drone records dual-spectral videos at 1920×1080 resolution for visible light and 640×512 resolution for thermal imaging. Image pairs are sampled from these videos, with visible light images cropped to 640×512 pixels for consistency. No post-processing is applied, preserving the raw data for analysis. The dataset includes diverse lighting conditions, covering both day and night, as well as varying temperatures from winter to summer in Wuhan, China. The drone captures images from heights of 50 to 80 meters at angles of 45° to 60°. In total, there are 6125 RGBT image pairs, divided into 4900 pairs for training and 1225 for validation. Compared to the VTUAV-det dataset, objects in the RGBTDronePerson dataset are smaller and more challenging to analyze.

All the experiments are conducted on a single NVIDIA RTX 3090 GPU, with models implemented using MMDetectionV2 [26] and PyTorch [27]. Unless specified otherwise, ResNet-50 [28] with FPN [7] is used as the backbone network. For training, the model employs the SGD optimizer with a learning rate of 0.01, momentum of 0.9, and weight decay of 0.0001. In contrast, DDQ R-CNN, AODet, and UOT utilizes the AdamW optimizer

with a learning rate of 0.0001 and weight decay of 0.05. UOT utilizes PVTv2 as the backbone. To address the issue of object size imbalance, the loss is reweighted for different levels accordingly [29].

4.2. Main Result and Analysis

Result on VTUAV-det: We convert some state-of-the-art methods [30–32] into two-stream RGBT detection networks. As shown in Table 1, mAP_s, mAP_m, and mAP_l represent the average accuracy of small, medium and large targets respectively. UOT achieves the state-of-the-art performance of 35.10% mAP on the VTUAV-det dataset, surpassing QFDet by 1.8 points and AODet by 0.7 points. It is worth noting that our method shows an obvious advantage in detecting small targets, surpassing QFDet by 2.5 points and AODet by 0.9 points in terms of mAP_s. At the same time, our method is also highly competitive for medium and large targets.

Robustness to Occlusion (Table 2): To further substantiate the effectiveness of our Quality–Occlusion Arbitration (QOA) mechanism, we conducted a fine-grained evaluation on the VTUAV-occlusion subset, which is specifically filtered according to the occlusion definitions in Figure 1. As shown in Table 2, our UOT model achieves a superior performance of 34.30% mAP, outperforming the previous best competitor AODet by 1.1%. It is particularly noteworthy that the performance gain is most significant for small targets under occlusion. Specifically, our method achieves 33.90% mAP_s, which is a substantial improvement of 2.0% over AODet and 3.7% over QFDet. This underscores that while small targets are easily lost in single-modality features due to foliage or shadows, our Occlusion Fusion Module (OFM) effectively recovers their discriminative features by dynamically arbitrating between RGB and thermal streams. The consistent lead across mAP_s, mAP_m, and mAP_l on this challenging subset demonstrates that UOT is more robust than existing methods when dealing with complex environmental obstructions.

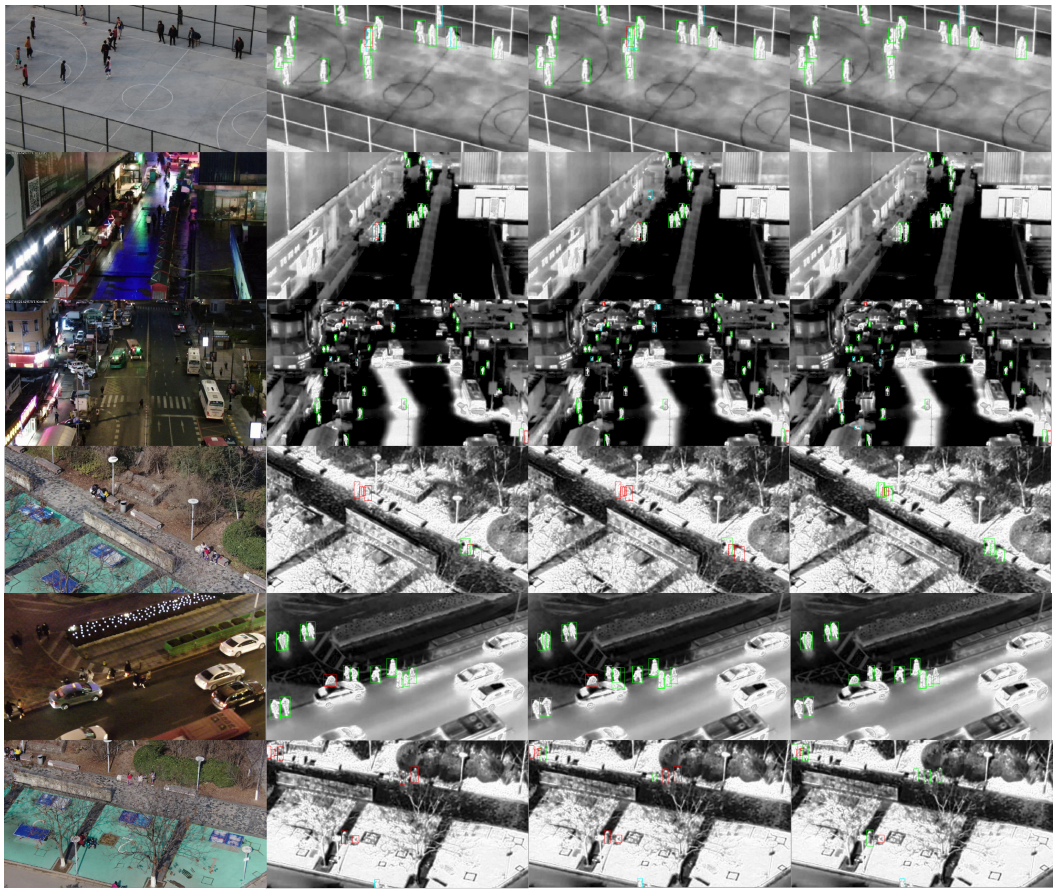
Table 1. Comparison results with state-of-the-art methods on the VTUAV-det dataset. Mark † indicates that the method has been adapted to the RGBT baseline detector version. The bold font demonstrates the best results.

Method	mAP	mAP_{50}	mAP_{75}	mAP_s	mAP_m	mAP_l	FPS
ProbEn [33]	12.70	36.40	5.20	8.50	14.20	14.80	-
CFT [34]	8.73	29.31	2.41	3.96	8.51	23.46	25.3
HRFuser [35]	25.90	55.90	20.10	2.70	27.90	51.90	14.3
TINet [36]	26.80	59.40	20.10	1.20	29.00	53.70	41.2
Faster R-CNN † [30]	26.50	58.30	20.50	1.20	28.50	54.50	39.9
ATSS † [31]	28.70	66.60	20.70	8.70	28.40	55.60	42.0
Cascade R-CNN † [32]	27.50	60.00	22.00	1.80	29.60	55.70	35.5
DetectoRS † [9]	29.80	63.20	24.20	3.10	32.30	57.90	9.5
RetinaNet † [37]	21.00	50.10	14.80	1.90	21.40	54.20	48.2
FCOS † [38]	25.20	60.60	17.20	6.00	25.30	51.80	48.4
GFL † [39]	29.70	67.30	22.00	9.50	29.60	56.50	44.6
DDQ R-CNN † [40]	32.50	74.60	23.00	17.30	31.80	56.00	18.7
QFDet [2]	33.30	75.50	24.20	18.10	32.40	57.20	27.4
AODet [24]	34.40	77.00	25.30	19.70	33.50	57.30	18.4
UOT (Ours)	**35.10**	**78.00**	**25.80**	**20.60**	**34.30**	**57.50**	14.2

Qualitative analysis on VTUAV-det: To provide an intuitive understanding of our performance gain, we present a visual comparison in Figures 5 and 6. While state-of-the-art algorithms perform well in general scenarios, they struggle under soft occlusion where

targets are partially hidden by dense tree trunks or foliage. In such cases, traditional detectors often produce fragmented features, leading to missed detections (indicated by red dashed boxes in SOTA results). Our UOT framework effectively bridges this perceptual gap. By leveraging the Occlusion–Quality Factor (OQF) within the Occlusion Fusion Module (OFM), the model identifies the modality less affected by the obstruction and uses it to guide the feature reconstruction of the occluded stream. As shown in the zoomed-in regions of Figure 6, even when the RGB features are severely contaminated by foreground foliage, the UOT model can still maintain a complete heat-map of the target by transferring reliable thermal cues, thereby ensuring consistent and reliable detection where baseline models fail.

Result on RGBTDronePerson: To further validate the generalizability of our method, we conducted comparative experiments with state-of-the-art object detection methods on the RGBTDronePerson dataset. Given that targets in this dataset exhibit significantly smaller scales and higher density compared to VTUAV-det, we implement a more granular classification protocol for small-scale objects to capture performance nuances across extreme scales. Specifically, as shown in Table 3, we define "small" targets as those with absolute pixel areas in the range of $[20, 32)$, while "tiny" targets encompass all objects within the $[0, 20)$ range. To further investigate the detection limits of our model under volatile UAV perspectives, we subdivide the "tiny" category into three sub-levels, "tiny1", "tiny2", and "tiny3", which denote absolute object sizes in the ranges of $[0, 8)$, $[8, 12)$, and $[12, 20)$ pixels, respectively.



(a)Visible image (b)QFDet (c)AODet (d)UOT

Figure 5. Comparison of visualization results on VTUAV-det with state-of-the-art methods. Green boxes represent True Positives (TP), blue boxes represent False Positives (FP), and red boxes represent False Negatives (FN).

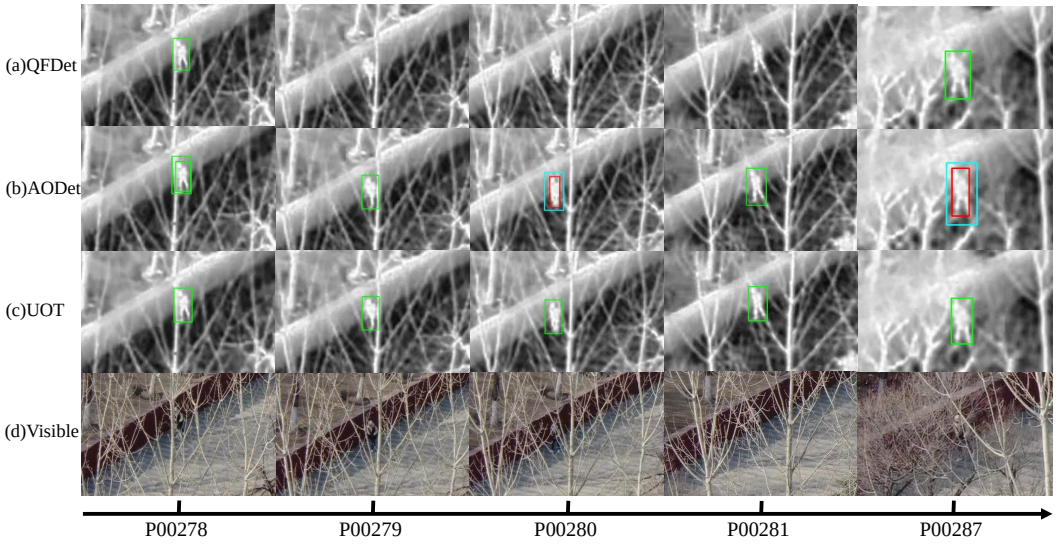


Figure 6. Continuous results for objects occluded by tree branches in the VTUAV-det dataset, green boxes represent True Positives (TP), blue boxes represent False Positives (FP), and red boxes represent False Negatives (FN). The absence of a box signifies that no target was detected. P00282–P00286 are not displayed due to weak occlusion and satisfactory detection performance.

Under this rigorous evaluation framework, UOT achieves a prominent mAP_{50}^{all} of 53.42% and an mAP_{50}^{tiny} of 54.70%, outperforming the second-best method by 5.76 points and 5.3 points, respectively. The significant performance gains across all tiny sub-categories (tiny1–3) underscore the efficacy of our cross-factor guidance mechanism. By effectively arbitrating between modalities, the OQF ensures that discriminative features are recovered even for the "tiny1" targets (less than 8 pixels), which are typically overwhelmed by background noise in conventional fusion architectures.

Qualitative analysis on RGBTDronePerson: Compared to the VTUAV-det dataset, the RGBTDronePerson dataset presents a more significant challenge due to higher occlusion density and smaller target scales. As illustrated in Figure 7, targets often occupy fewer than 20×20 pixels, making them easily overwhelmed by background noise or intermittent obstructions. Our qualitative results demonstrate that UOT excels at suppressing environmental interference. While competitive methods struggle to distinguish tiny persons from similar-sized clutter under erratic drone perspectives, UOT's Quality-Aware Factor (QAF) assigns higher importance to the most discriminative cross-modal regions. This fine-grained attention mechanism allows the model to "peek through" the gaps in soft occlusions, successfully identifying tiny targets that are nearly invisible in a single modality. This visual robustness directly explains the substantial 5.3-point gain observed in our mAP_{tiny} metrics, proving that the UOT framework is not merely fitting the data but is genuinely perceiving targets through the guidance of our cross-factor mechanism.

Table 2. Performance comparison on the curated 'Soft-Occlusion' subset (VTUAV-occlusion). Note: We focus on small object metrics (mAPs) as they are most sensitive to occlusion.

Method	mAP	mAP_s	mAP_m	mAP_l
Faster R-CNN [30]	26.80	1.00	27.50	45.60
ATSS [31]	27.00	15.00	24.80	48.00
Cascade R-CNN [32]	27.40	1.30	28.00	47.50
DDQ R-CNN [40]	32.40	29.00	30.00	48.90
QFDet [2]	32.70	30.20	30.20	49.20
AODet [24]	33.20	31.90	30.50	50.00
UOT (Ours)	**34.30**	**33.90**	**30.70**	**50.10**

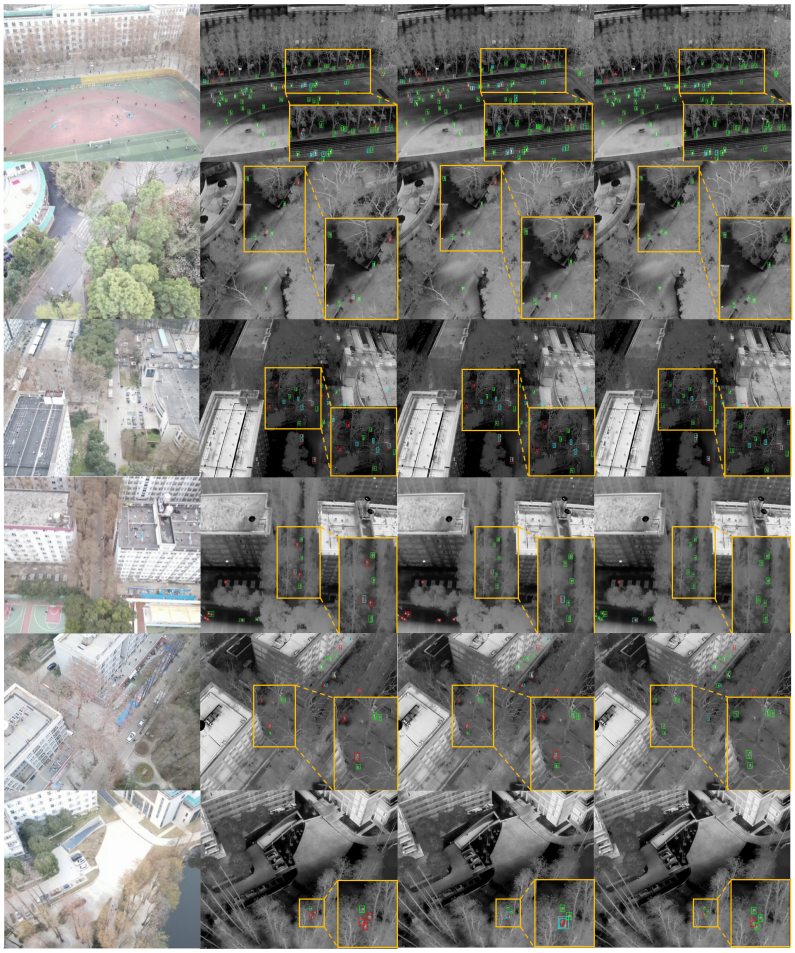


Figure 7. Comparison of visualization results on RGBTDronePerson with state-of-the-art methods. Green boxes represent True Positives (TP), blue boxes represent False Positives (FP), and red boxes represent False Negatives (FN).

Table 3. Comparison with state-of-the-art methods on the RGBTDronePerson dataset. Mark † indicates that the method has been adapted to the RGBT baseline detector version. The bold font demonstrates the best results.

Method	mAP_{50}^{all}	mAP_{50}^{tiny}	mAP_{50}^{tiny1}	mAP_{50}^{tiny2}	mAP_{50}^{tiny3}	mAP_{50}^{small}	mAP_{25}^{all}
ProbEn [33]	3.64	3.83	0.77	5.41	5.19	0.33	10.73
CFT [34]	22.69	22.83	16.72	18.14	27.52	8.14	37.32
HRFuser [35]	22.23	22.50	0.00	26.73	26.26	23.85	33.24
TINet [36]	28.30	28.60	0.00	24.12	34.99	**34.97**	40.34
Faster R-CNN † [30]	28.37	28.48	0.00	21.97	34.87	34.40	38.03
ATSS † [31]	36.24	37.47	16.92	24.16	43.59	24.62	53.14
Cascade R-CNN † [32]	31.55	31.99	0.00	29.43	37.77	33.61	42.47
DetectoRS † [9]	32.21	32.65	11.55	30.89	38.40	33.71	43.37
RetinaNet † [37]	22.87	23.34	4.69	13.57	32.66	15.77	38.92
FCOS † [38]	29.89	30.71	9.40	22.73	34.87	26.00	45.21
GFL † [39]	38.77	40.00	17.86	28.15	46.66	26.75	55.59
DDQ R-CNN † [40]	45.92	47.16	22.53	36.83	53.68	30.64	59.30
QFDet [2]	46.72	48.75	22.15	37.91	53.71	28.41	61.62
AODet [24]	47.66	49.40	22.66	37.14	56.43	29.39	60.99
UOT (Ours)	**53.42**	**54.70**	**24.70**	**44.48**	**59.92**	34.12	**64.46**

4.3. Ablation Study

To quantitatively assess the individual and collective contributions of the proposed components, we conduct a systematic ablation study on the RGBTDronePerson (RGBTD) dataset. The ablation of the backbone network is based on PVTv2.

As shown in Table 4, the comparative analysis reveals that the integration of the Occlusion–Quality Factor (OQF) yields a substantial performance leap over the baseline architecture. Specifically, on the RGBTD dataset, the inclusion of OQF elevates the mAP_{50}^{all} from 47.66% to 51.20%. Notably, the mAP_{50}^{tiny} exhibits a more pronounced escalation, rising from 49.40% to 52.63%. This absolute gain of 3.23 percentage points for tiny targets is particularly significant, as it demonstrates that our Quality–Occlusion Arbitration (QOA) mechanism effectively recovers subtle semantic signatures of distant pedestrians that are typically submerged in foliage-induced noise. Ultimately, the full deployment of our novel paradigm achieves a peak mAP_{50}^{all} of 53.42% and mAP_{50}^{tiny} of 54.70%, representing significant overall improvements of 5.76 and 5.3 percentage points, respectively.

Table 4. Ablation study of the proposed methods on RGBTDronePerson.

Method		mAP			
Backbone (PVTv2)	**OQF**	**mAP_{50}^{all}**	**mAP_{50}^{tiny}**	**mAP_{50}^{small}**	**mAP_{25}^{all}**
		47.66	49.40	29.39	60.99
✓		49.52	51.07	30.83	62.08
	✓	51.20	52.63	30.76	62.87
✓	✓	**53.42**	**54.70**	**34.12**	**64.46**

These quantitative gains provide robust empirical evidence that the OQF is not merely a supplementary enhancement but a fundamental driver for resilient detection in complex aerial scenarios. The superiority of the OQF mechanism originates from its dual-functional architecture: it serves as a perceptual gatekeeper that evaluates cross-modal reliability through quality-aware priors, while simultaneously executing pixel-wise arbitration to suppress "feature contamination" from occluded modalities. By dynamically re-weighting the information flow, this unified mechanism empowers the network to "perceive through" fragmented natural obstructions, effectively mitigating the feature degradation issues that typically plague baseline detectors in densely occluded UAV environments.

4.4. Analysis of Failure Cases

Although the proposed UOT framework has achieved significant progress in addressing soft occlusion within UAV-based search and rescue missions, an in-depth examination of failure cases reveals persistent limitations in certain extreme scenarios. Figure 8 illustrates several representative failure modes from the RGBTDronePerson dataset.

These failure instances are predominantly concentrated in "hard occlusion" conditions, where the target suffers from severe physical obstruction in both visible and infrared modalities simultaneously. For instance, when a pedestrian is entirely obscured by substantial buildings or completely opaque barriers, nearly all discriminative features of the target are lost. Under such circumstances, even the Pre-Discriminant Module (PDM) and Occlusion Fusion Module (OFM) struggle to derive meaningful recovery cues from either modality. The absence of accessible cross-modal complementary features prevents effective feature reconstruction, ultimately resulting in false negatives. This suggests that relying exclusively on the cross-modal complementarity of single-frame images remains inadequate for addressing extreme bi-modal failures.



(a) Thermal Image (b) Visible Image (c) UOT

Figure 8. Example of detection failure in RGBTDronePerson dataset.

To overcome this limitation, Future research could focus on temporal feature modeling. By leveraging information from consecutive frames in video sequences, the model could exploit the target's motion trajectory and historical visual context prior to entering the hard-occlusion area, thereby enabling the "memory" and prediction of occluded targets over short durations. Integrating spatial cross-modal fusion with temporal context holds the potential to further enhance detection robustness in complex and dynamic environments.

5. Limitation and Discussion

While the proposed UOT framework significantly enhances the reliability of UAV-based person detection by mitigating "soft occlusion," certain aspects offer opportunities for further exploration in practical deployments.

1. **Contextual Recovery under Extreme Occlusion**: The current study primarily focuses on "soft occlusion," where at least one modality provides discriminative cues. However, in scenarios of "hard occlusion"—where the target is simultaneously obstructed in both RGB and thermal spectra—relying solely on instantaneous, single-frame spatial information presents an inherent physical limitation. Since discriminative features are temporarily absent in both channels, future research could benefit from integrating temporal sequences and motion priors. By leveraging the continuity of UAV video streams, the model could potentially "hallucinate" or track the target's trajectory through periods of total obstruction, thereby maintaining detection consistency even when sensory data are momentarily unavailable.
2. **Balancing Robustness and Edge Efficiency**: The dual-factor arbitration mechanism (PDM and OFM) introduces additional computational layers to ensure precise feature recovery. While the inference speed (14.2 FPS) is lower than some lightweight detectors, it fulfills the operational requirements for drone-based emergency response, the erratic and resource-constrained nature of UAV edge terminals demands ongoing architectural optimization. Future iterations of this work could explore lightweight fusion operators or knowledge distillation techniques to further reduce the latency of multi-modal integration. This would ensure that the occlusion-aware advantages of UOT can be fully realized even on low-power airborne processing units without compromising real-time performance.

6. Conclusions

In this study, we identified the unique challenges posed by multimodal occlusions and introduced an anti-occlusion detector specifically designed for drone-based RGBT human detection to address these issues. We developed a quality- and occlusion-guided multimodal feature fusion architecture to enhance detection performance. Extensive experiments were conducted on two large-scale RGBT human detection datasets, RGBTDronePerson and VTUAV-det. The results demonstrate that our proposed method consistently outperforms existing approaches across various scenarios, highlighting its robustness and effectiveness.

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

Funding: This research was partially supported by the Fundamental Research Funds for the Central Universities (2042025KF0053).

Data Availability Statement: The code and datasets are available at https://github.com/ZimingGui/UOT (accessed on 26 February 2026).

Conflicts of Interest: Author Shaobo Liu, Dong Yang and Tongyuan Zou were employed by the Space Star Technology Co., Ltd. The remaining authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

References

1. Akshatha, K.R.; Karunakar, A.K.; Satish Shenoy, B.; Phani Pavan, K.; Dhareshwar, C.V.; Johnson. D.G. Manipal-UAV person detection dataset: A step towards benchmarking dataset and algorithms for small object detection. *ISPRS J. Photogramm. Remote Sens.* **2023**, *195*, 77–89. [CrossRef]
2. Zhang, Y.; Xu, C.; Yang, W.; He, G.; Yu, H.; Yu, L.; Xia, G.S. Drone-based RGBT tiny person detection. *ISPRS J. Photogramm. Remote Sens.* **2023**, *204*, 61–76. [CrossRef]
3. Zhang, Y.; Rui, X.; Song, W. A UAV-Based Multi-Scenario RGB-Thermal Dataset and Fusion Model for Enhanced Forest Fire Detection. *Remote Sens.* **2025**, *17*. [CrossRef]
4. Zhang, X.; Feng, Y.; Wang, N.; Lu, G.; Mei, S. Aerial Person Detection for Search and Rescue: Survey and Benchmarks. *J. Remote Sens.* **2025**, *5*, 0474. [CrossRef]
5. Chen, Y.; Zhang, P.; Li, Z.; Li, Y.; Zhang, X.; Qi, L.; Sun, J.; Jia, J. Dynamic Scale Training for Object Detection *arXiv* **2021**, arXiv:2004.12432. [CrossRef]
6. Kisantal, M.; Wojna, Z.; Murawski, J.; Naruniec, J.; Cho, K. Augmentation for small object detection *arXiv* **2019**, arXiv:1902.07296. [CrossRef]
7. Lin, T.Y.; Dollár, P.; Girshick, R.; He, K.; Hariharan, B.; Belongie, S. Feature Pyramid Networks for Object Detection. In Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, USA, 21–26 July 2017; pp. 936–944. [CrossRef]
8. Liu, S.; Huang, D.; Wang, Y. Learning Spatial Fusion for Single-Shot Object Detection. *arXiv* **2019**, arXiv:1911.09516. [CrossRef]
9. Qiao, S.; Chen, L.C.; Yuille, A. DetectoRS: Detecting Objects with Recursive Feature Pyramid and Switchable Atrous Convolution. In Proceedings of the 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Nashville, TN, USA, 20–25 June 2021; pp. 10208–10219. [CrossRef]
10. Chen, P.; Wang, J.; Zhang, Z.; He, C. FRLI-Net: Feature Reconstruction and Learning Interaction Network for Tiny Object Detection in Remote Sensing Images. *IEEE Signal Process. Lett.* **2025**, *32*, 2159–2163. [CrossRef]
11. Zhang, H.; Zhang, H.; Liu, K.; Gan, Z.; Zhu, G.N. UAV-DETR: Efficient End-to-End Object Detection for Unmanned Aerial Vehicle Imagery. In Proceedings of the 2025 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), Hangzhou, China, 19–25 October 2025; pp. 15143–15149. [CrossRef]
12. Peng, H.; Hu, Y.; Yu, B.; Zhang, Z. TCAINet an RGB T salient object detection model with cross modal fusion and adaptive decoding. *Sci. Rep.* **2025**, *15*, 14266. [CrossRef]

13. Zhang, X.; Feng, Y.; Wang, N.; Lu, G.; Mei, S. Transformer-Based Person Detection in Paired RGB-T Aerial Images with VTSaR Dataset. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2025**, *18*, 5082–5099. [CrossRef]
14. He, Z.; Zhao, B.; Wu, Y.; Jiang, Y.; Zhao, Q. A Progressive Target-Aware Network for Drone-Based Person Detection Using RGB-T Images. *Remote Sens.* **2025**, *17*, 3361. [CrossRef]
15. Tang, Z.; Xie, Y.; Xu, T.; Wu, X.J.; Kittler, J. Learning Bi-directional fusion and deformation-sensitive loss for RGB-T tiny object detection. *Inf. Fusion* **2026**, *128*, 103985. [CrossRef]
16. Zhang, X.; Feng, Y.; Zhang, S.; Wang, N.; Lu, G.; Mei, S. Robust Aerial Person Detection with Lightweight Distillation Network for Edge Deployment. *IEEE Trans. Geosci. Remote Sens.* **2024**, *62*, 5630616. [CrossRef]
17. 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]
18. Bodla, N.; Singh, B.; Chellappa, R.; Davis, L.S. Soft-NMS–improving object detection with one line of code. In Proceedings of the IEEE International Conference on Computer Vision, Venice, Italy, 22–29 October 2017; pp. 5561–5569.
19. Huang, X.; Ge, Z.; Jie, Z.; Yoshie, O. NMS by representative region: Towards crowded pedestrian detection by proposal pairing. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 13–19 June 2020; pp. 10750–10759.
20. Cui, Y.; Hou, B.; Wu, Q.; Ren, B.; Wang, S.; Jiao, L. Remote Sensing Object Tracking with Deep Reinforcement Learning Under Occlusion. *IEEE Trans. Geosci. Remote Sens.* **2022**, *60*, 5605213. [CrossRef]
21. Bai, Y.; Song, Y.; Zhao, Y.; Zhou, Y.; Wu, X.; He, Y.; Zhang, Z.; Yang, X.; Hao, Q. Occlusion and Deformation Handling Visual Tracking for UAV via Attention-Based Mask Generative Network. *Remote Sens.* **2022**, *14*, 4756. [CrossRef]
22. Zhang, W.; Liu, C.; Chang, F.; Song, Y. Multi-Scale and Occlusion Aware Network for Vehicle Detection and Segmentation on UAV Aerial Images. *Remote Sens.* **2020**, *12*, 1760. [CrossRef]
23. Li, X.; Diao, W.; Mao, Y.; Gao, P.; Mao, X.; Li, X.; Sun, X. OGMN: Occlusion-guided multi-task network for object detection in UAV images. *ISPRS J. Photogramm. Remote Sens.* **2023**, *199*, 242–257. [CrossRef]
24. Gui, Z.; Zhang, Y.; Lei, X.; Zhang, R.; Yang, W. AODet: Anti-Occlusion for Enhanced Small Object Detection in Drone-Based RGBT Imagery. In Proceedings of the IGARSS 2024—2024 IEEE International Geoscience and Remote Sensing Symposium, Athens, Greece, 7–12 July 2024; pp. 954–958. [CrossRef]
25. Zhang, P.; Zhao, J.; Wang, D.; Lu, H.; Ruan, X. Visible-thermal UAV tracking: A large-scale benchmark and new baseline. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, New Orleans, LA, USA, 18–24 June 2022; pp. 8886–8895.
26. Chen, K.; Wang, J.; Pang, J.; Cao, Y.; Xiong, Y.; Li, X.; Sun, S.; Feng, W.; Liu, Z.; Xu, J.; et al. MMDetection: Open mmlab detection toolbox and benchmark. *arXiv* **2019**, arXiv:1906.07155. [CrossRef]
27. Paszke, A.; Gross, S.; Massa, F.; Lerer, A.; Bradbury, J.; Chanan, G.; Killeen, T.; Lin, Z.; Gimelshein, N.; Antiga, L.; et al. PyTorch: An Imperative Style, High-Performance Deep Learning Library. *arXiv* **2019**, arXiv:1912.01703.
28. He, K.; Zhang, X.; Ren, S.; Sun, J. Deep Residual Learning for Image Recognition. In Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA, 27–30 June 2016; pp. 770–778. [CrossRef]
29. Yang, C.; Huang, Z.; Wang, N. QueryDet: Cascaded Sparse Query for Accelerating High-Resolution Small Object Detection. In Proceedings of the 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), New Orleans, LA, USA, 18–24 June 2022; pp. 13658–13667. [CrossRef]
30. Ren, S.; He, K.; Girshick, R.; Sun, J. Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks. *IEEE Trans. Pattern Anal. Mach. Intell.* **2017**, *39*, 1137–1149. [CrossRef]
31. Zhang, S.; Chi, C.; Yao, Y.; Lei, Z.; Li, S.Z. Bridging the gap between anchor-based and anchor-free detection via adaptive training sample selection. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 14–19 June 2020; pp. 9759–9768.
32. Cai, Z.; Vasconcelos, N. Cascade R-CNN: Delving Into High Quality Object Detection. In Proceedings of the 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018; pp. 6154–6162. [CrossRef]
33. Chen, Y.T.; Shi, J.; Ye, Z.; Mertz, C.; Ramanan, D.; Kong, S. Multimodal object detection via probabilistic ensembling. In *Proceedings of the European Conference on Computer Vision, Tel Aviv, Israel, 23–27 October 2022*; Springer: Cham, Switzerland, 2022; pp. 139–158.
34. Fang, Q.; Han, D.; Wang, Z. Cross-Modality Fusion Transformer for Multispectral Object Detection. *arXiv* **2022**, arXiv:2111.00273. [CrossRef]
35. Brödermann, T.; Sakaridis, C.; Dai, D.; Van Gool, L. HRFuser: A Multi-Resolution Sensor Fusion Architecture for 2D Object Detection. In Proceedings of the 2023 IEEE 26th International Conference on Intelligent Transportation Systems (ITSC), Bilbao, Spain, 24–28 September 2023; pp. 4159–4166. [CrossRef]
36. Zhang, Y.; Yu, H.; He, Y.; Wang, X.; Yang, W. Illumination-Guided RGBT Object Detection with Inter- and Intra-Modality Fusion. *IEEE Trans. Instrum. Meas.* **2023**, *72*, 2508013. [CrossRef]

37. Lin, T.Y.; Goyal, P.; Girshick, R.; He, K.; Dollár, P. Focal Loss for Dense Object Detection. In Proceedings of the 2017 IEEE International Conference on Computer Vision (ICCV), Venice, Italy, 22–29 October 2017; pp. 2999–3007. [CrossRef]
38. Tian, Z.; Shen, C.; Chen, H.; He, T. FCOS: Fully Convolutional One-Stage Object Detection. In Proceedings of the 2019 IEEE/CVF International Conference on Computer Vision (ICCV), Seoul, Republic of Korea, 27 October–2 November 2019; pp. 9626–9635. [CrossRef]
39. Li, X.; Wang, W.; Wu, L.; Chen, S.; Hu, X.; Li, J.; Tang, J.; Yang, J. Generalized focal loss: Learning qualified and distributed bounding boxes for dense object detection. *Adv. Neural Inf. Process. Syst.* **2020**, *33*, 21002–21012.
40. Zhang, S.; Wang, X.; Wang, J.; Pang, J.; Lyu, C.; Zhang, W.; Luo, P.; Chen, K. Dense Distinct Query for End-to-End Object Detection. In Proceedings of the 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Vancouver, BC, Canada, 17–24 June 2023; pp. 7329–7338. [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.