Fusion Meets Diverse Conditions: A High-diversity Benchmark and Baseline for UAV-based Multimodal Object Detection with Condition Cues

Supplementary Material

Contents

A. Details of ATR-UMOD dataset

A.1. UAV-borne RGB-IR Platform

Our dataset is acquired using a professional DJI Matrice 300 RTK equipped with either a Zenmuse H20T or H20N camera to ensure flexible and stable flight in complex conditions. The setup details of the DJI Matrice 300 RTK are provided in Tab. 5. The Zenmuse H20T and H20N cameras are advanced dual-sensor systems with vertically arranged RGB and IR lenses. Both lenses operate at a frame rate of 30 frames per second. The RGB lens captures images at a resolution of 1920 × 1080 pixels, while the IR lens operates within the 8 ∽ 14 μm wavelength range with a resolution of 640 × 512 pixels.

A.2. The Misalignment in UAV-based RGB-IR Images

The raw multimodal images collected by the UAV-borne RGB-IR platform exhibit inevitable misalignment due to the differences in imaging space and acquisition time of the two sensors, as depicted in Fig. 6. This misalignment disrupts spatial correspondence between RGB and IR representations, leading to erroneous feature fusion in downstream tasks. As shown in Fig. 7, this misalignment can be attributed to several key factors:

(1) **Field of View Mismatch.** Due to the hardware limitations of the multimodal sensor system, the RGB and IR images have discrepancies in the field of view, with the RGB camera capturing larger field of view than the IR camera. This discrepancy results in *discrepancies in spatial scope*, making direct pixel-level misalignment challenging.

(2) **Resolution Disparity.** RGB and IR cameras typically have different resolutions due to sensor characteristics, leading to a resolution mismatch between the two modalities. The RGB camera often provides higher-resolution images than the IR camera, resulting in *discrepancies in reso-*

Parameter	Specification
Aircraft Dimensions	Unfolded: 810 × 670 × 430 mm Folded: 430 × 420 × 430 mm
Weight	Approx. 6.3 kg (without payload)
Maximum Payload	2.7 kg
Max Flight Time	Up to 55 minutes (no payload)
Max Speed	23 m/s
Max Operating Altitude	7000 m
Maximum Flight Time	Approximately 55 minutes (with TB60 battery)
Maximum Ascend Speed	5 m/s (at sea level)
Maximum Descend Speed	4 m/s (vertical), 6 m/s (glide)
Maximum Horizontal Speed	23 m/s (at sea level, windless conditions)

Table 5. The setup of the DJI Matrice 300 RTK

Figure 6. An example of misalignment in raw RGB-IR image pairs.

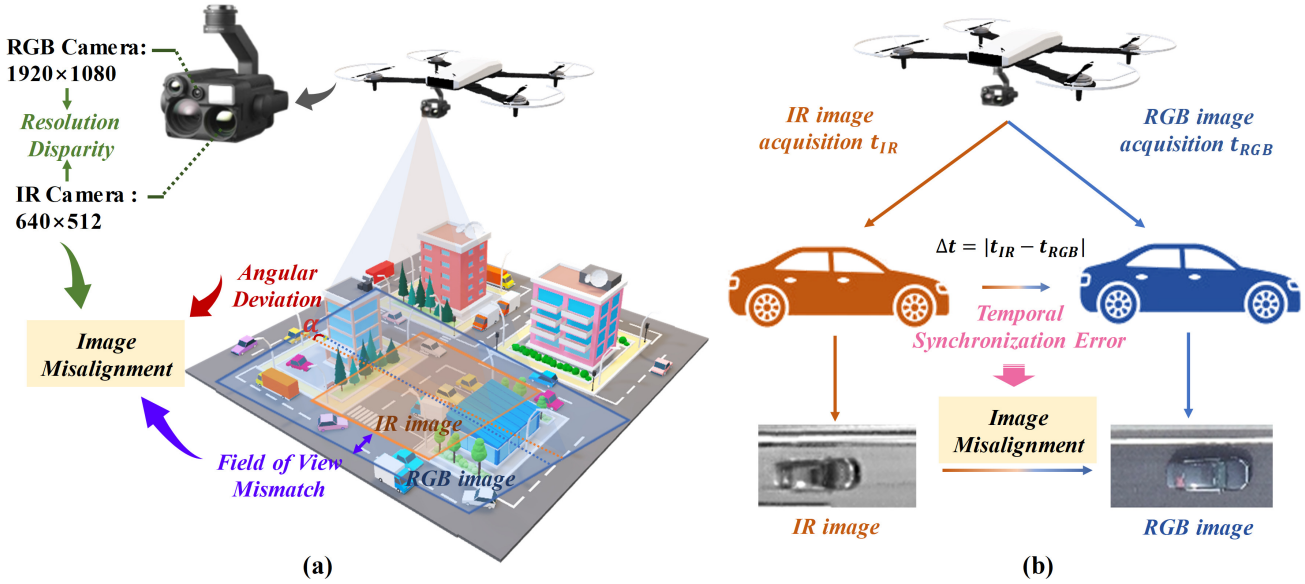

Figure 7. Illustration of the image misalignment reasons in UAV-based RGB-IR images. (a) Misalignment for field of view mismatch, resolution disparity, and angular deviation in RGB-IR sensors. (b) Misalignment for temporal synchronization error in RGB-IR sensors.

lution.

(3) **Angular Deviation.** Systematic errors in the manufacturing of RGB and IR sensors, primarily caused by len calibration discrepancies and sensor placement inaccuracies, can lead to *discrepancies in field angles*. This issue is particularly pronounced for distant objects since small differences in field angles translate into larger spatial offsets at greater distances.

(4) **Temporal Synchronization Error.** There exists a slight but inherent temporal delay between the multimodal images acquisition due to differences in sensor response times and the triggering mechanisms. This non-synchronicity means that any motion of the object or platform will lead to a *discrepancies in object positions* in the imagery. This time lag-induced displacements can significantly affect the accuracy of feature matching and fusion processes, especially in scenarios involving fast-moving objects.

Consequently, the fusion of these misaligned images necessitates the advanced preprocessing techniques to resolve both spatial and temporal discrepancies for subsequent analysis.

A.3. Data Preprocessing

To ensure high-quality and well-aligned RGB-IR image pairs for subsequent analysis, we developed a rigorous data preprocessing pipeline to address misalignment and image invalidity challenges, as shown in Fig. 8. The pipeline consists of the following key steps:

(1) **Video Frame Extraction.** We uniformly sampled frames from the original RGB and IR video streams to maintain a consistent temporal resolution across both modalities. This step ensures that each frame in one modality can be accurately matched with its counterpart in the other.

(2) **Timestamp Registration.** Due to inherent temporal discrepancies between the RGB and IR sensors, dynamic objects in the scene often exhibit positional shifts across multimodal images. To mitigate this issue, we performed timestamp registration by artificially minimizing the temporal offsets between corresponding frames. Specifically, initial frame pairs were selected by analyzing object positions across multimodal streams, ensuring that positional deviations of dynamic elements were minimized. Subsequently, unmatched frames in either modality were discarded to maintain strict one-to-one multimodal image correspondence.

(3) **Data Pruning.** To enhance the quality of the dataset and address image invalidity challenges, we manually selected representative RGB-IR image pairs and discarding redundant or anomalous frames. In detail, image pairs exhibiting minimal variation, significant artifacts or anomalies, as well as missing modality data were excluded. Such pruning ensures that only semantically meaningful and high-quality image pairs are retained for further processing.

(4) **Spatial Registration.** Given the inherent spatial misalignment between RGB and IR images, we developed an automated registration workflow in a coarse-to-fine manner.

- *Coarse Registration.* The coarse registration process aligns RGB images with IR images in field angles, spatial scopes, and resolution through a series of transformations. Firstly, RGB images were rotated to match the an-

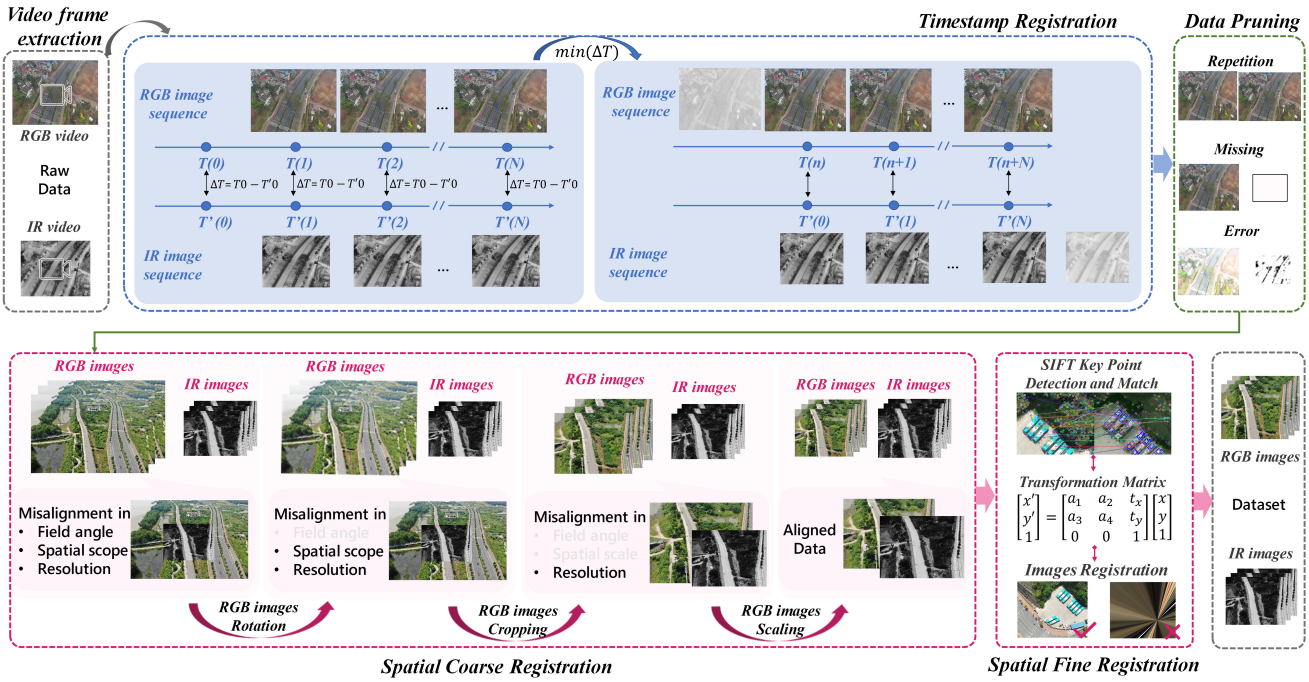

Figure 8. Visualization of the data preprocessing process in ATR-UMOD dataset.

gular orientation of their IR counterparts. Next, cropping was applied to ensure that the field of view remained spatially consistent across modalities. Finally, images were rescaled to match the IR resolution of 640×512 pixels. These transformations were parameterized using an affine matrix and a cropping matrix. Through extensive statistical analysis, we observed that the transformation parameters were approximately consistent under the same imaging height and angle. Leveraging this insight, we established a mapping function between imaging parameters (flying heights and camera angles) and the transformation parameters, enabling fully automated coarse registration.

- *Fine Registration.* While coarse registration establishes an initial alignment, only utilizing unified matrix parameters may not ensure precise registration across all cases. To address residual misalignment, fine registration employs feature-based methods for further refinement at the pixel level. Initially, distinctive features were extracted from the multimodal images using the Scale-Invariant Feature Transform (SIFT) [10]. Robust feature matching was then achieved through the Random Sample Consensus (RANSAC) algorithm [4]. Based on the matched features, an affine transformation was estimated to achieve precise pixel-level alignment. To prevent significant information loss due to potential SIFT failure, we implemented a fallback mechanism: if the proportion of invalid pixels (e.g., zero-value pixels) in the registered RGB images exceeded 30%, the final output reverted to the result of coarse registration.

A.4. Object Annotation

For object annotation, we independently annotated objects in both RGB and IR images. Each annotation underwent a thorough secondary review to enhance precision, establishing a reliable ground truth for the object detection task. To maintain cross-modal consistency, we performed manual cross-verification to align corresponding objects between RGB and IR images. This step ensures that object categories and spatial positions remain as consistent as possible across modalities. The final dataset provides annotations in both XML and JSON formats, supporting both horizontal bounding boxes and rotated bounding boxes. Horizontal bounding boxes are represented as (x, y, w, h), where (x, y) denotes the center coordinates, and w and h represent the width and height of the box, respectively. However, given the diverse orientations of objects in UAV-based images, it can be challenging to mark object extents accurately using only horizontal bounding boxes. Consequently, rotated bounding boxes are also available in our annotation files with two widely adopted formats: A parameterized representation as (x, y, w, h, θ), where θ denotes the orientation angle relative to the horizontal axis. A point-based representation defined by the coordinates of the four corners: $(x_1, y_1, x_2, y_2, x_3, y_3, x_4, y_4)$. This dual-format annotation strategy enhances the adaptability of the dataset and facilitates a wide range of research applications.

A.5. Visualization of Different Object Categories

To improve the practical applicability of the detection model, ATR-UMOD dataset encompasses a diverse set of

11 object categories, including car, SUV, van, bus, freight car, truck, motorcycle, trailer, excavator, crane, and tank truck categories, offering a comprehensive and fine-grained object categorization. Sample images of these categories are shown in Fig. 9. This diversity forces the model to learn high-level discriminative features for fine-grained detection, improving its ability to distinguish visually similar objects with greater precision.

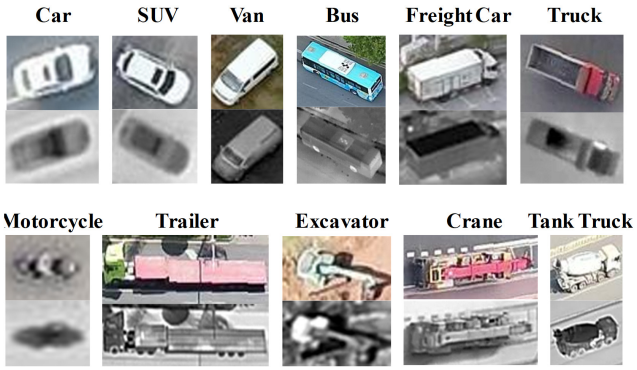

Figure 9. Examples of different object categories in ATR-UMOD dataset. The first row shows the RGB images, while the second row shows the corresponding IR images.

B. Implementation Details of our PCDF

B.1. Detailed Initial Prompt Construction

As illustrated in Eq. (1) in the main text, the initial prompt is constructed by formatting various condition attributes into a fixed template. Considering the standard text prompts for CLIP are typically structured as "A photo of a [CLASS]", we design a similar template to ensure compatibility. This template comprises a subject description and several condition prefixes. Specifically, the subject description consists of a global description, "An aerial image," and an ending symbol, ".", which provides a domain-specific context for the subsequent conditions. The condition prefixes are designed to specify specific condition attributes, including scenario, height, angle, time, weather, and illumination. For example, the prefix "over" specifies the scenario, while "at" refers to both the altitude and angle of the image. Consequently, a complete initial prompt can constructed as follows:

```
An aerial image, over {scenario}, at a
altitude of {altitude}, at an angle of
{angle}, at {Time}, in {Weather},
in {Illumination}.
```

B.2. Condition Merged in Experiments

The excessive variability of conditions results in limited sample sizes for each conditions, increasing the risk of overfitting and hindering model training. Therefore, conditions are appropriately merged to enhance data efficiency. In detail, altitude conditions are categorized into low altitude ($\leq 120m$) and high altitude ($> 120m$); angle conditions are grouped into low angle ($\leq 30°$) and high angle ($> 30°$); time conditions are merged into Morning ($7AM \backsim 2PM$), Afternoon ($2PM \backsim 7PM$), and Night ($7PM \backsim 7AM$); "After Rain" and "Rainy" combined into "Cloudy" under weather condition; "Overexposure" and "Twilight" merged into "Normal" and "Dim" under illumination condition, respectively.

B.3. The Pipeline of the Training and Testing Process

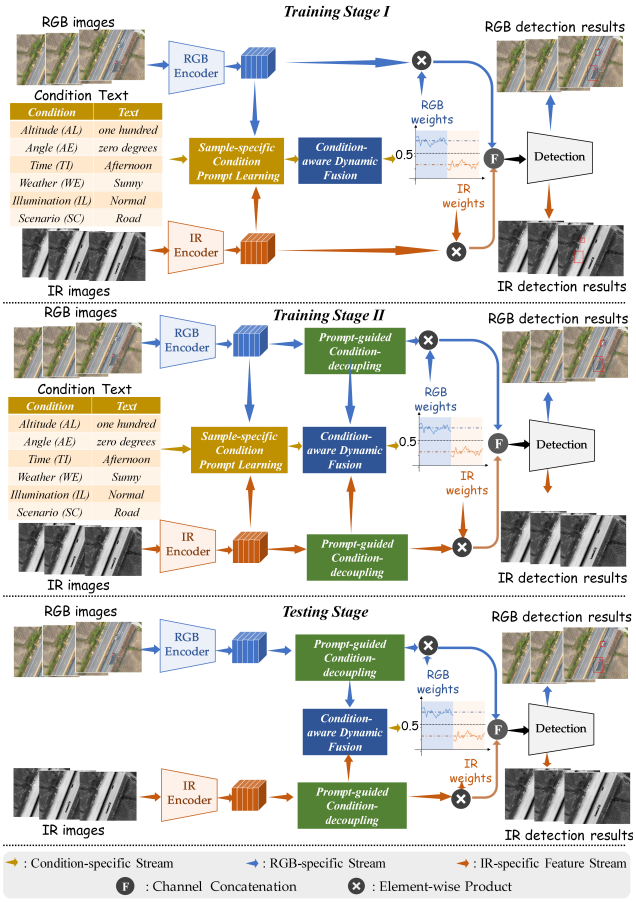

Figure 10. The training and testing processes of PCDF.

To ensure effective training of each component, our model adopts a two-stage training strategy. As illustrated in Fig. 10, the first stage focuses on learning the relationship between conditions and multimodal reliability by training only the sample-specific condition prompt learning (SCPL) module and the condition-aware dynamic fusion (CDF) module. Building on the pre-trained parameters in this stage, the second stage trains the prompt-guided condition-

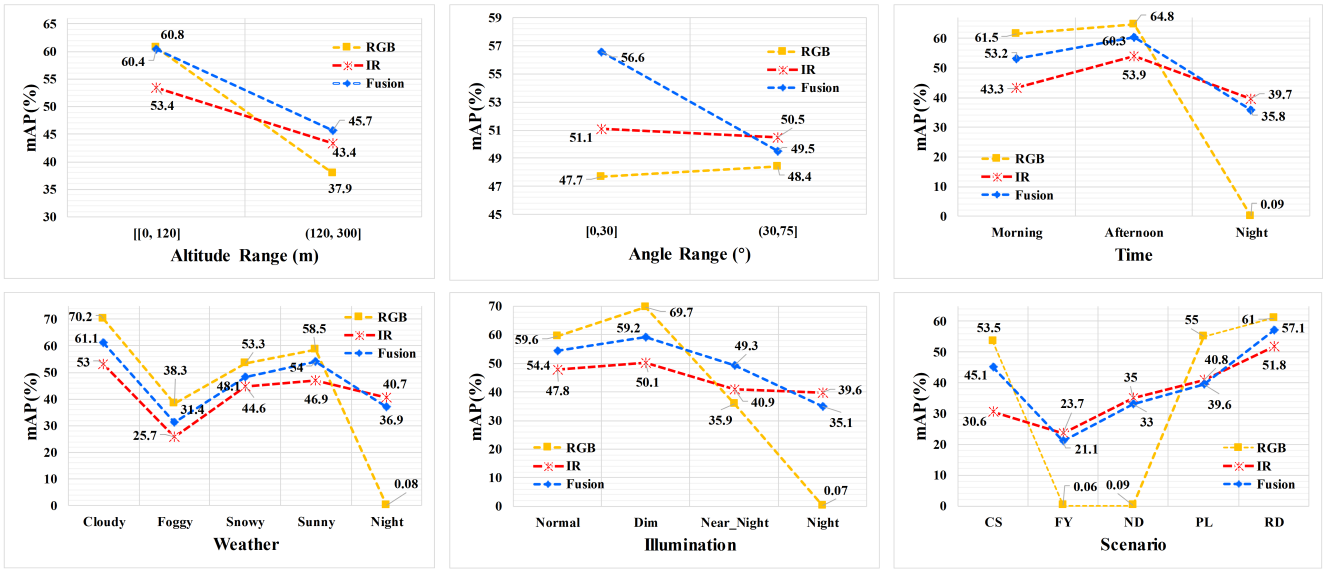

Figure 11. Visualization of the unimodal RGB model, unimodal IR model, and multimodal detector on ATR-UMOD dataset across different conditions. Noted that CS, FY, ND, PL, and RD represent Construction site, Factory, Neighborhood, Parking lot, and Road, respectively.

decoupling module (PCD) to decouple multimodal features into condition-specific and condition-invariant components, allowing testing without explicit condition annotations. As a result, in the testing stage, the PCD module, which relies on condition prompts as input, is unnecessary, thereby enhancing the practicality and deployment feasibility of our PCDF model in real-world applications.

C. Research Background of Our Method

Under diverse imaging conditions, object characteristics in multimodal images exhibit significant variations, leading to fluctuations in multimodal reliability [18]. For instance, under favorable illumination conditions, RGB images provide rich color information and detailed object textures, making them highly reliable for object detection. However, in poor illumination conditions, RGB images may suffer from noise and reduced visibility, severely degrading detection performance. In contrast, IR images remain unaffected by illumination changes, offering higher reliability in such conditions. To validate this issue, we evaluate the detection performance of the unimodal RGB, unimodal IR, and multimodal detectors across different conditions on the ATR-UMOD dataset, as shown in Fig. 11. For a fair comparison, unimodal detectors share the same architecture, while the multimodal detector integrates features from the unimodal detectors.

The results in Fig. 11 reveal that the performance gap between RGB and IR unimodal detectors fluctuates across conditions, demonstrating the dynamic reliability of each modality under varying conditions. Theoretically, a fusion model should outperform the best unimodal model in any given condition, as it integrates complementary multimodal information for a more robust and comprehensive object representations. However, as depicted in Fig. 11, the performance of the fusion models often falls short of the best unimodal model. This discrepancy arises because the fusion model fail to account for the fluctuating reliability of modalities across varying conditions, leading to suboptimal fusion that neither fully leverages the strengths of the more reliable modality nor effectively mitigates the noise introduced by less reliable ones. To overcome this limitation, we propose the PCDF method, which dynamically adjusts multimodal contributions based on conditions, enhancing the robustness and reliability of the fusion model across diverse scenarios.

D. Experiments in Dronevehicle

We evaluate our proposed PCDF against the state-of-the-art unimodal and multimodal fusion methods on the DroneVehicle dataset. It is worth noting that as the DroneVehicle dataset has no extra condition annotations, we manually relabelled "day" and "night" conditions for each image pair to complete the model training. Among unimodal approaches, we consider one-stage detectors including RetinaNet [8], S^2A-Net [5], and YOLOv5s [7]; as well as the two-stage methods Faster R-CNN [11], Oriented R-CNN [13], and RoITransformer [3]. To ensure a comprehensive evaluation, each method uses RGB and IR images independently as training data. For multimodal fusion, we compare PCDF with several advanced methods: UA-CMDet [12], Halfway Fusion [9], CIAN [16], AR-CNN [17], MBNet [19], TS-FADet [15], C^2Former [14], SLBAF-Net [2], CALNet [6], and OAFA [1]. All the multimodal detectors are trained

Detectors	Modality	Car	Truck	Freight-car	Bus	Van	mAP (%) ↑
RetinaNet [8]	RGB	78.5	34.4	24.1	69.8	28.8	47.1
Faster R-CNN [11]		79.0	49.0	37.2	77.0	37.0	55.9
Oriented R-CNN [13]		80.1	53.8	41.6	85.4	43.3	60.8
S^2A-Net [5]		80.0	54.2	42.2	84.9	43.8	61.0
RoITransformer [3]		61.6	55.1	42.3	85.5	44.8	61.6
YOLOv5s [7]		78.6	55.3	43.8	87.1	46.0	62.1
RetinaNet [8]	IR	88.8	35.4	39.5	76.5	32.1	54.5
Faster R-CNN [11]		89.4	53.5	48.3	87.0	42.6	64.2
Oriented R-CNN [13]		89.8	57.4	53.1	89.3	45.4	67.0
S^2A-Net [5]		89.9	54.5	55.8	88.9	48.4	67.5
RoITransformer [3]		90.1	60.4	58.9	89.7	52.2	70.3
YOLOv5s [7]		90.0	59.5	60.8	89.5	53.8	70.7
UA-CMDet [12]	RGB+IR	87.5	60.7	46.8	87.1	38.0	64.0
Halfway Fusion [9]		90.1	62.3	58.5	89.1	49.8	70.0
CIAN [16]		90.1	63.8	60.7	89.1	50.3	70.8
AR-CNN [17]		90.1	64.8	62.1	89.4	51.5	71.6
MBNet [19]		90.1	64.4	62.4	88.8	53.6	71.9
TSFADet [15]		89.9	67.9	63.7	89.8	54.0	73.1
C^2Former [14]		90.2	68.3	64.4	89.8	58.5	74.2
SLBAF-Net [2]		90.2	72.0	68.6	89.9	59.9	76.1
CALNet [6]		**90.3**	73.7	68.7	89.7	59.7	76.4
OAFA [1]		**90.3**	76.8	73.3	**90.3**	66.0	79.4
PCDF (Ours)		90.2	**77.7**	**74.4**	89.8	**66.2**	**79.7**

Table 6. Detection results (in %) on the DroneVehicle dataset. Note that all detectors locate and classify vehicles with OBB heads. Best results are highlighted in **bold**. And the second one is marked with underline.

with IR labels.

The results are listed in Tab. 6. It can be observed that our PCDF achieves the highest mAP score of 79.7%, demonstrating its superior performance compared to both unimodal and multimodal methods. In each category, PCDF achieves 90.2%, 77.7%, 74.4%, 89.8%, and 66.2% AP on the car, truck, freight-car, bus, and van categories respectively, showcasing its competitive performance across various categories. Additionally, the results indicate that even when considering only the 'day' and 'night' conditions, our method still achieves outstanding performance, underscoring the significance of condition-guided dynamic fusion.

E. Ablation Study of Different Conditions

To assess the contribution of each condition in our proposed method, we conduct an ablation study on our ATR-UMOD dataset. Fig. 12 illustrates the mAP performance under different conditions. The results reveal that each condition contributes to varying degrees of performance improvement, highlighting their respective influence on the learning of the multimodal reliability.

Among these conditions, the illumination condition has the most significant impact on performance. This observation aligns with the understanding that RGB images excel under good illumination due to their rich color and texture information, while IR images perform better under poor illumination conditions for their insensitivity to illumination variations, highlighting the critical role of illumination in multimodal fusion. Time conditions indirectly affect illumination levels and also have a positive influence on performance. Additionally, altitude leads to secondary im-

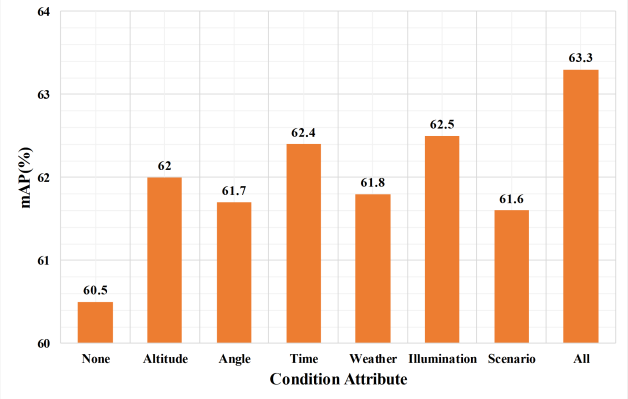

Figure 12. Ablation study of different conditions on the ATR-UMOD dataset. "None" means without any condition, and "All" means with all conditions.

provements. The underlying reason can be that at low altitude, RGB images provide more reliable information for their finer texture details. However, when comes to higher altitude, the reduced object size leads to a loss of texture details, resulting in decreased reliability. In contrast, IR images, which rely on thermal emissions, maintain a more stable detection ability, particularly for objects with notable temperature differences. For angle conditions, higher angles can lead to occlusions or reflections in RGB images, whereas IR images exhibit a certain degree of resistance to these effects. Therefore, angle conditions still affect multimodal reliability to some extent. Weather conditions moderately influence the reliability of RGB images by affecting object visibility, thereby leading to performance improvements. Lastly, scenario condition has the least impact, suggesting that spatial context alone is not a dominant factor for multimodal reliability.

The results provide insights into the interplay between different conditions and multimodal feature learning, offering guidance for future research on dynamic fusion methods. Moreover, combining all conditions yields the best performance, highlighting the importance of considering all factors comprehensively in multimodal fusion.

F. Visualization of the Intermediate Results

In PCDF, we decouple the original multimodal features into condition-specific and condition-invariant features, where the former captures condition information, and the latter is adaptively reassigned to achieve reliable feature fusion. To validate its effectiveness, we visualize intermediate feature maps in Fig. 13, including the original multimodal features, decoupled condition-invariant features, fixed fusion features with consistent multimodal contributions, and dynamic fusion features with reassigned multimodal contributions. It can be observed that the decoupled condition-irrelevant multimodal features retain more target-related information and less condition noise than the original multimodal features, which not only validates the effectiveness of our decoupling strategy but also demonstrates their capability to suppress condition-induced noise. Furthermore, the dynamic fusion features provide more reliable object representations with reduced interferences compared to their fixed-contribution counterparts. This highlights the effectiveness of our dynamic fusion method, which leverages discriminative information from the dominant modality while suppressing noise from the less-contributive modality.

G. Hyperparameter Analyses

The crucial hyperparameters in PCDF include the threshold τ in Eq. (2) in the main text. This parameter directly determines the number of selected condition attributes in

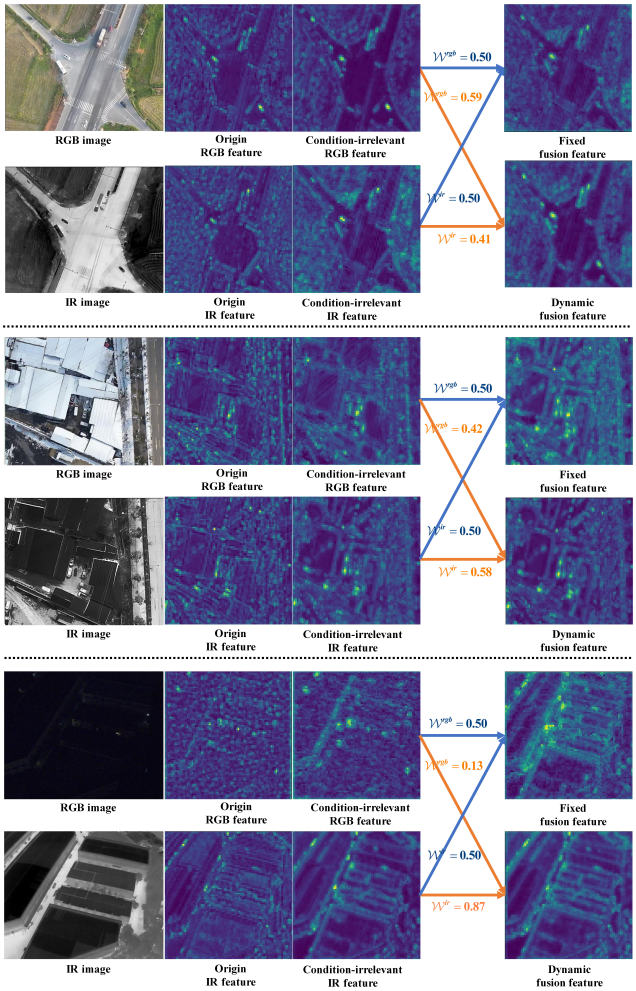

Figure 13. Two examples of the intermediate feature maps of the multimodal images, original multimodal features, decoupled condition-irrelevant features, fixed fusion features with consistent multimodal contributions, and dynamic fusion features with reassigned multimodal contributions. The blue line indicates the consistent multimodal contribution with 0.5 and 0.5, and the orange line indicates the reassigned multimodal contribution.

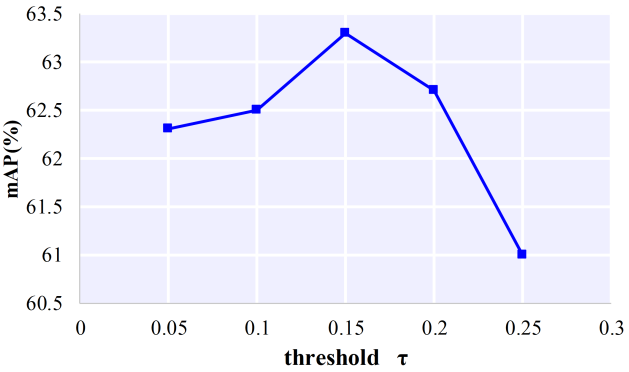

Figure 14. Performance for different values of threshold τ.

SCPL, affecting the available conditions for multimodal fusion. To investigate the impact of τ on the performance of PCDF, we conduct a series of experiments on the ATR-UMOD dataset, with results summarized in Fig. 14. It can be observed that the mAP performance initially improves before gradually declining as τ increases. This trend can be attributed to that a higher τ allows the model to focus on more relevant conditions, which can reduce the influence of noisy conditions. Despite this, excessively large values of τ may filter out valuable condition information, resulting in insufficient or even a complete lack of condition guidance, ultimately degrading performance. These findings suggest that careful tuning of this parameter is necessary to achieve optimal performance in practical applications.

H. More Visualization Results in Different Conditions

In this section, we present comprehensive visualization results under different conditions, including altitude, angle, time, weather, illumination, and scenario conditions, as displayed in Fig. 15 to Fig. 20. The confidence threshold is set to 0.25, and fusion-based results are displayed on IR images to align with supervisory labels. Missed detections and false positives are highlighted with blue and red dashed circles, respectively. These visualizations demonstrate the robustness and adaptability of our method across diverse conditions, showcasing its potential for complex real-world applications.

References

[1] Chen Chen, Jiahao Qi, Xingyue Liu, Kangcheng Bin, Ruigang Fu, Xikun Hu, and Ping Zhong. Weakly misalignment-free adaptive feature alignment for uavs-based multimodal object detection. In *CVPR*, pages 26826–26835, 2024. 5, 6

[2] Xiaolong Cheng, Keke Geng, Ziwei Wang, Jinhu Wang, Yuxiao Sun, and Pengbo Ding. Slbaf-net: Super-lightweight bimodal adaptive fusion network for uav detection in low recognition environment. *Multimedia Tools and Applications*, pages 1–20, 2023. 5, 6

[3] Jian Ding, Nan Xue, Yang Long, Gui-Song Xia, and Qikai Lu. Learning roi transformer for oriented object detection in aerial images. In *CVPR*, pages 2849–2858, 2019. 5, 6

[4] Martin A. Fischler and Robert C. Bolles. Random sample consensus: A paradigm for model fitting with applications to image analysis and automated cartography. *Communications of the ACM*, 24(6):381–395, 1981. 3

[5] Jiaming Han, Jian Ding, Jie Li, and Gui-Song Xia. Align deep features for oriented object detection. *IEEE TGRS*, 60: 1–11, 2022. 5, 6

[6] Xiao He, Chang Tang, Xin Zou, and Wei Zhang. Multispectral object detection via cross-modal conflict-aware learning. In *ACM MM*, pages 1465–1474, 2023. 5, 6

[7] Glenn Jocher. ultralytics/yolov5. https://github.com/ultralytics/yolov5, oct 2020. 5, 6

[8] Tsung-Yi Lin, Priya Goyal, Ross B. Girshick, Kaiming He, and Piotr Dollár. Focal loss for dense object detection. In *ICCV*, pages 2999–3007, 2017. 5, 6

[9] Jingjing Liu, Shaoting Zhang, Shu Wang, and Dimitris N. Metaxas. Multispectral deep neural networks for pedestrian detection. In *BMVC*, 2016. 5, 6

[10] David G. Lowe. Object recognition from local scale-invariant features. In *ICCV*, pages 1150–1157, 1999. 3

[11] Shaoqing Ren, Kaiming He, Ross B. Girshick, and Jian Sun. Faster R-CNN: towards real-time object detection with region proposal networks. *IEEE TPAMI*, 39(6):1137–1149, 2017. 5, 6

[12] Yiming Sun, Bing Cao, Pengfei Zhu, and Qinghua Hu. Drone-based rgb-infrared cross-modality vehicle detection via uncertainty-aware learning. *IEEE TCSVT*, 32(10):6700–6713, 2022. 5, 6

[13] Xingxing Xie, Gong Cheng, Jiabao Wang, Xiwen Yao, and Junwei Han. Oriented R-CNN for object detection. In *ICCV*, pages 3500–3509, 2021. 5, 6

[14] Maoxun Yuan and Xingxing Wei. C^2former: Calibrated and complementary transformer for rgb-infrared object detection. *IEEE TGRS*, 62:1–12, 2024. 5, 6

[15] Maoxun Yuan, Yinyan Wang, and Xingxing Wei. Translation, scale and rotation: Cross-modal alignment meets rgb-infrared vehicle detection. In *ECCV*, pages 509–525, 2022. 5, 6

[16] Lu Zhang, Zhiyong Liu, Shifeng Zhang, Xu Yang, Hong Qiao, Kaizhu Huang, and Amir Hussain. Cross-modality interactive attention network for multispectral pedestrian detection. *Information Fusion*, 50:20–29, 2019. 5, 6

[17] Lu Zhang, Zhiyong Liu, Xiangyu Zhu, Zhan Song, Xu Yang, Zhen Lei, and Hong Qiao. Weakly aligned feature fusion for multimodal object detection. *arXiv:2204.09848*, 2022. 5, 6

[18] Qingyang Zhang, Yake Wei, Zongbo Han, Huazhu Fu, Xi Peng, Cheng Deng, Qinghua Hu, Cai Xu, Jie Wen, Di Hu, and Changqing Zhang. Multimodal fusion on low-quality data: A comprehensive survey. *arXiv:2404.18947*, 2024. 5

[19] Kailai Zhou, Linsen Chen, and Xun Cao. Improving multispectral pedestrian detection by addressing modality imbalance problems. In *ECCV*, pages 787–803, 2020. 5, 6

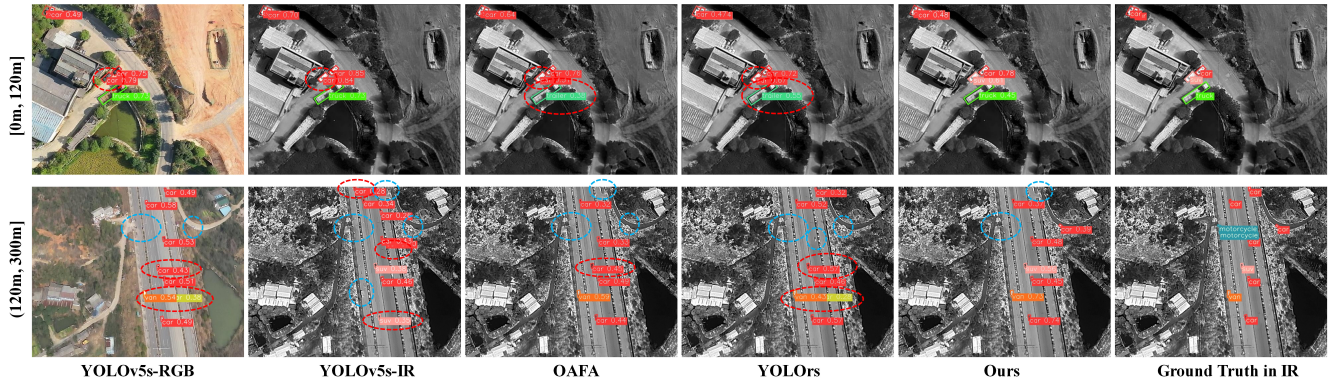

Figure 15. Results in different altitude conditions.

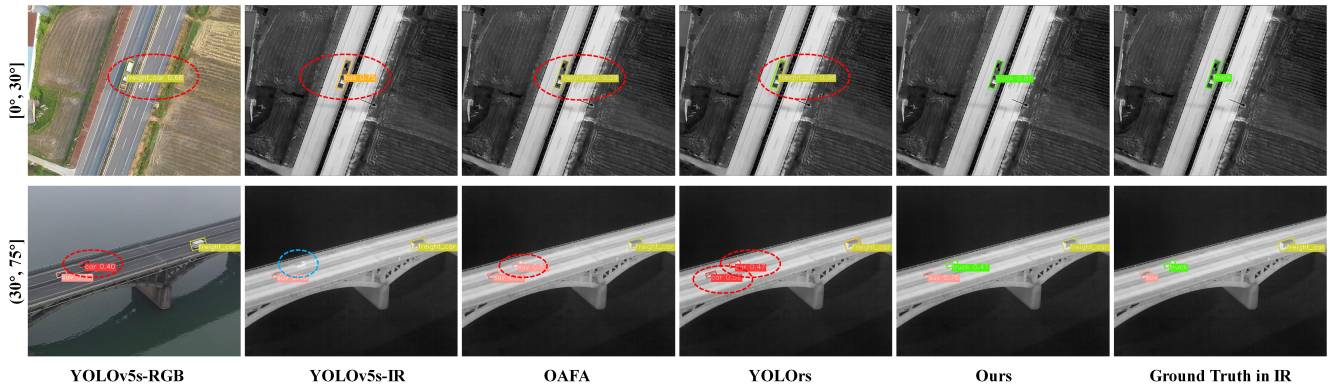

Figure 16. Results in different angle conditions.

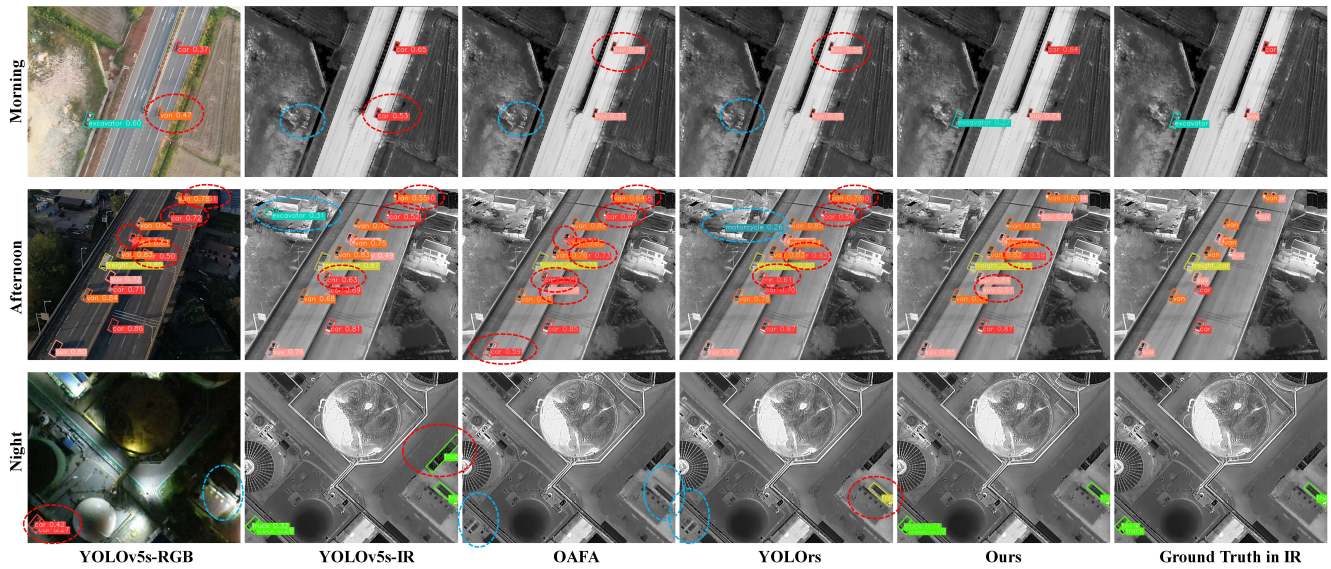

Figure 17. Results in different time conditions.

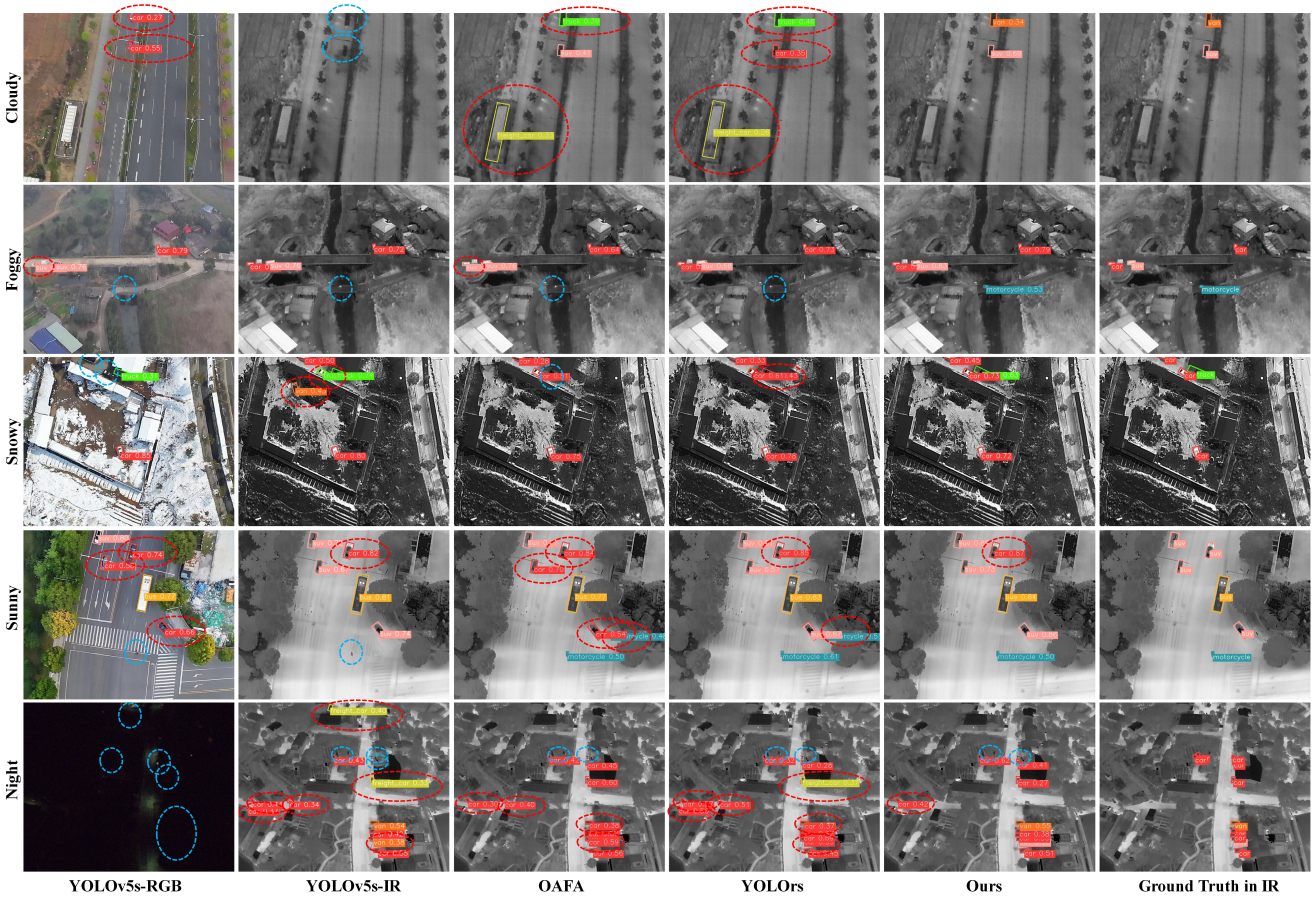

Figure 18. Results in different weather conditions.

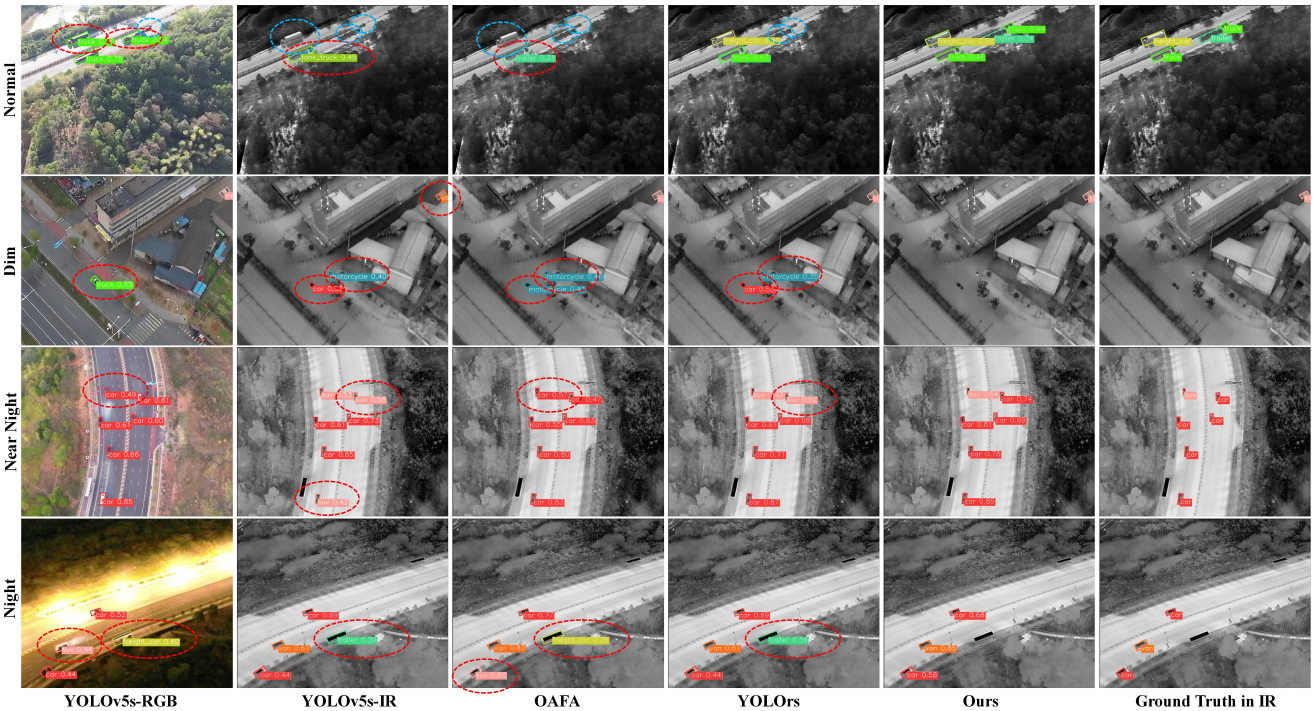

Figure 19. Results in different illumination conditions.

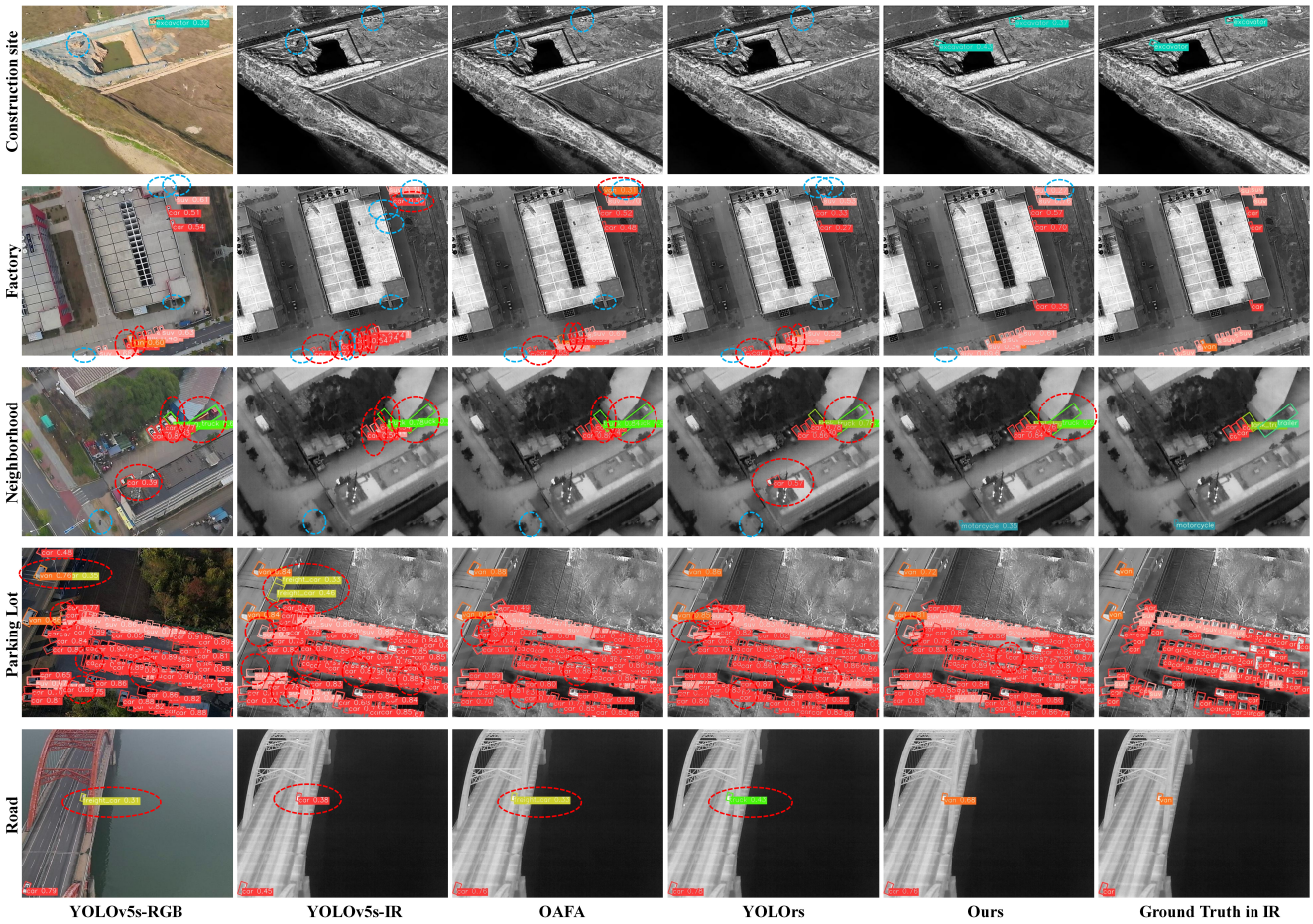

Figure 20. Results in different scenario conditions.