

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

Chen Chen Kangcheng Bin* Ting Hu Jiahao Qi Xingyue Liu
Tianpeng Liu Zhen Liu Yongxiang Liu* Ping Zhong*
National University of Defense Technology, China

{chenchen21c, binkc21, huting, qijiahao1996, liuxingyue18}@nudt.edu.cn,
{liutianpeng2004, zhen_liu, zhongping}@nudt.edu.cn, lyx_bible@sina.com

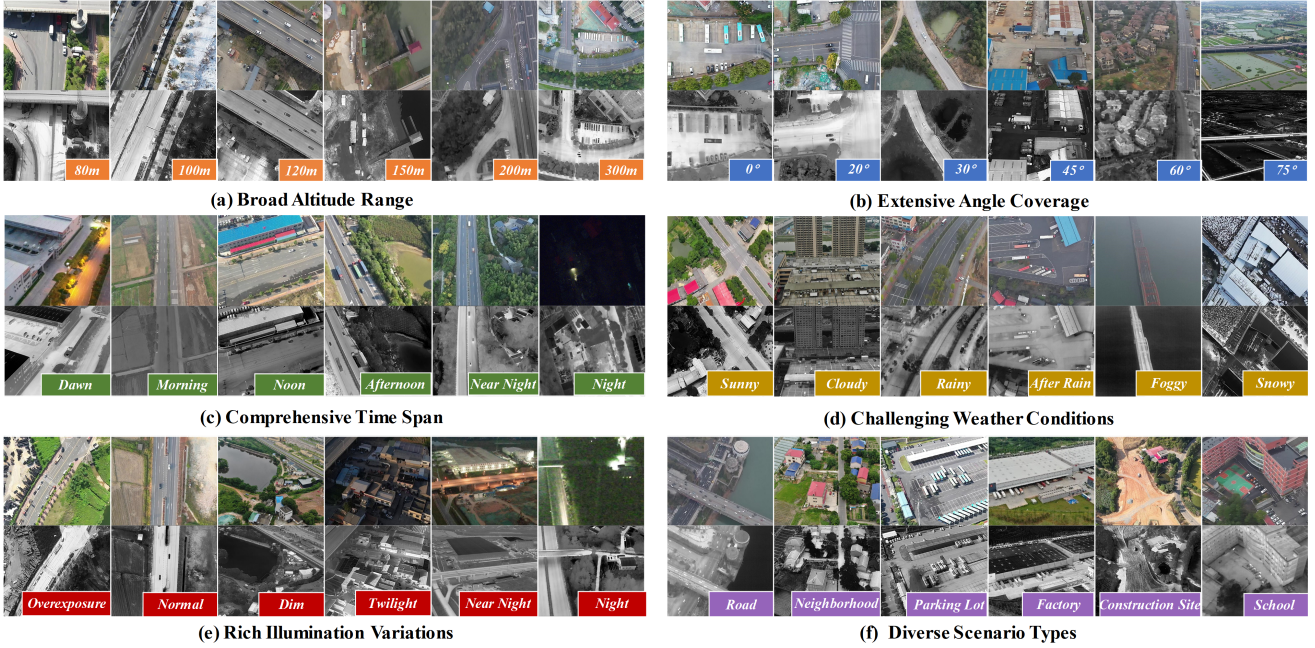


Figure 1. High-diversity imaging conditions in our ATR-UMOD. Some representative examples are shown for each condition. **(a) Broad Altitude Range:** It encompasses an altitude range from 80m to 300m, offering a rich resource for multi-scale object analysis. **(b) Extensive Angle Coverage:** Nearly full angular coverage from 0° to 75° ensures comprehensive object appearances from various viewpoints. **(c) Comprehensive Time Span:** All-day data collection captures fluctuations in light, shadows, and thermal characteristics over time. **(d) Challenging Weather Conditions:** Incorporating 7 typical and extreme weather conditions enhances robustness in real-world applications. **(e) Rich Illumination Variations:** It covers 6 illumination levels from lightless to high-light, improving adaptability to varying image qualities. **(f) Diverse Scenario Types:** Considering cross-scene generalization, it spans 11 scenarios types with complex backgrounds. These conditions are additionally annotated in each image pair, providing valuable high-level contextual insights and establishing a comprehensive benchmark for condition-specific performance evaluation.

Abstract

Unmanned aerial vehicles (UAV)-based object detection with visible (RGB) and infrared (IR) images facilitates robust around-the-clock detection, driven by advancements in deep learning techniques and the availability of high-quality dataset. However, the existing dataset struggles to fully capture real-world complexity for limited imaging conditions. To this end, we introduce a high-diversity dataset ATR-UMOD covering varying scenarios, spanning

altitudes from 80m to 300m, angles from 0° to 75°, and all-day, all-year time variations in rich weather and illumination conditions. Moreover, each RGB-IR image pair is annotated with 6 condition attributes, offering valuable high-level contextual information. To meet the challenge raised by such diverse conditions, we propose a novel prompt-guided condition-aware dynamic fusion (PCDF) to adaptively reassign multimodal contributions by leveraging annotated condition cues. By encoding imaging conditions as text prompts, PCDF effectively models the relationship between conditions and multimodal contributions through a

*Corresponding authors.

task-specific soft-gating transformation. A prompt-guided condition-decoupling module further ensures the availability in practice without condition annotations. Experiments on ATR-UMOD dataset reveal the effectiveness of PCDF.

1. Introduction

Unmanned aerial vehicle (UAV)-based object detection using visible (RGB) and infrared (IR) images (referred to as RGB-IR UOD) offers a promising solution for traffic monitoring, military reconnaissance, and so on [3, 7, 16, 23]. Its advancement heavily depends on comprehensive datasets, as modern computer vision techniques predominantly rely on a data-driven manner. DroneVehicle [30], the pioneer dataset for RGB-IR UOD, holds significant potential to facilitate progress in this field. However, it is constrained by a narrow variety of imaging conditions in altitude, angle, time, weather, illumination, and scenario, which struggles to fully represent the complexity in real-world scenarios.

To address this issue, we introduce ATR-UMOD, a novel dataset to provide more comprehensive data support for RGB-IR UOD and improve model robustness against complex real-world conditions. Compared to the existing dataset, it excels in some aspects: (1) **Diverse imaging conditions.** As illustrated in Fig. 1, it was built at flight altitudes ranging from 80m to 300m and camera angles from 0° to 75° covering all-day and all-year conditions. It also spans a wide variety of scenarios with richer weather and illumination variations, closely mirroring real-world complexities. (2) **Richer object types.** We provide 11 fine-grained object categories covering typical objects in real-world applications, supporting fine-grained detection from UAV perspectives. (3) **Extra condition annotations.** We additionally annotated 6 condition attributes for each image pair, as indicated in Fig. 2a, providing valuable high-level contextual insights and establishing a comprehensive benchmark for condition-sensitive performance evaluation.

ATR-UMOD captures the complexity of real-world conditions, but it also introduces new challenges. As shown in Fig. 2b, most existing methods underperform on ATR-UMOD, likely due to visual information bottlenecks in such complex conditions [42]. To this end, several studies have explored imaging condition cues, such as illumination, as auxiliary information [40, 44]. Motivated by this, we try to **leverage conditions as auxiliary contextual prompts** for improved detection performance across diverse conditions.

Studies in this area dynamically reassigned multimodal contributions based on imaging conditions for trustworthy fusion [37, 43]. They modeled the relationships between condition representations and multimodal contributions for dynamic fusion [11, 18, 31], enhancing effective information utilization from high-contribution modalities while mitigating noises from subordinate ones. Despite these advances, two challenges still remain: (1) **Inadequate Con-**

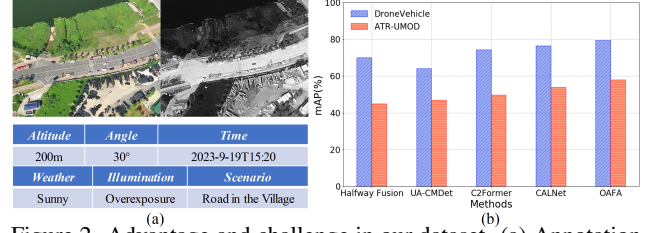


Figure 2. Advantage and challenge in our dataset. (a) Annotation example. (b) Performance degradation on ATR-UMOD.

dition Representation. They often focused solely on single condition attribute (e.g., illumination), neglecting others which also impact multimodal reliability [4]. Furthermore, condition representations are typically derived from a condition prediction model [18, 31]. In this situation, diverse condition representations require advanced multi-label prediction techniques, which are challenging due to the diversity and interdependence of condition attributes [25].

(2) **Task-irrelevant Condition-guided Pipeline.** Existing methods often depend on pretext tasks to model the relationships between conditions and multimodal contributions, such as utilizing an illumination prediction task to assign illumination values as RGB contributions [11]. The mismatch of the optimization objectives between pretext tasks and the detection task result in suboptimal multimodal contributions and ultimately compromise performance.

To this end, we propose **Prompt-guided Condition-aware Dynamic Fusion (PCDF)**, a novel method that adaptively reassigns multimodal contributions based on condition prompts, improving detection robustness across diverse conditions. Leveraging CLIP's powerful text-semantic representation capability [24], we encode multi-label conditions as text prompts to obtain expressive and robust condition representations. Considering the varying sensitivity of multimodal contributions to different condition attributes in each sample, a sample-specific condition prompt learning (SCPL) strategy is adopted to ensure relevant prompt construction. To establish task-specific relationships between conditions and multimodal contributions, we introduce a condition-aware dynamic fusion (CDF) module that refines feature reweighting through a detection-oriented normalized soft-gating transformation. Additionally, since explicit condition labels are unseen in practices, we design a prompt-guided condition-decoupling (PCD) module where condition-specific features generate prompts to dynamically modulate condition-invariant features. Extensive experiments on the ATR-UMOD dataset validate the effectiveness and robustness of PCDF under diverse conditions.

2. Related Work

2.1. RGB-IR UOD Dataset

RGB-IR UOD is a promising and emerging field, yet its datasets remain scarce with DroneVehicle [30] which has been instrumental in advancing research. Despite signifi-

Dataset	Categories	Conditions						Conditions Labeled	Pubilsh
		Altitude	Angle	Time	Weather	Illumination	Scenario		
DroneVehicle	5	80m 100m 120m	15° 30° 45°	Morning Afternoon Night	Sunny Cloudy Foggy Night	Day Night Darknight	Urban	✗	TCSVT 2022
ATR-UMOD	11	80m ~ 300m	0° ~ 75°	Dawn Morning Noon Afternoon Near Night Night	Sunny Cloudy Rainy After Rain Snowy Foggy Night	Overexposure Normal Dim Twilight Near Night Night	Urban Suburban Village	✓	ICCV 2025

Table 1. Comparison with the existing RGB-IR UOD dataset.

cant contributions, its imaging conditions are restricted by fixed flying altitudes and camera angles, limited imaging time, exclusive clear weathers, restricted illumination variations, and simple scenarios, which cannot fully capture the dynamic changes in object scales, viewpoints, and appearances, as well as the complexity of backgrounds. Additionally, only 5 object categories limits the range of potential applications and undermines the generalization ability of detection model. Finally, the lack of condition annotations prevents the exploration of conditional impacts on multimodal fusion and hinders comprehensive evaluation under diverse conditions. To address these issues, our dataset features annotated 11 object categories and 6 additional condition attributes, covering a broader range of imaging conditions across multiple dimensions, as detailed in Tab. 1, which better mirrors real-world complexities and provides a comprehensive benchmark for condition-sensitive, fine-grained detection from UAV perspectives.

2.2. Condition Representation Method

Leveraging condition representation as additional information has proven effective in computer vision tasks [1, 8, 11, 22, 31, 33]. For example, Chu *et al.* [8] pioneered use a fully connected network to model geolocation representations for fine-grained classification, yet it may fail to capture rich condition semantics due to the lack of explicit constraints. To solve this, Guan *et al.* [11] extracted illumination representations from a Day-Night prediction network for RGB-IR fusion. Wu *et al.* [31] introduced region-wise illumination prediction for finer representations. However, they only focus on a single condition, ignoring other effective condition attributes. Moreover, multi-condition representations with such prediction networks remain challenging due to the diversity and interdependence of the condition attributes. To this end, we propose a multi-condition-guided fusion method, leveraging CLIP’s robust and flexible semantic representations ability to encode multi-conditions as text prompts for effective condition representations.

2.3. Condition-guided Fusion Method

Since imaging conditions greatly affect multimodal reliabilities (*e.g.*, IR outperforms RGB in low-light conditions) [38], condition-guided fusion methods have gained increasing attention [6, 11, 18, 40]. They aimed to dynam-

cally reassign multimodal contributions based on condition-sensitive modality reliability for trustworthy fusion. Guan *et al.* [11] pioneered illumination-guided fusion by a Day-Night prediction network and directly treated Day probabilities as RGB reliabilities. To prevent modality imbalance under extreme illuminations, Zhang *et al.* [40] introduced a linear gate function to optimize reliability. IAF R-CNN [18] and IGT [6] further modeled nonlinear reliabilities with a Sigmoid function. However, all of them rely on condition prediction tasks that are misaligned with detection objectives, leading to suboptimal modality reliabilities. In contrast, we propose a detection-oriented soft-gating transformation that leverages rich-semantic condition representations to learn task-specific multimodal reliability.

3. ATR-UMOD Dataset

3.1. Dataset Construction

Data collection and object annotation. ATR-UMOD is built spanning diverse imaging conditions in flying altitude, camera angle, shooting time, weather, illumination and scenario. Due to hardware limitations, raw RGB-IR images suffer from inevitable cross-modal misalignment for differences in imaging space and time [35]. To this end, we employed homography transformation [41] and region cropping for spatial calibration and timestamp alignment for temporal calibration. For annotation, RGB and IR objects were labeled separately with oriented bounding boxes.

Attribute annotation. We enriched ATR-UMOD with detailed condition annotations, offering essential context to address visual bottlenecks and facilitate in-depth analysis of conditional impact on multimodal fusion. Specifically, we labeled 6 key condition attributes for each image pair: *Altitude, Angle, Time, Weather, Illumination, and Scenario.*

Training and testing sets. It is divided into training and testing sets with 11,850 and 1,503 image pairs, respectively. To ensure rigorous evaluation, the subsets are derived from non-overlapping scenarios. Additionally, as shown in Fig. 3a, the object distribution across each subset has been carefully balanced to minimize data bias.

3.2. Dataset Statistics

Object statistics. It contains 13,353 well-aligned RGB-IR image pairs at 640×512 resolution, covering 161,799 RGB objects and 162,253 IR objects across 11 categories.

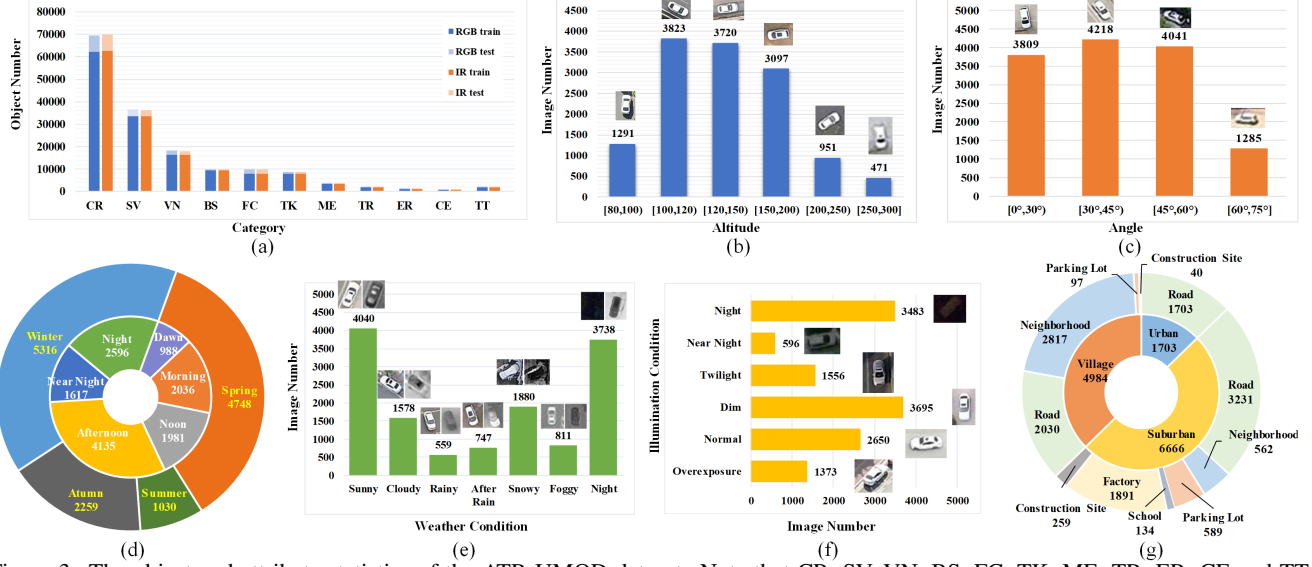


Figure 3. The object and attribute statistics of the ATR-UMOD dataset. Note that CR, SV, VN, BS, FC, TK, ME, TR, ER, CE and TT represent car, SUV, van, bus, freight car, truck, motorcycle, trailer, excavator, crane, and tank truck categories, respectively.

As depicted in Fig. 3a, it exhibits a pronounced long-tail distribution [39], with the car being the dominant category. This distribution closely reflects real-world situations but also introduces significant challenges for detection models.

Altitude statistics. Flying altitude of UAV significantly affects the object scales. According to Fig. 3b, the dataset spans altitudes from 80m to 300m, capturing substantial scale variations. This broad range of altitudes promotes detection generalization across different object scales.

Angle statistics. Angle is the camera pitch angle (from 0° to 90°) which impacts the object scale and viewpoint variation. As illustrated in Fig. 3c, the dataset spans an angle from 0° to 75°, achieving nearly full angular coverage excluding extreme situations. This wide scope enriches the dataset with comprehensive multi-view object information.

Time statistics. It records the timestamp of image acquisition including year, month, day, hour, and minutes. As shown in Fig. 3d, the dataset spans a broad temporal range from dawn to night throughout all seasons, capturing various object characteristics in all-day and all-year conditions.

Weather statistics. Textures in RGB images and thermal radiations in IR images are usually altered by varying weather conditions. As seen in Fig. 3e, the dataset contains 7 typical and extreme weather types, fostering improved detection availability in real-world applications.

Illumination statistics. As noted in Fig. 3f, images span 6 illumination levels from lightless to high-light. Since object characteristics and image quality are sensitive to illumination especially in RGB modality, this diverse illumination boosts model’s robustness in real-world situations.

Scenarios statistics. As shown in Fig. 3g, images were captured across 11 scenario types within Urban, Suburban, and Village, encompassing a wide range of environments such as Road, Neighborhood, Construction Site, Parking

Lot, and so on. High diversity of scenarios brings in complex interference from cluttered backgrounds.

3.3. Advances of ATR-UMOD Dataset

Compared with the existing RGB-IR UOD dataset, our ATR-UMOD has several unique advancements:

(1) **More diversified data distribution.** Considering limited imaging conditions, our dataset significantly enhances condition diversity in several dimensions, including broader altitude ranges, extended angle coverages, comprehensive time span, challenging weather conditions, richer illumination variations, and more complex backgrounds. These improvements allow the dataset to better reflect the complexity of real-world data distribution, making it a more comprehensive dataset for data-driven RGB-IR UOD.

(2) **Richer object types.** The ATR-UMOD dataset contains 11 object categories, whereas the existing dataset is limited to 5 categories. The increased diversity of object type not only facilitates models in capturing subtle features but also enhances their ability to recognize a wider variety of targets for more complex real-world applications.

(3) **Extra condition information.** Due to variations in multimodal image quality and object characteristics under different conditions, condition information is vital for the effectiveness of detection models. To this end, ATR-UMOD first annotates 6 key condition attributes for each image pair, enabling deeper exploration of conditional impacts on multimodal object detection and making it a comprehensive benchmark for condition-specific performance evaluation.

4. Method

4.1. Overview

Our method dynamically reassigns multimodal contributions based on multi-condition prompts. As shown in

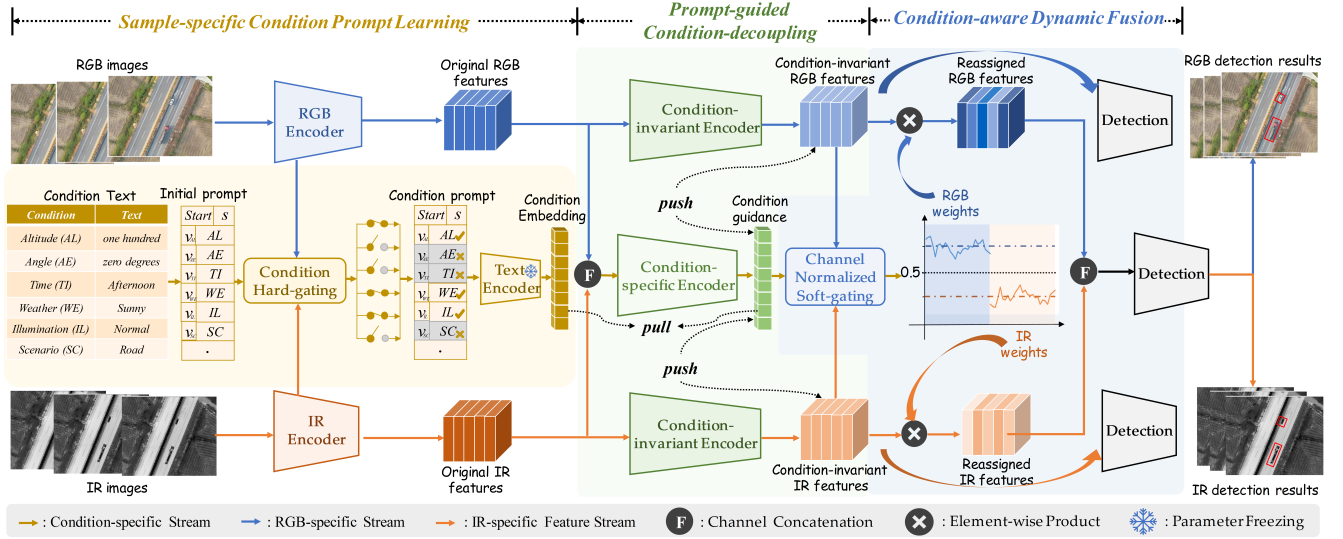


Figure 4. Overview structure of the proposed PCDF method.

Fig. 4, RGB-IR image pairs are processed through dual-branch encoders to extract unimodal original features. Simultaneously, condition texts are fed to SCPL to learn relevant condition prompts. To address inaccessible conditions in practice, unimodal original features are decoupled into condition-specific and -invariant features. The condition-specific features are aligned with condition embedding to gain condition guidance. Multimodal weights are finally obtained by this guidance for dynamically reassigning contributions of the condition-irrelevant features for detection.

4.2. Sample-specific Condition Prompt Learning

Leveraging CLIP’s powerful text representation ability and rich textual information, we encode condition semantics through prompt learning. However, different condition attributes affect individual samples to varying degrees [26], some may be negligible or even disruptive. For example, scenario is often irrelevant under night illumination. Thus, using all attributes indiscriminately as reliability cues is unreasonable. To address this, SCPL learns relevant and effective attributes for each sample.

Initial prompt construction (IPC). Given a set of condition attributes $\mathbb{A} = \{\mathcal{A}_1, \mathcal{A}_2, \dots, \mathcal{A}_N\}$, where N is the number of condition attributes and each attribute $\mathcal{A}_n$ comprises M_n distinct classes, represented as $\mathcal{A}_n = \{a_n^1, a_n^2, \dots, a_n^{M_n}\}$, we create an initial condition prompt by formatting these attributes into a fixed template. This template comprises a subject description s and several condition prefixes v_n . Details are provided in Supplementary material Sec. B.1. Taken a condition prefix with a condition attribute as a condition block $\mathcal{O}_i = \{o_1^i, o_2^i, \dots, o_N^i\}$, the initial prompt I_i for sample i can be expressed as:

$$I_i = s + \sum_{n=1}^N o_n^i, \quad o_n^i = v_n + a_n^i. \quad (1)$$

The initial condition embedding $\mathcal{I}_i$ is obtained by feeding I_i into the frozen text encoder of CLIP [24].

Sample-specific condition prompt-tuning (SCPT). To eliminate the effects of irrelevant attributes, we introduce a sample-specific prompt tuning mechanism based on hard-gating masks. Inspired by that experts assess the influence of condition attributes on multimodal reliability by observing specific patterns in each sample [21], we feed $\mathcal{I}_i$ with multimodal features into a condition hard-gating network to generate learnable sample-specific condition prompts. Precisely, multimodal features $\mathcal{F}_{rgb}^i, \mathcal{F}_{ir}^i \in \mathbb{R}^{C \times H \times W}$ are fused through nonlinear layers F_c and a Softmax function σ to yield attribute availability probabilities. The hard-gating masks $\mathcal{G}_i = \{g_1^i, g_2^i, \dots, g_N^i\}$ are then obtained by indicator function $\mathbb{1}$ with a predefined threshold τ :

$$\mathcal{G}_i = \mathbb{1}(\sigma(F_c(\text{Pool}(\mathcal{F}_{rgb}^i, \mathcal{F}_{ir}^i), \mathcal{I}_i)) \geq \tau), \quad (2)$$

where $(\cdot, \cdot)$ is concatenation, Pool is maxpooling, and τ is set to 0.15 (see in Supplementary Sec. G). $g_n^i \in \{0, 1\}$ determines whether the n -th attribute should be included or excluded. The adjusted condition block $o_n^{i'}$ is defined as:

$$o_n^{i'} = \begin{cases} o_n^i & \text{if } g_n^i = 1, \\ \emptyset, & \text{if } g_n^i = 0. \end{cases} \quad (3)$$

This gating mask is applied to the initial prompt to obtain the sample-specific condition prompt P_i :

$$P_i = I_i \times \mathcal{G}_i = s + \sum_{n=1}^N o_n^{i'}. \quad (4)$$

Finally, we transform P_i into condition embeddings $\mathcal{F}_t^i$ with CLIP. Noted that SCPL is only utilized in training.

4.3. Prompt-guided Condition-decoupling

As $\mathcal{F}_t^i$ is inaccessible in practice, condition guidance must be derived from visual features. However,

directly modeling it from original features may bring in interferences between condition and object information. Thus, we decouple original features into condition-specific and condition-invariant components, where the condition-specific features tied to condition semantics, while the condition-invariant features focus on robust object-discriminative representations.

To achieve this, we introduce a three-branch decoupling network. Specifically, the first branch is the condition-specific encoder $\mathcal{S}$ that extracts condition-specific features $\mathcal{F}^{s,i}$ from the visual features. Other branches consist of condition-invariant encoders $\mathcal{V}_m$ that independently extract condition-invariant features $\mathcal{F}_m^{v,i}$ from the unimodal features $\mathcal{F}_m^i$ ($m \in \{rgb, ir\}$). This can be formulated as:

$$\mathcal{F}^{s,i} = \mathcal{S}(F(\mathcal{F}_{rgb}^i, \mathcal{F}_{ir}^i); \theta^s), \mathcal{F}_m^{v,i} = \mathcal{V}_m(\mathcal{F}_m^i; \theta_m^v), \quad (5)$$

where θ_m^v and θ^s are the learnable parameters, $F(\cdot, \cdot)$ denotes the multimodal fusion function.

For $\mathcal{F}^{s,i}$, it is essential to ensure semantic consistency with the intended condition prompts $\mathcal{F}_t^i$. For this purpose, we adopt a prompt-guided distillation loss $\mathcal{L}_{dt}$ to minimize the distance between the $\mathcal{F}^{s,i}$ and $\mathcal{F}_t^i$, which is defined by a widely used distance metric named CMD [36]:

$$\mathcal{L}_{dt} = \frac{1}{|b-a|} \|E(\mathcal{F}^{s,i}) - E(\mathcal{F}_t^i)\|_2 + \sum_{k=2}^5 \frac{1}{|b-a|^k} \|C_k(\mathcal{F}^{s,i}) - C_k(\mathcal{F}_t^i)\|_2, \quad (6)$$

where $E(\cdot)$ is the empirical expectation vector, $C_k(\cdot)$ is the vector of k -th order sample central moments, and $[a, b]$ is the bound of the random variable $\mathcal{F}^{s,i}$ and $\mathcal{F}_t^i$.

For $\mathcal{F}_m^{v,i}$, the following properties must be satisfied: (1) it remains invariant to varying conditions; (2) it exhibits sufficient discrimination for effective object detection. As condition guidance has been modeled from $\mathcal{F}^{s,i}$, we present a irrelevant loss $\mathcal{L}_{irr}$ for property (1) that highlights the dissimilarity between $\mathcal{F}_m^{v,i}$ and $\mathcal{F}^{s,i}$. It is achieved by the squared Frobenius norm $\|\cdot\|_F^2$:

$$\mathcal{L}_{irr} = \left\| \left(F_{rgb}^{v,i} \right)^T \mathcal{F}^{s,i} \right\|_F^2 + \left\| \left(F_{ir}^{v,i} \right)^T \mathcal{F}^{s,i} \right\|_F^2. \quad (7)$$

For property (2), we introduce a discrimination loss $\mathcal{L}_{dc}$ by a detector to ensure the discriminative capacity of $\mathcal{F}_m^{v,i}$:

$$\mathcal{L}_{dc} = \sum_{m \in \{rgb, ir\}} (\mathcal{L}_{cls}(F_m^{v,i}) + \mathcal{L}_{reg}(F_m^{v,i}) + \mathcal{L}_{obj}(F_m^{v,i})), \quad (8)$$

where $\mathcal{L}_{cls}$, $\mathcal{L}_{reg}$, and $\mathcal{L}_{obj}$ are the classification, regression, and objectness loss, respectively. Finally, the decoupling loss $\mathcal{L}_{dec}$ can be formulated as:

$$\mathcal{L}_{dec} = \lambda_1 \mathcal{L}_{dt} + \lambda_2 \mathcal{L}_{irr} + \lambda_3 \mathcal{L}_{dc}, \quad (9)$$

where λ_i is the trade-off parameter that is experimentally set to 0.01, 0.003, and 0.01 respectively in this study.

4.4. Condition-aware Dynamic Fusion

The multimodal reliability is determined by the condition guidance $\mathcal{F}^{s,i}$. Given that different channels capture distinct semantic aspects [28], we introduce a channel-wise normalized soft-gating transformation to enhance model adaptability. In detail, it adaptively maps $\mathcal{F}^{s,i}$ to multimodal weights $\mathcal{W}_m^i \in \mathbb{R}^{1 \times C}$ via a nonlinear projection function $\mathbf{F}_t$ followed by a channel-wise normalized operation, ensuring information preservation in fusion features while constraining weights within $[0, 1]$:

$$\mathcal{W}_m^i = \frac{\exp([\mathbf{F}_t(\mathcal{F}^{s,i})]_m)}{\exp([\mathbf{F}_t(\mathcal{F}^{s,i})]_{rgb}) + \exp([\mathbf{F}_t(\mathcal{F}^{s,i})]_{ir})}, \quad (10)$$

where $[\cdot]_m$ represents the channels of m modality. These weights are applied to *condition-invariant features* $\mathcal{F}_m^{v,i}$ to dynamically adjust multimodal contributions. Notably, only $\mathcal{F}_m^{v,i}$ are reassigned for the fusion process, mitigating interference of condition-induced noise. The final fused feature $\mathcal{F}_f^i$ is obtained through a simple concatenation operation:

$$\mathcal{F}_f^i = \text{Concat}(\mathcal{W}_{rgb}^i \odot \mathcal{F}_{rgb}^{v,i}, \mathcal{W}_{ir}^i \odot \mathcal{F}_{ir}^{v,i}), \quad (11)$$

where $\odot$ denotes the element-wise multiplication. $\mathcal{F}_f^i$ is fed into detection head for task-oriented reliability learning. This dynamic fusion adaptively leverages discriminative information from the dominant modality while suppressing contributions from the suboptimal one.

5. Experiments

5.1. Implementation Details

Our method was implemented in PyTorch on an NVIDIA RTX 4090 GPU. The network parameters were updated using SGD [2] optimizer with an initial learning rate of 0.01 and decayed exponentially. Momentum and weight decay were set to 0.937 and 0.0005, respectively. We utilized the ViT-B/16 [10] pre-trained model from CLIP as a text encoder. Our model comprises two trainable processes, including a fusion network with SCPL and the full pipeline, both trained for 50 epochs with an 640×512 image size and a batch size of 16. All baseline methods were trained with their original parameter settings to ensure optimal performance. The Mean Average Precision (mAP) is adopted to evaluate the detection performance with an IoU of 0.5.

5.2. Results Comparisons

We evaluate PCDF on the ATR-UMOD dataset through comprehensive qualitative and quantitative analyses, benchmarking against 7 state-of-the-art (SOTA) *unimodal detectors*, including RetinaNet [19], S²A-Net [13], Faster R-CNN [27], ReDet [12], RoITransformer [9], Oriented R-CNN [32], and YOLOv5s [17], as well as 8 *multimodal detectors*, including IAF R-CNN [18], Halfway Fusion [20],

Detectors	Modality	CR	SV	VN	BS	FC	TK	TT	TR	CE	ER	ME	mAP (%) ↑
RetinaNet [19]	RGB	26.3	29.9	32.9	59.6	17.9	23.1	2.1	4.7	9.3	17.8	14.1	21.6
S ² A-Net [13]		34.2	44.9	45.9	69.2	24.4	37.4	5.6	22.5	49.5	31.2	25.6	35.5
Faster R-CNN [27]		35.2	49.0	48.7	69.4	26.8	44.0	17.6	35.0	55.3	36.6	22.6	40.0
ReDet [12]		36.9	52.5	51.6	74.8	33.5	48.1	16.7	40.7	61.4	36.8	32.9	41.1
RoI Transformer [9]		37.2	53.3	51.9	71.5	30.1	46.8	18.2	36.3	58.9	38.3	25.3	42.5
Oriented R-CNN [32]		36.9	52.5	51.6	74.8	33.5	48.1	16.7	40.7	61.7	36.8	32.9	44.2
YOLOv5s [17]		45.8	60.7	57.5	75.2	41.6	52.1	18.2	42.3	<u>68.7</u>	<u>47.5</u>	47.4	50.7
RetinaNet [19]	IR	39.3	29.1	20.8	48.9	24.7	13.1	6.4	1.1	6.2	1.0	18.4	18.9
S ² A-Net [13]		50.2	35.9	31.8	59.9	35.5	24.3	31.4	16.0	10.8	1.0	32.0	29.9
Faster R-CNN [27]		53.4	39.0	35.6	64.9	37.3	28.0	33.4	25.7	34.1	11.9	17.6	34.7
ReDet [12]		57.4	42.6	38.8	70.4	42.3	31.5	52.0	15.9	33.7	8.7	23.1	37.9
RoI Transformer [9]		54.6	41.8	38.7	64.0	43.1	33.6	61.0	23.4	32.8	7.0	23.4	38.5
Oriented R-CNN [32]		57.5	41.6	36.8	63.8	43.5	28.6	64.3	28.5	44.2	6.9	23.9	40.0
YOLOv5s [17]		65.8	51.2	51.6	75.3	53.1	38.9	<u>83.3</u>	46.2	57.9	12.0	42.7	52.5
IAF R-CNN [18]	RGB+IR	51.9	45.9	48.3	64.6	37.8	42.6	30.4	30.8	43.8	43.5	20.8	41.9
Halfway Fusion [20]		53.1	47.0	51.3	73.5	42.1	42.4	39.5	34.4	52.5	35.3	22.9	44.9
UA-CMDet [30]		50.9	43.3	47.9	75.8	51.4	44.5	42.8	40.1	54.8	39.6	23.2	46.8
C ² Former [34]		60.5	53.3	51.6	81.6	46.1	44.7	46.6	29.3	56.3	36.8	40.0	49.7
TINet [40]		60.2	51.4	54.4	74.5	50.2	46.0	44.6	39.7	59.0	<u>47.5</u>	27.0	50.4
CALNet [14]		<u>71.9</u>	65.5	71.0	78.4	53.6	51.2	37.7	35.3	56.3	31.9	38.6	53.8
OAFA [5]		70.4	59.6	63.1	81.5	60.1	47.5	80.1	32.4	59.0	33.0	50.1	57.9
YOLOrs [29]		73.2	<u>62.6</u>	<u>66.3</u>	<u>81.8</u>	<u>61.1</u>	48.2	70.3	37.6	64.3	41.8	52.9	60.0
PCDF (Ours)		70.8	60.6	65.4	84.3	62.1	<u>51.3</u>	86.1	<u>42.5</u>	71.1	49.0	<u>51.2</u>	63.1

Table 2. Detection results (in %) on the ATR-UMOD dataset. All detectors perform object localization and classification with OBB heads. Best results are marked with **bold**, while the second one is highlighted in underline.

UA-CMDet [30], C²Former [34], TINet [40], CALNet [14], OAFA [5], and YOLOrs [29]. Among these methods, IAF R-CNN and TINet are illumination-guided fusion methods. Our baseline is a one-stage dual-stream detector that integrates two modalities through concatenation fusion. Noted that the multimodal detectors are all trained with IR labels.

Quantitative comparison. The quantitative comparisons are presented in Tab. 2. The mAP results demonstrate that PCDF significantly outperforms the SOTA unimodal and multimodal methods, surpassing the second-best method by 3.1%. Moreover, PCDF consistently excels across multiple categories, achieving the best or second-best performance in most cases while maintaining competitive results in the remaining ones. This demonstrates the effectiveness of our approach in dynamically leveraging reliable information from both RGB and IR modalities, thereby enhancing detection performance.

Qualitative comparison. Fig. 5 provides qualitative comparisons across typical conditions among the SOTA unimodal model, SOTA multimodal models, and PCDF. In Overexposure (first row), Night (second row), and Snowy (third row) conditions, RGB and IR modalities exhibit distinct reliabilities. Unimodal methods struggle with excessive exposure, low visibility, and occlusion in RGB images, as well as insufficient information in IR images, leading to detection failures. Fusion methods also fail to handle these challenging conditions effectively due to their rigid fusion strategies. In contrast, our method dynamically exploits the complementarity of RGB and IR modalities, achieving superior detection performance.

5.3. Results on Different Conditions

To assess the effectiveness of our method across varying conditions, we conduct comprehensive experiments on

Conditions		Method			
		CALNet	OAFA	YOLOrs	Ours
AL	[0, 120]	55.3	59.4	61.8	67.0
	(120, 300]	41.4	47.3	48.3	50.3
AN	[0, 30]	51.0	58.2	58.5	60.4
	(30, 75]	43.1	51.7	55.0	57.2
TI	Morning	43.9	53.6	55.6	58.9
	Afternoon	53.8	60.0	62.7	66.3
	Night	31.2	38.2	35.3	41.7
WE	Cloudy	54.8	62.1	69.6	71.2
	Foggy	28.8	30.9	33.2	32.6
	Snowy	38.2	48.4	50.1	53.4
	Sunny	48.7	54.6	57.1	60.1
	Night	32.5	38.0	35.4	41.3
IL	Normal	49.8	55.0	56.9	60.6
	Dim	52.5	59.8	67.5	67.7
	Near Night	41.5	44.8	47.5	48.1
	Night	31.5	37.3	34.6	40.0
SC	Construction site	41.0	42.3	47.2	49.0
	Factory	17.6	23.0	22.4	23.0
	Neighborhood	33.7	34.6	31.9	35.3
	Parking Lot	31.9	39.5	38.0	40.8
	Road	49.6	57.0	59.3	62.3

Table 3. Detection results (in %) in different conditions. Noted that AL, AN, TI, WE, IL, and SC represent altitude, angle, time, weather, illumination, and scenario, respectively. Best results are marked with **bold**.

ATR-UMOD dataset. Tab. 3 presents the detection results under different conditions in the SOTA multimodal methods and our PCDF. Due to the excessive number of conditions, sample size for each condition was often insufficient, causing overfitting and impairing model training. Therefore, conditions were appropriately merged in Tab. 3. Details are provided in Supplementary material Sec. B.2. The results indicate that PCDF achieves superior performance across nearly all conditions, demonstrating its robustness and adaptability in diverse conditions. The suboptimal performance in “Foggy” condition may be attributed to inconsistencies in fog levels and visibility, which can be better addressed through fine classification in future work.

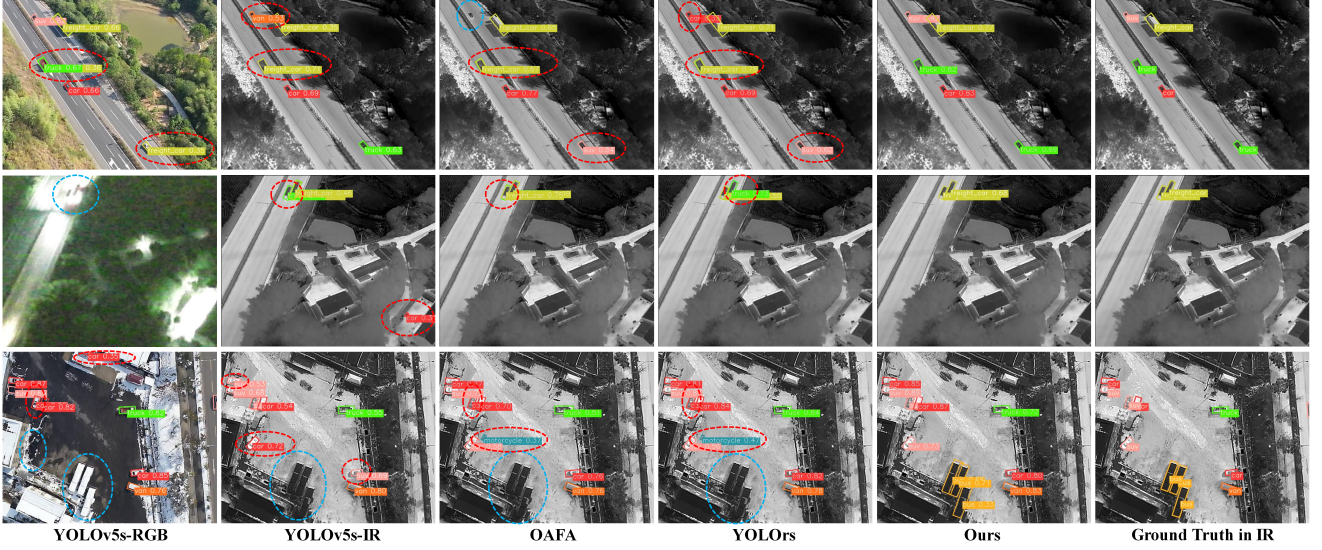


Figure 5. Qualitative comparison on ATR-UMOD dataset. Confidence threshold is 0.25. Fusion method results are displayed on IR images to align with the supervisory labels. Missed and incorrectly detected objects are indicated with blue and red dashed circles, respectively.

5.4. Ablation Study

Effectiveness of SCPL. This module is designed to adaptively construct effective prompts with relevant conditions. To assess its impact, we conduct two ablation experiments. (1) *w/o SCPT*: prompts are constructed solely using the initial prompts. The performance drop in Tab. 4 suggests that without SCPT, model captures unnecessary condition semantics while diluting the influence of the meaningful ones, leading to unreliable condition representations. (2) *w/o SCPL*: since SCPL is the foundation of PCD, we replace condition guidance with data guidance by applying channel attention [15] for dynamic fusion. The performance improvement over the baseline highlights the significance of dynamic fusion. However, its performance is still inferior to PCDF, underscoring the essential role of condition-based information in mitigating multimodal reliance bias.

Effectiveness of PCD. It enables PCDF to test with condition information without condition labels. Tab. 4 shows a mAP drop of 1.2% when PCD is removed. The reason lies in that PCD mitigates condition-induced noise interference by decoupling condition-irrelevant features, improving generalization across varying conditions. Moreover, *w/o L_{dt} , L_{irr} , or L_{dc}* in PCD also result in varying degrees of performance degradation, underscoring their roles in maintaining semantic consistency between the condition guidance and condition-specific features, separating condition-irrelevant and specific features, and enhancing the discriminability of condition-irrelevant features, respectively.

Effectiveness of CDF. It aims to dynamically reassign multimodal contributions in response to condition variations. Ablation studies were conducted by replacing CDF with simple fusion that integrate condition features into multimodal visual features via addition or concatenation. The results reveal a notable performance decline, which can

be attributed to the lack of direct relationship perception between conditions and multimodal contributions while introducing condition noise into the fusion process.

Module Name	Experimental Design	mAP (%) $\uparrow$
Baseline	N/A	58.4
SCPL	w/o SCPT	62.3
	w/o SCPL	60.5
PCD	w/o L_{dt}	61.6
	w/o L_{irr}	62.7
	w/o L_{dc}	62.0
	w/o PCD	62.1
CDF	w/o CDF (add)	62.0
	w/o CDF (concat)	61.5
PCDF	N/A	63.1

Table 4. Ablation study on PCDF. “w/o” means without.

6. Conclusion

In this paper, we built a high-diversity RGB-IR UOD dataset featuring fine-grained object types, broad altitude ranges, extensive angle coverage, comprehensive time span, challenging weather conditions, rich illumination variations, diverse scenario types, and additional condition annotations. Recognizing visual information bottlenecks in such diverse conditions, we incorporate conditions as contextual prompts for dynamically reassigning multimodal features. Leveraging CLIP’s powerful semantic representations, we construct sample-specific condition prompts and design a soft-gating transformation to establish task-specific relationships between prompts and multimodal contributions. A condition-decoupling mechanism enables testing without condition annotations. Experiments on ATR-UMOD dataset validate the SOTA performance of our method.

Acknowledgements. This work was supported by the National Natural Science Foundation of China (62201586).

References

- [1] Oisín Mac Aodha, Elijah Cole, and Pietro Perona. Presence-only geographical priors for fine-grained image classification. In *ICCV*, pages 9595–9605, 2019. 3
- [2] Léon Bottou. Stochastic gradient descent tricks. In *Neural Networks: Tricks of the Trade*, pages 421–436, 2012. 6
- [3] Ilker Bozcan and Erdal Kayacan. AU-AIR: A multi-modal unmanned aerial vehicle dataset for low altitude traffic surveillance. In *ICRA*, pages 8504–8510, 2020. 2
- [4] Tim Broedermann, Christos Sakaridis, Yuqian Fu, and Luc Van Gool. Condition-aware multimodal fusion for robust semantic perception of driving scenes. *arXiv:2410.10791*, 2024. 2
- [5] 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. 7
- [6] Keyu Chen, Jinqiang Liu, and Han Zhang. IGT: illumination-guided RGB-T object detection with transformers. *Knowledge-Based Systems*, 268:110423, 2023. 3
- [7] Xiwen Chen, Bryce Hopkins, Hao Wang, Leo O’Neill, Fatemeh Afghah, Abolfazl Razi, Peter Z. Fulé, Janice Coen, Eric Rowell, and Adam C. Watts. Wildland fire detection and monitoring using a drone-collected RGB/IR image dataset. In *IEEE AIPR*, pages 1–4, 2022. 2
- [8] Grace Chu, Brian Potetz, Weijun Wang, Andrew Howard, Yang Song, Fernando Brucher, Thomas Leung, and Hartwig Adam. Geo-aware networks for fine-grained recognition. In *ICCV*, pages 247–254, 2019. 3
- [9] 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. 6, 7
- [10] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In *ICLR*, 2021. 6
- [11] Dayan Guan, Yanpeng Cao, Jiangxin Yang, Yanlong Cao, and Michael Ying Yang. Fusion of multispectral data through illumination-aware deep neural networks for pedestrian detection. *Information Fusion*, 50:148–157, 2019. 2, 3
- [12] Jiaming Han, Jian Ding, Nan Xue, and Gui-Song Xia. Redet: A rotation-equivariant detector for aerial object detection. In *CVPR*, pages 2786–2795, 2021. 6, 7
- [13] Jiaming Han, Jian Ding, Jie Li, and Gui-Song Xia. Align deep features for oriented object detection. *IEEE TGRS*, 60: 1–11, 2022. 6, 7
- [14] 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. 7
- [15] Jie Hu, Li Shen, and Gang Sun. Squeeze-and-excitation networks. In *CVPR*, pages 7132–7141, 2018. 8
- [16] Yingying Ji, Kechen Song, Hongwei Wen, Xiaotong Xue, Yunhui Yan, and Qinggang Meng. UAV applications in intelligent traffic: RGBT image feature registration and complementary perception. *Advanced Engineering Informatics*, 63:102953, 2025. 2
- [17] Glenn Jocher. ultralytics/yolov5. <https://github.com/ultralytics/yolov5>, oct 2020. 6, 7
- [18] Chengyang Li, Dan Song, Ruofeng Tong, and Min Tang. Illumination-aware faster R-CNN for robust multispectral pedestrian detection. *PR*, 85:161–171, 2019. 2, 3, 6, 7
- [19] 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. 6, 7
- [20] Jingjing Liu, Shaoting Zhang, Shu Wang, and Dimitris N. Metaxas. Multispectral deep neural networks for pedestrian detection. In *BMVC*, 2016. 6, 7
- [21] Tianshan Liu, Kin-Man Lam, Rui Zhao, and Guoping Qiu. Deep cross-modal representation learning and distillation for illumination-invariant pedestrian detection. *IEEE TCSVT*, 32(1):315–329, 2022. 5
- [22] Ru Luo, Qishan He, Lingjun Zhao, Siqian Zhang, Gangyao Kuang, and Kefeng Ji. Geospatial contextual prior-enabled knowledge reasoning framework for fine-grained aircraft detection in panoramic SAR imagery. *IEEE TGRS*, 62:1–13, 2024. 3
- [23] Juncheng Ma, Binhui Liu, Lin Ji, Zhicheng Zhu, Yongfeng Wu, and Weihua Jiao. Field-scale yield prediction of winter wheat under different irrigation regimes based on dynamic fusion of multimodal UAV imagery. *International Journal of Applied Earth Observation and Geoinformation*, 118: 103292, 2023. 2
- [24] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *ICML*, pages 8748–8763, 2021. 2, 5
- [25] Jesse Read, Luca Martino, and Jaakko Hollmén. Multi-label methods for prediction with sequential data. *PR*, 63:45–55, 2017. 2
- [26] Lei Ren, Huilin Yin, Wancheng Ge, and Qian Meng. Environment influences on uncertainty of object detection for automated driving systems. In *International Congress on Image and Signal Processing, BioMedical Engineering and Informatics*, pages 1–5, 2019. 5
- [27] 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. 6, 7
- [28] Ramprasaath R. Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. Grad-cam: Visual explanations from deep networks via gradient-based localization. *IJCV*, 128(2):336–359, 2020. 6
- [29] Manish Sharma, Mayur Dhanaraj, Srivallabha Karnam, Dimitris G. Chachlakis, Raymond W. Ptucha, Panos P. Markopoulos, and Eli Saber. Yolors: Object detection in multimodal remote sensing imagery. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 14:1497–1508, 2021. 7

- [30] 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. [2](#), [7](#)
- [31] Jiawen Wu, Tao Shen, Qingwang Wang, Zhimin Tao, Kai Zeng, and Jian Song. Local adaptive illumination-driven input-level fusion for infrared and visible object detection. *Remote Sensing*, 15(3):660, 2023. [2](#), [3](#)
- [32] Xingxing Xie, Gong Cheng, Jiabao Wang, Xiwen Yao, and Junwei Han. Oriented R-CNN for object detection. In *ICCV*, pages 3500–3509, 2021. [6](#), [7](#)
- [33] Lingfeng Yang, Xiang Li, Renjie Song, Borui Zhao, Juntian Tao, Shihao Zhou, Jiajun Liang, and Jian Yang. Dynamic MLP for fine-grained image classification by leveraging geographical and temporal information. In *CVPR*, pages 10935–10944, 2022. [3](#)
- [34] Maoxun Yuan and Xingxing Wei. C²former: Calibrated and complementary transformer for rgb-infrared object detection. *IEEE TGRS*, 62:1–12, 2024. [7](#)
- [35] 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. [3](#)
- [36] Werner Zellinger, Thomas Grubinger, Edwin Lughofer, Thomas Natschläger, and Susanne Saminger-Platz. Central moment discrepancy (CMD) for domain-invariant representation learning. In *ICLR*, 2017. [6](#)
- [37] Qingyang Zhang, Haitao Wu, Changqing Zhang, Qinghua Hu, Huazhu Fu, Joey Tianyi Zhou, and Xi Peng. Provable dynamic fusion for low-quality multimodal data. In *ICML*, pages 41753–41769, 2023. [2](#)
- [38] 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. [3](#)
- [39] Yifan Zhang, Bingyi Kang, Bryan Hooi, Shuicheng Yan, and Jiashi Feng. Deep long-tailed learning: A survey. *IEEE TPAMI*, 45(9):10795–10816, 2023. [4](#)
- [40] Yan Zhang, Huai Yu, Yujie He, Xinya Wang, and Wen Yang. Illumination-guided RGBT object detection with inter- and intra-modality fusion. *IEEE Transactions on Instrumentation and Measurement*, 72:1–13, 2023. [2](#), [3](#), [7](#)
- [41] Zhengyou Zhang. A flexible new technique for camera calibration. *IEEE TPAMI*, 22(11):1330–1334, 2000. [3](#)
- [42] Zixiang Zhao, Lilun Deng, Haowen Bai, Yukun Cui, Zhipeng Zhang, Yulun Zhang, Haotong Qin, Dongdong Chen, Jianshe Zhang, Peng Wang, and Luc Van Gool. Image fusion via vision-language model. In *ICML*, 2024. [2](#)
- [43] Xiao Zheng, Chang Tang, Zhiguo Wan, Chengyu Hu, and Wei Zhang. Multi-level confidence learning for trustworthy multimodal classification. In *AAAI*, pages 11381–11389, 2023. [2](#)
- [44] Kailai Zhou, Linsen Chen, and Xun Cao. Improving multi-spectral pedestrian detection by addressing modality imbalance problems. In *ECCV*, pages 787–803, 2020. [2](#)