Infrared-Privileged UAV Detection via Cross-Modal Vector-Quantization

Zhibo Lou*, Ruijie Zhang*, Zeyu Luo, Qianxi Cao,
Feng Qian, Junjie Chen†, Yuming Fang

Jiangxi University of Finance and Economics, Nanchang, China
{2202320618, 2202303430, 2202320619, 2202426517}@stu.jxufe.edu.cn,
{qianfeng, chenjunjie}@jxufe.edu.cn, fa0001ng@e.ntu.edu.sg

Abstract

RGB and infrared images has shown remarkable robustness for object detection based on unmanned aerial vehicles (UAV). However, the primitive RGB and infrared (IR) images are inevitably misaligned due to the device gap between RGB and infrared cameras. Most existing methods rely on manually filtered and aligned images, and thus are limited in real-world application. Some recent methods tend to directly learn from misaligned images, which only weakly benefit from the multi-modality and may be misled by dramatically misaligned IR images. Considering that the manually aligned images are available during training while unavailable in inference, we explore a new learning paradigm using the IR modality as privileged information. In the training stage, our model learns to hallucinate the complementary knowledge in IR modality based on RGB modality. In inference, our model could hallucinate the complementary IR modality to facilitate UAV detection. Specifically, we propose to quantize the IR features and hallucinate the codebook-indices based on RGB features, which is more effective and robust than directly hallucinating features. In addition, we propose to hierarchically hallucinate multi-scale codebook-indices, which could further improve the hallucinating quality. Experiments on DroneVehicle and VisDrone datasets demonstrate the effectiveness of our method.

Code —
https://github.com/chenbys/InfraredPrivilegedUAV

Extended version — https://github.com/chenbys/InfraredPrivilegedUAV/Extended.pdf

1 Introduction

Multimodal UAV object detection seeks to leverage data from heterogeneous sensors to achieve precise target recognition and localization. Visible-light imagery offers rich, detailed information under normal illumination conditions, whereas infrared imagery can capture abundant object cues even in complete darkness. Consequently, the fusion of visible (RGB) and infrared (IR) images has been widely adopted across tasks such as classification (Liu et al. 2024), semantic segmentation (Guo et al. 2021), object detection (El Ahmar et al. 2023), and pedestrian detection (Rathinam et al. 2024).

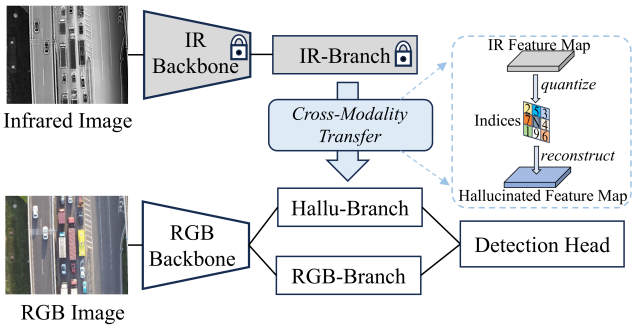

Figure 1: The overview of our framework. In the training stage, the pre-trained IR-branch produces codebook indices as hallucination target, which embeds the complementary information in IR modality. Our RGB detector learns to hallucinate the indices and obtain the complementary information in IR modality via Hallu-Branch. In the testing stage, without IR input, our RGB detector fuses the features from Hallu-Branch and RGB-Branch to predict results.

Despite the effectiveness of RGB-IR modality, the RGB-IR images are captured by different cameras, and thus suffer from the pixel misalignment issue. Existing datasets (Hwang et al. 2015; Zhang et al. 2019; Teledyne FLIR Systems 2018) usually employ manual filtering and fixed transformation to obtain aligned RGB-IR images, but the misalignment still exists due to the various capture attitudes and angles. Moreover, the manual filtering and fixed transformation may be available in the training stage, but not always available in extensive real-world applications. To this end, some recent methods (Zhang et al. 2019, 2025; Wang et al. 2019) tend to directly learn from misaligned images and focus on designing various weakly-align modules. Nevertheless, this paradigm only weakly benefits from the multi-modality, and may be misled by dramatically misaligned IR images.

Considering that the aligned IR images could be available in training stage while unavailable in inference, we explore a new paradigm for UAV detection, which employs IR modality as a kind of privileged information (Vapnik and Vashist 2009; Chen, Niu, and Zhang 2021; Zhang et al. 2021). In the

*These authors contributed equally.
†Corresponding author

task of infrared-privileged UAV detection, the major issue is to transfer complementary knowledge from IR modality to RGB modality in the training stage. As a consequence, in the test stage without IR modality, the model could better infer based on both RGB knowledge and transferred knowledge. Besides, the IR camera is far less prevailing than RGB camera, due to the device cost, imaging quality, etc. Therefore, the model trained in infrared-privileged paradigm has wider applications in real-world scenarios.

For the practical task of infrared-privileged UAV detection, we propose a novel framework based on cross-modal vector-quantization, as shown in Fig. 1. In the training stage, a pre-trained IR detector employs IR-branch to provide the vector-quantization indices, which represent rich semantic knowledge in IR modality. Our RGB detector consists of two branches, where RGB branch extracts RGB semantic features and Hallu-branch is supervised to hallucinate the vector-quantization indices in IR modality. Based on the hallucinated indices, the IR semantic features could be robustly reconstructed according to codebook. After that, our RGB detector fuses RGB features and reconstructed IR features to estimate detection results. In the test stage, without IR input and IR-branch, our RGB detector hallucinates VQ indices and could benefit from both RGB and IR semantic features. Besides, most modern detectors benefit from hierarchical multi-scale feature maps, and thus we propose to quantize them in a hierarchical residual manner, which could improve the hallucinating quality.

We conducted extensive experiments on two benchmark datasets, *i.e.*, DroneVehicle (Sun et al. 2021) and VisDrone (Zhu et al. 2018). The comprehensive comparison and in-depth analysis could demonstrate the effectiveness of our method. Our contributions could be summarized as follows: (1) We explore a new learning paradigm for UAV detection, which is more practical in the wide applications where aligned IR images are unavailable. (2) We propose a novel framework (named as CMVQ) based on cross-modality vector-quantization, which quantizes the IR features and hallucinates the codebook-indices based on RGB features. We also propose to hierarchically hallucinate multi-scale codebook-indices. (3) Extensive experiments on DroneVehicle and VisDrone datasets demonstrate the advantages of our method.

2 Related Works

2.1 RGB-IR UAV Detection

Multimodal fusion methods fully exploit the complementary advantages of infrared (IR) and visible (RGB) sensors to enhance UAV-based object detection performance. Over the years, RGB–IR UAV detection has played a pivotal role across various applications (Shen et al. 2023; Yang, Ma, and Zakhor 2022; Zhao et al. 2024; Goecks, Woods, and Valasek 2020). For example, (Hu et al. 2025) designed a dedicated fusion network for agricultural straw-fire scenarios, providing benchmarks and methodologies for disaster-specific monitoring; (Speth et al. 2022) integrated RGB and thermal imaging sensors for disaster monitoring, achieving high-recall detection in extreme events such as fires and floods; (Wang et al. 2024) introduced a lightweight cross-modal attention framework for real-time UAV detection in low-visibility conditions.Regarding how to combine different modalities to boost detection performance, many approaches (Li et al. 2017; Vipparla et al. 2024; Krishnan et al. 2023) fuse RGB and IR data at the input level. More recently, most studies (Lv and Lan 2025; Chen et al. 2024) have focused on feature-level fusion. For instance, (Jing et al. 2025) proposed MCFNet, which employs multi-scale feature fusion and context-enhancement modules to achieve accurate small-UAV detection. In our task, during inference with only a single RGB image, we improve detection performance by hallucinating infrared feature maps via cross-modal prediction.

2.2 Learning Using Privileged Information

The concept of privileged information was first introduced by (Vapnik and Vashist 2009), where it is defined as data available exclusively during the training phase and unavailable during testing. To date, Learning Using Privileged Information (LUPI) has been widely applied in many computer-vision tasks, including classification (Chen, Niu, and Zhang 2021), detection (Zhang et al. 2021; Motiian et al. 2016), semantic segmentation (Gu et al. 2020), keypoint detection (Bisla and Choromanska 2018), and pose estimation (Lee et al. 2023). To name a few, Chen, Niu, and Zhang proposed to transfer and hallucinate depth attention from depth-privileged scene recognition. Lambert, Sener, and Savarese proposed using privileged information to regulate the variance of Gaussian Dropout, improving model resilience. Hajavi and Etemad distilled video features as privileged information for audio representation learning. In contrast, our study achieves high-quality cross-modal reconstruction by aligning discrete indices, significantly reducing the difficulty associated with aligning continuous variables.

2.3 Vector-Quantization

Vector-Quantization is first proposed VQ-VAE (Van Den Oord, Vinyals et al. 2017), which discretizes the continuous latent space into a learnable codebook, laying the foundation for all subsequent VQ-VAE series variants and extensive applications (Esser, Rombach, and Ommer 2021; Bao et al. 2021; Li, Niu, and Zhang 2023; Liao et al. 2024; Hu et al. 2024, 2020). For example, Razavi, Van den Oord, and Vinyals introduced a hierarchical encoding structure based on the original VQ-VAE, significantly enhancing the quality and diversity of generated samples. Esser, Rombach, and Ommer combined VQ-VAE with GAN, achieving sharper and more detailed image generation by balancing the codebook reconstruction error and adversarial loss. Based on wav2vec 2.0, Baevski, Schneider, and Auli introduced vector quantization to discretize the original speech signal, preserving crucial time-frequency information while effectively compressing feature dimensions, thereby significantly enhancing downstream ASR performance. In contrast, our method performs cross-modal infrared feature map prediction by embedding infrared codebook entries based on the predicted discrete indices.

3 Method

3.1 Task Definition

Considering that the aligned RGB-IR images could be available in training stage while unavailable in inference, we explore the task of Infrared-Privileged UAV Detection. Formally, in the training stage, each sample contains aligned RGB image $\boldsymbol{x}^R \in \mathbb{R}^{3\times H\times W}$ and IR image $\boldsymbol{x}^I \in \mathbb{R}^{3\times H\times W}$. In the test stage, each sample only contains RGB image $\boldsymbol{x}^R \in \mathbb{R}^{3\times H\times W}$. The overall idea of our method is learning to hallucinate complementary information in $\boldsymbol{x}^I$ based on $\boldsymbol{x}^R$. Consequently, in the test stage without $\boldsymbol{x}^I$, our model could benefit from the hallucinated information in $\boldsymbol{x}^I$. For ease of description, we employ the phrase "hallucination modality" to summarize the information estimated from RGB modality to approaching to these in infrared modality. To avoid confusion, we use the superscripts R, I, and H to distinguish the variables in RGB modality, infrared modality and hallucination modality.

3.2 Framework Overview

Overall, our method transfers knowledge across the feature maps from IR modality to RGB modality. Therefore, our method is model-agnostic, and could be applied on various backbones and detector models. For ease of descriptions, we introduce our method based on YOLOv8m (Yaseen 2024), which could be easily extended to other models by replacing the processed feature maps. In particular, we process the multi-scale backbone feature maps (*aka.*, P3, P4, and P5), which are denoted as $\boldsymbol{F}_3$, $\boldsymbol{F}_4$, and $\boldsymbol{F}_5$. We refer to the network layers producing the processed feature maps as "branch" for brevity.

Our network framework in the training stage is illustrated in Fig. 2, which mainly consists of: (1) a pre-trained and frozen IR-branch, producing the hallucination targets; (2) a RGB detector, employing the RGB-branch to extract RGB features and employing the Hallu-branch to hallucinate IR features via vector-quantization. Specifically, given the training sample $\boldsymbol{x}^R$ and $\boldsymbol{x}^I$, we firstly feed $\boldsymbol{x}^I$ into the IR-branch to obtain the hallucination target, *i.e.*, codebook-index $\boldsymbol{E}_3^I$, $\boldsymbol{E}_4^I$, and $\boldsymbol{E}_5^I$, as shown in the top region of Fig. 2. Afterwards, we feed $\boldsymbol{x}^R$ into RGB-Branch to extract RGB feature maps, *i.e.*, $\boldsymbol{F}_3^R$, $\boldsymbol{F}_4^R$, and $\boldsymbol{F}_5^R$, as shown in the bottom region of Fig. 2. Meanwhile, we employ RGB feature maps to hallucinate the codebook-index $\boldsymbol{E}_3^H$, $\boldsymbol{E}_4^H$, and $\boldsymbol{E}_5^H$, which are employed to reconstruct the IR features $\boldsymbol{F}_3^H$, $\boldsymbol{F}_4^H$, and $\boldsymbol{F}_5^H$, as shown in Fig. 2 (highlighted with blue background). Finally, we fuse the RGB and hallucinated features, and feed into the detection head to obtain results.

3.3 Vector-Quantization in IR Modality

Instead of directly hallucinating IR features (Hoffman, Gupta, and Darrell 2016), we propose quantize IR features and hallucinate the codebook indices a hierarchical residual manner, which is more effective and robust. To this end, we conventionally pre-train a detector with IR image $\boldsymbol{x}^I$ as input, and then hierarchically quantize the processed feature maps as the hallucination target. Below, we omit some unnecessary superscripts and subscripts for brevity.

Basic Vector-Quantization. Given one processed feature map $\boldsymbol{F} \in \mathbb{R}^{H\times W\times C}$, we quantize each feature vector $\boldsymbol{f} \in \mathbb{R}^C$ in $\boldsymbol{F}$ via a nearest-neighbor lookup (Van Den Oord, Vinyals et al. 2017) in a codebook $\boldsymbol{\mathcal{C}}$:

$$z = \mathcal{Q}(\boldsymbol{f}, \boldsymbol{\mathcal{C}}) := \arg\min_k \|\boldsymbol{f} - \boldsymbol{\mathcal{C}}[k]\|_2, \tag{1}$$

where $\mathcal{Q}$ is the quantizing function, $z \in \{1, 2, \ldots, N\}$ is the codebook index of $\boldsymbol{f}$, and $\boldsymbol{\mathcal{C}}[k]$ is the k-th entry of the N size codebook $\boldsymbol{\mathcal{C}} \in \mathbb{R}^{N\times C}$. After quantizing, the reconstructed feature according the index z is:

$$\hat{\boldsymbol{f}} = \mathcal{R}(z, \boldsymbol{\mathcal{C}}) := \boldsymbol{\mathcal{C}}[z], \tag{2}$$

where $\mathcal{R}$ is the reconstructing function. To learn the quantization codebook, we follow (Van Den Oord, Vinyals et al. 2017) and employ the loss below:

$$\mathcal{L}_{VQ} = \sum_{\boldsymbol{f}\in\boldsymbol{F}} \|\mathrm{sg}[\boldsymbol{f}] - \hat{\boldsymbol{f}}\|_2^2 + \sum_{\boldsymbol{f}\in\boldsymbol{F}} \|\boldsymbol{f} - \mathrm{sg}[\hat{\boldsymbol{f}}]\|_2^2, \tag{3}$$

where sg denotes stop-gradient.

Hierarchical Residual Quantization. Considering the hierarchical relations of processed feature maps $\boldsymbol{F}_3$, $\boldsymbol{F}_4$, and $\boldsymbol{F}_5$, we propose to quantize them in a hierarchical residual manner. Firstly, we quantize the basal $\boldsymbol{F}_3$ and learn the codebook $\boldsymbol{\mathcal{C}}_3$ using Eqn. 3 Then, we quantize the residual feature of $\boldsymbol{F}_4$ against $\boldsymbol{F}_3$:

$$\tilde{\boldsymbol{F}}_4 = \boldsymbol{F}_4 - \mathcal{D}(\hat{\boldsymbol{F}}_3), \tag{4}$$

where $\hat{\boldsymbol{F}}_3$ is the quantized and reconstructed features of $\boldsymbol{F}_3$, and $\mathcal{D}$ denotes the downsampling and squeezing function for dimensional consistency. In this way, the codebook $\boldsymbol{\mathcal{C}}_4$ could focus on quantizing the incremental features in $\boldsymbol{F}_4$ against $\boldsymbol{F}_3$. Similarly, we learn the codebook $\boldsymbol{\mathcal{C}}_5$ for quantizing $\boldsymbol{F}_5$.

Summary. After this pre-training stage, we obtain a IR detector to extract processed features from $\boldsymbol{x}^I$, as well as the codebooks $\boldsymbol{\mathcal{C}}_3$, $\boldsymbol{\mathcal{C}}_4$, and $\boldsymbol{\mathcal{C}}_5$ for vector-quantization. Besides, we denote the codebook size of $\boldsymbol{\mathcal{C}}_5$ as $N = 64$, and employ $2N$ for $\boldsymbol{\mathcal{C}}_4$ and $4N$ for $\boldsymbol{\mathcal{C}}_3$, probably because that low-level feature maps are relatively more diverse and thus require more entries for quantizing.

3.4 Index Hallucination in RGB Modality

In the formal training stage as shown in Fig. 2, we firstly feed $\boldsymbol{x}^I$ into the pre-trained IR detector to obtain the IR feature maps (*i.e.*, $\boldsymbol{F}_3^I$, $\boldsymbol{F}_4^I$, and $\boldsymbol{F}_5^I$) and corresponding quantizing indices (*i.e.*, $\boldsymbol{Z}_3^I$, $\boldsymbol{Z}_4^I$, and $\boldsymbol{Z}_5^I$), which embed the complementary knowledge in IR modality and thus serve as the hallucination targets.

For the RGB input $\boldsymbol{x}^R$, we feed it into our RGB detector backbone and extract RGB feature maps, *i.e.*, $\boldsymbol{F}_3^R$, $\boldsymbol{F}_4^R$, and $\boldsymbol{F}_5^R$. To hallucinate the indices $\boldsymbol{Z}_3^I$, we feed $\boldsymbol{F}_3^R$ into a hallucination head, as $\boldsymbol{Y}_3^H = \mathcal{H}_3(\boldsymbol{F}_3^R)$, where $\mathcal{H}_3$ contains one 3x3 convolutional layer followed by pixel-level classifier (because the indices are integers), and $\boldsymbol{Y}_3^H$ is the hallucinated classification logits. As the conjugation of Eqn. 4, we feed the residual to hallucinate $\boldsymbol{Z}_4^I$, as:

$$\boldsymbol{Y}_4^H = \mathcal{H}_4(\boldsymbol{F}_4^R - \mathcal{D}(\boldsymbol{F}_3^R)), \tag{5}$$

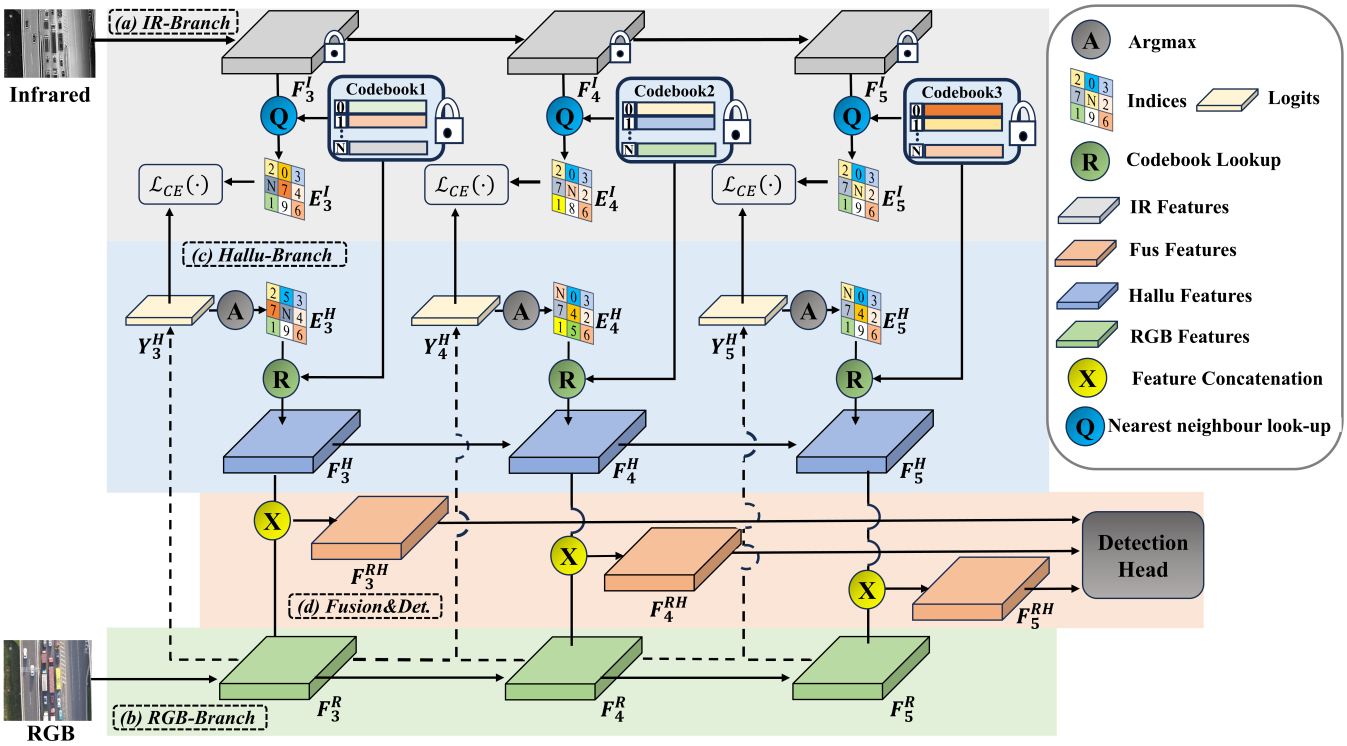

Figure 2: The detailed architecture of our framework in the training stage. In the top region (a), the pre-trained and frozen IR-branch in IR detector extracts IR feature maps and quantize them to codebook-indices as hallucination targets. In the bottom region (b), the RGB-branch in RGB detector extract RGBs feature maps as the foundations. In the mid region (c), the Hallu-branch hallucinates the indices and correspondingly reconstruct the IR feature maps. Finally (d), the fusion of RGB features and hallucinated features are fed into detection head to obtain results.

where $\mathcal{H}_4$ has the same architecture but different parameters with $\mathcal{H}_3$. Similarly, we obtain $\boldsymbol{Y}_5^R$ for hallucinating $\boldsymbol{Z}_5^I$. After that, we reconstruct the IR features by the hallucinated indices and pre-trained codebooks:

$$\boldsymbol{F}_3^H = \mathcal{R}(\mathcal{A}(\boldsymbol{Y}_3^H), \boldsymbol{\mathcal{C}}_3), \tag{6}$$

$$\boldsymbol{F}_4^H = \mathcal{R}(\mathcal{A}(\boldsymbol{Y}_4^H), \boldsymbol{\mathcal{C}}_4) + \mathcal{D}(\boldsymbol{F}_3^H), \tag{7}$$

$$\boldsymbol{F}_5^H = \mathcal{R}(\mathcal{A}(\boldsymbol{Y}_5^H), \boldsymbol{\mathcal{C}}_5) + \mathcal{D}(\boldsymbol{F}_4^H), \tag{8}$$

where $\mathcal{A}$ denotes the argmax function for deriving the classified indices from the hallucinated logits. In this way, we obtain the hallucination features (*i.e.*, $\boldsymbol{F}_3^H$, $\boldsymbol{F}_4^H$, and $\boldsymbol{F}_5^H$) containing complementary IR information with RGB input.

3.5 RGB-Hallucination Multi-Modality Detection

RGB-IR detection is effective due to the complementary feature fusion of RGB modality and IR modality. In our infrared-privileged scenario, the IR modality is unavailable in the test stage, and thus our RGB detector employs hallucination modality to bridge the deficiency of IR modality. In this way, our model can be regarded as a multi-modality detector, which only requires RGB modality in the test stage.

Specifically, our RGB-branch extracts $\boldsymbol{F}_3^R$, $\boldsymbol{F}_4^R$, and $\boldsymbol{F}_5^R$, which contains the semantic information from RGB modality. Meanwhile, our model hallucinates $\boldsymbol{F}_3^H$, $\boldsymbol{F}_4^H$, and $\boldsymbol{F}_5^H$, which contains the complementary information in the codebook from IR modality. We intuitively fuse the corresponding features from RGB and hallucination modalities by:

$$\boldsymbol{F}_i^{RH} = \mathcal{S}([\boldsymbol{F}_i^R; \boldsymbol{F}_i^H]), \tag{9}$$

where $i \in \{3, 4, 5\}$, $[\cdot;\cdot]$ indicates feature concatenation, and $\mathcal{S}$ is a 1x1 convolution for squeezing. After that, we feed $\boldsymbol{F}_3^{RH}$, $\boldsymbol{F}_4^{RH}$, and $\boldsymbol{F}_5^{RH}$ into the network layers after the processed feature maps (after P3, P4, and P5) to produce results.

3.6 Summary of Training and Inference

In the training stage, our RGB detection simultaneously learns to hallucinate the codebook indices and learns to detect objects using RGB-Hallucination features. Therefore, our fully training loss could be summarized as:

$$\mathcal{L}_{FULL} = \mathcal{L}_{DET} + \alpha \cdot \sum_{i \in \{3,4,5\}} \mathcal{L}_{CE}(\boldsymbol{Y}_i^H, \boldsymbol{Z}_i^I), \tag{10}$$

where $\mathcal{L}_{DET}$ denotes the original detection loss in YOLO-v8m, and α balances the trade-off with indices hallucination. $\mathcal{L}_{CE}$ indicates the cross-entropy loss, because the target indices $\boldsymbol{Z}_i^I$ are integers and $\boldsymbol{Y}_i^H$ are the classification logits. We employ cross-validation and set $\alpha = 10.0$ by default. In the test stage, the IR detector is removed due to lacking

IR input. Given the test sample $\boldsymbol{x}^R$, our RGB detector extracts RGB features and hallucinates IR features, and thus can employ complementary information to detect objects.

4 Experiments

4.1 Datasets, Metrics and Implementation Details

DroneVehicle Dataset. DroneVehicle (Sun et al. 2021) is a benchmark for RGB–infrared aerial vehicle detection, comprising paired visible and thermal images captured under diverse illumination conditions throughout the day. It features five vehicle categories and a total of 19,459 aligned image pairs, of which 17,990 are designated for training and 1,469 for testing. Objects in this dataset are typically small and subject to two key challenges: (1) diminished visibility and low contrast in the RGB images during nighttime hours, and (2) spurious thermal artifacts in the infrared channel that can be mistaken for vehicles. The pronounced appearance gap between the two modalities makes DroneVehicle an ideal testbed for evaluating cross-modal fusion and robustness under complementary imaging conditions.

VisDrone Dataset. The VisDrone2019 dataset (Zhu et al. 2018) is a large-scale RGB aerial benchmark for UAV vision tasks such as object detection, tracking, and crowd counting. The VisDrone2019 dataset includes 10,209 static images (plus 288 video clips covering 261,908 frames), manually annotated with over 2.6 million bounding boxes covering categories like pedestrians, cars, bicycles, and tricycles. The detection subset defines 6,471 training, 548 validation, and 1,610 test-dev images, and provides attributes including occlusion and truncation ratios for detailed evaluation. Compared to DroneVehicle, which focuses on paired RGB–IR vehicle detection under varying illumination, VisDrone is RGB-only and emphasizes densely annotated urban UAV scenes for multi-task benchmarks involving complex lighting, occlusion, and scale variation.

Metrics We follow (Lin et al. 2015) and primarily employ COCO-style metric of mAP for evaluation. For mAP calculations, an intersection over union threshold (*e.g.*, 0.5) is employ to determine true and false positives.

Implementation Details Our method is implemented based on the CALNet (He et al. 2023) code, which also employs a multimodal fusion scheme. During training, we apply data augmentation techniques such as random rotation and random cropping to enrich the dataset. Experiments are conducted on Ubuntu 20.04 with Python 3.7, PyTorch 1.10.1, Intel® Core™ i9-14900KF (32 cores), and dual NVIDIA RTX 4090 GPUs. Hyper-parameters are set following the YOLOv8m (Yaseen 2024) framework. We evaluated random seed values in the range 1–10 and observed that all settings yielded comparable performance.

4.2 Comparison with Prior Methods

Baseline Setting. We set three groups of baselines for comparison: (1) RGB-only group, including YOLOv8m (Yaseen 2024), CEASC (Du et al. 2023), DQ-DETR (Huang et al. 2024) and RemDet (Li et al. 2024). (2) IR-privileged group, including FeatHallu (Hoffman, Gupta, and Darrell 2016) and M2D-LIF† (Zhao et al. 2025). (3) RGB-IR group, including CALNet (He et al. 2023), ODAF (Chen et al. 2024) and M2D-LIF (Zhao et al. 2025). Specifically, we adapt FeatHallu (Hoffman, Gupta, and Darrell 2016) and M2D-LIF† (Zhao et al. 2025) using the same base detector (YOLOv8m) as the representative baselines in IR-privileged group. For FeatHallu (Hoffman, Gupta, and Darrell 2016), the hallucination network hallucinates the P3, P4, and P5 features in the IR detector. For M2D-LIF† (Zhao et al. 2025), we employ the RGB backbone and detector with Mono-Modality Distillation, which distills the knowledge from IR to RGB and thus belongs to IR-privileged Group. For our method, we implement our framework on four representative RGB detectors, by hallucinating the last three backbone feature maps. Besides, the baselines in RGB-IR group take extra IR images as test input, which are set to show the upper-bound of our method.

Comparison on RGB-IR Dataset. We firstly conduct the comparison on the RGB-IR dataset (*i.e.*, DroneVehicle), and summarize the results in Tab. 1. Firstly, we can see that the methods in IR-privileged group generally outperform the methods in RGB-only group, indicating the effectiveness of using IR modality as privileged information. Secondly, using the same base detector, "Ours+YOLOv8m" outperforms the FeatHallu and M2D-LIF† dramatically, indicating the advantages of hallucinating codebook-indices. Thirdly, using different base detectors, our methods consistently outperform the original detectors, showing the applicability of our framework by index hallucination. Besides, our IR-privileged methods could reach about 85% of the upper-bound revealed by RGB-IR methods, which is relatively satisfactory.

Applicability on RGB-Only Scenario. Considering that our method only requires RGB input in the test stage, we evaluate the applicability of our trained method on RGB-only scenario, *i.e.*, VisDrone (Gu et al. 2020) dataset. Specifically, we firstly train all methods on DroneVehicle and then fine-tune on VisDrone. In fine-tuning IR-privileged methods, we freeze the backbones, hallucination work, and IR-branch to prevent the degradation of transferred IR knowledge. We summarize the performances of overlapping classes on val set in Tab. 2, and have the following observations. Firstly, the RGB-IR baselines are inapplicable due to the deficiency of IR modality in the test stage. Secondly, the IR-privileged methods still outperform RGB-only methods generally, indicating that the transferred knowledge is also beneficial cross dataset. Finally, our methods consistently outperform the original detectors and other IR-privileged baselines, which demonstrates the well applicability of our proposed framework.

4.3 Ablation Study

In this section, we conduct ablation study based on YOLOv8m detector on DroneVehicle dataset to investigate the impacts of various modules. Specifically, in Row #1 of Tab. 3, we start from the original RGB-only architecture. In Row #2, we only add the architecture of Hallu-

Training Input	Test Input	Venue	Method	Backbone	Car	Truck	Freighter	Bus	Van	mAP@0.5
RGB	RGB	2023	YOLOv8m	CSPDarknet53	91.0	55.8	43.1	86.3	43.4	63.9
		CVPR2023	CEASC	ResNet50	91.4	54.6	42.5	88.7	45.1	64.5
		ECCV2024	DQ-DETR	ResNet50	89.5	57.7	45.6	89.1	43.8	65.1
		AAAI2025	RemDet	RemDet	90.1	56.4	46.3	89.4	47.5	65.9
RGB and IR	RGB	CVPR2016	FeatHalu	CSPDarknet53	92.7	54.8	45.2	88.2	45.6	65.3
		ICCV2025	M2D-LIF†	CSPDarknet53	92.4	57.4	45.3	88.2	46.5	66.0
		–	Ours+YOLOv8m	CSPDarknet53	91.9	62.4	47.3	88.5	47.7	67.6
		–	Ours+CEASC	ResNet50	92.0	59.4	51.3	89.1	46.6	67.7
		–	Ours+DQ-DETR	ResNet50	90.6	60.7	48.6	90.1	52.7	68.5
		–	Ours+RemDet	RemDet	91.0	68.4	57.9	88.8	50.6	71.3
RGB and IR	RGB and IR	MM2023	CALNet	CSPDarknet53	90.3	76.2	63.0	89.1	58.5	75.4
		CVPR2024	ODAF	CSPDarknet53	90.3	76.8	73.3	90.3	66.0	79.4
		ICCV2025	M2D-LIF	CSPDarknet53	97.8	81.0	67.9	96.0	64.6	81.4

Table 1: The performance comparison on the DroneVehicle dataset. The methods are divided into three groups according to different training input and test input.

Training Input	Test Input	Venue	Method	Backbone	Car	Bus	Truck	Van	AVG
RGB	RGB	2023	YOLOv8m	CSPDarknet53	76.7	60.6	30.5	40.1	52.0
		CVPR2023	CEASC	ResNet50	78.3	66.4	37.2	44.7	56.7
		ECCV2024	DQ-DETR	ResNet50	77.9	65.7	44.2	47.6	58.9
		AAAI2025	RemDet	RemDet	81.9	65.3	47.4	47.9	60.6
RGB and IR	RGB	CVPR2016	FeatHalu	CSPDarknet53	82.0	67.2	40.7	44.3	58.6
		ICCV2025	M2D-LIF†	CSPDarknet53	80.4	65.5	45.7	46.6	59.6
		–	Ours+YOLOv8m	CSPDarknet53	83.1	71.8	45.3	50.6	62.7
		–	Ours+CEASC	ResNet50	84.5	71.0	46.7	51.2	63.4
		–	Ours+DQ-DETR	ResNet50	82.7	71.2	50.0	52.6	64.1
		–	Ours+RemDet	RemDet	86.6	74.9	50.2	51.7	65.9

Table 2: The performance comparison of pre-training on the DroneVehicle dataset and then fine-tuning on the VisDrone dataset. The RGB-IR methods are inapplicable due to the deficiency of IR modality in test stage.

	Adding Hallu-Branch	Indices Hallu.	Hierarchical Quantization	mAP@0.5
#1				60.2
#2	✓			60.8
#3	✓	✓		65.7
#4	✓	✓	✓	67.6

Table 3: Ablation Study on the DroneVehicle Dataset.

branch to produce extract features, and feed the feature fusion of RGB-branch and Hallu-branch to obtain results. In Row #3, we employ Hallu-branch to respectively hallucinate the codebook-indices with $\mathcal{L}_{CE}$ enabled. In Row #4, we further hallucinate the indices using the hierarchical residual manner.

In particular, by comparing Row #2 to Row #1, we could see that adding extra network layers only leads to negligible performance gain. By hallucinating indices in Row #3, the performance of our RGB detector is dramatically improved (*i.e.*, +4.9%), indicating the direct benefit of transferring complementary information via index hallucination. In Row #4, the performance is further enhanced (*i.e.*, +1.9%), showing the effectiveness of hierarchical hallucination. Therefore, our proposed modules are effective and complementary.

4.4 Visualization Results

We firstly qualitatively compare our method with representative baselines, and then qualitatively analyse our method.

Visual Comparison As shown in Fig. 3, we compare with YOLOv8m, FeatHallu, and M2D-LIF†. In the challenging cases under fog weather or night, the RGB modality provide limited information, while IR modality provides robust and complementary information. Thus, the RGB-only method (*i.e.*, YOLOv8m) estimates inferior results. Without IR input, the IR-privileged methods can benefit from transferred IR information. As shown in Fig. 3, our method estimates the optimal results, indicating the advantages of hallucinating codebook-indices.

Visual Analysis To investigate the effectiveness of our method, we visualize the RGB features extracted by RGB-

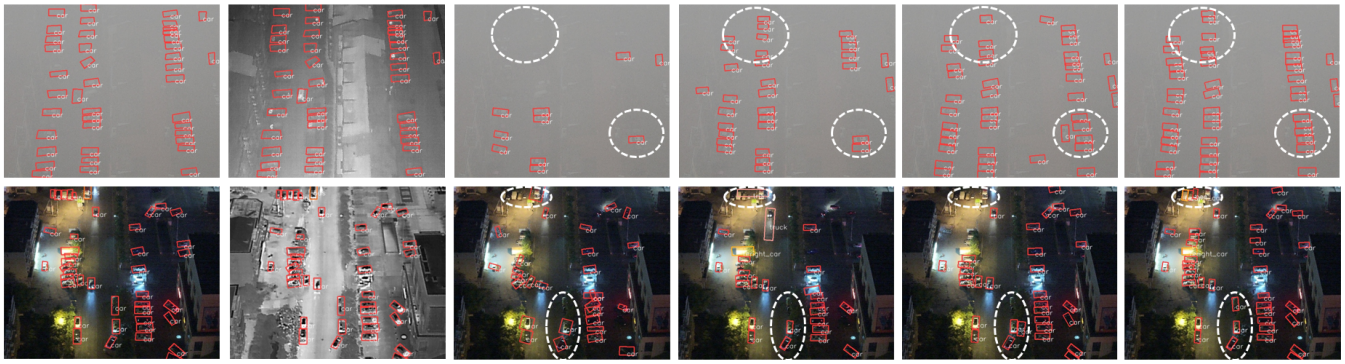

Figure 3: Qualitative comparison on the DroneVehicle dataset in two challenging cases. The white dotted circles highlight the significant regions for comparison. The IR images are presented only for visual reference, which are not input into the models.

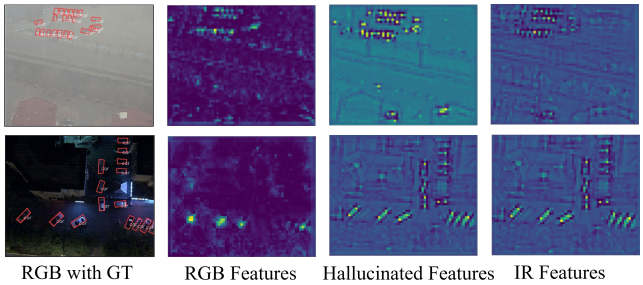

Figure 4: Qualitative analysis of our method on the DroneVehicle dataset in two challenging cases.

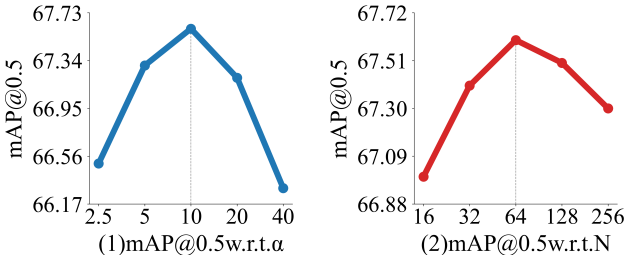

Figure 5: The effects of varying the values of α and N on DroneVehicle. The dotted lines indicate our default values.

branch, hallucinated features extracted by Hallu-branch, and IR features extracted by IR-branch. The IR features are only presented for visual reference, which are not available in the formal evaluation. Besides, we only present the visualizations for P3 features due to its representativeness. As shown in Fig. 4, we can see that the hallucinated features are close to IR features, and also complementary to RGB features. Specifically, in the case under night, the hallucinated features more accurately activate in the object regions. Therefore, our RGB detector could dramatically benefit from the complementary information in hallucinated features without IR input.

4.5 Hyper-Parameters Analysis

In this section, we analyse two major hyper-parameters in our method: (1) $\alpha = 10.0$, for balancing the trade-off between detection loss and hallucination loss in Eqn. 10. (2) $N = 64$, for setting the codebook size in vector-quantization. We determine the optimal values of hyper-parameters via cross-validation. In this hyper-parameter analysis, we vary each hyper-parameter within a specified range while keeping the other fixed, and present the obtained performances in Fig. 5. Lower α may lead to immature feature hallucination, while higher α may suppress the major supervision of detection loss. Smaller codebooks are insufficient to quantize the IR features, while larger codebooks contains more useless entries and also increase the difficulty of index hallucination. Overall, our hyper-parameter setting is relatively appropriate.

5 Conclusion

In this work, we have explored a novel paradigm for UAV detection that using infrared modality as privileged information. Specifically, we have proposed to quantize infrared features and hallucinate codebook-indices based on RGB features. We also have designed a hierarchical hallucination mechanism for multi-scale codebook indices. Extensive experiments on the DroneVehicle and VisDrone datasets have demonstrated the effectiveness of our method.

Acknowledgements

This work was supported in part by the National Natural Science Foundation of China under Grants 62402201, U24A20220, 62132006, in part by the Natural Science Foundation of Jiangxi Province of China under Grants 20252BAC230003, 20252BAC240197, 20242BAB21006, in part by the China Postdoctoral Science Foundation under Grant 2025M771495, and in part by the Early-Career Young Scientists and Technologists Project of Jiangxi Province under Grant 20244BCE52070, 20244BCE52072.

References

Baevski, A.; Schneider, S.; and Auli, M. 2019. vq-wav2vec: Self-supervised learning of discrete speech representations. *arXiv preprint arXiv:1910.05453*.

Bao, H.; Dong, L.; Piao, S.; and Wei, F. 2021. Beit: Bert pre-training of image transformers. *arXiv preprint arXiv:2106.08254*.

Bisla, D.; and Choromanska, A. 2018. VisualBackProp for learning using privileged information with CNNs. arXiv:1805.09474.

Chen, C.; Qi, J.; Liu, X.; Bin, K.; Fu, R.; Hu, X.; and Zhong, P. 2024. Weakly Misalignment-Free Adaptive Feature Alignment for UAVs-Based Multimodal Object Detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 26826–26835.

Chen, J.; Niu, L.; and Zhang, L. 2021. Depth privileged scene recognition via dual attention hallucination. *IEEE Transactions on Image Processing*, 30: 9164–9178.

Du, B.; Huang, Y.; Chen, J.; and Huang, D. 2023. Adaptive Sparse Convolutional Networks with Global Context Enhancement for Faster Object Detection on Drone Images. arXiv:2303.14488.

El Ahmar, W.; Massoud, Y.; Kolhatkar, D.; AlGhamdi, H.; Alja'Afreh, M.; Hammoud, R.; and Laganiere, R. 2023. Enhanced thermal-rgb fusion for robust object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 365–374.

Esser, P.; Rombach, R.; and Ommer, B. 2021. Taming transformers for high-resolution image synthesis. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 12873–12883.

Goecks, V. G.; Woods, G.; and Valasek, J. 2020. Combining Visible and Infrared Spectrum Imagery using Machine Learning for Small Unmanned Aerial System Detection. arXiv:2003.12638.

Gu, Z.; Niu, L.; Zhao, H.; and Zhang, L. 2020. Hard Pixel Mining for Depth Privileged Semantic Segmentation. arXiv:1906.11437.

Guo, Z.; Li, X.; Xu, Q.; and Sun, Z. 2021. Robust semantic segmentation based on RGB-thermal in variable lighting scenes. *Measurement*, 186: 110176.

Hajavi, A.; and Etemad, A. 2023. Audio Representation Learning by Distilling Video as Privileged Information. arXiv:2302.02845.

He, X.; Tang, C.; Zou, X.; and Zhang, W. 2023. Multispectral Object Detection via Cross-Modal Conflict-Aware Learning. In *Proceedings of the 31st ACM International Conference on Multimedia*, MM '23, 1465–1474. New York, NY, USA: Association for Computing Machinery. ISBN 9798400701085.

Hoffman, J.; Gupta, S.; and Darrell, T. 2016. Learning with side information through modality hallucination. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 826–834.

Hu, J.; Lin, J.; Gong, S.; and Cai, W. 2024. Relax image-specific prompt requirement in sam: A single generic prompt for segmenting camouflaged objects. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, 12511–12518.

Hu, J.; Tuo, H.; Wang, C.; Qiao, L.; Zhong, H.; Yan, J.; Jing, Z.; and Leung, H. 2020. Discriminative partial domain adversarial network. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXVII 16*, 632–648. Springer.

Hu, X.; Jiang, Y.; Xia, X.; Chen, C.; Liu, W.; Wan, P.; Bin, K.; and Zhong, P. 2025. UAV-StrawFire: A visible and infrared dataset for real-time straw-fire monitoring with deep learning and image fusion. *International Journal of Applied Earth Observation and Geoinformation*, 141: 104586.

Huang, Y.-X.; Liu, H.-I.; Shuai, H.-H.; and Cheng, W.-H. 2024. DQ-DETR: DETR with Dynamic Query for Tiny Object Detection. arXiv:2404.03507.

Hwang, S.; Park, J.; Kim, N.; Choi, Y.; and Kweon, I. 2015. Multispectral Pedestrian Detection: Benchmark Dataset and Baseline. In *Proceedings of the 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 1037–1045.

Jing, J.; Hu, J. F.; Zhao, Z. P.; and Liu, Y. 2025. MCFNet: Research on small target detection of RGB-infrared under UAV perspective with multi-scale complementary feature fusion. *IET Image Processing*, 19(1): e13320.

Krishnan, B. S.; Jones, L. R.; Elmore, J. A.; Samiappan, S.; Evans, K. O.; Pfeiffer, M. B.; Blackwell, B. F.; and Iglay, R. B. 2023. Fusion of visible and thermal images improves automated detection and classification of animals for drone surveys. *Scientific Reports*, 13(1): 10385.

Lambert, J.; Sener, O.; and Savarese, S. 2018. Deep learning under privileged information using heteroscedastic dropout. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 8886–8895.

Lee, S.; Rim, J.; Jeong, B.; Kim, G.; Woo, B.; Lee, H.; Cho, S.; and Kwak, S. 2023. Human Pose Estimation in Extremely Low-Light Conditions. arXiv:2303.15410.

Li, C.; Zhao, R.; Wang, Z.; Xu, H.; and Zhu, X. 2024. RemDet: Rethinking Efficient Model Design for UAV Object Detection. arXiv:2412.10040.

Li, H.; Ding, W.; Cao, X.; and Liu, C. 2017. Image Registration and Fusion of Visible and Infrared Integrated Camera for Medium-Altitude Unmanned Aerial Vehicle Remote Sensing. *Remote Sensing*, 9(5).

Li, J.; Niu, L.; and Zhang, L. 2023. Knowledge Proxy Intervention for Deconfounded Video Question Answering. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2782–2793.

Liao, Z.; Li, J.; Niu, L.; and Zhang, L. 2024. Align and Aggregate: Compositional Reasoning with Video Alignment and Answer Aggregation for Video Question-Answering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 13395–13404.

Lin, T.-Y.; Maire, M.; Belongie, S.; Bourdev, L.; Girshick, R.; Hays, J.; Perona, P.; Ramanan, D.; Zitnick, C. L.; and Dollár, P. 2015. Microsoft COCO: Common Objects in Context. arXiv:1405.0312.

Liu, K.; Li, H.; Lu, D.; and Huang, H. 2024. RGB-infrared image fusion and classification based on tensor decomposition. In *Conference on Spectral Technology and Applications (CSTA 2024)*, volume 13283, 928–936. SPIE.

Lv, K.; and Lan, P. 2025. DGE-YOLO: Dual-Branch Gathering and Attention for Accurate UAV Object Detection. arXiv:2506.23252.

Motiian, S.; Piccirilli, M.; Adjeroh, D. A.; and Doretto, G. 2016. Information Bottleneck Learning Using Privileged Information for Visual Recognition. In *Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 1496–1505.

Rathinam, A.; Pauly, L.; Rharbaoui, W.; Kacem, A.; Gaudillière, V.; Aouada, D.; et al. 2024. Hybrid Attention for Robust RGB-T Pedestrian Detection in Real-World Conditions. *IEEE Robotics and Automation Letters*.

Razavi, A.; Van den Oord, A.; and Vinyals, O. 2019. Generating diverse high-fidelity images with vq-vae-2. *Advances in neural information processing systems*, 32.

Shen, S.; Li, D.; Mei, L.; Xu, C.; Ye, Z.; Zhang, Q.; Hong, B.; Yang, W.; and Wang, Y. 2023. DFA-Net: Multi-Scale Dense Feature-Aware Network via Integrated Attention for Unmanned Aerial Vehicle Infrared and Visible Image Fusion. *Drones*, 7(8).

Speth, S.; Goncalves, A.; Rigault, B.; Suzuki, S.; Bouazizi, M.; Matsuo, Y.; and Prendinger, H. 2022. Deep learning with RGB and thermal images onboard a drone for monitoring operations. *Journal of Field Robotics*, 39(6): 840–868.

Sun, Y.; Cao, B.; Zhu, P.; and Hu, Q. 2021. Drone-based RGB-Infrared Cross-Modality Vehicle Detection via Uncertainty-Aware Learning. arXiv:2003.02437.

Teledyne FLIR Systems, I. 2018. FLIR ADAS Thermal–Visible Dataset. https://oem.flir.com/solutions/automotive/adas-dataset-form/. Accessed: July 30, 2025.

Van Den Oord, A.; Vinyals, O.; et al. 2017. Neural discrete representation learning. *Advances in neural information processing systems*, 30.

Vapnik, V.; and Vashist, A. 2009. A new learning paradigm: Learning using privileged information. *Neural networks*, 22(5-6): 544–557.

Vipparla, C.; Krock, T.; Nouduri, K.; Fraser, J.; AliAkbarpour, H.; Sagan, V.; Cheng, J.-R. C.; and Kannappan, P. 2024. Fusion of Visible and Infrared Aerial Images from Uncalibrated Sensors Using Wavelet Decomposition and Deep Learning. *Sensors*, 24(24).

Wang, J.; Tian, X.; Dai, S.; Zhuo, T.; Zeng, H.; Liu, H.; Liu, J.; Zhang, X.; and Zhang, Y. 2024. RGB-T Object Detection via Group Shuffled Multi-receptive Attention and Multi-modal Supervision. arXiv:2405.18955.

Wang, S.; Wang, R.; Yao, Z.; Shan, S.; and Chen, X. 2019. Cross-modal Scene Graph Matching for Relationship-aware Image-Text Retrieval. arXiv:1910.05134.

Yang, L.; Ma, R.; and Zakhor, A. 2022. Drone Object Detection Using RGB/IR Fusion. arXiv:2201.03786.

Yaseen, M. 2024. What is YOLOv8: An In-Depth Exploration of the Internal Features of the Next-Generation Object Detector. arXiv:2408.15857.

Zhang, L.; Liu, Z.; Zhu, X.; Song, Z.; Yang, X.; Lei, Z.; and Qiao, H. 2025. Weakly Aligned Feature Fusion for Multimodal Object Detection. *IEEE Trans. Neural Netw. Learn. Syst.*, 36(3): 4145–4159.

Zhang, L.; Zhu, X.; Chen, X.; Yang, X.; Lei, Z.; and Liu, Z. 2019. Weakly Aligned Cross-Modal Learning for Multispectral Pedestrian Detection. In *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, 5126–5136.

Zhang, Z.; Liu, Y.; Chen, J.; Niu, L.; and Zhang, L. 2021. Depth privileged object detection in indoor scenes via deformation hallucination. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, 3456–3464.

Zhao, T.; Liu, B.; Gao, Y.; Sun, Y.; Yuan, M.; and Wei, X. 2025. Rethinking Multi-Modal Object Detection from the Perspective of Mono-Modality Feature Learning. arXiv:2503.11780.

Zhao, T.; Yuan, M.; Jiang, F.; Wang, N.; and Wei, X. 2024. Removal then Selection: A Coarse-to-Fine Fusion Perspective for RGB-Infrared Object Detection. arXiv:2401.10731.

Zhu, P.; Wen, L.; Bian, X.; Ling, H.; and Hu, Q. 2018. Vision Meets Drones: A Challenge. arXiv:1804.07437.