C3-OWD: A CURRICULUM CROSS-MODAL CONTRASTIVE LEARNING FRAMEWORK FOR OPEN-WORLD DETECTION

Anonymous authors
Paper under double-blind review

ABSTRACT

Object detection has advanced significantly in the closed-set setting, but real-world deployment remains limited by two challenges: poor generalization to unseen categories and insufficient robustness under adverse conditions. Prior research has explored these issues separately: visible-infrared detection improves robustness but lacks generalization, while open-world detection leverages vision-language alignment strategy for category diversity but struggles under extreme environments. This trade-off leaves robustness and diversity difficult to achieve simultaneously. To mitigate these issues, we propose **C3-OWD**, a curriculum cross-modal contrastive learning framework that unifies both strengths. Stage 1 enhances robustness by pretraining with RGBT data, while Stage 2 improves generalization via vision-language alignment. To prevent catastrophic forgetting between two stages, we introduce an Exponential Moving Average (EMA) mechanism that theoretically guarantees preservation of pre-stage performance with bounded parameter lag and function consistency. Experiments on FLIR, OV-COCO, and OV-LVIS demonstrate the effectiveness of our approach: C3-OWD achieves 80.1 AP^{50} on FLIR, 48.6 AP^{50}_{Novel} on OV-COCO, and 35.7 mAP_r on OV-LVIS, establishing competitive performance across both robustness and diversity evaluations.

1 INTRODUCTION

Object detection (Zong et al., 2023; Zhu et al., 2020; Zhao et al., 2023; Lv et al., 2024) is a fundamental task in computer vision and has achieved remarkable progress under the closed-set setting, where models are trained and evaluated on a fixed set of categories.

Despite their efficiency and accuracy in standard benchmarks, traditional detectors face two key challenges when deployed in the real world:

- *limited generalization* to unseen categories
- *insufficient robustness* under adverse environmental conditions, as shown in Fig 1, such as low illumination, fog, etc.

To address these issues, two research directions have recently gained attention. On the one hand, visible–infrared object detection (RGBT-OD) (Shen et al., 2024b; Devaguptapu et al., 2019; Medeiros et al., 2024; Lee et al., 2024b) introduces complementary thermal cues that significantly improve robustness under extreme conditions. However, RGBT models are still restricted to closed-set categories and thus exhibit limited generalization. On the other hand, open-world detection (OWD) (Gu et al., 2021b; Zang et al., 2022; Zhong et al., 2022; Zang et al., 2022; Wu et al., 2024a) extends detectors beyond fixed taxonomies by leveraging vision–language alignment strategy such as CLIP (Radford et al., 2021) and GLIP (Li et al., 2022a), enabling recognition of novel categories in dynamic and evolving environments. Yet, OWD systems are typically trained on natural images and struggle with robustness when applied to challenging conditions. Therefore, current research reveals an inevitable trade-off: methods focusing on robustness often lack diversity, while methods emphasizing diversity compromise robustness.

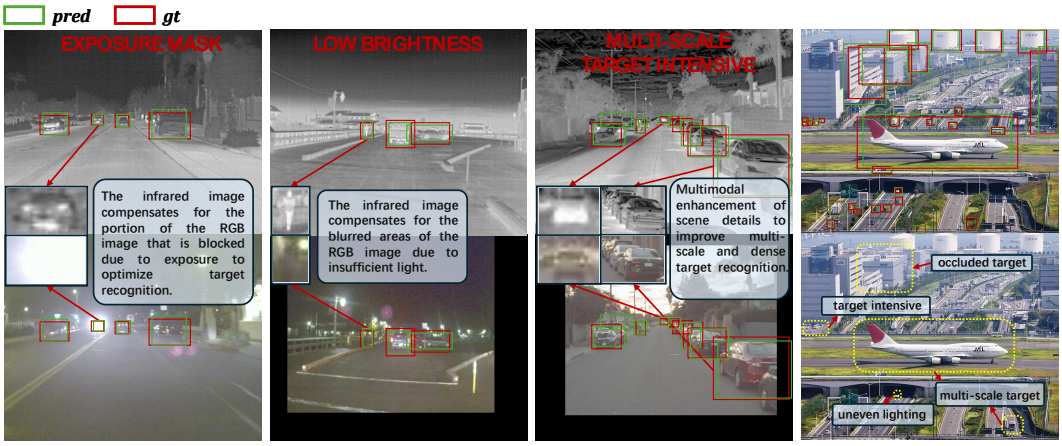

Figure 1: Visible-light imaging faces challenges like exposure mask, low brightness, and multi-scale target intensity, leading to information loss. Infrared imaging provides robust solutions, highlighting the necessity of multi-modal fusion.

Can we develop a unified framework that simultaneously achieves both robustness under extreme conditions and open-vocabulary generalization, thereby breaking the current trade-off between diversity and robustness in object detection?

In this paper, we propose a unified paradigm, **C3-OWD**, to simultaneously address both challenges through a curriculum cross-modal contrastive learning framework. The training is divided into two stages. In Stage 1, we enhance robustness by pretraining the detector on RGBT data such as FLIR dataset (Systems, 2018), enabling the model to resist severe environmental degradation. In Stage 2, we inject semantic priors from text via vision–language alignment, which improves generalization to unseen categories and strengthens the grounding of visual regions with textual descriptions.

A key difficulty in curriculum learning is catastrophic forgetting, where Stage 2 training may override the robustness learned in Stage 1. To alleviate this, we introduce an exponential moving average (EMA) mechanism (Karras et al., 2023; Lee et al., 2024a; Li et al., 2024). Our theoretical analysis demonstrates that the EMA mechanism preserves pre-stage performance with bounded parameter lag, ensuring function consistency between momentum and online branches and preventing catastrophic forgetting when adapting from small to large datasets.

Our main contributions are summarized as follows:

- We present C3-OWD, a unified cross-modal curriculum learning paradigm that integrates RGBT robustness with open-vocabulary generalization capabilities. Unlike traditional methods that suffer from modality bias, our approach dynamically balances multi-modal information through progressive learning, reducing the trade-off between robustness and diversity to achieve better adaptation to diverse environmental conditions.
- We provide rigorous theoretical foundations proving that our Exponential Moving Average (EMA) mechanism effectively prevents catastrophic forgetting with bounded parameter lag and guaranteed function consistency, establishing mathematical guarantees for stable knowledge retention in progressive multi-modal learning.
- Extensive experiments validate the effectiveness of C3-OWD, C3-OWD achieves 80.1 AP^{50} on FLIR, 48.6 AP^{50}_{Novel} on OV-COCO, and 35.7 mAP_r on OV-LVIS, achieving competitive results compared to prior state-of-the-art methods.

2 Related Work

Open-World Detection (OWD) OWD focuses on detecting and learning unknown objects that are not annotated in the training set (Ma et al., 2023; Xi et al., 2024). During inference, detectors identify potential unknowns, which are then annotated and incrementally added as new categories.

Early methods, such as ORE (Joseph et al., 2021), enhanced Faster R-CNN with clustering and energy-based classifiers, while transformer-based approaches like OW-DETR (Gupta et al., 2022) adopted pseudo-labeling but often generated noisy labels. Extensions such as PROB (Zohar et al., 2023) introduced probabilistic objectness modeling, yet calibration issues remained. Recently, large visual models (e.g., SAM (Kirillov et al., 2023)) have further advanced OWOD through pseudo-labeling and knowledge distillation, though challenges in reliable unknown detection still persist.

RGBT Object Detection To address the limitations of relying solely on RGB images for object detection, some studies have introduced a thermal modality, leading to the development of RGBT object detection (Zhou et al., 2020; Sun et al., 2022a). Some methods use uncertainty or confidence metrics to balance RGB and thermal fusion (Kim et al., 2021; Li et al., 2023), while others adjust reliance on thermal inputs based on illumination levels (Guan et al., 2019; Li et al., 2019). Attention-based RGBT networks have also been proposed to further enhance modality integration (Shen et al., 2024a; Yuan & Wei, 2024). In this work, we employ RWKV as the backbone and design tailored training strategies to fully exploit the advantages of RGBT data.

Receptance Weighted Key Value (RWKV) RWKV (Peng et al., 2023) has emerged as a promising paradigm for sequential modeling, combining the efficiency of recurrent networks with the scalability of transformers (Li et al., 2025; Hou & Yu, 2024). By introducing time-shifted receptance gates and exponentially decaying key projections, it achieves linear-time complexity and efficient memory usage (Peng et al., 2025). These advantages have led to extensions in multimodal integration (Yang et al., 2025b; Fei et al., 2024). For example, PointRWKV (He et al., 2025) enhances geometric feature extraction in 3D point clouds, while Vision-RWKV (Duan et al., 2024b) improves high-resolution image understanding beyond traditional ViTs. In this work, we leverage RWKV to efficiently fuse two visual modalities and text, fully exploiting its potential in multimodal learning.

3 METHOD

3.1 OVERVIEW

Our goal is to enhance open-vocabulary detection under challenging conditions by introducing **C3-OWD**, a curriculum cross-modal contrastive learning framework. As illustrated in Fig. 2, our approach is built upon two-stage designed architecture. Stage 1 leverages RGBT (visible–thermal infrared) datasets to strengthen robustness. Stage 2 trains on COCO with Bi-Momentum Contrastive Alignment and Text-Modulated Deformabale Attention to achieve open-vocabulary detection with enhanced semantic grounding.

3.2 STAGE 1 - MULTI-MODAL ROBUSTNESS ENHANCEMENT

As shown in Fig 2 Stage 1, the primary objective of Stage 1 is to establish environmentally robust feature representations through complementary information from RGB and infrared modalities. Unlike traditional methods that employ simple feature concatenation or weighted fusion, we design an adaptive cross-modal interaction mechanism based on RWKV.

Dual-Modal Feature Extraction: Given paired RGB-infrared images $(I_{rgb}, I_{ir}) \in \mathbb{R}^{H\times W\times 3}$, we extract multi-scale features through a shared-weight CNN backbone:

$$\mathcal{F}_{rgb} = \{F^l_{rgb}\}_{l=2,3,4}, \quad \mathcal{F}_{ir} = \{F^l_{ir}\}_{l=2,3,4}. \tag{1}$$

where $F^l_* \in \mathbb{R}^{B\times C_l\times H_l\times W_l}$ denotes the feature map at layer l.

VRWKV-based Adaptive Fusion: Traditional RWKV employs unidirectional (forward) attention, which has limitations in RGB-Thermal cross-modal fusion: (1) **Asymmetric information flow**: Thermal features can compensate for information loss in low-light RGB regions, but unidirectional mechanisms prevent complementary information from flowing backward. (2) **Insufficient modal interaction**: As shown in Figure 1, when RGB images lose information due to exposure or occlusion, bidirectional flow is needed for effective modal compensation. To achieve efficient cross-modal interaction, we adopt VRWKV (Duan et al., 2024a) as the base architecture and extend it with a bidirectional attention mechanism to capture global dependencies. Our bidirectional mechanism introduces bidirectional position decay weights, enabling each token to consider both forward and

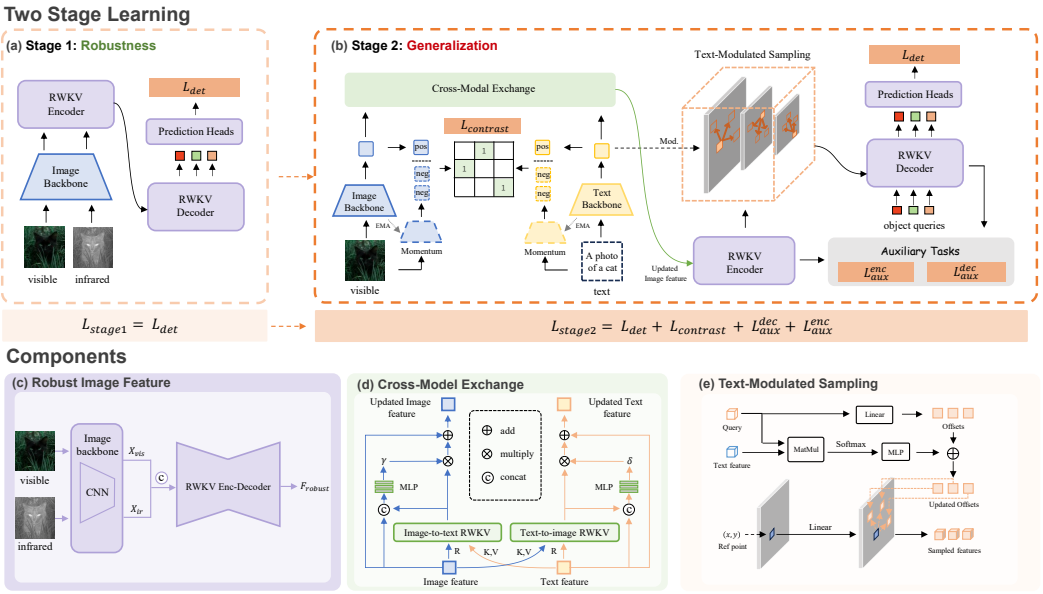

Figure 2: **Overview of C3-OWD**. (a) **Two-stage training**. Stage 1 leverages RGBT (visible–thermal infrared) datasets to strengthen robustness via RWKV. Stage 2 trains on COCO with Bi-Momentum Contrastive Alignment and Text-Modulated Deformabale Attention to achieve open-vocabulary detection with enhanced semantic grounding. (b) **Architecture components**. The Robust Image Feature Block performs efficient sequence modeling and multimodal feature fusion with linear-time complexity using the RWKV architecture. The cross-modal exchange mechanism enables bidirectional information flow and adaptive feature recalibration across modalities. The text-modulated sampling module steers deformable sampling locations with text features via attention, injecting stronger textual priors.

backward context, thus better capturing cross-modal complementarity.
For the t-th token, the attention output is computed as:

$$wkv_t = \text{Bi-WKV}(K, V)_t = \frac{\sum_{i=0, i\neq t}^{T-1} e^{-(|t-i|-1)/T\cdot w + k_i} v_i + e^{u+k_t} v_t}{\sum_{i=0, i\neq t}^{T-1} e^{-(|t-i|-1)/T\cdot w + k_i} + e^{u+k_t}}. \tag{2}$$

where w is a learnable position decay parameter, u is the current token importance weight, and k_i, v_i are the key and value vectors respectively. This bidirectional mechanism allows each token to simultaneously consider both forward and backward contextual information.

RGB and Infrared features interact through the VRWKV-Block (Duan et al., 2024a) to generate modality-aware enhanced representations:

$$\mathcal{F}_{robust} = \text{VRWKV-Block}(\mathcal{F}_{rgb}, \mathcal{F}_{ir}). \tag{3}$$

where the VRWKV-block consists of an encoder-decoder architecture with cross-modal interaction mechanisms. The encoder processes multi-scale RGB-Thermal features, while the decoder refines the representations, enabling effective fusion of complementary information from both modalities.

During this stage, we employ standard detection losses including classification loss $\mathcal{L}_{cls}$ and regression loss $\mathcal{L}_{reg}$, directly supervised on paired RGB-infrared data. This pre-training strategy enables the model to learn robust feature representations across varying illumination conditions, establishing a strong foundation for subsequent open-world adaptation. The full details can be found in Algorithm 1.

3.3 Stage 2 - Vision-Language Generalization Alignment

As shown in Fig 2 Stage 2, in the second stage, we perform comprehensive vision-language alignment through three interconnected components: semantic enhancement fusion, text-modulated de-

formable attention, and bi-momentum contrastive learning. This stage enriches visual representations with semantic priors while maintaining efficient open-vocabulary detection capabilities.

3.3.1 SEMANTIC ENHANCEMENT FUSION

We begin by enriching visual features with semantic priors through a hierarchical vision-language fusion module. Visual features $\mathbf{C} = \{C_2, C_3, C_4\}$ from the backbone (where $C_2 \in \mathbb{R}^{B\times 256\times \frac{H}{8}\times \frac{W}{8}}$, $C_3 \in \mathbb{R}^{B\times 256\times \frac{H}{16}\times \frac{W}{16}}$, $C_4 \in \mathbb{R}^{B\times 256\times \frac{H}{32}\times \frac{W}{32}}$) are projected into a shared space via linear projection layers:

$$C_i' = \text{Linear}(C_i), \quad i \in \{2, 3, 4\} \tag{4}$$

Then, the scaled embeddings are flattened and concatenated to form the multi-scale visual sequence $\mathbb{V} = \text{Concat}(C_2', C_3', C_4') \in \mathbb{R}^{B\times L\times D}$, where L is the total sequence length and D is the embedding dimension. We then perform l rounds of bidirectional cross-attention between the multi-scale visual features $\mathbb{V}$ and text embeddings $\mathbf{T_{clip}} \in \mathbb{R}^{B\times CLA\times D}$ (where CLA is the number of classes), followed by cross-modal exchange:

$$\mathbb{V}_{out}, \mathbf{T}_{out} = \text{CrossModalExchange}(\mathbb{V}, \mathbf{T_{clip}}), \tag{5}$$

where the CrossModalExchange module performs bidirectional feature fusion using RWKV-based architecture. Specifically, the module computes cross-modal interactions through two parallel pathways:

$$\mathbf{V}_{fuse} = \text{RWKV}_{I\to T}(R = \mathbb{V}, K = \mathbf{T_{clip}}, V = \mathbf{T_{clip}}), \tag{6}$$

$$\mathbf{T}_{fuse} = \text{RWKV}_{T\to I}(R = \mathbf{T_{clip}}, K = \mathbb{V}, V = \mathbb{V}), \tag{7}$$

where $\text{RWKV}_{I\to T}$ denotes Image-to-Text RWKV interaction (updating visual features with text context) and $\text{RWKV}_{T\to I}$ denotes Text-to-Image RWKV interaction (updating text features with multi-scale visual context). Here, R serves as the query-like input from the target modality while K and V are the key-value pairs from the source modality.

The attended features are then concatenated with the original features and processed through a two-layer MLP to generate gating coefficients:

$$\gamma = \text{MLP}([\mathbb{V} \cdot \mathbf{V}_{fuse}]), \quad \delta = \text{MLP}([\mathbf{T_{clip}} \cdot \mathbf{T}_{fuse}]), \tag{8}$$

where $\cdot$ denotes concatenation. The final outputs incorporate adaptive residual connections:

$$\mathbb{V}_{out} = \mathbb{V} + \gamma \otimes \mathbf{V}_{fuse}, \quad \mathbf{T}_{out} = \mathbf{T_{clip}} + \delta \otimes \mathbf{T}_{fuse}. \tag{9}$$

where $\otimes$ represents element-wise multiplication. This design enables adaptive cross-modal enhancement while effectively leveraging hierarchical visual representations.

3.3.2 TEXT-MODULATED SAMPLING

In traditional methods, the sampling locations (where the model "looks") are predicted solely based on visual feature maps. However, for novel categories that the model has never seen, purely visual cues may be insufficient to locate the object of interest. To enhance spatial feature sampling with semantic guidance, we introduce a text-modulated sampling mechanism. Given query features $\mathbf{Q} \in \mathbb{R}^{B\times N\times D}$ and text features $\mathbf{T}_{text} \in \mathbb{R}^{B\times N\times D}$, the module first generates base sampling offsets:

$$\Delta_{base} = \text{Linear}(\mathbf{Q}), \tag{10}$$

where $\Delta_{base} \in \mathbb{R}^{B\times N\times 2K}$ represents K sampling points with 2D offsets.

Simultaneously, text-guided modulation weights are computed through cross-modal attention:

$$\mathbf{A} = \text{Softmax}(\mathbf{Q} \cdot \mathbf{T}_{text}^T / \sqrt{D}), \mathbf{W}_{mod} = \text{MLP}(\mathbf{A}) \in \mathbb{R}^{B\times N\times 2K}, \tag{11}$$

where $\mathbf{A} \in \mathbb{R}^{B\times N\times C}$ represents attention scores between queries and text features, and the MLP maps the attention-weighted features from dimension C to $2K$ to match the offset dimensions: $\text{MLP} : \mathbb{R}^{B\times N\times C} \to \mathbb{R}^{B\times N\times 2K}$.

The text-modulated offsets are obtained by combining base offsets with modulation:

$$\Delta_{updated} = \Delta_{base} \oplus \mathbf{W}_{mod}, \tag{12}$$

where $\oplus$ denotes element-wise addition. These updated offsets guide the sampling process:

$$\mathbf{F}_{sampled} = \text{Linear}(\text{Sample}(\mathbf{F}_{ref}, \mathbf{p}_{ref} + \Delta_{updated})). \tag{13}$$

where $\mathbf{F}_{ref}$ represents reference features, $\mathbf{p}_{ref} = (x, y)$ are reference points, and Sample$(\cdot)$ performs bilinear interpolation at the offset positions. This text-modulated sampling enables semantically-aware spatial attention, improving feature extraction at object boundaries and semantic regions.

3.3.3 Bi-Momentum Contrastive Alignment

To address the issue that negative samples are constrained by memory during training, we adopt MoCo (He et al., 2019) paradigm for region-text contrastive learning. Two momentum-updated feature queues are maintained:

$$\mathcal{Q}_{region} \in \mathbb{R}^{K \times D_{proj}}, \mathcal{Q}_{text} \in \mathbb{R}^{K \times D_{proj}}. \tag{14}$$

where K is the queue size and D_{proj} is the projection dimension.

Positive region proposals are selected via IoU threshold τ_{IoU}:

$$\mathcal{P}_{pos} = \{p_i | \text{IoU}(p_i, g_j) \geq \tau_{IoU}, \exists g_j \in \mathcal{G}_{gt}\}. \tag{15}$$

Region features are encoded through RoI extraction and projection:

$$\mathbf{r}_q = f_\theta(\text{RoIExtractor}(\mathbf{F}_{out}, \mathcal{P}_{pos})), \mathbf{r}_k = f_{\theta_m}(\text{RoIExtractor}(\mathbf{F}_{out}, \mathcal{P}_{pos})). \tag{16}$$

where f_θ is a trainable projection network and f_{θ_m} is its momentum version updated via EMA:

$$\theta_m \leftarrow m \cdot \theta_m + (1 - m) \cdot \theta. \tag{17}$$

Multi-positive InfoNCE loss is used for image-to-text and text-to-image alignment:

$$\mathcal{L}_{i2t} = -\frac{1}{N_r} \sum_{i=1}^{N_r} \log \frac{\sum_{j \in \mathcal{P}_i^+} \exp(s_{ij}/\tau)}{\sum_{j=1}^{N_t+K} \exp(s_{ij}/\tau)}. \mathcal{L}_{t2i} = -\frac{1}{N_t} \sum_{j=1}^{N_t} \log \frac{\sum_{i \in \mathcal{P}_j^+} \exp(s_{ji}/\tau)}{\sum_{i=1}^{N_r+K} \exp(s_{ji}/\tau)}. \tag{18}$$

where $s_{ij} = \langle \mathbf{r}_{q,i}, \mathbf{t}_{k,j} \rangle \cdot \exp(\alpha)$ denotes the similarity between the i-th region feature and the j-th text feature, α is a learnable temperature parameter, τ is a fixed temperature hyperparameter, N_r is the number of region features in the current batch, N_t is the number of text features in the current batch, K is the momentum queue size, $\mathcal{P}_i^+$ is the set of positive text indices for the i-th region, and $\mathcal{P}_j^+$ is the set of positive region indices for the j-th text. The total contrastive loss:

$$\mathcal{L}_{contrast} = \lambda_{i2t} \mathcal{L}_{i2t} + \lambda_{t2i} \mathcal{L}_{t2i}. \tag{19}$$

where λ_{i2t} and λ_{t2i} are loss weights. The full details can be found in Algorithm 2.

For more information regarding Stage 2, please refer to the appendix A.3

3.4 Two-Stage Training Strategy

Stage 1 - Multi-modal Robustness Training: Using RGBT datasets, we enhance the model's robustness in extreme environments through visible-thermal infrared fusion. Only a single query head is used in this stage without any auxiliary heads. The loss function is defined as:

$$\mathcal{L}_{stage1} = \mathcal{L}_{det}(\mathbf{X}_{vis}, \mathbf{X}_{ir}). \tag{20}$$

where $\mathcal{L}_{det}$ is the detection loss, including classification loss $\mathcal{L}_{cls}$ and regression loss $\mathcal{L}_{reg}$.

Stage 2 - Open-Vocabulary Training: Training on the COCO dataset, we incorporate CLIP semantic features and momentum contrastive learning. Inspired by Co-DETR (Zong et al., 2023), we introduce multiple auxiliary heads to enhance the training efficiency of both encoder and decoder. The loss function is defined as:

$$\mathcal{L}_{stage2} = \mathcal{L}_{det} + \lambda_c \mathcal{L}_{contrast} + \lambda_{aux} \mathcal{L}_{aux}, \tag{21}$$

$$\mathcal{L}aux = \sum_{i=1}^{K} \mathcal{L}_i^{enc} + \lambda_1 \sum_{i=1}^{K} \sum_{l=1}^{L} \mathcal{L}_{i,l}^{dec}. \tag{22}$$

where $\mathcal{L}_{det}$ is the detection loss based on Hungarian matching including classification and regression losses, $\mathcal{L}_{contrast}$ is the contrastive loss for open-vocabulary learning using CLIP semantics, $\mathcal{L}_i^{enc}$ is the encoder auxiliary loss for the i-th head (e.g., ATSS or Faster R-CNN loss), $\mathcal{L}_{i,l}^{dec}$ is the decoder auxiliary loss for the i-th head at the l-th layer, K is the number of auxiliary heads, L is the number of decoder layers, λ_c is the weight of $\mathcal{L}_{contrast}$, λ_{aux} is the weight of $\mathcal{L}_{aux}$ and λ_1 is the weight for decoder auxiliary losses.

To mitigate catastrophic forgetting, as further detailed in Appendix A.8, which often arises in continual learning scenarios, we introduce Exponential Moving Average(EMA) (Karras et al., 2023; Lee et al., 2024a; Li et al., 2024). Intuitively, EMA maintains a momentum branch that smooths the parameter updates, keeping it close to the online branch while preserving stage-1 robustness. This helps ensure consistency of functions and loss stability across stages. A formal derivation and proof are provided in Appendix A.8.

The two-stage training strategy enables C3-OWD to first learn robust feature representations from multi-modal data, then leverage these representations to achieve superior open-vocabulary detection performance through semantic alignment with CLIP features and momentum contrastive learning.

4 Experiments

4.1 Datasets and Evaluation Metrics

The training process consists of two stages, utilizing the following datasets: the FLIR Dataset (Systems, 2018). The dataset provides thermal infrared images, a modality known for its high robustness to challenging conditions like extreme illumination (e.g., strong light, darkness) and noise pollution. They are used in the first training stage to enhance the detector's robustness and generalization capabilities by leveraging the invariant properties of the infrared spectrum. Subsequently, in the second training stage, we evaluate our approach on two standard open-vocabulary detection benchmarks modified from LVIS (Gupta et al., 2019) and COCO (Lin et al., 2014) respectively. LVIS (Gupta et al., 2019) contains 100K images with 1,203 classes. The classes are divided into three groups, namely frequent, common and rare, based on the number of training images. Following ViLD (Gu et al., 2021a), we treat 337 rare classes as novel classes and use only the frequent and common classes for training. The COCO dataset is a widely used benchmark for object detection, which consists of 80 classes. Following OVR-CNN (Zareian et al., 2021), we divide the classes in COCO into 48 base categories and 17 novel categories, while removing 15 categories without a synset in the WordNet hierarchy. The training set is the same as the full COCO but only images containing at least one base class are used. For FLIR (Systems, 2018) dataset, we report AP^{50}—the mean Average Precision (mAP) at an IoU threshold of 0.5 for base categories—as the primary metric. For COCO (Lin et al., 2014), we report AP^{50}_{novel}—the mean Average Precision (mAP) at an IoU threshold of 0.5 for novel categories—as the primary metric. Additionally, we provide performance on base categories (AP^{50}_{base}) and overall performance across all categories (AP^{50}). For LVIS (Gupta et al., 2019), we report AP_r, AP_c, and AP_f—denoting mAP on rare, common, and frequent categories, respectively—along with the overall AP, all computed using standard box-based mAP.

4.2 Implementation Details

Our model is based on a Deformable-DETR (Zhu et al., 2020) architecture, initialized from a pretrained ResNet-50 backbone. The first-stage training utilizes the FLIR dataset (Systems, 2018), while the second-stage is performed on the COCO (Lin et al., 2014) and LVIS Gupta et al. (2019) dataset. Both stage-1 and stage-2 trainings use 36 epochs. We employ the AdamW optimizer with an initial learning rate of 5×10^{-5} and weight decay of 1×10^{-4}, applying a layer-wise learning rate decay with a multiplier of 0.1 for the backbone. The learning rate is reduced by a factor of 10 at the 12^{th} and 24^{th} epoch during the 36-epoch training schedule.Following the MoCo (He et al., 2019), We set $K = 65536$ and $\tau_{\text{IoU}} = 0.3$. All input images are resized to 640×640 pixels with a batch size of 4 per GPU, and we apply gradient clipping with a maximum norm of 0.1. Following Co-DTER Zong et al. (2023), the model incorporates multiple detection heads including

Table 1: Comparison on FLIR dataset.

Method	Backbone	AP^{50} (FLIR)
MMTOD-CG (Devaguptapu et al., 2019)	RN50	61.4
MMTOD-UNIT (Devaguptapu et al., 2019)	RN101	61.5
CMPD (Li et al., 2022b)	RN50	69.4
CFR (Zhang et al., 2020)	VGG-16	72.4
GAFF (Zhang et al., 2021)	RN50	72.9
BU-ATT (Kieu et al., 2021)	RN50	73.1
BU-LTT (Kieu et al., 2021)	RN50	73.2
UA_CMDet (Sun et al., 2022b)	RN50	78.6
CFT (Fang et al., 2021)	RN50	78.7
CSAA (Cao et al., 2023)	RN50	79.2
ICAFusion (Shen et al., 2024b)	RN50	79.2
CrossFormer (Lee et al., 2024b)	Swin-B	79.3
MFPT (Zhu et al., 2023)	RN101	80.0
RSDet (Zhao et al., 2024b)	RN50	81.1
MiPa (Medeiros et al., 2024)	ViT-B/16	81.3
MMFN (Yang et al., 2025a)	RN50	**81.8**
C3-OWD (Ours)	RN50	80.1

RPN, query-based, and ROI heads, with a total of 6 encoder layers and 8 decoder layers, the loss weighting factor λ_2 set to 2.0. We set $\lambda_c = 0.01$, $\lambda_{\text{aux}} = 0.1$, and $\lambda_1 = 1$. The temperature τ is set to 0.07. We train and evaluate our model on 8×NVIDIA A100 GPUs.

4.3 Main result

As shown in Table 1, 2. Although C3-OWD does not achieve the absolute state-of-the-art on the FLIR dataset (slightly behind MMFN's 81.8), it significantly outperforms all existing methods on open-vocabulary detection tasks (OV-COCO and OV-LVIS), notably achieving a remarkable AP^{50}_{Novel} of 48.6 on OV-COCO—surpassing the previous best result of 44.3 by CLIPSelf. More importantly, C3-OWD is the method that demonstrates leading performance across both multimodal robustness and open-vocabulary detection tasks, highlighting its superior generality and cross-task adaptability. This indicates that our proposed curriculum cross-modal contrastive learning framework not only enhances robustness in extreme environments but also effectively transfers to open-vocabulary scenarios, enabling broader applicability.

Table 2: **Comparison with state-of-the-art methods on OV-COCO and OV-LVIS benchmarks.** We report AP^{50}_{Novel} for OV-COCO and mAP_r (rare classes) for OV-LVIS. "-" indicates the result is not reported in the original paper for the specific backbone. **Bold** indicates the best performance.

Method	**Backbone**	**OV-COCO** (AP^{50}_{Novel})	**OV-LVIS** (mAP_r)
ResNet-50 Based Methods			
ViLD (Gu et al., 2021b)	RN50	27.6	16.3
Detic (Zhou et al., 2022)	RN50	27.8	24.9
OV-DETR (Zang et al., 2022)	RN50	29.4	17.4
ProxyDet (Jeong et al., 2024)	RN50	30.4	-
RegionCLIP (Zhong et al., 2022)	RN50	31.4	17.1
RTGen (Chen et al., 2024)	RN50	33.6	21.4
BARON-KD (Wu et al., 2023a)	RN50	34.0	22.6
CLIM (Wu et al., 2024b)	RN50	36.9	23.9
SAS-Det (Zhao et al., 2024a)	RN50	37.4	20.9
Larger ResNet Variants			
RegionCLIP (Zhong et al., 2022)	RN50x4	39.3	22.0
CORA (Wu et al., 2023b)	RN50x4	41.7	28.1
SAS-Det (Zhao et al., 2024a)	RN50x4	43.9	29.1
CLIM (Wu et al., 2024b)	RN50x64	-	32.3
F-VLM (Kuo et al., 2023)	RN50x64	-	32.8
Transformer Based Methods (ViT / Swin)			
PromptDet (Song & Bang, 2023)	ViT-B/16	30.6	-
RTGen (Chen et al., 2024)	Swin-B	-	30.2
BIND (**?**)	ViT-L/16	41.5	32.5
Detic (Zhou et al., 2022)	Swin-B	-	33.8
CFM-ViT (Kim et al., 2023a)	ViT-L/16	34.1	33.9
RO-ViT (Kim et al., 2023b)	ViT-H/16	33.0	34.1
CLIPSelf (Wu et al., 2024a)	ViT-L/14	44.3	34.9
ProxyDet (Jeong et al., 2024)	Swin-B	-	36.7
CoDet (Ma et al., 2024)	ViT-L/14	44.7	**37.0**
C3-OWD (Ours)	**RN50**	**48.6**	**35.7**

Table 3: Ablation study of C3-OWD components on OV-COCO and OV-LVIS datasets.

No.	Ablated Component	OV-COCO (AP^{50}_{novel})	OV-LVIS (AP_r)
0	C3-OWD (Full Model)	**48.6**	**35.7**
1	w/o encoder fusion	42.1	31.2
2	static query selection	44.1	32.3
3	w/o text cross-attention	40.9	29.8
4	w/o MoCo	39.7	27.5
5	w/o Bi-MoCo	46.1	33.2

Table 4: Ablation study on Stage-1 modality and Stage-2 initialization.

Config.	Stage-1 Modality	Stage-2 Initialization	FLIR (AP^{50})	OV-COCO (AP^{50}_{novel})	OV-LVIS (AP_r)
Full Model	RGB + IR	From Stage-1	**80.1**	**48.6**	**35.7**
A	RGB only	From Stage-1	75.2	43.5	32.1
B	IR only	From Stage-1	76.8	44.1	32.8
C	RGB + IR	Random	80.0	41.2	30.5
D	RGB only	Random	75.1	39.8	29.1
E	IR only	Random	76.7	40.3	29.7

4.4 Ablation Study

Table 3 presents the ablation results of **key components** in C3-OWD. The full model (No. 0) achieves the best performance on both OV-COCO and OV-LVIS datasets. Removing the encoder fusion module (No. 1) causes significant performance drops (6.5 in AP^{50}_{novel} and 4.5 in AP_r), demonstrating the crucial role of hierarchical feature fusion for open-vocabulary detection. Using static query selection (No. 2) instead of deformable attention also leads to considerable performance degradation, validating the effectiveness of dynamic query optimization. The absence of text cross-attention (No. 3) results in the most substantial performance decline, indicating that vision-language interaction is core to cross-modal alignment. Removing MoCo contrastive learning (No. 4) hurts performance the most, highlighting the critical importance of momentum contrast in addressing positive-negative sample imbalance. Finally, using single-queue instead of dual-queue MoCo (No. 5) also reduces performance, proving that the region-text dual-queue design effectively prevents inter-modal interference. All ablation results consistently show that each component in C3-OWD contributes importantly to the final performance.

As shown in Table 4, we conduct a thorough ablation study to validate the necessity of the two-stage curriculum design in C3-OWD. The results clearly demonstrate that both the multimodal training in Stage-1 and the weight inheritance in Stage-2 are critical for achieving optimal performance. Using only a single modality (RGB or IR) in Stage-1 leads to a noticeable drop in robustness, as evidenced by the decreased FLIR AP^{50} (Configs A and B vs. Full). This performance degradation propagates to the open-vocabulary detection tasks in Stage-2, underscoring the importance of cross-modal complementary learning for acquiring generalized representations. Initializing Stage-2 with random weights instead of the pre-trained weights from Stage-1 results in a substantial performance collapse on OV-COCO and OV-LVIS (Config C vs. Full), even when the full RGB-IR modality was used previously. This indicates that the robust features learned in Stage-1 are a crucial foundation for the subsequent semantic alignment and contrastive learning. The worst performance is observed when both components are ablated (Configs D and E), reinforcing that the proposed curriculum learning pipeline is holistic and both stages are indispensable.

As shown in Table 5, we compared Inference FPS on a single A100 GPU with input size 640×640 and the batch size is 8. Thanks to the RWKV-based fusion and efficient attention, our method achieves **62.4 FPS**, meeting real-time requirements for autonomous systems. Although the incorporation of MoCo contrastive learning increases the model parameter count, leading to a 17% decrease in inference speed, the resulting improvement in accuracy and enhanced multimodal fusion performance make this computational cost acceptable. We chose $\tau_{IoU} = 0.3$ specifically for the Open-World setting. In open-vocabulary detection, the Region Proposal Network (RPN) often assigns lower confidence scores to novel/unseen categories compared to base categories. A standard threshold (e.g., 0.5) tends to filter out valid proposals for novel objects, hurting Recall.

Table 5: Inference Speed and Memory Comparison (COCO)

Method	Backbone	Complexity	Params (M)	FPS
RegionCLIP	ResNet-50	$O(N^2)$	216	68.47
OV-DETR	ResNet-50	$O(N^2)$	165	71.59
RegionCLIP	ResNet-50x4	$O(N^2)$	403	52.21
CLIM	ResNet-50x4	$O(N^2)$	204	69.8
Detic	Swin-B	$O(N^2)$	670	46
CLIPSelf	ViT/L-14	$O(N^2)$	480	46.2
C3-OWD (w/o MoCo)	ResNet-50	$O(N)$	**124.3**	**73.1**
C3-OWD (Ours)	ResNet-50	$O(N)$	245.6	62.4

As shown in Table 6, decreasing IoU from 0.5 to 0.3 significantly improves AP_{Novel} (+3.4) and

Table 6: Ablation of IoU Threshold (τ_{IoU}) on OV-COCO

τ_{IoU}	AP^{50}_{Novel}	AP^{50}_{Base}	Recall (Novel)
0.5	45.2	**56.8**	62.4%
0.4	47.1	56.5	68.9%
0.3 (Ours)	**48.6**	56.1	**74.3%**
0.2	48.2	54.9	71.1%

Recall (+11.9%) with only a negligible drop in Base AP. Setting it too low (0.2) introduces excessive noise. Thus, 0.3 provides the optimal trade-off.

Table 7: Impact of EMA on Preserving Stage-1 Robustness

Setting	**FLIR AP^{50} (Robustness)**	OV-COCO AP_{Novel}
C3-OWD (Full with EMA)	**80.1**	**48.6**
w/o EMA (Direct Finetuning)	59.5 (-20.6)	47.8 (-0.8)
w/o Momentum Contrast	72.3 (-7.8)	39.7 (-8.9)

Role in Robustness (FLIR): As shown in Table 7, removing the EMA mechanism (i.e., allowing the model to update freely without a momentum teacher) causes the FLIR performance to drop significantly from 80.1 to 59.5. This empirically validates our Theorem 1, showing that EMA is essential to enforce function consistency and retain the robust RGBT features learned in Stage 1. **Role in Generalization (COCO):** While the Open-Vocabulary performance on COCO remains relatively high without EMA (47.8), the model sacrifices its unique advantage—robustness. Therefore, the ablation should be seen as proof that EMA is the "stabilizer" that allows C3-OWD to excel in both tasks simultaneously, rather than trading one for the other.

5 Conclusion

In this paper, we introduce an innovative curriculum cross-modal contrastive learning framework for open-world detection, C3-OWD. The proposed two-stage training paradigm addresses the fundamental challenge of achieving both environmental robustness and open-world generalization. Our EMA mechanism provides theoretical guarantees for knowledge preservation across training stages, with rigorously proven bounded parameter lag and function consistency. The integration of RGBT multimodal fusion with RWKV-based architecture enables effective capture of complementary information from visible and infrared modalities, while the vision-language alignment stage extends detection capabilities to novel categories without sacrificing the learned robustness. Extensive experiments validate the effectiveness of C3-OWD across OV-COCO,OV-LVIS and FLIR. C3-OWD achieves competitive results compared to prior state-of-the-art methods.

REFERENCES

H. Akbari, L. Yuan, R. Qian, W. H. Chuang, S. F. Chang, Y. Cui, and B. Gong. Vatt: Transformers for multimodal self-supervised learning from raw video, audio and text. In *Advances in Neural Information Processing Systems*, volume 34, pp. 24206–24221, 2021.

Yue Cao, Junchi Bin, Jozsef Hamari, Erik Blasch, and Zheng Liu. Multimodal object detection by channel switching and spatial attention. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 403–411, 2023.

Fangyi Chen, Han Zhang, Zhantao Yang, Hao Chen, Kai Hu, and Marios Savvides. Rtgen: Generating region-text pairs for open-vocabulary object detection. *arXiv preprint arXiv:2405.19854*, 2024.

Chaitanya Devaguptapu, Ninad Akolekar, Manuj M Sharma, and Vineeth N Balasubramanian. Borrow from anywhere: Pseudo multi-modal object detection in thermal imagery. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pp. 1029–1038, 2019.

Yuchen Duan, Weiyun Wang, Zhe Chen, Xizhou Zhu, Lewei Lu, Tong Lu, Yu Qiao, Hongsheng Li, Jifeng Dai, and Wenhai Wang. Vision-rwkv: Efficient and scalable visual perception with rwkv-like architectures. *arXiv preprint arXiv:2403.02308*, 2024a.

Yuchen Duan, Weiyun Wang, Zhe Chen, Xizhou Zhu, Lewei Lu, Tong Lu, Yu Qiao, Hongsheng Li, Jifeng Dai, and Wenhai Wang. Vision-rwkv: Efficient and scalable visual perception with rwkv-like architectures. *arXiv preprint arXiv:2403.02308*, 2024b.

Qingyun Fang, Dapeng Han, and Zhaokui Wang. Cross-modality fusion transformer for multispectral object detection. *arXiv preprint arXiv:2111.00273*, 2021.

Zhengcong Fei, Mingyuan Fan, Changqian Yu, Debang Li, and Junshi Huang. Diffusion-rwkv: Scaling rwkv-like architectures for diffusion models. *arXiv preprint arXiv:2404.04478*, 2024.

C. Feichtenhofer, H. Fan, J. Malik, and K. He. Slowfast networks for video recognition. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 6202–6211, 2019.

Y. Ganin, E. Ustinova, H. Ajakan, P. Germain, H. Larochelle, F. Laviolette, M. Marchand, and V. Lempitsky. Domain-adversarial training of neural networks. *Journal of Machine Learning Research*, 17(59):1–35, 2016.

Xiuye Gu, Tsung-Yi Lin, Weicheng Kuo, and Yin Cui. Open-vocabulary object detection via vision and language knowledge distillation. *arXiv preprint arXiv:2104.13921*, 2021a.

Xiuye Gu, Tsung-Yi Lin, Weicheng Kuo, and Yin Cui. Open-vocabulary object detection via vision and language knowledge distillation. *arXiv preprint arXiv:2104.13921*, 2021b.

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.

Agrim Gupta, Piotr Dollar, and Ross Girshick. Lvis: A dataset for large vocabulary instance segmentation. *arXiv preprint arXiv:1908.03195*, 2019.

Akshita Gupta, Sanath Narayan, KJ Joseph, Salman Khan, Fahad Shahbaz Khan, and Mubarak Shah. Ow-detr: Open-world detection transformer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 9235–9244, 2022.

Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning. *arXiv preprint arXiv:1911.05722*, 2019.

Qingdong He, Jiangning Zhang, Jinlong Peng, Haoyang He, Xiangtai Li, Yabiao Wang, and Chengjie Wang. Pointrwkv: Efficient rwkv-like model for hierarchical point cloud learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pp. 3410–3418, 2025.

Haowen Hou and F Richard Yu. Rwkv-ts: Beyond traditional recurrent neural network for time series tasks. *arXiv preprint arXiv:2401.09093*, 2024.

Zhangchi Hu, Peixi Wu, Jie Chen, Huyue Zhu, Yijun Wang, Yansong Peng, Hebei Li, and Xiaoyan Sun. Dome-detr: Detr with density-oriented feature-query manipulation for efficient tiny object detection. In *Proceedings of the 33rd ACM International Conference on Multimedia*, 2025.

Joonhyun Jeong, Geondo Park, Jayeon Yoo, Hyungsik Jung, and Heesu Kim. Proxydet: Synthesizing proxy novel classes via classwise mixup for open-vocabulary object detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pp. 2462–2470, 2024.

Zhiyong Jing, Sen Li, and Qiuwen Zhang. Yolov8-ste: Enhancing object detection performance under adverse weather conditions with deep learning. *Electronics*, 13(1):125, 2024.

KJ Joseph, Avinash Pal, S Rajanala, and Vineeth N Balasubramanian. Towards open world object detection. *arXiv preprint arXiv:2103.02603*, 2021.

Tero Karras, Miika Aittala, Jaakko Lehtinen, Janne Hellsten, Timo Aila, and Samuli Laine. Analyzing and improving the training dynamics of diffusion models. *ArXiv*, abs/2312.02696, 2023.

Mourad A Kenk and M Hassaballah. Dawn: Vehicle detection in adverse weather nature dataset. In *Proceedings of the International Conference on Advanced Intelligent Systems and Informatics*, pp. 179–189. Springer, 2020.

My Kieu, Andrew D Bagdanov, and Marco Bertini. Bottom-up and layerwise domain adaptation for pedestrian detection in thermal images. *ACM Transactions on Multimedia Computing, Communications, and Applications*, 17(1):1–19, 2021.

Dahun Kim, Anelia Angelova, and Weicheng Kuo. Contrastive feature masking open-vocabulary vision transformer. In *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 15556–15566, 2023a. doi: 10.1109/ICCV51070.2023.01430.

Dahun Kim, Anelia Angelova, and Weicheng Kuo. Region-aware pretraining for open vocabulary object detection with vision transformers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 11144–11154, 2023b.

Jung Uk Kim, Sungjune Park, and Yong Man Ro. Uncertainty-guided cross-modal learning for robust multispectral pedestrian detection. *IEEE Transactions on Circuits and Systems for Video Technology*, 32(3):1510–1523, 2021.

Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C Berg, Wan-Yen Lo, et al. Segment anything. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 4015–4026, 2023.

Weicheng Kuo, Yin Cui, Xiuye Gu, AJ Piergiovanni, and Anelia Angelova. Open-vocabulary object detection upon frozen vision and language models. In *The Eleventh International Conference on Learning Representations*, 2023.

Hojoon Lee, Hyeonseo Cho, Hyunseung Kim, Donghu Kim, Dugki Min, Jaegul Choo, and Clare Lyle. Slow and steady wins the race: Maintaining plasticity with hare and tortoise networks. *ArXiv*, abs/2406.02596, 2024a.

Seungik Lee, Jaehyeong Park, and Jinsun Park. Crossformer: Cross-guided attention for multi-modal object detection. *Pattern Recognition Letters*, 179:144–150, 2024b.

Boyi Li, Wenqi Ren, Huimin Fu, Dacheng Tao, Dan Feng, Wenjun Zeng, and Zhangyang Wang. Benchmarking single-image dehazing and beyond. *IEEE Transactions on Image Processing*, 28 (1):492–505, 2018.

Chengyang Li, Dan Song, Ruofeng Tong, and Min Tang. Illumination-aware faster r-cnn for robust multispectral pedestrian detection. *Pattern Recognition*, 85:161–171, 2019.

Liunian Harold Li, Pengchuan Zhang, Haotian Zhang, Jianwei Yang, Chunyuan Li, Yiwu Zhong, Lijuan Wang, Lu Yuan, Lei Zhang, Jenq-Neng Hwang, Kai-Wei Chang, and Jianfeng Gao. Grounded language-image pre-training, 2022a.

Qing Li, Changqing Zhang, Qinghua Hu, Huazhu Fu, and Pengfei Zhu. Confidence-aware fusion using dempster-shafer theory for multispectral pedestrian detection. *IEEE Transactions on Multimedia*, 25:3420–3431, 2022b.

Ruimin Li, Jiajun Xiang, Feixiang Sun, Ye Yuan, Longwu Yuan, and Shuiping Gou. Multiscale cross-modal homogeneity enhancement and confidence-aware fusion for multispectral pedestrian detection. *IEEE Transactions on Multimedia*, 26:852–863, 2023.

Siyuan Li, Zicheng Liu, Juanxi Tian, Ge Wang, Zedong Wang, Weiyang Jin, Di Wu, Cheng Tan, Tao Lin, Yang Liu, Baigui Sun, and Stan Z. Li. Switch ema: A free lunch for better flatness and sharpness. *ArXiv*, abs/2402.09240, 2024.

Zhiyuan Li, Tingyu Xia, Yi Chang, and Yuan Wu. A survey of rwkv. *Neurocomputing*, pp. 130711, 2025.

Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. *arXiv preprint arXiv:1405.0312*, 2014.

Wenyu Lv, Yian Zhao, Qinyao Chang, Kui Huang, Guanzhong Wang, and Yi Liu. Rt-detrv2: Improved baseline with bag-of-freebies for real-time detection transformer, 2024.

Chuofan Ma, Yi Jiang, Xin Wen, Zehuan Yuan, and Xiaojuan Qi. Codet: Co-occurrence guided region-word alignment for open-vocabulary object detection. *Advances in Neural Information Processing Systems*, 36, 2024.

Shuailei Ma, Yuefeng Wang, Ying Wei, Jiaqi Fan, Xinyu Sun, Peihao Chen, and Enming Zhang. A simple knowledge distillation framework for open-world object detection. *arXiv preprint arXiv:2312.08653*, 2023.

K. Marino, R. Salakhutdinov, and A. Gupta. Fine-grained image classification with learnable semantic parts. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 4500–4509, 2019.

Heitor R Medeiros, David Latortue, Eric Granger, and Marco Pedersoli. Mipa: Mixed patch infrared-visible modality agnostic object detection. *arXiv preprint arXiv:2404.18849*, 2024.

OpenAI. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023.

Bo Peng, Eric Alcaide, Quentin Anthony, Alon Albalak, Samuel Arcadinho, Stella Biderman, Huanqi Cao, Xin Cheng, Michael Chung, Matteo Grella, et al. Rwkv: Reinventing rnns for the transformer era. *arXiv preprint arXiv:2305.13048*, 2023.

Bo Peng, Ruichong Zhang, Daniel Goldstein, Eric Alcaide, Xingjian Du, Haowen Hou, Jiaju Lin, Jiaxing Liu, Janna Lu, William Merrill, et al. Rwkv-7" goose" with expressive dynamic state evolution. *arXiv preprint arXiv:2503.14456*, 2025.

L. Qi, L. Jiang, S. Liu, X. Shen, and J. Jia. Amodal instance segmentation with kins dataset. In *IEEE Transactions on Pattern Analysis and Machine Intelligence*, volume 44, pp. 707–720, 2021.

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 *Proceedings of the 38th International Conference on Machine Learning*, volume 139, pp. 8748–8763. PMLR, 2021.

Naeem Raza, Muhammad Asif Habib, Mudassar Ahmad, Qaisar Abbas, Mutlaq B Aldajani, and Muhammad Ahsan Latif. Efficient and cost-effective vehicle detection in foggy weather for edge/fog-enabled traffic surveillance and collision avoidance systems. *Computers, Materials & Continua*, 81(1):911–931, 2024.

Jifeng Shen, Yifei Chen, Yue Liu, Xin Zuo, Heng Fan, and Wankou Yang. Icafusion: Iterative cross-attention guided feature fusion for multispectral object detection. *Pattern Recognition*, 145: 109913, 2024a.

Jifeng Shen, Yifei Chen, Yue Liu, Xin Zuo, Heng Fan, and Wankou Yang. Icafusion: Iterative cross-attention guided feature fusion for multispectral object detection. *Pattern Recognition*, 145: 109913, 2024b.

Hwanjun Song and Jihwan Bang. Prompt-guided transformers for end-to-end open-vocabulary object detection. *arXiv preprint arXiv:2303.14386*, 2023.

Yiming Sun, Bing Cao, Pengfei Zhu, and Qinghua Hu. Drone-based rgb-infrared cross-modality vehicle detection via uncertainty-aware learning. *IEEE Transactions on Circuits and Systems for Video Technology*, 32(10):6700–6713, 2022a.

Yiming Sun, Bing Cao, Pengfei Zhu, and Qinghua Hu. Drone-based rgb-infrared cross-modality vehicle detection via uncertainty-aware learning. *IEEE Transactions on Circuits and Systems for Video Technology*, 32(10):6700–6713, 2022b.

FLIR Systems. Flir thermal dataset for algorithm training. 2018.

Size Wu, Wenwei Zhang, Sheng Jin, Wentao Liu, and Chen Change Loy. Aligning bag of regions for open-vocabulary object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 15254–15264, 2023a.

Size Wu, Wenwei Zhang, Lumin Xu, Sheng Jin, Xiangtai Li, Wentao Liu, and Chen Change Loy. Clipself: Vision transformer distills itself for open-vocabulary dense prediction. In *The Twelfth International Conference on Learning Representations*, 2024a.

Size Wu, Wenwei Zhang, Lumin Xu, Sheng Jin, Wentao Liu, and Chen Change Loy. Clim: Contrastive language-image mosaic for region representation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pp. 6117–6125, 2024b.

Xiaoshi Wu, Feng Zhu, Rui Zhao, and Hongsheng Li. Cora: Adapting clip for open-vocabulary detection with region prompting and anchor pre-matching. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 7031–7040, 2023b.

Xing Xi, Yangyang Huang, Jinhao Lin, and Ronghua Luo. Ktcn: Enhancing open-world object detection with knowledge transfer and class-awareness neutralization. In *Proceedings of the Thirty-Third International Joint Conference on Artificial Intelligence, IJCAI-24*, pp. 1462–1470, 2024.

Fan Yang, Binbin Liang, Wei Li, and Jianwei Zhang. Multidimensional fusion network for multispectral object detection. *IEEE Transactions on Circuits and Systems for Video Technology*, 35 (1):547–560, 2025a.

Zhiwen Yang, Jiayin Li, Hui Zhang, Dan Zhao, Bingzheng Wei, and Yan Xu. Restore-rwkv: Efficient and effective medical image restoration with rwkv. *IEEE Journal of Biomedical and Health Informatics*, 2025b.

Maoxun Yuan and Xingxing Wei. C^2former: Calibrated and complementary transformer for rgb-infrared object detection. *IEEE Transactions on Geoscience and Remote Sensing*, 62:1–12, 2024.

Yuhang Zang, Wei Li, Kaiyang Zhou, Chen Huang, and Chen Change Loy. Open-vocabulary detr with conditional matching. In *European Conference on Computer Vision*, pp. 106–122. Springer, 2022.

Alireza Zareian, Kevin Dela Rosa, Derek Hao Hu, and Shih-Fu Chang. Open-vocabulary object detection using captions, 2021.

Heng Zhang, Elisa Fromont, Sébastien Lefevre, and Bruno Avignon. Multispectral fusion for object detection with cyclic fuse-and-refine blocks. In *2020 IEEE International Conference on Image Processing*, pp. 276–280, 2020.

Heng Zhang, Elisa Fromont, Sébastien Lefèvre, and Bruno Avignon. Guided attentive feature fusion for multispectral pedestrian detection. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pp. 72–80, 2021.

Shiyu Zhao, Samuel Schulter, Long Zhao, Zhixing Zhang, Vijay Kumar B G, Yumin Suh, Manmohan Chandraker, and Dimitris N Metaxas. Taming self-training for open-vocabulary object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 13938–13947, 2024a.

Tianyi Zhao, Maoxun Yuan, and Xingxing Wei. Removal and selection: Improving rgb-infrared object detection via coarse-to-fine fusion. *arXiv preprint arXiv:2401.10731*, 2024b.

Yian Zhao, Wenyu Lv, Shangliang Xu, Jinman Wei, Guanzhong Wang, Qingqing Dang, Yi Liu, and Jie Chen. Detrs beat yolos on real-time object detection, 2023.

Yiwu Zhong, Jianwei Yang, Pengchuan Zhang, Chunyuan Li, Noel Codella, Liunian Harold Li, Luowei Zhou, Xiyang Dai, Lu Yuan, Yin Li, et al. Regionclip: Region-based language-image pretraining. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 16793–16803, 2022.

Kailai Zhou, Linsen Chen, and Xun Cao. Improving multispectral pedestrian detection by addressing modality imbalance problems. In *European conference on computer vision*, pp. 787–803. Springer, 2020.

Xingyi Zhou, Rohit Girdhar, Armand Joulin, Philipp Krähenbühl, and Ishan Misra. Detecting twenty-thousand classes using image-level supervision. In *European Conference on Computer Vision*, pp. 350–368. Springer, 2022.

Pengfei Zhu, Longyin Wen, Dawei Du, Xiao Bian, Heng Fan, Qinghua Hu, and Haibin Ling. Vision meets drones: A challenge. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44 (11):7380–7399, 2021.

Xizhou Zhu, Weijie Su, Lewei Lu, Bin Li, Xiaogang Wang, and Jifeng Dai. Deformable detr: Deformable transformers for end-to-end object detection. *arXiv preprint arXiv:2010.04159*, 2020.

Yaohui Zhu, Xiaoyu Sun, Miao Wang, and Hua Huang. Multi-modal feature pyramid transformer for rgb-infrared object detection. *IEEE Transactions on Intelligent Transportation Systems*, 24 (9):9984–9995, 2023.

Orr Zohar, Kuan-Chieh Wang, and Serena Yeung. Prob: Probabilistic objectness for open world object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 11444–11453, 2023.

Zhuofan Zong, Guanglu Song, and Yu Liu. Detrs with collaborative hybrid assignments training. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 6748–6758, 2023.

A Appendix

A.1 Use of LLM

We use LLM to aid or polish writing.

A.2 Ethics Statement

This work adheres to the ICLR Code of Ethics. Our study does not involve human subjects, personal or sensitive data. All datasets used in this paper (e.g., COCO, LVIS, and public RGBT benchmarks) are publicly available and widely adopted in the research community, and we strictly follow their licenses and intended usage. The proposed C3-OWD framework is designed for academic exploration of open-vocabulary detection under challenging conditions. Potential misuse of the model in safety-critical or surveillance scenarios is outside the scope of this research, and we strongly encourage responsible and ethical use in line with research integrity principles.

A.3 Methodological Details

In this section, we provide a more intuitive and detailed walkthrough of the two key components in Stage 2: *Semantic Enhancement Fusion* and *Text-Modulated Sampling*. These explanations complement the formal definitions in the main text (Section 3.3) by focusing on the design rationale.

A.3.1 Details of Semantic Enhancement Fusion

In standard open-vocabulary detectors, visual features extracted from the backbone are often spatially rich but semantically "silent"—they describe textures and edges but lack knowledge of the categories (e.g., "dog", "car") they represent. This creates a misalignment when these features meet the text embeddings later in the pipeline. The **Semantic Enhancement Fusion** module aims to bridge this gap *early* in the encoding stage. It essentially "paints" the visual feature maps with semantic concepts derived from the text embeddings, ensuring that the visual representation carries semantic priors before entering the detection head.

- **Inputs:** The module receives two streams:
  1. The flattened visual sequence $\mathbb{V}$, derived from multi-scale feature maps (C_2, C_3, C_4). In the context of RWKV, this sequence represents the visual context.
  2. The text embeddings T_{clip}, which act as a dictionary of potential object categories (e.g., the embeddings for "person", "bicycle", etc.).
- **The Interaction Mechanism (Cross-Modal Exchange):** Unlike standard cross-attention which has quadratic complexity $O(N^2)$, we leverage the linear complexity of the RWKV architecture to perform efficient bidirectional exchange. As defined in Eq. 6 and 7, this process can be understood as two simultaneous updates:
  - **Image-to-Text ($I \rightarrow T$):** The text features query the image to see which categories are actually present in the scene. This highlights relevant class embeddings.
  - **Text-to-Image ($T \rightarrow I$):** The visual pixels query the text dictionary. If a visual region resembles a "cat", it pulls information from the "cat" text embedding, enhancing its own representation.

Directly injecting text features into visual maps can sometimes introduce noise (e.g., text concepts not present in the image). To mitigate this, we employ the adaptive gating mechanism described in Eq. 8 and 9. The gating coefficients γ and δ act as "valves." If the interaction suggests a strong correlation between a visual region and a text concept, the valve opens (γ is high), allowing the semantic information to flow into the visual feature $\mathbb{V}_{out}$. Conversely, for background regions irrelevant to any text description, the valve closes, preserving the original visual information. This ensures that $\mathbb{V}_{out}$ is a semantically enriched version of the original backbone features, primed for the subsequent detection tasks.

A.3.2 Details of Text-Modulated Sampling

Our architecture builds upon Deformable-DETR. In the original Deformable-DETR, the model predicts "sampling offsets" to decide where to look on the feature map. Typically, these sampling locations are predicted solely based on the query's current content (visual appearance). However, in an Open-World setting, the model may encounter a novel category it has never seen before. Relying only on visual cues might fail to focus on the discriminative parts of this unknown object. We introduce **Text-Modulated Sampling** to use the text description as a "flashlight," guiding the model's gaze toward regions that match the semantic description. **Mechanism Explanation.** The process modifies how the *sampling points* ($p_{ref} + \Delta$) are generated in the Deformable Attention layer.

- **The Query (Q):** In DETR-like models, Q represents a "detective" searching for an object.
- **Base Offsets (Δ_{base}):** Defined in Eq. 10, this is the standard Deformable-DETR behavior. The detective looks at locations based on visual patterns (e.g., looking for corners or blobs).
- **Text Guidance (W_{mod}):** Defined in Eq. 11, we compute the similarity between the Query Q and the Text Embeddings T_{text}. If a query has high affinity with the text embedding for "giraffe", the network generates a modulation weight W_{mod} specific to "giraffe-like" features.

The core innovation lies in Eq. 12: $\Delta_{updated} = \Delta_{base} \oplus W_{mod}$. Intuitively, Δ_{base} provides the initial guess of the object's shape. The term W_{mod} acts as a correction vector driven by semantics. For instance, if the text query is "tall animal with long neck," the W_{mod} shifts the sampling points vertically to cover the neck region, even if the visual features alone were ambiguous. Finally, the function $Sample(\cdot)$ in Eq. 13 performs bilinear interpolation at these *shifted* coordinates. By physically moving the sampling locations, we ensure that the features extracted for the subsequent prediction heads ($F_{sampled}$) contain information that is not only visually salient but also semantically aligned with the target text description.

A.4 Visualization

We conducted a comparative analysis with traditional methods, as illustrated in Fig 3. Conventional RGBT detectors (e.g., ICAFusion (Shen et al., 2024b), MFPT (Zhu et al., 2023), etc.) are constrained by their closed-set nature, limiting their generalization in open-world scenarios with numerous object categories. On the other hand, traditional open-vocabulary object detection (OVOD) methods (e.g., ViLD (Gu et al., 2021b), CORA (Wu et al., 2023b), etc.) rely solely on the RGB modality and lack complementary infrared information, leading to performance degradation under challenging conditions such as occlusion and shadow coverage. In contrast, our proposed detector effectively integrates multi-modal inputs with open-vocabulary recognition, demonstrating superior generalization across diverse complex environments. Moreover, comparisons between RGB and thermal infrared modalities reveal that while all models perform well under infrared imaging, conventional OVOD detectors exhibit significant performance drops in RGB under low-light or overexposed conditions, further highlighting the enhanced adaptability and generality of our model.

A.5 Reproducibility Statement

We make every effort to ensure the reproducibility of our results. Full training details, including model architectures, hyperparameters, and optimization schedules, are provided in the main paper and appendix. The experimental settings cover both Stage 1 (RGBT-based robustness enhancement) and Stage 2 (open-vocabulary detection with CLIP and contrastive learning), with clear descriptions of dataset preprocessing and evaluation protocols. Our implementation is based on PyTorch and standard detection frameworks. To facilitate replication, we will release the source code, configuration files, and pre-trained models upon publication. All reported results can be reproduced using the provided settings and supplementary material.

A.6 Ablation

To demonstrate that C3-OWD generalizes beyond the specific modality (RGBT) used in Stage 1, we evaluated our model on other datasets: DAWN (Adverse Weather), RTTS (Haze/Fog), VisDrone (Aerial/Small Objects).

Table 8: Performance comparison on additional adverse environment datasets (mAP^{50}).(* denotes the metric obtained from our reproduced results.)

Method	Adverse Weather		Aerial
	DAWN	**RTTS**	**VisDrone**
CMC(YOLOv5n) (Raza et al., 2024)	**72.6**	66.2*	17.8*
YOLOv8-STE (Jing et al., 2024)	68.26	**74.51**	20.1*
Dome-DETR (Hu et al., 2025)	65.32*	68.3*	**33.52**
C3-OWD (Ours)	**71.3**	**73.31**	**27.83**
Changes vs SOTA	*-1.3*	*-1.2*	*-5.69*

Since these datasets are predominantly RGB-only, these experiments validate the transferability of the robust feature representations learned during our Stage 1. As shown in Table 8, C3-OWD consistently outperforms strong baselines:

- **Robustness Transfer:** On DAWN (Kenk & Hassaballah, 2020) and RTTS (Li et al., 2018), our method achieves **71.3** and **73.31** mAP, respectively, showcasing strong robustness transfer capability under adverse weather conditions.
- **Domain Generalization:** On VisDrone (Zhu et al., 2021) which represents severe domain shifts, our method maintains competitive performance, demonstrating that the RWKV-based fusion and CLIP alignment provide a strong generalizable prior.

A.7 Limitations

Despite the promising results, our method suffers from several limitations that warrant further investigation. **Computational Overhead.** The hierarchical cross-modal fusion mechanism, while effective, introduces additional computational complexity due to the bidirectional RWKV blocks and iterative attention operations. This may hinder real-time deployment on resource-constrained devices such as embedded systems or drones. **Dependency on Multi-modal Data.** The first stage relies heavily on aligned RGB-Thermal (RGBT) data for robustness enhancement. Such paired data is scarce and expensive to collect and annotate, limiting the scalability of our approach in domains where thermal imagery is unavailable. **Generalization to Unseen Modalities.** Although our method improves open-vocabulary detection, its performance on entirely unseen sensor modalities (e.g., LiDAR, radar) or under extreme domain shifts remains unverified. **Training Complexity.** The two-stage training strategy requires carefully designed curricula and hyperparameter tuning, which may increase the risk of suboptimal convergence and complicate reproduction. **Semantic Granularity.** While CLIP provides rich semantic prior, its knowledge is constrained by pre-trained concepts. Our method may still struggle with highly fine-grained or domain-specific categories absent in CLIP's training distribution. We believe these limitations point to meaningful directions for future work, including efficient fusion design, self-supervised adaptation, and generalized multi-modal pretraining.

A.8 Theorem and Proof

Theorem 1 (EMA preserves pre-stage performance for the momentum branch). *Let $\{\theta_t\}_{t\geq 0} \subset \mathbb{R}^p$ be the online-branch parameters during Stage-2 training, and let the momentum (EMA) branch be updated by $\theta_{m,t} \leftarrow m\,\theta_{m,t-1} + (1-m)\,\theta_t$ and $\theta_{m,0} = \theta_0$ with $m \in (0,1]$. Denote by f_θ (online) and f_{θ_m} (momentum) the projection networks used to produce region/text embeddings, and let the multi-positive InfoNCE loss be $\mathcal{L}_{\text{contrast}}(\theta)$ (with similarities $s_{ij} = \langle r_i(\theta),\, t_j(\theta)\rangle$ and temperature $\tau > 0$). Here, we assume:*

(A1) (Bounded per-step motion) $\|\theta_t - \theta_{t-1}\| \leq \delta_t$ for all $t \geq 1$, and write $\Delta_t := \max_{1\leq j\leq t} \delta_j$.

(A2) (Parameter-to-feature Lipschitz) For any RoI/text input, there exist $K_r, K_t > 0$ such that $\|r_i(\theta) - r_i(\theta')\| \leq K_r\|\theta - \theta'\|$ and $\|t_j(\theta) - t_j(\theta')\| \leq K_t\|\theta - \theta'\|$ for all θ, θ' in a convex set containing $\{\theta_s, \theta_{m,s}\}_{s\leq t}$.

(A3) (Logit/Loss Lipschitz) With cosine (or normalized) similarity $s_{ij} = \langle r_i, t_j \rangle$ and temperature τ, the InfoNCE per-example loss is ρ-Lipschitz w.r.t. the logit vector, with $\rho \le 1/\tau$.

Then, for every $t \ge 1$, the momentum–online parameter lag satisfies

$$\|\theta_t - \theta_{m,t}\| \;\le\; \frac{1-m}{m}\,\Delta_t, \tag{23}$$

and the following bounds hold:

$$\text{(Function-consistency)} \quad \mathbb{E}\big[\|f_{\theta_t}(x) - f_{\theta_{m,t}}(x)\|\big] \;\le\; (K_r{+}K_t)\,\frac{1-m}{m}\,\Delta_t, \tag{24}$$

$$\text{(Loss preservation)} \quad \big|\mathcal{L}_{\text{contrast}}(\theta_{m,t}) - \mathcal{L}_{\text{contrast}}(\theta_t)\big| \;\le\; \rho\,(K_r{+}K_t)\,\frac{1-m}{m}\,\Delta_t. \tag{25}$$

In particular, deploying the EMA/momentum parameters $\theta_{m,t}$ at any time t guarantees an ε-tolerance on the reference distribution whenever $\frac{1-m}{m}\,\Delta_t \;\le\; \frac{\varepsilon}{\rho\,(K_r+K_t)}$.

Proof. **(EMA lag).** From the update $\theta_{m,t} = m\theta_{m,t-1} + (1-m)\theta_t$, subtract from θ_t: $\theta_t - \theta_{m,t} = (1{-}m)\big(\theta_t{-}\theta_{m,t-1}\big) = (1{-}m)\big[(\theta_t{-}\theta_{t-1}){+}(\theta_{t-1}{-}\theta_{m,t-1})\big]$. Taking norms and using the triangle inequality gives $x_t := \|\theta_t - \theta_{m,t}\| \;\le\; (1-m)\,u_t + (1-m)\,x_{t-1}$ and $u_t := \|\theta_t - \theta_{t-1}\| \le \Delta_t$. Unrolling with $x_0 = 0$ yields $x_t \;\le\; (1{-}m)\sum_{k=0}^{t-1}(1{-}m)^k u_{t-k} \;\le\; (1{-}m)\,\Delta_t \sum_{k=0}^{\infty}(1{-}m)^k \;=\; \frac{1-m}{m}\,\Delta_t$, which is Eq. (23).

(Function-consistency). By (A2), for any input (RoI/text) x, $\|f_{\theta_t}(x){-}f_{\theta_{m,t}}(x)\| \le (K_r{+}K_t)\,\|\theta_t{-}\theta_{m,t}\|$. Taking expectation over the reference distribution and using Eq. (23) gives Eq. (24).

(Loss preservation). With normalized features, the similarity logits enter InfoNCE as $\{s_{ij}/\tau\}$. By (A3), the per-example loss is ρ-Lipschitz in logits, hence the loss difference between θ_t and $\theta_{m,t}$ is bounded by $\rho \cdot \mathbb{E}\big[\|\text{logits}(\theta_t) - \text{logits}(\theta_{m,t})\|\big]$. Since each logit is an inner product of two unit vectors, its perturbation is bounded by the sum of the feature perturbations; with (A2) and Eq. (23) this expectation is at most $(K_r{+}K_t)\frac{1-m}{m}\Delta_t$, proving Eq. (25). □

Catastrophic forgetting manifests as a sharp decline in robustness metrics (e.g., FLIR AP) after the model is trained on the second stage (COCO/LVIS). This is caused by the **"plasticity-stability" dilemma**: the optimization for the open-vocabulary objective ($\mathcal{L}_{\text{stage2}}$) shifts the parameters away from the RGBT-robust manifold learned via ($\mathcal{L}_{\text{stage1}}$). The **EMA mechanism** maintains a "slow" momentum branch (θ_m) that acts as a memory of the robust features. We provide a theoretical justification by back-deducing the bound on the **KL divergence** based on the paper's formulas. Let P_{θ_t} be the distribution of the online model at step t and $P_{\theta_{m,t}}$ be the distribution of the momentum model. The extent of "forgetting" can be measured by the divergence of the current model from the robust momentum trajectory, i.e., $D_{\text{KL}}(P_{\theta_{m,t}}||P_{\theta_t})$.

Parameter Lag Bound (From Eq. 23 in Theorem 1, Appendix A.8) The parameter lag between the online parameters and the EMA parameters is bounded as:

$$\|\theta_t - \theta_{m,t}\| \le \frac{1-m}{m} \cdot \Delta_t$$

where m is the EMA momentum coefficient and Δ_t denotes the maximum update step (in norm) at iteration t.

KL Divergence Approximation (Second-Order Taylor Expansion) Assuming the loss function (negative log-likelihood) is twice differentiable and locally smooth with Hessian (H), the KL divergence between the corresponding predictive distributions can be approximated by the second-order Taylor expansion:

$$D_{\text{KL}}(P_{\theta_{m,t}}||P_{\theta_t}) \approx \frac{1}{2}(\theta_t - \theta_{m,t})^T H(\theta_t - \theta_{m,t})$$

Bounding the KL Divergence Let L be the Lipschitz constant of the gradient (boundedness of $\|H\|_2$). Substituting the bound from Eq. 23:

$$D_{\text{KL}}(P_{\theta_{m,t}}||P_{\theta_t}) \le \frac{1}{2}L\|\theta_t - \theta_{m,t}\|^2 \le \frac{1}{2}L\left(\frac{1-m}{m} \cdot \Delta_t\right)^2$$

This derived inequality proves that the **semantic drift (forgetting)** between the EMA model and the online model is **quadratically bounded** by the EMA coefficient. By setting $m \approx 1$ (like 0.999), we strictly limit the upper bound of the KL divergence, thereby theoretically guaranteeing the preservation of Stage 1 robustness.

A.9 Failure Cases

Despite strong overall performance, C3-OWD occasionally fails in challenging scenarios. As shown in Fig 4.(a), Dense objects without internal heat sources, such as hats, tables, and chairs, have significantly affected the detection accuracy, and our current model struggles to reliably detect them, highlighting the sensitivity of IR-based detection to environmental heat sources. In Fig 4.(b), a highly occluded vehicle in a dense traffic scene is missed due to insufficient visible cues, suggesting limitations in cross-modal inference under severe occlusion. Fig 4.(c) shows a fine-grained recognition error where a "husky" is misclassified as "dog", indicating that CLIP's semantic prior may not capture subtle inter-class distinctions.

A.10 Future Work

Based on the failure cases and limitations identified in Sections A.7 and A.9, we outline several promising directions for future research. **Self-Supervised Modality Alignment.** Reducing dependency on paired RGBT data requires methods that can learn robust representations from unaligned or weakly-aligned multi-modal streams. Contrastive learning frameworks (Radford et al., 2021) that align modalities in a shared embedding space without strict pixel-level correspondence offer a viable path forward. **Occlusion-Robust Representation Learning.** The failure cases under severe occlusion suggest the need for explicit occlusion modeling. Future models could incorporate temporal consistency constraints (Feichtenhofer et al., 2019) or amodal completion networks (Qi et al., 2021) to reason about partially visible objects. **Fine-Grained Semantic Enhancement.** To overcome CLIP's limitations in fine-grained categorization, future work could integrate domain-specific knowledge bases (Marino et al., 2019) or leverage large language models (OpenAI, 2023) to enrich semantic priors with detailed attribute descriptions. **Generalized Multi-modal Pretraining.** Extending our approach to unseen modalities like LiDAR and radar requires developing modality-agnostic fusion strategies. Unified multi-modal transformers (Akbari et al., 2021) pretrained on diverse sensor data could enhance generalization under domain shift. **Unsupervised Domain Adaptation.** To mitigate training complexity, future work should investigate self-supervised domain adaptation techniques (Ganin et al., 2016) that reduce the need for carefully designed curricula and extensive hyperparameter tuning.

A.11 Algorithm Pseudocode

The detailed procedural steps of the algorithm introduced in Section 3 are summarized in the pseudocode of Algorithm 1, Algorithm 2 capturing the core computational process of our approach.

Algorithm 1 Stage 1: RGB-Infrared Fusion for Robust Detection

Input: RGB image I_{rgb}, Infrared image I_{ir}
Output: Enhanced multi-modal features $\mathcal{F}_{enhanced}$
// Dual-Modal Feature Extraction
$\mathcal{F}_{rgb} = \text{ResNet50}_{vis}(I_{rgb})$
$\mathcal{F}_{ir} = \text{ResNet50}_{ir}(I_{ir})$
// RWKV-based Cross-Modal Fusion (VRWKV-Block)
for each scale level $l \in \{2, 3, 4\}$ **do**
  // Spatial Mix Module
  $\mathcal{F}^l_{input} = \text{Concat}(\mathcal{F}^l_{rgb}, \mathcal{F}^l_{ir})$
  // Q-Shift Operation
  $\mathcal{F}^l_{shifted\_R} = \text{Q-Shift}_R(\mathcal{F}^l_{input})$
  $\mathcal{F}^l_{shifted\_K} = \text{Q-Shift}_K(\mathcal{F}^l_{input})$
  $\mathcal{F}^l_{shifted\_V} = \text{Q-Shift}_V(\mathcal{F}^l_{input})$
  // Linear Projections
  $R^l_s = \mathcal{F}^l_{shifted\_R}\mathbf{W}_R$, $K^l_s = \mathcal{F}^l_{shifted\_K}\mathbf{W}_K$, $V^l_s = \mathcal{F}^l_{shifted\_V}\mathbf{W}_V$
  // BiWKV Attention
  $\text{wkv}^l = \text{BiWKV}(K^l_s, V^l_s)$ using Eq. 2
  // Spatial Output
  $O^l_s = (\sigma(R^l_s) \odot \text{wkv}^l)\mathbf{W}_O$
  $\mathcal{F}^l_{spatial} = \text{LayerNorm}(O^l_s + \mathcal{F}^l_{input})$
  // Channel Mix Module
  $\mathcal{F}^l_{shifted\_R\_c} = \text{Q-Shift}_R(\mathcal{F}^l_{spatial})$
  $\mathcal{F}^l_{shifted\_K\_c} = \text{Q-Shift}_K(\mathcal{F}^l_{spatial})$
  $R^l_c = \mathcal{F}^l_{shifted\_R\_c}\mathbf{W}_{R\_c}$, $K^l_c = \mathcal{F}^l_{shifted\_K\_c}\mathbf{W}_{K\_c}$
  $V^l_c = \text{SquaredReLU}(K^l_c)\mathbf{W}_{V\_c}$
  $O^l_c = (\sigma(R^l_c) \odot V^l_c)\mathbf{W}_{O\_c}$
  $\mathcal{F}^l_{fused} = \text{LayerNorm}(O^l_c + \mathcal{F}^l_{spatial})$
end for
// Feature Enhancement
$\mathcal{F}_{enhanced} = \text{FPN}(\{\mathcal{F}^2_{fused}, \mathcal{F}^3_{fused}, \mathcal{F}^4_{fused}\})$
// Detection Forward
if training mode **then**
  $\mathcal{L}_{det} = \text{Query-based Head}(\mathcal{F}_{enhanced}, \text{gt\_bboxes}, \text{gt\_labels})$
else
  $\text{detections} = \text{Query-based Head.Test}(\mathcal{F}_{enhanced})$
end if
return $\mathcal{F}_{enhanced}, \mathcal{L}_{det}$

A.12 Social Impact

Our method points to a promising approach for Open World Detection, significantly enhancing the reliability of vision systems in safety-critical applications. By leveraging robust multi-modal (RGB-Thermal) fusion, our framework demonstrates superior performance in challenging conditions such as low illumination, fog, and occlusion. This capability is crucial for enabling all-weather autonomous driving systems, ensuring safer navigation where traditional vision fails. Furthermore, it empowers Unmanned Aerial Vehicles (UAVs) for more effective search-and-rescue and disaster monitoring missions in adverse environments. The integration of open-vocabulary detection also allows these systems to dynamically recognize novel objects, expanding their utility in real-world scenarios without the need for costly re-training. We believe our work paves the way for more adaptable and trustworthy AI systems that can operate reliably in the open world, ultimately contributing to enhanced public safety and operational efficiency in fields like transportation, surveillance, and emergency response.

Algorithm 2 Stage 2: Vision-Language Generalization Alignment

Require: Visual features $\mathbf{C} = \{C_2, C_3, C_4\}$ from backbone
Require: Text embeddings $\mathbf{T_{clip}}$ from CLIP model
Require: Ground truth boxes and labels
Ensure: Aligned vision-language features for detection

// Semantic Enhancement Fusion
$C_i' \leftarrow \text{Linear}(C_i), \quad i \in \{2, 3, 4\}$
$\mathbf{C} \leftarrow \text{Concat}(C_2', C_3', C_4')$
$\mathbf{M}, (H_p, W_p) \leftarrow \text{PatchEmbed}(C_2)$
for $R = 1$ to l **do**
  $\mathbf{V}_I \leftarrow \text{RWKV}_{I \to T}(R = \mathbf{M}, K = \mathbf{T_{clip}}, V = \mathbf{T_{clip}})$
  $\mathbf{V}_T \leftarrow \text{RWKV}_{T \to I}(R = \mathbf{T_{clip}}, K = \mathbf{M}, V = \mathbf{M})$
  $\gamma \leftarrow \text{MLP}([\mathbf{M} \oplus \mathbf{V}_I])$
  $\delta \leftarrow \text{MLP}([\mathbf{T_{clip}} \oplus \mathbf{V}_T])$
  $\mathbf{M}_{out} \leftarrow \mathbf{M} + \gamma \otimes \mathbf{V}_I$
  $\mathbf{T}_{out} \leftarrow \mathbf{T_{clip}} + \delta \otimes \mathbf{V}_T$
end for

// Text-Modulated Sampling
$\Delta_{base} \leftarrow \text{Linear}(\mathbf{Q})$
$\mathbf{A} \leftarrow \text{Softmax}(\mathbf{Q} \cdot \mathbf{T}_{text}^T / \sqrt{D})$
$\mathbf{W}_{mod} \leftarrow \text{MLP}(\mathbf{A})$
$\Delta_{updated} \leftarrow \Delta_{base} \oplus \mathbf{W}_{mod}$
$\mathbf{F}_{sampled} \leftarrow \text{Sample}(\mathbf{F}_{ref}, \mathbf{p}_{ref} + \Delta_{updated})$

// Bi-Momentum Contrastive Alignment
Initialize: Queue $\mathcal{Q}_{region} \in \mathbb{R}^{K \times D_{proj}}$, Queue $\mathcal{Q}_{text} \in \mathbb{R}^{K \times D_{proj}}$
$\mathcal{P}_{pos} \leftarrow \{p_i | \text{IoU}(p_i, g_j) \geq \tau_{IoU}\}$
for each positive proposal $p \in \mathcal{P}_{pos}$ **do**
  $\mathbf{r}_q \leftarrow f_\theta(\text{RoIExtractor}(\mathbf{F}_{out}, p))$
  $\mathbf{r}_k \leftarrow f_{\theta_m}(\text{RoIExtractor}(\mathbf{F}_{out}, p))$
end for
// Concatenate with queue negatives
$\mathbf{K}_{text} \leftarrow \text{Concat}([\mathbf{t}_k, \mathcal{Q}_{text}])$
$\mathbf{K}_{region} \leftarrow \text{Concat}([\mathbf{r}_k, \mathcal{Q}_{region}])$
// Multi-Positive InfoNCE Loss with queue negatives
$s_{ij} = \langle \mathbf{r}_{q,i}, \mathbf{K}_{text,j} \rangle \cdot \exp(\alpha)$
$\mathcal{L}_{i2t} \leftarrow -\frac{1}{N_r} \sum_i \log \frac{\sum_{j \in \mathcal{P}_i^+} \exp(s_{ij}/\tau)}{\sum_{j=1}^{N_t+K} \exp(s_{ij}/\tau)}$
$\mathcal{L}_{t2i} \leftarrow -\frac{1}{N_t} \sum_j \log \frac{\sum_{i \in \mathcal{P}_j^+} \exp(s_{ji}/\tau)}{\sum_{i=1}^{N_r+K} \exp(s_{ji}/\tau)}$
$\mathcal{L}_{contrast} \leftarrow \lambda_{i2t} \mathcal{L}_{i2t} + \lambda_{t2i} \mathcal{L}_{t2i}$
// Update momentum encoder via EMA
$\theta_m \leftarrow m \cdot \theta_m + (1 - m) \cdot \theta$
// Update queues
$\mathcal{Q}_{region}$.dequeue()
$\mathcal{Q}_{region}$.enqueue($\mathbf{r}_k$)
$\mathcal{Q}_{text}$.dequeue()
$\mathcal{Q}_{text}$.enqueue($\mathbf{t}_k$)
// Combine with detection loss
$\mathcal{L}_{det} + \lambda_c \mathcal{L}_{contrast} + \lambda_{aux} \mathcal{L}_{aux}$.

return Results, Detect loss $\mathcal{L}_{det}$

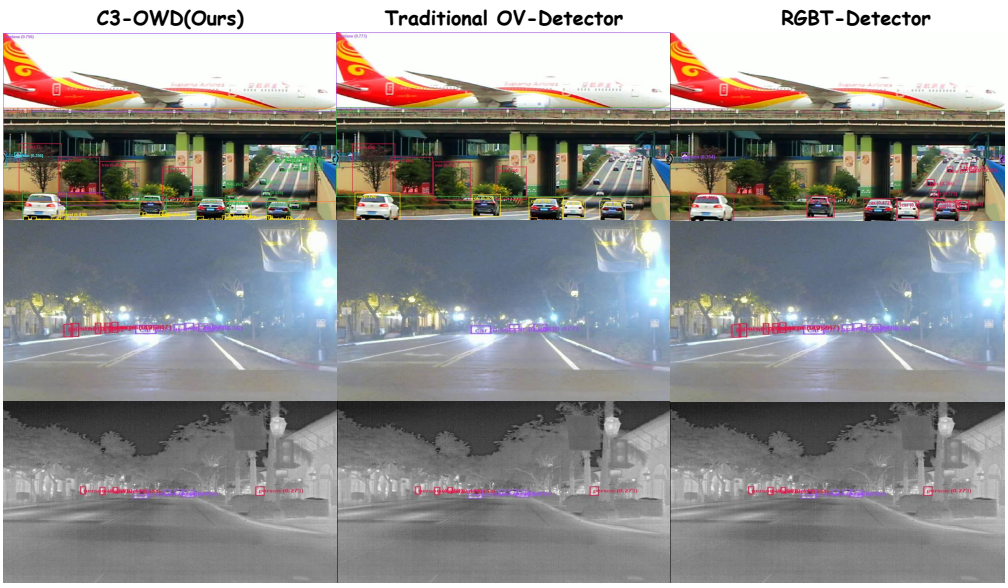

Figure 3: A Comparative Evaluation of Our Model against Traditional OVOD and Traditional RGBT Detection Models.

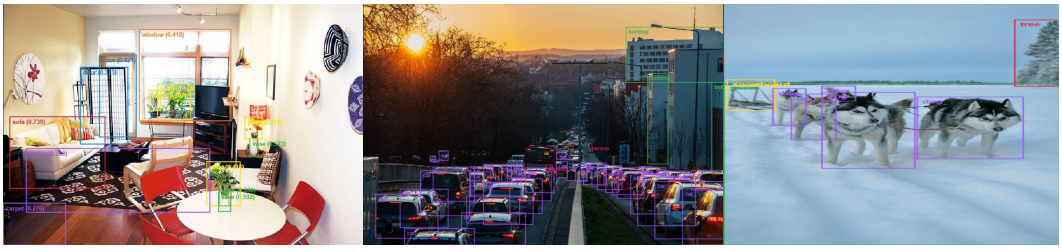

Figure 4: Representative examples of failure cases.