Rethinking Pre-Training and Augmentation for Zero-Shot Cross-City Object Detection

Long Hoang Pham*, Quoc Pham-Nam Ho*, Huy-Hung Nguyen*, Duong Nguyen-Ngoc Tran, Ngoc Doan-Minh Huynh, Cu Quoc Le, Hoang Khang Nguyen, Hyung-Min Jeon, Chi Dai Tran, Son Hong Phan, Duong Khac Vu, Trinh Le Ba Khanh, and Jae Wook Jeon†

Automation Lab,
Department of Electrical and Computer Engineering,
Sungkyunkwan University, South Korea
{phlong,hpnquoc,huyhung91,jwjeon}@skku.edu

Abstract. Real-world deployment of traffic surveillance systems is bottlenecked by geographic domain shift, in which models trained in one city underperform when applied to an unseen target city. Conventional domain adaptation relies on hyperparameter-sensitive architectures or direct profiling of target data. Both are fundamentally precluded in privacy-conscious ecosystems that require completely blind training and evaluation loops. In this setting, we explore the effects of pre-training and augmentation in addressing the domain shift problem. Specifically, we propose a new modular training pipeline for object detection structured around two core orthogonal pillars: (1) a multi-dataset pre-training strategy featuring a class-agnostic objectness distillation to decouple structural vehicle geometry from semantic taxonomies, and (2) a domain-resilient augmentation stream featuring a novel Grayworld transformation that forces global attention heads to strip volatile chromatic shortcuts in favor of robust shape priors. When evaluated with the real-time transformer-based detector RF-DETR, our framework bridges cross-city distribution gaps while using limited GPU memory (16GB). Our optimized variants, RF-DETR-HR and RF-DETR-Grayworld, deliver a substantial empirical gain of +24.29 over the baseline, achieving 1st place (47.53 mAP) on the AI City Challenge Track 6 leaderboard. Code and data are available at: SKKUAutoLab/aic26_cross_city.

Keywords: Cross-city object detection · Domain adaptation · Privacy-preserving learning · Training-as-a-Service · Traffic surveillance system

1 Introduction

Real-world deployment of traffic surveillance systems (TSS) faces severe challenges from geographic domain shift. Object detectors perform well when trained and tested in the same location, but their accuracy drops sharply when applied to an unseen city with different road layouts, camera viewpoints, vehicle distributions, and environmental conditions [31, 45]. Despite the importance of this

problem, robust cross-city generalization remains underexplored because large-scale surveillance data collection is constrained by privacy concerns.

To address this research gap, the 10th AI City Challenge (AIC2026) [35] Track 6 introduces a privacy-preserving benchmark using the Milestone Systems' Hafnia Training-as-a-Service (TaaS) platform [22]. In this setup, neither the source-city training data nor the target-city data can be accessed. Instead, methods must be containerized and run in fully blind training and inference loops. This setting requires practical, robust techniques rather than architecture scaling [30, 39, 44, 47, 48] or direct profiling of target data [7, 9, 16, 36, 41].

In this work, we develop a robust, modular training pipeline for Zero-Shot Cross-City Object Detection (ZCOD). To address geographic domain shift without heavy parameter overhead, our framework focuses on two key stages: data preparation and multi-stage optimization. First, we introduce a multi-dataset pre-training strategy governed by a class-agnostic objectness distillation pass. By collapsing heterogeneous label taxonomies into a single binary foreground representation, the model extracts rich geometric inductive biases about vehicle structural boundaries and avoids class alignment conflicts across domains. Second, we propose a domain-resilient chromatic augmentation stream called Grayworld augmentation. This method forces global self-attention heads to remove volatile, sensor-dependent photometric shortcuts and favor robust, shape-centric representation learning. Finally, we formulate a platform-optimized hyperparameter calibration strategy to enable efficient training of the state-of-the-art (SOTA) transformer detector, RF-DETR [30], within the strict 16GB GPU memory limit. Our optimized variants, RF-DETR-HR and RF-DETR-Grayworld, bridge cross-city distribution gaps and achieve significant mAP scores of 47.53 and 46.63, respectively. We also achieve SOTA results and secure 1st place on the AIC2026-Track 6 leaderboard [1].

Our core technical contributions are summarized as follows:

- We introduce a multi-dataset pre-training strategy with a class-agnostic objectness distillation pass to efficiently learn robust vehicle geometry.
- We present a domain-resilient augmentation pillar using a novel Grayworld transformation that forces global self-attention heads to remove transient chromatic shortcuts and favor robust shape priors.
- We construct two improved methods: RF-DETR-HR and RF-DETR-Grayworld, which achieve **SOTA, 1st place performance (mAP 47.53) on the hidden evaluation benchmark**.

2 Related Work

2.1 Cross-Domain Object Detection

Object detectors are primarily split into CNN-based (e.g., Faster R-CNN [29], YOLOs [15, 27, 28, 39, 40]) and transformer-based methods (e.g., DETR [44], Deformable-DETR [47], Co-DETR [48], RF-DETR [30]). While CNNs overfit to

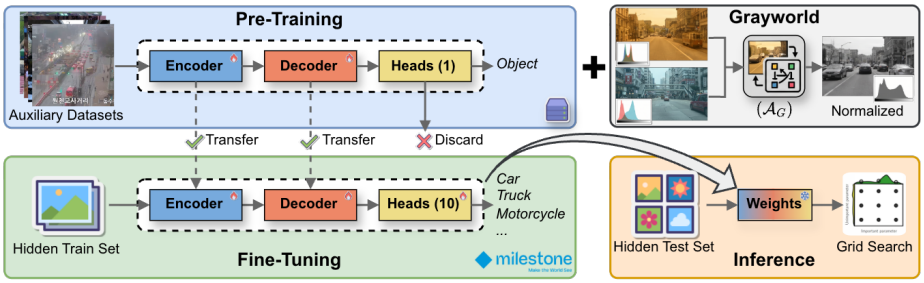

Fig. 1: Overview of the proposed pre-training and Grayworld augmentation framework.

local source-domain features like lighting and asphalt style, transformers struggle with severe performance degradation under geographic domain shifts due to style-sensitive positional queries and attention maps. To bridge the gap between source and target distributions, existing literature relies on three main approaches: Unsupervised Domain Adaptation (UDA), Fully Test-Time Adaptation (F-TTA), and Vision-Language Model (VLM) prompting. UDA strategies [9, 16, 36] use adversarial learning, feature statistic alignment, or student-teacher pseudo-labeling to align feature distributions. F-TTA methods [7, 41] adapt model parameters online during inference via entropy minimization or batch-normalization updates. VLM-based approaches [4, 19, 25, 26] use textual prompts to guide visual representations toward a conceptual target domain.

Despite their success in standard benchmarks, these techniques fail when deployed under **true zero-shot, completely blind, and hardware-constrained settings**. UDA frameworks require offline access to unlabeled target-domain images during training. This assumption breaks in containerized or blind evaluation pipelines where target images remain completely hidden during the entire process. F-TTA methods execute backpropagation and gradient updates during runtime, which introduces severe constraints on resource-limited hardware and risks total model collapse if test-time pseudo-labels become noisy. Finally, VLM methods depend on prior knowledge of target conditions, which is unavailable in completely blind deployments. These disadvantages demonstrate the necessity of a purely source-side generalization strategy that enforces color-invariant, domain-agnostic feature representations during pre-training without requiring target data, test-time gradient updates, or prior domain assumptions.

2.2 Traffic Surveillance Datasets

We describe the datasets used in this study and summarize them in Table 1:

Hafnia Dataset [21] comprises 27.8K images (13K training and 15K testing) from a large-scale traffic dataset with diverse viewpoints, roadway types, camera perspectives, and imaging conditions. The test set contains images from the training city and a new target domain. The dataset contains 10 fine-grained object classes across two main categories: Vehicle and Person.

Table 1: Comparison of traffic surveillance datasets.

Dataset	Viewpoint	Images	Labels	Classes	Resolution	Characteristics
Hafnia [21]	Coarse→Fine	27.8K	151K	10	720–3648p	Cross-city domain shift
TSBOW [14]	Coarse→Fine	48K	1.1M	8	720p	Weather, time-of-day
TrafficCAM [12]	Medium	4.3K	84.2K	10	288–1080p	Complex traffic flow
FishEye8K [13]	Medium	8K	157K	5	1080–1280p	Distortion, omnidirection
VisDrone [46]	Coarse	8.6K	2.6M	10	2160p	Small objects
MOT20 [11]	Medium, Fine	13K	2.2M	1	1080p	Pedestrian

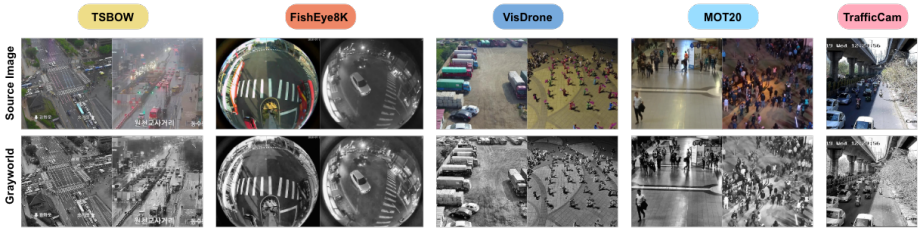

Fig. 2: Visual comparisons of raw pre-training dataset images alongside their corresponding Grayworld-augmented variants. By neutralizing sensor-specific white-balance variances and volatile RGB distributions, the Grayworld transformation effectively bridges color and style discrepancies across disparate source domains.

TSBOW [14] is a diverse urban dataset featuring 48K manually labeled images, with an average of 122 objects per frame. It includes various weather conditions (sunny, rainy, snowy, and lighting) and scenarios (straight roads and intersections). This dataset provides models with useful prior knowledge and is well-suited to robust detection applications.

TrafficCAM [12] is a collection depicting complex traffic flow in urban India. It contains top-down and side-frontal views at resolutions from 240p to FHD. The dataset includes over 4.3K images extracted from more than 2K videos under environmental conditions such as fog and nighttime.

FishEye8K [13] is a benchmark providing omnidirectional coverage using fisheye cameras. This dataset contains 8K images and 157K annotations under four illumination conditions (sunrise, sunny, sunset, and night) in Taiwan. The dataset has been widely used in previous AIC challenges [24, 34, 42]. Thus, it serves as an ideal training dataset to equip our proposed framework to detect objects under extreme lighting changes and diverse camera perspectives.

VisDrone [46] consists of around 8.6K images at 4K resolution. This dataset is commonly used to improve detection and tracking models for small and medium objects captured by drone-mounted cameras [8]. It features highly crowded scenes with significant variations in object scale, occlusion, and perspective.

MOT20 [11] is a benchmark for multi-person tracking in crowded scenes. This dataset contains 8 sequences with over 13K images, and the mean crowd density is 246 pedestrians per frame. Leveraging this data trains our framework to better distinguish overlapping person objects in dense crowds.

3 Preliminary: Hafnia Training-aaS Platform

Unlike conventional benchmark evaluations that test models locally on open-source data, AIC2026-Track 6 [35] uses a privacy-preserving Milestone Systems' Hafnia TaaS platform [22]. It imposes the following constraints:

- **No Data Access:** The training and benchmarking datasets are completely hidden. Algorithms must be packaged as standalone deployment units (i.e., Docker containers), and the data can only be accessed through anonymized training or inference jobs without debugging capability.
- **Blind Evaluation:** The benchmarking process is conducted in two steps. First, the Hafnia platform [22] performs inference on the testing images and generates prediction files. Then, these files are submitted to the AIC2026 evaluation server [1] to be evaluated against the hidden ground-truth.
- **Hardware Constraints:** Training and inference jobs are forced to run within deterministic resource, typically limited to x1–x4 NVIDIA Tesla T4 GPU (16GB VRAM) with a certain amount of training credits.

This TaaS platform changes the workflow from local open-loop training to a blind end-to-end training-inference pipeline. The hardware constraints also introduce a trade-off: VLM distillation [4, 19, 25, 26] becomes a computational bottleneck. Consequently, maximizing the mean Average Precision (mAP) requires localized feature calibration and a memory-efficient training pipeline rather than expanding backbones [30, 39, 44, 47, 48] or complex F-TTA [7, 41]. Our proposed pipeline leverages this constraint as a design feature, optimizing training and inference pipelines to achieve maximum performance.

4 Methodology

4.1 Problem Formulation

We view the domain-robust object detection pipeline as a constrained optimization problem targeting generalization under strict geographic domain shift. Specifically, we systematically investigate the performance impact of different training and inference factors. We also need to consider memory consumption in practice. Let $\mathcal{F}$ denote the detection network parameterized by weights θ. The objective is to maximize the mean Average Precision (mAP) over an unseen target domain distribution $\mathcal{D}_T$ without accessing its visual features or label spaces during both training and inference. Hence, we formulate this multi-stage optimization problem as:

$$\begin{aligned} \max_{\mathcal{D}_S, \mathcal{A}, E, R_T, R_I, \tau_c} \quad & \text{mAP}(\mathcal{F}(\mathcal{D}_S, \mathcal{D}_T; \mathcal{A}, E, B, R_T, R_I, \tau_c)) \\ \text{s.t.} \quad & \text{Memory}(\mathcal{F}(B, R_T, R_I)) \leq M_{\text{GPU}} \end{aligned} \tag{1}$$

where:

- $\mathcal{D}_S \in \{\mathcal{D}_1, \mathcal{D}_2, ...\}$ is our custom pre-training source dataset.

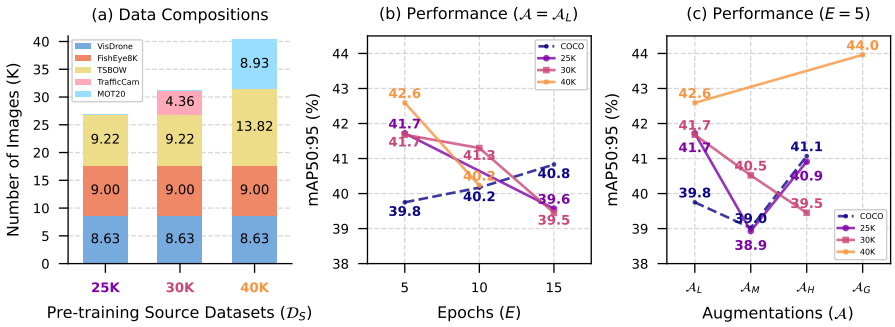

Fig. 3: Compositions and performance comparison of auxiliary pre-training datasets ($\mathcal{D}_S$). Performance is evaluated on a 50% subset of the hidden test set under identical inference configurations.

- $\mathcal{A}$ is the data augmentation strategy.
- E is the number of training epochs.
- B is the batch size during training.
- R_T and R_I are the training and inference resolutions, respectively.
- τ_c is the confidence threshold during inference.

It is obvious that either a larger model or a higher resolution increase both mAP and memory consumption. In practice, to fully utilize memory, we can attempt to calibrate $\mathcal{F}$, E, B, R_T, and R_I to make the total memory consumption close to the maximum GPU memory ($M_{\text{GPU}} = 16\text{GB}$). In the following sections, we focus on $\mathcal{D}_S$, $\mathcal{A}$, and the grid search approach.

4.2 Multi-Dataset Pre-Training ($\mathcal{D}_S$)

Training detectors from scratch is time-consuming. Standard workflows rely on transfer learning from a large-scale distribution. However, generic distributions (e.g., COCO [18] or Objects365 [33]) have broader taxonomies than the specific, fine-grained categories in Table 1. Even when using traffic-specific datasets (in Section 2.2) for better initialization, class definitions still conflict across distributions. For example, a vehicle labeled as "Car" or "Van" in one dataset might be merged into a generic "Vehicle" class. To bridge this taxonomic gap, we discuss two main approaches: **pseudo-label** and **objectness distillation**.

Pseudo-label/Relabeling aims to generate annotations for external data aligned with the target domain distribution. This process can be done via semi-supervised labeling [17,32,37] or VLM-based protocols [4,10,23,25]. Specifically, we explored the zero-shot grounding and localization capabilities of Qwen3-VL [25], denoted as $\mathcal{M}_{\text{VLM}}$, for processing unannotated images. By feeding images through $\mathcal{M}_{\text{VLM}}$ with structured prompt matrices, we dynamically generated high-confidence

bounding box coordinates and fine-grained category logits. This automated relabeling strategy acts as a domain-specific filter, mapping open-vocabulary semantics onto the target category space to isolate relevant object features from out-of-distribution background noise. Ultimately, this technique harmonizes heterogeneous datasets into a synchronized label space. *However, the generated pseudo-labels often require extensive manual refinement to reach the precision needed for fine-grained traffic surveillance. This makes the approach computationally feasible but labor-intensive.*

Objectness Distillation aims to maximize the detector's capacity to identify fine-grained traffic primitives under significant distribution shifts. This method separates the learning of spatial geometry from category-specific classification. A class-agnostic structural distillation pass is implemented by collapsing all semantic labels from auxiliary datasets into a single binary objectness identifier (i.e., Class 0: object):

$$\mathcal{Y}_{\text{agnostic}} = \{y_i \to 0 \mid \forall y_i \in \mathcal{Y}_{\text{source}}\} \tag{2}$$

This formulation redefines the detection task as a binary foreground object localization problem. Applying this reduction over a warm-up schedule (3 to 5 epochs) forces the backbone encoder and transformer cross-attention layers to prioritize distilling high-fidelity object geometries through joint L_1 and GIoU loss minimization. This strategy prevents the classification head from overfitting early to the source city's specific vehicle distribution. By anchoring the model's queries on universal structural priors, we preserve the architectural plasticity required to rapidly adapt to the fine-grained 10-class taxonomy during the subsequent main training phases on the Hafnia platform.

To analyze the scaling behavior of our framework, we construct three pretraining datasets $\mathcal{D}_S \in \{25\text{K}, 30\text{K}, 40\text{K}\}$, aggregated from the heterogeneous traffic surveillance distributions detailed in Section 2.2. The exact composition and source breakdown for each subset are shown in Fig. 3(a). We then conduct an empirical evaluation to assess the downstream generalization impact of each pretraining scale via a grid-search exploration in Section 4.4. Based on observations from Figs. 3(b, c), we use $\mathcal{D}_S = 40\text{K}$ for the rest of the experiments in this paper.

4.3 Data Augmentation ($\mathcal{A}$)

Under geographic domain shift, transformer-based architectures [30] often exhibit severe style memorization [43], where the encoder overfits to source-specific chromatic profiles, illumination artifacts, and sensor-level textures. To decouple vehicle geometry from environmental styles, we implement several standard augmentation strategies ($\mathcal{A}_L$, $\mathcal{A}_M$, $\mathcal{A}_H$) and propose a novel **Grayworld ($\mathcal{A}_G$)** approach. We follow Albumentations guideline [6] to build augmentation pipelines. The detailed steps of each pipeline are listed in Fig. 4. Examples of augmented images are shown in Fig. 2.

Spatial-Geometric Augmentations ($\mathcal{A}_L$, $\mathcal{A}_M$, $\mathcal{A}_H$) handle diverse camera viewpoints and scale distributions by applying randomized affine transformations

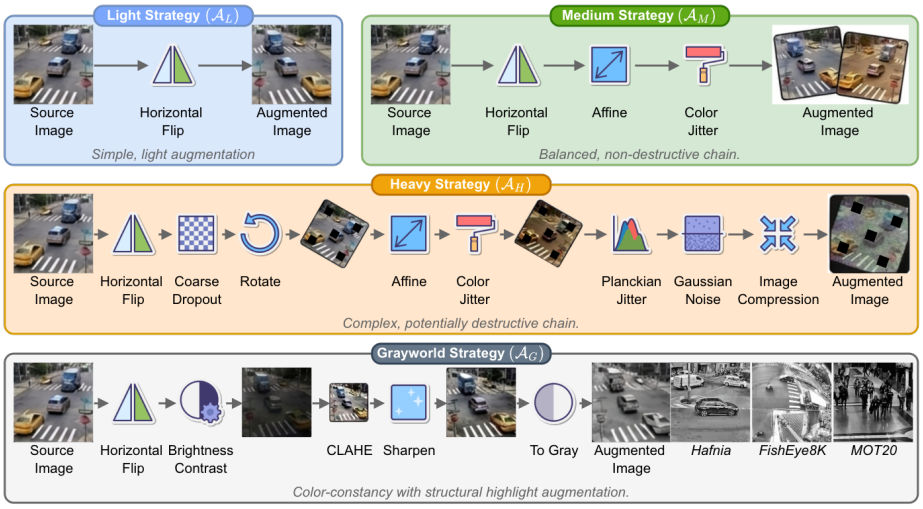

Fig. 4: Illustration of augmentation strategies structured around order-specific sequential operations.

such as rotation, horizontal flipping, and random cropping. These augmentations ensure the model is robust to layout variations typical of cross-city deployment.

Grayworld Augmentation ($\mathcal{A}_G$): Geographic domain shift in TSS is driven by photometric variance across heterogeneous camera sensors (Fig. 2). Local illumination conditions, such as the cool-toned overcast light or the yellow chromaticity of nighttime sodium-vapor lamps, drastically perturb the RGB distributions of target objects. Transformer architectures [30] often lack intrinsic color constancy and are susceptible to shortcut learning. Instead of distilling domain-invariant geometric primitives, the global self-attention heads may overfit to source-specific chromatic footprints, like asphalt reflectance or camera white-balance biases. When deployed in an unseen target city, this learned dependency causes performance degradation as the target's chromatic profile diverges.

To address this problem, we draw inspiration from the robustness of grayscale thermal imaging [2, 3, 20, 38], where object detection relies exclusively on structural intensity and spatial gradients, thereby bypassing the volatility of chromatic signatures. This principle is mathematically aligned with the Gray-World Assumption [5], which states that the average reflectance of a complex scene is achromatic. Motivated by these concepts, we propose chromatic decoupling via grayscale intensity mapping and contrast enhancement. Since vehicle identifiers (e.g., chassis contours and wheel arches) are primarily geometric, they are preserved within high-frequency spatial gradients rather than color channels. Consequently, we collapse the RGB space into a single-channel intensity map throughout all training and inference phases. To compensate for the loss of color-based cues, we apply Contrast Limited Adaptive Histogram Equalization (CLAHE) [49] and edge sharpening, effectively forcing the network to priori-

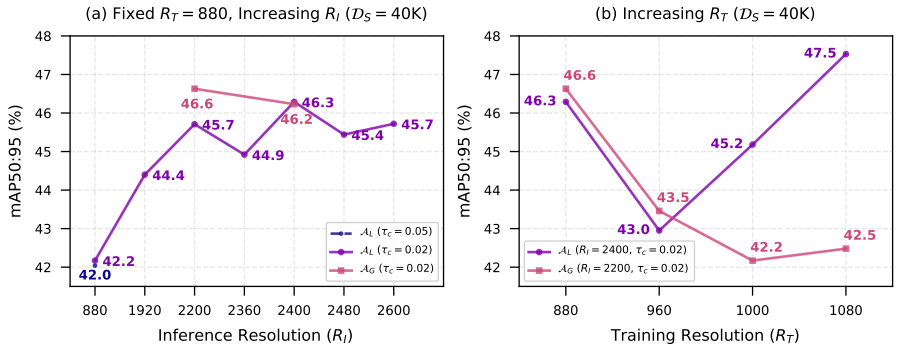

Fig. 5: Impact of training resolution (R_T) and inference resolution (R_I) calibration on object detection performance. Performance is evaluated on the full hidden test set.

tize robust shape-based semantics over transient, domain-specific color styles. Note that the Grayworld transformation must be applied consistently across pre-training, fine-tuning, and inference routines. Maintaining this strict grayscale consistency from initial weight anchoring to final blind deployment stabilizes the latent representation space, preventing feature drift and consistently increasing mAP on hidden benchmark streams, as verified in Section 5.2.

4.4 Hyperparameter Calibration

Solving the global optimization objective in Eq. 1 is challenging due to the high-dimensional parameter space and complex cross-domain dynamics. To find effective training and inference configurations under platform constraints, we conduct a systematic grid-search exploration. We evaluate the sensitivity and cumulative impact of seven core factors: model size ($\mathcal{F}$), pre-training data ($\mathcal{D}_S$), augmentation strategy ($\mathcal{A}$), training epochs (E), training resolution (R_T), inference resolution (R_I), and confidence threshold (τ_c). The complete experimental matrix and empirical trends are summarized in Table 2 and Figs. 5–6. The specific factor configurations and key findings are detailed as follows.

Model Size ($\mathcal{F}$): Scaling the detector's backbone capacity is the most direct way to improve raw mAP under severe domain shift. However, larger variants consume a large part of the platform's 16GB VRAM budget. To find the optimal compute allocation, we stress-tested all RF-DETR [30] variants within the Hafnia platform. We find that optimizing the largest model variant (RF-DETR-2XLarge) is feasible by limiting the physical batch size per GPU to $B = 2$. To reduce high gradient variance and match the optimization trajectory of a larger batch, we combine this with an 8-step gradient accumulation strategy, keeping an effective batch size of 16.

Pre-training ($\mathcal{D}_S$) and Augmentation ($\mathcal{A}$): We conduct a multi-factorial grid search between auxiliary pre-training $\mathcal{D}_S \in \{25\text{K}, 30\text{K}, 40\text{K}\}$ and augmen-

tation strategy $\mathcal{A} \in \{\mathcal{A}_L, \mathcal{A}_M, \mathcal{A}_H, \mathcal{A}_G\}$. Our observations show that increasing pre-training volume provides richer spatial priors and diverse vehicle viewpoint geometries, yielding steady mAP improvements on unseen target cities (Fig. 3). Crucially, we identify a distinct coupling: aggressive augmentation schemes ($\mathcal{A}_H$) are effective mainly when transferring from generic COCO initializations to prevent early feature over-consolidation. Conversely, when the model is pre-trained on our 40K traffic dataset, light augmentations ($\mathcal{A}_L$) and Grayworld ($\mathcal{A}_G$) achieve optimal convergence, preserving distilled vehicle structural features while neutralizing domain-specific style drift.

Training Epochs (E): In ZCOD deployment, the training schedule controls the trade-off between domain feature adaptation and source-domain style memorization. Our grid search shows that a compact schedule of $E = 5$ epochs yields optimal cross-city mAP performance. Extending fine-tuning beyond this (e.g., $E \in \{10, 15\}$) causes a noticeable decline in target-domain accuracy (Fig. 3(b)).

Training and Inference Resolutions (R_T, R_I): Image resolution directly dictates feature map granularity and positional query alignment in Vision Transformers. We evaluate training resolutions $R_T \in \{880, 960, 1000, 1080\}$ and inference resolutions $R_I \in \{880, 1920, 2200, 2360, 2400, 2480, 2600\}$. Our empirical grid search reveals three key insights about resolution dynamics:

- **Inference Scale Saturation:** mAP increases with inference resolution R_I as higher pixel densities restore fine-grained spatial cues for small vehicles. However, performance reaches a saturation point, peaking at 2400×2400 for the light augmentation ($\mathcal{A}_L$) and 2200×2200 for the Grayworld augmentation ($\mathcal{A}_G$) before over-smoothing artifacts and logit noise reduce precision.
- **Augmentation-Scale Coupling:** Increasing the training resolution R_T to 1080×1080 mainly benefits $\mathcal{A}_L$ by expanding spatial feature resolution. In contrast, models trained with $\mathcal{A}_G$ show intrinsic scale-invariant shape expression and maintain strong generalization without high training resolutions.
- **Platform Computation:** Scaling training resolution beyond the 880×880 baseline causes significant memory overhead, forcing batch size reduction to $B = 1$ to avoid out of memory within the 16GB VRAM limit. Increasing inference scale to $R_I = 2600$ remains viable on the Hafnia deployment container by optimizing the model's forward execution graph via TorchScript Just-In-Time (JIT) tracing with dynamic precision casting.

Confidence Threshold (τ_c): The confidence threshold controls the trade-off between false-positive filtering and target recall during test-time evaluation. In ZCOD deployment, models often show under-confident prediction logits due to out-of-distribution. Our grid search shows that lowering the post-processing threshold ($\tau_c = 0.02$) consistently increases overall mAP. By relaxing τ_c, the detector keeps valid, low-confidence object queries such as heavily occluded, small, or distant vehicles that standard thresholds ($\tau_c \geq 0.25$) would prune aggressively. Since mAP evaluates the entire area under the Precision-Recall curve across IoU thresholds, recovering these true positive detections at lower confidence levels significantly boosts overall recall without large precision penalties.

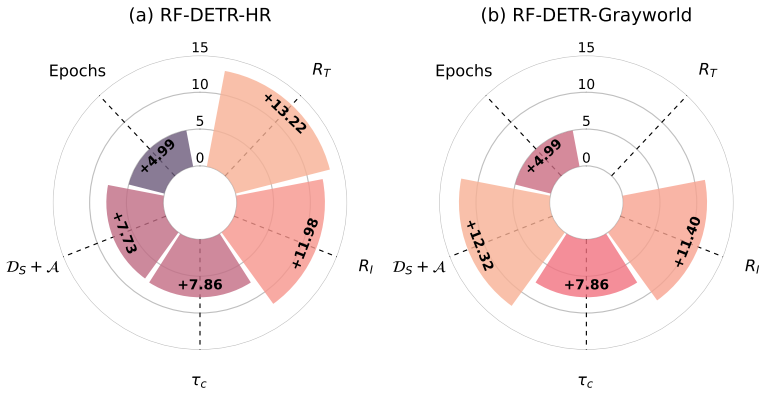

Fig. 6: Incremental mAP performance contributions of each hyperparameter for **RF-DETR-HR** and **RF-DETR-Grayworld** relative to the COCO pre-trained RF-DETR-2XLarge baseline. Axes are arranged in counter-clockwise order.

4.5 Improved Methods

Building upon the empirical findings and hyperparameter dynamics analyzed in Sections 4.2–4.4, we construct two model configurations tailored for ZCOD:

- **RF-DETR-HR:** Integrates domain-aligned 40K pre-training with light data augmentation ($\mathcal{A}_L$), high-resolution training ($R_T = 1080 \times 1080$), and test-time inference scaling ($R_I = 2400 \times 2400$), maximizing spatial feature density to resolve small and distant traffic targets.
- **RF-DETR-Grayworld:** Combines domain-aligned 40K pre-training and Grayworld chromatic transformation ($\mathcal{A}_G$) alongside test-time inference scaling ($R_I = 2200 \times 2200$), forcing the model to prioritize domain-invariant structural geometry over volatile, location-specific color statistics.

The relative performance contributions and empirical ablation trajectories of each component are visually summarized in Fig. 6.

5 Experimental Results

5.1 Implementation Details

Local Pre-training is conducted on an NVIDIA RTX A6000 GPU (48GB VRAM). For each pre-training dataset $\mathcal{D}_S \in \{25\text{K}, 30\text{K}, 40\text{K}\}$, we optimize two sets of weights: an unaugmented baseline and a Grayworld variant. Furthermore, whenever the training resolution R_T is adjusted, we need to pre-train dedicated checkpoint weights to preserve scale-dependent positional embeddings. This warm-up phase runs for a brief 3–5 epoch schedule, after which the top-performing Exponential Moving Average (EMA) weights are selected as the initialization checkpoint for downstream fine-tuning.

Table 2: Experimental results on the full test set of AIC2026-Track 6 [1]. The best and the second-best results are in **bold** and underline, respectively.

#	**Model** ($\mathcal{F}$)	$\mathcal{D}_S$	E	$\mathcal{A}$	R_T	R_I	τ_c	**mAP**	**mAP50**	**mAP75**
0	RF-DETR-Nano (Baseline)	COCO	8	-	384	384	0.05	23.24	33.77	24.13
1	RF-DETR-Nano	COCO	1	-	384	384	0.05	26.72	38.10	27.11
2	RF-DETR-Large	COCO	1	-	704	704	0.05	30.77	44.92	32.24
3	RF-DETR-XLarge	COCO	1	-	704	704	0.05	31.06	43.39	33.00
4	RF-DETR-2XLarge	COCO	1	-	880	880	0.05	34.31	46.99	36.61
5	RF-DETR-2XLarge	COCO	5	$\mathcal{A}_L$	880	880	0.05	39.30	53.18	41.31
6	RF-DETR-2XLarge	COCO	5	$\mathcal{A}_M$	880	880	0.05	38.43	52.88	40.47
7	RF-DETR-2XLarge	COCO	5	$\mathcal{A}_H$	880	880	0.05	40.37	56.11	43.04
8	RF-DETR-2XLarge	25K	5	$\mathcal{A}_L$	880	880	0.05	42.28	56.85	43.95
9	RF-DETR-2XLarge	25K	5	$\mathcal{A}_M$	880	880	0.05	40.50	55.03	43.28
10	RF-DETR-2XLarge	25K	5	$\mathcal{A}_H$	880	880	0.05	38.43	54.68	39.90
11	RF-DETR-2XLarge	30K	5	$\mathcal{A}_L$	880	880	0.05	41.18	55.26	43.18
12	RF-DETR-2XLarge	40K	5	$\mathcal{A}_L$	880	880	0.05	42.04	55.58	44.10
13	RF-DETR-2XLarge	40K	5	$\mathcal{A}_L$	880	880	0.02	42.17	55.80	44.23
14	RF-DETR-2XLarge	40K	5	$\mathcal{A}_L$	880	1920	0.02	44.40	57.71	47.65
15	RF-DETR-2XLarge	40K	5	$\mathcal{A}_L$	880	2200	0.02	45.71	58.86	48.86
16	RF-DETR-2XLarge	40K	5	$\mathcal{A}_L$	880	2400	0.02	46.29	59.51	49.84
17	RF-DETR-2XLarge	40K	5	$\mathcal{A}_L$	960	2400	0.02	42.95	57.41	45.71
18	RF-DETR-2XLarge	40K	5	$\mathcal{A}_L$	1000	2400	0.02	45.18	59.19	48.13
19	**RF-DETR-2XLarge** (RF-DETR-HR)	**40K**	**5**	$\mathcal{A}_L$	**1080**	**2400**	**0.02**	**47.53**	**62.07**	**50.31**
20	RF-DETR-2XLarge (RF-DETR-Grayworld)	40K	5	$\mathcal{A}_G$	880	2200	0.02	46.63	59.46	49.96
21	RF-DETR-2XLarge	40K	5	$\mathcal{A}_G$	880	2400	0.02	46.23	59.39	49.23
22	RF-DETR-2XLarge	40K	5	$\mathcal{A}_G$	960	2200	0.02	43.46	57.01	47.14
23	RF-DETR-2XLarge	40K	5	$\mathcal{A}_G$	1000	2200	0.02	42.17	55.62	44.87
24	RF-DETR-2XLarge	40K	5	$\mathcal{A}_G$	1080	2200	0.02	42.48	55.89	45.12

Online Fine-tuning is performed on the hidden training split within the Hafnia Training-aaS environment [22] using an NVIDIA Tesla T4 GPU (16GB VRAM). We load the localization-anchored pre-trained weights and adapt the classification prediction heads to the full 10-class target. To ensure spatial feature adaptation while maintaining semantic stability, we use a differential learning rate strategy. We assign a base learning rate of $\mathrm{lr} = 10^{-4}$ to the transformer decoder and classification heads, and a slightly higher rate of $\mathrm{lr}_{\mathrm{enc}} = 1.5 \times 10^{-4}$ to the transformer encoder. This asymmetrical setup prioritizes spatial context adaptation, which is essential for accommodating novel camera viewpoints and road layouts in unseen target cities. At the same time, it conservatively updates the classification heads to reduce catastrophic forgetting of pre-trained spatial priors. The complete set of fine-tuning hyperparameters is detailed in Section 4.4.

Benchmarking process is conducted in two steps. First, the containerized model runs forward-pass inference on the concealed target-city test split hosted on the Hafnia platform infrastructure [22], producing standardized prediction files. Then, these prediction files are submitted to the official AIC2026 evaluation server [1] for hidden ground-truth benchmarking and leaderboard ranking.

The primary evaluation is based on standard COCO-Eval [18] metrics, with the final ranking determined by mAP50:95.

5.2 Benchmark Results

In this section, we present our empirical benchmarking framework (results shown in Table 2), detailing the iterative experimental setup, calibration progression, and key methodological findings derived across each evaluation cycle:

Model Scale Exploration (#1–4): We evaluated all RF-DETR model variants by fine-tuning each architecture for one epoch using standard COCO pre-trained weights under default training settings. Our initial benchmark showed that the largest variant, RF-DETR-2XLarge, achieved the highest target-domain mAP (34.31) among the baselines. Furthermore, our ablation study (in Fig. 3(b)) confirms that restricting fine-tuning to $E \leq 5$ epochs significantly outperformed longer schedules (26.72 vs. 23.24 mAP), effectively preventing transformer attention heads from memorizing source-city visual clutter and overfitting to domain-specific artifacts.

Pre-training and Augmentation (#5–12): We systematically evaluated the combinations of auxiliary pre-training datasets $\mathcal{D}_S$ and augmentation strategies $\mathcal{A}$. Our findings reveal a key paradigm: scaling the domain-aligned pre-training dataset to 40K instances yields a significantly higher cross-city mAP (42.04) than applying aggressive synthetic augmentations ($\mathcal{A}_H$) on smaller initializations (40.37). Heavy synthetic transformations risk corrupting feature geometries, whereas expanding real-world pre-training volume exposes the model to natural environmental, lighting, and viewpoint diversity. Thus, expanding the pre-training data acts as a powerful macro-level augmentation technique, providing realistic domain invariance while preserving structural fidelity.

Hyperparameter Calibration (#13–19): After identifying the optimal pre-training and augmentation strategy, we calibrated image resolutions (R_T, R_I) and post-processing thresholds (τ_c). First, lowering the confidence threshold to $\tau_c = 0.02$ recovered under-confident target predictions caused by domain-shift logit suppression, yielding a slight mAP increase (+0.13). Second, increasing test-time inference resolution R_I to 2400×2400 significantly boosted small-target recall (46.29 vs. 42.17). Finally, increasing training resolution R_T to 1080×1080 further expanded spatial feature map capacity for the $\mathcal{A}_L$ augmentation (+1.24 mAP), though this required reducing the physical batch size ($B = 1$) to strictly fit the 16GB VRAM limit.

Grayworld Augmentation (#20–24): We also explored the impact of the Grayworld chromatic transformation ($\mathcal{A}_G$) applied during local pre-training and online fine-tuning. Our results show that by removing volatile RGB color distributions and neutralizing camera white-balance variances, Grayworld augmentation prevents the Vision Transformer from exploiting transient photometric shortcuts. Specifically, at the same training resolution of $R_T = 880$, the Grayworld variant outperforms light augmentation baselines (46.63 vs. 45.71 mAP), confirming its robustness against unpredictable lighting and sensor style shifts.

Table 3: Official public leaderboard for AIC2026-Track 6 [1]. Our team, SKKU-AL-T1 (Team ID 34), secures 1st place with an absolute mAP margin of +4.72 over the second-place entry.

Rank	Team ID	Team Name	Model	mAP	mAP50	mAP75
1	**34**	**SKKU-AL-T1**	**RF-DETR-HR**	**47.53**	**62.07**	**50.31**
			RF-DETR-Grayworld	46.63	59.46	49.96
2	256	BIT-ODL	detr	42.81	55.39	45.68
3	152	Chisinau	RF-DETR Large	41.76	57.27	45.21
4	261	BK2TheFuture	RFDETR2XLarge	41.69	52.89	44.07
5	162	S2 Detection	RF-DETR	41.14	54.47	44.14
6	58	VisionOps	detr ensemble	40.22	53.33	42.59
7	273	Team IPCV	DEIMv2X, Dinov3S+	38.82	50.21	40.68
8	101	bfc	detr	37.41	50.97	39.89
9	247	NextITS	RF-DETR Large, square 1120	36.54	48.79	38.57
10	188	Team United	RF-DETR Large	35.29	46.55	36.56

5.3 Comparison & Final Ranking

The final rankings of AIC2026-Track 6 [1] are summarized in Table 3. Our solutions, RF-DETR-HR and RF-DETR-Grayworld, rank 1st (47.53 mAP) and 2nd nominally (46.63 mAP) according to the reported results. Furthermore, our methods outperform the subsequent competitor entry by substantial absolute margins (+4.7 and +3.8 mAP).

6 Conclusion

In this work, we addressed zero-shot cross-city object detection under strict hardware and privacy constraints on the Milestone System' Hafnia platform. To tackle geographic domain shifts without changing complex Vision Transformer architectures, we introduced a modular framework based on two pillars: a class-agnostic 40K pre-training strategy that separates vehicle structural geometry from changing taxonomies, and a Grayworld chromatic transformation that removes sensor-dependent color shortcuts. Combined with platform-optimized techniques such as differential learning rates, gradient accumulation, and JIT tracing, our optimized models (RF-DETR-HR and RF-DETR-Grayworld) operate efficiently within a 16GB VRAM limit. These innovations yielded substantial mAP gains over standard baselines and secured the **1st place on the official AI City Challenge (AIC2026) Track 6 evaluation leaderboard**.

Acknowledgements

This work was supported by Institute of Information & Communications Technology Planning & Evaluation (IITP) grant funded by the Korea government (MSIT) (No.2021-0-01364, An intelligent system for 24/7 real-time traffic surveillance on edge devices)

References

1. AI City Challenge Evaluation System (2026), https://eval.aicitychallenge.org/aicity2026/submission/leaderboard
2. Ahmar, W.A.E., Kolhatkar, D., Nowruzi, F.E., Laganiere, R.: Enhancing Thermal MOT: A Novel Box Association Method Leveraging Thermal Identity and Motion Similarity. In: European Conference on Computer Vision (ECCV) Workshops. pp. 103–120 (2024)
3. Ahmar, W.E., Sappa, A., Hammoud, R.: Thermal Pedestrian Multiple Object Tracking Challenge (TP-MOT). In: IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops (2025)
4. Bai, J., Bai, S., Yang, S., et al.: Qwen-VL: A Versatile Vision-Language Model for Understanding, Localization, Text Reading, and Beyond. arXiv preprint arXiv:2308.12966 (2023)
5. Buchsbaum, G.: A Spatial Processor Model for Object Colour Perception. Journal of the Franklin Institute **310**(1), 1–26 (1980)
6. Buslaev, A., Iglovikov, V.I., Khvedchenya, E., et al.: Albumentations: Fast and Flexible Image Augmentations. Information **11**(2) (2020)
7. Chang, W.G., You, T., Seo, S., et al.: Domain-Specific Batch Normalization for Unsupervised Domain Adaptation. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (2019)
8. Chen, G., Wang, W., He, Z., et al.: VisDrone-MOT2021: The Vision Meets Drone Multiple Object Tracking Challenge Results. In: IEEE/CVF International Conference on Computer Vision (ICCV) Workshops. pp. 2839–2846 (2021)
9. Chen, Y., Li, W., Sakaridis, C., et al.: Domain Adaptive Faster R-CNN for Object Detection in the Wild. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (2018)
10. Cheng, T., Song, L., Ge, Y., et al.: YOLO-World: Real-Time Open-Vocabulary Object Detection. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (2024)
11. Dendorfer, P., Rezatofighi, H., Milan, A., et al.: MOT20: A Benchmark for Multi-object Tracking in Crowded Scenes. arXiv preprint arXiv:2003.09003 (2020)
12. Deng, Z., Cheng, Y., Liu, L., et al.: TrafficCAM: A Versatile Dataset for Traffic Flow Segmentation (2022)
13. Gochoo, M., Otgonbold, M.E., Ganbold, E., et al.: FishEye8K: A Benchmark and Dataset for Fisheye Camera Object Detection. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops. pp. 5304–5312 (2023)
14. Huynh, N.D.M., Tran, D.N.N., Pham, L.H., et al.: TSBOW: Traffic Surveillance Benchmark for Occluded Vehicles Under Various Weather Conditions. In: AAAI Conference on Artificial Intelligence. vol. 40, pp. 5239–5247 (2026)
15. Jocher, G.: YOLOv5 by Ultralytics (2020)
16. Khanh, T.L.B., Nguyen, H.H., Pham, L.H., et al.: Dynamic Retraining-Updating Mean Teacher for Source-Free Object Detection. In: European Conference on Computer Vision (ECCV). pp. 328–344 (2024)
17. Lee, D.H.: Pseudo-label: The Simple and Efficient Semi-supervised Learning Method for Deep Neural Networks. In: Workshop on challenges in representation learning, ICML. vol. 3, p. 896 (2013)
18. Lin, T.Y., Maire, M., Belongie, S., et al.: Microsoft COCO: Common Objects in Context. In: European Conference on Computer Vision (ECCV). pp. 740–755 (2014)

19. Liu, S., Zeng, Z., Ren, T., et al.: Grounding DINO: Marrying DINO with Grounded Pre-Training for Open-Set Object Detection. In: European Conference on Computer Vision (ECCV) (2024)
20. Manjunath, D., Sikdar, A., Gurunath, P., et al.: SAGA: Semantic-Aware Gray Color Augmentation for Visible-to-Thermal Domain Adaptation Across Multi-View Drone and Ground-Based Vision Systems. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops. pp. 4578–4588 (2025)
21. Milestone Systems: Hafnia Dataset: ECCV Cross-City Object Detection Dataset (2024), part of the Hafnia project
22. Milestone Systems: Project Hafnia: A Game-Changer in AI Model Training (2025)
23. Pham, L.H., Ho, Q.P.N., Tran, D.N.N., et al.: Improving Object Detection to Fisheye Cameras with Open-Vocabulary Pseudo-Label Approach. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops. pp. 7100–7109 (2024)
24. Pham, L.H., Ho, Q.P.N., Vu, D.K., et al.: Data Augmentation Is All You Need For Robust Fisheye Object Detection. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV) Workshops (2025)
25. Qwen Team: Qwen3 Technical Report (2025)
26. Radford, A., et al.: Learning Transferable Visual Models From Natural Language Supervision. In: 38th International Conference on Machine Learning (ICML). pp. 8748–8763 (2021)
27. Redmon, J., et al.: You Only Look Once: Unified, Real-Time Object Detection. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 779–788 (2016)
28. Redmon, J., et al.: YOLO9000: Better, Faster, Stronger. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 6517–6525 (2017)
29. Ren, S., He, K., Girshick, R., Sun, J.: Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks. IEEE Transactions on Pattern Analysis and Machine Intelligence **39**(6), 1137–1149 (2017)
30. Robinson, I., Robicheaux, P., Popov, M., et al.: RF-DETR: Real-Time Detection Transformer. In: International Conference on Learning Representations (ICLR) (2026)
31. Saunier, N., et al.: An Embedded Computer-Vision System for Multi-Object Detection in Traffic Surveillance. IEEE Transactions on Intelligent Transportation Systems (2018)
32. Sekachev, B., et al.: Computer Vision Annotation Tool (CVAT) (2020)
33. Shao, S., Li, Z., Zhang, T., et al.: Objects365: A Large-Scale, High-Quality Dataset for Object Detection. In: IEEE/CVF International Conference on Computer Vision (ICCV). pp. 8429–8438 (2019)
34. Tang, Z., Wang, S., Anastasiu, D.C., Chang, M.C., et al.: The 9th AI City Challenge. In: IEEE/CVF International Conference on Computer Vision (ICCV) Workshops. pp. 5526–5535 (2025)
35. Tang, Z., Wang, S., Anastasiu, D.C., Chang, M.C., et al.: The 10th AI City Challenge. In: European Conference on Computer Vision (ECCV) Workshops (2026)
36. Tarvainen, A., Valpola, H.: Mean Teachers Are Better Role Models: Weight-averaged Consistency Targets Improve Semi-supervised Deep Learning Results. In: Advances in Neural Information Processing Systems. vol. 30 (2017)
37. Tkachenko, M., et al.: Label Studio: Data Labeling Software (2020)

38. Tran, D.N.N., Pham, L.H., Ho, Q.P.N., et al.: A Hybrid Data-Centric Framework for Thermal Multiple-Object Tracking with Complex Motion Patterns. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops. pp. 7200–7209 (2026)
39. Ultralytics: Ultralytics YOLO26: Unified Real-Time End-to-End Vision Models (2026)
40. Wang, C.Y., et al.: YOLOv7: Trainable Bag-of-Freebies Sets New State-of-the-Art for Real-Time Object Detectors. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 7464–7475 (2023)
41. Wang, D., Shelhamer, E., Liu, S., et al.: Tent: Fully Test-Time Adaptation by Entropy Minimization. In: International Conference on Learning Representations (2021)
42. Wang, S., Anastasiu, D.C., Tang, Z., Chang, M.C., et al.: The 8th AI City Challenge. In: The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops (2024)
43. Yang, X., Xie, Q.: StyleProto: Style-Augmented Prototype Learning for Cross-Domain Few-Shot Object Detection. In: AAAI Conference on Artificial Intelligence. vol. 40, pp. 11748–11756 (2026)
44. Zhao, Y., Lv, W., Xu, S., et al.: DETRs Beat YOLOs on Real-time Object Detection. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 12739–12748 (2024)
45. Zhou, W., Yang, L., Zhao, L., et al.: Vision Technologies with Applications in Traffic Surveillance Systems: A Holistic Survey. ACM Computing Surveys **58**(2), 1–39 (2026)
46. Zhu, P., Wen, L., Du, D., et al.: Detection and Tracking Meet Drones Challenge. IEEE Transactions on Pattern Analysis and Machine Intelligence **44**(11), 7380–7399 (2021)
47. Zhu, X., Su, W., Lu, L., et al.: Deformable DETR: Deformable Transformers for End-to-End Object Detection. In: International Conference on Learning Representations (2021)
48. Zong, Z., Song, G., Liu, Y., et al.: DETRs with Collaborative Hybrid Assignments Training. In: IEEE/CVF International Conference on Computer Vision (ICCV). pp. 6748–6758 (2023)
49. Zuiderveld, K.: Contrast-Limited Adaptive Histogram Equalization. In: Graphics gems IV, pp. 474–485. Academic Press Professional, Inc. (1994)