remote sensing

MDPI

Article

RDF-Det: Robust Dual-Path Fusion for CNN-Based Vehicle Fire Detection in Ground-Based RGB Imagery

Wenyi Lu, Xiang Zhang *, Boyu Zhang, Zhiqing Li, Zeqiang Chen and Nengcheng Chen

National Engineering Research Center for Geographic Information System, China University of Geosciences, Wuhan 430074, China; wenyilu@cug.edu.cn (W.L.); boyuzhang@cug.edu.cn (B.Z.); lizhiqing97@cug.edu.cn (Z.L.); chenzeqiang@cug.edu.cn (Z.C.); chennengcheng@cug.edu.cn (N.C.)
* Correspondence: zhangxiang76@cug.edu.cn

Highlights

What are the main findings?

- RDF-Det integrates RDF-Neck to fuse parallel semantic and spatial paths, achieving 0.6352 AP and 0.9290 AP50 (+7.24 AP system-level gain over RTM-Base), while RDF-Neck yields +0.65–1.14 AP on the DINOv3 test subset.
- The main RDF-Det configuration (RTM-RDF-GLR-DINO) outperforms the latest YOLO26 detector by 1.17 AP and 2.19 AP50 points with over fourfold lower cross-seed variance, while global shared routing and fixed-uniform fusion outperform unconstrained level-wise routing.

What are the implications of the main findings?

- The method provides a domain-targeted visual detector for vehicles with visible fire in cluttered public-transport and hazardous-material road incidents, supporting safety-oriented ground-based transportation sensing.
- The multi-seed statistical evaluation and topology–routing decoupling analysis provide rigorous empirical guidance on balancing dual-path feature aggregation, fusion complexity, and detector-family compatibility.

Abstract

Road incidents involving public-transport vehicles and hazardous-material carriers require ground-based RGB sensing capable of localizing full vehicle extent while identifying visible fire cues under severe background clutter, scale variation, and illumination artifacts. To address the tension between preserving macro-level vehicle semantics and capturing localized fire patterns, we propose the Robust Dual-Path Fusion Detector (RDF-Det), centered on a parallel feature-fusion neck (RDF-Neck) that integrates top–down semantic propagation with bidirectional spatial aggregation. By decoupling the dual-path topology from branch-mixing parameterization, we systematically investigate four routing strategies—level-wise learned, globally shared, fixed-uniform, and parameter-matched frozen routing—and evaluate cross-family instantiations across RTMDet, YOLOX, and Faster R-CNN. Across a 100-run five-seed benchmark, the main RDF-Det configuration (RTM-RDF-GLR-DINO) achieves mean validation scores of 0.6352 AP and 0.9290 AP50, delivering a system-level improvement of 7.24 AP points over RTM-Base and outperforming the YOLO26 detector (0.6235 AP and 0.9071 AP50). Controlled evaluations on both the validation split and a screened test subset demonstrate that lightweight global and fixed-uniform fusion remain highly competitive with complex level-wise routing, while cross-family comparisons clarify the architectural conditions governing one-stage and two-stage adaptation. Supported by repeated runs, confidence intervals, and Holm-corrected paired tests, this study provides

Check for updates

Academic Editor: Kevin Tansey

Received: 30 September 2026
Revised: 2 November 2026
Accepted: 14 November 2026
Published: 17 November 2026

Copyright: © 2026 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license.

both an effective vehicle-fire detection framework and systematic empirical insights into dual-path multi-scale fusion.

Keywords: ground-based RGB sensing; vehicle fire detection; convolutional neural networks; robust dual-path fusion; statistical evaluation

1. Introduction

Road incidents involving public-transport vehicles and hazardous-material carriers motivate ground-based RGB sensing for safety-critical Intelligent Transportation Systems (ITS). Disaster-risk research emphasizes that early warning becomes useful only when risk information is translated into actionable evidence for operators and response systems [1–3]. In road monitoring, computer vision has already supported vehicle detection, intersection surveillance, and traffic-event understanding [4–6]. Hazardous-material transportation studies have quantified route-level, accident-stage, and driver-level risk [7–9]. These studies explain why the problem matters, but they do not provide the fine-grained visual answer needed at the incident scene: which vehicle shows visible fire. Operational incident perception therefore requires formulating vehicle fire detection as an object-level visual task whose output is directly inspectable by traffic operators—namely, a vehicle-associated class label, a bounding box, and a confidence score. At this individual-vehicle observation scale, ground-view road imagery provides fine-grained vehicle-associated fire-state information for urban transportation monitoring. Within a coordinated ground–air–space monitoring system, such vehicle-level image interpretation can provide a local observation layer complementary to wider-area aerial and satellite coverage. Where overhead structures restrict an aerial line of sight, a ground camera can retain a local view of the road segment.

The practical urgency of such vehicle-level fire monitoring is underscored by highway incident statistics. The U.S. Fire Administration reported annual averages of approximately 171,500 highway vehicle fires, 345 deaths, 1300 injuries, and USD 1.1 billion in property loss in the United States during 2014–2016, with collisions involved in 60% of fatal highway vehicle fires [10].

Fire-risk objects differ from ordinary traffic objects because the discriminative evidence is unstable. Flames and smoke are deformable, scale-varying, illumination-sensitive, and easily confused with headlights, reflections, fog, dust, and night-time artifacts. Fire and smoke detection studies have repeatedly reported false alarms and localization degradation under complex visual conditions [11–13]. Video-based flame detectors show that color and motion cues are sensitive to scene changes [14,15]. Deep smoke/flame detectors further show that texture and temporal cues become unreliable when targets are small or mixed with structured backgrounds [16–18]. Road-incident scenes intensify these difficulties through metallic vehicle reflections, partial occlusion, high-density traffic, and open-road background clutter. The practical goal is therefore simple but demanding: detect a small number of fire-risk objects early, without allowing dominant traffic-scene distractors to control the representation.

Modern CNN-based detectors often improve accuracy by combining stronger backbones, redesigned multi-scale necks, customized losses, and refined training strategies. Such combinations are effective in general object detection [19–21] and are consistent with recent traffic-incident and small-object perception studies, where feature quality and multi-scale fusion strongly affect robustness [22,23]. However, when several components are changed simultaneously, the source of improvement becomes difficult to attribute. This is especially problematic in safety-oriented ITS, where single-run results can be distorted by initialization variance and benchmark noise [24–26]. It remains unclear how a parallel

semantic–spatial fusion structure performs under different routing choices and detector architectures. To address this question, we propose RDF-Det for vehicle fire-state detection and examine three related contributions. First, RDF-Neck combines an FPN-style semantic path and a CSPNeXtPAFPN-style spatial path through corresponding-level fusion and post-convolution. Second, the RTMDet study compares four routing choices and analyzes their accuracy, cross-seed variation, and optimization behavior. Third, the YOLOX and Faster R-CNN instantiations examine architecture-dependent outcomes, including configurations that do not improve over their backbone-enhanced references. Together, these contributions establish the complete RDF-Det framework and clarify the role of RDF-Neck through the 100-run benchmark study and recent-detector comparisons.

Section 2 reviews related work. Section 3 describes RDF-Det and its routing variants. Section 4 presents the experimental setup. Section 5 analyzes detection performance, computational cost, and qualitative results. Section 6 concludes the study.

2. Related Work

This section reviews fire and smoke perception, road-scene conditions affecting vehicle-level fire detection, and the feature-fusion and evaluation methods relevant to this study.

2.1. Vision-Based Fire and Smoke Detection

Vision-based fire and smoke detection has progressed from handcrafted color-motion rules to deep representation learning. Early video-surveillance systems used flame shape, color, and motion, while showing sensitivity to background motion and illumination changes [13,14]. Deep CNN-based systems improved fire localization and classification in surveillance video [11]. Recent reviews and attention-based models identify smoke, fog, and clutter as persistent sources of error [12,16,17]. These methods operate on local visual evidence, whereas regional fire-susceptibility and disaster-risk studies address broader spatial conditions [1,27]. For vehicle fires, local evidence must further be associated with a specific vehicle, which yields the vehicle-level output used here: a bounding box and a vehicle/fire-state class.

Recent vehicle-fire work also distinguishes the output required by the task. Kwon et al. compared YOLO-based segmentation models, Mask R-CNN, and Cascade Mask R-CNN for electric-vehicle fire imagery, with flame and smoke regions as the annotated outputs [28]. RDF-Det instead predicts vehicle-associated boxes and fire states. This distinction motivates feature aggregation that retains vehicle extent while using the surrounding fire-related context.

2.2. Vehicle-Level Fire Detection in Road Scenes

Vehicle-level fire detection shares the scale variation, illumination changes, and occlusion of traffic monitoring [4,5]. Traffic-participant and all-day vehicle detectors address feature extraction under mixed traffic and nighttime conditions [29,30], and multi-perspective accident-video studies address rare events across camera viewpoints [31]. These works target the sensing conditions, but they do not output the visible fire state of an individual vehicle. Studies on hazardous-material transportation [7,8] and accident risk [9,32,33] motivate such object-level fire detection from a safety perspective. The key feature-fusion problem is therefore to retain vehicle localization cues while aggregating the semantic context needed to separate fire-related evidence from road-scene distractors.

2.3. Multi-Scale Feature Fusion and Comparative Analysis

Multi-scale representation is central to modern object detection. Generic object-detection surveys and two-stage detector studies show the value of hierarchical features and

proposal-based localization [19–21]. Previous feature-fusion research further demonstrates that high-resolution pyramids and cross-level aggregation can strengthen small-object perception [23]. Anchor-free and one-stage designs, including focal-loss and keypoint-based detectors, reduce dependence on proposal generation but remain sensitive to feature alignment and scale balance [34,35]. COCO-style detection evaluation, feature pyramids, and weighted bidirectional fusion provide the evaluation and architectural context for the neck modules considered here [36–38].

2.4. Statistical Reliability in Safety-Critical Perception

A persistent limitation in application-driven detection papers is the reliance on one training run per configuration. Statistical-testing guides for NLP and machine learning show that significance claims can be fragile when variance, repeated comparisons, and random initialization are not controlled [24–26]. Holm correction provides a principled way to control family-wise error in multiple hypothesis testing [39]. Paired effect sizes and confidence intervals further make the magnitude of improvement easier to interpret than raw p-values alone [40,41]. For a safety-critical ITS detector, these tools are not cosmetic; they determine whether a reported improvement is likely to survive a different seed or implementation.

3. Methodology

This section defines the vehicle fire detection task and describes how backbone features pass through the two RDF-Neck branches to the detection head. It then presents the routing variants and their instantiations in the three detector families.

3.1. Problem Formulation and Overview

In the context of Intelligent Transportation System (ITS) safety monitoring, vehicle fire detection is formulated as a supervised object-detection problem on ground-based road-scene RGB images. Given an input RGB image $x \in \mathbb{R}^{H\times W\times 3}$ with spatial size H × W and three color channels, captured from a public-transport or hazardous-material road incident, the detector predicts a set of class-labeled bounding boxes:

$$\hat{DP} = \left\{\left(\hat{b}_i \middle| \hat{c}_i \middle| \hat{s}_i\right)\right\}_{i=1}^{N} \tag{1}$$

where $\hat{b}_i$ denotes the spatial coordinates of the bounding box, $\hat{c}_i$ represents the predicted vehicle/fire-state class, $\hat{s}_i$ is the associated confidence score, and N is the number of detected instances. Unlike generic object detection, optimizing $\hat{DP}$ in road-incident scenarios is strictly constrained by sharp illumination variations, severe background clutter, and extreme spatial scale variance typical of vehicle fires. To explicitly address these constraints, we propose the Robust Dual-Path Fusion Detector (RDF-Det). The design retains two feature-aggregation paths so that semantic context and localization cues can be combined and different routing choices can be examined empirically.

As shown in Figure 1, the main RDF-Det configuration combines an optimized DINOv3 ConvNeXt-B backbone, an RDF-Neck using GLR, and an RTMDet detection head. The backbone, instantiated by either CSPNeXt-P5 or DINOv3 ConvNeXt-B in this study, extracts multilevel features, $\{C_3, C_4, C_5\}$. These features are then processed by the proposed RDF-Neck, which consists of four explicitly defined components: a Semantic Cascade Path (SCP), a Spatial Reinforcement Path (SRP), Reliability-Calibrated Fusion (RCF), and a Post-Fusion Alignment Block (PAB).

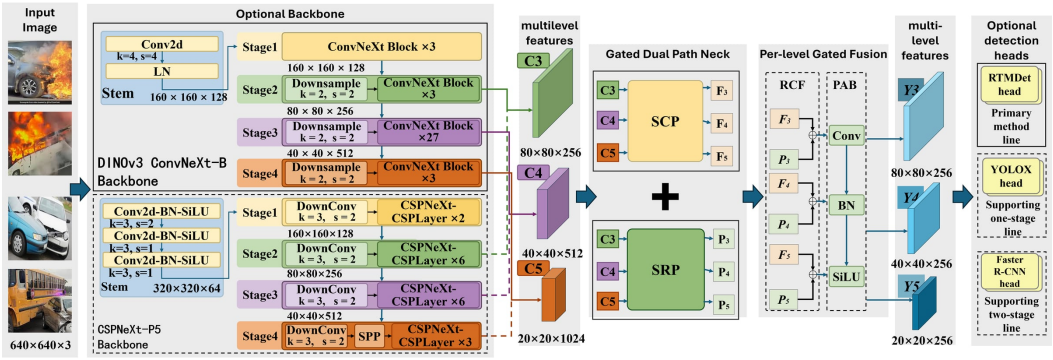


Figure 1. Macroscopic system architecture of RDF-Det. The RDF-Neck consists of an SCP (top–down FPN path), an SRP (CSPNeXtPAFPN path), an RCF (per-level fusion of F_l and P_l), and a PAB (Conv-BN-SiLU PostConv), followed by family-specific detection heads, including the RTMDet, YOLOX, and Faster R-CNN heads.

3.2. Overall Architecture of RDF-Det

RDF-Det is formulated as a three-stage CNN-based detection framework comprising a feature-extraction backbone, the proposed RDF-Neck, and a task-specific detection head. Given an input image, the backbone extracts a multi-scale feature set, $\{C_3, C_4, C_5\}$, which is fed into RDF-Neck to aggregate complementary semantic context and spatial localization cues across scales. The resulting aligned pyramid features $\{Y_3, Y_4, Y_5\}$ are then delivered to the detection head.

Figure 2 further details the microscopic topology of the two parallel feature paths. Specifically, the upper branch implements the SCP, while the lower branch implements the SRP. The level-wise RCF and PAB operations are applied to the corresponding outputs of these two paths formally defined in Section 3.3. This design separates the global detector pipeline from the local fusion operator, allowing the topology and routing mechanisms of RDF-Neck to be analyzed independently in Section 3.4.

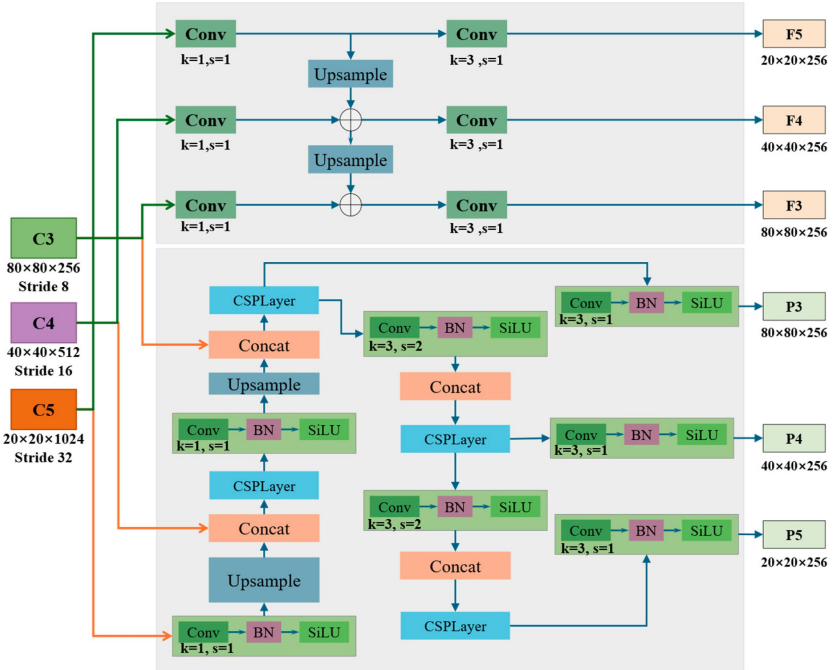


Figure 2. Microscopic topology of RDF-Neck.

RDF-Neck organizes semantic and spatial aggregation as two parallel paths whose outputs are combined at corresponding feature levels. The SCP applies top–down FPN aggregation, whereas the SRP applies the top–down and bottom–up aggregation of CSPNeXtPAFPN. Both branches receive the same backbone feature levels, and their outputs are fused at corresponding levels before post-convolution. This design differs from a single sequential path-aggregation neck and from the repeated fusion of adjacent levels in a weighted bidirectional pyramid. The routing comparison varies how the scalar mixing coefficients are shared and optimized while keeping the two branches fixed. These coefficients are image-independent model parameters shared across all inputs.

3.3. Robust Dual-Path Fusion Neck

Let $\mathcal{L} = \{3, 4, 5\}$ be the set of pyramid levels adopted in our detector. The corresponding backbone feature pyramid is represented as $\boldsymbol{C} = \{C_l\}_{l\in\mathcal{L}}$, where C_l denotes the feature map at level l. The RDF-Neck then processes $\boldsymbol{C}$ through two parallel structural pathways:

$$\boldsymbol{F} = \mathrm{SCP}(\boldsymbol{C}),\ \boldsymbol{P} = \mathrm{SRP}(\boldsymbol{C}). \tag{2}$$

where $\boldsymbol{F} = \{F_l\}_{l\in\mathcal{L}}$ denotes the semantic-cascade features generated by the top–down FPN-style path, and $\boldsymbol{P} = \{P_l\}_{l\in\mathcal{L}}$ denotes the spatial-reinforcement features generated by the CSPNeXtPAFPN-style path. For each level, l, F_l and P_l are aligned to the same spatial resolution and channel dimension before fusion. At each pyramid level, the two branch outputs are combined by RCF:

$$\hat{\boldsymbol{Y}} = \mathrm{RCF}(\boldsymbol{F}, \boldsymbol{P}; \boldsymbol{\alpha}) \tag{3}$$

where $\hat{\boldsymbol{Y}} = \{\alpha_l F_l + (1-\alpha_l)P_l\}_{l\in\mathcal{L}}$ is the pre-alignment fused feature, $\alpha_l \in [0, 1]$ is the branch-reliability coefficient for the SCP output, and $1-\alpha_l$ weights the SRP output. Here, reliability denotes the degree to which a feature source is trusted within the neck. For learnable-routing variants, the reliability coefficients in Equation (3) are parameterized by trainable routing logits:

$$\boldsymbol{\alpha} = \sigma(\boldsymbol{w}) \tag{4}$$

where $\mathbf{w} = \{w_l\}_{l\in\mathcal{L}}$ is the set of trainable scalar logits, and $\sigma(\cdot)$ denotes the sigmoid function. The constrained routing variants, including globally shared, fixed, and frozen coefficients, are defined separately in Section 3.4.

Finally, the fused feature pyramid in Equation (3) is further processed by PAB for post-fusion alignment:

$$\boldsymbol{Y} = PAB\big(\hat{\boldsymbol{Y}}\big) \tag{5}$$

where $\boldsymbol{Y} = \mathrm{SiLU}\Big(\mathrm{BN}\Big(\mathrm{Conv}_{3\times3}\Big(\big\{\hat{Y}_l\big\}_{l\in\mathcal{L}}\Big)\Big)\Big)$ is the final output feature delivered to the detection head. The PAB standardizes the feature distribution after SCP-SRP fusion and projects each pyramid level to a unified 256-channel representation.

3.4. Fusion and Routing Variants

To decouple the structural contribution of the dual-path topology from the parameterization of branch mixing, we formulate four RDF-Neck routing variants by progressively constraining the fusion coefficient α_l across pyramid levels, $l \in \{3, 4, 5\}$ (where α_l weights the SCP, and $1-\alpha_l$ weights the SRP). In Level-wise Learned Routing (LLR), each pyramid level is assigned an independent trainable logit, w_l, producing a level-specific coefficient, $\alpha_l = \sigma(w_l)$. In Global Learned Routing (GLR), the routing freedom is reduced to a single trainable logit, w_g, shared by all pyramid levels, such that $\alpha_l = \sigma(w_g)$ for every l. In Fixed Uniform Fusion (FUF), learnable routing is removed entirely, and the two branches are uniformly combined with $\alpha_l = 0.5$ at all levels, while retaining the dual-path topology. In Parameter-Matched Frozen Routing (PMFR), the level-wise routing-logit graph is retained

but its logits are frozen at values corresponding to $\alpha_l = 0.5$, preserving the stored routing structure while disabling routing optimization. Together, these four variants compare level-specific adaptive routing, globally shared routing, fixed-uniform fusion, and graph-matched frozen routing.

By monotonically reducing the degrees of routing freedom from three trainable parameters (LLR) to one shared parameter (GLR) and finally zero trainable routing parameters (FUF and PMFR) under an identical two-branch architecture, these four variants provide a rigorous control group for isolating the impact of routing flexibility on multi-scale representation learning. Figure 3 illustrates their computational graphs, and Table 1 summarizes the structural parameterization of each variant.

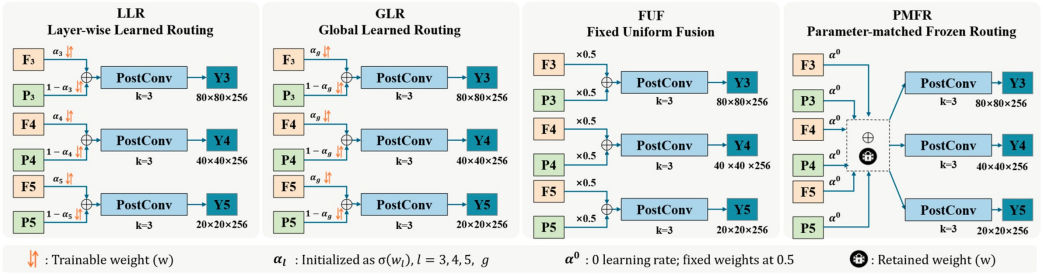


Figure 3. Computational graphs of the four fusion and routing variants. Each output combines the corresponding F and P feature levels.

Table 1. Fusion and routing variants of RDF-Neck.

Variant	Fusion Coefficient	Coefficient Sharing	Stored Routing Logits	Optimized Routing Logits	Ablation Role
LLR	$\alpha_l = \sigma(w_l)$	Level-specific	3	3	Fully adaptive reference
GLR	$\alpha_l = \sigma(w_g), \forall l$	Shared across pyramid levels	1	1	Level-sharing control
FUF	$\alpha_l = 0.5, \forall l$	Fixed uniform sharing	0	0	Fixed-fusion control
PMFR	$\alpha_l = \sigma(w_l^0) = 0.5$	Level-specific frozen values	3	0	Graph-matched frozen control

Note: Here, α_l denotes the SCP mixing coefficient at pyramid level $l \in \{3, 4, 5\}$, while $1 - \alpha_l$ weights the SRP. "Coefficient sharing" refers to sharing across pyramid levels rather than spatial locations.

3.5. Cross-Family Instantiations

To examine how the same dual-path interface behaves across detection frameworks, we instantiate it in three CNN-based detector families. RTMDet is the main detector for mechanism analysis, as it provides a stable anchor-free framework [42]. YOLOX represents one-stage detectors with a decoupled head [43], and Faster R-CNN represents two-stage proposal-based detectors [21]. To improve feature representation under severe road-incident degradation, DINOv3 ConvNeXt-B is adopted as an optional backbone [44,45].

4. Experimental Setup

This section describes the dataset, experimental configurations, training settings, and evaluation metrics, including the measures used to assess computational cost.

4.1. Dataset and Task Definition

The experiments are conducted on a four-class Transportation Fire-Risk (TFR) dataset benchmark explicitly tailored to road incidents involving ordinary vehicles, coaches, and oil tankers. We employ a domain-specific image collection designed to reflect the severe visual constraints of real-world transportation monitoring, including extreme background clutter, amorphous smoke structures, unpredictable illumination, vehicle reflections, and complex occlusions. The benchmark isolates fine-grained fire-risk perception into four standardized target categories: Car-Fire, Car-NoFire, LKYW-Fire, and LKYW-NoFire, where Car denotes ordinary vehicles excluding coaches and oil tankers, and LKYW denotes the coach and oil-tanker group. TFR was specifically curated to include fire images involving coaches and oil tankers, for which publicly available examples are scarce. Images showing smoke without visible flames are assigned the NoFire label. The original image-level split comprises 968 training, 277 validation, and 141 test images, approximately following a 7:2:1 ratio. The same split is used across all experimental configurations. Figure 4 shows representative challenging scenes, and Table 2 reports class frequencies, split sizes, and image-resolution statistics.

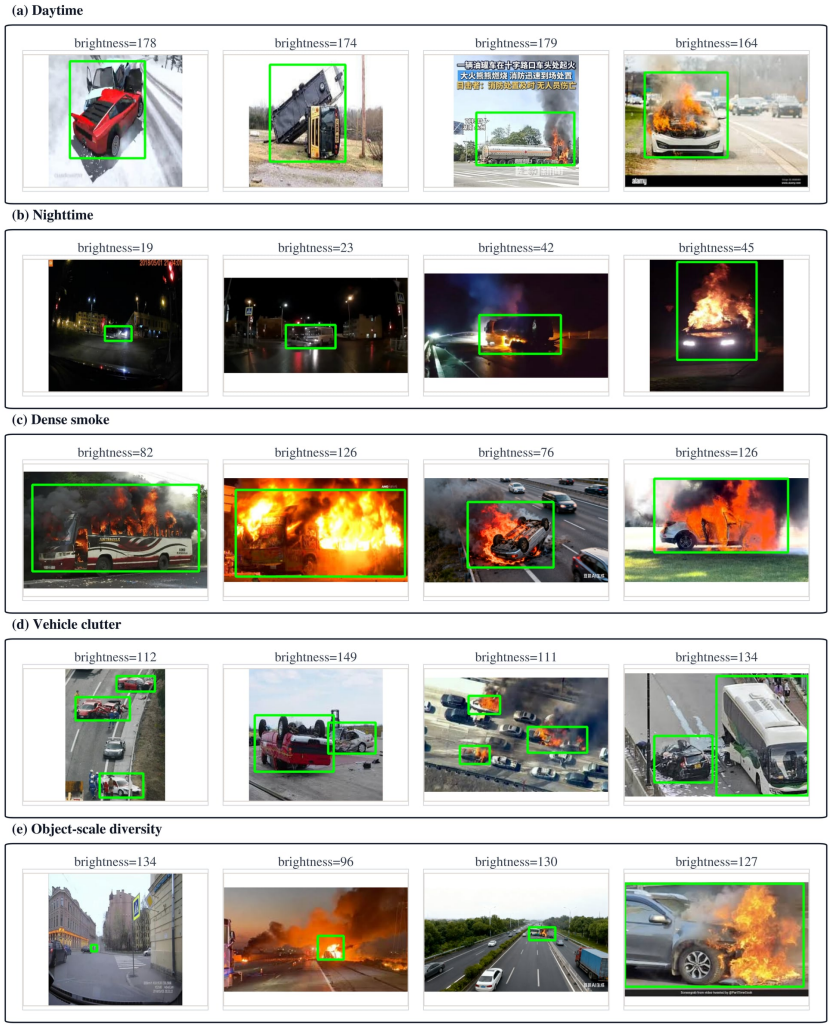


Figure 4. A montage of challenging scenarios covering (**a**) daytime, (**b**) nighttime, (**c**) dense smoke, (**d**) vehicle clutter, and (**e**) object-scale diversity.

Table 2. Details of the Transportation Fire-Risk (TFR) dataset.

Split	Images	Annotations	Car-F	Car-NF	LKYW-F	LKYW-NF	Mean Width	Mean Height
Training	968	988	204	249	236	299	801.8	516.1
Validation	277	283	55	68	73	87	749	489.3
Testing	141	142	33	39	36	34	805.7	512.1

Note: Car-F, Car-NF, LKYW-F, and LKYW-NF denote Car-Fire, Car-NoFire, LKYW-Fire, and LKYW-NoFire, respectively.

TFR combines public image resources and incident-related web pages, including image searches, traffic-incident news, and the Car Accident Severity resource on Roboflow Universe [46]. Bounding boxes cover the vehicle extent rather than separate flame or smoke regions, and all vehicle boxes and four-class labels were manually reviewed. The collection includes 135 identified AI-generated images: 98 in training, 22 in validation, and 15 in testing. These images are disclosed as synthetic samples.

No scene is shared across the training, validation, and test splits. In addition, a screened test subset of 110 images (containing 111 vehicle boxes: 20 Car-Fire, 37 Car-NoFire, 23 LKYW-Fire, and 31 LKYW-NoFire) is established prior to model evaluation to provide a unified held-out test benchmark across the representative configurations.

4.2. Experimental Configurations

The main study comprises 100 training runs for 20 configurations across RTMDet, YOLOX, and Faster R-CNN. RTMDet is the main setting for comparing the RDF routing variants. Backbone, neck, loss, and training variants provide context for interpreting the complete configurations. The YOLOX and Faster R-CNN blocks evaluate whether the dual-path interface transfers to one-stage and two-stage detection, respectively. Furthermore, five-seed runs of the recent YOLO26 detector are included as a reference.

In the main 20-configuration study, RTM-RDF-GLR-DINO is the main RDF-Det configuration reported in this paper. RTM-Base, YOLOX-Base, and FRCNN-Base are the family baselines. The remaining configurations are backbone, training, neck, loss, or routing controls, as specified in Table 3.

Table 3. Evaluated configurations and their roles across detector families.

Family	Configuration	Backbone	Neck/Fusion Module	Control Setting	Experimental Role
RTMDet	RTM-Base	CSPNeXt-P5	CSPNeXtPAFPN	—	Family baseline
	RTM-DINO	DINOv3 ConvNeXt-B	CSPNeXtPAFPN	Strict pretrained loading	Backbone-only control
	RTM-DINO-Opt	DINOv3 ConvNeXt-B	CSPNeXtPAFPN	Backbone lr multiplier = 0.1	Optimized backbone control
	RTM-DINO-Opt-WD0	DINOv3 ConvNeXt-B	CSPNeXtPAFPN	Backbone wd = 0	Regularization control
	RTM-FPN	CSPNeXt-P5	FPN	—	Neck-only control
	RTM-Focal	CSPNeXt-P5	CSPNeXtPAFPN	Focal-loss variant	Loss-function control
	RTM-RDF-LLR-CSP	CSPNeXt-P5	RDF-Neck	LLR	Dual-path control without DINOv3
	RTM-RDF-LLR-DINO	DINOv3 ConvNeXt-B	RDF-Neck	LLR	Adaptive-routing reference
	RTM-RDF-GLR-DINO (RDF-Det)	DINOv3 ConvNeXt-B	RDF-Neck	GLR	Main RDF-Det configuration
	RTM-RDF-FUF-DINO	DINOv3 ConvNeXt-B	RDF-Neck	FUF	Fixed-fusion control
	RTM-RDF-PMFR-DINO	DINOv3 ConvNeXt-B	RDF-Neck	PMFR	Frozen-routing control

Table 3. *Cont.*

Family	Configuration	Backbone	Neck/Fusion Module	Control Setting	Experimental Role
YOLOX	YOLOX-Base	CSPDarknet-L	YOLOXPAFPN	—	Family baseline
	YOLOX-DINO	DINOv3 ConvNeXt-B	YOLOXPAFPN	Strict pretrained loading	Backbone-transfer control
	YOLOX-CSPPAFPN	CSPDarknet-L	CSPNeXtPAFPN	—	Neck-transfer control
	YOLOX-GIoU	CSPDarknet-L	YOLOXPAFPN	GIoU-loss variant	Loss-function control
	YOLOX-RDF-LLR-DINO	DINOv3 ConvNeXt-B	RDF-Neck adapter	LLR	Supporting RDF instantiation
Faster R-CNN	FRCNN-Base	ResNet-50	FPN	—	Family baseline
	FRCNN-Neck	ResNet-50	Two-stage neck control	—	Neck-only control
	FRCNN-DINO	DINOv3 ConvNeXt-B	FPN	Strict pretrained loading	Backbone-transfer control
	FRCNN-DINO-Neck	DINOv3 ConvNeXt-B	Two-stage neck control	—	Combined transfer control

4.3. Training Settings and Experimental Controls

To reduce the confounding effect of random initialization, the 20 main configurations are trained independently with the same set of five random seeds: 0, 1, 42, 1234, and 3407. This matched-seed design allows paired performance differences to be computed between each proposed variant and its corresponding baseline, providing a controlled basis for the subsequent significance analysis. RTMDet uses Mosaic/MixUp, random resizing and cropping, HSV augmentation, and horizontal flips; YOLOX uses Mosaic/MixUp, random affine transforms, HSV augmentation, horizontal flips, and multiscale resizing; and Faster R-CNN uses resizing and horizontal flips. The pretrained DINOv3 ConvNeXt-B backbone is jointly fine-tuned with the neck and detection head. The SCP mixing coefficients are initialized to 0.85 for LLR and 0.5 for GLR. For each run, both the best validation checkpoint and the final-epoch checkpoint are retained. Table 4 reports the input resolution, training and validation batch sizes, number of epochs, and optimizer settings used for each detector family.

Table 4. Training configurations across detector families.

Family	Input Size	Train Batch	Val Batch	Epochs	Optimizer
RTMDet	640 × 640	8	4	300	AdamW (lr = 0.004, wd = 0.05)
YOLOX	640 × 640	8; 4 × 2 accum. (DINO)	8; 4 (DINO)	300	SGD (lr = 0.01, wd = 5×10^{-4}, momentum = 0.9)
Faster R-CNN	896 × 896	2	1	70	AdamW (lr = 2×10^{-4}, wd = 0.05)

For the recent-detector reference, the pretrained YOLO26m from Ultralytics 8.4.19 [47] is trained with five seeds (0, 1, 42, 205, and 3407). Training uses a 640 × 640 input, a batch size of 16, and SGD with an initial learning rate of 0.01 and a weight decay of 0.0005, for up to 300 epochs with an early-stopping patience of 50 epochs. Cosine learning-rate decay, Mosaic, MixUp, and standard geometric and color augmentations are applied.

4.4. Evaluation Metrics and Statistical Analysis

The primary metric for joint localization and classification performance is the COCO-style mean average precision over IoU thresholds from 0.50 to 0.95 with a step size of 0.05, denoted as AP in the following sections. As a secondary application-oriented metric, AP at an IoU threshold of 0.50, denoted as AP50, is reported to measure coarse localization robustness in cluttered incident scenes. For each configuration, the reported AP and AP50 are computed separately for each metric as the mean of its best validation scores over five repeated runs. To characterize training stability and avoid conclusions driven by transient

validation peaks, the best–final gap is defined as the performance difference between the best validation checkpoint and the final training epoch. The convergence behavior is further documented by recording the epoch at which the validation performance peaks.

To strengthen statistical support for empirical comparisons, performance differences are analyzed using a paired statistical analysis. For each baseline–variant comparison in the main configuration matrix, paired differences are computed across the same matched initialization seeds. In addition to raw p-values, we report paired t-statistics, 95% confidence intervals, and paired Cohen's (d_z) effect sizes. To account for multiple comparisons across all 17 baseline–variant pairs for each metric, raw p-values are adjusted using the Holm–Bonferroni correction procedure. Statistical significance labels are assigned only according to the Holm-corrected p-values, thereby reducing the risk of attributing stochastic training variation to architectural improvement.

Beyond the validation benchmark, we further conduct a held-out test evaluation on the screened test subset defined in Section 4.1 across all five seeds (0, 1, 42, 1234, and 3407) for four representative RTMDet configurations: RTM-Base, RTM-DINO-Opt, RTM-RDF-GLR-DINO, and RTM-RDF-FUF-DINO. To strictly prevent test-set leakage, each run is evaluated directly using the checkpoint selected by the highest validation AP under identical preprocessing and standard COCO evaluation settings, without any retraining or test-guided model selection.

4.5. Computational Complexity and Inference Efficiency

The computational performance of the instantiated architecture is evaluated from two complementary perspectives: static model complexity and runtime inference speed. Static complexity is quantified by parameter count and estimated floating-point operations (FLOPs), both normalized to a 640 × 640 input resolution and computed using the MMEngine analysis utility under the validation-data preprocessing setting. Runtime efficiency is assessed using batch-1 server-side inference latency and frames per second (FPS) on an NVIDIA RTX 4090D GPU. All speed measurements are conducted with PyTorch 2.1.2 (CUDA 11.8) eager execution, using 20 warmup iterations followed by 100 CUDA-synchronized timed iterations to reduce measurement noise and improve timing stability.

5. Results and Discussion

This section presents the benchmark results, statistical comparisons, routing analysis, and optimization behavior, followed by computational cost and qualitative analysis of successes and failures.

5.1. Overall Benchmark Performance

Table 5 and Figure 5 summarize the main benchmark results obtained from the 100-run evaluation. The results lead to three main observations regarding model performance. First, the highest mean AP scores in the main matrix are achieved by RTMDet-based configurations, providing the main setting for the proposed fire-risk detection task. The highest mean AP is obtained by RTM-RDF-GLR-DINO, with 0.6352 AP and 0.9290 AP50. Its 95% confidence interval for AP, [0.6308, 0.6396], together with a small best-final AP gap of 0.18 points, indicates high accuracy and low cross-seed variability.

Table 5. Main benchmark results over 100 runs with mean AP/AP50, AP confidence intervals, and best-final gaps.

Configuration ID	Runs	AP (Mean ± std.)	95% CI (AP)	AP50 (Mean ± std.)	Final AP	Final AP50	Best-Final Gap (AP Points)	Best-Final Gap (AP50 Points)
RTM-RDF-GLR-DINO (RDF-Det)	5	0.6352 ± 0.0036	[0.6308, 0.6396]	0.9290 ± 0.0040	0.6334	0.9238	0.18	0.52
RTM-RDF-FUF-DINO	5	0.6348 ± 0.0077	[0.6252, 0.6444]	0.9288 ± 0.0037	0.6332	0.9278	0.16	0.10
RTM-RDF-PMFR-DINO	5	0.6314 ± 0.0059	[0.6241, 0.6387]	0.9312 ± 0.0079	0.6298	0.9214	0.16	0.98
RTM-DINO-Opt	5	0.6312 ± 0.0064	[0.6233, 0.6391]	0.9214 ± 0.0059	0.6300	0.9212	0.12	0.02
RTM-RDF-LLR-DINO	5	0.6294 ± 0.0054	[0.6227, 0.6361]	0.9214 ± 0.0044	0.6270	0.9152	0.24	0.62
RTM-RDF-LLR-CSP	5	0.5840 ± 0.0043	[0.5787, 0.5893]	0.8912 ± 0.0061	0.5828	0.8912	0.12	0.00
RTM-FPN	5	0.5784 ± 0.0038	[0.5736, 0.5832]	0.8898 ± 0.0043	0.5778	0.8898	0.06	0.00
RTM-DINO	5	0.5742 ± 0.0063	[0.5664, 0.5820]	0.8630 ± 0.0060	0.5592	0.8528	1.50	1.02
RTM-DINO-Opt-WD0	5	0.5730 ± 0.0032	[0.5690, 0.5770]	0.8628 ± 0.0040	0.5620	0.8554	1.10	0.74
RTM-Base	5	0.5628 ± 0.0145	[0.5447, 0.5809]	0.8608 ± 0.0127	0.5628	0.8608	0.00	0.00
RTM-Focal	5	0.5506 ± 0.0049	[0.5445, 0.5567]	0.8470 ± 0.0064	0.5500	0.8466	0.06	0.04
YOLOX-RDF-LLR-DINO	5	0.6240 ± 0.0051	[0.6176, 0.6304]	0.9314 ± 0.0082	0.6192	0.8990	0.48	3.24
YOLOX-DINO	5	0.6218 ± 0.0090	[0.6106, 0.6330]	0.9308 ± 0.0067	0.6198	0.8978	0.20	3.30
YOLOX-CSPPAFPN	5	0.5098 ± 0.0105	[0.4968, 0.5228]	0.8196 ± 0.0097	0.5098	0.8196	0.00	0.00
YOLOX-Base	5	0.4676 ± 0.0117	[0.4531, 0.4821]	0.8046 ± 0.0142	0.4676	0.8046	0.00	0.00
YOLOX-GIoU	5	0.4610 ± 0.0192	[0.4372, 0.4848]	0.8086 ± 0.0192	0.4610	0.8086	0.00	0.00
FRCNN-DINO	5	0.5864 ± 0.0062	[0.5787, 0.5941]	0.9072 ± 0.0151	0.5728	0.8760	1.36	3.12
FRCNN-DINO-Neck	5	0.5772 ± 0.0070	[0.5685, 0.5859]	0.9098 ± 0.0133	0.5542	0.8580	2.30	5.18
FRCNN-Base	5	0.3864 ± 0.0093	[0.3748, 0.3980]	0.7368 ± 0.0044	0.3578	0.6724	2.86	6.44
FRCNN-Neck	5	0.2792 ± 0.0572	[0.2082, 0.3502]	0.6092 ± 0.0714	0.2678	0.5798	1.14	2.94

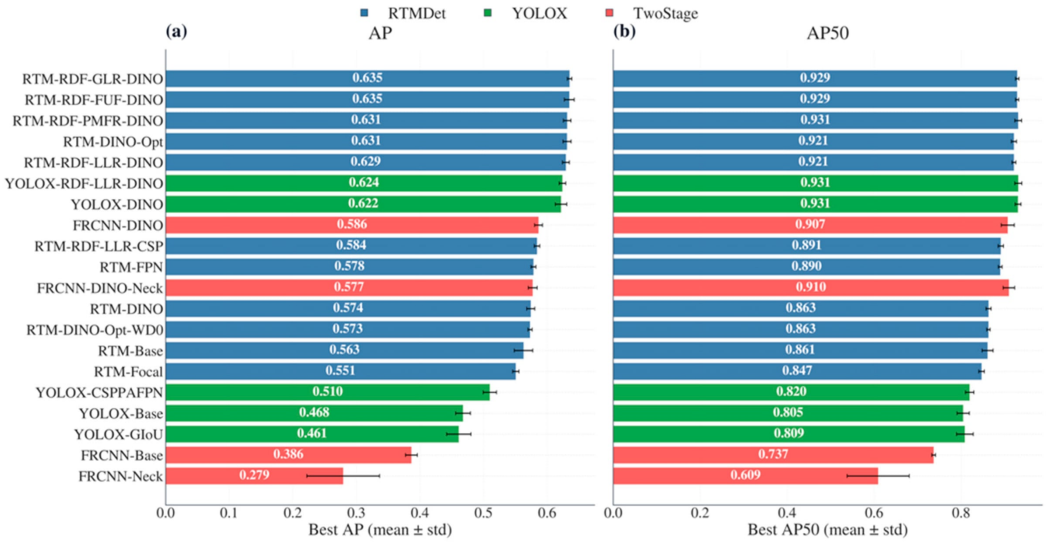


Figure 5. Mean AP and AP50 benchmark performance across the 100-run evaluation.

Second, RTM-RDF-FUF-DINO and RTM-RDF-PMFR-DINO achieve 0.6348 and 0.6314 AP, respectively, close to the 0.6352 AP of RTM-RDF-GLR-DINO. Across the RTMDet routing variants, global shared routing (GLR) and fixed-uniform fusion (FUF) both outperform unconstrained level-wise routing (LLR), indicating that the dual-path representation does not rely on complex per-level weight parameterization to achieve high validation accuracy.

Third, upgrading the backbone to DINOv3 ConvNeXt-B substantially elevates baseline performance across all three detector families (reaching 0.6312 AP in RTM-DINO-Opt,

0.6218 AP in YOLOX-DINO, and 0.5864 AP in FRCNN-DINO). Building upon these strong backbone-enhanced references, incorporating RDF-Neck yields further accuracy and stability gains in the one-stage anchor-free detectors (RTMDet and YOLOX) while exhibiting different adaptation behaviors in the two-stage proposal-based Faster R-CNN framework.

Table 6 adds the standard YOLO26 reference. Its five recorded runs yield 0.6235 ± 0.0163 AP and 0.9071 ± 0.0079 AP50. RTM-RDF-GLR-DINO has a higher reported mean AP by 1.17 points. This comparison shows that RDF-Det has a higher mean AP than the latest YOLO-series architecture on the same vehicle-fire benchmark.

Table 6. Validation results with the YOLO26 reference.

Configuration	Runs	AP	AP50
YOLO26	5	0.6235 ± 0.0163	0.9071 ± 0.0079
RTM-DINO-Opt	5	0.6312 ± 0.0064	0.9214 ± 0.0059
RTM-RDF-GLR-DINO (RDF-Det)	5	0.6352 ± 0.0036	0.9290 ± 0.0040

Note: Values are means ± sample standard deviations over five runs on the fixed TFR validation split.

RTM-RDF-GLR-DINO has a cross-seed AP standard deviation of 0.0036, approximately 22.1% of that of YOLO26 (0.0163). At IoU = 0.50, it obtains 0.9290 AP50, exceeding YOLO26 (0.9071) and RTM-Base (0.8608) by 2.19 and 6.82 AP50 points, respectively. These results demonstrate that RDF-Det improves both localization reliability at IoU = 0.50 and run-to-run training consistency compared with standard single-path detectors.

5.2. Family-Specific Analysis

RTMDet provides the main setting for the structure and routing analysis. RTM-RDF-GLR-DINO reaches 0.6352 AP, compared with 0.5628 for RTM-Base. The 7.24-point increase compares complete configurations, including backbone, optimization, and neck changes. RTM-DINO-Opt already reaches 0.6312 AP, leaving a descriptive difference of 0.40 AP points to GLR. These configurations also differ in backbone weight decay, so this difference is not presented as an isolated neck effect. Under the same CSPNeXt backbone without DINOv3, replacing the native neck with RDF-Neck (RTM-RDF-LLR-CSP) increases the mean AP by 2.12 points (from 0.5628 to 0.5840 AP) and AP50 by 3.04 points (from 0.8608 to 0.8912 AP50). Together, these comparisons distinguish the 7.24-point system-level gain from the consistent local improvements provided by RDF-Neck under both CSPNeXt and DINOv3 backbones.

In YOLOX, the baseline reaches 0.4676 AP, the DINO-backed reference reaches 0.6218, and YOLOX-RDF-LLR-DINO reaches 0.6240. The large gain over YOLOX-Base includes the effect of the backbone change, whereas the 0.22-point gain over YOLOX-DINO isolates the contribution of RDF and indicates a modest improvement. This confirms that the parallel dual-path representation transfers positively to a decoupled-head one-stage detector under the same DINOv3 backbone.

The Faster R-CNN results show a contrasting behavior in a two-stage detector. FRCNN-Base reaches 0.3864 AP, FRCNN-DINO reaches 0.5864, and FRCNN-DINO-Neck reaches 0.5772, which is 0.92 AP points below the DINO-backed reference. The adaptation also reduces the pyramid from five outputs to four and adjusts the RPN strides accordingly. Because the pyramid feeds both proposal generation and RoI processing, these interface changes may account for the drop. This contrasting outcome highlights that two-stage detectors are more sensitive to pyramid-level truncation and RPN stride alignment than one-stage dense predictors. Table 7 summarizes these local comparisons across all three detector families.

Table 7. Local comparisons for interpreting structure, fusion, and detector adaptation.

Reference Configuration	Compared Configuration	Reference AP	AP	ΔAP (Points)
RTM-Base	RTM-RDF-LLR-CSP	0.5628	0.5840	+2.12
RTM-DINO-Opt	RTM-RDF-GLR-DINO (RDF-Det)	0.6312	0.6352	+0.40
RTM-RDF-LLR-DINO	RTM-RDF-FUF-DINO	0.6294	0.6348	+0.54
YOLOX-DINO	YOLOX-RDF-LLR-DINO	0.6218	0.6240	+0.22
FRCNN-DINO	FRCNN-DINO-Neck	0.5864	0.5772	−0.92

5.3. Statistical Reliability with Corrected Inference

Table 8 summarizes the main family-level statistical outcomes, and Table 9 reports the complete Holm-corrected paired comparisons relative to the corresponding family baselines. The paired statistical analysis further supports the repeatability of differences between the complete configurations under the matched-seed evaluation design (Figure 6). Within the main RTMDet evaluation branch, the paired comparison between RTM-RDF-GLR-DINO and RTM-Base yields an improvement of 7.24 AP points. This gain is supported by a strictly positive 95% confidence interval of [5.61, 8.87] AP points, a large paired effect size of Cohen's $d_z = 5.516$, and a Holm-corrected p-value of 0.0031. The non-learned routing variant, RTM-RDF-FUF-DINO, also remains statistically significant after multiplicity correction, with an AP difference of 7.20 points and a paired effect size of $d_z = 3.728$. These paired tests confirm that the system-level gains of the dual-path configurations over the baseline are statistically robust against random seed initialization.

Table 8. Summary of Holm-corrected principal baseline comparisons across detector families.

Family	Comparison	Seeds	ΔAP (Points)	95% CI (AP Points)	dz (AP)	Holm p	Holm sig.
RTMDet	RTM-Base vs. RTM-RDF-GLR-DINO (RDF-Det)	5	7.24	[5.61, 8.87]	5.516	0.0031	**
YOLOX	YOLOX-Base vs. YOLOX-RDF-LLR-DINO	5	15.64	[14.21, 17.07]	13.597	9.76×10^{-5}	***
Faster R-CNN	FRCNN-Base vs. FRCNN-DINO	5	20.00	[18.51, 21.49]	16.667	4.64×10^{-5}	***

Note: **, $p < 0.01$; ***, $p < 0.001$.

Table 9. Holm-corrected paired comparisons relative to main family baselines under five seeds.

Variant	ΔAP (Points)	dz (AP)	Holm sig. (AP)	ΔAP50 (Points)	dz (AP50)	Holm sig. (AP50)
RTM-RDF-GLR-DINO (RDF-Det)	7.24	5.516	**	6.82	5.282	**
RTM-RDF-FUF-DINO	7.20	3.728	*	6.80	5.068	**
RTM-RDF-PMFR-DINO	6.86	5.596	**	7.04	6.215	**
RTM-DINO-Opt	6.84	3.792	*	6.06	4.551	**
RTM-RDF-LLR-DINO	6.66	5.089	**	6.06	4.657	**
RTM-RDF-LLR-CSP	2.12	1.756	ns	3.04	2.008	ns
RTM-FPN	1.56	0.942	ns	2.90	2.734	*
RTM-DINO	1.14	1.093	ns	0.22	0.223	ns
RTM-DINO-Opt-WD0	1.02	0.733	ns	0.20	0.155	ns
RTM-Focal	−1.22	−0.863	ns	−1.38	−1.010	ns
YOLOX-RDF-LLR-DINO	15.64	13.597	***	12.68	8.555	***
YOLOX-DINO	15.42	22.211	***	12.62	9.819	***
YOLOX-CSPPAFPN	4.22	2.041	ns	1.50	0.671	ns
YOLOX-GIoU	−0.66	−0.273	ns	0.40	0.189	ns
FRCNN-DINO	20.00	16.667	***	17.04	9.327	***
FRCNN-DINO-Neck	19.08	21.508	***	17.30	12.668	***
FRCNN-Neck	−10.72	−1.813	ns	−12.76	−1.748	ns

Note: ns, $p \geq 0.05$; *, $p < 0.05$; **, $p < 0.01$; ***, $p < 0.001$.

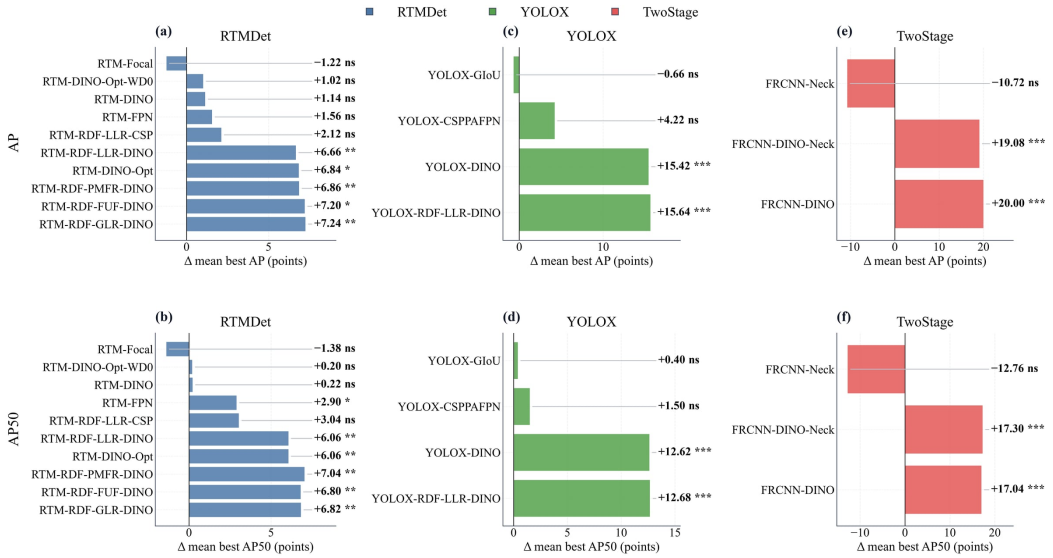


Figure 6. Differences between complete configurations and their family baselines in AP/AP50 points, with statistical significance based on Holm-adjusted p-values from paired t-tests across five matched random seeds (ns, $p \geq 0.05$; *, $p < 0.05$; **, $p < 0.01$; ***, $p < 0.001$).

A similar pattern is observed in the cross-family evaluations. YOLOX-RDF-LLR-DINO produces a large corrected improvement over YOLOX-Base, with a paired effect size of $d_z = 13.597$ and a Holm-corrected p-value of approximately 9.76×10^{-5}. The two-stage line shows a similar direction of evidence, where FRCNN-DINO provides a substantial corrected improvement over FRCNN-Base. At the same time, the Holm correction plays an important filtering role by preventing weak component-level changes from being over-interpreted. Marginal interventions such as RTM-FPN on AP and YOLOX-CSPPAFPN do not remain significant after correction, indicating that not every isolated backbone or neck substitution should be treated as a reliable architectural gain.

5.4. Fusion and Routing Analysis

The RTMDet control block compares adaptive and fixed routing within the same dual-path topology (Figure 7). Within this setting, the main question is whether more flexible learned mixing is needed to obtain competitive performance from the two branches. The fully adaptive level-wise routing reference, RTM-RDF-LLR-DINO, achieves 0.6294 AP and 0.9214 AP50. Removing the learnable routing mechanism entirely with the fixed uniform fusion variant, RTM-RDF-FUF-DINO, changes AP by only +0.54 points relative to this reference. Its 95% confidence interval, $[-0.86, 1.94]$ AP points, crosses zero, and the Holm-corrected p-value of 0.6862 indicates that this AP difference is not statistically significant. On the secondary AP50 metric, however, RTM-RDF-FUF-DINO improves by 0.74 AP50 points, with a Holm-corrected p-value of 0.0168. This indicates that unconstrained level-specific routing weights are not strictly required within the dual-path topology, as uniform branch combination already preserves the complementary semantic and spatial representations while yielding a statistically significant gain in AP50.

The global-routing variant, RTM-RDF-GLR-DINO, provides a globally shared mixing alternative. Compared with the RTM-RDF-LLR-DINO reference, it improves AP by 0.58 points, reaches Holm-corrected significance, and yields an effect size of $d_z = 1.966$. In contrast, the parameter-matched frozen-routing variant, RTM-RDF-PMFR-DINO, maintains competitive accuracy without demonstrating a statistically significant gain over the fully adaptive reference. While PMFR achieves the highest mean AP50, GLR attains the highest overall AP (0.6352) alongside the lowest cross-seed standard deviation (0.36 vs.

0.77 AP points for FUF), justifying its selection as the main RDF-Det configuration and demonstrating that compact global routing or simple uniform fusion is more effective than over-parameterized level-wise routing once the parallel dual-path topology is established.

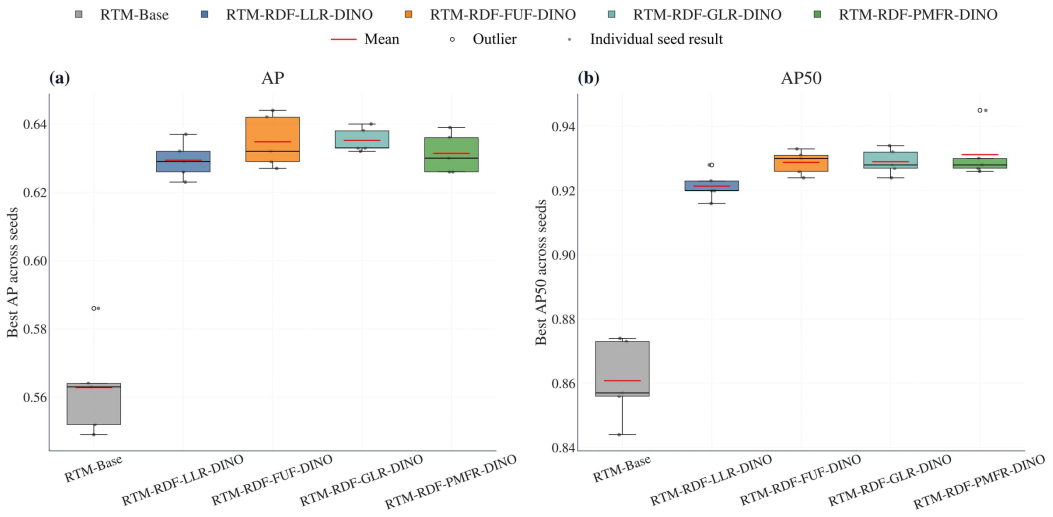


Figure 7. RTMDet-family comparison of RDF-Neck fusion and routing variants in terms of AP and AP50.

5.5. Optimization Dynamics and Convergence Stability

Optimization stability is further analyzed through cross-seed variability, best-final validation gaps and best-epoch convergence statistics. Figure 8 positions representative models in an accuracy-stability space using the mean and standard deviation of best AP and AP50, while Figure 9 shows when the best AP and AP50 checkpoints are reached. Table 10 reports the corresponding mean peak-epoch statistics. The strongest RTMDet dual-path variants exhibit very small AP best-final gaps, ranging from 0.16 to 0.18 points. This narrow range indicates that the models do not rely on isolated transient validation spikes; instead, their final training states remain close to their best validation checkpoints, suggesting stable performance retention until the end of the training schedule. In contrast, the DINO-based YOLOX variants and two-stage variants show larger best-final gaps, especially on AP50, indicating greater sensitivity to temporary validation peaks and less stable late-stage optimization behavior.

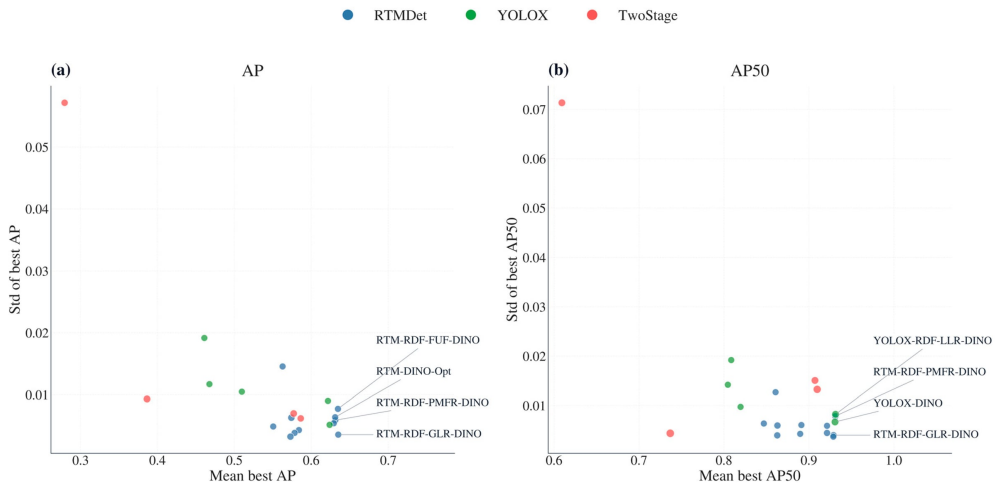


Figure 8. Accuracy–stability profiles of representative models.

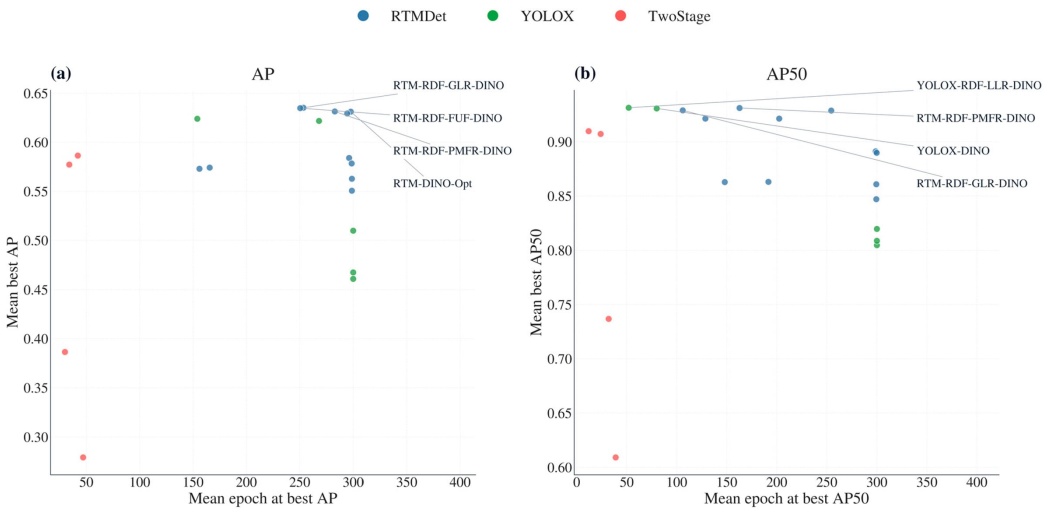


Figure 9. Best-epoch convergence dynamics across detector families.

Table 10. Peak-epoch statistics for representative models in terms of AP and AP50.

Configuration	Mean Best Epoch (AP)	Mean Best Epoch of (AP50)
RTM-RDF-GLR-DINO (RDF-Det)	253	106
RTM-RDF-FUF-DINO	250	254
RTM-RDF-LLR-DINO	294	129
RTM-Base	299	299
RTM-DINO-Opt	298	202
FRCNN-Base	30	32
FRCNN-DINO-Neck	34	12
FRCNN-DINO	42	24
YOLOX-RDF-LLR-DINO	154	52
YOLOX-Base	300	300

The best-epoch statistics further distinguish the optimization profiles of different variants. RTM-Base reaches its best AP very late, at epoch 299, which is consistent with its near-zero best-final gap. In comparison, RTM-RDF-GLR-DINO reaches its best AP50 around epoch 106, while its best COCO-style AP appears much later, around epoch 253. RTM-RDF-FUF-DINO shows a more synchronized convergence pattern, with its best AP and AP50 appearing near epochs 250 and 254, respectively. These different temporal trajectories suggest that routing parameterization affects the optimization path even when the final AP values remain close. Therefore, peak accuracy alone is insufficient for evaluating architectural robustness; convergence stability should also be considered.

5.6. Computational Cost and Runtime Efficiency

RDF-Det introduces a measurable but practical accuracy–efficiency trade-off. At a 640 × 640 input resolution, RTM-Base requires approximately 53 M parameters and 80 G FLOPs. After integrating the DINOv3 backbone and the RDF dual-path fusion neck, RTM-RDF-GLR-DINO increases the static complexity to approximately 118 M parameters and 182 G FLOPs. This increase is mainly driven by the larger DINOv3 backbone and the added dual-path fusion neck.

However, the increase in inference latency is smaller than the increase in parameter count and FLOPs. Under the standardized server-side benchmarking setting on an NVIDIA RTX 4090D GPU with PyTorch eager execution, batch size 1, and 100 synchronized timed

iterations, RTM-Base achieves 75.5 FPS, corresponding to 13.2 ms per image. RTM-RDF-GLR-DINO achieves 42.8 FPS, corresponding to 23.4 ms per image. Thus, the proposed model retains 56.7% of the baseline throughput with a system-level difference of 7.24 AP points from RTM-Base. These measurements characterize the server-side accuracy–cost trade-off under the recorded hardware and execution settings. Table 11 and Figure 10 summarize this Pareto trade-off.

Table 11. Complexity and throughput summary for representative models.

Configuration	Params (M)	FLOPs (G)	FPS	ΔAP (Points) vs. RTM-Base	Speed Retained (RTM-Base)
RTM-Base	53	80	75.5	0.00	100%
RTM-DINO-Opt	114	171	42.0	6.84	55.6%
RTM-RDF-LLR-DINO	118	182	43.4	6.66	57.5%
RTM-RDF-GLR-DINO (RDF-Det)	118	182	42.8	7.24	56.7%
RTM-RDF-FUF-DINO	118	182	41.6	7.20	55.1%

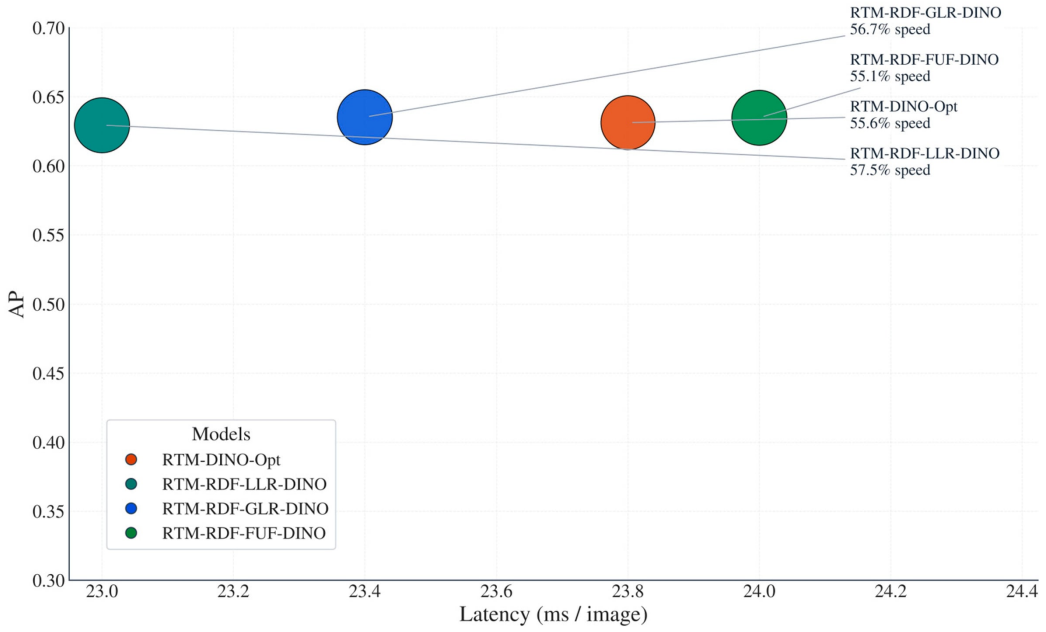


Figure 10. Accuracy–efficiency trade-off of representative configurations in terms of mean AP, parameter count, FLOPs, and server-side throughput.

Relative to RTM-DINO-Opt, RTM-RDF-GLR-DINO uses 118 M parameters (vs. 114 M) and 182 G FLOPs (vs. 171 G), increases of approximately 3.5% and 6.4%, respectively. Measured throughput is comparable (42.8 vs. 42.0 FPS) under the same server-side settings. Compared with RTM-Base, the GLR configuration improves the mean AP by 7.24 points while throughput decreases from 75.5 to 42.8 FPS.

5.7. Qualitative Assessment

The qualitative results are consistent with the quantitative detection results. As shown in Figure 11, RDF-Det more effectively suppresses background distractors and maintains tighter bounding-box localization under transportation-specific visual degradation.



Figure 11. Qualitative comparison between the baseline detector and RDF-Det under challenging transportation incident conditions, including false-positive suppression, localization refinement, and clutter suppression.

In false-positive suppression, the baseline produces redundant and partially misplaced predictions around flame-like textures and surrounding vehicle structures, whereas RDF-Det reduces these spurious responses and focuses on the actual fire-related regions. In the localization refinement, RDF-Det generates bounding boxes that are more closely aligned with the ground-truth object extent, particularly under smoke occlusion and small flame regions. In the clutter suppression, RDF-Det remains stable in dense traffic scenes, where nearby vehicles, reflections, and complex road backgrounds may otherwise trigger ambiguous detections.

Figure 11 illustrates localization and background-response differences in selected validation images. Figure 12 complements these examples with the observed failure modes.

Figure 12 shows four failure cases of RTM-RDF-GLR-DINO from the same qualitative validation analysis, with predictions displayed at a score threshold of 0.3. In Figure 12a, the small annotated non-fire vehicle is missed. In Figure 12b, the predicted box covers only part of the annotated burning vehicle in dense smoke. Figure 12c shows redundant NoFire detections extending beyond the annotated vehicle, which are localization/background errors and do not trigger fire alarms. In Figure 12d, low illumination and smoke are accompanied by oversized boxes and a Car-Fire prediction for an LKYW-Fire target.

These examples show that small vehicle extent, smoke occlusion, clutter, and low illumination can impair localization or vehicle-group classification. Figure A1 compares RTM-Base and RDF-Det on weak flame cues, dense smoke, nighttime illumination, vehicle reflections, and severe occlusion. With weak flame cues, the baseline produces competing LKYW-Fire and LKYW-NoFire predictions, whereas RDF-Det retains an LKYW-Fire detection with a score of 0.88. In the dense smoke scene, RDF-Det removes the baseline's additional LKYW-NoFire prediction and improves localization. In the nighttime illumination scene, RDF-Det retains an LKYW-Fire detection without the baseline's additional Car-Fire prediction. In the vehicle reflection scene, both models predict LKYW-NoFire, while RDF-Det more closely matches the annotated vehicle extent. In the severe occlusion scene, the baseline produces no detection above the display threshold, whereas RDF-Det detects the LKYW-NoFire target with a score of 0.85.

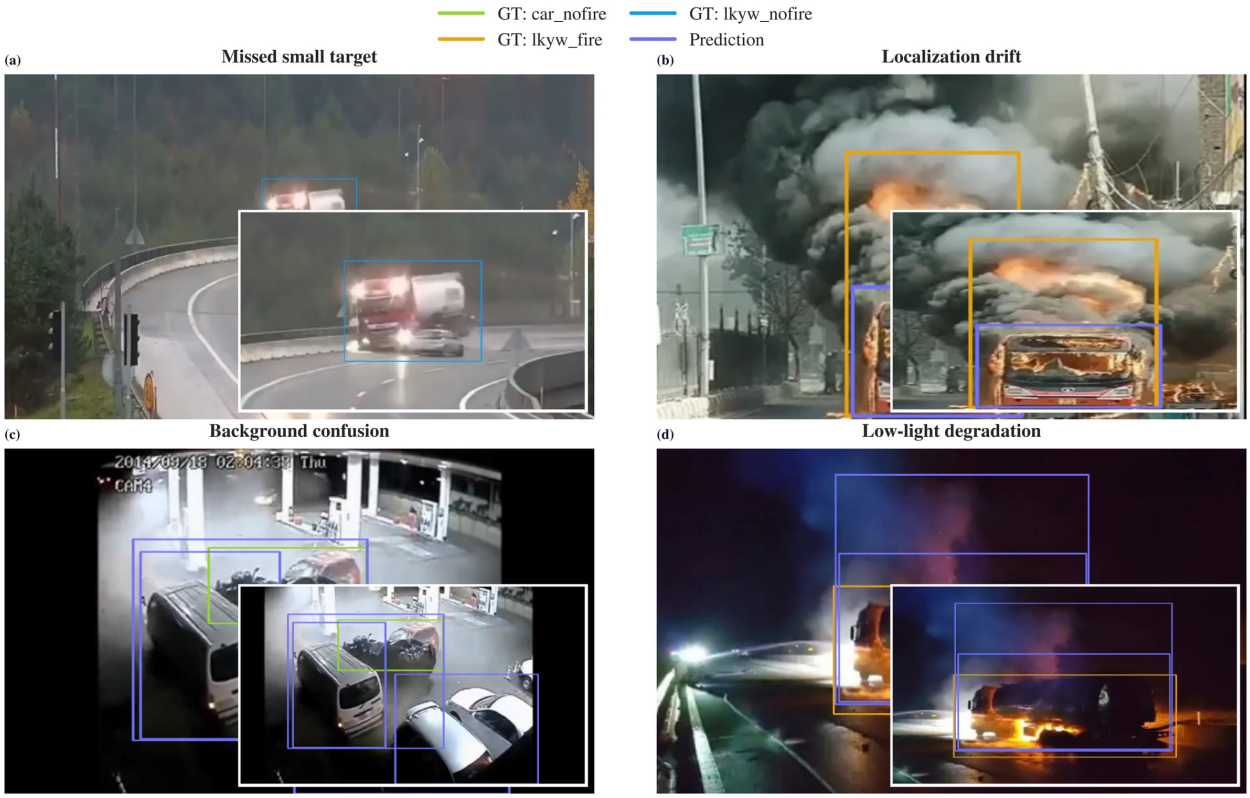


Figure 12. Failure cases of RTM-RDF-GLR-DINO: (**a**) missed small non-fire vehicle; (**b**) partial localization of a burning vehicle under dense smoke; (**c**) redundant NoFire detections in a cluttered scene; and (**d**) localization and vehicle-group classification errors in low light.

5.8. Discussion and Limitations

The contribution of RDF-Det is the parallel semantic–spatial structure together with an analysis of how that structure is fused and adapted. The RTMDet results favor examining simple fusion carefully: GLR and FUF have a similar mean AP, while their seed variation and optimization trajectories differ. An increasingly flexible routing parameterization is therefore not automatically a better choice. Figure 13 places the routing variants beside the backbone-enhanced references and the two supporting detector families, making both the positive and negative outcomes visible.

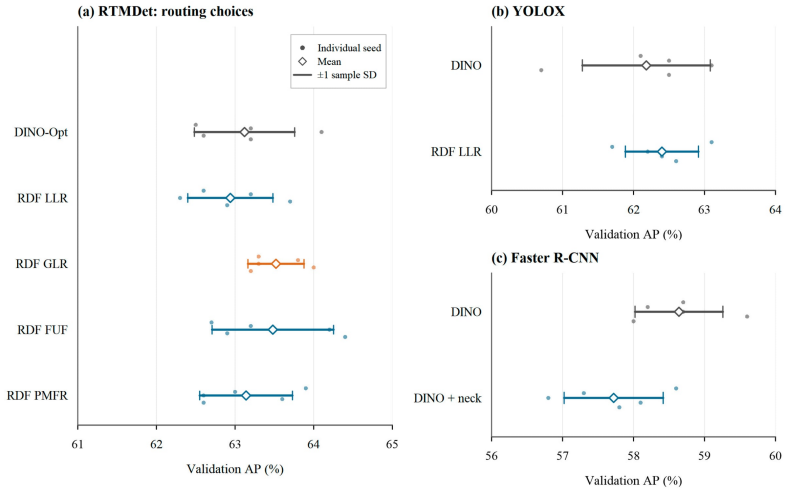


Figure 13. Five-run AP results across architectures and routing choices. All configurations use the DINO backbone. (**a**) RTMDet with optimized-backbone and RDF routing; (**b**) YOLOX with DINO and RDF-LLR; (**c**) Faster R-CNN with DINO and DINO-Neck.

Table 12 reports the results on the screened test subset. The mean AP is 0.5431 for RTM-Base, 0.6460 for RTM-DINO-Opt, 0.6525 for GLR, and 0.6574 for FUF. The GLR and FUF configurations exceed RTM-Base by 10.94 and 11.43 AP points, respectively. Relative to the backbone-enhanced RTM-DINO-Opt reference, GLR and FUF further improve the mean AP by 0.65 and 1.14 points, respectively.

Table 12. Evaluation on the screened test subset.

Configuration	Runs	AP	AP50
RTM-Base	5	0.5431 ± 0.0169	0.8420 ± 0.0176
RTM-DINO-Opt	5	0.6460 ± 0.0123	0.9294 ± 0.0061
RTM-RDF-GLR-DINO (RDF-Det)	5	0.6525 ± 0.0143	0.9338 ± 0.0125
RTM-RDF-FUF-DINO	5	0.6574 ± 0.0060	0.9360 ± 0.0055

Note: Values are means ± sample standard deviations over five runs on the same 110 images and 111 boxes.

FUF has a mean AP of 0.49 points above GLR on this subset, with AP standard deviations of 0.60 and 1.43 points, respectively. The ordering therefore differs from the main validation summaries, while both evaluations support examining simple fusion as a competitive choice. Together, these results motivate considering fusion complexity, accuracy, and run-to-run variation jointly when choosing a routing design.

The current study provides a controlled comparison of structural and fusion choices within TFR, and the supplementary evaluation on the screened test subset further examines these choices within the same collection. Regarding external ground-view validation, direct four-class AP evaluation requires vehicle-level bounding boxes and labels compatible with the four TFR vehicle/fire-state categories on imagery disjoint from the TFR sources; however, existing public resources target different tasks—the Car Accident Severity resource provides accident-severity annotations and already contributes images to TFR [46], whereas the vehicle-fire segmentation work discussed in Section 2.1 annotates flame and smoke regions [28].

Beyond ground-view datasets, extending the framework to broader remote sensing platforms also faces domain-shift constraints. Ground, UAV, and satellite observations differ in viewing geometry, object scale, and spectral inputs; cross-platform and cross-incident validation will therefore require source-traceable, event-disjoint vehicle-fire collections with matched vehicle-associated boxes and fire-state labels.

Future research will therefore proceed in several directions. The first is to expand the scale and environmental diversity of ground-view transportation incident datasets, including more night scenes, reflective surfaces, smoke densities, camera viewpoints, and hard-negative non-fire samples. The second is to deploy and profile lighter RDF-Neck variants on low-power edge-computing backends tailored for roadside units and vehicle-mounted platforms, potentially with model compression and deployment-oriented acceleration, such as TensorRT. The third is to incorporate temporal information from video streams so that fire-risk evidence can be accumulated across frames rather than inferred from isolated images. A complementary direction is UAV-assisted urban traffic monitoring using oblique RGB views, with matched vehicle/fire annotations to study cross-view transfer of the dual-path representation. Finally, deep feature-level visual results will be introduced to examine gradient flow and activation patterns within the dual-path fusion mechanism, while retaining the topology–routing ablation design established in this study.

6. Conclusions

This study presents RDF-Det, a vehicle fire-state detector for ground-view RGB imagery, built around RDF-Neck, a dual-path feature-fusion neck that aggregates top–down semantic context and bidirectional spatial cues in parallel and combines them at corresponding pyramid levels. Within this fixed dual-path topology, we compared level-wise learned, globally shared, fixed-uniform, and frozen routing and instantiated the neck in RTMDet, YOLOX, and Faster R-CNN. Three findings emerge. First, the main RDF-Det configuration (RTM-RDF-GLR-DINO) achieves 0.6352 AP and 0.9290 AP50 over five runs, a system-level gain of 7.24 AP points over RTM-Base, and exceeds the standard YOLO26 detector by 1.17 AP and 2.19 AP50 points with lower cross-seed variation; under the same CSPNeXt backbone, replacing the original neck with RDF-Neck improves AP and AP50 by 2.12 and 3.04 points, respectively. Second, globally shared and fixed-uniform fusion are as competitive as level-wise learned routing (0.6352 vs. 0.6348 AP), indicating that the benefit stems mainly from the parallel dual-path structure rather than from increasingly flexible routing. Third, the effect of the neck depends on the detector architecture: it yields a small gain in the one-stage YOLOX and a decrease in the two-stage Faster R-CNN, where the pyramid also serves proposal generation.

These results are obtained on the TFR collection with repeated runs and Holm-corrected paired tests, and a supplementary evaluation on a screened test subset shows the same tendency (0.6525 and 0.6574 AP for GLR and FUF). Future work will validate RDF-Det on source-traceable, event-disjoint incident data, extend it to video-level temporal reasoning, and explore lightweight deployment on roadside and UAV-based monitoring platforms.

Author Contributions: Conceptualization, W.L. and X.Z.; methodology, W.L., B.Z., and X.Z.; software, W.L. and B.Z.; validation, W.L.; formal analysis, W.L., X.Z., and Z.L.; visualization, W.L. and Z.L.; investigation, W.L.; resources, X.Z., Z.C., and N.C.; data curation, W.L. and X.Z.; writing—original draft preparation, W.L.; writing—review and editing, W.L., X.Z., B.Z., Z.L., Z.C., and N.C.; supervision, X.Z., Z.C., and N.C.; project administration, X.Z.; funding acquisition, X.Z. and N.C. All authors have read and agreed to the published version of the manuscript.

Funding: This work was supported by the Key Research and Development Program of Hubei Province (Grant No. 2024BCB101).

Data Availability Statement: The data presented in this study, together with the trained model weights, are available on request from the corresponding author. The source code is publicly available at https://github.com/luckyLuWen/RobustDF-Detv0 (accessed on 29 September 2026).

Conflicts of Interest: The authors declare no conflicts of interest.

Appendix A

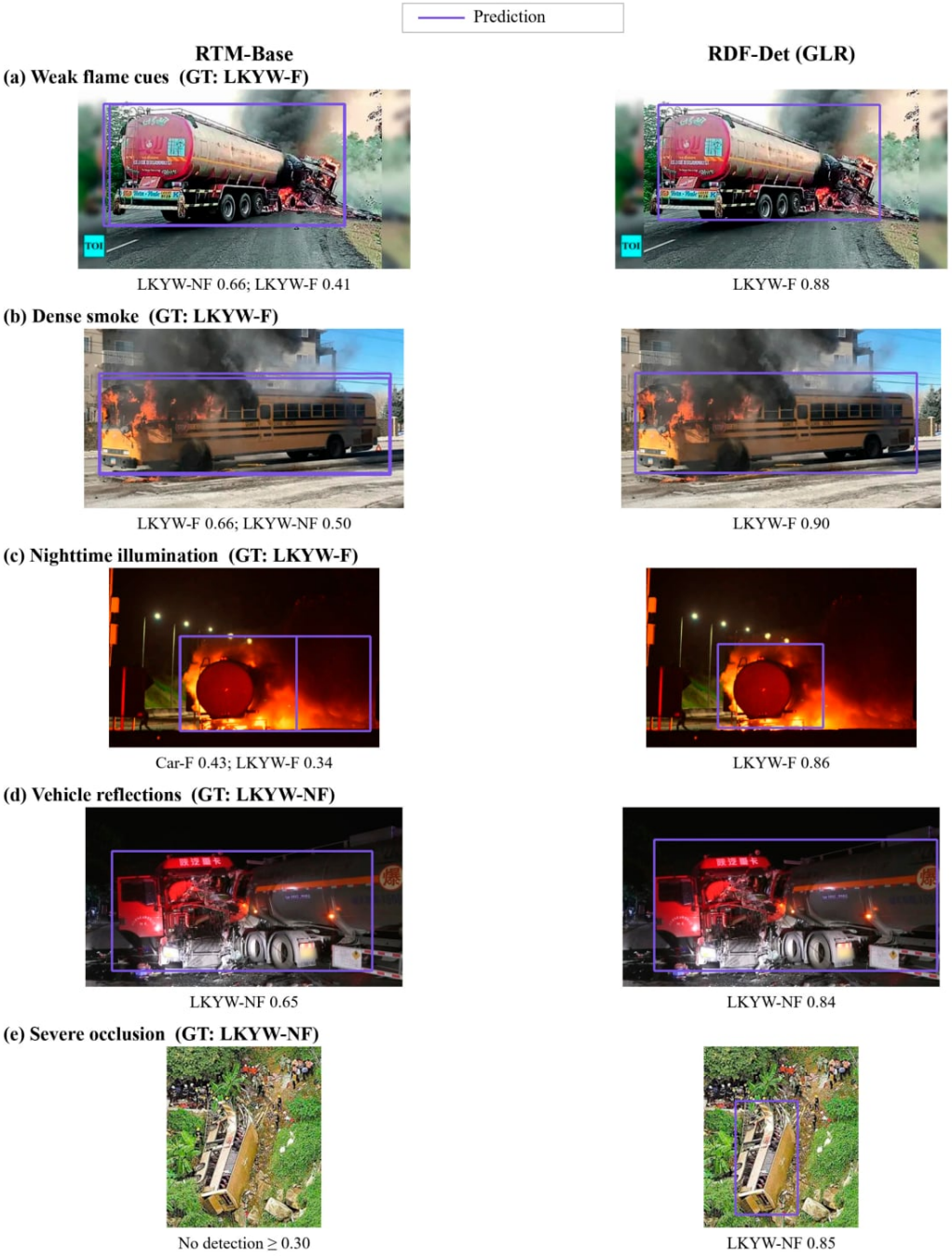


Figure A1. Qualitative comparison of RTM-Base and RDF-Det (GLR) under (**a**) weak flame cues, (**b**) dense smoke, (**c**) nighttime illumination, (**d**) vehicle reflections, and (**e**) severe occlusion. Detections are shown at a score threshold of 0.30.

References

1. Shi, P.; Ye, T.; Wang, Y.; Zhou, T.; Xu, W.; Du, J.; Wang, J.; Li, N.; Huang, C.; Liu, L.; et al. Disaster Risk Science: A Geographical Perspective and a Research Framework. *Int. J. Disaster Risk Sci.* **2020**, *11*, 426–440. [CrossRef]
2. Aitsi-Selmi, A.; Murray, V.; Wannous, C.; Dickinson, C.; Johnston, D.; Kawasaki, A.; Stevance, A.-S.; Yeung, T. Reflections on a Science and Technology Agenda for 21st Century Disaster Risk Reduction. *Int. J. Disaster Risk Sci.* **2016**, *7*, 1–29. [CrossRef]
3. Baudoin, M.-A.; Henly-Shepard, S.; Fernando, N.; Sitati, A.; Zommers, Z. From Top-Down to "Community-Centric" Approaches to Early Warning Systems: Exploring Pathways to Improve Disaster Risk Reduction Through Community Participation. *Int. J. Disaster Risk Sci.* **2016**, *7*, 163–174. [CrossRef]

4. Sivaraman, S.; Trivedi, M.M. Looking at Vehicles on the Road: A Survey of Vision-Based Vehicle Detection, Tracking, and Behavior Analysis. *IEEE Trans. Intell. Transp. Syst.* **2013**, *14*, 1773–1795. [CrossRef]
5. Datondji, S.R.E.; Dupuis, Y.; Subirats, P.; Vasseur, P. A Survey of Vision-Based Traffic Monitoring of Road Intersections. *IEEE Trans. Intell. Transp. Syst.* **2016**, *17*, 2681–2698. [CrossRef]
6. Boukerche, A.; Hou, Z. Object Detection Using Deep Learning Methods in Traffic Scenarios. *ACM Comput. Surv. CSUR* **2021**, *54*, 1–35. [CrossRef]
7. Toumazis, I.; Kwon, C. Routing Hazardous Materials on Time-Dependent Networks Using Conditional Value-at-Risk. *Transp. Res. Part C Emerg. Technol.* **2013**, *37*, 73–92. [CrossRef]
8. Guo, J.; Luo, C. Risk Assessment of Hazardous Materials Transportation: A Review of Research Progress in the Last Thirty Years. *J. Traffic Transp. Eng. Engl. Ed.* **2022**, *9*, 571–590. [CrossRef]
9. Ren, C.; Yang, M. Risk Assessment of Hazmat Road Transportation Accidents before, during, and after the Accident Using Bayesian Network. *Process Saf. Environ. Prot.* **2024**, *190*, 760–779. [CrossRef]
10. U.S. Fire Administration. Highway Vehicle Fires (2014–2016). In *Topical Fire Report Series*. USFA: Emmitsburg, MD, USA, 2018; Volume 19. Available online: https://www.usfa.fema.gov/downloads/pdf/statistics/v19i2.pdf (accessed on 27 October 2026).
11. Muhammad, K.; Ahmad, J.; Lv, Z.; Bellavista, P.; Yang, P.; Baik, S.W. Efficient Deep CNN-Based Fire Detection and Localization in Video Surveillance Applications. *IEEE Trans. Syst. Man Cybern. Syst.* **2019**, *49*, 1419–1434. [CrossRef]
12. Gragnaniello, D.; Greco, A.; Sansone, C.; Vento, B. Fire and Smoke Detection from Videos: A Literature Review under a Novel Taxonomy. *Expert Syst. Appl.* **2024**, *255*, 124783. [CrossRef]
13. Foggia, P.; Saggese, A.; Vento, M. Real-Time Fire Detection for Video-Surveillance Applications Using a Combination of Experts Based on Color, Shape, and Motion. *IEEE Trans. Circuits Syst. Video Technol.* **2015**, *25*, 1545–1556. [CrossRef]
14. Ko, B.C.; Ham, S.J.; Nam, J.Y. Modeling and Formalization of Fuzzy Finite Automata for Detection of Irregular Fire Flames. *IEEE Trans. Circuits Syst. Video Technol.* **2011**, *21*, 1903–1912. [CrossRef]
15. Li, X.; Chen, Z.; Wu, Q.M.J.; Liu, C. 3D Parallel Fully Convolutional Networks for Real-Time Video Wildfire Smoke Detection. *IEEE Trans. Circuits Syst. Video Technol.* **2020**, *30*, 89–103. [CrossRef]
16. Gu, K.; Xia, Z.; Qiao, J.; Lin, W. Deep Dual-Channel Neural Network for Image-Based Smoke Detection. *IEEE Trans. Multimed.* **2020**, *22*, 311–323. [CrossRef]
17. Majid, S.; Alenezi, F.; Masood, S.; Ahmad, M.; Gündüz, E.S.; Polat, K. Attention Based CNN Model for Fire Detection and Localization in Real-World Images. *Expert Syst. Appl.* **2022**, *189*, 116114. [CrossRef]
18. Khan, Z.A.; Hussain, T.; Ullah, F.U.M.; Gupta, S.K.; Lee, M.Y.; Baik, S.W. Randomly Initialized CNN with Densely Connected Stacked Autoencoder for Efficient Fire Detection. *Eng. Appl. Artif. Intell.* **2022**, *116*, 105403. [CrossRef]
19. Liu, L.; Ouyang, W.; Wang, X.; Fieguth, P.; Chen, J.; Liu, X.; Pietikäinen, M. Deep Learning for Generic Object Detection: A Survey. *Int. J. Comput. Vis.* **2020**, *128*, 261–318. [CrossRef]
20. Zou, Z.; Chen, K.; Shi, Z.; Guo, Y.; Ye, J. Object Detection in 20 Years: A Survey. *Proc. IEEE* **2023**, *111*, 257–276. [CrossRef]
21. Ren, S.; He, K.; Girshick, R.; Sun, J. Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks. *IEEE Trans. Pattern Anal. Mach. Intell.* **2017**, *39*, 1137–1149. [CrossRef]
22. Fang, J.; Qiao, J.; Xue, J.; Li, Z. Vision-Based Traffic Accident Detection and Anticipation: A Survey. *IEEE Trans. Circuits Syst. Video Technol.* **2024**, *34*, 1983–1999. [CrossRef]
23. Chen, Z.; Ji, H.; Zhang, Y.; Zhu, Z.; Li, Y. High-Resolution Feature Pyramid Network for Small Object Detection on Drone View. *IEEE Trans. Circuits Syst. Video Technol.* **2024**, *34*, 475–489. [CrossRef]
24. Dror, R.; Baumer, G.; Shlomov, S.; Reichart, R. The Hitchhiker's Guide to Testing Statistical Significance in Natural Language Processing. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*; Gurevych, I., Miyao, Y., Eds.; Association for Computational Linguistics: Melbourne, Australia, 2018; pp. 1383–1392.
25. Henderson, P.; Islam, R.; Bachman, P.; Pineau, J.; Precup, D.; Meger, D. Deep Reinforcement Learning That Matters. *Proc. AAAI Conf. Artif. Intell.* **2018**, *32*, 3207–3214. [CrossRef]
26. Bouthillier, X.; Delaunay, P.; Bronzi, M.; Trofimov, A.; Nichyporuk, B.; Szeto, J.; Sepah, N.; Raff, E.; Madan, K.; Voleti, V.; et al. Accounting for Variance in Machine Learning Benchmarks. In Proceedings of the MLSys 2021, Virtual, 5–9 April 2021.
27. Zhang, G.; Wang, M.; Liu, K. Forest Fire Susceptibility Modeling Using a Convolutional Neural Network for Yunnan Province of China. *Int. J. Disaster Risk Sci.* **2019**, *10*, 386–403. [CrossRef]
28. Kwon, H.; Choi, S.; Woo, W.; Jung, H. Evaluating Segmentation-Based Deep Learning Models for Real-Time Electric Vehicle Fire Detection. *Fire* **2025**, *8*, 66. [CrossRef]
29. Zhou, W.; Wang, C.; Xia, J.; Qian, Z.; Wu, Y. Monitoring-Based Traffic Participant Detection in Urban Mixed Traffic: A Novel Dataset and A Tailored Detector. *IEEE Trans. Intell. Transp. Syst.* **2024**, *25*, 189–202. [CrossRef]
30. Zhou, W.; Wang, C.; Ge, Y.; Wen, L.; Zhan, Y. All-Day Vehicle Detection From Surveillance Videos Based on Illumination-Adjustable Generative Adversarial Network. *IEEE Trans. Intell. Transp. Syst.* **2024**, *25*, 3326–3340. [CrossRef]

31. Vijay, T.K.; Dogra, D.P.; Choi, H.; Nam, G.; Kim, I.-J. Detection of Road Accidents Using Synthetically Generated Multi-Perspective Accident Videos. *IEEE Trans. Intell. Transp. Syst.* **2023**, *24*, 1926–1935. [CrossRef]
32. Niu, S.; Ukkusuri, S.V. Risk Assessment of Commercial Dangerous -Goods Truck Drivers Using Geo-Location Data: A Case Study in China. *Accid. Anal. Prev.* **2020**, *137*, 105427. [CrossRef]
33. Hong, J.; Tamakloe, R.; Park, D. Application of Association Rules Mining Algorithm for Hazardous Materials Transportation Crashes on Expressway. *Accid. Anal. Prev.* **2020**, *142*, 105497. [CrossRef]
34. Lin, T.-Y.; Goyal, P.; Girshick, R.; He, K.; Dollár, P. Focal Loss for Dense Object Detection. *IEEE Trans. Pattern Anal. Mach. Intell.* **2020**, *42*, 318–327. [CrossRef]
35. Law, H.; Deng, J. CornerNet: Detecting Objects as Paired Keypoints. *Int. J. Comput. Vis.* **2020**, *128*, 642–656. [CrossRef]
36. Lin, T.-Y.; Maire, M.; Belongie, S.; Hays, J.; Perona, P.; Ramanan, D.; Dollár, P.; Zitnick, C.L. Microsoft COCO: Common Objects in Context. In *Proceedings of the Computer Vision—ECCV 2014*; Fleet, D., Pajdla, T., Schiele, B., Tuytelaars, T., Eds.; Springer International Publishing: Cham, Switzerland, 2014; pp. 740–755.
37. Lin, T.-Y.; Dollár, P.; Girshick, R.; He, K.; Hariharan, B.; Belongie, S. Feature Pyramid Networks for Object Detection. In Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, USA, 21–26 July 2017; pp. 936–944.
38. Tan, M.; Pang, R.; Le, Q.V. EfficientDet: Scalable and Efficient Object Detection. In Proceedings of the 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Virtual, 14–19 June 2020; pp. 10778–10787.
39. Holm, S. A Simple Sequentially Rejective Multiple Test Procedure. *Scand. J. Stat.* **1979**, *6*, 65–70.
40. Cohen, J. *Statistical Power Analysis for the Behavioral Sciences*, 2nd ed.; Routledge: New York, NY, USA, 2013; ISBN 978-0-203-77158-7.
41. Dietterich, T.G. Approximate Statistical Tests for Comparing Supervised Classification Learning Algorithms. *Neural Comput.* **1998**, *10*, 1895–1923. [CrossRef]
42. Lyu, C.; Zhang, W.; Huang, H.; Zhou, Y.; Wang, Y.; Liu, Y.; Zhang, S.; Chen, K. RTMDet: An Empirical Study of Designing Real-Time Object Detectors. *arXiv* **2022**, arXiv:2212.07784.
43. Ge, Z.; Liu, S.; Wang, F.; Li, Z.; Sun, J. YOLOX: Exceeding YOLO Series in 2021. *arXiv* **2021**, arXiv:2107.08430.
44. Siméoni, O.; Vo, H.V.; Seitzer, M.; Baldassarre, F.; Oquab, M.; Jose, C.; Khalidov, V.; Szafraniec, M.; Yi, S.; Ramamonjisoa, M.; et al. DINOv3 2025. *arXiv* **2025**, arXiv:2508.10104.
45. Liu, Z.; Mao, H.; Wu, C.-Y.; Feichtenhofer, C.; Darrell, T.; Xie, S. A ConvNet for the 2020s. In Proceedings of the 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Virtual, 19–24 June 2022; pp. 11966–11976.
46. Alotaibi, S. Car Accident Severity Dataset. Roboflow Universe. 2024. Available online: https://universe.roboflow.com/shahad-alotaibi-ujdrb/car-accident-severity (accessed on 31 October 2026).
47. Ultralytics Ultralytics YOLO26. Available online: https://docs.ultralytics.com/models/yolo26/ (accessed on 27 October 2026).

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.