TF-DETR: TASK-FACTOR REPRESENTATION FOR RGB-X OBJECT DETECTION

Anonymous authors
Paper under double-blind review

ABSTRACT

Existing RGB-X detectors commonly maintain substantial semantic representations for both modalities, which can introduce representational and computational redundancy when RGB already provides a strong primary semantic stream. Meanwhile, the task-relevant contribution of the auxiliary modality X varies across sensing modalities and detection objectives. Motivated by these observations, we propose TF-DETR (Task-Factor DETR), which preserves RGB as the primary semantic stream while representing X through Task Factors that combine explicit detection-oriented states with flexible sensor-dependent latent evidence, rather than constructing a second full semantic stream. To exploit these factors effectively, TF-DETR establishes cross-modal correspondence through global-anchored multi-scale task alignment, and incorporates aligned auxiliary evidence into the RGB encoder through reliability-aware query guidance and bounded low-rank feature refinement. This asymmetric design allows X to guide and refine RGB representations through a compact auxiliary pathway, thereby avoiding redundant semantic encoding and reducing computational overhead. Experiments on six RGB-X benchmarks spanning thermal, event, and depth sensing demonstrate strong detection performance and computational efficiency with a lightweight auxiliary pathway. Code will be released publicly.

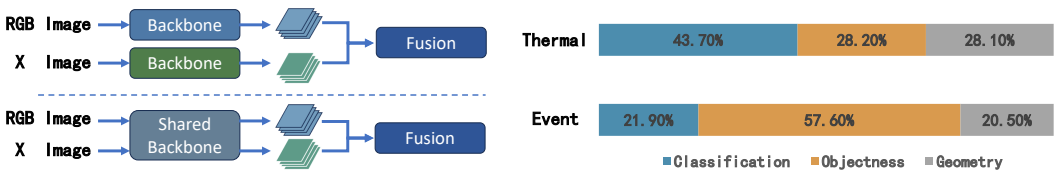
1 INTRODUCTION

Object detection in unconstrained environments increasingly benefits from sensing modalities beyond conventional RGB cameras. Thermal imagers, event cameras, and depth sensors capture physical signals that complement visible appearance, providing additional observations under low illumination, rapid motion, occlusion, and severe appearance degradation (Dong et al., 2025; Gehrig & Scaramuzza, 2024; Tan et al., 2025). Existing RGB-X detectors commonly construct feature representations for both modalities and then focus on how these representations should be aligned, interacted with, and fused. Representative designs either maintain modality-specific parallel backbones or employ a shared backbone before multimodal fusion, as illustrated in Fig. 1a (Cheng et al., 2024; Cho et al., 2025; Shrivastava & Owens, 2025; Shen et al., 2024).

However, maintaining a full semantic representation for each modality can be unnecessarily redundant when RGB already provides a strong primary detection representation. Recent RGB-IR studies have observed substantial homogeneity between features extracted by dual backbones, suggesting that repeatedly encoding similar high-level semantics across modalities may introduce considerable representational and computational redundancy (Yang et al., 2026). This motivates a shift in perspective: rather than asking how two full feature streams should be fused, we ask what task-relevant information the auxiliary modality should contribute to an already strong RGB detector. Different auxiliary sensors capture different physical signals, and their incremental utility need not be distributed uniformly across classification, objectness estimation, and geometry prediction. Characterizing this conditional task utility can therefore provide a more direct basis for deciding how X should be represented (Yang et al., 2024b).

We study this question through a controlled conditional task utility analysis on RGB-Event detection using PKU-DAVIS-SOD and RGB-Thermal detection using M3FD. RGB-only and X-only encoders are trained independently and then frozen, after which lightweight probes measure the incremental benefit of adding the paired auxiliary representation for classification, objectness, and geometry

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.



(a) Feature-centric RGB-X paradigms. (b) Conditional task utility of X.

Figure 1: **Motivation for task-factor representation.** **(a)** Representative feature-centric RGB-X paradigms with parallel or shared backbones. **(b)** Given a strong RGB representation, RGB-Thermal shows a stronger classification contribution, whereas RGB-Event exhibits an objectness-dominant utility profile. Task-wise gains are normalized within each sensing setting.

prediction. As shown in Fig. 1b, the resulting utility profiles differ substantially across sensing settings: Event exhibits an objectness-dominant profile, whereas the Thermal setting shows the largest relative gain for classification.

These observations suggest that the useful contribution of X is both task-dependent and sensor-dependent (Madaan et al., 2026). A common auxiliary representation should therefore expose task variables with stable detection semantics while retaining enough flexibility to preserve evidence whose role varies across sensors (Dufumier et al., 2025; Wang et al., 2024). This motivates our **Task Factor** representation, which combines explicit detection-oriented states with learnable latent factors rather than prescribing a fixed semantic role for all auxiliary information.

Building on these considerations, we propose TF-DETR (**T**ask-**F**actor DETR), which retains RGB as the primary semantic stream and represents the auxiliary modality through compact Task Factors combining task-grounded states with sensor-adaptive latent evidence. To exploit these Task Factors effectively, we design Global-Anchored Multi-Scale Task Alignment to establish coarse-to-fine cross-modal correspondence in task space. We then develop a reliability-aware task-guided encoder that incorporates the aligned auxiliary evidence through query guidance and bounded low-rank feature refinement. This design provides a common task-level interface for heterogeneous sensors without maintaining a second full semantic stream.

Experiments on RGB-Thermal and RGB-Event detection benchmarks demonstrate state-of-the-art performance with a lightweight auxiliary pathway. Our contributions are summarized as follows:

1. We propose **TF-DETR** and introduce a **Task-Factor representation** for RGB-X object detection. It combines explicit detection-oriented states with flexible sensor-adaptive latent factors, providing a common task-level interface while retaining sensor-specific evidence.

2. We introduce **Global-Anchored Multi-Scale Task Alignment**, which estimates cross-modal correspondence from task-grounded and positional cues, refines residual discrepancies across scales, and transfers the complete auxiliary task state.

3. We develop a **reliability-aware task-guided encoder** that incorporates aligned Task Factors through reliability-gated query guidance and bounded low-rank feature refinement, explicitly controlling when and how strongly auxiliary evidence influences the RGB representation.

2 Related Work

2.1 RGB-X Object Detection

RGB-X object detection exploits heterogeneous sensing signals to complement conventional RGB observations. For RGB-T detection, recent studies mainly improve cross-modal representation and interaction under spectral heterogeneity. CALNet explicitly addresses semantic conflicts between RGB and infrared features (He et al., 2023), while ICAFusion employs iterative cross-attention to aggregate complementary multispectral features (Shen et al., 2024). C^2Former further combines feature calibration and complementary interaction through inter-modality cross-attention (Yuan & Wei, 2024), and Scene-Specific Fusion explores a lightweight alternative by adapting pretrained unimodal detectors with compact fusion modules (Deevi et al., 2024).

RGB-Event detection has developed along a distinct direction to exploit the asynchronous and high-temporal-resolution nature of event sensing. SODFormer introduces spatiotemporal Transformer modeling and asynchronous event-frame fusion (Li et al., 2023a); HFRN performs hierarchical cross-modal feature refinement (Cao et al., 2024); and ACGR constructs an asynchronous collaborative graph representation rather than relying on conventional frame-synchronized two-stream processing (Li et al., 2025). More recently, Beyond Duality explicitly separates shared and modality-private representations for RGB-Event detection (Wang et al., 2026). Depth provides another complementary sensing case, with classical RGB-D detection exploring both modality-correlated and modality-specific representations for objectness and recognition (Xu et al., 2017). Despite their different treatment of thermal, event, and depth signals, these methods predominantly formulate multimodal detection around the construction and interaction of modality features. Our work instead begins by examining the conditional task utility of X given an already strong RGB representation, and uses this evidence to design a task-grounded, sensor-adaptive auxiliary interface rather than assigning heterogeneous sensors a fixed representational role.

2.2 Auxiliary-Modality Representation and Modality Roles

Recent multimodal research has increasingly investigated how heterogeneous auxiliary modalities should be represented within a unified model. TokenFusion selectively transfers informative cross-modal tokens while largely preserving the underlying single-modal Transformer (Wang et al., 2022). CMX develops a unified rectification-and-fusion framework for RGB paired with depth, thermal, polarization, event, and LiDAR modalities (Zhang et al., 2023a), while CMNeXt further supports an arbitrary number of supplementary modalities through lightweight information extraction before fusion with RGB (Zhang et al., 2023b). These methods substantially improve multimodal generality, but mainly focus on common mechanisms for constructing and interacting multimodal features, rather than explicitly structuring the auxiliary representation according to downstream task utility.

Another line of work considers the relative roles and commonality of different modalities. ASPnet and ShaSpec decompose multimodal representations into shared and modality-specific components (van Amsterdam et al., 2023; Wang et al., 2023). OneTracker adopts a primary–auxiliary formulation by adapting a pretrained RGB foundation tracker to additional modalities through lightweight prompting (Hong et al., 2024), whereas Un-Track maps depth, thermal, and event inputs into a common latent representation for unified RGB-X tracking (Wu et al., 2024). Any2Seg similarly pursues modality-agnostic representations for arbitrary modality combinations (Zheng et al., 2024). In contrast, Keep the Balance argues that excessive RGB dominance may underutilize X, and therefore develops parameter-efficient symmetric adaptation for both streams (Cai et al., 2025). These studies indicate that modality roles remain an open design choice rather than a settled architectural prior. Our formulation differs from both modality-invariant and shared/private representation learning: motivated by sensor-dependent conditional utility, we explicitly ground detection-relevant variables that admit stable task interpretations, while preserving free residual capacity for complementary evidence whose role may vary across sensors.

3 Methods

3.1 Overview

We propose TF-DETR, which preserves RGB as the primary semantic stream and represents the auxiliary modality through compact Task Factors. As shown in Fig. 2, TF-DETR comprises three components: **Task-Factor representation** for constructing task-space states, **Global-Anchored Multi-Scale Task Alignment** for cross-modal correspondence, and a **Reliability-Aware Task-Guided Encoder** for incorporating aligned auxiliary evidence into RGB features.

3.2 Task-Factor representation

Given an RGB image R and an auxiliary observation X, TF-DETR assigns the two modalities different representational roles. The RGB stream retains its multi-scale semantic features $\{F_R^l\}_{l=3}^5$ as the primary representation for detection, whereas X is processed by a lightweight **Re-parameterizable Auxiliary Task Encoder (RepATE)**, and mapped into task-oriented states. To establish a task-space reference for cross-modal alignment, a lightweight RGB task probe projects the RGB features into the same factorized space:

$$T_X^l = \Phi_X\left(\mathcal{E}_X^l(X)\right), \qquad T_R^l = \Phi_R\left(\psi_R(F_R^l)\right), \quad l \in 3,4,5. \tag{1}$$

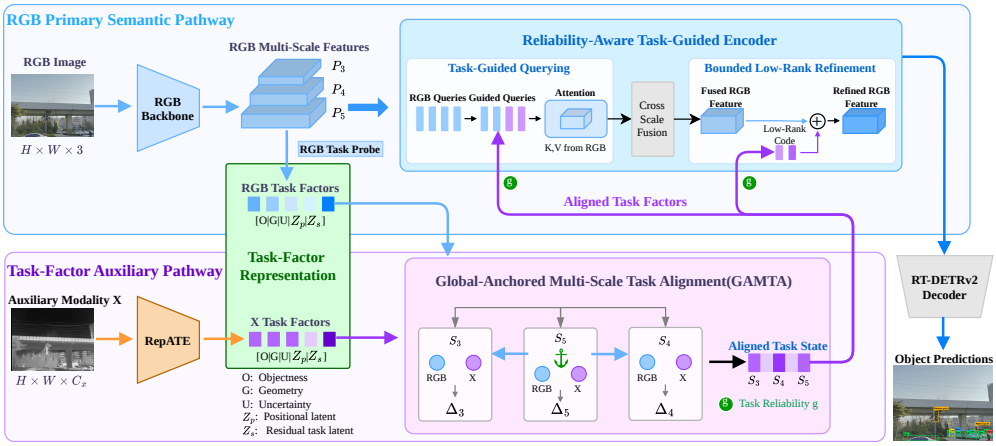


Figure 2: Overview of TF-DETR, which models auxiliary observations with lightweight Task Factors for task-space alignment and reliability-aware RGB interaction.

Both task states share the same factorized form,

$$T_m^l = \underbrace{[O_m^l, G_m^l, U_m^l]}_{\text{task-grounded states}} \oplus \underbrace{Z_{p,m}^l}_{\text{correspondence latent}} \oplus \underbrace{Z_{s,m}^l}_{\text{residual task latent}}, \qquad m \in \{R, X\}. \tag{2}$$

The factorization combines explicit task states with learnable latent capacity. O_m^l captures class-agnostic evidence of object presence, which we term *objectness*. $G_m^l = (\Delta x, \Delta y, \log \bar{w}, \log \bar{h})$ describes local object geometry. U_m^l parameterizes the corresponding geometric uncertainty in log-scale form. These variables provide detection-oriented anchors with stable interpretations across sensing modalities.

The remaining factors are learned rather than assigned fixed semantics. $Z_{p,m}^l$ serves as a positional latent for cross-modal correspondence and is not explicitly parameterized as optical flow or displacement. $Z_{s,m}^l$ is a residual task latent that retains useful information not captured by the explicit factors. In the auxiliary stream, $Z_{s,X}^l$ can therefore adapt to sensor-dependent evidence, such as appearance, motion, or structural cues, without imposing shared semantics across sensors.

Importantly, the RGB task state T_R^l is used only as a task-space reference for subsequent correspondence and reliability reasoning. It does not replace or compress the original RGB feature F_R^l, which continues through the primary semantic pathway. In this way, Task Factors provide a common task-level interface for cross-modal interaction while retaining the flexibility needed to preserve sensor-specific auxiliary evidence. See Appendix C.3 for architecture details.

3.3 Global-Anchored Multi-Scale Task Alignment

Even when paired RGB-X observations are approximately registered in image space, heterogeneous sensing and modality-specific encoding can produce residual discrepancies between their task responses. Additional inconsistency may arise from calibration or synchronization errors in practical settings. Directly comparing T_R^l and T_X^l would mix task discrepancy with spatial correspondence. We introduce **Global-Anchored Multi-Scale Task Alignment (GAMTA)** to establish task-space correspondence from a shared global anchor followed by scale-specific residual refinement.

Global task anchor.

We estimate a shared coarse alignment from the lowest-resolution task level (i.e., P5 in Fig. 2), whose larger receptive field provides contextual support for cross-modal correspondence. Using the alignment cues $A_m^l = [O_m^l, G_m^l, Z_{p,m}^l]$, the global transform and multi-scale warping are defined as

$$\Theta^5 = \mathcal{G}^5(A_R^5, A_X^5), \qquad \bar{T}_X^l = \mathcal{W}(T_X^l; \Theta^5), \quad l \in \{3, 4, 5\}. \tag{3}$$

Here, Θ^5 parameterizes a low-degree-of-freedom similarity sampling transform from RGB task coordinates to auxiliary sampling coordinates. The same transform is shared across scales, providing a common spatial anchor before local refinement.

Multi-scale residual transport. The shared global transform captures coarse correspondence but may leave scale-dependent residual discrepancies. At each level, GAMTA therefore predicts a small set of bounded residual offsets $\{\delta_k^l\}_{k=1}^K$ and their correspondence weights. The aligned auxiliary state is obtained by

$$\widetilde{T}_X^l(p) = \sum_{k=1}^{K} \pi_k^l(p)\, \bar{T}_X^l\left(p + \delta_k^l(p)\right). \tag{4}$$

The candidate search range decreases toward coarser feature levels. An additional rejection state $\pi_\varnothing^l$ allows the transport to reduce auxiliary contribution when no valid sampling support is available; correspondingly, the non-rejection probability mass attenuates the transported state.

Alignment confidence. Using the rejection probability $\pi_\varnothing^l$ introduced above, we normalize the non-rejection candidate probabilities as

$$\hat{\pi}_k^l = \frac{\pi_k^l}{1 - \pi_\varnothing^l + \epsilon}, \tag{5}$$

and define the alignment confidence as

$$C_{\text{align}}^l = (1 - \pi_\varnothing^l)\left(1 - H_{\text{norm}}(\hat{\boldsymbol{\pi}}^l)\right). \tag{6}$$

Thus, alignment confidence reflects both the availability of valid auxiliary support and the concentration of the candidate correspondence distribution. GAMTA thus uses task-grounded and positional cues to determine cross-modal correspondence, while transporting the complete auxiliary Task-Factor representation.

Task reliability. Spatial correspondence does not by itself determine whether the aligned auxiliary evidence should strongly influence the RGB representation. We use a lightweight task-state estimator to assess reliability from alignment confidence, RGB-X objectness and geometry discrepancies, and the predicted geometric uncertainties U_R^l and $\widetilde{U}_X^l$. Let D_O^l and D_G^l denote the pre-gating objectness and geometry discrepancies between the RGB and aligned auxiliary task states, respectively. The predicted reliability is further constrained by spatial validity and modality presence:

$$\hat{g}^l = \sigma\Big(\mathcal{S}^l\left(C_{\text{align}}^l,\, D_O^l,\, D_G^l,\, U_R^l,\, \widetilde{U}_X^l\right)\Big), \qquad g^l = \hat{g}^l C_{\text{align}}^l V_X^l M_X. \tag{7}$$

Here, V_X^l is the aligned spatial-validity mask and M_X indicates auxiliary-modality availability. The resulting g^l conditions the two encoder interaction paths.

3.4 Reliability-Aware Task-Guided Encoder

After cross-modal alignment, TF-DETR uses the aligned auxiliary Task Factors to guide the RGB encoder while preserving RGB as the primary semantic stream. The interaction is applied in two stages. At the lowest-resolution level, aligned positional and geometric cues modulate attention queries. After cross-scale fusion, residual task evidence provides a bounded low-rank correction to the RGB features.

Given the aligned auxiliary state $\widetilde{T}_X^l$ and the RGB task reference T_R^l, we construct reliability-conditioned task residuals using the estimated state g^l. For geometry, we additionally account for the relative uncertainty of the two modalities. Let $b_R^l = \text{Mean}_c[\exp(U_R^l)]$ and $\widetilde{b}_X^l = \text{Mean}_c[\exp(\widetilde{U}_X^l)]$. We define

$$\Delta Z_s^l = g^l\left(\widetilde{Z}_{s,X}^l - Z_{s,R}^l\right), \qquad \Delta G^l = g^l \kappa^l\left(\widetilde{G}_X^l - G_R^l\right), \quad \kappa^l = \frac{b_R^l}{b_R^l + \widetilde{b}_X^l + \epsilon}. \tag{8}$$

Here, ΔZ_s^l is treated as a learned residual task discrepancy rather than an explicitly defined channel-wise semantic difference. **Query-side task guidance.** At the lowest-resolution encoder level, aligned positional and geometric cues together with alignment confidence and task reliability form a task context C_Q. The resulting task-conditioned residual modulates only the attention query:

$$Q = Q_R + \alpha g^5 \odot \Psi_Q\left(\left[\widetilde{Z}_{p,X}^5,\, \widetilde{G}_X^5,\, \Delta G^5,\, C_{\text{align}}^5,\, g^5\right]\right). \tag{9}$$

Thus, auxiliary task evidence guides *where* RGB semantic information is retrieved, while the keys and values remain RGB-derived.

Bounded low-rank feature refinement. Following cross-scale fusion, aligned task evidence locally refines the RGB features at each scale. For the fused representation F_C^l, the auxiliary residual correction is generated via a bottleneck low-rank mapping:

$$\Delta F^l = W_{\text{up}}^l \phi\left(W_{\text{down}}^l\left[\Delta Z_s^l,\, \Delta G^l,\, \sigma(\widetilde{O}_X^l),\, C_{\text{align}}^l,\, g^l\right]\right), \tag{10}$$

where W_{down}^l projects the task context into a compact rank-r subspace ($r \ll d$). To prevent feature distortion from misaligned or extreme auxiliary activations, the update magnitude is strictly bounded relative to the local RGB feature norm:

$$\widehat{\Delta F}^{\,l}(p) = \Delta F^l(p) \min\left(1,\, \frac{\rho \|F_C^l(p)\|_2}{\|\Delta F^l(p)\|_2 + \epsilon}\right), \qquad F_{\text{out}}^l = F_C^l + \widehat{\Delta F}^{\,l}. \tag{11}$$

Consequently, query modulation is explicitly gated by task reliability, whereas feature refinement integrates reliability-conditioned task discrepancies under dual low-rank capacity and norm bounds. Together, these complementary interactions enable auxiliary evidence to enrich the RGB encoder without relying on a resource-intensive parallel auxiliary semantic branch.

3.5 Training Objectives

TF-DETR is trained end-to-end using a joint objective that couples the standard RT-DETRv2 detection loss $\mathcal{L}_{\text{det}}$ (including its auxiliary and denoising supervision) with dedicated auxiliary objectives for task grounding, cross-modal alignment, and intervention consistency:

$$\mathcal{L} = \mathcal{L}_{\text{det}} + \lambda_{\text{task}}\mathcal{L}_{\text{task}} + \mathcal{L}_{\text{align}} + \lambda_{\text{inv}}\mathcal{L}_{\text{inv}}. \tag{12}$$

Task-factor grounding. The explicitly defined task variables are supervised at all three scales for both modalities. Objectness is trained with detection-oriented supervision, while geometry is optimized with uncertainty-aware regression. In compact form,

$$\mathcal{L}_{\text{task}} = \sum_{m\in\{R,X\}} \sum_{l=3}^{5} \left(\lambda_O^m \mathcal{L}_{\text{obj}}^l + \lambda_G^m \mathcal{L}_{\text{geo}}^l\right), \qquad \ell_{\text{geo}} = e^{-U}|G - G^*| + U. \tag{13}$$

The latent factors Z_p and Z_s receive no explicit regression targets and are learned through alignment and downstream detection.

Alignment supervision. GAMTA is trained with complementary global, local, rejection, and alignment consistency:

$$\mathcal{L}_{\text{align}} = \lambda_{\text{G}}\mathcal{L}_{\text{G}} + \lambda_{\text{L}}\mathcal{L}_{\text{L}} + \lambda_{\text{N}}\mathcal{L}_{\text{N}} + \lambda_{\text{C}}\mathcal{L}_{\text{cons}}. \tag{14}$$

These terms supervise the global transform, scale-specific residual correspondence, invalid-support rejection, and correspondence consistency under controlled auxiliary perturbations, respectively.

Intervention consistency. We further apply controlled perturbations to the auxiliary observation and encourage the explicit objectness and geometry states to remain stable.

$$\mathcal{L}_{\text{inv}} = D_O(O_X, O_{X^{\text{int}}}) + \lambda_G^{\text{inv}} D_G(G_X, G_{X^{\text{int}}}). \tag{15}$$

This objective is applied only to O_X and G_X, leaving the latent factors unconstrained by intervention consistency. See Appendix C.4 for supervision details.

4 Experiments

4.1 Experimental Setup

Datasets. We conduct the main evaluation on four RGB-X object detection benchmarks covering RGB-Thermal and RGB-Event sensing. Specifically, M3FD (Liu et al., 2022) and DroneVehicle (Sun et al., 2022) are used for RGB-Thermal detection, while DSEC-Detection (Gehrig & Scaramuzza, 2024) and PKU-DAVIS-SOD (Li et al., 2023a) are used for RGB-Event detection. We

Table 1: Comparison with state-of-the-art RGB-Event object detection methods on DSEC-Det and PKU-DAVIS-SOD. The best results are shown in **bold**, and the second-best results are underlined.

Method	Backbone	DSEC-Det		PKU-DAVIS-SOD		Params
		mAP_{50}	mAP	mAP_{50}	mAP	
FPN-fusion (Tomy et al., 2022)	ResNet-50	28.9	15.2	36.6	19.5	65.6M
SODFormer (Li et al., 2023a)	ResNet-50	51.3	28.4	50.4	20.7	87.8M
SFNet (Liu et al., 2024)	CSPDarkNet	41.2	26.8	59.6	31.9	57.5M
FAOD (Zhang et al., 2024)	CSPDarkNet	63.5	42.5	57.5	30.5	20.3M
ACGR (Li et al., 2025)	ResNet-50	51.9	31.7	51.9	–	41.4M
SFDNet (Fan et al., 2025)	YOLOv8	73.3	51.3	60.7	31.3	58.1M
SPFD (Wang et al., 2026)	ResNet-50	74.1	49.2	62.4	32.0	78.6M
Hyper-FEOD (Bao et al., 2026)	YOLO11s	73.2	52.5	58.5	32.3	19.42M
TF-DETR	ResNet-50	**81.7**	**59.3**	**68.5**	**39.2**	43.8M

Table 2: Efficiency and speed comparisons with state-of-the-art detectors. **(a)** Evaluation on the RGB-Event DSEC-Detection benchmark. **(b)** Evaluation on the M3FD benchmark. All models are evaluated on an RTX 4090 with a batch size of 1.

(a) DSEC-Detection (RGB-Event)				(b) M3FD (RGB-Thermal)			
Method	Backbone	Params	FPS	Method	Backbone	Params	FPS
SODFormer (Li et al., 2023a)	ResNet-50	87.8M	16.2	CAFF-DINO (Helvig et al., 2024)	Swin-Large	196.1M	7.8
DAGr (Gehrig & Scaramuzza, 2024)	ResNet-50	34.6M	22.7	CFT (Qingyun et al., 2021)	YOLOv5	44.2M	12.4
ACGR (Li et al., 2025)	ResNet-50	41.4M	33.1	ICAFusion (Shen et al., 2024)	YOLOv5	48.7M	10.9
SFNet (Liu et al., 2024)	CSPDarkNet	57.5M	35.3	GM-DETR (Xiao et al., 2024)	HGNetv2	70.0M	24.5
CAFR (Cao et al., 2024)	ResNet-50	82.4M	13.4	MM-DETR (Han et al., 2025)	ResNet-50	50.0M	25.0
TF-DETR (Ours)	ResNet-50	**43.8M**	**39.4**	**TF-DETR (Ours)**	ResNet-50	**43.8M**	**36.4**

additionally evaluate on SUN RGB-D (Song et al., 2015) and KITTI (Geiger et al., 2012) under a 2D RGB-Depth detection setting to assess cross-sensor generalization. Further dataset details are provided in Appendix A.

Evaluation Metrics and Implementation Details. We evaluate detection performance using standard 2D object detection metrics, primarily AP and AP_{50}.TF-DETR uses RT-DETRv2 (Zhao et al., 2024a) with a ResNet-50 RGB backbone (He et al., 2016) and the lightweight RepATE encoder for the auxiliary modality. Further details on modality-specific input processing, model implementation, and training configurations are provided in Appendices B and C.

4.2 Comparison with SOTA Results

RGB-Event detection. As shown in Table 1, TF-DETR consistently outperforms existing methods on both DSEC-Det and PKU-DAVIS-SOD. On DSEC-Det, it improves over the strongest competing results by 6.8 mAP and 7.6 mAP_{50}, while on PKU-DAVIS-SOD the corresponding gains are 6.9 mAP and 6.1 mAP_{50}. The consistent improvements across the two benchmarks suggest that representing event observations through compact Task Factors remains effective under different event-data distributions and acquisition conditions. Notably, these gains are achieved with a compact 43.8M-parameter model, without requiring a larger auxiliary branch.

RGB-Thermal detection. As shown in Table 3, TF-DETR achieves strong performance on both RGB-Thermal benchmarks. On M3FD, it obtains 64.6 AP and 93.3 AP_{50}, improving the previous best results by 0.2 AP and 1.2 AP_{50}, respectively. On DroneVehicle, TF-DETR achieves the best AP_{50} of 80.1 and the second-best AP of 58.8, remaining close to the highest AP while outperforming most existing methods. The consistent performance across the two benchmarks supports the effectiveness of the Task-Factor representation under different RGB-Thermal detection settings. TF-DETR maintains these results with the same compact 43.8M-parameter architecture, further highlighting the efficiency of its auxiliary representation.

RGB-Depth detection. We further evaluate TF-DETR on SUN RGB-D and KITTI to examine whether the proposed Task-Factor formulation extends beyond thermal and event sensing. Under the 2D RGB-Depth detection setting, TF-DETR achieves 54.7 AP and 69.1 AP_{50} on SUN RGB-D,

Table 3: Comparison with state-of-the-art RGB-Thermal detectors on M3FD and DroneVehicle. Best results are shown in **bold**, and second-best results are underlined.

(a) M3FD					(b) DroneVehicle				
Method	**Backbone**	**mAP$_{50}$**	**mAP**	**Params**	**Method**	**Backbone**	**mAP$_{50}$**	**mAP**	**Params**
TarDAL (Liu et al., 2022)	YOLOv5	80.5	54.1	–	TSFADet (Yuan et al., 2022)	ResNet-50	73.1	44.1	104.7M
IGNet (Li et al., 2023b)	YOLOv5	81.5	54.5	–	C^{2}Former (Yuan & Wei, 2024)	ResNet-50	74.2	47.3	100.8M
DAMSDet (Guo et al., 2024)	ResNet-50	80.2	52.9	79.1M	CSOM-ODAF (Chen et al., 2024)	YOLOv5	77.1	50.1	–
EMMA (Zhao et al., 2024b)	YOLOv5	82.9	55.4	–	DMM (Zhou et al., 2025)	ResNet-50	77.2	55.8	88.0M
MMFN (Yang et al., 2024a)	YOLOv5	86.2	57.4	176M	IV-YOLO (Tian et al., 2024)	YOLOv8	74.6	56.8	4.31M
Fusion-Mamba (Dong et al., 2025)	YOLOv8	88.0	61.9	287.6M	Fusion-Mamba (Dong et al., 2025)	YOLOv8	79.2	56.0	287.6M
WaveMamba (Zhu et al., 2025)	YOLOv8	92.1	64.4	69.1M	WaveMamba (Zhu et al., 2025)	ResNet-50	79.3	**59.9**	193.2M
SFEDet (Tian et al., 2026)	YOLOv8	89.8	61.0	24.5M	MFE-DETR (Yan & Wang, 2026)	ResNet-50	78.6	57.8	45.8M
TF-DETR (Ours)	ResNet-50	**93.3**	**64.6**	43.8M	**TF-DETR (Ours)**	ResNet-50	**80.1**	58.8	43.8M

Table 4: Ablation on auxiliary representation and capacity. The upper block compares auxiliary encoder capacity and representation type, while the lower block studies the residual latent capacity of Task Factors.

ID	X Encoder	Aux. Representation	Dim.	Task Loss.	mAP$_{50}$	mAP	Params
Representation and Encoder Capacity							
A0	ResNet-50	Task Factor	33	✓	92.9	64.4	66.4M
A1	ResNet-50	Generic Feature	256	×	93.0	**64.8**	67.1M
A2	RepATE	Generic Feature	32	×	92.5	64.0	43.3M
A3	RepATE	Task Factor	33	✓	**93.3**	64.6	43.8M
Residual Latent Capacity							
A4	RepATE	Task Factor	25	✓	92.4	64.0	43.3M
A3	RepATE	Task Factor	33	✓	**93.3**	**64.6**	43.8M
A5	RepATE	Task Factor	49	✓	92.4	64.1	44.4M

and 79.6 AP and 95.9 AP$_{50}$ on KITTI. These results demonstrate that the same Task-Factor interface can accommodate depth observations without changing its representation formulation, providing additional evidence of its applicability across heterogeneous auxiliary sensors.

4.3 Efficiency and Inference Speed

As shown in Table 2, TF-DETR reaches 39.4 FPS on DSEC-Detection and 36.4 FPS on M3FD with 43.8M parameters. For a consistent efficiency comparison, all models are evaluated on a single NVIDIA RTX 4090 GPU with an input resolution of 640×640 pixels and a batch size of 1. Under this setting, TF-DETR provides the highest throughput among the compared methods. These results indicate that the lightweight auxiliary encoder and compact Task-Factor representation avoid much of the computational overhead associated with maintaining a second full semantic stream, while preserving high detection throughput. With TensorRT 10.4 FP16 deployment under the same setting, throughput further increases to 240.3 FPS. See Appendix D for measurement details.

4.4 Ablation Study

Auxiliary representation and capacity. As shown in Table 4, the ResNet-50 auxiliary branch with generic features remains competitive, whereas the lightweight RepATE encoder with a 33D Task Factor achieves comparable detection accuracy with approximately 35% fewer parameters. This result shows that a full high-dimensional auxiliary semantic stream is not required to match the ResNet-50 baseline under the current setting. Under the same RepATE encoder and nearly matched representation dimensionality, Task Factors outperform compact generic features, indicating that low-dimensional compression alone is insufficient and that task-oriented structure together with explicit task grounding is beneficial for lightweight auxiliary modeling. Moreover, replacing ResNet-50 with RepATE under the same Task-Factor formulation does not degrade performance, suggesting that once auxiliary evidence is organized into compact task states, the additional capacity of a heavy auxiliary encoder provides limited benefit. The residual latent capacity study further shows that increasing the Task Factor from 25D to the default 33D yields a modest improvement, while expanding it to 49D brings no additional gain.

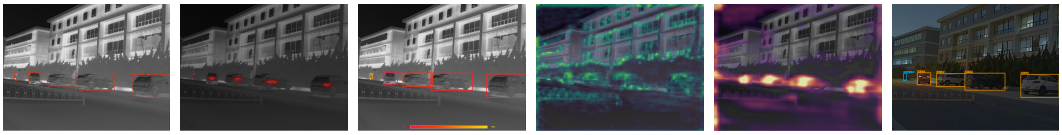
Table 5: Ablation studies on M3FD. **(a)** Effect of query guidance and feature refinement for integrating aligned Task Factors into the RGB encoder. **(b)** Effect of individual training objectives, evaluated by removing each objective from the full TF-DETR model.

(a) Encoder interaction

Query Guidance	Feature Refinement	mAP_{50}	mAP
×	×	91.7	63.5
×	✓	92.7	64.1
✓	×	92.1	64.0
✓	✓	**93.3**	**64.6**

(b) Training objectives

Setting	mAP_{50}	mAP
Full supervision	**93.3**	**64.6**
w/o Task Factor Loss	91.8	63.6
w/o Alignment Supervision	92.0	63.8
w/o Intervention Consistency	92.5	64.2



(a) Thermal + GT (b) Objectness O_X (c) G_X / U_X (d) Position Z_p (e) Structure Z_s (f) Detection

Figure 3: **Qualitative visualization of structured Task Factors.** O_X highlights object regions; G_X and U_X represent local object geometry and its predicted uncertainty; Z_p and Z_s exhibit correspondence-related positional responses and cross-modal residual task structure, respectively. Geometry boxes are color-coded from red (low uncertainty, LU) through orange to yellow (high uncertainty, HU). Final detections show the integrated outputs of these structured representations.

Encoder interaction ablation. As shown in Table 5a, constructing and aligning Task Factors alone is insufficient to fully exploit auxiliary evidence; effective interaction is still required to incorporate it into the primary RGB stream. Either query guidance or feature refinement consistently improves detection when used independently, indicating that both paths can make use of the aligned Task Factors. Their combination achieves the best performance, suggesting that the two mechanisms are complementary in where and how auxiliary task evidence affects the RGB encoder. Query guidance influences semantic retrieval during attention, whereas feature refinement introduces bounded residual corrections after multi-scale fusion. Together, they enable Task Factors to interact with the RGB representation at different stages while preserving RGB as the primary semantic stream.

4.5 Qualitative analysis of Task Factors.

As shown in Fig. 3, the learned Task Factors exhibit distinct task-relevant spatial behaviors rather than collapsing into a generic auxiliary representation. O_X highlights object regions, while G_X and U_X capture object geometry and its associated uncertainty, with more challenging small-scale targets showing higher uncertainty. Meanwhile, Z_p exhibits spatially organized correspondence-related responses across RGB and Thermal, whereas the cross-modal residual of Z_s concentrates mainly on detection-relevant regions. These observations suggest that the learned Task Factors do not collapse into redundant generic features; instead, they form functionally differentiated intermediate representations for object presence, geometry, correspondence, and residual local task structure. Additional qualitative results are provided in Appendix E.

5 Conclusion

We propose TF-DETR for RGB-X object detection, representing auxiliary observations through compact Task Factors rather than a full parallel semantic stream. TF-DETR combines task-space alignment with reliability-aware interaction to effectively incorporate complementary auxiliary evidence into the RGB representation. Experiments across RGB-Thermal, RGB-Event, and RGB-Depth settings demonstrate strong performance and cross-sensor applicability, supporting task-oriented auxiliary representation as an effective and efficient design for RGB-X detection.

AI USE STATEMENT

Generative AI tools were used only to assist with language polishing, manuscript organization, and the presentation of figures and tables. The research ideas, methodology, implementation, experiments, analysis, and conclusions were developed and verified by the authors. The authors reviewed all AI-assisted content and take full responsibility for the final manuscript.

REPRODUCIBILITY STATEMENT

We provide detailed descriptions of the proposed method, model architecture, training objectives, and inference procedure in the main paper. The Appendix further documents the implementation details, model configurations, hyperparameter settings, modality-specific preprocessing, and dataset protocols used in our experiments. All six benchmarks used in this work are based on publicly available datasets, with their corresponding preprocessing and evaluation procedures described to facilitate consistent reproduction of the reported results. Source code and configuration files will be released upon publication.

ETHICS STATEMENT

This work is based on publicly available object-detection benchmarks and does not introduce new data collection involving human participants. The proposed method is intended for general multimodal perception research. The authors declare no conflicts of interest and received no external financial support for this work.

REFERENCES

Wei Bao, Yuehan Wang, Tianhang Zhou, Siqi Li, and Yue Gao. Sparse hypergraph-enhanced frame-event object detection with fine-grained moe. *arXiv preprint arXiv:2604.11140*, 2026.

Jiaxin Cai, Jingze Su, Qi Li, Wenjie Yang, Shu Wang, Tiesong Zhao, Shengfeng He, and Wenxi Liu. Keep the balance: A parameter-efficient symmetrical framework for rgb+ x semantic segmentation. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 10587–10598. IEEE, 2025.

Hu Cao, Zehua Zhang, Yan Xia, Xinyi Li, Jiahao Xia, Guang Chen, and Alois Knoll. Embracing events and frames with hierarchical feature refinement network for object detection. In *European Conference on Computer Vision*, pp. 161–177. Springer, 2024.

Chen Chen, Jiahao Qi, Xingyue Liu, Kangcheng Bin, Ruigang Fu, Xikun Hu, and Ping Zhong. Weakly misalignment-free adaptive feature alignment for uavs-based multimodal object detection. In *2024 IEEE/CVF conference on computer vision and pattern recognition (CVPR)*, pp. 26826–26835. IEEE, 2024.

Zhiyuan Cheng, Hongjun Choi, Shiwei Feng, James Liang, Guanhong Tao, Dongfang Liu, Michael Zuzak, and Xiangyu Zhang. Fusion is not enough: Single modal attacks on fusion models for 3d object detection. In *International Conference on Learning Representations*, volume 2024, pp. 23905–23929, 2024.

Minkyoung Cho, Yulong Cao, Jiachen Sun, Qingzhao Zhang, Marco Pavone, Jeong Joon Park, Heng Yang, and Zhuoqing Mao. Cocoon: Robust multi-modal perception with uncertainty-aware sensor fusion. In *International Conference on Learning Representations*, volume 2025, pp. 40998–41021, 2025.

Sri Aditya Deevi, Connor Lee, Lu Gan, Sushruth Nagesh, Gaurav Pandey, and Soon-Jo Chung. Rgb-x object detection via scene-specific fusion modules. In *2024 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pp. 7351–7360. IEEE, 2024.

Wenhao Dong, Haodong Zhu, Shaohui Lin, Xiaoyan Luo, Yunhang Shen, Guodong Guo, and Baochang Zhang. Fusion-mamba for cross-modality object detection. *IEEE Transactions on Multimedia*, 2025.

Benoit Dufumier, Javiera Castillo Navarro, Devis Tuia, and Jean-Philippe Thiran. What to align in multimodal contrastive learning? In *International Conference on Learning Representations*, volume 2025, pp. 5408–5432, 2025.

Liangwei Fan, Jingjun Yang, Lei Wang, Jinpu Zhang, Xiangkai Lian, and Hui Shen. Efficient spiking neural network for rgb–event fusion-based object detection. *Electronics*, 14(6):1105, 2025.

Daniel Gehrig and Davide Scaramuzza. Low-latency automotive vision with event cameras. *Nature*, 629(8014):1034–1040, 2024.

Andreas Geiger, Philip Lenz, and Raquel Urtasun. Are we ready for autonomous driving? the kitti vision benchmark suite. In *2012 IEEE conference on computer vision and pattern recognition*, pp. 3354–3361. IEEE, 2012.

Junjie Guo, Chenqiang Gao, Fangcen Liu, Deyu Meng, and Xinbo Gao. Damsdet: Dynamic adaptive multispectral detection transformer with competitive query selection and adaptive feature fusion. *arXiv e-prints*, pp. arXiv–2403, 2024.

Jianhong Han, Yupei Wang, Yuan Zhang, and Liang Chen. Mm-detr: An efficient multimodal detection transformer with mamba-driven dual-granularity fusion and frequency-aware modality adapters. *arXiv preprint arXiv:2512.00363*, 2025.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 770–778, 2016.

Xiao He, Chang Tang, Xin Zou, and Wei Zhang. Multispectral object detection via cross-modal conflict-aware learning. In *Proceedings of the 31st ACM International Conference on Multimedia*, pp. 1465–1474, 2023.

Kevin Helvig, Baptiste Abeloos, and Pauline Trouvé-Peloux. Caff-dino: Multi-spectral object detection transformers with cross-attention features fusion. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pp. 3037–3046. IEEE, 2024.

Lingyi Hong, Shilin Yan, Renrui Zhang, Wanyun Li, Xinyu Zhou, Pinxue Guo, Kaixun Jiang, Yiting Chen, Jinglun Li, Zhaoyu Chen, et al. Onetracker: Unifying visual object tracking with foundation models and efficient tuning. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 19079–19091. IEEE, 2024.

Dianze Li, Yonghong Tian, and Jianing Li. Sodformer: Streaming object detection with transformer using events and frames. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45 (11):14020–14037, 2023a.

Dianze Li, Jianing Li, Xu Liu, Xiaopeng Fan, and Yonghong Tian. Asynchronous collaborative graph representation for frames and events. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1655–1666. IEEE, 2025.

Jiawei Li, Jiansheng Chen, Jinyuan Liu, and Huimin Ma. Learning a graph neural network with cross modality interaction for image fusion. In *Proceedings of the 31st ACM international conference on multimedia*, pp. 4471–4479, 2023b.

Jinyuan Liu, Xin Fan, Zhanbo Huang, Guanyao Wu, Risheng Liu, Wei Zhong, and Zhongxuan Luo. Target-aware dual adversarial learning and a multi-scenario multi-modality benchmark to fuse infrared and visible for object detection. In *2022 IEEE/CVF conference on computer vision and pattern recognition (CVPR)*, pp. 5792–5801. IEEE, 2022.

Zhanwen Liu, Nan Yang, Yang Wang, Yuke Li, Xiangmo Zhao, and Fei-Yue Wang. Enhancing traffic object detection in variable illumination with rgb-event fusion. *IEEE Transactions on Intelligent Transportation Systems*, 25(12):20335–20350, 2024.

Divyam Madaan, Varshan Muhunthan, Kyunghyun Cho, and Sumit Chopra. Multi-modal data spectrum: Multi-modal datasets are multi-dimensional. In *International Conference on Learning Representations*, volume 2026, pp. 131287–131305, 2026.

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

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

Ayush Shrivastava and Andrew Owens. Self-supervised spatial correspondence across modalities. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 6383–6393. IEEE, 2025.

Shuran Song, Samuel P Lichtenberg, and Jianxiong Xiao. Sun rgb-d: A rgb-d scene understanding benchmark suite. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 567–576, 2015.

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

Yuedong Tan, Zongwei Wu, Yuqian Fu, Zhuyun Zhou, Guolei Sun, Eduard Zamfir, Chao Ma, Danda Paudel, Luc Van Gool, and Radu Timofte. Xtrack: Multimodal training boosts rgb-x video object trackers. In *2025 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 5734–5744. IEEE, 2025.

Chao Tian, Zikun Zhou, Chao Yang, Guoqing Zhu, and Zhenyu He. Efficient rgb-t object detection via sparse cross-modality fusion. In *European Conference on Computer Vision*, pp. 541–559. Springer, 2026.

Dan Tian, Xin Yan, Dong Zhou, Chen Wang, and Wenshuai Zhang. Iv-yolo: A lightweight dual-branch object detection network. *Sensors*, 24(19):6181, 2024.

Abhishek Tomy, Anshul Paigwar, Khushdeep S Mann, Alessandro Renzaglia, and Christian Laugier. Fusing event-based and rgb camera for robust object detection in adverse conditions. In *2022 International conference on robotics and automation (ICRA)*, pp. 933–939. IEEE, 2022.

Beatrice van Amsterdam, Abdolrahim Kadkhodamohammadi, Imanol Luengo, and Danail Stoyanov. Aspnet: Action segmentation with shared-private representation of multiple data sources. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2384–2393. IEEE, 2023.

Chenyu Wang, Sharut Gupta, Xinyi Zhang, Sana Tonekaboni, Stefanie Jegelka, Tommi Jaakkola, and Caroline Uhler. An information criterion for controlled disentanglement of multimodal data. *arXiv preprint arXiv:2410.23996*, 2024.

Hu Wang, Yuanhong Chen, Congbo Ma, Jodie Avery, Louise Hull, and Gustavo Carneiro. Multimodal learning with missing modality via shared-specific feature modelling. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 15878–15887. IEEE, 2023.

Keyao Wang, Shuai Liu, Hengda Shi, Lukui Shi, and Haiyong Chen. Beyond duality: A hybrid framework of leveraging shared and private features for rgb-event object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 4415–4424, 2026.

Yikai Wang, Xinghao Chen, Lele Cao, Wenbing Huang, Fuchun Sun, and Yunhe Wang. Multimodal token fusion for vision transformers. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 12176–12185. IEEE, 2022.

Zongwei Wu, Jilai Zheng, Xiangxuan Ren, Florin-Alexandru Vasluianu, Chao Ma, Danda Pani Paudel, Luc Van Gool, and Radu Timofte. Single-model and any-modality for video object tracking. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 19156–19166. IEEE, 2024.

Yiming Xiao, Fanman Meng, Qingbo Wu, Linfeng Xu, Mingzhou He, and Hongliang Li. Gm-detr: Generalized muiltispectral detection transformer with efficient fusion encoder for visible-infrared detection. In *2024 IEEE/CVF conference on computer vision and pattern recognition workshops (CVPRW)*, pp. 5541–5549. IEEE, 2024.

Xiangyang Xu, Yuncheng Li, Gangshan Wu, and Jiebo Luo. Multi-modal deep feature learning for rgb-d object detection. *Pattern Recognition*, 72:300–313, 2017.

Zekai Yan and Mu-Jiang-Shan Wang. Mfe-detr: Multimodal feature-enhanced detection transformer for rgb–infrared object detection in aerial imagery. *Symmetry*, 18(3):417, 2026.

Fan Yang, Binbin Liang, Wei Li, and Jianwei Zhang. Multidimensional fusion network for multi-spectral object detection. *IEEE Transactions on Circuits and Systems for Video Technology*, 35 (1):547–560, 2024a.

Haitian Yang, Juan Fang, Yiren Zhu, Xudong Zhao, Yufei Guo, Xiaohan Zhang, Xiaoxing Hu, Xue Yang, and Qi Ming. Crossweaver: Towards efficient cross-modal interweaving and decoupling for weakly-aligned multispectral object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 6361–6370, 2026.

Zequn Yang, Yake Wei, Ce Liang, and Di Hu. Quantifying and enhancing multi-modal robustness with modality preference. In *International Conference on Learning Representations*, volume 2024, pp. 15860–15872, 2024b.

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

Maoxun Yuan, Yinyan Wang, and Xingxing Wei. Translation, scale and rotation: Cross-modal alignment meets rgb-infrared vehicle detection. In *European conference on computer vision*, pp. 509–525. Springer, 2022.

Haitian Zhang, Xiangyuan Wang, Chang Xu, Xinya Wang, Fang Xu, Huai Yu, Lei Yu, and Wen Yang. Frequency-adaptive low-latency object detection using events and frames. *arXiv preprint arXiv:2412.04149*, 2024.

Jiaming Zhang, Huayao Liu, Kailun Yang, Xinxin Hu, Ruiping Liu, and Rainer Stiefelhagen. Cmx: Cross-modal fusion for rgb-x semantic segmentation with transformers. *IEEE Transactions on intelligent transportation systems*, 24(12):14679–14694, 2023a.

Jiaming Zhang, Ruiping Liu, Hao Shi, Kailun Yang, Simon Reiß, Kunyu Peng, Haodong Fu, Kaiwei Wang, and Rainer Stiefelhagen. Delivering arbitrary-modal semantic segmentation. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1136–1147. IEEE, 2023b.

Yian Zhao, Wenyu Lv, Shangliang Xu, Jinman Wei, Guanzhong Wang, Qingqing Dang, Yi Liu, and Jie Chen. Detrs beat yolos on real-time object detection. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 16965–16974. IEEE, 2024a.

Zixiang Zhao, Haowen Bai, Jiangshe Zhang, Yulun Zhang, Kai Zhang, Shuang Xu, Dongdong Chen, Radu Timofte, and Luc Van Gool. Equivariant multi-modality image fusion. In *2024 IEEE/CVF conference on computer vision and pattern recognition (CVPR)*, pp. 25912–25921. IEEE, 2024b.

Xu Zheng, Yuanhuiyi Lyu, and Lin Wang. Learning modality-agnostic representation for semantic segmentation from any modalities. In *European Conference on Computer Vision*, pp. 146–165. Springer, 2024.

Minghang Zhou, Tianyu Li, Chaofan Qiao, Dongyu Xie, Guoqing Wang, Ningjuan Ruan, Lin Mei, Yang Yang, and Heng Tao Shen. Dmm: Disparity-guided multispectral mamba for oriented object detection in remote sensing. *IEEE Transactions on Geoscience and Remote Sensing*, 63:1–13, 2025.

Haodong Zhu, Wenhao Dong, Linlin Yang, Hong Li, Yuguang Yang, Yangyang Ren, Qingcheng Zhu, Zichao Feng, Changbai Li, Shaohui Lin, et al. Wavemamba: Wavelet-driven mamba fusion for rgb-infrared object detection. In *2025 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 11219–11229. IEEE, 2025.

A Datasets and Evaluation Protocols

We evaluate TF-DETR on six RGB-X object detection benchmarks covering three complementary auxiliary modalities: RGB-Event, RGB-Thermal, and RGB-Depth. Table A.1 summarizes the dataset characteristics and the evaluation protocols adopted in our experiments. Unless otherwise specified, all experiments are evaluated using the COCO bounding-box protocol, with AP averaged over IoU thresholds from 0.50 to 0.95 ($\mathrm{AP}_{50:95}$) as the primary metric.

DSEC-Detection. DSEC-Detection extends the DSEC driving dataset with object-level annotations for event-based automotive perception. The benchmark contains 60 sequences, 70,379 annotated frames, and 390,118 bounding boxes spanning eight categories: pedestrian, rider, car, bus, truck, bicycle, motorcycle, and train. The official DSEC-Detection protocol defines 41 training, 6 validation, and 13 test sequences.

A particular characteristic of DSEC-Detection is that its object annotations are defined in the distorted event-camera view. The original rectified image frames therefore need to be remapped to the event-camera view before constructing aligned RGB–Event pairs. This spatial convention is retained in the processed data used in our experiments.

For data preparation, we follow the FAOD-style processing protocol, which organizes DSEC-Detection using an RVT-compatible representation and supports frequency-controlled pairing between RGB frames and Event observations. Since the DSEC image stream operates at 20 Hz in this processing protocol, this setting associates RGB and Event observations at a nominal frequency of 20 Hz.

We further convert the processed DSEC-Detection data offline into the same HDF5/COCO metadata organization used by our RGB-E data pipeline. The conversion changes only the storage and indexing format: each detection sample retains its paired RGB frame, Event representation, and object annotations. We train and evaluate all eight categories without category filtering.

Importantly, while our data organization follows the FAOD-style DSEC processing protocol, evaluation is performed using our unified COCO bounding-box evaluator. We therefore report $\mathrm{AP}_{50:95}$ under the same evaluation implementation used for the remaining RGB-X datasets, rather than relying on a dataset-specific metric implementation.

PKU-DAVIS-SOD. PKU-DAVIS-SOD is a large-scale neuromorphic object detection benchmark designed for challenging traffic scenarios involving illumination changes, high-speed motion, and motion blur. The dataset is captured using a DAVIS346 camera, which simultaneously provides conventional frames and asynchronous events at a spatial resolution of 346×260. It contains 220 driving sequences, with conventional frames recorded at 25 Hz.

The benchmark provides annotations for three traffic-related categories: car, pedestrian, and two-wheeler. In total, approximately 276k timestamps are manually annotated, resulting in about 1.0801M bounding boxes. The official dataset statistics report approximately 671.3k, 194.7k, and 214.1k bounding-box annotations for the training, validation, and test subsets, respectively.

For challenging samples in which conventional frames suffer from severe low-light degradation or motion blur, the dataset creators additionally reconstruct grayscale images from the event stream using E2VID to assist manual annotation. These reconstructed images are used only during the annotation process and are not used as model inputs in our experiments. TF-DETR receives the original RGB frame together with its corresponding Event representation.

Following the FAOD-style preprocessing protocol, we use the `freq_1_1` version of PKU-DAVIS-SOD. The frequency construction of this protocol uses the original 25 Hz acquisition rate as its reference; therefore, the 1:1 configuration corresponds to nominally 25 Hz RGB and 25 Hz Event observations.

For our implementation, the processed sequences are further organized into a unified HDF5/COCO-style representation. Each detection sample is indexed by its sequence identity, RGB-frame index, Event-representation index, and associated object annotations. Separate COCO-format annotation files are used for training and validation. This conversion standardizes the data interface while preserving the original RGB–Event pairing and the three-category detection label space.

As for DSEC-Detection, evaluation is performed using our unified COCO bounding-box evaluator. We report AP averaged over IoU thresholds from 0.50 to 0.95 ($\mathrm{AP}_{50:95}$), ensuring a consistent evaluation implementation across all RGB-X benchmarks.

DroneVehicle. DroneVehicle is a large-scale aerial RGB–Thermal benchmark designed for cross-modal vehicle detection from unmanned aerial platforms. The public dataset contains 56,878 images, corresponding to 28,439 paired RGB and infrared observations. It provides oriented bounding-box annotations for five vehicle categories: car, truck, bus, van, and freight car. The downloaded images have a spatial resolution of 840×712, including the border region introduced by the original dataset during annotation.

For our experiments, the RGB and thermal streams are organized as paired image directories and converted into a COCO-style detection representation. The COCO annotation files define the training and validation samples, while the corresponding thermal image is retrieved from the paired infrared directory according to the RGB file name. All five vehicle categories are retained in our experiments.

Although the original DroneVehicle annotations are provided as oriented bounding boxes, our detector is trained and evaluated using the converted COCO bounding-box annotations adopted by the unified TF-DETR pipeline. Accordingly, DroneVehicle is evaluated using the same COCO bounding-box evaluator as the remaining benchmarks, with $\mathrm{AP}_{50:95}$ as the primary metric.

M3FD. M3FD is a multi-scenario visible–infrared benchmark containing 4,200 registered RGB–infrared image pairs for image-fusion and object-detection tasks. Most image pairs have a spatial resolution of 1024×768. The visible and infrared observations are spatially registered, and the dataset covers diverse illumination, weather, and scene conditions.

The detection annotations contain 34,407 manually labeled object instances distributed across six categories: people, car, bus, motorcycle, lamp, and truck. This broader category composition makes M3FD complementary to DroneVehicle, which focuses exclusively on aerial vehicle categories.

For our experiments, the visible images are stored in the `vi` directory and the corresponding infrared images in the `ir` directory. The paired modalities share the same relative file names. Rather than assuming a dataset split not explicitly encoded in the original image pairs, we define the training and validation subsets through separate COCO-format annotation files. The resulting detector therefore operates on the six-category label space defined by the M3FD annotations.

Evaluation follows the same unified COCO bounding-box protocol, using $\mathrm{AP}_{50:95}$ as the primary metric.

KITTI. KITTI is a widely used autonomous-driving benchmark collected from real-world urban traffic environments. Its official 2D object-detection benchmark contains 7,481 training images and 7,518 test images, together with synchronized visual and range-sensing measurements.

The original KITTI benchmark is not directly used in its native evaluation form in our experiments. Instead, we use an offline-converted RGB–Depth version in which RGB images and their corresponding precomputed depth maps are organized as paired inputs. The training and validation subsets are represented by separate COCO-format annotation files, providing a unified data interface with the other RGB-X benchmarks.

Our converted KITTI RGB-D annotation set defines a 16-category detection label space, and the detector is therefore configured with 16 output categories. This should be distinguished from the native KITTI object benchmark and its original category/evaluation protocol: the reported results in our work correspond to our converted RGB-D COCO protocol rather than the official KITTI leaderboard setting.

The paired RGB-D inputs are evaluated using our common COCO bounding-box evaluator, with $\mathrm{AP}_{50:95}$ as the primary metric.

SUN RGB-D. SUN RGB-D is a large-scale indoor RGB-D scene-understanding benchmark. SUN RGB-D V1 contains 10,335 RGB-D images acquired using four different RGB-D sensing systems.

Table A.1: Summary of the six RGB-X benchmarks used in our experiments. The *Scale* column reports statistics of the corresponding public dataset, whereas *#Cls.* denotes the detection label space used in our experiments. All reported results use our unified COCO bounding-box evaluation implementation.

Modality	Dataset	Scale	#Cls.	Experimental Protocol
RGB-E	DSEC-Detection	60 seq. / 70,379 labeled frames	8	FAOD-style 1:1 + HDF5/COCO conversion
RGB-E	PKU-DAVIS-SOD	220 seq. / 276k labeled timestamps	3	FAOD-style 1:1 + HDF5/COCO conversion
RGB-T	DroneVehicle	28,439 RGB-IR pairs	5	Paired COCO conversion
RGB-T	M3FD	4,200 RGB-IR pairs	6	Paired COCO conversion
RGB-D	KITTI	7,481 train / 7,518 test images	16	Paired RGB-D COCO conversion
RGB-D	SUN RGB-D	10,335 RGB-D images	16	Paired RGB-D COCO conversion

Table B.1: Modality-specific preprocessing of the auxiliary X input. The RGB stream is converted to `float32` and scaled by $1/255$ for all experiments.

X Modality	Channels	Numerical Processing	X Resize	Validity Handling	X-Specific Training Augmentation
Event	20	Float conversion only	Nearest-exact	Paired observation validity	None
Thermal	1	Datatype-based scaling	Bilinear	Modality presence / valid region	Occlusion ($p = 0.20$), dropout ($p = 0.05$)
Depth	1	Fixed-range scaling $[0, 20480]$	Nearest	Valid-depth mask	None

The dataset integrates multiple indoor RGB-D sources and provides dense scene annotations, including 2D object annotations and 3D bounding boxes, covering a wide range of indoor environments.

For our experiments, we convert SUN RGB-D into the same paired RGB-D COCO-style representation used for KITTI. RGB images and depth maps are stored in separate directories, while corresponding samples are associated through their file names. Separate COCO-format annotation files define the training and validation subsets.

Our converted SUN RGB-D detection protocol uses a 16-category label space, and the detector is configured accordingly. As with KITTI, this 16-category 2D detection setting is specific to our converted COCO annotations and should not be confused with the original SUN RGB-D semantic-segmentation or 3D object-detection evaluation protocols.

All SUN RGB-D results are evaluated with the same COCO bounding-box implementation used throughout our experiments, reporting $\mathrm{AP}_{50:95}$ as the primary metric.

B Modality-Specific Input Processing

TF-DETR uses a unified RGB-X architecture across Event, Thermal, and Depth inputs. The RGB stream is processed identically in all experiments, while only the representation and preprocessing of the auxiliary X modality are adapted to its sensing characteristics. Table B.1 summarizes these modality-specific operations.

B.1 Unified RGB-X Input Interface

All three auxiliary modalities are converted into a common RGB-X sample interface before entering TF-DETR. Each sample contains an RGB tensor I_{RGB}, an auxiliary tensor X, a modality-presence indicator, and an auxiliary validity mask. The same interface additionally supports alignment-related supervision when required.

The RGB stream always contains three channels and is converted to `float32` with intensity scaling by $1/255$. Appearance augmentation is restricted to RGB, whereas geometric transformations that affect spatial coordinates are synchronously applied to RGB, X, and the corresponding detection targets.

The auxiliary modality is then processed according to its sensing characteristics, as described below.

Table B.2: Common training configuration of TF-DETR.

Hyperparameter	Value	Hyperparameter	Value
Optimizer	AdamW	Gradient clipping	0.1
LR (new modules / backbone)	10^{-4} / 10^{-5}	Mixed precision	FP16 (AMP)
AdamW (β_1, β_2)	$(0.9, 0.999)$	EMA decay / warmup	0.9999 / 2k
Weight decay (base / norm)	1×10^{-4} / 0	Warmup steps / Seed	300 / 0
Hungarian cost (cls / L_1 / GIoU)	2.0 / 5.0 / 2.0	Task-factor loss	1.0
Detection loss (VFL / L_1 / GIoU)	1.0 / 5.0 / 2.0	Global-alignment loss	1.0
Null-match loss	0.2	Local-alignment loss	0.5
batch size (train / val)	16 / 32	Input resolution	640×640

B.2 Event Input

For RGB-E experiments, the Event modality is represented using a 20-channel stacked-histogram representation. The precomputed representation is stored using the configuration

$$\texttt{dt} = 50, \qquad \texttt{nbins} = 10, \tag{B.1}$$

and each Event observation is loaded as a tensor of size

$$X_E \in \mathbb{R}^{20 \times H \times W}. \tag{B.2}$$

Unlike conventional images, the Event representation is not treated as an 8-bit intensity image. The stored Event tensor is converted to `float32`, but no division by 255, per-sample min–max normalization, or additional image-style intensity normalization is applied. This preserves the numerical statistics of the precomputed Event histogram.

Spatial resizing is also adapted to the discrete histogram representation. RGB images are resized using bilinear interpolation, whereas Event tensors use nearest-exact interpolation whenever available, with nearest-neighbor interpolation as fallback. The same modality-specific rule is retained during batch-level multi-scale training.

The standard RGB-E training setting does not introduce additional X-specific corruption. Artificial affine misalignment, Event occlusion, and Event modality dropout are disabled in the main RGB-E experiments.

B.3 Thermal Input

For RGB-T experiments, the thermal modality is represented as a single-channel image,

$$X_T \in \mathbb{R}^{1 \times H \times W}. \tag{B.3}$$

Thermal images are explicitly converted to single-channel grayscale form before numerical conversion. They are subsequently converted to `float32` using datatype-based scaling. For an unsigned integer thermal image with datatype maximum $v_{\max}$, the conversion can be written as

$$\tilde{T} = \frac{T}{v_{\max}}. \tag{B.4}$$

Under the adopted 8-bit grayscale representation, this corresponds to scaling the thermal intensity approximately to the range $[0, 1]$.

Unlike Event and Depth, thermal images are resized using bilinear interpolation. RGB and thermal inputs therefore share the same interpolation scheme during batch-level multi-scale resizing, while their validity masks are resized using nearest-neighbor interpolation.

Table B.3: Detailed model configuration of TF-DETR. RepATE uses 20 input channels for Event and one input channel for Thermal and Depth. The task projection width denotes the intermediate task feature width before the factor-specific heads.

Component	Configuration	Value
RGB stream and hybrid encoder		
RGB backbone	Backbone	PResNet-50vd
	Input feature channels	$[512, 1024, 2048]$
	Feature strides	$[8, 16, 32]$
Hybrid encoder	Hidden dimension	256
	AIFI encoder level	P5 only
	Encoder layers	1
	Attention heads	8
	FFN dimension	1024
	Encoder activation	GELU
Auxiliary task encoder and Task Factors		
RepATE	Input channels	20 (Event), 1 (Thermal/Depth)
	Stage channels	$[24, 32, 48, 64, 96]$
	Stem / TaskRep blocks	1 stem conv. / $[2, 2, 2, 1]$
	Task projection width	32
Task Factors	Objectness O	1
	Geometry G	4
	Geometric uncertainty$U$	4
	Positional latent Z_{p}	8
	Residual task latent Z_{s}	16
	Total Task-Factor dimension	33
Task-space alignment and reliability		
Global alignment	Estimation level	P5
	Projected dimension	16
	Transformation	Translation + rotation + isotropic scale
	Maximum translation	0.15
	Maximum rotation	15°
	Scale range	$[0.8, 1.25]$
Local transport	Candidate correspondences	4 + Null Match
	Maximum offsets (P3/P4/P5)	$[3, 2, 1]$ grid units
Task-state estimation	Reliability cues	Auxiliary quality, alignment confidence, U_R, $\widetilde{U}_X$, D_O, D_G
Task-conditioned RGB encoding		
TQ-AIFI	Applied level	P5
	Information flow	Task-conditioned Q; RGB-only K, V
	Initial task coefficient α	0
Task refinement	Position	After CCFF, at P3/P4/P5
	Low-rank dimension	24
	Maximum residual ratio ρ	0.10
RT-DETRv2 decoder		
Decoder	Feature levels	3
	Hidden dimension	256
	Decoder layers	6
	Object queries	300
	Attention heads	8
	FFN dimension	1024
	Sampling points	$[4, 4, 4]$
	Query selection	Encoder Top-K, default
Denoising	Number of denoising queries	100
	Label noise ratio	0.5
	Box noise scale	1.0
Post-processing	Maximum predictions	300

We additionally apply two X-specific robustness augmentations during RGB-T training. Random thermal occlusion is activated with probability 0.20. When applied, a rectangular thermal region covering between 2% and 20% of the image area is invalidated, and the corresponding auxiliary validity mask is updated consistently.

Thermal modality dropout is applied with probability 0.05. In this case, the entire thermal tensor and its validity mask are set to zero, and the modality-presence indicator is disabled. Artificial cross-modal affine misalignment is not used in the standard RGB-T experiments.

B.4 Depth Input

For RGB-D experiments, depth is represented as a single-channel input,

$$X_D \in \mathbb{R}^{1\times H\times W}. \tag{B.5}$$

In contrast to thermal preprocessing, the original numerical precision of the depth map is preserved during loading. In particular, higher-precision integer depth representations are not first converted into 8-bit grayscale images.

Depth values are then converted to `float32` using a fixed normalization range:

$$\tilde{D} = \mathrm{clip}\left(\frac{D}{20480}, 0, 1\right). \tag{B.6}$$

This fixed-range normalization is used consistently for both RGB-D benchmarks.

Depth sensing also requires explicit handling of invalid measurements. After numerical conversion, we construct a binary depth-validity mask M_D according to

$$M_D(p) = \begin{cases} 1, & \tilde{D}(p) > 0 \text{ and } \tilde{D}(p) \text{ is finite,} \\ 0, & \text{otherwise.} \end{cases} \tag{B.7}$$

Thus, zero-valued and non-finite measurements are not treated as reliable auxiliary observations.

Depth maps are resized using nearest-neighbor interpolation rather than bilinear interpolation. This avoids synthesizing intermediate depth values across sparse measurements, invalid regions, or depth discontinuities. The corresponding depth-validity mask is resized using the same nearest-neighbor principle.

No artificial affine perturbation, synthetic depth occlusion, or complete depth-modality dropout is applied in the standard RGB-D experiments. Consequently, the auxiliary validity information in RGB-D primarily reflects naturally occurring missing or invalid depth measurements rather than synthetic corruption.

Overall, the three modality settings differ only in the representation and sensor-appropriate preprocessing of the auxiliary input. Event uses a 20-channel histogram representation without image-style normalization, Thermal uses a normalized one-channel intensity representation, and Depth uses a one-channel precision-preserving representation with explicit invalid-measurement handling. The remaining TF-DETR architecture is shared across all three modalities.

C Implementation and Training Details

Unless otherwise specified, all six benchmarks use the same TF-DETR architecture and training configuration. The modality-dependent differences are restricted primarily to the X-input representation described in Appendix B and the dataset-specific batch sizes. Table B.3 summarizes the common model configuration.

C.1 Model Initialization

The RGB detection pathway preserves the RT-DETRv2 R50vd topology, including the RGB backbone, hybrid encoder, CCFF pathway, and transformer decoder. We initialize the shared RGB detection subgraph from the official RT-DETRv2 R50vd pretrained parameters. Parameters belonging

exclusively to TF-DETR, including RepATE, RGB Task probes, task-space alignment, reliability estimation, task-query modulation, and low-rank task refinement, are initialized by their corresponding module definitions.

RepATE is designed as a lightweight task encoder rather than a second semantic backbone. It generates three task maps corresponding to P3, P4, and P5 and converts each map into the explicit Task Factors O, G, U, Z_p, Z_s. No high-dimensional X-semantic feature is forwarded directly to the RT-DETRv2 decoder.

At the lowest-resolution level, global correspondence is first estimated from the P5 Task Factors and is shared across the three scales. Residual local correspondence is subsequently estimated independently at P3–P5. The resulting alignment confidence, auxiliary quality, uncertainty, and cross-modal task discrepancies are used to estimate a reliability gate.

Task information affects the RGB semantic pathway in two constrained locations. First, TQ-AIFI modulates only the attention query at P5, while the keys and values remain derived from the RGB representation. Second, after the original CCFF pathway, aligned task evidence produces a bounded low-rank residual correction at each scale. The correction magnitude is limited to at most $\rho = 0.10$ of the corresponding RGB feature magnitude.

C.2 Training Configuration

The common optimization settings are summarized in Table B.

The detection objective retains the original RT-DETRv2 Varifocal, ℓ_1 bounding-box, and GIoU losses with Hungarian matching. Task-space supervision is applied to the category-agnostic objectness, geometry, and uncertainty-aware geometry predictions. Global correspondence is supervised at the transformation level, while local correspondence is optimized using the expected residual displacement. An additional Null-Match objective supervises locations for which reliable X correspondence is unavailable.

During training, the input resolution is sampled from

$$480, 512, 544, 576, 608, 640, 640, 640, 672, 704, 736, 768, 800, \tag{C.1}$$

with the modality-specific interpolation rules described in Appendix B.3. All experiments use synchronized batch normalization and automatic mixed-precision training.

C.3 RepATE Architecture

RepATE is a lightweight re-parameterizable task encoder used exclusively for the auxiliary modality. Rather than constructing a second full semantic backbone, it progressively extracts multi-scale task maps at P3, P4, and P5, which are subsequently projected into the Task Factors $\{O, G, U, Z_p, Z_s\}$.

The encoder consists of a convolutional stem followed by four stages of TaskRep blocks with channel dimensions $[24, 32, 48, 64, 96]$ and block depths $[2, 2, 2, 1]$. Each TaskRep block adopts a multi-branch structure during training and can be structurally re-parameterized into an equivalent single convolutional branch for deployment. This design retains sufficient local representation capacity for task-factor extraction while keeping the auxiliary pathway lightweight.

For all modalities, the intermediate task projection width is 32. RepATE takes 20-channel Event representations or single-channel Thermal/Depth inputs, while the remaining architecture is shared across sensing settings. Detailed RepATE configurations are summarized in Table B.3.

C.4 Task-Factor and Alignment Supervision Details

This section provides additional details on the supervision used for the Task-Factor representation and Global-Anchored Multi-Scale Task Alignment (GAMTA). The overall objective follows the decomposition introduced in the main paper:

$$\mathcal{L} = \mathcal{L}_{\text{det}} + \lambda_{\text{task}}\mathcal{L}_{\text{task}} + \lambda_{\text{G}}\mathcal{L}_{\text{G}} + \lambda_{\text{L}}\mathcal{L}_{\text{L}} + \lambda_{\text{N}}\mathcal{L}_{\text{N}} + \lambda_{\text{C}}\mathcal{L}_{\text{cons}} + \lambda_{\text{inv}}\mathcal{L}_{\text{inv}}. \tag{C.2}$$

Here, $\mathcal{L}_{\text{det}}$ denotes the original RT-DETRv2 detection objective, including its auxiliary and denoising supervision. The remaining terms provide task grounding, global and local correspondence supervision, invalid-support rejection, alignment consistency, and modality-intervention consistency.

Multi-scale Task-Factor targets. The explicitly grounded Task Factors are supervised at P3–P5. Each ground-truth box $b = (c_x, c_y, w, h)$ is first assigned to one task level according to its pixel-scale size

$$s(b) = \sqrt{(wW)(hH)}, \tag{C.3}$$

where H and W denote the image height and width. We use thresholds of 64 and 192 pixels: objects with $s \leq 64$ are assigned to P3, those with $64 < s \leq 192$ to P4, and the remaining objects to P5.

At the assigned level l, with spatial resolution $H_l \times W_l$, the continuous object center is mapped to

$$u = c_x W_l, \qquad v = c_y H_l, \tag{C.4}$$

and the corresponding discrete task location is $(i, j) = (\lfloor u \rfloor, \lfloor v \rfloor)$. A Gaussian objectness target $O^{l,*}$ is placed at this location, with its radius determined from the object size at the corresponding feature level. The center location is also used to construct the geometry target

$$G^{l,*} = (u - (i + 0.5),\ v - (j + 0.5),\ \log w,\ \log h)\,, \tag{C.5}$$

where the first two terms describe the sub-cell center offset and the last two encode normalized box size.

For the RGB branch, these targets are constructed directly from the detection annotations. For the auxiliary branch, targets are expressed in the corresponding X coordinate system whenever modality-specific box annotations or alignment supervision are available. In particular, explicit X-space boxes are used when provided; otherwise, available global-transform and dense-flow supervision are used to map the RGB annotations into the auxiliary coordinate system before constructing the task targets.

Objectness and uncertainty-aware geometry grounding. Let $p_m^l = \sigma(O_m^l)$ denote the predicted objectness probability for modality $m \in \{R, X\}$. We use a Gaussian heatmap focal objective

$$\ell_{\text{obj}}(p, y) = \begin{cases} -(1-p)^{\alpha_o} \log p, & y = 1, \\ -(1-y)^{\beta_o} p^{\alpha_o} \log(1-p), & y < 1, \end{cases} \tag{C.6}$$

with $\alpha_o = 2$ and $\beta_o = 4$. The loss is evaluated only over valid locations and normalized by the number of positive centers when positives are present.

Geometry is supervised only at positive task locations. The four-dimensional uncertainty factor U_m^l parameterizes a heteroscedastic error scale for the corresponding components of G_m^l. For each geometry component, we use

$$\ell_{\text{geo}} = \exp(-U_m^l)\left|G_m^l - G^{l,*}\right| + U_m^l. \tag{C.7}$$

Thus, large predicted uncertainty reduces the contribution of a large regression residual but is itself penalized by the additive log-variance term. The geometry loss is averaged over positive locations and geometry channels.

The losses are first averaged over the three task levels for each modality. The resulting task-grounding objective is

$$\mathcal{L}_{\text{task}} = \mathcal{L}_{\text{obj}}^X + \frac{1}{2}\mathcal{L}_{\text{obj}}^R + 2\mathcal{L}_{\text{geo}}^X + \mathcal{L}_{\text{geo}}^R. \tag{C.8}$$

Auxiliary-modality losses are additionally masked by modality presence and validity. The latent factors Z_p and Z_s receive no direct regression targets: Z_p is shaped through correspondence-related objectives and downstream detection, whereas Z_s is learned entirely through downstream task optimization.

Alignment-target construction. For the standard experiments, no synthetic X-only geometric perturbation is applied. Each paired sample is therefore initialized with an identity global alignment target and zero residual-flow target,

$$\Theta^* = I, \qquad \mathbf{d}^*(p) = \mathbf{0}. \tag{C.9}$$

Accordingly, $\mathcal{L}_{\mathrm{G}}$ and $\mathcal{L}_{\mathrm{L}}$ regularize the predicted correspondence around the nominal paired geometry. The data pipeline additionally supports controlled X-only geometric perturbations; when enabled, the exact inverse sampling transform and the corresponding dense residual displacement are recorded and used as alignment targets.

Global alignment supervision. GAMTA first estimates a shared similarity transform from P5. Let $\Theta \in \mathbb{R}^{2\times 3}$ denote the predicted affine sampling matrix constructed from translation, rotation, and isotropic scale, and let Θ^* denote the available ground-truth transform. We supervise the predicted global transform using a modality-presence-masked ℓ_1 loss,

$$\mathcal{L}_{\mathrm{G}} = \frac{\sum_b q_b \left\|\Theta_b - \Theta_b^*\right\|_1}{6\sum_b q_b + \epsilon}, \tag{C.10}$$

where $q_b \in \{0, 1\}$ indicates whether the auxiliary modality is present. The resulting transform is shared when globally warping the Task Factors at P3–P5.

Local correspondence supervision. After global warping, the local transport module predicts K residual sampling offsets $\{\mathbf{d}_k^l\}_{k=1}^K$ and their matching weights $\{w_k^l\}_{k=1}^K$ together with an explicit Null Match state. In our experiments, $K = 4$. The expected residual displacement used for supervision is

$$\widehat{\mathbf{d}}^{\,l} = \sum_{k=1}^{K} w_k^l \mathbf{d}_k^l. \tag{C.11}$$

The ground-truth residual flow is resized to the spatial resolution of each task level, with its displacement components scaled consistently with the feature-grid resolution. Local correspondence is then supervised primarily around detection-relevant locations:

$$\mathcal{L}_{\mathrm{L}} = \frac{1}{3}\sum_{l=3}^{5} \frac{\sum_p M_R^l(p) \left\|\widehat{\mathbf{d}}^{\,l}(p) - \mathbf{d}^{l,*}(p)\right\|_1}{\sum_p M_R^l(p) + \epsilon}, \tag{C.12}$$

where M_R^l is the RGB objectness target heatmap at level l. This concentrates local correspondence supervision on regions relevant to object detection rather than uniformly weighting the complete image plane.

Null-Match supervision. The Null Match state is used to reject locations for which the candidate sampling set provides insufficient valid auxiliary support. Let $v_k^l(p) \in [0, 1]$ denote the validity of the k-th sampled auxiliary location. The rejection target is constructed as

$$y_{\mathrm{N}}^l(p) = 1 - \max_k v_k^l(p), \tag{C.13}$$

so the target approaches one when no candidate provides valid support and approaches zero when at least one valid correspondence is available. The predicted Null probability p_{N}^l is optimized with binary cross-entropy:

$$\mathcal{L}_{\mathrm{N}} = \frac{1}{3}\sum_{l=3}^{5} \mathrm{BCE}\left(p_{\mathrm{N}}^l, y_{\mathrm{N}}^l\right). \tag{C.14}$$

This objective gives the local transport module an explicit rejection mechanism rather than forcing every RGB location to select an auxiliary correspondence.

Alignment consistency. When a consistency view of the auxiliary observation is provided during training, both views are independently passed through the same task-space alignment pipeline. Let $\widetilde{T}_X^l$ and $\widetilde{T}_{X,c}^l$ denote the resulting aligned task states. Consistency is imposed on objectness, geometry, and the positional latent. For Z_p, we first apply channel-wise ℓ_2 normalization,

$$\bar{Z}_p = \frac{Z_p}{\|Z_p\|_2 + \epsilon}, \tag{C.15}$$

and define the per-location discrepancy

$$D^l_{\text{cons}} = \left|\sigma(\widetilde{O}^l_X) - \sigma(\widetilde{O}^l_{X,c})\right| + \frac{1}{4}\left\|\widetilde{G}^l_X - \widetilde{G}^l_{X,c}\right\|_1 + \frac{1}{d_p}\left\|\bar{Z}^l_{p,X} - \bar{Z}^l_{p,X,c}\right\|_1, \tag{C.16}$$

where $d_p = 8$. The discrepancy is averaged over the intersection of the two valid-support masks and over the three task levels. We do not directly impose this consistency constraint on Z_s, allowing the residual latent to retain sensor- and perturbation-dependent task evidence.

Modality-intervention consistency. A separate intervention objective is applied when an intervened auxiliary view is available. In contrast to alignment consistency, this term constrains only the explicitly grounded objectness and geometry states after task-space alignment. Let $\widetilde{T}^l_X$ and $\widetilde{T}^l_{X,\text{int}}$ denote the aligned original and intervened states. Their discrepancy is

$$D^l_{\text{inv}} = \left|\sigma(\widetilde{O}^l_X) - \sigma(\widetilde{O}^l_{X,\text{int}})\right| + \lambda^{\text{inv}}_G \frac{1}{4}\left\|\widetilde{G}^l_X - \widetilde{G}^l_{X,\text{int}}\right\|_1. \tag{C.17}$$

The loss is evaluated only over foreground locations using the RGB objectness target together with the intersection of the two validity masks:

$$\mathcal{L}_{\text{inv}} = \frac{1}{3}\sum_{l=3}^{5} \frac{\sum_p M^l_{\text{fg}}(p) D^l_{\text{inv}}(p)}{\sum_p M^l_{\text{fg}}(p) + \epsilon}. \tag{C.18}$$

Here, M^l_{fg} combines the RGB target heatmap, auxiliary validity, and modality-presence mask. We use $\lambda^{\text{inv}}_G = 1$. Neither Z_p nor Z_s is directly constrained by this intervention objective.

Loss weights. Unless otherwise specified, the same loss weighting is used across all six benchmarks. The standard RT-DETRv2 VFL, ℓ_1, and GIoU terms use weights 1.0, 5.0, and 2.0, respectively. The TF-DETR-specific terms use

$$\lambda_{\text{task}} = 1.0, \qquad \lambda_{\text{G}} = 1.0, \qquad \lambda_{\text{L}} = 0.5, \qquad \lambda_{\text{N}} = 0.2, \qquad \lambda_{\text{C}} = 0.2, \qquad \lambda_{\text{inv}} = 0.2. \tag{C.19}$$

D Detailed Efficiency Measurement Protocol

We evaluate inference efficiency under two complementary settings: native-framework execution in PyTorch and deployment-oriented inference with TensorRT. The two protocols are reported separately because they characterize different aspects of runtime efficiency.

Native-framework inference. We first evaluate architecture-level inference efficiency in a unified PyTorch environment. All models are benchmarked on a single NVIDIA RTX 4090 GPU using paired RGB-X inputs, with both modalities resized to 640×640 pixels and a batch size of 1. Inference is performed with PyTorch 2.1.2 using `torch.compile` with the TorchInductor backend. Computation is conducted in FP32 with TF32 enabled.

All input tensors are pre-allocated on the GPU before timing to eliminate host-to-device transfer overhead. After 10 warm-up iterations, we measure 100 consecutive forward passes with explicit CUDA synchronization before and after each inference call. The complete benchmark is repeated independently three times. The reported runtime measures only model forward execution and excludes disk I/O, image decoding, resizing, normalization, host-to-device data transfer, and post-processing.

TensorRT deployment inference. We further evaluate deployment-oriented inference efficiency on the same NVIDIA RTX 4090 GPU using TensorRT 10.4 with FP16 optimization. The deployable model is first exported to ONNX and subsequently compiled into a TensorRT FP16 engine. All measurements use paired RGB-X inputs at 640×640 resolution with a batch size of 1.

For this evaluation, we use a fixed set of 1,000 real paired multimodal samples. All samples are decoded and prepared before timing begins. Following 50 warm-up iterations, engine execution is

measured over all 1,000 samples using CUDA events, and the complete experiment is repeated three times. Timing includes only TensorRT engine execution and excludes disk access, image decoding, input preprocessing, host-to-device transfer, engine construction, and post-processing.

Average per-sample latency is converted to throughput as

$$\mathrm{FPS} = \frac{1000}{\mathrm{Latency\ (ms)}}. \tag{D.1}$$

The two protocols characterize different deployment regimes. The PyTorch benchmark reflects runtime efficiency under the native training and evaluation framework, whereas the TensorRT benchmark measures throughput after deployment-specific graph and precision optimization. We therefore report PyTorch and TensorRT throughput separately and do not combine results from the two backends when comparing model efficiency.

E Cross-Sensor Qualitative Analysis of Task Factors

To examine how the proposed Task-Factor representation behaves across heterogeneous auxiliary sensors, we visualize the learned factors on all six benchmarks. We focus on their functional spatial patterns rather than assigning fixed visual semantics to the latent factors. Specifically, O_X, G_X, and U_X denote the explicitly grounded objectness, geometry, and geometric uncertainty, respectively, while Z_p and Z_s are learned latent factors associated with correspondence-related positional information and residual task-related structure.

The geometry boxes visualized in the corresponding panels are decoded from selected positive geometry probes and therefore are not expected to correspond one-to-one with the final decoder predictions. For RGB-D examples, where X denotes the depth modality, the depth-derived factor responses are overlaid on the spatially aligned RGB image only for visualization clarity, while the original depth observation is shown separately.

M3FD (RGB-Thermal). As shown in Fig. E.2, the thermal modality retains informative foreground cues in scenes where visible appearance is weakened by low illumination or other appearance degradation. The objectness factor O_X produces localized responses around detection-relevant regions such as vehicles and pedestrians, while G_X and U_X provide local geometric states with different levels of predicted uncertainty across targets. Meanwhile, Z_p exhibits a spatially organized, correspondence-related positional response around objects and their neighboring regions, whereas Z_s emphasizes additional residual structures associated with the detection task. These patterns indicate that the thermal observation is decomposed into factors with different functional roles rather than represented by a single undifferentiated auxiliary feature.

DroneVehicle (RGB-Thermal). As shown in Fig. E.1, DroneVehicle presents a substantially different RGB-Thermal setting characterized by aerial viewpoints, small targets, and densely distributed vehicles. Despite this change in scene geometry, O_X still produces localized responses around numerous vehicle instances, while G_X and U_X describe target-dependent geometric states and their uncertainty. In dense regions, Z_p exhibits a more continuous correspondence-related positional response that reflects the spatial organization of nearby targets, whereas Z_s captures broader residual task-related structures around vehicle groups and their local context. Together with the M3FD results, these visualizations suggest that the functional roles of the Task Factors remain qualitatively consistent from ground-level scenes to dense aerial scenes, although their detailed spatial patterns adapt to the imaging configuration.

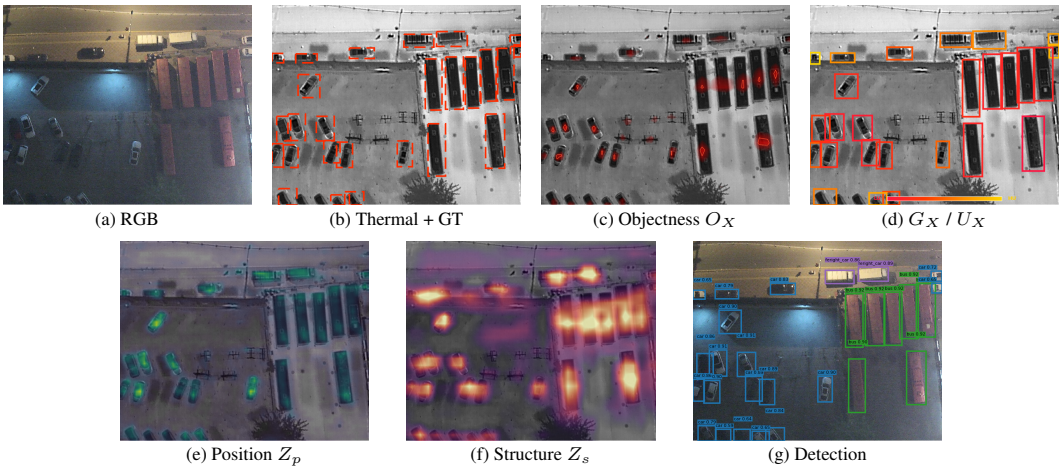
DSEC-Detection (RGB-Event). As shown in Fig. E.3, the Event modality is characterized by sparse brightness-change and event-edge responses rather than complete appearance information. Nevertheless, O_X remains concentrated around vehicles, pedestrians, and riders, suggesting that the learned objectness state is not simply a direct visualization of raw event activity but exhibits a more task-oriented spatial response. The latent factor Z_p shows a spatially organized, correspondence-related positional response around object neighborhoods and event-active scene structures, while Z_s emphasizes residual structures associated with brightness changes and local event patterns. These observations are consistent with the intended role of the compact Event pathway in supplying complementary task-oriented evidence.

PKU-DAVIS-SOD (RGB-Event). As shown in Fig. E.4, the learned Task Factors remain functionally differentiated across low-light, motion-blurred, and normally exposed examples. When the RGB appearance becomes degraded, the Event observation still contains brightness-change boundaries and dynamic structural cues. Correspondingly, O_X retains localized responses around detection-relevant regions, while Z_p exhibits a correspondence-related positional response associated with target locations and local scene organization. The residual factor Z_s further highlights task-related structures associated with locally active Event regions. Similar functional patterns are also observed under normal appearance conditions, suggesting that the Event Task Factors retain complementary structural cues under degraded RGB appearance without changing their factorized roles.

KITTI (RGB-Depth). As shown in Fig. E.6, the depth-derived Task Factors exhibit a noticeably different spatial pattern because the auxiliary observation is sparse and strongly constrained by the distribution of valid depth measurements. O_X still produces responses around a subset of vehicles and their neighboring regions, but these responses are generally more spatially sparse than those observed for Thermal inputs. The latent factor Z_p exhibits a correspondence-related positional response influenced by road geometry and the distribution of valid depth samples, while Z_s forms more regional and directional residual task-related responses rather than isolated instance-wise activations. The difference between the intermediate geometry probes and the final detections is particularly visible on KITTI because of the sparsity of the depth observation. Despite this sparsity, the factors retain qualitatively distinct roles in objectness, geometry, positional correspondence, and residual task structure.

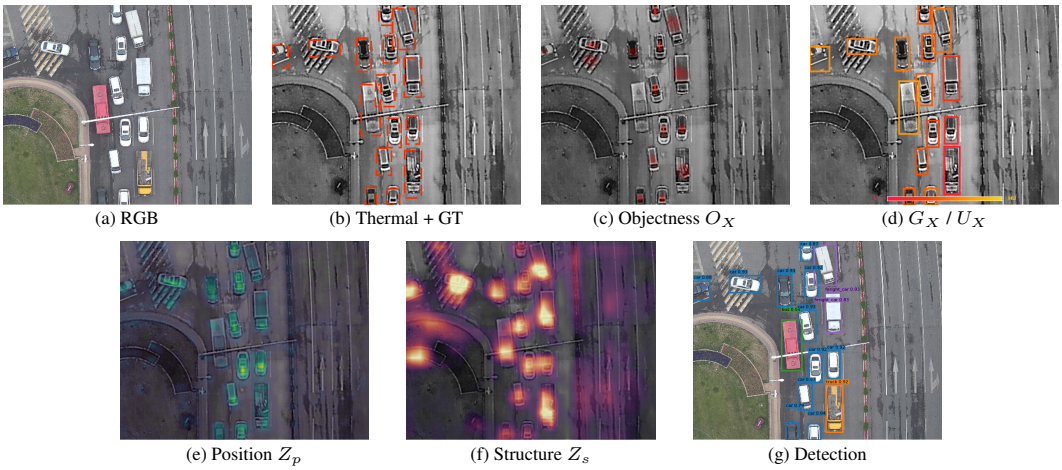
SUN RGB-D (RGB-Depth). As shown in Fig. E.5, the denser indoor depth observations lead to spatial responses that differ noticeably from those on KITTI. O_X primarily responds around foreground objects such as furniture and other indoor instances, whereas Z_p exhibits broader correspondence-related positional responses along object boundaries, furniture groups, and locally coherent depth regions. Similarly, Z_s forms region-level residual task-related responses rather than isolated peaks associated with individual detections. Such patterns are consistent with the geometric nature of dense depth observations and need not correspond one-to-one with individual object instances. In particular, Z_s should be interpreted as a residual task latent that preserves task-related information beyond the explicitly grounded factors, rather than as an instance-wise semantic activation map. These results illustrate how the same Task-Factor formulation can retain its functional organization while adapting its spatial expression to dense scene geometry.

Cross-sensor observations. Across RGB-Thermal, RGB-Event, and RGB-Depth settings, the learned Task Factors exhibit a qualitatively consistent functional organization despite the substantial differences in the physical properties of the auxiliary modalities. O_X primarily provides localized object-presence evidence; G_X describes intermediate object geometry and U_X its associated uncertainty; Z_p exhibits spatially organized, correspondence-related positional responses; and Z_s captures residual task-related structure beyond the explicitly grounded factors. At the same time, their detailed spatial patterns remain sensor-dependent. Thermal-derived factors exhibit localized foreground and structural responses; Event-derived factors emphasize sparse brightness-change boundaries and dynamic structures; and Depth-derived factors are shaped more strongly by scene geometry and measurement density. These qualitative results are therefore consistent with the central motivation of Task-Factor representation: heterogeneous auxiliary modalities need not be forced into an identical full semantic representation, but can instead communicate with the detector through a compact task-oriented interface whose functional roles remain shared while its latent evidence adapts to the sensor.



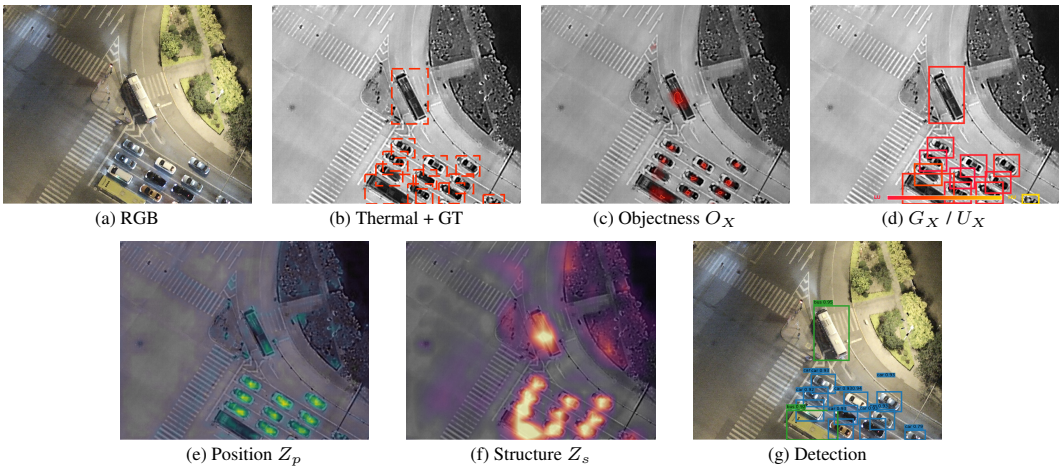
(a) RGB (b) Thermal + GT (c) Objectness O_X (d) G_X / U_X

(e) Position Z_p (f) Structure Z_s (g) Detection



(a) RGB (b) Thermal + GT (c) Objectness O_X (d) G_X / U_X

(e) Position Z_p (f) Structure Z_s (g) Detection



(a) RGB (b) Thermal + GT (c) Objectness O_X (d) G_X / U_X

(e) Position Z_p (f) Structure Z_s (g) Detection

Figure E.1: **Qualitative Task-Factor visualizations on the DroneVehicle dataset.**For each sample: (a) input RGB image; (b) auxiliary-modality observation with ground-truth bounding boxes; (c) objectness response O_X; (d) local geometry G_X and its predicted uncertainty U_X; (e) correspondence-related positional latent Z_p; (f) residual task-related structural latent Z_s; and (g) final detection predictions. O_X highlights detection-relevant object regions, while G_X/U_X characterize intermediate object geometry and its uncertainty. Geometry boxes are color-coded from red (low uncertainty, LU), through orange, to yellow (high uncertainty, HU).

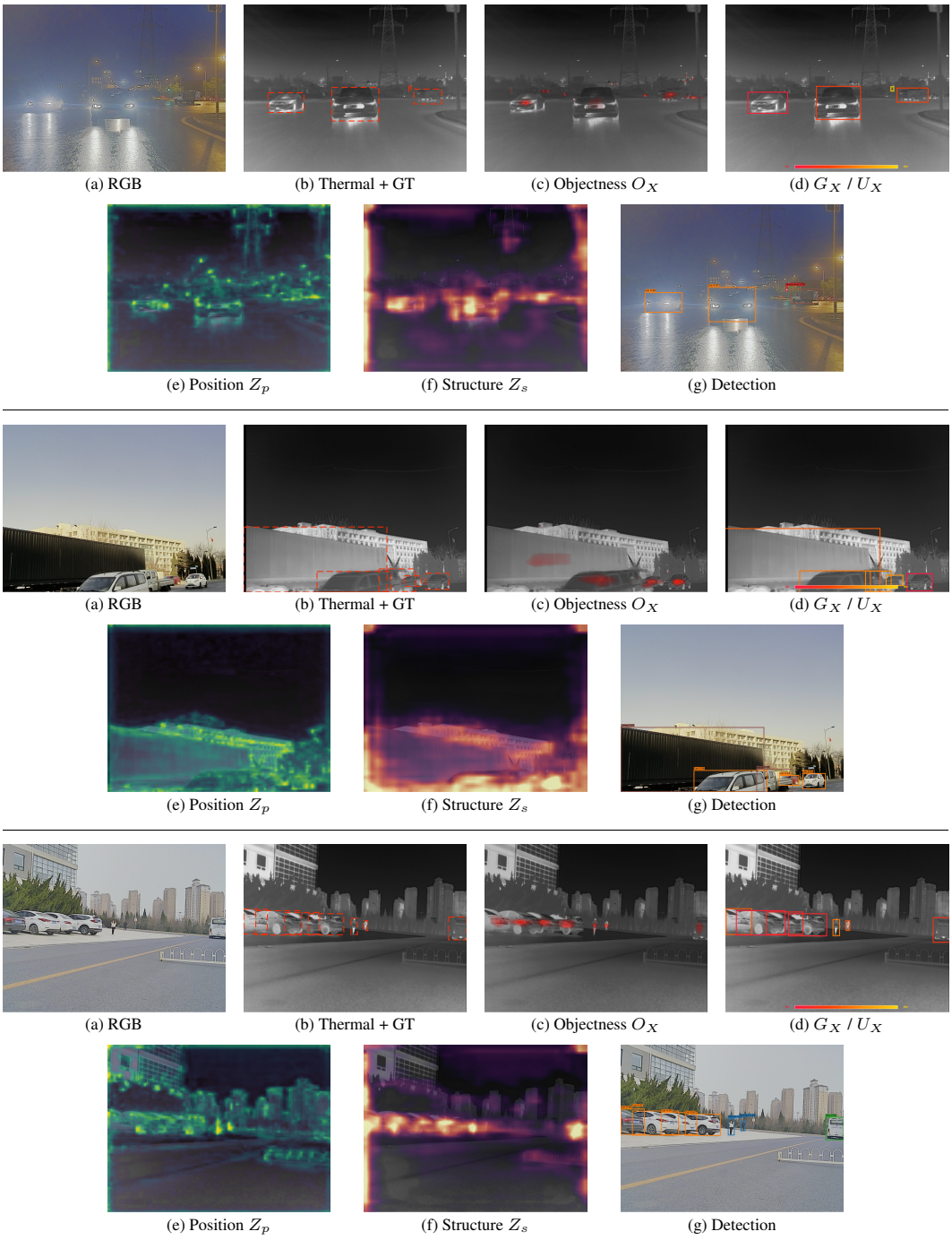


Figure E.2: **Qualitative Task-Factor visualizations on the M3FD benchmark.** For each sample: (a) input RGB image; (b) auxiliary-modality observation with ground-truth bounding boxes; (c) objectness response O_X; (d) local geometry G_X and its predicted uncertainty U_X; (e) correspondence-related positional latent Z_p; (f) residual task-related structural latent Z_s; and (g) final detection predictions. O_X highlights detection-relevant object regions, while G_X/U_X characterize intermediate object geometry and its uncertainty. Geometry boxes are color-coded from red (low uncertainty, LU), through orange, to yellow (high uncertainty, HU).

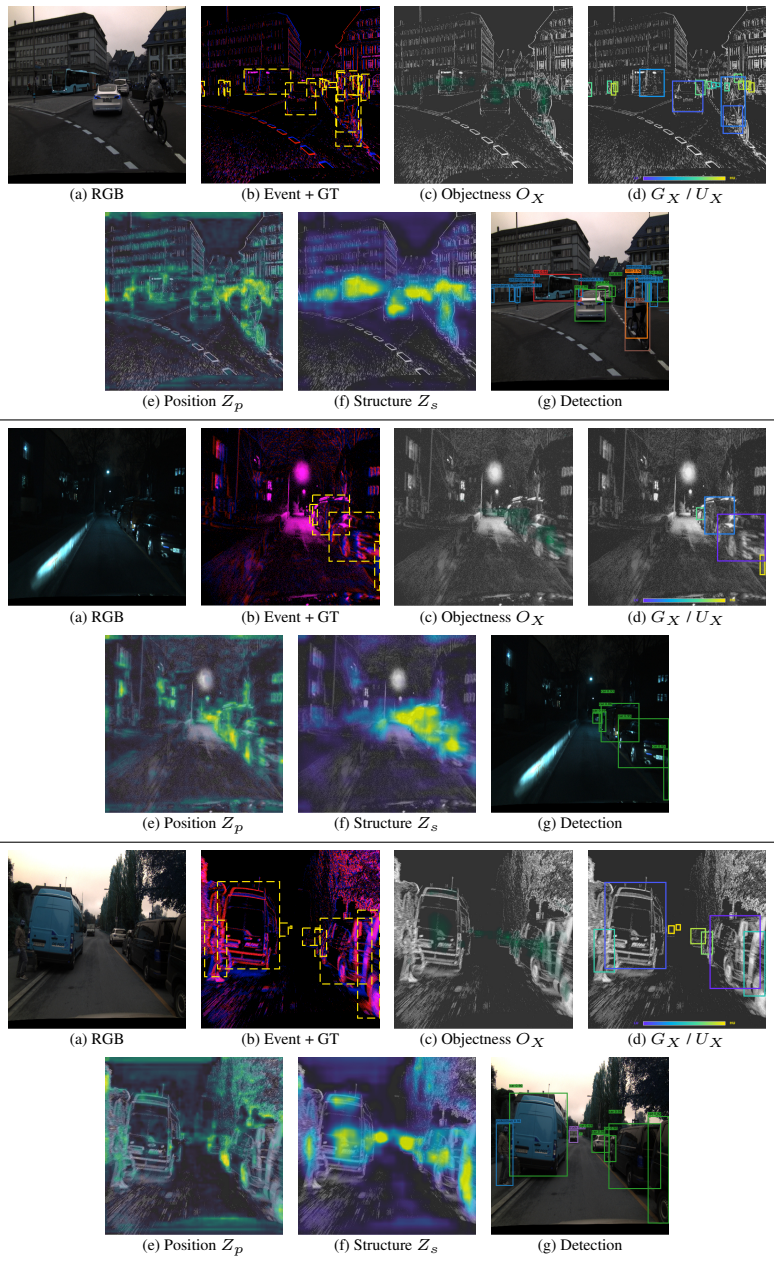


Figure E.3: **Qualitative Task-Factor visualizations on the DSEC-Detection benchmark.** For each sample: (a) input RGB image; (b) auxiliary-modality observation with ground-truth bounding boxes; (c) objectness response O_X; (d) local geometry G_X and its predicted uncertainty U_X; (e) correspondence-related positional latent Z_p; (f) residual task-related structural latent Z_s; and (g) final detection predictions. O_X highlights detection-relevant object regions, while G_X/U_X characterize intermediate object geometry and its uncertainty. Geometry boxes are color-coded from red (low uncertainty, LU), through orange, to yellow (high uncertainty, HU).

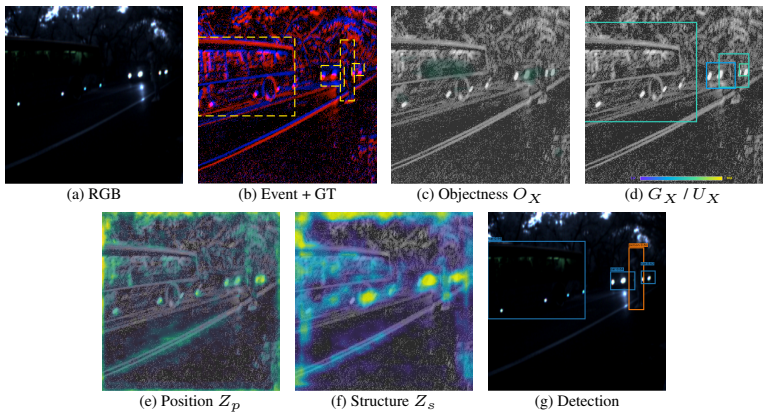


Figure E.4: **Qualitative Task-Factor visualizations on the PKU-DAVIS-SOD benchmark.** For each sample: (a) input RGB image; (b) auxiliary-modality observation with ground-truth bounding boxes; (c) objectness response O_X; (d) local geometry G_X and its predicted uncertainty U_X; (e) correspondence-related positional latent Z_p; (f) residual task-related structural latent Z_s; and (g) final detection predictions. O_X highlights detection-relevant object regions, while G_X/U_X characterize intermediate object geometry and its uncertainty. Geometry boxes are color-coded from red (low uncertainty, LU), through orange, to yellow (high uncertainty, HU).

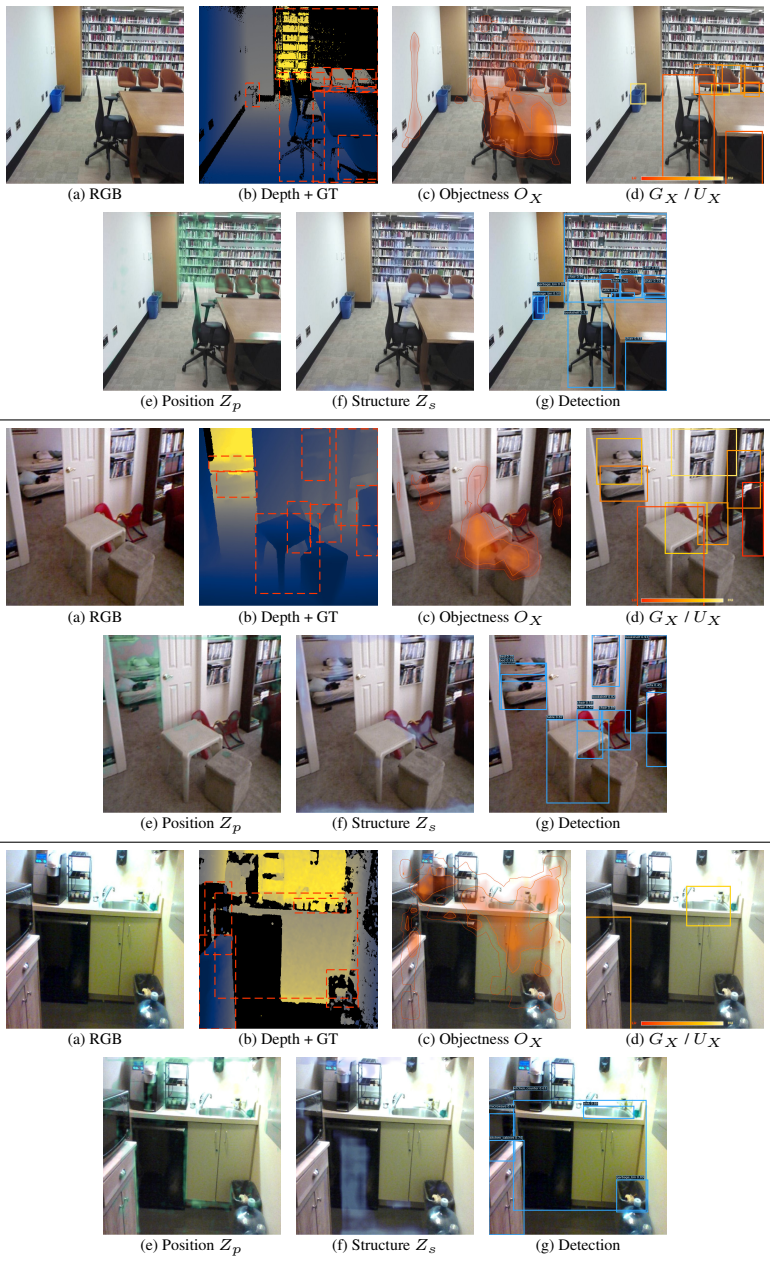


Figure E.5: **Qualitative Task-Factor visualizations on the SUN RGB-D benchmark.** For each sample: (a) input RGB image; (b) auxiliary-modality observation with ground-truth bounding boxes; (c) objectness response O_X; (d) local geometry G_X and its predicted uncertainty U_X; (e) correspondence-related positional latent Z_p; (f) residual task-related structural latent Z_s; and (g) final detection predictions. O_X highlights detection-relevant object regions, while G_X/U_X characterize intermediate object geometry and its uncertainty. Geometry boxes are color-coded from red (low uncertainty, LU), through orange, to yellow (high uncertainty, HU).

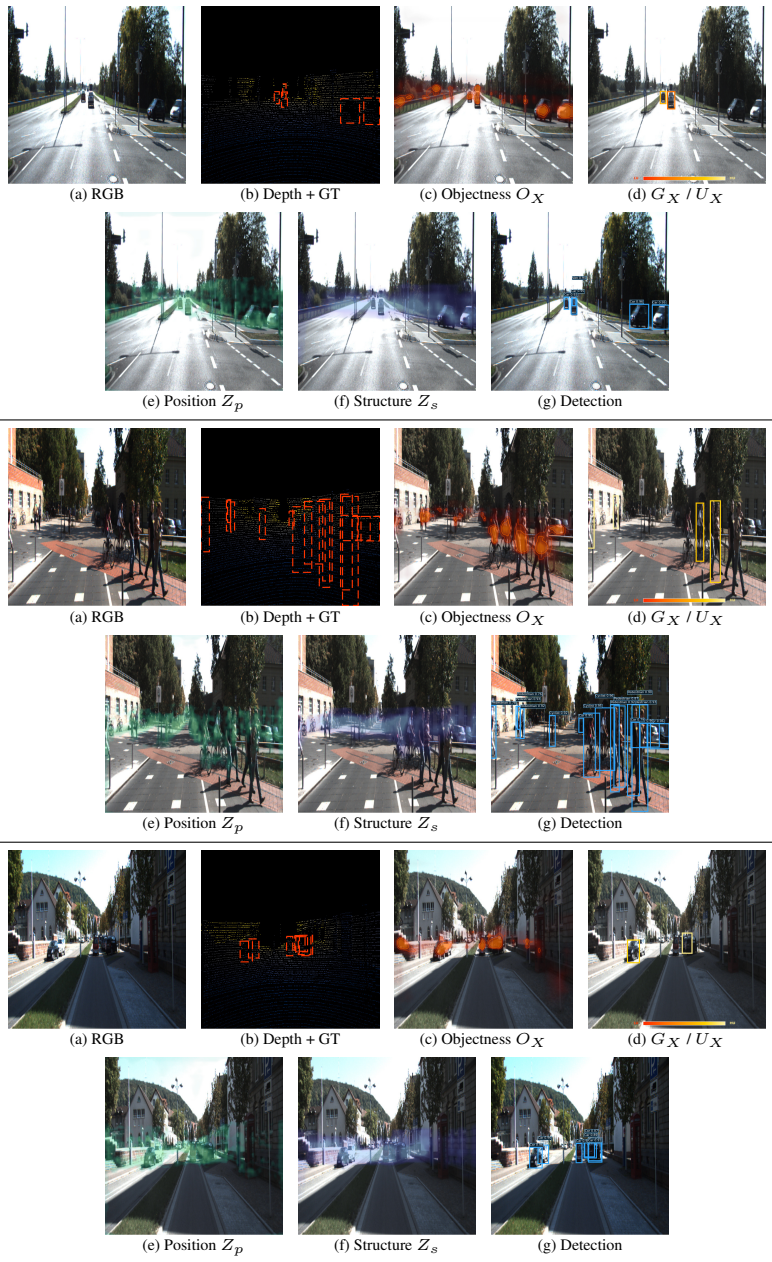


Figure E.6: **Qualitative Task-Factor visualizations on the KITTI benchmark.** For each sample: (a) input RGB image; (b) auxiliary-modality observation with ground-truth bounding boxes; (c) objectness response O_X; (d) local geometry G_X and its predicted uncertainty U_X; (e) correspondence-related positional latent Z_p; (f) residual task-related structural latent Z_s; and (g) final detection predictions. O_X highlights detection-relevant object regions, while G_X/U_X characterize intermediate object geometry and its uncertainty. Geometry boxes are color-coded from red (low uncertainty, LU), through orange, to yellow (high uncertainty, HU).