中图法分类号:TP391.41　　**文献标识码**:A　　**文章编号**:1006-8961(XXXX)XX-0001-16

论文引用格式:Li Jiandong, Kang Xuhe, Qu Haicheng, Wu Cong. Robust visible-infrared object detection with incomplete modalities[J/OL]. Journal of Image and Graphics, XXXX:1-16. DOI: 10.11834/jig.260530. (李建东, 康栩赫, 曲海成, 武聪. 模态缺失条件下的可见光-红外鲁棒目标检测[J/OL]. 中国图象图形学报, XXXX:1-16. DOI: 10.11834/jig.260530.)[DOI:10.11834/jig.260530]

模态缺失条件下的可见光-红外鲁棒目标检测

李建东[1,2],康栩赫[1*],曲海成[1],武聪[1,3]

1. 辽宁工程技术大学　智能科学与工程学院,辽宁　葫芦岛　125105; 2. 辽宁工程技术大学　矿业学院,辽宁　阜新　123000; 3. 辽宁工程技术大学　创新实践学院,辽宁　阜新　123000

摘　要:**目的**　可见光-红外(RGB-T)目标检测依赖两路传感器的互补信息,但可能因传感器失效、遮挡或通信中断等导致单路模态暂不可用,进而引发完整双模态检测器出现特征污染和检测性能下降。为此,本文提出一种面向模态缺失的三分支受保护主路径检测器TPB-DETR(Tri-branch Protected-Base DETR)。**方法**　TPB-DETR以分支保护融合编码器(Branch-Protected Fusion Encoder, BPF Encoder)为核心,结合缺失辅助上下文补偿(Missing-Auxiliary Context Compensation, MACC)和三分支课程训练(Three-Branch Curriculum Training, TBCT)进行检测。BPF Encoder根据RGB+IR、RGB-only和IR-only三种分支状态选择真实可靠特征作为受保护主路径,并将辅助路径信息限制为门控残差更新和高层辅助注意力,降低强制融合对主路径表征的污染;MACC从真实可用分支生成伪辅助上下文,避免直接使用零占位缺失分支;TBCT在同一检测器中交替采样并优化完整RGB+IR、RGB-only和IR-only三种状态,动态课程权重作为其内部更新频率调度机制。**结果**　在FLIR数据集上,TPB-DETR在RGB+IR、RGB-only和IR-only输入下分别取得47.32、34.99和46.17 mAP,缺失平均mAP为40.58;在LLVIP数据集上,三种输入状态分别取得70.30、53.71和69.25 mAP,缺失平均mAP为61.48。FLIR消融结果显示,保留MACC与相同TBCT,将BPF Encoder替换为GM-DETR式MSFI/CMSF融合后,缺失平均mAP由40.58降至38.86。LLVIP上的对应消融进一步验证了BPF Encoder的有效性。**结论**　本文方法在完整输入精度与缺失输入鲁棒性之间取得稳定平衡,验证了受保护主路径、伪辅助上下文和三分支训练调度提升可见光-红外检测器在单路模态不可用条件下的部署适应性。

关键词:可见光-红外目标检测;模态缺失;保护融合;分支状态感知;三分支课程训练

Robust visible-infrared object detection with incomplete modalities

Li Jiandong[1,2], Kang Xuhe[1*], Qu Haicheng[1], Wu Cong[1,3]

1. School of Intelligent Science and Engineering, Liaoning Technical University, Huludao 125105, Liaoning, China; 2. College of Mining, Liaoning Technical University, Fuxin 123000, Liaoning, China; 3. College of Innovation Practice, Liaoning Technical University, Fuxin 123000, Liaoning, China

Abstract: **Objective** Visible-infrared (RGB-T) object detection combines the texture, color, edge, and scene cues of visible imagery with the illumination-insensitive thermal responses of infrared imagery. This complementarity is useful for traffic monitoring, nighttime surveillance, and mobile perception, but most fusion detectors are developed for a complete and reliable RGB+IR pair. Their encoders generally perform symmetric or nearly equal-weight interaction through addition, concatenation, bidirectional attention, or unrestricted cross-modal updates. Such a design implicitly assumes that both

收稿日期:2026-08-23;**修回日期**:2026-10-31

*通信作者:康栩赫,通信作者,男,硕士研究生,主要研究方向为目标检测与识别。E-mail:2124287010@qq.com

基金项目:辽宁省教育厅基本科研项目(LJ242610147001,LJ242610147002)

Supported by: Basic Scientific Research Project of the Liaoning Provincial Department of Education (LJ242610147001, LJ242610147002)

© 中国图象图形学报版权所有

streams are simultaneously available and similarly trustworthy. When one sensor is unavailable because of hardware failure, occlusion, exposure breakdown, synchronization error, transmission loss, or bandwidth restriction, a zero-filled, copied, or generated substitute may still be treated as valid evidence. The unreliable branch can then overwrite or contaminate the representation extracted from the truly available sensor. Therefore, detection with incomplete modalities is not merely an input-replacement or augmentation problem; it reveals a structural weakness in the conventional symmetric fusion paradigm. This study aims to establish a structurally protected fusion formulation in which branch availability determines the roles and write permissions of the two feature paths. The goal is to use one detector for complete RGB+IR, RGB-only, and IR-only operation while preserving competitive complete-modality accuracy and improving the lower-bound performance under single-sensor input. **Method** A unified detector, Tri-branch Protected-Base DETR (TPB-DETR), is constructed on an RT-DETR-style framework with a PResNetX-50 feature extractor. Given the known branch-state identifier, the network activates only the sensor inputs declared available and assigns explicit roles before fusion. TPB-DETR contains three coordinated components, each addressing a different source of degradation. First, the Branch-Protected Fusion Encoder (BPF Encoder) reformulates cross-modal fusion as an asymmetric base-and-auxiliary update. The reliable real feature is designated as the protected base path, whereas the other stream is restricted to an auxiliary path. Under complete RGB+IR input, infrared features are used as the protected base according to their relative stability in the targeted low-light setting, and real RGB features supply complementary appearance details. Under RGB-only input, real RGB features become the base and a pseudo-infrared context becomes the auxiliary input; under IR-only input, real infrared features remain the base and a pseudo-RGB context becomes auxiliary. At every fused scale, the base feature is retained through an identity residual with coefficient one. Auxiliary information can enter only through spatial and channel gates, which regulate where and how strongly auxiliary residuals are written, and through auxiliary cross-attention limited to high-level semantic scales. The auxiliary path can thus enrich the base representation but cannot directly replace it. This role assignment separates the reliability of real and pseudo features and prevents missing-side artifacts from being amplified by an otherwise symmetric fusion block. Second, Missing-Auxiliary Context Compensation (MACC) supplies the BPF Encoder with bounded task-oriented context when a sensor is missing. MACC maps the real available feature to a weak pseudo auxiliary representation through a lightweight residual transformation with learnable scaling and bias. It is designed as a task-oriented context generator rather than an image reconstruction module, and its output does not represent a recovered missing modality. Its function is to preserve a unified fusion interface and provide limited contextual hints instead of a pure zero placeholder. Because the pseudo representation remains on the auxiliary path, its influence is controlled by the same gates and residual coefficients used by the BPF Encoder. Third, Three-Branch Curriculum Training (TBCT) optimizes the shared model across the three deployment states by scheduling their update frequencies over 25 epochs. The early stage emphasizes RGB-only and IR-only samples to establish usable single-sensor branches; the middle stage balances the three states; and the late stage increases complete RGB+IR sampling to recover full-modality accuracy. This curriculum coordinates robustness and accuracy without defining separate models or allowing losses with different scales to compete within a simultaneous three-branch forward pass. Together, BPF Encoder, MACC, and TBCT implement structural protection, controlled context compensation, and state-balanced optimization, respectively. **Result** TPB-DETR is evaluated on the FLIR-aligned traffic dataset and the LLVIP low-light pedestrian dataset using one parameter set for all branch states. Average mAP and mAP50 are the arithmetic means of the corresponding RGB-only and IR-only results, summarizing robustness under either sensor loss. On FLIR, complete RGB+IR input achieves 47.32 mAP and 84.33 mAP50. RGB-only and IR-only inputs achieve 34.99 and 46.17 mAP, respectively, yielding an average mAP of 40.58; their corresponding mAP50 values are 71.68 and 82.80, with an average mAP50 of 77.24. On LLVIP, complete RGB+IR input reaches 70.30 mAP and 98.14 mAP50. The RGB-only and IR-only branches reach 53.71 and 69.25 mAP, producing an average mAP of 61.48; their mAP50 values are 91.52 and 97.70, and the average mAP50 is 94.61. The stronger IR-only results on both datasets are consistent with the thermal modality′s advantage in low-illumination scenes, while the RGB-only branch remains usable without a separate visible-only detector. A controlled FLIR ablation isolates the value of protected fusion. When the BPF Encoder is replaced by a GM-DETR-style MSFI/CMSF fusion module while MACC, three-branch training, and the same TBCT schedule are retained, average mAP decreases from 40.58 to 38.86. The 1.72-point difference is the independent

© 中国图象图形学报版权所有

improvement associated with the protected fusion structure under matched training conditions; RGB-only mAP falls from 34.99 to 32.44 and IR-only mAP from 46.17 to 45.28. Removing MACC reduces average mAP to 40.16, indicating that controlled weak context is preferable to a zero missing-side placeholder. Retaining three-branch training but removing the dynamic curriculum lowers average mAP to 39.35 and complete RGB+IR mAP to 46.86, showing that staged sampling helps balance robustness and full-input performance. These results distinguish the structural effect of BPF Encoder from the benefit of merely exposing the detector to samples with incomplete modalities. The corresponding ablation on LLVIP further confirms the effectiveness of the BPF Encoder. **Conclusion** TPB-DETR reframes RGB-T detection with incomplete modalities as a branch-state-conditioned structural fusion problem. Its central contribution is an asymmetric protected-fusion paradigm: reliable real features retain an uninterrupted base path, whereas real or pseudo auxiliary information receives only bounded write access. MACC and TBCT support this principle by providing controlled context and coordinating the three deployment states. The experiments demonstrate that a single detector can preserve competitive complete-modality performance and remain effective when either sensor is absent. More broadly, the results show that reliability-aware role assignment is a stronger basis for robust multimodal fusion than treating two streams as permanently equivalent.

Key words: visible-infrared object detection; incomplete modalities; protected fusion; branch-state awareness; three-branch curriculum training

论文引用格式：DOI:10.11834/jig.260530

0 引　言

可见光图像提供纹理、颜色与场景语义，红外图像在夜间和弱光环境下保持较稳定的热辐射响应。RGB-T目标检测利用两者的互补信息，服务于交通与安防监控。现有方法通过特征交互、跨模态注意力和动态特征选择提升完整双模态检测性能（Xiao等，2024；Fu等，2024b；许可等，2024）。实际部署中，传感器故障、遮挡或通信中断可能使单路模态暂不可用。此时，缺失侧的零填充、复制或生成特征若仍按真实模态参与融合，可能干扰另一侧有效表征。

RGB-T检测包括早期、中期和后期融合（别倩等，2023）。早期融合拼接输入或生成融合图像，结构紧凑，但易忽略模态差异（唐霖峰等，2023；孙旭辉等，2023）；中期融合在特征层交换互补信息，兼顾模态特异性与互补语义；后期融合组合检测框或置信度，便于独立调整分支，但深层交互不足（张荣等，2026）。

在特征级融合方面，GM-DETR通过模态内特定特征交互（modality-specific feature interaction，MSFI）与跨模态跨尺度特征融合（cross-modality-scale feature fusion，CMSF）编码多光谱特征；CAFF-DINO和ICAFusion分别通过交叉注意力和迭代交互增强两路信息交换（Xiao等，2024；Helvig等，2024；Shen等，2024）。CF-Deformable DETR利用跨模态可变形注意力适应弱配准输入（Fu等，2024b）。UniRGB-IR、红外中心传感器融合Transformer和FD2-Net则分别通过基础模型适配、红外引导和频域分解改善表征（Yuan等，2025；Hwang等，2025；Li等，2025）。这些设计主要面向完整输入，单路缺失时仍需区分真实与替代特征的融合角色。

针对模态缺失，已有研究主要包括模态丢弃与模态无关学习。模态丢弃在训练或评估时遮蔽一路输入，使模型接触单模态状态。GM-DETR分析了单模态与双模态输入的泛化差异（Xiao等，2024）。模态无关学习降低对固定模态组合的依赖，MiPa通过混合patch和模态无关模块学习共享表征（Medeiros等，2025），AUNet验证模态可用性并自适应启停交互机制（Bie等，2026）。

模态缺失下的可靠检测需协调三个层面：结构上，将模态可用性落实为主、辅路径的角色划分与更新规则；特征上，区分真实可用特征与生成辅助上下文，避免不可靠信息干扰主路径；训练上，为缺失分支提供持续、均衡的监督，兼顾完整输入精度与单路输入鲁棒性。三者相互关联，需要在同一检测器中协同设计。

针对上述三个层面的问题，本文提出三分支受保护主路径检测器（Tri-branch Protected-Base DETR，TPB-DETR），以分支保护融合编码器（Branch-Protected Fusion Encoder，BPF Encoder）为核心，结合缺失辅助上下文补偿（Missing-Auxiliary Context Compensation，MACC）与三分支课程训练

© 中国图象图形学报版权所有

(Three-Branch Curriculum Training, TBCT),统一支持 RGB+IR、RGB-only 和 IR-only 检测。主要工作如下:

1) 提出分支保护融合编码器(BPF Encoder)。该模块依据输入状态划分主路径(base)与辅助路径(aux),通过受保护融合块(Protected Fusion Block, PFB)保留真实主路径的恒等通路。空间、通道门控及高层注意力调节辅助残差,使真实与伪辅助特征均按主辅角色参与融合,并在三种输入状态下保持相同的保护规则。

2) 构建缺失辅助上下文补偿(MACC)。MACC 从真实可用分支生成检测导向的弱辅助上下文,用于替代缺失侧的零占位或复制输入;该补偿不重建缺失图像,而是作为 BPF Encoder 的配套机制提供可控强度的辅助语义。

3) 构建三分支课程训练(TBCT)。在同一检测器中交替采样并优化完整 RGB+IR、RGB-only 和 IR-only 三种状态,使模型具备三种输入状态下的检测能力;动态课程权重作为 TBCT 内部的更新频率调度,按照早期缺失保护、中期三分支平衡和后期完整分支恢复的顺序,平衡完整双模态精度与单路传感器可用时的鲁棒性。

1 方 法

1.1 总体框架

如图 1 所示,TPB-DETR 面向完整 RGB+IR 输入和单路模态缺失输入建立统一检测流程。该流程以分支状态 m 为控制变量,将输入、主干前向、缺失辅助上下文生成、受保护融合以及训练采样调度连接起来。给定可见光图像 $\boldsymbol{I}^{\mathrm{r}}$ 和红外图像 $\boldsymbol{I}^{\mathrm{t}}$,三种输入状态分别表示为完整双模态、仅 RGB 真实可用和仅 IR 真实可用。为避免在缺失状态下仍将不存在的模态视为有效输入,本文首先将状态输入显式定义为:

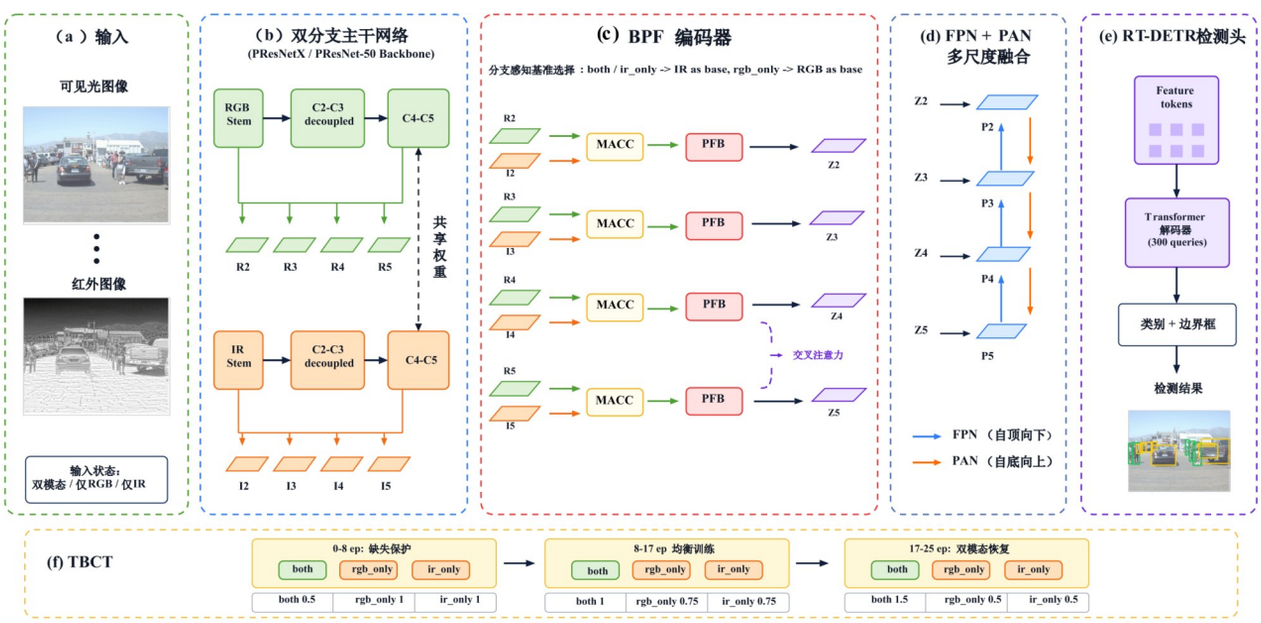


图 1 TPB-DETR 整体框架

Fig. 1 Overall architecture of TPB-DETR

$$\begin{cases} \mathcal{M} = \{\mathrm{both}, \mathrm{rgb}_{_\mathrm{only}}, \mathrm{ir}_{_\mathrm{only}}\}, \\ \boldsymbol{X}^{\mathrm{both}} = (\boldsymbol{I}^{\mathrm{r}}, \boldsymbol{I}^{\mathrm{t}}), \boldsymbol{X}^{\mathrm{rgb}} = (\boldsymbol{I}^{\mathrm{r}}, \varnothing), \boldsymbol{X}^{\mathrm{ir}} = (\varnothing, \boldsymbol{I}^{\mathrm{t}}). \end{cases} \tag{1}$$

模型按式(1)定义的三种输入状态,在同一参数集合下完成检测推理。图 1(a)给出输入状态,图 1(b)通过双分支 PResNetX-50 提取多尺度特征,图 1(c)在 BPF Encoder 中先由 MACC 生成缺失侧伪辅助上下文,再通过 PFB 完成受保护融合,图 1(d)将融合特征送入 FPN/PAN 多尺度路径,图 1(e)由 RT-DETR 检测头输出类别和边界框预测,图 1(f)表示训练阶段的 TBCT 采样调度。总体前向过程可表

© 中国图象图形学报版权所有

示为：

$$\begin{aligned}\boldsymbol{E}^m &= \mathrm{B}_\theta(\boldsymbol{X}^m),\\ \boldsymbol{Z}^m &= \mathrm{BPF}_\theta(\boldsymbol{E}^m,m),\\ \boldsymbol{P}^m &= \mathrm{FPN/PAN}_\omega(\boldsymbol{Z}^m),\\ \hat{\boldsymbol{Y}}^m &= \mathrm{D}_\psi(\boldsymbol{P}^m),m\in\mathbf{M}.\end{aligned}\tag{2}$$

式(2)中，$\mathrm{B}_\theta(\cdot)$为PResNetX-50主干网络，$\boldsymbol{E}^m$为状态m下的多尺度主干特征；$\mathrm{BPF}_\theta(\cdot)$为分支保护融合编码器，$\boldsymbol{Z}^m$为其输出的多尺度融合特征；$\mathrm{FPN/PAN}_\omega(\cdot)$表示检测头前的多尺度特征聚合路径；$\mathrm{D}_\psi(\cdot)$表示RT-DETR解码器，$\hat{\boldsymbol{Y}}^m$为第$m$种状态下的类别与边界框预测。

分支状态m不仅表示输入是否完整，也控制真实模态前向、MACC伪辅助上下文生成以及融合阶段base/aux角色选择。both表示RGB和IR均真实可用，rgb_only表示仅RGB真实可用，ir_only表示仅IR真实可用。

主干网络采用PResNetX-50，在stem和前两个stage为RGB和IR分别维护独立卷积参数(conv1_rgb/conv1_ir及对应残差层)，高层stage共享语义表示。对于完整输入，两路模态独立提取低层特征后在共享stage汇合；对于缺失输入，仅真实可用模态参与特征提取，缺失侧主干输出保持零张量，完全跳过缺失分支的前向计算。零占位特征不直接用于融合，缺失辅助上下文由真实可用特征经过MACC生成。融合编码器根据分支状态选择受保护主路径和辅助路径，输出步长为4、8、16和32的多尺度融合特征，再输入FPN/PAN路径和RT-DETR解码器(Zhao等，2024)，得到类别概率和边界框预测。

作为三分支训练与统一推理的状态约定，表1给出了三种输入状态下真实输入、缺失处理、受保护主路径和辅助上下文的对应关系。该状态约定并非独立模块，而是后续MACC、BPF Encoder和TBCT共享的输入语义。完整输入下，RGB和IR均为真实模态，当前实现以IR作为受保护主路径、RGB作为真实辅助信息；RGB-only输入下，RGB为真实可用主路径，IR侧仅生成伪红外辅助上下文；IR-only输入下，IR为真实可用主路径，RGB侧仅生成伪可见光辅助上下文。为使后续模块的输入来源和融合角色保持一致，本文进一步对三种分支状态下的特征流进行约定。具体而言，真实输入决定主干网络中实际参与编码的模态；缺失侧决定是否需要引入MACC生成伪辅助上下文；受保护主路径和辅助上下文则分别对应PFB中的base路径和aux路径。基于该约定，完整输入、RGB-only输入和IR-only输入可以在同一检测框架下共享后续融合与检测结构。

表1　三种分支状态的结构语义

Table 1　Structural semantics of three branch states

分支状态	真实输入	缺失侧	受保护主路径	辅助上下文
both	RGB+IR	无	IR特征	RGB特征
rgb_only	RGB	伪IR上下文	RGB特征	伪IR上下文
ir_only	IR	伪RGB上下文	IR特征	伪RGB上下文

这种状态定义具有双重作用。其一，缺失模态下缺失侧骨干前向被跳过，模型不会将零输入经过卷积后得到的偏置响应当作真实模态特征。其二，融合模块始终明确哪一路为可靠base，哪一路为辅助aux，从而可对辅助路径使用更保守的门控和尺度控制。完整双模态状态采用IR作为base，主要考虑红外成像在低照度和热目标响应中的稳定性；但该设计并不固定所有状态都偏向IR，因为在RGB-only状态下，可见光特征被设为受保护主路径。需要说明的是，本文的base选择服务于缺失模态鲁棒性目标，而非证明IR在所有场景下均优于RGB。完整分支采用IR作为protected base，是基于低照度RGB-T检测中红外响应更稳定的任务先验；RGB信息仍通过受控辅助路径参与纹理、边缘和外观补充，且在RGB-only状态下RGB本身被设为protected base。

1.2　分支保护融合编码器(BPF Encoder)

BPF依据输入状态建立非对称融合：保留真实主路径，并将真实辅助模态或MACC生成的上下文以受控残差形式加入主路径，具体结构如图2所示。

对于第i个尺度，$\boldsymbol{B}_i^m$表示状态m下的受保护主路径特征，$\boldsymbol{A}_i^m$表示辅助路径特征。PFB首先由$\boldsymbol{B}_i^m$和$\boldsymbol{A}_i^m$生成空间门控$\boldsymbol{G}_i^s$和通道门控$\boldsymbol{G}_i^c$，并用二者共同调制辅助残差，而不直接缩放主路径特征。

$$\begin{cases}\boldsymbol{G}_i^s=\sigma\left(f_i^s([\boldsymbol{B}_i^m,\boldsymbol{A}_i^m])\right),\\ \boldsymbol{G}_i^c=\sigma\left(f_i^c(\mathrm{GAP}([\boldsymbol{B}_i^m,\boldsymbol{A}_i^m]))\right),\\ \boldsymbol{D}_i^g(\boldsymbol{B}_i^m,\boldsymbol{A}_i^m)=\boldsymbol{G}_i^s\odot\boldsymbol{G}_i^c\odot\boldsymbol{P}_i(\boldsymbol{A}_i^m).\end{cases}\tag{3}$$

式(3)中，$\sigma(\cdot)$为Sigmoid函数，$f_i^s(\cdot)$为空间门控卷积，$f_i^c(\cdot)$为SE风格通道门控，$\boldsymbol{P}_i(\cdot)$为辅助特征

© 中国图象图形学报版权所有

的1×1残差投影，⊙表示逐元素乘法。$\boldsymbol{D}_i^{\mathrm{g}}(\cdot)$只由辅助路径产生，并作为写入base的门控残差项。式(4)定义高层交叉注意力更新，其他尺度更新置零：池化后base token作为query，aux token为key和value，使辅助信息以增量形式写入主路径。

$$\boldsymbol{D}_i^{\mathrm{a}}(\boldsymbol{B}_i^m,\boldsymbol{A}_i^m)=\begin{cases}\mathrm{Up}(\mathrm{Attn}_i(\boldsymbol{T}_{\mathrm{B}},\boldsymbol{T}_{\mathrm{A}})-\boldsymbol{T}_{\mathrm{B}}), & i\in\mathbf{I}_{\mathrm{a}},\\ 0, & i\notin\mathbf{I}_{\mathrm{a}},\end{cases}$$
$$\boldsymbol{T}_{\mathrm{B}}=\mathrm{Pool}(\boldsymbol{B}_i^m),\boldsymbol{T}_{\mathrm{A}}=\mathrm{Pool}(\boldsymbol{A}_i^m). \tag{4}$$

最终融合输出为：

$$\boldsymbol{Z}_i^m=\boldsymbol{B}_i^m+\boldsymbol{\gamma}_{\mathrm{g},i}^{\rho(m)}\odot\boldsymbol{D}_i^{\mathrm{g}}(\boldsymbol{B}_i^m,\boldsymbol{A}_i^m)+\boldsymbol{\gamma}_{\mathrm{a},i}^{\rho(m)}\odot\boldsymbol{D}_i^{\mathrm{a}}(\boldsymbol{B}_i^m,\boldsymbol{A}_i^m). \tag{5}$$

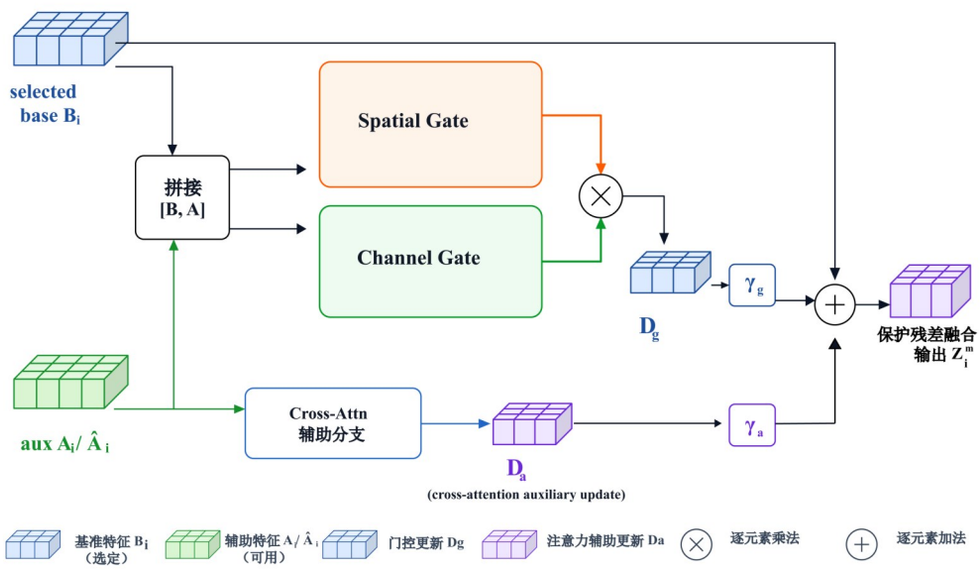


图2　受保护融合块结构

Fig. 2　Structure of the Protected Fusion Block

式(5)中，$\boldsymbol{D}_i^{\mathrm{g}}(\cdot)$为空间、通道门控后的辅助残差更新，$\boldsymbol{D}_i^{\mathrm{a}}(\cdot)$为池化交叉注意力产生的辅助更新项，$\mathcal{I}_{\mathrm{a}}$为启用交叉注意力的高层尺度集合。$\rho(\cdot)$为分支状态到尺度系数的映射：$\rho(\mathrm{both})=\mathrm{both}$，$\rho(\mathrm{rgb_only})=\rho(\mathrm{ir_only})=\mathrm{miss}$；$\boldsymbol{\gamma}_{\mathrm{g},i}^{\rho(m)}$和$\boldsymbol{\gamma}_{\mathrm{a},i}^{\rho(m)}$为第$i$尺度的可学习通道系数，并广播至空间维。

上述融合将分支状态落实为主辅更新规则：主路径特征$\boldsymbol{B}_i^m$的系数恒为1，门控和注意力产生的更新量仅以残差方式叠加到主路径。门控掩码$\boldsymbol{G}_i^s$和$\boldsymbol{G}_i^c$只作用于辅助残差$\boldsymbol{P}_i(\boldsymbol{A}_i^m)$，不直接作用于$\boldsymbol{B}_i^m$；交叉注意力项也采用更新后的base token减去原base token的形式形成辅助增量。门控与注意力共同调节辅助更新，恒等通路保留真实主路径特征。为区分真实与伪辅助上下文的写入强度，完整双模态分支使用独立的$\boldsymbol{\gamma}_{\mathrm{g},i}^{\mathrm{both}}$和$\boldsymbol{\gamma}_{\mathrm{a},i}^{\mathrm{both}}$，两个缺失分支共享$\boldsymbol{\gamma}_{\mathrm{g},i}^{\mathrm{miss}}$和$\boldsymbol{\gamma}_{\mathrm{a},i}^{\mathrm{miss}}$；完整分支的门控和注意力系数以$10^{-3}$初始化，缺失分支的注意力系数初始化为0，门控系数同样以10^{-3}起步。

三种分支状态下base和aux的定义为：both状态采用$\boldsymbol{B}_i^m=\boldsymbol{E}_i^{\mathrm{ir}}$、$\boldsymbol{A}_i^m=\boldsymbol{E}_i^{\mathrm{rgb}}$；rgb_only状态采用$\boldsymbol{B}_i^m=\boldsymbol{E}_i^{\mathrm{rgb}}$、$\boldsymbol{A}_i^m=\hat{\boldsymbol{A}}_i^{\mathrm{ir}}$；ir_only状态采用$\boldsymbol{B}_i^m=\boldsymbol{E}_i^{\mathrm{ir}}$、$\boldsymbol{A}_i^m=\hat{\boldsymbol{A}}_i^{\mathrm{rgb}}$。在缺失分支中，辅助上下文只作为受控残差进入融合结果；在完整分支中，两路真实模态通过同一融合结构进行互补交互。当前实现采用4个多尺度特征，交叉注意力主要作用于高层语义尺度，以减少低层噪声对缺失辅助特征的放大。

1.3　缺失辅助上下文补偿(MACC)

缺失状态下不直接使用零占位，MACC从真实可用特征生成伪辅助上下文，结构如图3所示。

对于第i个尺度，记$a(m)$为状态m下真实可用的模态，$\bar{a}(m)$为与其相对的缺失侧模态，其中$a(\mathrm{rgb_only})=\mathrm{rgb}$、$\bar{a}(\mathrm{rgb_only})=\mathrm{ir}$，$a(\mathrm{ir_only})=\mathrm{ir}$、$\bar{a}(\mathrm{ir_only})=\mathrm{rgb}$。MACC以真实可用特征$\boldsymbol{F}_i^{a(m)}$为输入，生成供PFB辅助路径使用的伪辅助上下文：

$$\hat{\boldsymbol{A}}_i^{\bar{a}(m)}=\boldsymbol{F}_i^{a(m)}+\boldsymbol{s}_i^{\bar{a}(m)}\odot\boldsymbol{R}_i^{\bar{a}(m)}(\boldsymbol{F}_i^{a(m)})+\boldsymbol{b}_i^{\bar{a}(m)},\quad m\in\mathrm{rgb}_{_\mathrm{only}},\mathrm{ir}_{_\mathrm{only}} \tag{6}$$

© 中国图象图形学报版权所有

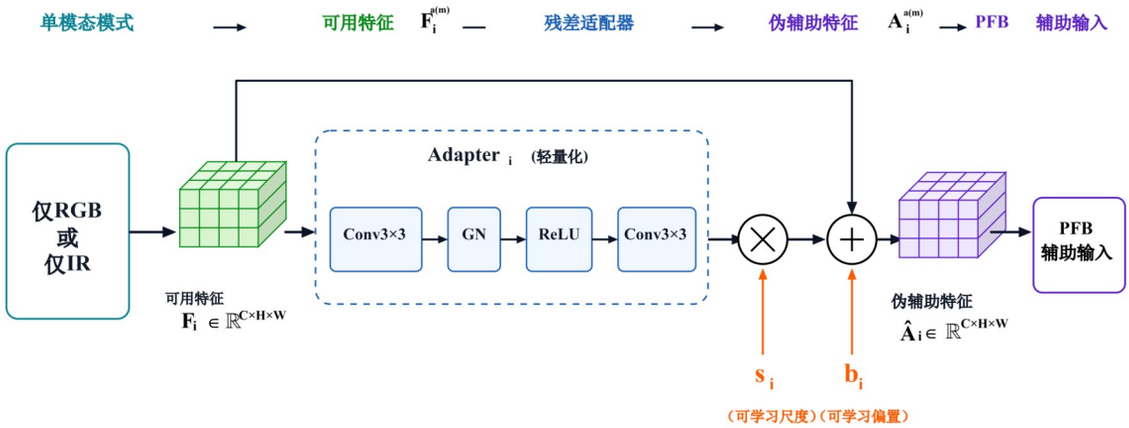


图3　缺失辅助上下文补偿模块结构

Fig. 3　Structure of the Missing-Auxiliary Context Compensation module

式(6)中，$\boldsymbol{R}_i(\cdot)$为生成缺失侧上下文的轻量卷积残差适配器，由Conv3×3、GN、ReLU和Conv3×3组成；$\boldsymbol{s}_i$为可学习通道尺度，$\boldsymbol{b}_i$为全局偏置。当前为伪IR和伪RGB上下文分别维护一套MACC适配器。$\boldsymbol{s}_i$初始化为10^{-3}，适配器最后一层卷积采用零初始化，使训练初期伪辅助上下文主要退化为真实可用特征上的弱残差扰动。

需要强调的是，上述公式并不表示重建缺失图像，也不表示将零占位分支复制为真实模态。MACC的输入始终来自真实可用分支，其输出仅作为BPF Encoder中辅助路径的输入，用于提供轻量、检测导向且可控强度的辅助上下文。

1.4　三分支课程训练(TBCT)

TBCT用于训练阶段的分支状态采样与更新频率调度，整体流程如图4所示。图中双模态、仅RGB和仅IR分别对应RGB+IR、RGB-only和IR-only三种输入状态，均由同一参数模型完成前向推理，而非三套独立检测器。

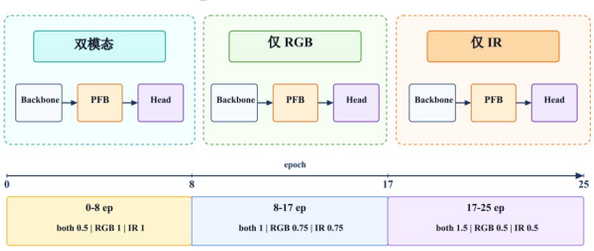


图4　三分支课程训练机制

Fig. 4　Three-Branch Curriculum Training

具体实现上，训练阶段分别维护both、rgb_only和ir_only三个分支数据流。每个epoch开始时，根据当前阶段的课程权重w_e^m构造一个分支交替序列；例如早期阶段(0.5, 1.0, 1.0)对应完整分支和两个缺失分支约为1:2:2的更新频率。训练时每个step只从该序列指定的一个分支取batch，并只执行该分支状态的一次前向、反向传播和参数更新。因此，该方法的分支权重不是单个batch内同时计算三路损失后的线性加权系数，而是当前epoch内三种分支数据加载器(loader)的交替采样和参数更新贡献比例。设第e个epoch中分支m的课程权重为w_e^m，对应的采样概率为：

$$p_e^m = \frac{w_e^m}{\sum_{m'\in \mathbf{M}} w_e^{m'}}. \tag{7}$$

则训练过程可写为如式(8)的期望目标：

$$\min_{\theta,\omega,\psi} \mathrm{E}_{(I^r,I^t,Y)\ \ \boldsymbol{D},m\ \ p_e}[\mathcal{L}_{\mathrm{det}}^m(\hat{\boldsymbol{Y}}^m,\boldsymbol{Y})] \tag{8}$$

式(8)中，θ、ω和ψ分别为主干/BPF、FPN/PAN和RT-DETR解码器的参数，$\boldsymbol{\mathcal{D}}$为训练数据分布，$\mathcal{L}_{\mathrm{det}}^m$为第$m$种状态下的RT-DETR检测损失，包括分类损失、边界框L_1损失和GIoU损失。上述目标表示由分支交替更新流近似的经验期望，而不是三路分支在同一batch内的联合损失。采用更新频率实现课程权重可以避免三路同时前向带来的显存和计算开销，同时降低不同分支损失尺度差异对加权优化的影响。具体动态课程权重如表2所示。

早期提高RGB-only和IR-only权重，使单模态路径获得充分监督，避免完整分支过早主导共享参数；

© 中国图象图形学报版权所有

表2　动态课程权重

Table 2　Dynamic curriculum weights

epoch	阶段	both	rgb_only	ir_only	训练目的
0－7	early	0.50	1.00	1.00	强化缺失分支
8－16	middle	1.00	0.75	0.75	平衡三种状态
17－24	late	1.50	0.50	0.50	恢复完整分支

中期提高完整分支权重并保留缺失分支训练；后期进一步强化RGB-T输入，恢复完整输入检测精度。三种状态在同一训练周期内交替优化，分支权重随epoch调整。图4中的0－8、8－17和17－25为左闭右开区间，对应表2的0－7、8－16和17－24。后续固定课程消融仅去除动态调度，仍保留三分支训练。

2　实验与分析

2.1　实验设置

本文在FLIR aligned和LLVIP两个RGB-T检测数据集上进行实验。FLIR aligned包含车载交通场景下的可见光与热红外图像，目标类别包括person、car和bicycle；实验使用的COCO格式划分包含4129对训练图像和1013对验证图像（Teledyne FLIR，2018）。LLVIP包含低照度监控场景下配准的可见光-红外行人图像（Jia等，2021）；实验使用12025对训练图像和3463对验证图像，类别为person。实验环境为Ubuntu 20.04.6 LTS，硬件采用AMD EPYC 7601 CPU、251 GiB内存和NVIDIA GeForce RTX 3090 GPU；软件环境采用Python 3.9.25、PyTorch 2.0.1和torchvision 0.15.2。完整双模态检测采用COCO风格mAP、mAP50和mAP75作为评价指标；缺失模态评估分别报告RGB-only和IR-only分支结果，并以两种缺失状态的算术平均衡量缺失输入下的总体表现。

模型采用PResNetX-50主干、隐藏维度为256的BPF Encoder以及6层RT-DETR解码器（Zhao等，2024；Carion等，2020；Zhu等，2021；Zhang等，2023），解码器查询数为300。PResNetX在stem和前2个stage采用RGB/IR解耦分支，高层stage共享语义表示；融合编码器使用4个尺度特征，步长为4、8、16和32，并在高层尺度使用交叉注意力。模型参数量约为67.1M。训练输入基础尺寸为640×640；训练增强包括随机水平翻转、随机IoU裁剪、随机缩放外扩和RGB侧光度扰动。优化器为AdamW，基础学习率为1.0×10^{-4}，主干学习率为1.0×10^{-5}，权重衰减为1.0×10^{-4}，batch size为4，采用余弦退火学习率、指数移动平均（exponential moving average，EMA）及0.1的梯度裁剪。两数据集均训练25轮。为综合衡量两种缺失输入状态下的检测性能，本文进一步计算缺失平均mAP和缺失平均mAP50（式中分别记为Avg mAP和Avg mAP50）：

$$\begin{aligned}\mathrm{Avg}_{\mathrm{mAP}} &= \frac{\mathrm{mAP}_{\mathrm{rgb_only}}+\mathrm{mAP}_{\mathrm{ir_only}}}{2},\\ \mathrm{Avg}_{\mathrm{mAP50}} &= \frac{\mathrm{mAP50}_{\mathrm{rgb_only}}+\mathrm{mAP50}_{\mathrm{ir_only}}}{2}.\end{aligned}\tag{9}$$

式（9）中，下标rgb_only和ir_only分别表示RGB-only与IR-only输入状态。缺失平均mAP和缺失平均mAP50分别为上述两种状态对应指标的算术平均。

2.2　三分支检测结果

表3给出了本文方法在FLIR和LLVIP数据集上的三分支检测结果。完整RGB+IR输入在两个数据集上分别达到47.32 / 84.33和70.30 / 98.14的mAP / mAP50。缺失模态条件下，FLIR上RGB-only与IR-only分支分别取得34.99 / 71.68和46.17 / 82.80，LLVIP上分别取得53.71 / 91.52和69.25 / 97.70。由式（9）可得，FLIR和LLVIP的缺失平均mAP / mAP50分别为40.58 / 77.24和61.48 / 94.61。可以看出，模型在单路模态可用时仍具备较好的检测稳定性，其中IR-only分支表现更优，符合红外模态在低照度场景中的成像优势。

2.3　完整双模态公开对比

公开RGB-T检测论文近年来通常同时报告COCO mAP、mAP50和mAP75，其中mAP能够反映

表3　三分支检测结果

Table 3　Detection results under three branch states

数据集	分支状态	mAP	mAP50	mAP75
FLIR	RGB+IR	47.32	84.33	45.43
	RGB-only	34.99	71.68	28.67
	IR-only	46.17	82.80	43.39
LLVIP	RGB+IR	70.30	98.14	81.37
	RGB-only	53.71	91.52	57.77
	IR-only	69.25	97.70	79.54

© 中国图象图形学报版权所有

多IoU阈值下的整体定位与分类质量，因此本文方法以mAP作为主要比较指标，并保留mAP50和mAP75作为辅助指标。为避免按结果选择指标，完整双模态比较固定采用公开论文最常见的mAP/mAP50/mAP75三指标，不因某一方法在单项指标上更优而切换主指标。表4汇总了近年来与本文研究检测协议相近、且同时报告FLIR-aligned与LLVIP完整RGB+IR结果的代表性方法（Medeiros 等，2025；Xiao 等，2024；Yuan 等，2025；Fang 等，2021；Cao 等，2023；Shen 等，2024；Fu 等，2024a；Zhao 等，2026；Lee 等，2024），其中GM-DETR作为该研究结构消融中的直接baseline被单独列出。需要说明的是，不同论文在骨干网络、预训练数据、输入尺寸、训练轮数和数据清洗协议上存在差异，表4用于公开性能定位，不作为严格同协议排序；未公开的指标以“-”表示。

表4　完整RGB+IR输入下的公开mAP对比

Table 4　Public mAP comparison under complete RGB+IR inputs

方法	FLIR-aligned			LLVIP		
	mAP	mAP50	mAP75	mAP	mAP50	mAP75
CFT	40.2	78.7	35.5	63.6	97.5	72.9
CSSA	41.3	79.2	37.4	59.2	94.3	66.6
ICAFusion	41.4	79.2	36.9	62.3	96.3	71.7
LRAF-Net	42.8	80.5	-	66.3	97.9	-
RSDet	43.8	83.9	40.1	61.3	95.8	70.4
CrossFormer	42.1	79.3	38.5	65.1	97.4	75.4
UniRGB-IR	44.1	81.4	40.2	63.2	96.1	72.2
MiPa	44.8	81.3	41.8	66.5	98.2	78.1
GM-DETR(fus)	45.3	83.6	42.0	69.6	97.5	80.6
GM-DETR(iso+fus)	45.8	83.9	42.6	70.2	97.4	81.4
本文方法	47.32	84.33	45.43	70.30	98.14	81.37

从完整输入结果看，本文方法在FLIR-aligned上达到47.32 mAP和84.33 mAP50，高于GM-DETR公开结果的45.8 mAP；在LLVIP上达到70.30 mAP和98.14 mAP50，与GM-DETR的70.2 mAP处于同一水平，同时在同一25个epoch训练中支持RGB+IR、RGB-only和IR-only三种输入状态。鉴于本文的核心目标是兼顾完整输入与缺失模态输入，完整双模态性能并不作为唯一评价依据，而需要结合下文缺失状态结果进行综合分析。

2.4　缺失模态公开对比

缺失模态场景中的公开协议目前尚不统一：GM-DETR和MS2Fusion主要报告FLIR-aligned上输入替换或单模态模型的mAP50，MiPa在补充材料中报告mAP50/mAP75/mAP，YOLOv11-RGBT给出单模态与完整RGB-T模型基线（Medeiros 等，2025；Xiao 等，2024；Wan 等，2025；Shen 等，2026）。为避免根据结果选择评价指标，表5和表6同时保留mAP与mAP50；未报告指标以“-”标记，不换算、不估计。表中“单模态”表示两个独立单模态模型，“输入替换”（drop-input）表示完整双模态模型在测试时用同一模态替换缺失输入，“统一模型”表示同一参数模型支持不同输入状态。由于三类设置的训练目标和部署开销并不等价，表5和表6只在同一指标列内作公开参考，不混合比较mAP和mAP50；DINO/MiPa结果引自MiPa补充材料，GM-DETR结果来自GM-DETR（Xiao 等，2024），MS2Fusion和YOLOv11x结果分别来自MS2Fusion（Shen 等，2026）与YOLOv11-RGBT（Wan 等，2025）。

FLIR-aligned结果表明，直接将完整双模态模型改为输入替换会明显降低RGB-only结果，例如ICAFusion和GM-DETR iso+fus的RGB-only mAP50分别为57.80和64.30。在同时公开mAP的DINO、MiPa

© 中国图象图形学报版权所有

表5 FLIR-aligned缺失模态公开对比

Table 5 Public comparison with incomplete modalities on FLIR-aligned

方法	设置	RGB mAP	RGB mAP50	IR mAP	IR mAP50	Avg mAP	Avg mAP50
YOLOv5	单模态	-	67.80	-	73.90	-	70.85
RT-DETR	单模态	-	69.70	-	80.50	-	75.10
DINO-RGB	模态无关	32.33	66.07	26.30	56.60	29.32	61.33
DINO-IR	模态无关	24.30	56.47	38.97	70.40	31.63	63.43
DINO-Both	混合训练	27.37	60.50	35.90	68.93	31.63	64.72
MiPa+MI	模态无关	30.60	64.80	37.50	70.43	34.05	67.62
ICAFusion	输入替换	-	57.80	-	66.00	-	61.90
GM-DETR fus	输入替换	-	57.60	-	78.70	-	68.15
GM-DETR iso+fus	输入替换	-	64.30	-	80.60	-	72.45
MS2Fusion	输入替换	-	68.40	-	82.10	-	75.25
YOLOv11x	单模态	34.36	68.65	44.43	80.27	39.40	74.46
本文方法	统一模型	34.99	71.68	46.17	82.80	40.58	77.24

和YOLOv11x单模态基线中，TPB-DETR在RGB-only、IR-only和缺失平均mAP上均取得最优结果；在mAP50对比中，两种缺失状态的结果分别达到71.68和82.80，同样高于表中公开mAP50基线。这些结果表明，本文方法在mAP与mAP50两个指标上均取得一致的改善，验证了模型在FLIR缺失模态检测中的有效性。模态无关学习方法MiPa+MI在RGB-only和IR-only下的mAP分别为30.60和37.50，本文方法分别提高4.39和8.67个百分点，在两种缺失状态下均保持了较好的检测能力。与YOLOv11x单模态基线相比，本文方法在RGB-only和IR-only下的mAP分别提高0.63和1.74个百分点，mAP50分别提高3.03和2.53个百分点。两种缺失状态均取得提升，其中IR-only的mAP改善更为突出。这些结果由同一套模型参数获得，输入状态改变时无需切换到另一套单模态检测器。对于传感器暂时不可用的场景，统一模型能够根据实际输入继续完成检测，并在双路输入恢复后使用模态互补信息。

LLVIP由低照度行人场景构成，红外模态通常占主导，因此IR-only结果普遍高于RGB-only。YOLOv11x单模态模型在LLVIP的RGB mAP、IR mAP和缺失平均mAP上略优于本文方法，但它对应两个独立单模态检测器；本文方法仅以同一套参数统一支持RGB+IR、RGB-only和IR-only三种状态，在IR-only mAP50上达到97.70，且在平均mAP和平均mAP50上均接近强单模态基线。该结果表明，本文方法以同一套参数支持三种输入状态，在完整输入性能、缺失输入可用性和统一部署之间取得平衡。

表6 LLVIP缺失模态公开对比

Table 6 Public comparison with incomplete modalities on LLVIP

方法	设置	RGB mAP	RGB mAP50	IR mAP	IR mAP50	Avg mAP	Avg mAP50
DINO-RGB	模态无关	51.87	90.87	59.43	94.23	55.65	92.55
DINO-IR	模态无关	29.03	66.87	64.27	96.87	46.65	81.87
DINO-Both	混合训练	46.43	82.40	64.83	96.50	55.63	89.45
MiPa+MI	模态无关	48.10	89.10	63.17	96.83	55.63	92.97
YOLOv11x	单模态	54.67	91.53	69.93	97.41	62.30	94.47
本文方法	统一模型	53.71	91.52	69.25	97.70	61.48	94.61

© 中国图象图形学报版权所有

2.5 消融实验

2.5.1 核心模块消融

表7汇总FLIR与LLVIP的核心模块消融。各组使用相应数据集的相同划分、主干预训练权重、检测头、预处理和优化设置，以batch size为4训练25轮。每组只改变对应模块；课程对照保持相同的逐轮优化更新次数。

表7 FLIR与LLVIP核心模块消融结果

Table 7 Core-module ablation results on FLIR and LLVIP

方法	FLIR					LLVIP				
	RGB+IR mAP	RGB-only mAP	IR-only mAP	RGB-only mAP_{50}	IR-only mAP_{50}	RGB+IR mAP	RGB-only mAP	IR-only mAP	RGB-only mAP_{50}	IR-only mAP_{50}
替换BPF	47.56	32.44	45.28	68.40	82.21	70.11	53.33	67.76	90.13	97.28
去除MACC	47.06	34.16	46.15	70.47	82.91	68.49	52.93	67.94	90.49	97.65
固定课程	46.86	33.16	45.53	69.06	82.29	68.27	53.34	67.60	91.04	97.70
本文方法	47.32	34.99	46.17	71.68	82.80	70.30	53.71	69.25	91.52	97.70

以MSFI/CMSF融合替换BPF，并保持MACC和动态课程不变。采用BPF后，FLIR的RGB-only和IR-only mAP分别提高2.55和0.89个百分点，LLVIP的对应提升为0.38和1.49个百分点。FLIR在红外缺失时改善更突出，LLVIP则在可见光缺失时提升更大。两组FLIR完整输入mAP为47.56和47.32，相差0.24个百分点，BPF主要改善单路输入检测。

去除MACC后，以全零特征作为缺失侧辅助输入，并保留BPF和TBCT。加入MACC使FLIR的RGB-only mAP由34.16提高至34.99，提升0.83个百分点；LLVIP的RGB-only与IR-only mAP分别提高0.78和1.31个百分点，说明可用模态生成的辅助上下文有助于缺失输入检测。

固定课程保留三种输入状态，区别在于不随训练阶段调整比例。动态课程使FLIR的RGB-only和IR-only mAP分别提高1.83和0.64个百分点，LLVIP的对应提升为0.37和1.65个百分点。完整输入mAP也分别提高0.46和2.03个百分点，说明阶段性调度能够兼顾完整输入精度与缺失输入检测性能。

2.5.2 缺失辅助上下文对比

表8比较四种辅助上下文构造方式，均保留BPF与动态课程。

去除MACC时采用全零辅助特征；特征复制直接复用可用特征；简单卷积在各尺度采用恒等初始化的单层1×1卷积；MACC采用第1.3节的卷积残差映射。

表8 FLIR缺失辅助上下文对比

Table 8 Auxiliary-context comparison on FLIR

方法	RGB+IR mAP	RGB-only mAP	IR-only mAP	RGB-only mAP_{50}	IR-only mAP_{50}
去除MACC	47.06	34.16	46.15	70.47	82.91
特征复制	47.08	34.71	45.67	70.62	82.57
简单卷积	47.14	33.86	46.47	70.29	83.37
MACC	47.32	34.99	46.17	71.68	82.80

红外缺失（RGB-only）时，MACC取得34.99的mAP，较去除MACC、特征复制和简单卷积分别提高0.83、0.28和1.13个百分点。IR-only下四组mAP处于45.67至46.47之间，MACC为46.17。MACC在RGB-only和完整输入下均取得最高mAP，说明残差上下文映射有助于可见光单路检测。

2.5.3 受保护主路径选择

表9比较FLIR与LLVIP完整输入下的IR和RGB主路径，其余配置一致，缺失时以可用模态为主路径。IR主路径使FLIR的RGB-only和IR-only mAP分别提高1.16和3.33个百分点，LLVIP的IR-only由67.19提高至69.25。LLVIP的RGB-only在IR、RGB主路径下分别为53.71、54.11；RGB主路径在FLIR完整输入下高0.17个百分点。三种状态共享参数，完整输入的主辅分配影响共享特征学习，因此缺失状态的结果也会变化。两数据集均支持以

© 中国图象图形学报版权所有

IR为主路径改善可见光缺失时的检测。

表9 FLIR与LLVIP受保护主路径对比

Table 9 Protected-base comparison on FLIR and LLVIP

方法	FLIR					LLVIP				
	RGB+IR mAP	RGB-only mAP	IR-only mAP	RGB-only mAP_{50}	IR-only mAP_{50}	RGB+IR mAP	RGB-only mAP	IR-only mAP	RGB-only mAP_{50}	IR-only mAP_{50}
IR主路径	47.32	34.99	46.17	71.68	82.80	70.30	53.71	69.25	91.52	97.70
RGB主路径	47.49	33.83	42.84	69.43	79.55	69.83	54.11	67.19	91.66	97.22

2.5.4 课程权重与模块组合

课程对照按RGB+IR、RGB-only和IR-only的顺序设置分支更新比例。FLIR固定课程在25轮内保持1:2:2;温和课程在表2的三个阶段依次采用0.20:0.40:0.40、0.30:0.35:0.35和0.40:0.30:0.30。完整动态课程采用表2所列权重,其归一化比例依次为0.20:0.40:0.40、0.40:0.30:0.30和0.60:0.20:0.20。

表10 FLIR课程与上下文组合对比

Table 10 Curriculum and context combinations on FLIR

方法	RGB+IR mAP	RGB-only mAP	IR-only mAP	RGB-only mAP_{50}	IR-only mAP_{50}
去除MACC+固定	46.73	34.77	45.51	71.49	82.78
去除MACC+动态	47.06	34.16	46.15	70.47	82.91
MACC+固定	46.86	33.16	45.53	69.06	82.29
MACC+温和	46.95	34.25	45.96	70.49	82.68
MACC+动态	47.32	34.99	46.17	71.68	82.80

表10中,MACC配合固定、温和及完整动态课程时,RGB-only mAP依次为33.16、34.25和34.99,IR-only也逐步提高。去除MACC时,动态课程提高IR-only mAP,RGB-only则由34.77变为34.16。MACC配合完整动态课程在三种状态下均取得表中最高mAP,其中RGB-only较固定课程提高1.83个

百分点。这表明辅助上下文与课程调度共同影响缺失输入检测。

2.6 可视化分析

图5(a)—(d)与图5(e)—(h)分别对应RGB-only和IR-only;两组依次为替换BPF模型的结果、去除MACC模型的结果、本文方法的结果和GT。替换BPF的模型更易出现背景误检、定位偏移和小目标漏检,说明常规融合在缺失输入下易受不可靠辅助分支干扰。去除MACC的模型仍保留受保护融合路径,但缺失侧由零占位,缺少与可用模态匹配的辅助上下文,故远距离目标和弱纹理区域的检测仍不稳定。本文方法预测框更接近GT,在多目标区域保留更多检测结果。这表明,受保护主路径有助于维持可用模态的表征稳定性,MACC生成的伪辅助上下文可缓解缺失侧置零造成的信息断裂,从而提升单路模态输入下的检测可靠性。

为解释该差异,图6(a)—(d)与图6(e)—(h)分别为RGB-only和IR-only;两组依次为BPF Encoder的P4、P5空间门控响应,以及去除MACC模型和本文方法的P4辅助上下文。空间门控调节辅助残差的写入区域和强度,而非直接削弱或覆盖主路径特征;不同尺度响应表明,低层与高层语义特征对辅助信息的利用各有侧重。本文方法的伪辅助上下文相比去除MACC模型的对应响应更集中于目标邻域和结构化场景区域,说明MACC并非简单的零占位替代,而是与检测任务相关的弱辅助语义。

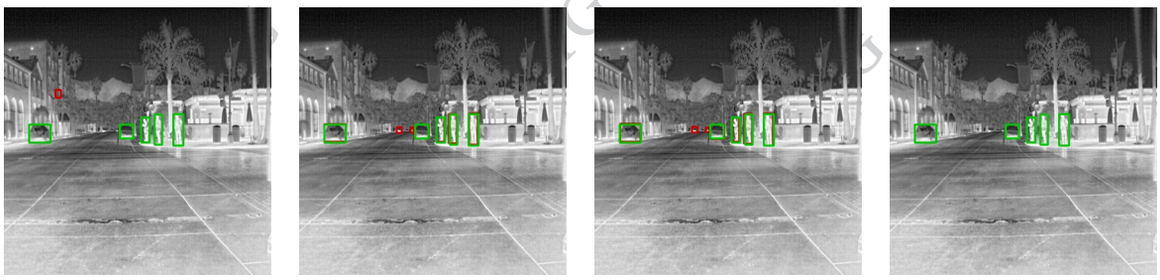
2.7 计算效率分析

在NVIDIA GeForce RTX 3090(标称显存24 GB)上测试本文方法,输入为640×640,单张输入、FP32计算。各状态固定一个验证样本,预热20次后

© 中国图象图形学报版权所有



（a） RGB-only，替换BPF　（b） RGB-only，去除MACC　（c） RGB-only，本文方法　（d） RGB-only，真实标注



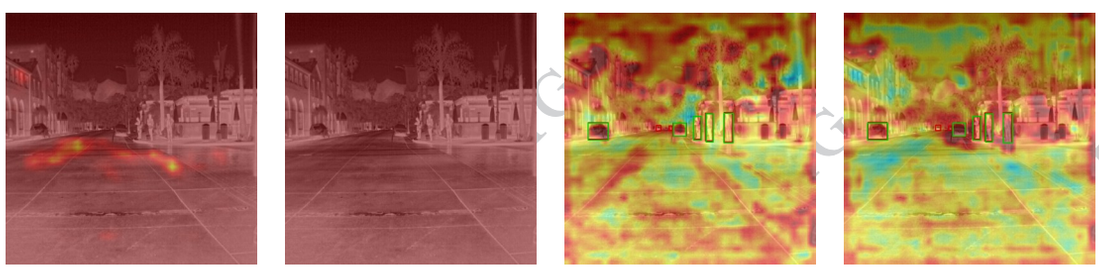
（e） IR-only，替换BPF　（f） IR-only，去除MACC　（g） IR-only，本文方法　（h） IR-only，真实标注

图5　缺失模态输入下的检测结果可视化

Fig. 5　Detection visualization with incomplete modalities. (a) - (d) RGB-only; (e) - (h) IR-only; within each group: with BPF replaced, without MACC, TPB-DETR, and ground truth



（a） RGB-only，P4门控响应　（b） RGB-only，P5门控响应　（c） RGB-only，去除MACC的P4图　（d） RGB-only，本文方法P4辅助图



（e） IR-only，P4门控响应　（f） IR-only，P5门控响应　（g） IR-only，去除MACC的P4图　（h） IR-only，本文方法P4辅助图

图6　BPF Encoder空间门控与MACC辅助上下文可视化

Fig. 6　Visualization of BPF gates and MACC auxiliary context (a) - (d) RGB-only; (e) - (h) IR-only; P4/P5 gates, P4 aux without MACC, and TPB-DETR P4 aux

© 中国图象图形学报版权所有

重复

测试100次。表11报告前向平均耗时、帧率（frames per second，FPS）及PyTorch峰值已分配显存，不计数据读取、预处理、数据传输及后处理。

表11　TPB-DETR模型前向计算效率

Table 11　Model-forward efficiency of TPB-DETR

数据集	输入状态	参数量/M	耗时/ms	FPS	显存/GiB
FLIR	RGB+IR	67.09	43.36	23.06	0.774
FLIR	RGB-only	67.09	36.55	27.36	0.773
FLIR	IR-only	67.09	36.78	27.19	0.773
LLVIP	RGB+IR	67.08	43.49	22.99	0.774
LLVIP	RGB-only	67.08	36.79	27.18	0.773
LLVIP	IR-only	67.08	36.88	27.11	0.773

FLIR与LLVIP模型的参数量分别为67.09 M和67.08 M。双模态前向速度分别为23.06和22.99 FPS，两种缺失状态均达到27 FPS以上，峰值显存占用约为0.77 GiB。缺失侧跳过骨干前向，使模型在沿用同一套参数的同时减少单模态推理开销。

公开模型的参数量可作为模型规模的参考：YOLOv11x单模态模型为54.24 M参数，YOLOv11x-Midfusion为80.04 M参数（Wan等，2025）；若部署两个独立YOLOv11x单模态检测器，则参数量约为108.48 M。本文方法使用同一套参数统一处理三种输入状态，参数量低于公开YOLOv11x-Midfusion和分别部署两个单模态模型的方案。

3　结　论

本文提出面向单路模态缺失的TPB-DETR，将缺失输入处理融入统一的检测框架，无需重建缺失图像，也无需针对不同输入状态分别部署检测器。BPF与MACC在统一融合接口下处理真实或生成的辅助特征，TBCT协调三种输入状态的训练，使同一套模型参数支持完整双模态与两种单模态检测。

公开结果对比中，TPB-DETR在FLIR的RGB-only和IR-only状态下，mAP较MiPa+MI分别提高4.39和8.67个百分点；在LLVIP上以单一模型取得接近两个独立YOLOv11x检测器的缺失平均mAP，缺失平均mAP50略高。保持MACC和课程设置不变时，BPF较MSFI/CMSF融合在两个数据集的RGB-only和IR-only状态下均取得更高mAP，提升为0.38～2.55个百分点。辅助上下文与课程组合对照也支持了MACC与阶段性调度的配合作用。

本文方法的不足在于MACC生成的辅助上下文与主路径特征均来源于可用模态，因此补偿效果依赖可用特征的质量。当剩余模态受到噪声或模糊干扰时，辅助上下文可能同时受到影响。后续将设置不同强度的噪声和模糊，对比辅助上下文与检测结果的变化，进一步评估这种情况下的补偿效果。

现有实验覆盖完整输入和单路完全缺失，尚未系统考察传感器连续退化及可用状态随时间变化的场景。后续将模拟图像质量逐步下降和模态间歇缺失，结合图像质量评估研究特征增强与融合方法，并在更多实际场景中验证。同时，将开展模型轻量化研究，以提高传感器状态变化下的检测稳定性和部署效率。

参考文献（References）

Bie Q, Wang X, Xu X, Zhao Q J, Wang Z, Chen J, et al. 2023. Visible-infrared cross-modal pedestrian detection: a summary. Journal of Image and Graphics, 28(5): 1287-1307 (别倩, 王晓, 徐新, 赵启军, 王正, 陈军, 等. 2023. 红外—可见光跨模态的行人检测综述. 中国图象图形学报, 28(5): 1287-1307) [DOI: 10.11834/jig.220670]

Bie Q, Wang X, Yang B, Yu Z X, Chen J and Xu X. 2026. Robust pedestrian detection with uncertain modality//Proceedings of the AAAI Conference on Artificial Intelligence. Singapore: AAAI Press: 2462-2470 [DOI: 10.1609/aaai.v40i4.37232]

Cao Y, Bin J, Hamari J, Blasch E and Liu Z. 2023. Multimodal object detection by channel switching and spatial attention//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops. Vancouver: IEEE: 403-411 [DOI: 10.1109/

© 中国图象图形学报版权所有

CVPRW59228.2023.00046]

Carion N, Massa F, Synnaeve G, Usunier N, Kirillov A and Zagoruyko S. 2020. End-to-end object detection with transformers//Proceedings of the European Conference on Computer Vision. Cham: Springer: 213-229 [DOI: 10.1007/978-3-030-58452-8_13]

Fang Q, Han D and Wang Z. 2021. Cross-modality fusion transformer for multispectral object detection[EB/OL]. [2026-06-16]. https://arxiv. org/abs/2111.00273 [DOI: 10.48550/arXiv. 2111.00273]

Fu H L, Wang S X, Duan P H, Xiao C Y, Dian R W, Li S T, et al. 2024a. LRAF-Net: long-range attention fusion network for visible-infrared object detection. IEEE Transactions on Neural Networks and Learning Systems, 35(10): 13232-13245 [DOI: 10.1109/TNNLS.20 23.3266452]

Fu H L, Yuan J, Zhong G J, He X, Lin J C and Li Z Y. 2024b. CF-Deformable DETR: an end-to-end alignment-free model for weakly aligned visible-infrared object detection//Proceedings of the Thirty-Third International Joint Conference on Artificial Intelligence. Jeju: IJCAI: 758-766 [DOI: 10.24963/ijcai.2024/84]

Helvig K, Abeloos B and Trouvé-Peloux P. 2024. CAFF-DINO: multispectral object detection transformers with cross-attention features fusion//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops. Seattle: IEEE: 3037-3046 [DOI: 10.1109/CVPRW63382.2024.00309]

Hwang S, Han D and Jeon M. 2025. Multispectral detection transformer with infrared-centric sensor fusion[EB/OL]. [2026-06-16]. https://arxiv. org/abs/2505.15137v1 [DOI: 10.48550/arXiv. 2505.15137]

Jia X Y, Zhu C, Li M Z, Tang W Q and Zhou W L. 2021. LLVIP: a visible-infrared paired dataset for low-light vision//Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops. Montreal: IEEE: 3496-3504 [DOI: 10.1109/ICCVW54120. 2021.00389]

Lee S, Park J H and Park J S. 2024. CrossFormer: cross-guided attention for multi-modal object detection. Pattern Recognition Letters, 179: 144-150 [DOI: 10.1016/j.patrec.2024.02.012]

Li K, Wang D, Hu Z Y, Li S F, Ni W P, Zhao L, et al. 2025. FD2-Net: frequency-driven feature decomposition network for infrared-visible object detection//Proceedings of the AAAI Conference on Artificial Intelligence. Philadelphia: AAAI Press: 4797-4805 [DOI: 10.1609/aaai.v39i5.32507]

Medeiros H R, Latortue D, Granger E and Pedersoli M. 2025. Mixed patch visible-infrared modality agnostic object detection//Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision. Tucson: IEEE: 9023-9032 [DOI: 10.1109/WACV61041.2025.00874]

Shen J F, Chen Y F, Liu Y, Zuo X, Fan H and Yang W K. 2024. ICAFusion: iterative cross-attention guided feature fusion for multispectral object detection. Pattern Recognition, 145: 109913 [DOI: 10.1016/j.patcog.2023.109913]

Shen J F, Zhan H B, Dong S H, Zuo X, Yang W K and Ling H B. 2026. Multispectral state-space feature fusion: bridging shared and cross-parametric interactions for object detection. Information Fusion, 127: 103895 [DOI: 10.1016/j.inffus.2025.103895]

Sun X H, Guan Z and Wang X. 2023. Vision Transformer for fusing infrared and visible images in groups. Journal of Image and Graphics, 28(1): 166-178 (孙旭辉,官铮,王学. 2023. 红外与可见光图像分组融合的视觉Transformer. 中国图象图形学报,28(1): 166-178) [DOI: 10.11834/jig.220515]

Tang L F, Zhang H, Xu H and Ma J Y. 2023. Deep learning-based image fusion: a survey. Journal of Image and Graphics, 28(1): 3-36 (唐霖峰,张浩,徐涵,马佳义. 2023. 基于深度学习的图像融合方法综述. 中国图象图形学报,28(1): 3-36) [DOI: 10.11834/jig.220422]

Teledyne FLIR. 2018. FLIR ADAS thermal dataset[EB/OL]. [2026-06-16]. https://www.flir.com/oem/adas/adas-dataset-form/

Wan D H, Lu R S, Fang Y, Lang X L, Shu S B, Chen J J, et al. 2025. YOLOv11-RGBT: towards a comprehensive single-stage multispectral object detection framework[EB/OL]. [2026-06-16]. https://arxiv. org/abs/2506.14696 [DOI: 10.48550/arXiv. 2506.14696]

Xiao Y M, Meng F M, Wu Q B, Xu L F, He M Z and Li H L. 2024. GM-DETR: generalized multispectral detection transformer with efficient fusion encoder for visible-infrared detection//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops. Seattle: IEEE: 5541-5549 [DOI: 10.1109/CVPRW63382.2024.00563]

Xu K, Liu X P, Wang H Y, Wan J W and Guo Y L. 2024. Infrared-visible image object detection algorithm using feature dynamic selection. Journal of Image and Graphics, 29(8): 2350-2363 (许可,刘心溥,汪汉云,万建伟,郭裕兰. 2024. 红外与可见光图像特征动态选择的目标检测网络. 中国图象图形学报,29(8): 2350-2363) [DOI: 10.11834/jig.230495]

Yuan M X, Cui B, Zhao T Y, Wang J Y, Fu S, Yang X, et al. 2025. UniRGB-IR: a unified framework for visible-infrared semantic tasks via adapter tuning//Proceedings of the 33rd ACM International Conference on Multimedia. Dublin: ACM: 2409-2418 [DOI: 10.1145/3746027.3754806]

Zhang H, Li F, Liu S L, Zhang L, Su H, Zhu J, et al. 2023. DINO: DETR with improved denoising anchor boxes for end-to-end object detection//Proceedings of the Eleventh International Conference on Learning Representations. Kigali: OpenReview

Zhang R, Yao L, Zhang Y X, Wang Y J, Zhang C Y and Liu F. 2026. Non-spatial registration decision fusion for multimodal object detection. Journal of Image and Graphics, 31(2): 541-555 (张荣,姚亮,张奕欣,王翌骏,张传一,刘凡. 2026. 非空间配准的多模态目标检测决策融合策略. 中国图象图形学报,31(2): 541-555)

© 中国图象图形学报版权所有

［DOI：10.11834/jig.250326］

Zhao T Y，Yuan M X，Jiang F，Wang N and Wei X X. 2026. Removal then selection：a coarse-to-fine fusion perspective for RGB-infrared object detection. IEEE Transactions on Intelligent Transportation Systems，27(2)：2504-2519［DOI：10.1109/TITS.2025.3638627］

Zhao Y A，Lv W Y，Xu S L，Wei J M，Wang G Z，Dang Q Q，et al. 2024. DETRs beat YOLOs on real-time object detection//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle：IEEE：16965-16974［DOI：10.1109/CVPR52733.2024.01605］

Zhu X Z，Su W J，Lu L W，Li B，Wang X G and Dai J F. 2021. Deformable DETR：deformable transformers for end-to-end object detection//Proceedings of the Ninth International Conference on Learning Representations. Virtual：OpenReview

作者简介

李建东，男，副教授，研究方向为图像与视觉信息计算、智能数据处理。E-mail：lijiandong1226@163.com

曲海成，男，副教授，主要研究方向为遥感影像高性能计算、视觉信息计算、目标检测与识别。E-mail：quhaicheng@lntu.edu.cn；

武聪，男，副教授，主要研究方向人工智能、智能信息处理。E-mail：wucong@lntu.edu.cn；

© 中国图象图形学报版权所有