/TemplateVersion (2026.1)

DOD-SA: 红外-可见光解耦对象检测与单模态标注

Hang Jin[1], Chenqiang Gao[1*], Junjie Guo[2], Fangcen Liu[2], Kanghui Tian[1], Qinyao Chang[3]
[1] Sun Yat-sen University
[2] Chongqing University of Posts and Telecommunications
[3] University of Electronic Science and Technology of China

Abstract

红外-可见物体检测在实际应用中显示出巨大潜力，它利用红外和可见图像的互补信息，实现全天候的稳健感知。然而，现有方法通常需要双模态标注，以在预测时输出两种模态的检测结果，这会带来高昂的标注成本。为了解决这一挑战，我们提出了一种创新的红外-可见物体检测框架，使用单模态标注，称为 DOD-SA。DOD-SA 的架构建立在一个单模态和双模态协作的教师-学生网络（CoSD-TSNet）之上，包含一个单模态分支（SM-Branch）和一个双模态解耦分支（DMD-Branch）。教师模型为未标注模态生成伪标签，同时支持学生模型的训练。该协作设计实现了从标注模态到未标注模态的跨模态知识转移，并促进了有效的 SM 到 DMD 分支监督。为了进一步提高模型的解耦能力和伪标签质量，我们引入了一种渐进和自调节的训练策略（PaST），该策略分三个阶段训练模型：（1）预训练 SM-Branch，（2）由 SM-Branch 指导 DMD-Branch 的学习，以及（3）细化 DMD-Branch。此外，我们设计了一个伪标签分配器（PLA），以对齐和配对跨模态的标签，在训练期间明确解决模态不对齐的问题。在 DroneVehicle 数据集上的大量实验表明，我们的方法优于最先进的（SOTA）方法。

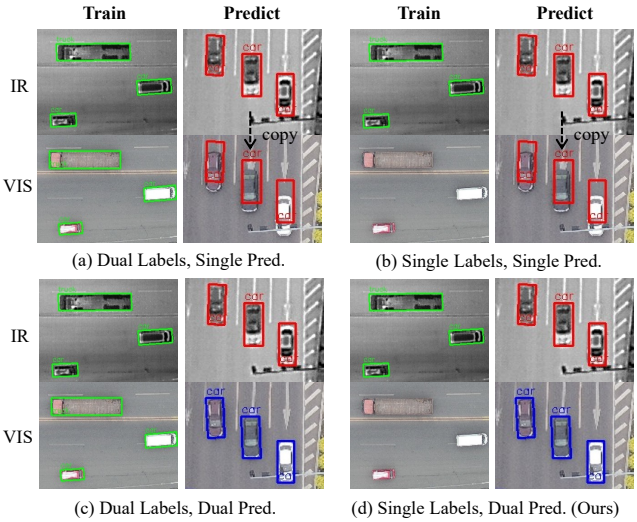

Figure 1: 多光谱目标检测方法的比较。绿色框：真实值；红色/蓝色框：预测值。

介绍

红外-可见光（IR-VIS）目标检测由于其能够利用互补的光谱信息实现全天候的稳健感知而在近年来引起了越来越多的关注 (?????) 。然而，在实际应用中，红外和可见光图像由于成像时间或物体运动的不同而不可避免地出现错位 (?) ，即使它们已经通过图像裁剪 (?) 和仿射变换 (?) 等操作进行人为预对齐。这种错位扰乱了两种模态的特征学习，从而影响了模型的检测性能。

为了解决这个问题，一些方法 (???) 利用来自两种模态的标签，在图像或特征层面显式学习跨模态偏移，促进信息融合。同时，其他方法 (????) 通过利用单一模态的标注使模型能够隐式学习这种偏移。然而，这些方法在两种模态中共享同一组边界框预测，如图 1（a）和（b）所示。

在实际应用中，目标检测通常作为下游处理的初步步骤。为了确保任务的可靠性，许多应用不能仅依靠单一模态的检测结果。例如，在红外和可见光场景中跟踪目标时，如果目标在某种模态中定位不佳或不可见，仅依赖一种模态的检测结果可能导致跟踪偏移或目标丢失。在这种情况下，结合其他模态的补充结果可以有效减轻这些问题。因此，为每种模态输出解耦的目标位置是至关重要的，特别是对于多模态目标跟踪、行为分析和无人机（UAVs）的自主导航。最近，DPDETR (?) 被提出来解决解耦问题。如图 1 (c) 所示，该模型可以在预测过程中输出每种模态中每个对象的位置。然而，该方法需要来自两种模态的注释，并且手动配对模态间的框的过程显著增加了注释成本和复杂性。

为此，我们提出了一种新的红外-可见光解耦目标检测框架，使用单模态标注（DOD-SA）。如图 1 (d) 所示，DOD-SA 在训练期间只需要一种模态的标签即可实现解耦预测。具体来说，受半监督对象检测 (????) 进展的启发，我们提出了一种单模态和双模态协作教师-学生网络（CoSD-TSNet）。该框架利用教师模型为未标记模态生成伪标签，解决了模态注释缺失的问题。此外，教师与学生之间的互学习机制增强了特征表示，同时引入单模态检测分支（SM-Branch）来增强双模态解耦分支（DMD-Branch）的训练，以确保稳健的协作学习。为了进一步提高模型的解耦能力和伪标签的质量，我们提出了一种渐进和自调节训练策略（PaST），该策略包括三个训练阶段：首先对 SM-Branch 进行预训练，然后由 SM-Branch 指导 DMD-Branch 的训练，最后通过自我

*Corresponding author.

增强学习对 DMD-Branch 进行精炼。为了过滤和校正伪标签，我们设计了一个专用的伪标签分配器（PLA）。PLA 模块使用感知形状的启发式方法，通过搜索区域将未标记模态的伪标签与标记模态的真实值匹配，同时维持一个动态伪标签包，以在训练过程中保存高质量的标签对。该模块通过直接的边框位置校正显式地解决了模态未对齐的问题，使模型能够实现更好的互补表示学习。在 DroneVehicle 数据集上的大量实验 (?) 表明，即使仅使用单模态注释，我们的方法也能达到 SOTA 的性能表现，甚至超过了完全监督的双模态 SOTA 方法。总之，我们的贡献可以总结如下：

- 据我们所知，我们是第一个提出一种新型框架，称为 DOD-SA，该框架使用单模态注释在红外和可见光模态上实现精确解耦的目标检测。
- 我们设计了一种单模态和双模态协作教师-学生网络（CoSD-TSNet），利用教师模型为未标记的模态生成伪标签。该框架还使单模态分支（SM-Branch）能够在统一的渐进和自调谐训练策略（PaST）中引导双模态解耦检测分支（DMD-Branch），进一步提高模型的解耦检测性能。
- 我们引入伪标签分配器（PLA），用于将未标记模态中的伪标签与已标记模态的真实标签匹配，获得高质量的标签对，这也明确解决了模态不一致的问题。
- 我们在 DroneVehicle 基准上验证了 DOD-SA，并展示了它在降低标注成本和提高鲁棒性方面达到了最先进的性能。

相关工作

弱对准条件下的红外-可见光目标检测

为了解决模态失配问题，现有的方法主要集中在学习跨模态偏移，这可以根据两种模态是否都有注释分为两类：双模态监督和单模态监督。

双模态监督。AR-CNN (?) 首先通过使用配对标注显式地学习几何偏移来解决模态失配问题，而 TSFADet (?) 和 CAGTDet (?) 在考虑比例/角度偏移的情况下提高了对齐。DPDETR (?) 通过模态特定查询实现了解耦预测。然而，这些方法依赖于耗费高昂的配对标注，在实际应用中限制了可扩展性。

单模态监督。C ^{2} 前 (?) 计算参考和感知模态特征之间的注意力进行融合。DAMSDet (?) 引入了一个可变形的跨注意力模块以提取细粒度的互补信息。OAFA (?) 通过分离公共和私人特征来学习偏移。Zhang et al. (?) 提出了一个基于平均教师的框修正模块以增强表示。这些方法使用单模态标签进行监督，但它们仅预测一个模态中的位置。我们的方法能够在单模态标签的情况下进行双模态解耦预测，创建了一种新的实用设置。

半监督目标检测方法 (SSOD)

SSOD 方法旨在通过利用无标签数据来降低标注成本，通常通过一致性正则化或伪标签的方法，其中伪标签的方法更为广泛采用。STAC (?) 引入了一个两阶段训练范式。Unbiased Teacher (?) 引入了焦点损失 (?) 和指数移动平均 (EMA) 以解决类别不平衡问题并提高伪标签质量。Soft Teacher (?) 使用伪标签的得分作为损失的权重。Efficient Teacher (?) 将伪标签分为确定和不确定标签。尽管取得了显著进展，现有的 SSOD 方法几乎只专注于单模态目标检测。本文中，我们将教师-学生网络应用于双模态目标检测，以解决单模态标注问题。

经典的标签分配方法，例如 ATSS (?) 、PAA (?) 、AutoAssign (?) 、OTA (?) 和 TaskAlignedAssigner (?) ，在完全监督的设置下显示了强大的性能。然而，最近的研究 (??) 表明，直接将这些策略应用于半监督目标检测可能导致性能下降，主要原因是存在噪声伪标签和模态不一致。在本文中，我们提出了一种伪标签分配器 (PLA)，以便在单模态标注条件下实现对双模态目标检测的应用。

方法

整体架构

如图 ?? 所示，我们的 DOD-SA 基于单模态与双模态协同师生网络（CoSD-TSNet），旨在有效解决缺失模态注释的问题。训练过程遵循由三个阶段组成的渐进和自调节训练策略（PaST）。在每个阶段，模型的特定分支被激活，并使用不同的损失函数来进一步提高伪标签的质量。此外，我们在第 2 阶段和第 3 阶段设计了一个伪标签分配器（PLA），用于生成解耦的标签对。此外，无论训练集中仅包含可见光注释还是仅包含红外注释，我们的方法都具有普适性。在本节中，我们将红外模态视为有标记模态。进一步的细节将在下面介绍。

单模态和双模态协作教师-学生网络 (CoSD-TSNet)

如图 ?? 所示，CoSD-TSNet 是一个师生框架，其中教师/学生模型由两个参数不共享的流（IR 流和 RGB 流）、单模态解码器（SP-Decoder）和双模态解耦位置解码器（DP-Decoder）组成。具体而言，IR/RGB 流由一个主干组成，并可能添加一个编码器用于特征提取。SP-Decoder 包含一个解码器和检测头，而 DP-Decoder 在这些组件之前进一步包括一个特征融合模块。对于解码过程，SP-Decoder 处理单模态特征，而 DP-Decoder 整合来自两种模态的特征。此外，通过 SP-Decoder 的分支称为单模态分支（SM-Branch），而通过 DP-Decoder 的分支称为双模态分支（DMD-Branch）。

对于 SM-Branch，它类似于一个单模态检测器，一次只能输入来自单个模态 $x_m \in R^{H\times W\times 3}, m \in \{rgb, ir\}$ 的图像，并输出一个分类结果 $\hat{c}$ 和该模态的位置结果 $\hat{b}_m$ ，其中 H × W 代表空间分辨率。对于 DMD-Branch，它作为一个解耦位置的双模态检测器运作，能够同时处理来自两个模态的图像 $\{x_{ir}, x_{rgb}\}$ 并输出一个分类结果 $\hat{c}$ 和两个模态的解耦位置结果 $\{\hat{b}_{ir}, \hat{b}_{rgb}\}$ 。

在我们的框架中，教师模型负责为无标注模态挖掘伪标签，以解决缺少一个模态标注的问题。通过指数移动平均（EMA）更新机制，随着学生模型的训练，教师模型的结果得到了改善。我们设计了这种单模态和双模态协作结构，利用单模态检测器来指导双模态检测器的训练。此外，它通过有效利用现有的标注数据，促进了从标注模态到非标注模态的知识转移。

渐进和自调训练策略（PaST）

为了逐步挖掘无标记模态的高质量伪标签并持续提高模型性能，我们提出了一种渐进自调整训练策略（PaST）。如图 ?? 所示，PaST 包括三个阶段：(1) 阶段 1。该阶段旨在初步训练 SM 分支，以发展生成可

见伪标签的初步能力。(2) 阶段 2。SM 分支引导 DMD 分支的训练，使其具备一定的解耦检测能力。(3) 阶段 3。我们优化整个 DMD 分支以细化红外和可见模态的检测结果。每个阶段具体解释如下。

阶段 1。此阶段仅激活 SM-分支，由两个部分组成：预热和教师-学生互相学习。预热部分旨在对标注数据进行模型训练，以建立标注模态的初步检测能力。在教师-学生互相学习部分，知识从标注模态传递到未标注模态，从而提高模型在未标注模态上的检测性能。

预热阶段：将标记的 IR 数据 x_{ir} 输入到学生模型的 SM-Branch 中，以获得 IR 预测 $\hat{y}^s_{ir} = \{\hat{c}, \hat{b}_{ir}\}$ ，这些预测由 IR 的真实值 $y_{ir} = \{c, b_{ir}\}$ 监督：

$$\mathcal{L}_{sup} = \mathcal{L}_{cls}(\hat{c}, c) + \mathcal{L}_{box}(\hat{b}_{ir}, b_{ir}), \tag{1}$$

其中 $\mathcal{L}_{cls}$ 是分类损失，$\mathcal{L}_{box}$ 是旋转边界框回归损失。这个部分会在 $\mathcal{K}_1$ 个 epochs 中继续进行，使用的损失函数为 $\mathcal{L}_{sup}$ 。

师生互学习：学生模型的 SM 分支处理经过强增强的无标签 RGB 数据 x^*_{rgb} ，而教师模型的 SM 分支接收同一数据的弱增强版本 x'_{rgb} 。教师模型输出 RGB 伪标签 $\tilde{y}^s_{rgb} = \{\tilde{c}, \tilde{b}_{rgb}\}$ ，而学生模型输出 RGB 预测 $\hat{y}^s_{rgb} = \{\hat{c}, \hat{b}_{rgb}\}$ 。为了获得更可靠和确定的伪标签，我们过滤掉低于批次自适应阈值的伪标签，该阈值在提出的伪标签分配器（PLA）中进行了说明。然后，我们使用过滤后的伪标签来监督学生模型的 RGB 预测：

$$\mathcal{L}_{unsup} = \mathcal{L}_{cls}(\hat{c}, \tilde{c}) + \mathcal{L}_{box}(\hat{b}_{rgb}, \tilde{b}_{rgb}). \tag{2}$$

这一部分进行了 $\mathcal{K}_2$ 个周期，使用的损失函数是 $\mathcal{L}_{sup} + \lambda\mathcal{L}_{unsup}$ 。在这里，λ 是一个自适应权重，随着训练的进行从 0 线性增加到 1.0。

第二阶段。我们保留第一阶段激活的模块，并额外激活学生模型中的 DMD-分支，在此我们将 IR 流的参数复制到 RGB 流。输入 x_{ir} and x^*_{rgb} 同时输入到学生模型的 DMD-分支，输出解耦后的检测结果 $\hat{y}^d = \{\hat{c}, \hat{b}_{ir}, \hat{b}_{rgb}\}$ 。输入 x'_{rgb} 被送入教师模型的 SM-分支，输出 RGB 伪标签 $\tilde{y}^s_{rgb} = \{\tilde{c}, \tilde{b}_{rgb}\}$ 。我们提出 PLA 以匹配 IR 真实值 $y_{ir} = \{c, b_{ir}\}$ 和 RGB 伪标签 $\tilde{y}^s_{rgb} = \{\tilde{c}, \tilde{b}_{rgb}\}$ 以形成可用的监督标签 $y^d = \{c, b_{ir}, \tilde{b}_{rgb}\}$ 。此 PLA 将在下面详细介绍。随后，我们使用匹配的标签来监督 DMD-分支：

$$\mathcal{L}_{paired} = \mathcal{L}_{cls}(\hat{c}, c) + \mathcal{L}_{boxir}(\hat{b}_{ir}, b_{ir}) + \mathcal{L}_{boxvis}(\hat{b}_{rgb}, \tilde{b}_{rgb}). \tag{3}$$

此外，我们还采用了阶段 1 中提到的 $\mathcal{L}_{sup}$ 和 $\mathcal{L}_{unsup}$ ，只是 $\mathcal{L}_{unsup}$ 中的 $\tilde{b}_{rgb}$ 现在是 PLA 的 RGB 结果。

第二阶段继续 $\mathcal{K}_3$ 个周期，该阶段的总体损失函数为 $\mathcal{L}_{sup} + \mathcal{L}_{unsup} + \mathcal{L}_{paired}$ 。第三阶段。此阶段仅激活教师和学生模型中的 DMD 分支。输入 x_{ir} 和 x'_{rgb} 同时输入教师模型的 DMD 分支，输出解耦的伪标签 $\tilde{y}^d = \{\tilde{c}, \tilde{b}_{ir}, \tilde{b}_{rgb}\}$ 。类似地，学生模型输出解耦的预测结果 $\hat{y}^d = \{\hat{c}, \hat{b}_{ir}, \hat{b}_{rgb}\}$ 。我们从 $\tilde{y}^d$ 中提取 RGB 伪标签 $\tilde{y}_{rgb} = \{\tilde{c}, \tilde{b}_{rgb}\}$ 。然后通过 PLA 与 IR 的真实值 $y_{ir} = \{c, b_{ir}\}$ 匹配，以为学生模型生成监督 $y^d = \{c, b_{ir}, \tilde{b}_{rgb}\}$ 。损失函数 $\mathcal{L}_{paired}$ 与方程（3）相似，除了 $\tilde{b}_{rgb}$ 来自教师模型的 DP 解码器。本阶段继续 $\mathcal{K}_4$ 个周期，损失函数为 $\mathcal{L}_{paired}$ 。

在模型的三阶段训练完成后，仅保留教师网络的 DMD-Branch 结构，以确保在红外和可见光模态上实现稳健有效的解耦目标检测。

如图 2 所示，我们的伪标签分配器（PLA）由三个子模块组成：伪标签过滤器（PLF）、形状感知双模态标签匹配器（SDLM）和动态标签修正（DLC）。这些子模块结合起来，通过匹配红外地面真值与可见伪标签，为 DMD 分支提供双模态监督信息。伪标签过滤器（PLF）。我们将每个 RGB 伪标签的分类概率视为评分：

$$\text{score}_i = \max_k p_{i,k}(\hat{c}), \tag{4}$$

其中 $p_{i,k}(\hat{c})$ 表示 i 样本被分类为类别 k 的概率。随后，类似于 (?) ，我们使用批自适应阈值来过滤掉低质量样本，其计算如下：

$$\tau = \mu - \sigma, \tag{5}$$

$$\mu = \frac{1}{N}\sum_{i=1}^{N}\text{score}_i, \tag{6}$$

$$\sigma = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(\text{score}_i - \mu)^2}, \tag{7}$$

其中 τ 、μ 和 σ 分别表示评分阈值、均值和方差，N 表示一批中的样本数量。

形状感知双模态标签匹配器（SDLM）。在滤除高质量 RGB 伪标签之后，我们提出了一种形状感知双模态标签匹配器（SDLM），用于将 IR 真实边界框集 $\{b_i^{\text{ir}}\}_{i=1}^{N}$ 与 RGB 伪标签集 $\{b_j^{\text{rgb}}\}_{j=1}^{M}$ 配对，其中 N 和 M 分别表示它们的数量。每个边界框 $b_i = (ct_{i,x}, ct_{i,y}, w_i, h_i, \theta_i)$ 由中心坐标 $ct_i = (ct_{i,x}, ct_{i,y})$ 、宽度 w_i 、高度 h_i 和旋转角度 θ_i 组成。对于每个 IR 真实框 b_i^{ir} ，我们定义一个搜索区域：

$$\mathcal{S}_i = \left\{ (x,y) \,\middle|\, (x,y) = (ct_{i,x}, ct_{i,y}) + R(\theta_i) \cdot \begin{bmatrix} \Delta x \\ \Delta y \end{bmatrix} \right\}, \tag{8}$$

$$(\Delta x, \Delta y) \in \left\{ \left(\pm\frac{\beta w_i}{2}, \pm\frac{\beta h_i}{2} \right) \right\}, \tag{9}$$

$$R(\theta_i) = \begin{bmatrix} \cos\theta_i & -\sin\theta_i \\ \sin\theta_i & \cos\theta_i \end{bmatrix}, \tag{10}$$

其中 $R(\theta_i)$ 是一个逆时针旋转矩阵。β 表示调整搜索区域长宽比的超参数。我们使用该搜索区域来保留具有内部中心的候选者，同时排除那些已经匹配的，获得 RGB 候选集 C_i^{rgb} ：

$$C_i^{\text{rgb}} = \left\{ b_j^{\text{rgb}} : ct_j \in \mathcal{S}_i \text{ and } b_j^{\text{rgb}} \notin P \right\}, \tag{11}$$

其中 P 是已配对的框集。然后我们计算每个 IR 真实框与 C_i^{rgb} 中的所有 RGB 框之间的交并比（IoU），然后将 IR 框与具有最高 IoU 的 RGB 候选者配对。由于跨模态的物体形状几乎相同，最佳重叠的 RGB 伪标签可能对应于相同的物体。

$$b_j^* = \arg\max_{b_j^{\text{rgb}} \in C_i^{\text{rgb}}} \text{IoU}\left(b_i^{\text{ir}}, b_j^{\text{rgb}}\right), \tag{12}$$

其中 b_j^* 表示成功匹配的可见伪标签框。

动态标签校正 (DLC)。考虑到仅依赖匹配的标签对会导致监督不足，我们引入不匹配的情况，并维护一个动态更新的标签对集合，以获得更多高质量的标签对。

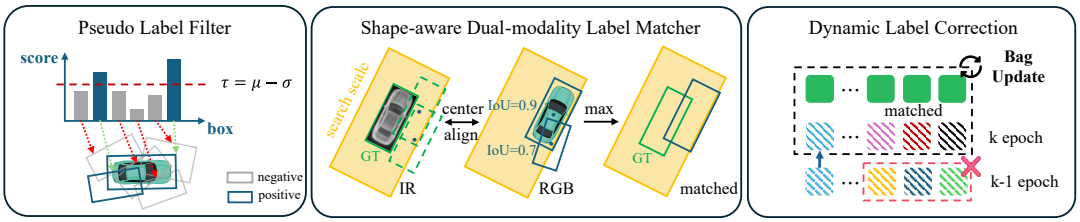

Figure 2: 伪标签分配器。它由三个组件组成：(1) 伪标签过滤器。用于筛选出高质量的伪标签。(2) 形状感知双模态标签匹配器。将每个红外地面实况边界框与其对应的可见伪标签进行匹配。(3) 动态标签校正。我们采用动态更新的标签对集合来不断校正伪标签。

在第 2 阶段的第一个时期，我们将未匹配的红外真实边界框 U_i^{ir} 复制到可见光模式。这些与匹配的标签对 P_0 结合，以初始化双模态解耦的真实值 G_0 。

$$U_i^{\text{ir}} = \left\{ b_j^{\text{ir}} : b_j^{\text{ir}} \notin P_0 \right\}, \tag{13}$$

$$G_0 = P_0 \cup (U_i^{\text{ir}}, U_i^{\text{ir}}). \tag{14}$$

在第 k 个周期中，令 (b_k^{ir} 和 b_k^{rgb}) 为 G_k 中的一个标签对，其中 G_k 表示第 k 个周期中动态更新的红外-可见光标签对集合。如果在下一个周期中，b_k^{ir} 成功与一个新的 RGB 伪标签 b_k^{rgb*} 匹配，则该标签对将更新为 (b_k^{ir} , b_k^{rgb*}) 。否则，标签对保持不变。

实验

数据集和评价指标

数据集。我们在 DroneVehicle 数据集 (?) 上进行实验，遵循最近的 SOTA 方法 (?????) 。该数据集是一个大规模的目标检测基准，专注于无人机捕获的配对红外和可见光图像。它包含 28,439 对 RGB-IR 图像。每种模态都有自己的一组带方向的边界框标注，覆盖五种车辆类型：汽车、公交车、卡车、货车和货运车。为了标准化实验协议，我们遵循 (?) 中描述的方法来处理标注。在预处理之后，数据集包括 17,990 对用于训练的图像和 1,469 对用于评估的图像。

评价指标。我们采用广泛使用的 IoU 阈值为 0.5 的平均精度均值（mAP）作为 DroneVehicle 数据集的检测准确性指标。

我们选择 DPDETR (?) 作为我们的基线模型，它具有出色的解耦检测能力。因此，我们模型中的红外和 RGB 流由一个预训练的 ResNet50 (?) 和一个特征图语义级别为 L=3 的高效编码器组成。SP-Decoder 遵循 RT-DETR (?) 的解码器，而 DP-Decoder 遵循 DPDETR 的解码器。两个解码器都包含六层，我们将注意力头的数量和选择的查询数分别设置为 H=8 和 N=300。对于模型训练，我们使用带有权重衰减为 $1 \times e^{-4}$ 的 AdamW 优化器。初始学习率设置为 $1 \times e^{-7}$ ，通过 LinearWarmup 升温到 $1 \times e^{-4}$ ，最终通过 PiecewiseDecay 减小到 $1 \times e^{-5}$ 。弱数据增强包括随机旋转和翻转。同时，强增强包括光度变换、空间遮挡和模糊增强。我们的模型输入大小设置为 640 × 640，批量大小为 16。EMA 衰减权重为 0.9999。训练轮次设置为：$\mathcal{K}_1 = 20$ 、$\mathcal{K}_2 = 10$ 、$\mathcal{K}_3 = 15$ 和 $\mathcal{K}_4 = 20$ 。在伪标签分配器中的超参数 β 设置为 1.0。我们在 Nvidia RTX A6000 GPU 上进行实验。

Model	Train image	Train label	Infrared-Test	Visible-Test
FSSD	IR	IR	72.72	-
YOLOX-X	IR	IR	73.24	-
Oriented RepPoints	IR	IR	68.00	-
RoITransformer	IR	IR	72.86	-
RTDETR(OBB)	IR	IR	77.93	-
FSSD	RGB	RGB	-	62.81
YOLOX-X	RGB	RGB	-	61.20
Oriented RepPoints	RGB	RGB	-	62.30
RoI Transformer	RGB	RGB	-	61.60
RTDETR(OBB)	RGB	RGB	-	72.92
AR-CNN (OBB)	IR+RGB	IR+RGB	71.58	-
TSFADet	IR+RGB	IR+RGB	73.06	-
C ^{2} former	IR+RGB	IR+RGB	74.20	-
CAGTDet	IR+RGB	IR+RGB	74.57	-
DPDETR	IR+RGB	IR+RGB	79.90	79.81
Halfway Fusion(OBB)	IR+RGB	IR	68.19	-
CIAN(OBB)	IR+RGB	IR	70.23	-
MC-DETR	IR+RGB	IR	76.90	-
DDCINet	IR+RGB	IR	78.40	-
CCLDet	IR+RGB	IR	79.40	-
DOD-SA(Ours)	IR+RGB	IR	80.41	78.87
CALNet	IR+RGB	RGB	-	76.41
E2E-MFD	IR+RGB	RGB	-	77.40
M2FP	IR+RGB	RGB	-	78.70
DOD-SA(Ours)	IR+RGB	RGB	77.96	78.19

Table 1: 无人机车辆数据集上的检测结果（mAP，以%表示）。最佳结果以粗体显示。Infrared-Test 表示训练模型在红外图像上的测试结果，而 Visible-Test 代表在可见光图像上的测试结果。

与最新技术的比较

我们将提出的 DOD-SA 与五种最先进的单模态检测器进行比较，包括 FSSD (?) 、YOLOX-X (?) 、Oriented RepPoints (?) 、RoI Transformer (?) 和 RTDETR (?) 。我们还与十三种多光谱物体检测方法进行比较，包括 AR-CNN (?) 、TSFADet (?) 、C ^{2} former (?) 、CAGTDet (?) 、DPDETR (?) 、Halfway Fusion (?) 、CIAN (?) 、MC-DETR (?) 、DDCINet (?) 、CCLDet (?) 、CALNet (?) 、E2E-MFD (?) 和 M2FP (?) 。

定量比较。如表 1 所示，DOD-SA 在相同训练标签条件下优于单模态方法。与双模态网络相比，它在仅使用红外标签训练时实现了最先进的性能（红外为 80.41 %，RGB 为 78.87 %），在红外测试中超过 CLDet 1 % mAP。即使在 RGB 标签上进行训练，我们的方法也能产生强大的解耦结果（红外为 77.96 %，RGB 为 78.19 %），展示了有效的模态泛化。此外，我们的方法优于 DPDETR 和其他双注释方法，因为教师-学生架构充当一致性正则化器，丰富了学习表示。

定性比较。图 3 展示了不同方法的一些代表性检测结果，可以观察到 C ^{2} Former 仅对一种模态产生结果，

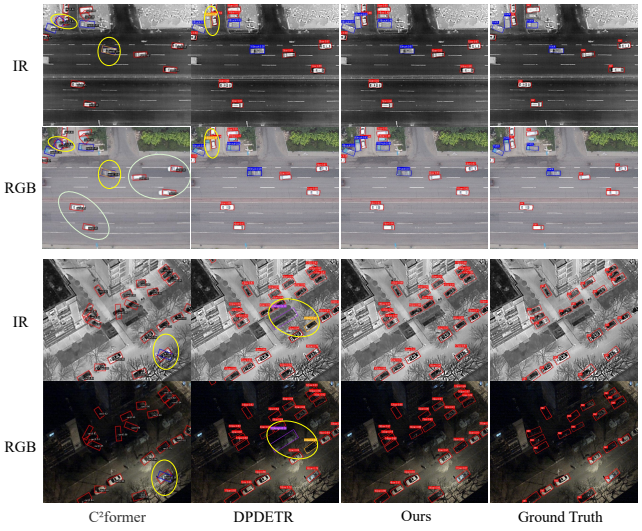

Figure 3: 在 DroneVehicle 数据集上的不同方法的代表性结果：白天场景（上）与夜间场景（下）。黄色圆圈表示误报，浅绿色圆圈表示定位错误（解耦失败）。

导致另一模态出现较大的位置误差。相比之下，即使在黑暗或高度错位的场景中，我们的 DOD-SA 能够准确预测红外和可见光模态中的解耦对象位置。我们的方法匹配或优于 DPDETR，后者偶尔会产生预测错误。这是因为 DOD-SA 有效地生成伪标签并利用互补特征，实现了在两种模态中的强大性能。

消融研究

PaST 的效果。如表所示 2，我们通过消除不同的训练阶段（每阶段 50 个 epochs）来分析每个阶段在 PaST 中的必要性。比较我们的实验和 Exp. IV 显示阶段 1 提高了 mAP 0.67 %，因为它帮助单模态检测器开发初始伪标签挖掘能力，为阶段 2 提供更好的指导。Exp. III vs. Exp. IV 验证了 SM-Branch 在指导阶段 2 中 DMD-Branch 训练的重要性。最后，Exp. V vs. 我们的实验表明阶段 3 通过师生互相学习提高了 mAP 0.86 %，进一步精炼了 DMD-Branch。

PLF 的效果。我们评估了伪标签过滤器（PLF）中阈值 τ 变化的影响。如表 3 所示，没有伪标签过滤（无 PLF）时，模型性能显著下降。mAP 值随着不同阈值的变化而变化，这表明固定的 τ 无法平衡伪标签的数量和质量。通过分布自适应动态调整 τ 能产生最佳结果，这显示了 PLF 对于模型训练的重要性。

SDLM 的效果。为了验证我们形状感知双模态标签匹配器（SDLM）的有效性，我们进行了若干对比实验。IoU 匹配策略使用了一种类似于应用于 DroneVehicle 数据集的标签配对方法。结果如表 4 所示。使用标注模态中的真实标签作为伪标签（没有 DLM）相比完全不使用伪标签将 mAP 提高了 0.64 %。我们的基线，即 IoU 匹配策略，通过解决标签分配问题，将 mAP 提高到 79.55 %。当我们应用 DLM 进行标签匹配时，mAP 在 IR 上进一步提高了 0.86 %，在 RGB 上提高了 1.35 %。这些结果证实了我们方法的有效性。

动态标签修正（DLC）的效果。我们进行了一项消融研究，以评估动态标签修正（DLC）的影响。如表 5 所示，仅使用匹配的标签对（不包括 DLC）在红外（IR）上达到了 79.18 % 的 mAP。包括不匹配的标签（不包括 DU）提升了性能，突出了它们的重要性。跨模态边框修正（CBC）(?) 使用逐步中心修正策略来调整未标记模态中的偏移真实框，这可能会引入标签模糊，因为平滑处理的标签可能被错置，从而在训练中误导模型。相比之下，我们的方法动态更新标签包以减轻这一问题，相较于 CBC 在 IR 上提升了 0.76 %，在 RGB 上提升了 0.75 %。伪标签的可视化。为了说明从未标记模态修正伪标签的有效性，我们对模型挖掘出的伪标签进行了可视化。图 ?? （a）和（b）阐释了我们的模型能够修正由物体高速运动引起的偏移。图 ?? （c）显示即使物体静止，也可能由于拍摄角度和无人机运动导致未对齐。然而，我们的方法有效地修正了这一问题。图 ?? （d）、（e）和（f）展示了我们的方法在低光或遮挡等困难条件下依然保持有效性。为了评估我们的方法对模态失配问题的鲁棒性，我们对测试图像集引入了位置偏移。我们固定红外（IR）图像，同时对 RGB 图像沿 x 轴和 y 轴应用空间偏移。像素值的变化范围定义在 $\{(\Delta x, \Delta y) \mid \Delta x, \Delta y \in [15, -15]; \Delta x, \Delta y \in \mathbb{Z}\}$ 。如图 ?? 所示，与基线相比，我们的方法在 IR 测试集上的性能在这些偏移下更稳定。对于 RGB 测试集，两个模型的表现相似，但在小幅度偏移下我们的方法略微更稳定。这表明我们的方法增强了模型对两种模态下每对物体之间对应关系的理解，并使其对位置失配更加稳健。

在本文中，我们提出了 DOD-SA，一种具有单模态标注的新型红外-可见光解耦目标检测框架。我们的协作单模态和双模态师生网络（CoSD-TSNet）利用单模态分支（SM-Branch）通过渐进和自调节训练策略（PaST）来增强双模态解耦分支（DMD-Branch）的训练，从而进一步提升模型性能。此外，我们引入了一种伪标签分配器（PLA），用于将标注模态中的真实标签与从未标注模态中挖掘的伪标签进行匹配。在 DroneVehicle 数据集上的实验表明，DOD-SA 相对于最新的方法具有优越性。消融实验验证了每个组件的必要性，并且我们的方法表现出对位置偏移的强鲁棒性。

	Stage 1	Stage 2	Stage 3	IR	RGB
I	✓			75.75	48.70
II		✓		79.44	77.68
III			✓	79.19	65.59
IV		✓	✓	79.74	78.18
V	✓	✓		79.55	77.77
Ours	✓	✓	✓	80.41	78.87

Table 2: PaST 的消融研究。

	Method	IR	RGB
I	w/o DLC	79.18	77.78
II	w/o DU	79.27	78.57
III	CBC	79.65	78.12
Ours	DLC	80.41	78.87

Table 5: DLC 的消融研究。

	method	IR	RGB
I	w/o PLF	78.99	77.85
II	τ =0.4	79.01	77.72
III	τ =0.6	79.51	78.47
IV	τ =0.8	79.21	78.24
Ours	PLF	80.41	78.87

Table 3: PLF 的消融研究。

	Method	IR	RGB
I	w/o SDLM	79.14	77.47
II	w/o pseudo labels	78.50	12.63
III	IOU Match Strategy	79.55	77.52
Ours	SDLM	80.41	78.87

Table 4: SDLM 的消融研究。