DATA: 域和时间对齐以实现高质量的协同感知特征融合

Chengchang Tian[1*], Jianwei Ma[1*], Yan Huang[1†], Zhanye Chen[1†], Honghao Wei[2], Hui Zhang[1], Wei Hong[1]
[1] Southeast University, Nanjing, China
[2] Washington State University, Pullman, WA, USA
{ chengchang_tian, jianwei_ma, yan_huang, chenzhanye, huizhang, weihong } @seu.edu.cn
honghao.wei@wsu.edu

Abstract

特征级别融合通过平衡性能和通信带宽折衷在协作感知（CP）中显示出前景。然而，其有效性关键依赖于输入特征的质量。高质量特征的获取面临硬件多样性和部署条件带来的领域差距，以及传输延迟导致的时间不对齐。这些挑战导致特征质量在协作网络中不断下降。在本文中，我们提出了域与时间对齐（DATA）网络，旨在系统地对齐特征，同时最大化其用于融合的语义表示。具体而言，我们提出了一种保持一致性的领域对齐模块（CDAM），通过邻近区域的层次化下采样和观察性约束判别器来减少领域差距。我们进一步提出了渐进式时间对齐模块（PTAM），通过多尺度运动建模和双阶段补偿来处理传输延迟。在对齐后的特征基础上，开发了一个实例聚焦的特征聚合模块（IFAM）来增强语义表示。大量实验证明，DATA在三个典型数据集上实现了最先进的表现，并且在严重的通信延迟和姿态误差下保持了鲁棒性。代码将在 https://github.com/ChengchangTian/DATA 发布。

1. 介绍

协同感知 (CP) [? ? ?] 已成为克服单代理感知 [? ?] 固有局限性的重要解决方案，如感知范围有限和遮挡区域。通过使多个代理共享自己的感知信息，CP 能够促进对环境的更全面理解。

对于 CP，中间融合 [? ? ? ? ? ? ?] 在特征层面进行信息共享和集成，以其感知性能和通信带宽之间的平衡权衡而被广泛研究。然而，保持高质量的输入特征进行融合是困难的。在实际世界部署中，在特征获取阶段获得高质量特征充满了显著的挑战 [?] 。硬件异构性 [? ? ? ?] （如 e.g. ，具有不同数量激光束的激光雷达和用于获取点云的不同模式）以及代理条件的差异（如 e.g. ，传感器安装高度和角度）导致代理之间产生不同的数据分布。这种代理之间的数据差异在 CP 中形成了域间差距，如图 1 （a）所示。此外，在

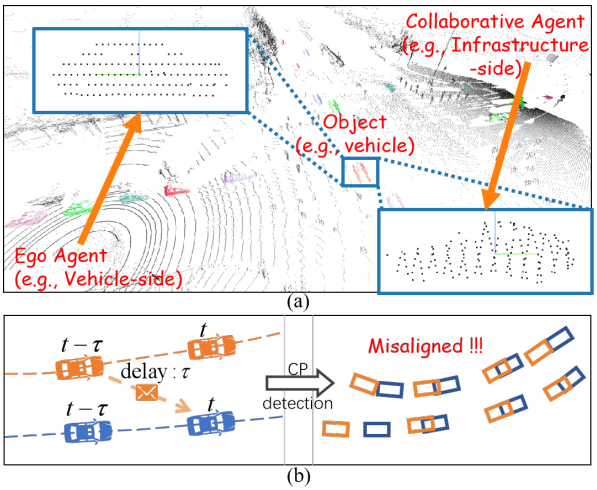

Figure 1. （a）领域差距。例如，不同主体具有不同结构和分散的前景模式。(b）时间错位。例如，在协作过程中的通信延迟以及随后主体之间的观察错位。

代理之间的通信传输期间引入了时间延迟 [? ? ? ?] ，如图 1 （b）所示，导致同一对象的特征错位。域和时间的不对齐不仅显著降低了获取特征的质量和可靠性，还在整个协作系统中表现出累积效应。因此，在特征获取期间进行域和时间对齐对于 CP 的精确性和鲁棒性至关重要。

在特征获取过程中，人们进行了多种尝试来实现领域和时间对齐。针对领域对齐，DI-V2X [?] 通过以随机混合不同来源实例构成的参考领域来实现领域不变的表示。然而，这种混合方法通过扰乱参考领域中对象之间的遮挡关系，损害了物理有效性。针对时间对齐，像 FFNet [?] 这样的基于全局运动流的方法，模型化整个场景以构建未来帧特征。然而，由于对主导背景区域的偏置，它们的全场景优化削弱了细粒度前景运动模式的建模。相反，基于 RoI 的方法，如 CoBEVFlow [?] ，采用局部化的运动预测进行集中建模。然而，它们对区域提案的依赖降低了遮挡处理和召回能力，限制了对整个场景运动动态的理解。

为了提高智能交通系统中跨域任务（CP）的性能，我们提出了域与时间对齐（DATA）网络，以生成高质量的 3D 目标检测特征。具体而言，该网络旨在学习域不

[1*] Equal contribution
[2†] Corresponding authors

www.xueshuxiangzi.com

变和时间一致的表示，以实现稳健的特征获取。DATA 网络由三个主要模块组成：（i）一致性保持域对齐模块（CDAM），用于在训练阶段减少域间隙。通过两种互补的方法来最小化单个代理内部和代理之间的域间隙：通过邻近区域的分层下采样来实现具有物理有效性的一致的距离自适应点密度，并通过在共享区域的一致观察条件下的特征级对抗学习来缓解真实的域间隙；（ii）渐进时间对齐模块（PTAM），用于解决时间错位。它通过多尺度特征逐级捕捉运动模式，并通过两阶段的补偿来建模复杂运动，以实现场景范围的表示，并且多窗口自监督训练策略同时实现有效的前景物体运动学习和维护全局场景的一致性；（iii）实例聚焦特征聚合模块（IFAM），用于有效聚合来自多个代理的对齐特征。为了验证 DATA 的有效性，我们在三个典型的 CP 数据集上进行了广泛的实验：DAIR-V2X-C、V2XSET 和 V2XSIM。综合实验结果表明，我们的方法在 DAIR-V2X-C 上比现有的最先进方法高出 2.36 个百分点的 AP，在 V2XSIM 和 V2XSET 上分别高出 1.84 个百分点和 2.85 个百分点的 AP。DATA 还表现出卓越的延迟鲁棒性，在 500 毫秒通信延迟下保持 75.58 个百分点的 AP，超过最先进方法 2.61 个百分点。本文的主要贡献可以总结如下：

- 我们提出了 DATA，这是一种新的 CP 框架，主要通过领域和时间对齐来解决特征获取的挑战，并辅以实例级特征优化，以最大化对齐特征的语义表示。
- 我们设计了 CDAM、PTAM 和 IFAM，以减少域差距，实现时间特征的一致性，并充分利用对齐特征的语义表示。

2. 相关工作

Object Detection in Collaborative Perception. 协作感知是自动驾驶系统中的一个关键部分。最近，各种研究 [? ? ? ? ?] 探索了提高感知性能的多种方法。Where2comm [?] 通过置信图选择性地传输感知关键特征。CodeFilling [?] 基于码本编码进一步压缩特征以降低传输成本。HM-ViT [?] 提出异质 3D 图变换器来处理代理之间的传感器配置差异。HEAL [?] 介绍了一种向后对齐训练机制来构建统一的特征空间。MRCNet [?] 通过运动感知鲁棒通信框架解决位置噪声、感知噪声和运动模糊问题。

Domain Alignment in Collaborative Perception. 解决域间差距问题 [?] 是提高协作感知性能的关键步骤。最近的方法通过不同的机制来解决域间差距。V2X-ViT [?] 通过在其异构多智能体自注意力模块中通过专门的嵌入编码不同的智能体组合来解决域间差距。MPDA [?] 通过可学习的特征缩放器和稀疏的跨域转换器来解决域间差距。DI-V2X [?] 提出了一种通过域混合实例增强和渐进式蒸馏的蒸馏框架。在本文中，我们提出了密度感知和区域感知的训练机制，以更好地弥合各种异构智能体之间的域间差距。

Time Alignment in Collaborative Perception. 时间同步对于确定现实世界中的部署性能也很重要。最初努力解决这一挑战的方法包括 V2VNet [?] 和 V2X-ViT [?] ，这些方法率先使用卷积网络和考虑延迟的定位编码进行延迟补偿的神经网络方法。此外，SyncNet [?] 通过金字塔 LSTM 架构结合多个历史帧来扩展这些方法。FFNet [?] 引入了一种基于流的框架来预测对齐融合的未来特征。同时，CoBEVFlow [?] 结合基于 RoI 的匹配和基于转换器的方法进行运动预测。在本文中，我们提出了两阶段补偿方法，以捕获细粒度和全局运动模式。此外，我们引入了一种多窗口自监督策略，以更好地学习局部区域内每个对象的运动模式。

3. 方法

3.1. 问题表述和总体架构

我们的框架在一个具有 N 个代理的 CP 系统中运行，其中每个代理同时作为数据接收者和数据提供者。在整篇论文中，我们定义当前正在接收和传输数据的代理为自我代理和协作代理。下标 i 和 j 分别表示自我代理和协作代理，其中 $i \neq j$ 。在当前时间 t ，自我代理处理其最新数据 $\mathcal{X}_i(t)$ ，同时整合协作代理在时间戳 $t-\tau$ 传输的数据，其中 τ 代表传输延迟。为了补偿这种传输延迟，协作代理在传输前处理其两个最新帧 $\mathcal{X}_j(t-\tau)$ 和 $\mathcal{X}_j(t-\tau-\Delta T)$ ，提供感知和运动信息。

DATA 的整体架构如图 ?? 所示。首先，自我代理的点云通过 CDAM 的 PHD（ 红色部分 ）生成多尺度特征。在协作分支中，它处理点云以生成两个时间戳的多尺度特征。两个代理的特征被转换为 BEV 特征，然后输入到 CDAM 的 OD 以促进两者之间的域对齐。协作代理的多尺度特征通过 PTAM 的第一个阶段（蓝色部分 ）。然后，输出特征连同最新帧（$t-\tau$ ）的特征和运动信息一起被压缩并传输到自我代理。自我代理解压接收到的数据以恢复多尺度特征，并实施 PTAM 的第二阶段以根据传输延迟 τ 进一步调整特征，实现完全的时间对齐。时间对齐后的多尺度特征被转换为 BEV 特征，随后与自我的 BEV 特征一起输入到 IFAM（ 绿色部分 ）中，以将互补信息融合成一个全面的表示。这种融合后的表示作为检测头（解码器）的输入，以生成最终检测结果。硬件异质性和部署变化引起的域差异在原始数据和特征层面上表现出来。因此，我们通过引入 CDAM 的 PHD 进行原始数据处理和 OD 的特征级对齐来解决这些问题。

单个代理的点云在不同观察距离下表现出显著的密度变化，导致模型学习偏向于高密度区域，并影响特征提取。为了解决这个问题，提出了 PHD 方法以平衡自车代理处的点云密度分布。PHD 主要包括以下三个步骤。

对于输入场景，我们定义自车代理 i 可观测到的所有对象的集合为 $\mathcal{O}_i = \{o_1, o_2, ..., o_{N_i^{\text{total}}}\}$ ，其中 N_i^{total} 是可观测对象的总数。基于与自车代理的距离，识别临近区域内的对象为

其中 $d_i(o_k)$ 是自车代理 i 到第 k 个对象的距离，d_{th} 是距离阈值。为了在下采样后保持与原场景的分布相似性，我们从 $\mathcal{O}_i^{\text{prox}}$ 中选择 N_{proc} 个对象进行后续处理。如果 $|\mathcal{O}_i^{\text{prox}}| > N_{\max}$ ，则从 $\mathcal{O}_i^{\text{prox}}$ 中随机选择

www.xueshuxiangzi.com

$N_{\text{proc}} = N_{\max}$ 个对象，否则选择 $\mathcal{O}_i^{\text{prox}}$ 中的所有对象。此处，$|\cdot|$ 表示集合的基数，$N_{\max}$ 是预定义的对象最大数量。

对于每个选定的对象，其点云通过使用不同尺度的同心包围盒分割为内部和外部区域。令 $\mathcal{B}_{i,k}^{\text{out}}$ 表示第 k 个对象的定向包围盒，其由中心坐标 (x_k, y_k, z_k)、尺寸 (h_k, w_k, l_k) 和方向 θ_k 参数化。我们还定义了一个缩放因子 $\alpha \in (0,1)$，以调整包围盒的高度、宽度和长度。这就创建了一个与 $\mathcal{B}_{i,k}^{\text{out}}$ 、i.e. 、$\mathcal{B}_{i,k}^{\text{in}} = \{(x_k, y_k, z_k, \alpha h_k, \alpha w_k, \alpha l_k, \theta_k)\}$ 具有相同中心和方向的内部包围盒，其中 $k = 1, \ldots, N_{\text{proc}}$。然后，分布在内外区域的点集为

$$\mathcal{R}_{i,k}^{\text{in}} = \{p | p \in \mathcal{B}_{i,k}^{\text{in}}\}, \quad \mathcal{R}_{i,k}^{\text{out}} = \{p | p \in \mathcal{B}_{i,k}^{\text{out}} \setminus \mathcal{B}_{i,k}^{\text{in}}\}. \quad (1)$$

。该分割有效地将稀疏内部点与对象轮廓上的密集点分开。

步骤 3：密度平衡。接下来，在分区区域中应用不同比率的最远点采样（FPS）[?]。对于内区域，使用高降采样比率 β_{in}，而对于外区域，应用保守降采样比率 β_{out}，保留更多点，从而捕捉物体轮廓并保留其几何细节。降采样可以被公式化为

$$\tilde{\mathcal{R}}_{i,k}^{\text{in}} = \text{FPS}(\mathcal{R}_{i,k}^{\text{in}}, \beta_{\text{in}}), \quad \tilde{\mathcal{R}}_{i,k}^{\text{out}} = \text{FPS}(\mathcal{R}_{i,k}^{\text{out}}, \beta_{\text{out}}). \quad (2)$$

。最后，结合内外区域的 k -th 物体的点云可以被公式化为 $\tilde{\mathcal{P}}_{i,k} = \tilde{\mathcal{R}}_{i,k}^{\text{in}} \cup \tilde{\mathcal{R}}_{i,k}^{\text{out}}, k = 1, \ldots, N_{\text{proc}}$。通过这种分层降采样，PHD 增强了不同距离的一致密度表示，并保留了原始数据的关键信息，特别是遮挡关系和几何结构。

3.1.1. 受观测性约束的鉴别器 (OD)

在 CP 中，代理之间的域差距源于传感器固有属性和观察特性。为了对齐这些域差距，识别出两个代理都保持有效观察的区域是至关重要的。为了解决这个问题，我们提出了一个 OD 模块，以显式地将可观察性并入域对齐过程中。

首先，我们将自我代理 i 的观测定义为自我域，而将 $N-1$ 协作代理的观测定义为协作域。为了进行域区分，随机选择一个协作代理进行域对齐。这使得可以实现多样的代理组合，从而促进域不变特征学习。

然后，通过一个前景估计器 $\Phi(\cdot)$，OD 使用自车和协作代理的鸟瞰视图（BEV）特征 H_i 和 H_j 生成可观测性地图 $M_i = \Phi(F_i)$ 和 $M_j = \Phi(F_j)$，其中 $M_i, M_j \in \mathbb{R}^{1\times H\times W}$，用于指示每个空间位置的可观测性。

随后，我们通过空间变换将协作特征投影到自我代理的坐标上，创建 $H_{j\to i}$ 和 $M_{j\to i}$。然而，这可能会潜在地产生空域。令 $\mathcal{V}$ 表示在变换后的特征图中有效网格的集合，并且空域被填充为

$$H_j^{\text{comp}} = \mathcal{I}_{\mathcal{V}} \cdot H_{j\to i} + (1 - \mathcal{I}_{\mathcal{V}}) \cdot H_i, \quad (3)$$

$$M_j^{\text{comp}} = \mathcal{I}_{\mathcal{V}} \cdot M_{j\to i} + (1 - \mathcal{I}_{\mathcal{V}}) \cdot M_i, \quad (4)$$

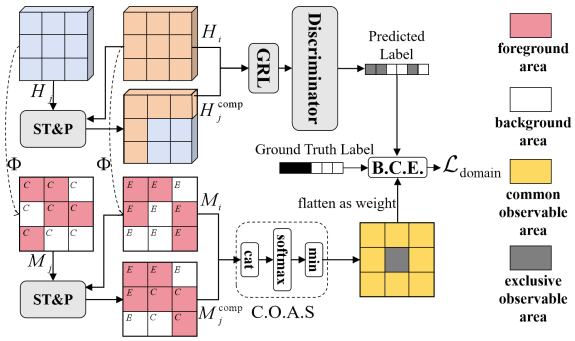

Figure 2. OD 的管道。ST & P 表示空间变换和填充，C.O.A.S. 指代公共可观察区域选择。通过公共可观察区域定位真实的领域差距，从而实现固有的领域不变特征学习。

，其中 H_j^{comp} 和 M_j^{comp} 分别表示补充的特征和可观察性图，并且指示函数 $\mathcal{I}_{\mathcal{V}}$ 对于 V 中的点为 1，对于其他点为 0。为了确保领域对齐关注于两个代理共享可观察性的区域，计算出可观察性加权图 $W \in \mathbb{R}^{1\times H\times W}$ 为 $W = \min(\text{softmax}([M_i, M_j^{\text{comp}}]))$，其中 $[\cdot]$ 表示拼接，并且所有操作沿第一个维度执行。最后，领域对齐目标为

$$\max_{\theta} \min_{\mu} \mathcal{L}_{\text{domain}} = \frac{1}{\sum W_{\text{flat}}^{sp}} \sum_{sp\in\mathcal{S}} W_{\text{flat}}^{sp} \cdot \mathcal{L}_{\text{BCE}}(D_{\mu}^{sp}(\Psi_{\theta}), Z^{sp}), \quad (5)$$

，其中 $\mathcal{S}$ 表示包含所有空间位置的集合，sp 表示一个位置，W_{flat} 表示展平的可观察性加权图，Ψ_{θ} 是特征提取器（点云为输入，BEV 特征为输出），D_{μ} 表示判别器，$\mathcal{L}_{\text{BCE}}$ 是二元交叉熵损失，Z 是领域标签（自我代理为 0，协作代理为 1）。

为联合优化这个最小最大的问题，在判别器前插入了一个梯度反转层（GRL）[?]。GRL 在前向传播中保持输入不变，并在反向传播中应用负缩放因子 $\gamma = -0.1$。这使得端到端的对抗训练成为可能，其中判别器学习在具有共享可观测性的区域区分域，特征提取器学习生成符合多智能体感知物理约束的域不变特征。

3.2. 渐进时序对齐模块 (PTAM)

3.2.1. 时间对齐的运动学视角

时间异步导致代理之间的特征未对齐，这对多代理协同感知（CP）构成了一个基本挑战。为了补偿在历史时间点 $t - \Delta t$ 合作特征的时间未对齐，我们利用了一个运动学视角 [? ?]。协作代理中任意尺度特征的时间演化被公式化为

$$F_j(t, \boldsymbol{x} + \boldsymbol{v}(t - \Delta t, \boldsymbol{x})\Delta t) = F_j(t - \Delta t, \boldsymbol{x}), \quad (6)$$

，其中 Δt 是一个时间间隔（在 CP 中典型传输延迟范围内 [? ? ?]），$F_j(t - \Delta t, \boldsymbol{x})$ 表示时间 $t - \Delta t$ 时位置 $\boldsymbol{x}$ 的特征，$\boldsymbol{v}(t - \Delta t, \boldsymbol{x})$ 表示描述时间 $t - \Delta t$ 时特征运动的速度场。这表明可以通过沿速度场追溯到时间 $t - \Delta t$ 的对应位置来获得时间 t 下某一位置的特征。

根据运动学公式，PTAM 的核心机制通过三个基本组件实现特征的时间演化：表示 $\boldsymbol{v}$ 的运动场 $\Delta p \in$

www.xueshuxiangzi.com

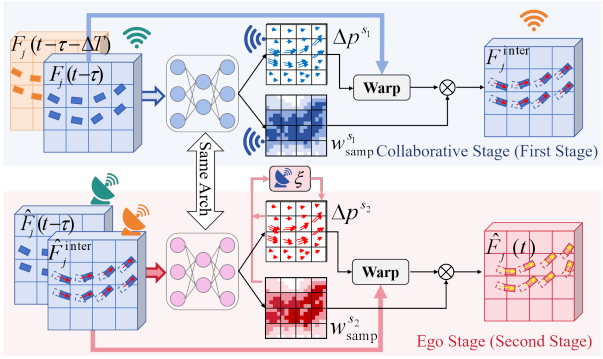

Figure 3. PTAM 的流水线（Pipeline）。此处展示了一个空间尺度。三对不同颜色的“WiFi”和“雷达”符号表示代理之间的信息传输和接收。通过两阶段预测，利用历史运动趋势和场景动态，实现精确的延迟补偿。

$\mathbb{R}^{2\times H\times W}$，表示 Δt 的时间缩放因子 ξ，以及通过双线性采样实现的位移补偿操作。为了简化，我们在下面省略 x 。

3.2.2. 两步实现方案针对两个代理

基于这一核心机制，PTAM 的渐进时间对齐策略由两个串行执行的阶段性预测组成。每个阶段对应一个代理。(i) 第一阶段。协作代理通过利用其最新的两帧特征 $F_j(t-\tau-\Delta T)$ 和 $F_j(t-\tau)$ 来捕捉历史运动模式，以预测中间特征 F_j^{inter} 。

(二) 第二阶段。自主代理基于接收到的运动信息和传输延迟特性执行自适应时间对齐。由于两个代理之间的潜在通信错误，接收到的特征被定义为 $\hat{F}_j(t-\tau)$ 和 $\hat{F}_j^{\text{inter}}$ ，并用于生成时间对齐的特征 $\hat{F}_j(t)$ 。为了捕捉不同粒度的层次运动模式，PTAM 同时在三个空间尺度上处理特征。

每个阶段的实施。在具有每个空间尺度的每个阶段，我们通过两个步骤执行核心机制，i.e. ，运动估计和特征扭曲。

步骤 1：运动估计。这个步骤在两个阶段中共享相同的架构。为了统一表示这两个阶段，我们将输入的相邻时序特征记为最新特征 F_{latest} 及其先前特征 F_{prev} 。运动估计从计算时序特征差异开始，定义为 $\Delta F = F_{\text{latest}} - F_{\text{prev}}$ ，然后它分别与这两个输入特征串联。这些串联的特征经过一系列操作，生成运动场 Δp 和采样权重 $w_{\text{samp}} \in \mathbb{R}^{1\times H\times W}$ 。

步骤 2：特征扭曲。这个步骤在两个阶段中的处理方式略有不同。对于在协作代理上执行的第一阶段，使用单位时间缩放因子 i.e. ，$\xi = 1$ ，以保持与传感器（例如 LIDAR）的采样周期一致，从而生成中间特征表示 $F_j^{\text{inter}} = w^{s_1}_{\text{samp}} \odot f_{\text{warp}}\big(F_j(t-\tau), \Delta p^{s_1}\big)$ ，其中 s_1 表示第一阶段，$f_{\text{warp}}(\cdot)$ 表示双线性采样操作。对于自我代理上的第二阶段，需要一个自适应时间缩放因子 ξ ，以处理由于传输延迟导致的可变时间间隔。运动差场 ΔM 定义为

$$\Delta M = (\Delta p^{s_2} \odot w^{s_2}_{\text{samp}}) - (\Delta p^{s_1} \odot w^{s_1}_{\text{samp}}), \quad (7)$$

，其中上标 s_2 表示第二阶段，而元素级的乘积通过结合 Δp 和相应的 w_{samp} 来捕获每个阶段的有效运动。然后，为了获得全局上下文向量 f_M ，我们使用级联卷积层、残差块和全局池化操作来处理 ΔM 。通过将捕获传输延迟特性的正弦位置嵌入添加到 f_M ，我们获得时间编码 f_T 。最后，这些特征被连接并通过 MLP 处理，以预测时间缩放因子为

$$\xi = \text{ReLU}(\text{MLP}([f_M, f_T])). \quad (8)$$

，并通过 $\hat{F}_j(t) = w^{s_2}_{\text{samp}} \odot f_{\text{warp}}\big(\hat{F}_j^{\text{inter}}, \Delta p^{s_1}\big)$ 获得最终的时间对齐特征。

为了增强 PTAM 在通过渐进-并行对齐捕捉多样运动模式的能力，我们提出了一种多窗口自监督训练策略。

在每个空间尺度 s ，大小为 $h_s \times w_s$ 的特征平面使用两种互补的窗口划分策略进行划分。这两个窗口集合定义为

$$W_1 = \{w_{m',n'} \mid 0 \le m' < \frac{h_s}{l}, 0 \le n' < \frac{w_s}{l}\}, \quad (9)$$

$$W_2 = \{w_{p',q'} \mid 0 \le p' < \frac{h_s}{l} - 1, 0 \le q' < \frac{w_s}{l} - 1\}, \quad (10)$$

，其中 $w_{m',n'}$ 和 $w_{p',q'}$ 分别表示左上角在位置 (m',n') 和 (p',q') 、大小为 $l\times l$ 的窗口。在此，W_1 实际从边界开始精确地划分特征，而与 W_1 相比，W_2 有一个 $l/2$ 的偏移。对于两个集合中的每个窗口 w ，我们计算预测和真实特征 F_j^{gt} 之间的余弦相似度 $\cos(\cdot,\cdot)$ 。尺度 s 处中间和最终预测的损失函数可以写成

$$\mathcal{L}^s_{\text{inter}} = \frac{1}{N_{\text{window}}} \sum_{w\in W_1\cup W_2} \|1 - \cos(F^{\text{inter},s}_{j,w}, F^{\text{gt},s}_{j,w}(t))\|_2^2, \quad (11)$$

$$\mathcal{L}^s_{\text{final}} = \frac{1}{N_{\text{window}}} \sum_{w\in W_1\cup W_2} \|1 - \cos(\hat{F}^{s}_{j,w}(t), F^{\text{gt},s}_{j,w}(t))\|_2^2, \quad (12)$$

，其中 N_{window} 表示窗口的总数。最后，计算所有三个尺度上的总时间对齐损失，由 $\mathcal{L}_{\text{temporal}} = \sum_{s=1,2,3}(\mathcal{L}^s_{\text{inter}} + \mathcal{L}^s_{\text{final}})$ 给出。

3.3. 实例聚焦特征聚合模块 (IFAM)

假设通过 CDAM 和 PTAM 获得了域和时间对齐的特征，但如何充分利用这些对齐特征的语义信息仍然对于提高感知性能至关重要。在此，设计了 IFAM 以增强前景物体的结构表示，以实现稳健的 CP。假设 $H_a \in \mathbb{R}^{C\times H\times W}$ 是第 a 个代理的 BEV 特征，其前景和背景特征标识为 $H_a^{\text{fore}} = H_a \odot M_a$, $H_a^{\text{back}} = H_a \odot (\mathbf{1} - M_a)$ ，其中 $M_a = \Phi(H_a)$ 是前景掩码，$\Phi(\cdot)$ 是第 3.1.1 节 (OD 模块) 中提到的前景估计器。为了加强前景区域的结构细节，应用了一组 3×3 卷积核。然后我们有 $H_a^{\text{enh}} = \text{StructConv}(H_a^{\text{fore}})$ ，其中 StructConv 由一个普通卷积分支构成，以保持基本特征强度，以及四个专门的卷积分支：中心差分卷积、水平和垂直差分卷积以及角度差分卷积。尽管这加强了结构细节，但可能会出现干扰后续检测的虚假目标。

www.xueshuxiangzi.com

为了抑制这些虚假的前景特征，同时保留增强的结构细节，本文提出了一种基于柱体编码 [? ?] 的前景验证机制，因为其每个通道维度固有地包含耦合的高度语义信息，可用于选择正确的前景特征。具体来说，给定原始和增强后的前景特征（H_a^{fore} 和 H_a^{enh}），它们沿通道维度的串联结果为 $H_a^{\text{cat}} = \left[H_a^{\text{fore}}, H_a^{\text{enh}}\right] \in \mathbb{R}^{2C\times H\times W}$。然后并行使用空间和通道注意模块。对于空间注意 W_a^{s}，在通道维度上执行最大和平均池化操作，然后进行串联和卷积。对于通道注意 W_a^{c}，应用空间平均池化，接着是两个卷积。然后初始注意权重 W_a^{init} 被推导为 $W_a^{\text{init}} = W_a^{\text{s}} \oplus W_a^{\text{c}}$。而验证权重为

$$W_a^{\text{verif}} = \text{Sigmoid}\left(\text{GConv}\left(\text{CS}\left(\left[H_a^{\text{cat}}, W_a^{\text{init}}\right]\right)\right)\right), \tag{13}$$

，其中 CS 表示通道洗牌 [?]，它促进跨组信息交换，打破固定的通道组合以进行全面的特征验证。组卷积 GConv(·) [?] 使得不同特征组可以独立生成权重，允许对高度语义模式进行专门验证。通过这种验证机制，展示高度语义关系（e.g.，与典型车辆形状和高度一致的那些）的区域得到保留，而包含虚假前景特征非自然表示组合的区域得到缓解。然后 k 号代理的验证前景特征最终表现为

$$H_a^{\text{verif}} = \text{Conv}_{1\times1}((W_a^{\text{verif}} \odot H_a^{\text{fore}} \oplus (\mathbf{1} - W_a^{\text{verif}}) \odot H_a^{\text{enh}}) \oplus H_a^{\text{fore}} \oplus H_a^{\text{enh}}). \tag{14}$$

。单独被细化的 BEV 特征通过组合背景特征获得，如 $H_a^{\text{refined}} = H_a^{\text{verif}} \oplus \epsilon H_a^{\text{back}}$，其中 ϵ 是一个学习参数，用于平衡背景特征的贡献。所有 N 代理的细化特征通过共享的 1×1 卷积逐步融合，以生成最终的 CP 表示。

4. 实验

4.1. 数据集和评估指标

实验在三个数据集上进行：DAIR-V2X-C [?]、V2XSET [?] 和 V2XSIM [?]。详细的数据集信息在补充材料的第 2.1 节中呈现。为了评估，所有数据集的汽车类别在交并比（IoU）阈值为 0.50 和 0.70 时测量平均精度（AP）。

4.2. 实现

骨干网络采用 PointPillar [?] 架构，网格大小为 0.4m × 0.4m。在 CDAM 中，距离阈值 $d_{\text{th}} = 50$ m，并且 $N_{\max} = 2$ 辆车辆被下采样。两种下采样比例为 $\beta_{\text{in}} = 0.6$ 和 $\beta_{\text{out}} = 0.8$。对于 PTAM，多窗口自监督训练采用窗口大小为 $l = 16$。训练过程包括三个连续阶段：检测模型训练、PTAM 训练和传输模块训练，这在补充材料的第 2 节中介绍。

4.3. 定量结果

性能无延迟。表 ?? 展示了在零延迟设置下与 SOTA 方法的性能比较。在真实世界的 DAIR-V2X-C 验证集

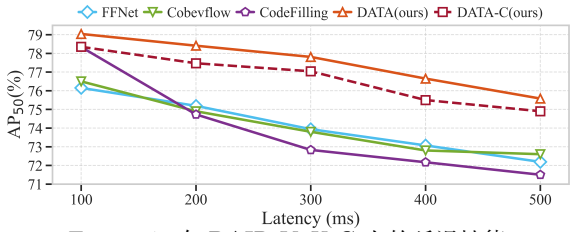

Figure 4. 在 DAIR-V2X-C 上的延迟性能。

上，与 HEAL [?] 相比，DATA 在 AP $_{50}$ 和 AP $_{70}$ 上分别实现了 0.94 % 和 2.36 % 的提升。在模拟数据集 V2XSIM 和 V2XSET 中，DATA 在模拟现实世界条件下（加入了高斯噪声：位置 $\mu_{\text{pos}} = 0, \sigma_{\text{pos}} = 0.2\ m$，方向 $\mu_{\text{rot}} = 0, \sigma_{\text{rot}} = 0.2°$）也展示了显著的提升。具体来说，在 V2XSIM 上，比 MRCNet 在 AP $_{50}$ 和 AP $_{70}$ 上分别领先 3.58 % 和 7.87 %，在 V2XSET 上则分别领先 4.53 % 和 8.67 %。这些在复杂模拟和真实场景中的一致性提升展示了 DATA 的有效性，首先通过 CDAM 实现域不变的特征提取以增强特征质量，然后利用 IFAM 有效地突出前景特征语义，以提升 CP 系统性能。

延迟下的性能。如图 4 所示，DATA 在 DAIR-V2X-C 数据集上的时间异步情况下有效地保持了高质量的 CP。在零延迟场景中强大的特征提取和融合能力的基础上，PTAM 在传输过程中确保通过不同的延迟实现可靠的特征预测，使延迟从 100ms 增加到 500ms 时 AP 仅减少了 3.46 %。在 500ms 延迟时，DATA 维持了 75.58 % AP，而基线方法，如 FFNet (72.19 %) 和 CodeFilling (71.50 %) 下降更加明显。同时，即使在压缩传输情况下，DATA 也能实现稳健的性能（在 500ms 延迟时 DATA-C 的 AP 为 74.90 %）。这些结果表明，DATA 通过获取和利用高质量特征，有效地解决了真实世界 CP 中的时间挑战。

姿态误差的鲁棒性。为了研究不同姿态噪声分量对 CP 性能的影响，在 V2XSET 和 V2XSIM 上进行实验。分别对合作者的位置和方向施加不同标准差的高斯噪声，定位噪声 σ_{loacl} 的范围从 0.2 m 到 0.6 m，航向噪声 σ_{head} 的范围从 0.2° 到 1.0°。如图 5 所示，所比较的方法在性能上明显下降。随着噪声量的增加，它们的 AP $_{70}$ 迅速下降，当 σ_{local} 和 σ_{head} 分别在 V2XSIM 和 V2XSET 达到 0.6 m 和 1.0° 时，下降超过 30 % 和 9 %。这表明空间特征对齐错误严重影响了它们的 CP 准确性。DATA 在两个数据集上的姿态不确定性方面表现出强大的弹性：在 V2XSIM 上，它在 $\sigma_{\text{local}} = 0.6$ m 达到 66.53 % AP $_{70}$，在 $\sigma_{\text{head}} = 1.0°$ 达到 75.23 % AP $_{70}$；同样在 V2XSET 上，它在 $\sigma_{\text{local}} = 0.6$ m 达到 67.22 % AP $_{70}$，在 $\sigma_{\text{head}} = 1.0°$ 达到 71.17 % AP $_{70}$。这种鲁棒性来自于两个互补的方面。首先，可观测性引导的域对齐从一致的观察中学习域不变特征，从嘈杂观察中提取基本几何模式。同时，通过前景-背景分离和语义细化的实例集中特征增强，加强了对象表示的结构完整性，确保在姿态扰动下的可靠检测。

www.xueshuxiangzi.com

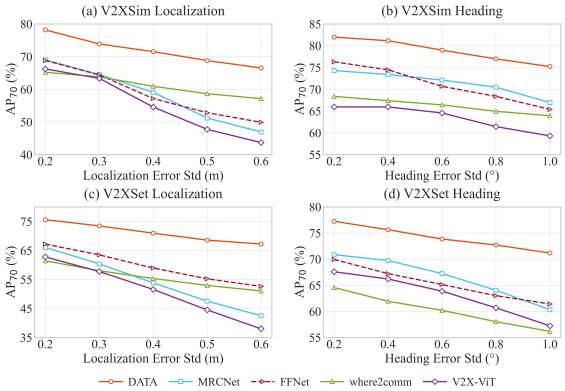

Figure 5. 带有姿态误差的性能比较。

4.4. 消融研究

域对齐和特征聚合的消融实验。为了研究每个模块的贡献，在没有延迟的情况下，在 DAIR-V2X-C 验证集上进行了实验。如表 1 所示，与基线模型相比，引入 IFAM 带来了显著的性能提升，即在 AP $_{70}$ 上提升了 3.70 % ，因为它的几何感知特征增强加强了结构完整性，并实现了更精确的目标定位。添加 OD 进一步将检测性能提升了 1.96 % AP $_{50}$ 和 0.62 % AP $_{70}$ ，因为它引导特征对齐关注于两个代理保持相当观测能力的区域，从根本上解决了跨域特征差异问题。集成 PHD 在 AP $_{50}$ 和 AP $_{70}$ 上分别带来了 0.96 % 和 1.25 % 的提升。这表明，保持不同空间区域的点云分布一致，同时保留固有的场景结构，对于提取可靠的域不变特征至关重要。

邻近区域阈值消融实验。PHD 模块旨在在不破坏场景结构的同时平衡不同空间区域的点云密度。表 3 的结果表明，50 m 阈值实现了最佳性能（79.94 % AP $_{50}$ ），代表了密度平衡和分布保留之间的良好折衷。较小的阈值（10 m, 30 m ）对邻近物体的覆盖不足，而较大的阈值（70 m, 90 m ）可能引入过多的下采样，从而破坏原始场景结构。

渐进时序对齐消融实验。为了研究窗口大小对时序对齐性能的影响，实验在具有 300 毫秒通信延迟的 DAIR-V2X-C 验证集上进行，比较了一阶段和两阶段模型。表 2 显示，对于所有窗口大小，两阶段模型均优于一阶段模型，其中两阶段模型在窗口大小为 16 时达到最佳性能（77.81 % AP $_{50}$ ），分别超过整个地图监督和一阶段方法 0.67 % 和 1.01 % 。当增加或减少窗口大小时，性能下降，这表明该中等窗口大小在局部运动捕捉和上下文信息保留之间实现了最佳平衡。特别是使用小窗口大小，比如 4 时，由于小窗口无法充分捕捉到车辆对象，两种模型都面临挑战。两阶段模型显示出更显著的性能下降（窗口大小为 4 时表现为 76.64 % AP $_{50}$ ），相比于单阶段方法（76.54 % AP $_{50}$ ），这是由于预测阶段错误积累的额外影响。这些发现证实两阶段渐进对齐可以通过适当的窗口分区训练，在中间预测状态中实现时间一致性。

在本文中，我们提出了一种新颖的框架 DATA，该框架解决了特征级协作感知中的基础领域和时间错位问题。我们开发了特征对齐的系统性解决方案，以在特征获取阶段实现对齐。所提出的 CDAM 通过密度感知的点云采样和可观测性引导的领域对齐来减少特征提取中的领域差距，而 PTAM 通过渐进运动优化确保特征传输过程中的时间一致性。在良好对齐的特征基础上，IFAM 进一步增强其语义表达能力，以最大化融合性能。大量关于真实世界和模拟数据集的实验证明了 DATA 的有效性，在各种挑战条件下达到了 SOTA 性能，验证了我们在协作感知中特征对齐和增强系统方法的鲁棒性。该工作部分得到了中国国家自然科学基金（Grant 62271142 和 U2341206）的支持，部分得到了南京 U35 坚基础工程（Grant 4204002501）的支持，以及江苏省重点研发计划（Grant BE2023021 和 BE2023011-2）的支持。

Config	AP $_{50}$	AP $_{70}$	Config	AP $_{50}$	AP $_{70}$
Baseline	76.80	59.91	+ P. + O.	78.96 $^{+2.16}$	62.68 $^{+2.77}$
+ P.	77.54 $^{+0.74}$	62.16 $^{+2.25}$	+ P. + I.	78.30 $^{+1.50}$	63.65 $^{+3.74}$
+ O.	78.52 $^{+1.72}$	62.58 $^{+2.67}$	+ O. + I.	78.98 $^{+2.18}$	64.23 $^{+4.32}$
+ I.	77.02 $^{+0.22}$	63.61 $^{+3.70}$	+ P. + O. + I.	79.94 $^{+3.14}$	65.48 $^{+5.57}$

Table 1. 对 DAIR-V2X-C 上的 CDAM 和 IFAM 进行全面消融研究。P.: PHD, O.: OD, I.: IFAM。

Model Type (AP $_{50}$ %)	w/o PTAM	Window Size l (AP $_{50}$ %)				
		w.m.	32	16	8	4
First-Stage (Collab.)	71.65	76.23	76.68	76.80	76.62	76.54
Two-Stage (Ego & Collab.)	71.65	77.14	77.62	77.81	77.59	76.64

Table 2. 在不同窗口大小下，两种模型类型在 300 毫秒延迟上的表现。“w.m.” 表示全图监督。

Distance Threshold d_{th}	10m	30m	50m	70m	90m
Metric (AP $_{50}$ %)	79.33	79.44	79.94	79.18	78.69

Table 3. 近端区域距离阈值的消融（d_{th} ）。

www.xueshuxiangzi.com

DATA: 域和时间对齐以实现高质量的协同感知特征融合

Supplementary Material

1. 模块设计的详细信息

1.1. 受可观测性约束的判别器实现

可观测性约束判别器（OD）采用轻量级卷积架构来区分自我和协作领域。判别器由两层卷积层组成：第一层有 256 个通道，核大小为 1×1 ，然后是 ReLU 激活；最后一层有一个输出通道，核大小为 1×1 ，输出预测的领域标签。

1.2. 用于鸟瞰图表示的多尺度特征处理

多尺度特征处理模块将不同空间分辨率的特征转换为全面的鸟瞰视图（BEV）表示。

假设从骨干网络中提取出三个尺度的特征，分别表示为 $F_{\text{large}} \in \mathbb{R}^{64\times H\times W}$ 、$F_{\text{middle}} \in \mathbb{R}^{128\times \frac{H}{2}\times \frac{W}{2}}$ 和 $F_{\text{small}} \in \mathbb{R}^{256\times \frac{H}{4}\times \frac{W}{4}}$ ，然后转换为 BEV 特征遵循一个系统的过程。

每个尺度都经过一次反卷积操作，以统一空间和通道维度，这个操作可以表示为

$$U_{\text{large}} = \Theta_1(F_{\text{large}}), \tag{15}$$

$$U_{\text{middle}} = \Theta_2(F_{\text{middle}}), \tag{16}$$

$$U_{\text{small}} = \Theta_3(F_{\text{small}}), \tag{17}$$

其中 Θ_1 、Θ_2 和 Θ_3 分别代表步幅为 1、2 和 4 的反卷积网络。每个反卷积网络由一个转置卷积层、批量归一化和 ReLU 激活构成，将特征投影到一个统一的空间分辨率 $H \times W$ 和通道维度 128。

最终的 BEV 表示通过沿通道维度连接这些上采样特征形成，公式为 $H_{\text{BEV}} = [U_{\text{large}}, U_{\text{middle}}, U_{\text{small}}]$ 。生成的 BEV 特征图 $H_{\text{BEV}} \in \mathbb{R}^{384\times H\times W}$ 整合了所有尺度的信息，同时保持一致的空间维度。该方法对于在 BEV 空间中进行准确的 3D 目标检测至关重要，它保留了来自大尺度特征的细粒度细节和来自小尺度特征的语义上下文。

1.3. 计算复杂性分析：分块计算与全局计算

我们分析了在计算特征图之间余弦相似度时，分块计算与全局计算的计算复杂度。考虑具有 C 个通道的尺寸为 $H \times W$ 的特征图，对于分块方法，我们使用窗口大小 $l \times l$ ，其中为了简化计算，l 可以被 H 和 W 同时整除。在此，$\boldsymbol{f}_{\text{p}}$ 表示预测的特征图，$\boldsymbol{f}_{\text{gt}}$ 表示真实值特征图。

对于整个特征图的全局计算，计算成本如下：总的来说，全局计算大约需要 $(3C + 2) \times H \times W$ 次乘法，$(3C - 1) \times H \times W$ 次加法，$2 \times H \times W$ 次开方运算和 $H \times W$ 次除法，结果得出计算复杂度为 $\mathcal{O}(C \times H \times W)$ 。

我们的方法使用了两种互补的窗口分割策略：

第一种策略（W_1 ，描述在方程（11）中）应用标准分区，产生 $\lfloor H/l \rfloor \times \lfloor W/l \rfloor$ 个窗口。第二种策略（W_2 ，描述在方程（12）中）采用偏移分区（相对于 W_1 的偏移为 $l/2$ ），由于在特征图的顶部、底部、左侧和右侧边界引入了偏移，因此产生 $\lfloor (H-l)/l \rfloor \times \lfloor (W-l)/l \rfloor$ 个窗口。在我们的实现中对于典型的维度（$H = 256, W = 128, l = 16$ ），我们有 $|W_1| = 16 \times 8 = 128$ 个窗口和 $|W_2| = 15 \times 7 = 105$ 个窗口。

对于每个策略中的每个窗口，计算点积需要 $C \times l \times l$ 次乘法和 $(C - 1) \times l \times l$ 次加法。ℓ_2 范数计算需要 $2 \times C \times l \times l$ 次乘法、$2 \times (C - 1) \times l \times l$ 次加法和 $2 \times l \times l$ 次平方根。计算余弦相似度需要 $l \times l$ 次除法和 $l \times l$ 次乘法，而 MSE 损失计算需要 $l \times l$ 次减法和 $l \times l$ 次乘法。

在这两种窗口策略下，总的计算量是

$$\left[\lfloor \frac{H}{l} \rfloor \times \lfloor \frac{W}{l} \rfloor + \lfloor \frac{H-l}{l} \rfloor \times \lfloor \frac{W-l}{l} \rfloor\right] \times [(3C + 2) \times l \times l]. \tag{18}$$

。对于我们的典型维度，这大约是 $1.8\times(3C+2)\times H\times W$ ，给出了计算复杂度为 $\mathcal{O}(C \times H \times W)$ 。这在渐近上等同于全局方法。

1.3.1. 逐块计算的优点

双窗口策略的分块方法在时间特征对齐中提供了几个显著的优势，如下所示：

首先，窗口化计算为捕捉局部运动模式提供了更强的学习信号，使模型能够更好地区分同一场景中的不同运动行为。交通场景展示了结构化模式和个体物体运动，这些都受益于这种局部化分析方法。

其次，基于窗口的计算重新平衡了前景对象在相似度计算中的影响，有效地抵消了通常主导特征空间的背景区域的影响。考虑一个对象跨越了 $m \times n$ 个像素，其中 $m, n < l$ 。在全局相似性计算中，前景对象的计算在计算整体余弦相似度指标时将与 $(m \times n)/(H \times W)$ 成比例。然而，在基于窗口的方法中，相似性会在计算均方误差损失之前独立为每个窗口计算，前景对象在相似性中的贡献更高，达到 $(m \times n)/(l \times l)$ 。当窗口大小 l 小于特征维度 H 和 W 时，这种贡献将更加显著。前景元素的这种增强表示加强了显著物体的运动模式建模，因为它们通常占据比背景更小的空间区域。

第三，互补的窗口划分策略确保了全面的空间覆盖。给定窗口大小 l 和偏移量 $l/2$ ，任何大小不超过 $l \times l$ 的对象将在至少一个窗口中完全包含在其中，无论其在特征图中的位置如何。该特性防止前景对象被分割到多个窗口中，从而为学习对象运动模式提供连贯的监督信号。

第四，基于块的方法解决了前景和背景区域注意力平衡的挑战。通过以窗口为单位处理特征图，模型在特

www.xueshuxiangzi.com

征空间中分配了更平衡的计算注意力，而不是被通常占据场景大部分的背景区域主导。正如我们的实验结果所示，这种方法显著改善了时间对齐质量和检测性能，特别是对于自动驾驶场景中感兴趣的前景对象。

1.4. 结构卷积：多方向结构增强

StructConv 操作采用五种专门的 3×3 卷积，这些卷积共同增强前景特征中的结构细节。每种卷积针对特定的几何模式，同时通过参数组合保持计算效率。我们详细描述这些专门的卷积如下。

特征保留卷积。这种卷积通过计算邻域像素的标准加权和来保持原始特征表示。这是结构增强构建的基础，同时保留了基本的强度信息。

中心-周围对比卷积。此卷积通过将核的中心权重修改为所有周围权重的负和来增强局部对比度。此操作强调特征边界处的强度过渡并增强对象轮廓。

水平边缘卷积。这种卷积通过在卷积核的左侧列放置正权重、中间列为零、右侧列放置左侧权重的相反数，来检测水平结构。这种对称的权重模式创建了一个方向梯度检测器，能够突出车辆外观中常见的水平边缘。

垂直边缘卷积。该卷积通过一种类似于水平边缘卷积的原理来捕获垂直特征，但其核的顶部行是正权重，中间行是零，而底部行则是顶部行权重的相反值。这种排列方式增强了场景中的垂直边界和表面过渡。

对角结构卷积。该卷积通过一种特殊的权重排列识别角特征和角度过渡。其实现是将原始核的旋转版本从自身中减去，创建一种模式，对物体拐角处存在的对角强度梯度有强烈响应。

虽然这些卷积有效地增强了结构细节，但它们可能会引入虚假的前景特征，这可能会对检测产生负面影响。为了解决这个问题，增强后的特征会经过一个耦合的高度-语义验证过程，该过程有选择性地保留结构增强，同时抑制虚假特征。此验证过程利用了柱体编码特征中固有的通道内高度和语义相关性，以确保仅保留结构上合理的增强。

通过将它们的权重和偏置合并到一个操作中，五次卷积得以高效实现，从而在不增加单独卷积计算负担的情况下提供全面的结构增强。

1.5. 前景估计器实现

在 CDAM 和 IFAM 中使用的前景估计器 $\Phi(\cdot)$ 被实现为一个轻量级网络，用于识别前景区域。网络架构包括一个 3×3 卷积，能够将输入通道维度减半，紧随其后的是批量归一化和 ReLU 激活层，最后是带有 sigmoid 激活的 1×1 卷积以产生单通道输出。

在这里，$\Phi(\cdot)$ 由前景占据损失函数监督，该函数计算预测的前景图和前景标签在空间位置上的加权聚焦损失。此损失函数强调在考虑前景和背景像素之间的类别不平衡时，正确识别占据区域的重要性。

那么前景占用损失 $\mathcal{L}_{foreground}$ 可以被表述为

$$\mathcal{L}_{foreground} = \sum_{s=1}^{S} w_s \cdot \mathcal{L}_{focal}(M_s, M_s^{\text{gt}}), \tag{19}$$

，其中 S 是特征图中的空间位置总数，M_s 表示位置 s 处前景估计器预测的前景值，而 M_s^{gt} 是从边界框派生出的对应的真实前景标签。

权重因子 w_s 定义为

$$w_s = \frac{M_s^{\text{gt}} \cdot w_{pos} + (1 - M_s^{\text{gt}}) \cdot w_{neg}}{\max\left(\sum_{s=1}^{S} M_s^{\text{gt}}, 1\right)}, \tag{20}$$

，其中 $w_{pos} = 2$ 是正样本的权重，$w_{neg} = 1$ 是负样本的权重。分母通过正样本的总数量对权重进行归一化，最小值为 1 以防止除以零。

可以计算出焦点损失 $\mathcal{L}_{focal}$ ，如下所示

$$\begin{aligned}\mathcal{L}_{focal}(p, y) = &- (y \cdot 0.25 \cdot (1-p)^2 \cdot \log(p) \\ &+ (1-y) \cdot (1-\alpha) \cdot p^2 \cdot \log(1-p)).\end{aligned} \tag{21}$$

此监督信号指导前景估计器准确预测前景的占用情况，同时处理场景中前景和背景区域之间固有的不平衡问题。

2. 详细的数据集信息和实验配置

2.1. 详细数据集信息

在三个协同感知数据集上进行了实验：DAIR-V2X-C [?] 、V2XSET [?] 和 V2XSIM [?] 。(i) DAIR-V2X-C 专注于车-基础设施协同数据，是 DAIR-V2X 的一个子集，包含 39,000 帧。根据 [?] 中的官方基准，一个包含 9,311 帧对的同步子集 VIC-Sync 以 5:2:3 的比例拆分为训练/验证/测试集。我们采用补充标注 [?] ，感知范围设置为 $x \in [-102.4m, 102.4m]$ 和 $y \in [-51.2m, 51.2m]$ 。(ii) V2XSIM 是由 SUMO [?] 和 CARLA [?] 共同模拟的协同感知数据集。该数据集由 10,000 帧和 501K 标注框组成，并以 8,000/1,000/1,000 的比例拆分为训练/验证/测试集。感知范围设置为 $x \in [-32m, 32m]$ 和 $y \in [-32m, 32m]$ 。(iii) V2XSet 由 OpenCDA [?] 和 CARLA 共同模拟，是一个具有真实噪声模拟的 V2X 数据集。该数据集包含 11,447 帧，并以 6,694/1,920/2,833 的比例拆分为训练/验证/测试集。感知范围设置为 $x \in [-140m, 140m]$ 和 $y \in [-40m, 40m]$ 。

2.2. 三阶段训练与损失函数

DATA 框架采用一个顺序的三阶段训练策略，每个阶段都有专门的损失函数，以适应模型的特定方面。

在第一阶段，我们通过使用同步数据共同优化编码器、CDAM、IFAM 和检测头。该阶段的总体损失函数被表述为

$$\mathcal{L}_{\text{stage1}} = \mathcal{L}_{\text{det}} + \lambda_{\text{fore}} \cdot \mathcal{L}_{\text{foreground}} + \lambda_{\text{domain}} \cdot \mathcal{L}_{\text{domain}} \tag{22}$$

www.xueshuxiangzi.com

，其中 $\mathcal{L}_{\text{det}}$ 是由三个组成部分构成的检测损失，如

$$\mathcal{L}_{\text{det}} = \mathcal{L}_{\text{cls}} + \lambda_{\text{reg}} \cdot \mathcal{L}_{\text{reg}} + \lambda_{\text{dir}} \cdot \mathcal{L}_{\text{dir}}, \tag{23}$$

，$\mathcal{L}_{\text{cls}}$ 表示用于分类的 Sigmoid Focal Loss，$\mathcal{L}_{\text{reg}}$ 表示用于回归的加权平滑 ℓ_1 Loss，$\mathcal{L}_{\text{dir}}$ 表示用于方向预测的加权 Softmax 分类损失。加权因子被设置为 $\lambda_{\text{reg}} = 2.0$ 、$\lambda_{\text{dir}} = 0.2$ 、$\lambda_{\text{fore}} = 0.4$ 和 $\lambda_{\text{domain}} = 1.0$ 。

第二阶段着重于使用异步数据训练 PTAM，同时冻结第一阶段的所有参数。损失函数被定义为

$$\mathcal{L}_{\text{stage2}} = \mathcal{L}_{\text{det}} + \lambda_{\text{temporal}} \cdot \mathcal{L}_{\text{temporal}}, \tag{24}$$

，其中检测损失 $\mathcal{L}_{\text{det}}$ 保持与第一阶段相同的形式和参数，而 $\mathcal{L}_{\text{temporal}}$ 是 3.3.3 节末尾描述的多尺度和多窗口时间对齐损失。在这里，我们设置 $\lambda_{\text{temporal}} = 1.0$ 来平衡检测和时间对齐目标的影响。

2.2.1. 阶段 3：压缩网络

最终阶段通过使用异步数据训练传输压缩和恢复网络，并冻结前几个阶段的所有参数。损失函数结合了检测目标与重建保真度，如

$$\mathcal{L}_{\text{stage3}} = \mathcal{L}_{\text{det}} + \lambda_{\text{recon}} \cdot \mathcal{L}_{\text{recon}}, \tag{25}$$

，其中检测损失 $\mathcal{L}_{\text{det}}$ 与前几个阶段保持一致，而 $\mathcal{L}_{\text{recon}}$ 被实现为一个均方误差（MSE）损失，用于量化压缩后的特征重建质量。压缩网络采用 UNet 架构以在压缩和解压缩过程中维护空间特征关系。在此，我们设置 $\lambda_{\text{recon}} = 1.0$ 。

对于网络优化，所有三个阶段都采用了 Adam 优化器 [?] 。在第一阶段，初始学习率设定为 0.002，15 和 30 轮时衰减 10 × ，模型训练 40 轮。第二和第三阶段都采用固定学习率 0.001 训练 10 轮。整个训练过程中使用的批量大小为 2。这个连续的训练方法使每个组件能够在其预期功能上进行专门化，同时保持整个系统的一致性。

值得注意的是，在下面所有的可视化结果中，红色框表示预测结果，绿色框表示真实值。图 6 展示了 DATA 和 FFNet 在 300 ms 和 500 ms 延迟下转弯场景的延迟补偿性能。由于转弯通常涉及变速运动，对网络建模复杂运动的要求更高。在图 6 中，显示 DATA 在这些严重延迟的条件下表现出良好的预测性能，而 FFNet 的预测结果显示在这种严重延迟条件下准确补偿不匹配的能力有限。此外，PTAM 的消融结果显示在使用 PTAM 时检测性能显著提升，证明 PTAM 在获取高质量时序对齐特征方面的稳健性。图 7 进一步展示了在交叉交通流条件下 DATA 和 FFNet 的性能，这要求模型能够同时捕捉不同区域的运动趋势。从图 7 的右下子图可以观察到，由于延迟引起的不对齐可以达到一个车辆长度，这对自动驾驶中的可靠感知构成了重大挑战。在实现 PTAM 之后，DATA 成功地在双向（图中水平和垂直方向）实现了位移补偿，展现了本地和全局运动建模能力。在这种情况下，FFNet 能够有效补偿垂直方向的交通流，但在水平位移补偿方面表现不佳，反映出其全局监督方法可能导致的局部运动建模潜在缺陷。

2.3. 同步条件下的定性比较

图 8 展示了来自 DAIR-V2X 数据集的一个场景，其中的物体群被分布在三个不同的区域：靠近路边基础设施（协作方）、围绕自车、以及它们之间的中间区域。这样的配置形成了三个不同的感知区域：自车主要观测的区域、协作方主要观测的区域以及两个代理共同观测的重叠区域。这种混合观测模式对感知系统提出了重大的挑战，即提取领域不变特征以实现稳健的感知。DATA、FFNet 和 Where2comm 都在自车和协作方点云同时存在的区域表现出有效的检测。然而，FFNet 和 Where2comm 在自车或协作代理点云占主导的区域产生误检，而 DATA 则保持了准确的检测性能。这表明 DATA 通过密度一致性和可观测性一致性的领域对齐有效地学习了领域不变特征，在存在显著领域差距的场景中展现出稳定的检测性能。

图 10 展示了一个要求严格的协同感知场景，具有 280 m 范围和来自 V2XSET 数据集的噪声干扰。该场景包含密集的交通条件，有众多遮挡和在延伸距离上的点云稀疏，提供了评估协同感知能力的理想测试环境。FFNet 和 Where2comm 在协同代理提供可靠信息的区域实现了良好的感知结果，但由于点云稀疏和噪声干扰，在更远的距离上出现检测不准确和错误识别。DATA 通过学习的领域不变特征和前景增强，有效地抵消了噪声干扰和稀疏的远距离点云挑战，在整个扩展范围上实现了强大的检测性能。图 9 同样展示了来自 V2XSIM 的噪声场景，具有遮挡和稀疏点云。DATA 实现了有效的感知性能，而 FFNet 和 Where2comm 在稀疏点云区域表现出漏检和错误识别，进一步验证了 DATA 的强大和可靠的感知能力。

www.xueshuxiangzi.com

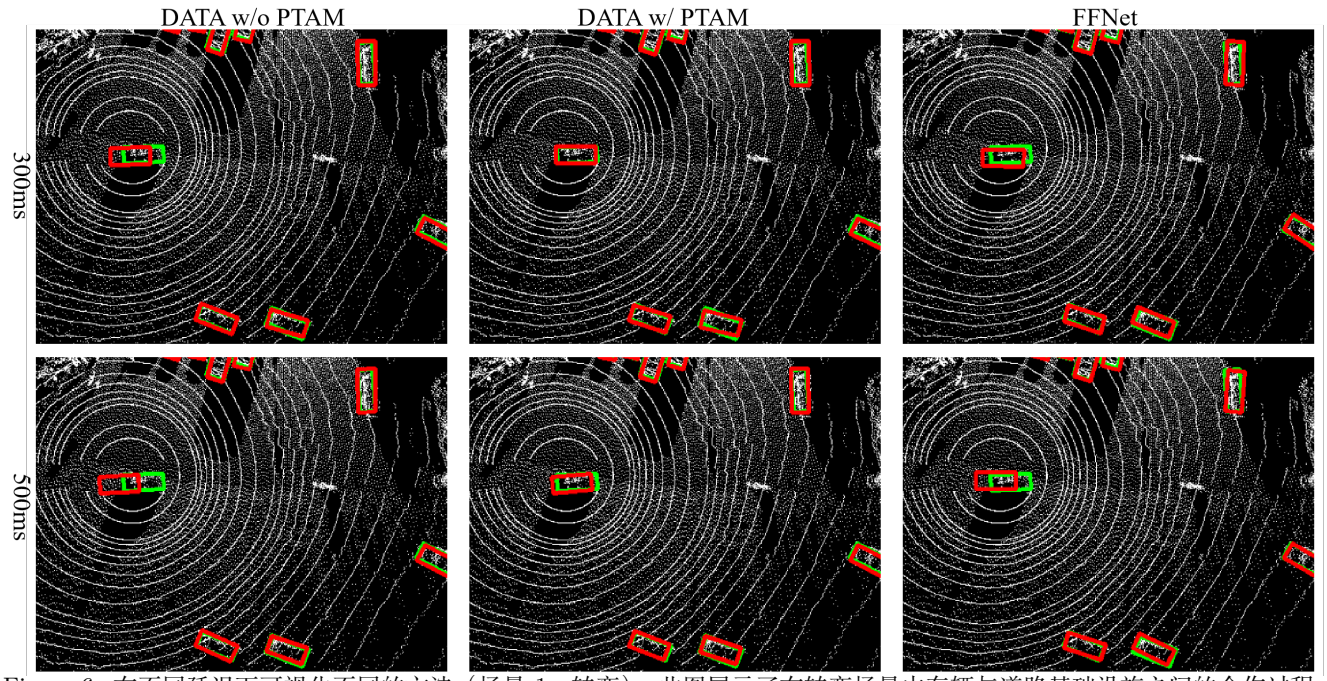

Figure 6. 在不同延迟下可视化不同的方法（场景 1：转弯）。此图展示了在转弯场景中车辆与道路基础设施之间的合作过程。DATA 展示了其对可变速度运动进行建模的卓越能力，这在严重延迟下通过精确补偿得到了证明。通过添加 PTAM 来对齐时间不同步，可以观察到显著的改进。

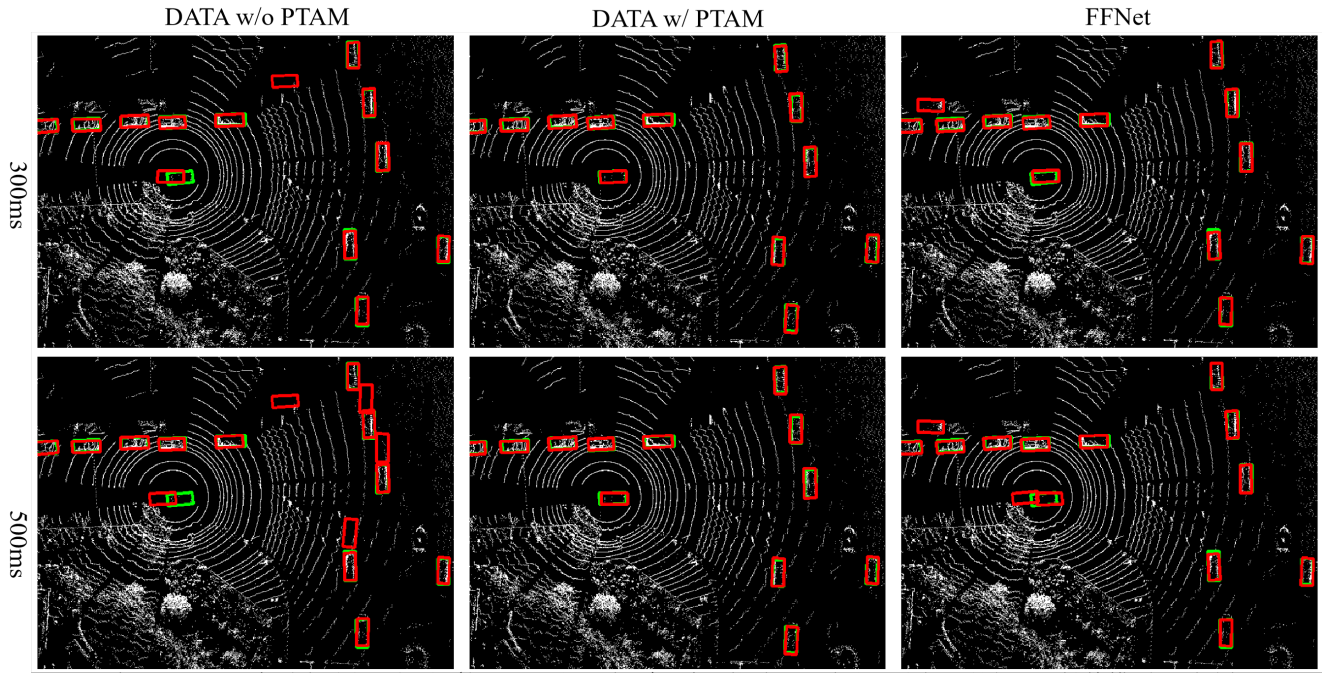

Figure 7. 在不同延迟下各种方法的可视化（场景 2：交叉路口）。该图描绘了一个交通流相交的场景，挑战模型同时捕捉不同区域的运动趋势。结果展示了 DATA 能够正确对齐两个不同方向的运动，证明了其在处理场景范围内运动信息方面的可靠能力。

www.xueshuxiangzi.com

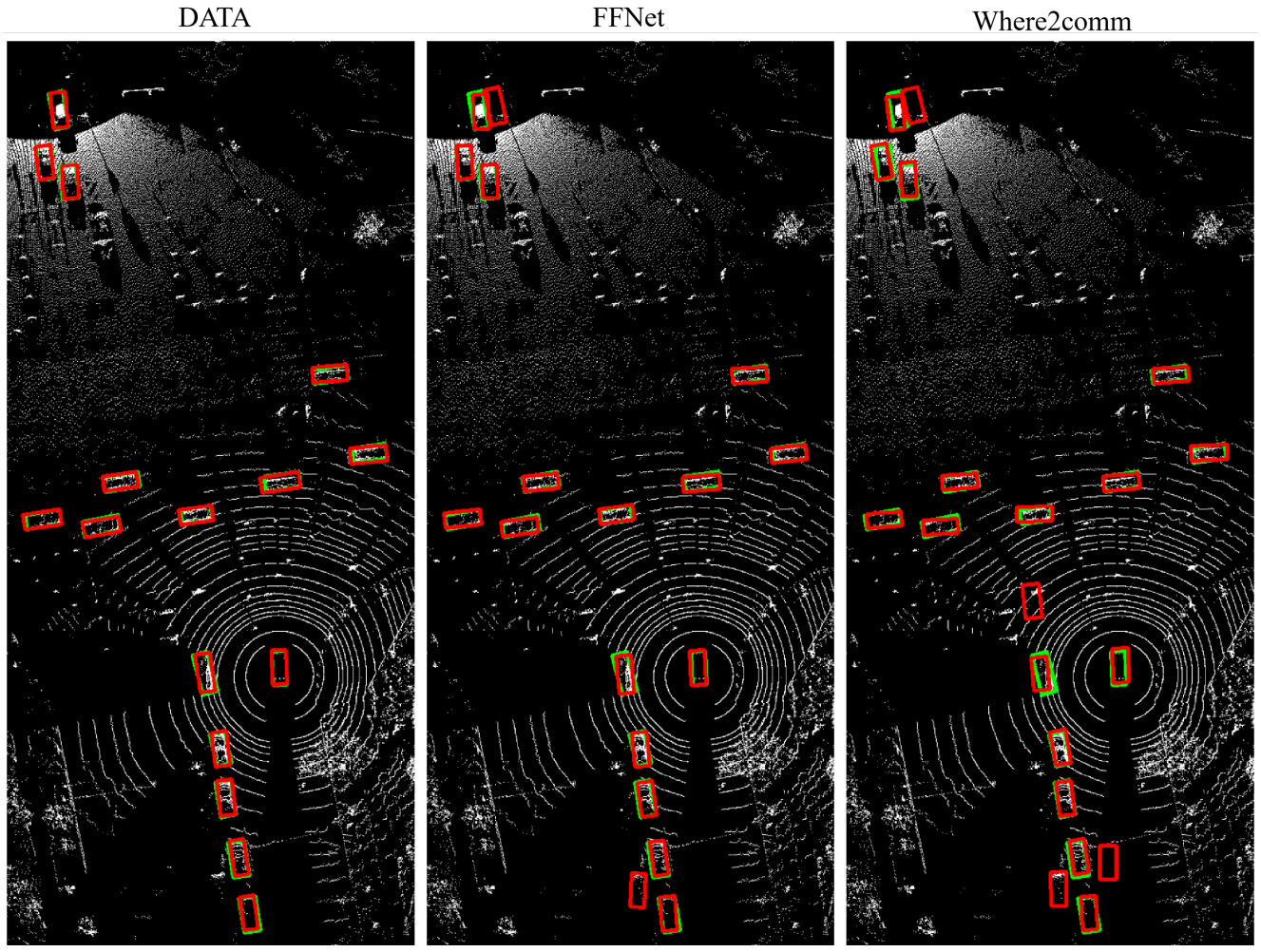

Figure 8. DAIR-V2X 上不同方法的可视化。DAIR-V2X 的车-基础设施协同感知场景显示，尽管存在域间差异，但 DATA 在保持所有区域的准确检测方面优于 FFNet 和 Where2comm，在混合可观测性挑战下展现出卓越的域不变特征提取能力。

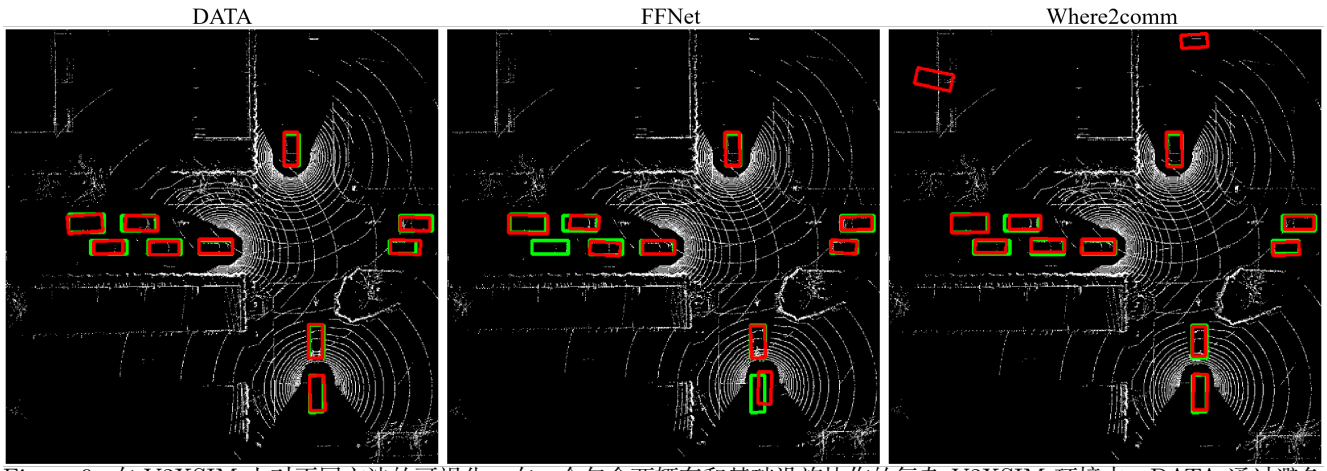

Figure 9. 在 V2XSIM 上对不同方法的可视化。在一个包含两辆车和基础设施协作的复杂 V2XSIM 环境中，DATA 通过避免漏检和假阳性，在处理遮挡物体和稀疏点云时，提供了卓越的感知性能，优于竞争对手。

www.xueshuxiangzi.com

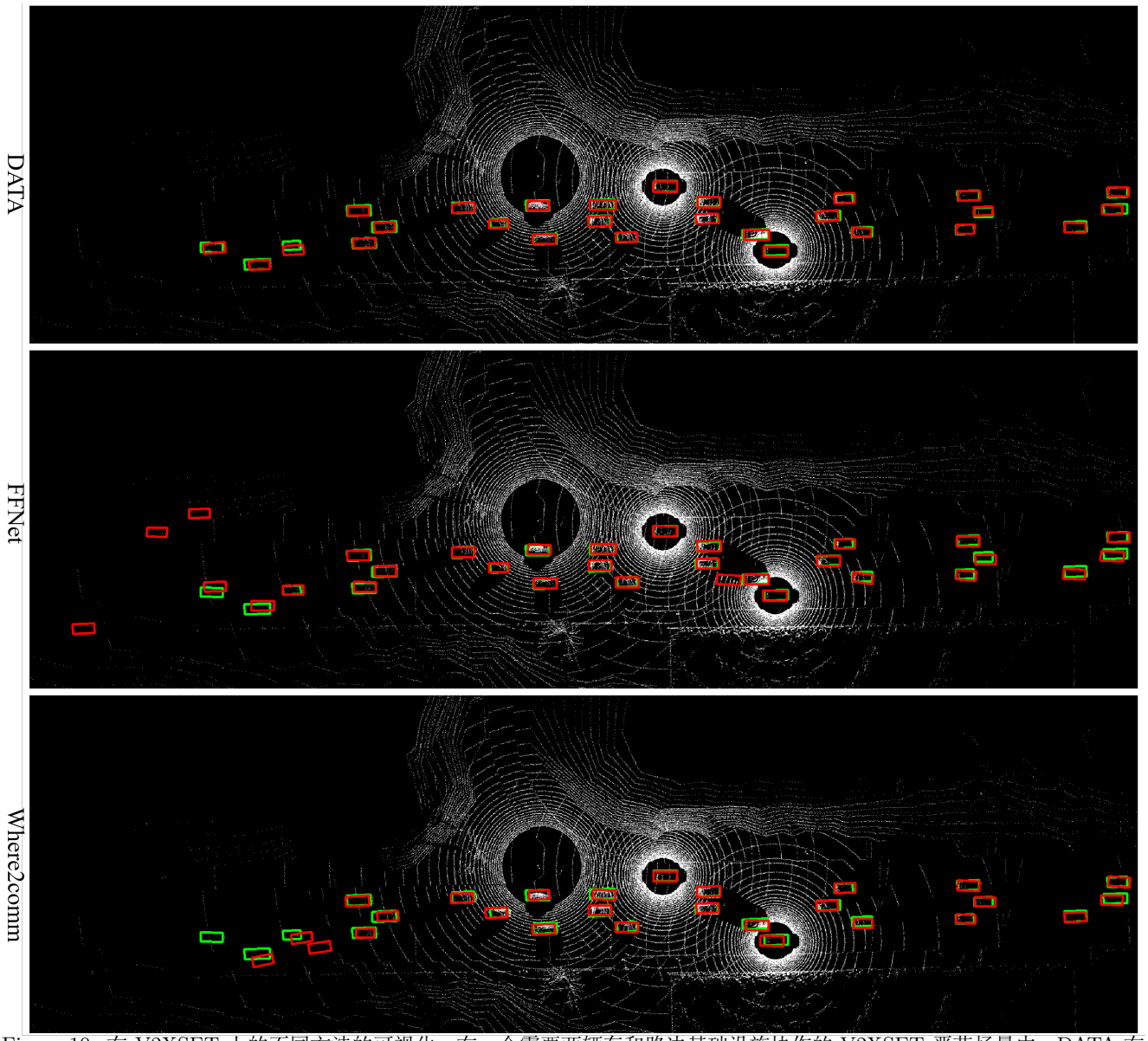

Figure 10. 在 V2XSET 上的不同方法的可视化。在一个需要两辆车和路边基础设施协作的 V2XSET 严苛场景中，DATA 在密集交通、遮挡和稀疏的远距离点云的情况下，依然能准确检测物体，表现优于竞争对手，展示了卓越的噪声抑制能力和领域不变特征学习能力。

www.xueshuxiangzi.com

Symbol	Definition	Section/Module
3.1 Problem Formulation and Overall Architecture		
N	Number of agents in collaborative perception system	Problem Formulation
i, j	Subscripts denoting ego agent and collaborative agents, where $i \neq j$	Problem Formulation
$X_i(t)$	Latest data of ego agent at current time t	Problem Formulation
$X_j(t-\tau)$	Data transmitted by collaborative agents at timestamp $t-\tau$	Problem Formulation
$X_j(t-\tau-\Delta T)$	Data of collaborative agent at timestamp $t-\tau-\Delta T$	Problem Formulation
τ	Transmission delay	Problem Formulation
ΔT	Time interval	Problem Formulation
3.2.1 Proximal-region Hierarchical Downsampling (PHD)		
$\mathcal{O}_i$	Set of all objects observable to ego agent i	CDAM-PHD
N_i^{total}	Total number of observable objects	CDAM-PHD
o_k	k -th object	CDAM-PHD
$\mathcal{O}_i^{prox}$	Objects within the proximal region	CDAM-PHD
$d_i(o_k)$	Distance from ego agent i to k -th object	CDAM-PHD
d_{th}	Distance threshold	CDAM-PHD
N_{proc}	Number of objects selected from $\mathcal{O}_i^{prox}$ for subsequent processing	CDAM-PHD
N_{max}	Predefined maximum number of objects	CDAM-PHD
$B_{i,k}^{out}$	Oriented bounding box of k -th object	CDAM-PHD
$B_{i,k}^{in}$	Inner bounding box with scaling factor α	CDAM-PHD
(x_k, y_k, z_k)	Center coordinates of k -th object	CDAM-PHD
(h_k, w_k, l_k)	Dimensions (height, width, length) of k -th object	CDAM-PHD
θ_k	Orientation of k -th object	CDAM-PHD
α	Scaling factor to adjust bounding box dimensions, $\alpha \in (0, 1)$	CDAM-PHD
$R_{i,k}^{in}, R_{i,k}^{out}$	Point sets in inner and outer regions	CDAM-PHD
β_{in}, β_{out}	High and conservative downsampling ratios for inner and outer regions	CDAM-PHD
$\mathrm{FPS}(\cdot, \cdot)$	Farthest Point Sampling operation	CDAM-PHD
$\tilde{R}_{i,k}^{in}, \tilde{R}_{i,k}^{out}$	Downsampled inner and outer regions	CDAM-PHD
$\tilde{P}_{i,k}$	Point clouds of k -th object after hierarchical downsampling	CDAM-PHD

Table 4. DATA 论文符号表 - 第一部分: 问题表述和 PHD

www.xueshuxiangzi.com

Symbol	Definition	Section/Module
3.2.2 Observability-constrained Discriminator (OD)		
H_i, H_j	BEV features of ego and collaborative agents	CDAM-OD
C, H, W	Channel dimension, height, and width of features	CDAM-OD
M_i, M_j	Observability maps, $M_i, M_j \in \mathbb{R}^{1\times H\times W}$	CDAM-OD
$\Phi(\cdot)$	Foreground estimator	CDAM-OD
$H_{j\to i}, M_{j\to i}$	Collaborative features and observability map projected onto ego coordinates	CDAM-OD
$\mathcal{V}$	Set of valid grids in transformed feature map	CDAM-OD
H_j^{comp}, M_j^{comp}	Complemented feature and observability map	CDAM-OD
$I_{\mathcal{V}}$	Indicator function (equals 1 for points in $\mathcal{V}$ and 0 for others)	CDAM-OD
W	Observability weighting map, $W \in \mathbb{R}^{1\times H\times W}$	CDAM-OD
$\mathcal{L}_{domain}$	Domain alignment objective	CDAM-OD
$\mathcal{S}$	Set of all spatial positions	CDAM-OD
sp	One spatial position	CDAM-OD
W_{flat}	Flattened observability weighting map	CDAM-OD
Ψ_θ	Feature extractor (point clouds as input and BEV features as output)	CDAM-OD
D_μ	Discriminator	CDAM-OD
$\mathcal{L}_{BCE}$	Binary cross-entropy loss	CDAM-OD
Z	Domain label (0 for ego agent and 1 for collaborative agent)	CDAM-OD
γ	Negative scaling factor in gradient reversal layer ($\gamma = -0.1$)	CDAM-OD

Table 5. DATA 论文符号表 - 第二部分：观测约束判别器

www.xueshuxiangzi.com

Symbol	Definition	Section/Module
3.3 Progressive Temporal Alignment Module (PTAM)		
$F_j(t, x)$	Features at collaborative agent at time t and position x	PTAM
$v(t - \Delta t, x)$	Velocity field describing motion of features at time $t - \Delta t$	PTAM
Δt	Time interval within range of typical transmission delay	PTAM
Δp	Motion field representing velocity, $\Delta p \in \mathbb{R}^{2\times H\times W}$	PTAM
ξ	Temporal scaling factor corresponding to Δt	PTAM
F_j^{inter}	Intermediate feature predicted by collaborative agent	PTAM
$\hat{F}_j(t-\tau)$	Feature from collaborative agent j at time $t-\tau$, as received by ego agent	PTAM
$\hat{F}_j^{inter}$	Intermediate feature from collaborative agent j , as received by ego agent	PTAM
$\hat{F}_j(t)$	Temporally aligned features	PTAM
F_{latest}, F_{prev}	Latest feature and its previous feature	PTAM
ΔF	Temporal feature difference, $\Delta F = F_{latest} - F_{prev}$	PTAM
w_{samp}	Sampling weight, $w_{samp} \in \mathbb{R}^{1\times H\times W}$	PTAM
$f_{warp}(\cdot)$	Bilinear sampling operation	PTAM
$s1, s2$	Superscripts indicating first stage and second stage	PTAM
ΔM	Motion difference field	PTAM
f_M	Global context vector	PTAM
f_T	Temporal encoding with sinusoidal positional embeddings	PTAM
3.3.3 Multi-window Self-supervised Training Strategy		
s	Spatial scale index	PTAM Training
h_s, w_s	Feature plane size at spatial scale s	PTAM Training
l	Window size	PTAM Training
$\mathcal{W}_1, \mathcal{W}_2$	Two complementary window partitioning strategies	PTAM Training
$w_{m',n'}$	Window with top-left corner at position (m', n')	PTAM Training
$w_{p',q'}$	Window with top-left corner at position (p', q')	PTAM Training
N_{window}	Total number of windows	PTAM Training
F_j^{gt}	Ground truth features	PTAM Training
$\cos(\cdot, \cdot)$	Cosine similarity between predictions and ground truth features	PTAM Training
$\mathcal{L}_{inter}^s, \mathcal{L}_{final}^s$	Loss functions for intermediate and final predictions at scale s	PTAM Training
$\mathcal{L}_{temporal}$	Total temporal alignment loss computed across all three scales	PTAM Training

Table 6. DATA 论文符号表 - 第三部分：渐进时间对齐模块

www.xueshuxiangzi.com

Symbol	Definition	Section/Module
3.4 Instance-focused Feature Aggregation Module (IFAM)		
H_a	BEV feature of a -th agent, $H_a \in \mathbb{R}^{C\times H\times W}$	IFAM
M_a	Foreground mask, $M_a = \Phi(H_a)$	IFAM
H_a^{fore}, H_a^{back}	Foreground and background features	IFAM
H_a^{enh}	Enhanced foreground features after structural convolution	IFAM
StructConv$(\cdot)$	Structural convolution with specialized convolutions	IFAM
H_a^{cat}	Concatenated features, $H_a^{cat} \in \mathbb{R}^{2C\times H\times W}$	IFAM
W_a^s, W_a^c	Spatial attention and channel attention	IFAM
W_a^{init}	Initial attention weights, $W_a^{init} = W_a^s \oplus W_a^c$	IFAM
W_a^{verif}	Verification weights	IFAM
CS	Channel shuffle operation	IFAM
GConv$(\cdot)$	Group convolution for independent weight generation	IFAM
H_a^{verif}	Verified foreground feature	IFAM
$H_a^{refined}$	Individually refined BEV features	IFAM
ϵ	Learnable parameter balancing contribution of background features	IFAM
4. Experiments and General Symbols		
$\text{AP}_{50}, \text{AP}_{70}$	Average Precision at IoU thresholds of 0.50 and 0.70	Evaluation Metrics
IoU	Intersection-over-Union	Evaluation Metrics
μ_{pos}, σ_{pos}	Mean and standard deviation for position noise	Noise Modeling
μ_{rot}, σ_{rot}	Mean and standard deviation for orientation noise	Noise Modeling
σ_{local}	Standard deviation for localization noise	Robustness Testing
σ_{head}	Standard deviation for heading noise	Robustness Testing
$\odot$	Element-wise product	General Operations
$\oplus$	Concatenation operation	General Operations
$[\cdot]$	Concatenation along specified dimension	General Operations
$\lvert\cdot\rvert$	Cardinality of a set	General Operations
softmax$(\cdot)$	Softmax function	Activation Functions
ReLU$(\cdot)$	ReLU activation function	Activation Functions
Sigmoid$(\cdot)$	Sigmoid function	Activation Functions
MLP$(\cdot)$	Multi-layer perceptron	Network Structures
$\text{Conv}_{1\times1}(\cdot)$	1×1 convolution operation	Network Structures

Table 7. DATA 论文的符号表 - 第四部分：IFAM 和常用符号

www.xueshuxiangzi.com