arXiv:2304.01534v2 [cs.CV] 8 Sep 2023

FedBEVT：道路交通系统中的联邦学习鸟瞰视角感知 Transformer

Rui Song*、Runsheng Xu*、Andreas Festag、Jiaqi Ma 和 Alois Knoll,

摘要——**鸟瞰视角（BEV）感知在自动驾驶领域正变得日益重要。它利用多视角摄像头数据训练 Transformer 模型，直接将道路环境的感知结果投射到 BEV视角。然而，训练 Transformer 模型通常需要大量数据，而道路交通摄像头数据往往具有私密性，因此通常不会共享。联邦学习提供了一种解决方案，使客户端能够协作训练模型，而无需交换数据，只需交换模型参数。本文介绍 FedBEVT，这是一种用于鸟瞰视角感知的联邦 Transformer 学习方法。为了解决 FedBEVT 中两种常见的数据异质性问题：（i）传感器位姿多样；（ii）感知系统中的传感器数量不一，我们分别提出两种方法——基于摄像头注意力个性化的联邦学习（FedCaP）和自适应多摄像头掩蔽（AMCM）。为在真实场景中评估我们的方法，我们构建了一个包含四种典型联邦学习应用场景的数据集。研究结果表明，FedBEVT 在四种用例中均优于基线方法，展现了我们的方法在提升自动驾驶鸟瞰视角感知方面的潜力。**

索引词——**联邦学习、鸟瞰图、道路环境感知、视觉 Transformer、协同智能交通系统**

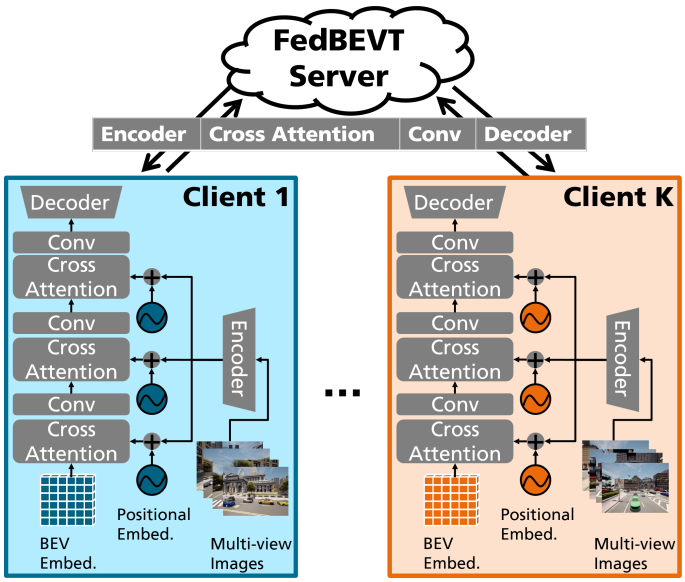


图 1. 采用摄像头注意力个性化的 FedBEVT。位置嵌入被视为每个客户端的私有参数。BEVT 的其他部分（灰色区域）则共享给服务器进行聚合。

I. 引言

近年来，自动驾驶领域对鸟瞰图（BEV）感知的关注显著增加。交通场景的BEV表征有几个特别吸引人的优点。首先，由于所有自动驾驶车辆（AV）都位于地面 [1], [2],, BEV表示可以省略z轴，从而提高感知结果的效率 [3]。其次，它提供了丰富的上下文和几何信息，可直接用于规划等下游任务 [4], [5]。尤其是在首篇全面探讨自动驾驶车辆及网联自动驾驶汽车控制系统的综述中 [5]，作者深入阐述了感知系统在实现低层控制方面的关键作用。最后，BEV表示可以作为不同传感器模态（如摄像头、激光雷达）和不同时间戳数据的统一空间，从而无需太多额外工作即可进行融合 [6]。传统上，实现多源传感器融合中的时间对齐需要额外的鲁棒算法，例如 [7]中提出的流行的估计-预测一体化框架，以及 [8]中针对全球定位系统信号受挑战场景的计算轻量型多传感器融合定位算法。

由于成本较低，鸟瞰图感知研究已开始探索使用多视角摄像头数据来预测鸟瞰图地图 [9]–[13]。然而，从二维数据推断三维信息颇具挑战，因为单目摄像头只能提供二维信息。因此，近期研究利用视觉 Transformer（以其推理不同数据之间相关性的能力而闻名）来解决二维到三维问题 [3], [9],[11], [14]–[16]。尽管取得了令人鼓舞的成果，这些基于 Transformer 的方法是在有限的公开数据集（如 NuScenes [17], ）上训练的，可能无法很好地泛化。汽车原始设备制造商（OEM）、传感器供应商、软件解决方案提供商和研究机构等公司，需要在多辆车辆上安装多传感器系统，并在长时间驾驶过程中收集大量数据，

* 前两位作者贡献相同。本研究得到德国联邦数字与交通部（BMVI）资助，项目包括“KIVI——英戈尔施塔特交通人工智能项目”和“5GoIng——英戈尔施塔特 5G 创新概念项目”。Rui Song 和 Andreas Festag 任职于德国英戈尔施塔特 85051 的弗劳恩霍夫运输与基础设施系统研究所 IVI，电子邮箱：{rui.song, andreas.festag}@ivi.fraunhofer.de。Rui Song 和 Alois Knoll 任职于德国加兴 85748 的慕尼黑工业大学机器人、人工智能与实时系统教席，电子邮箱：rui.song@tum.de,knoll@in.tum.de。Runsheng Xu 和 Jiaqi Ma 任职于美国加利福尼亚大学洛杉矶分校土木与环境工程系，邮编 CA 90024，电子邮箱：rxx3386@ucla. edu, majiaqimark@gmail.com。Andreas Festag 任职于德国英戈尔施塔特应用科学大学 CARISSMA 电动、网联与安全出行研究所（C-ECOS），地址为德国英戈尔施塔特 85049，电子邮箱：andreas.festag@carissma.eu。

这些数据可用作训练数据集。然而，此类数据通常属于私有数据，获取成本可能高得令人难以承受，而且有些数据完全不共享。在这种情况下，联邦学习的出现提供了一种解决方案，使各方能够在不交换数据的情况下开展协作训练，并解决数据隐私问题。

联邦学习方法具有诸多优势，但在多个客户端之间进行训练时可能会导致数据异质性。在基于摄像头的BEV感知场景中，数据异质性这一挑战源于传感器配置的多样性，包括车辆或交通基础设施上安装的摄像头数量以及传感器姿态的差异。 [18]。这些差异可能导致各客户端的数据特征存在显著差别，而原生联邦学习方法或许无法克服这一问题。本文旨在探讨一个实际问题：如何在联邦学习框架中利用外部数据的优势，同时最大限度地降低数据异质性的影响，从而尽可能提升本地模型的性能。

为应对上述数据异质性挑战，我们提出了一种新的鸟瞰图感知 Transformer（BEVT）联邦学习框架，称为 FedBEVT（如图 1 所示）。该框架支持有选择地与服务器共享 BEV Transformer 的特定部分。遵循大多数基于 Transformer 的鸟瞰图感知研究的做法 [3], [11]‚[12],，我们将一个 BEV视觉模型划分为五个组成部分：用于图像特征提取的编码器；包含摄像头几何信息的位置嵌入；将前视图投影到鸟瞰图的交叉注意力模块；用于优化 Transformer 中特征的卷积层；以及将 BEV 表示转换为最终预测的解码器。

鉴于不同客户端之间的摄像头位姿存在差异，共享整个模型进行全局聚合会导致性能不理想。为此，我们在 FedBEVT 中采用摄像头关注个性化，通过将包含摄像头内参和外参信息的位置嵌入设为私有来实现。此外，为解决每个客户端摄像头数量不同的问题，我们引入自适应多摄像头掩码（AMCM），通过重叠掩码训练多视图数据，使 BEV 尺寸保持一致；这些掩码根据所有摄像头的合并视场构建。这确保了联邦学习过程中 BEV 嵌入的尺寸一致，并使采用不同摄像头系统的客户端能够共同学习。为评估所提方法的有效性，我们创建了一个包含这些差异的数据集，并在联邦设置下将其分配给多个客户端，以模拟智能交通系统中的真实联邦学习场景。实验结果表明，对于大多数客户端，我们的方法在降低通信成本的同时取得了更高的测试准确率。

我们的贡献可概括如下：

- 我们针对道路交通感知应用中的 BEV Transformer 提出了一种联邦学习框架。据我们所知，这是首个专门为鸟瞰图感知任务设计的联邦 Transformer 训练框架。
- 我们为联邦设置下交通场景中的鸟瞰图感知提供了一个基准多视图摄像头数据集。为应对数据异质性这一挑战，我们考虑了 C-ITS 领域联邦学习应用场景中的两种常见数据变化：（*i*）不同的传感器位姿，以及（*ii*）感知系统中传感器数量的变化。
- 我们提出了一种基于摄像头关注个性化（FedCaP）的联邦学习框架，用于训练 BEV Transformer 模型。通过考虑训练数据与 Transformer 之间的关系，我们调整了联邦学习中 Transformer 的个性化方式，为每个客户端提供更适合其本地数据的定制模型，从而提高本地鸟瞰图感知的准确率。
- 我们提出了一种自适应多摄像头掩码方法，使拥有不同数量摄像头的客户端能够通过联邦学习训练模型。
- 我们针对道路交通系统提出了四种典型用例，并据此在联邦设置下将基准数据划分给不同客户端。这些分布式数据集用于展示通过我们的联邦学习系统和基线方法训练 BEVT 的性能。我们开源的 FedBEVT 实现已在 GitHub 上公开发布 [1]。

II. 相关工作

A. 道路交通系统中的鸟瞰图感知

鸟瞰图感知涉及将输入图像序列从透视视图转换为鸟瞰图，从而支持三维边界框检测或语义地图分割等感知任务。鸟瞰图感知主要有两种方法：基于几何的方法和基于 Transformer 的方法。基于几何的方法利用摄像头外参与内参之间天然的几何投影关系，将透视视图投影到鸟瞰图。例如，LSS [18] 通过对深度使用类别分布，并结合上下文向量，将二维图像提升为视锥形点云；随后利用摄像头外参和内参将其散射到 BEV 平面上。BEVDet [10] 采用了类似的框架，并增加了一个 BEV 编码器，以进一步优化投影后的 BEV 表征。另一方面，基于 Transformer 的方法能够通过可学习嵌入隐式利用摄像头几何信息。BEVFormer [9] 利用几何信息获取初始化的采样偏移，并应用可变形注意力，将图像特征查询到可学习的 BEV 嵌入中。尽管该方法准确率较高，但速度较慢。CVT [11]根据每个摄像头的内外参标定结果，为其分别生成位置嵌入。这些嵌入会添加到图像特征和可学习的 BEV 嵌入中，随后应用标准的跨视图注意力 Transformer，将图像特征提取至 BEV 嵌入中。在 CVT 的基础上，CoBEVT [3] 以一种新颖的融合轴向注意力取代标准注意力，从而大幅

[1]https://github.com/rruisong/FedBEVT

降低计算成本，并构建层次结构以探索多尺度特征。由于 CoBEVT 在推理速度和准确率之间实现了良好的权衡，本文选用它作为联邦学习研究中部署的模型。

B. 面向智能交通系统的联邦学习

联邦学习为智能交通系统带来了巨大的机遇和潜力。它能够在遵守隐私法规的同时，利用更多数据训练性能更佳的模型 [19]–[22]。此前的研究已开始探索将联邦学习用于训练适用于智能交通系统相关应用的模型，例如二维目标识别与检测 [23]–[25],行为预测 [26], 定位 [27], [28], 以及基于区块链的自动驾驶车辆 [29]–[32]。然而，联邦学习在鸟瞰图感知中的应用仍未得到探索。此外，现有研究 [33]–[35] 主要关注联邦学习中由系统多样性和标签偏移导致的异质性挑战，而尚未研究异构传感器配置这一数据异质性的主要成因。为弥补这些空白，我们聚焦于鸟瞰图感知的应用，并致力于解决由传感器配置造成的异质性问题。

C. 个性化联邦学习

联邦学习是一种协作式机器学习方法，网络中的大量客户端共同参与学习 [36]。通常，服务器会启动学习过程，客户端则下载全局模型，并使用各自的本地数据进行训练。随后，客户端将训练后的本地模型上传回服务器，通过聚合更新全局模型。经过多个通信轮次后，即可获得一个在全局数据上表现良好的模型。然而，这类模型在本地数据上的表现可能不佳，因为在大多数联邦学习应用中，本地数据分布往往与全局数据分布不同。

为了解决联邦学习中本地数据分布差异这一挑战，研究者提出了个性化联邦学习作为解决方案 [37]–[39]。这种方法会针对每个客户端定制模型，以适应其本地数据的独特特征。实现个性化联邦学习最流行且有效的方法之一是基于架构的方法 [40]。该方法将模型参数解耦，只共享并在客户端之间聚合一部分参数，而私有参数则仅使用本地数据进行学习。以往研究尝试根据模型架构 [41]–[44] 或数据相似性 [45]–[47]来选择这些私有参数。然而，这些方法都不是专为训练个性化 Transformer 模型而提出的，因为与其他机器学习模型相比，Transformer 模型的结构截然不同。因此，我们进一步研究如何开发针对 Transformer 模型独特特性量身定制的个性化联邦学习方法。

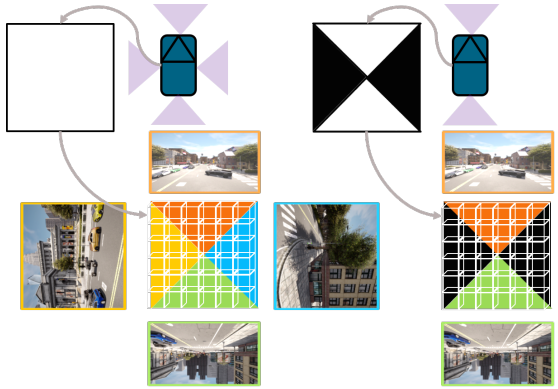


图 2. AMCM 在两辆配备不同数量摄像头的车辆中的应用示意图。通过使掩码适应各感知系统的总视场，可以保持鸟瞰图嵌入的大小，并对其进行调整以用于 FedBEVT。

D. 面向 *Transformer* 的联邦学习

Transformer 是一种采用自注意力机制的神经网络架构，最初是为自然语言处理任务设计的 [48]。它在计算机视觉任务中也表现出色，例如目标识别（ViT） [49]–[51]。正如 [52], 所展示的，与传统卷积神经网络相比，Transformer 架构在大型数据集上训练时性能更佳。此外，该架构在小目标检测领域也展现出良好潜力，无人机图像便是一个例证。联邦学习有望通过整合多个客户端的数据进一步改进训练过程，因此是训练 Transformer 模型的一种颇具前景的方法。 [53] 中的最新实验结果表明，Transformer 架构对分布偏移的鲁棒性更强，并能提升联邦学习在目标检测任务中的效果。

然而，将其应用于鸟瞰图语义分割任务仍有待进一步研究；该任务对自动驾驶和智能交通系统至关重要。我们的观察表明，在用于鸟瞰图语义分割任务的联邦学习中，摄像头配置的差异是造成数据异质性的主要来源。为应对这一挑战，我们的联邦学习框架采用个性化位置嵌入，并结合交叉注意力和卷积神经网络层，以提升 Transformer 的性能并利用更多数据。

III. 方法

A. 问题形式化

在联邦学习训练场景中，我们考虑 K个客户端，每个客户端都有自己的本地数据集，其中包含 N_k 个数据点及其对应的鸟瞰图语义掩码Y_k。每个数据点 $X_{k,\ i}$ 包含来自多视角摄像头系统的 L_k 张图像，因此其格式为（X_k，Y_k）。其中，$X_k \in \mathbb{R}^{N_k \times L_k \times H \times W \times 3}$ 和 $Y_k \in \mathbb{R}^{N_k \times h \times w \times 2}$，$H$ 和W 分别表示单张图像的高度和宽度，h 和w 分别表示鸟瞰图语义掩码的高度和宽度。语义掩码包含两个类别：背景和车辆。

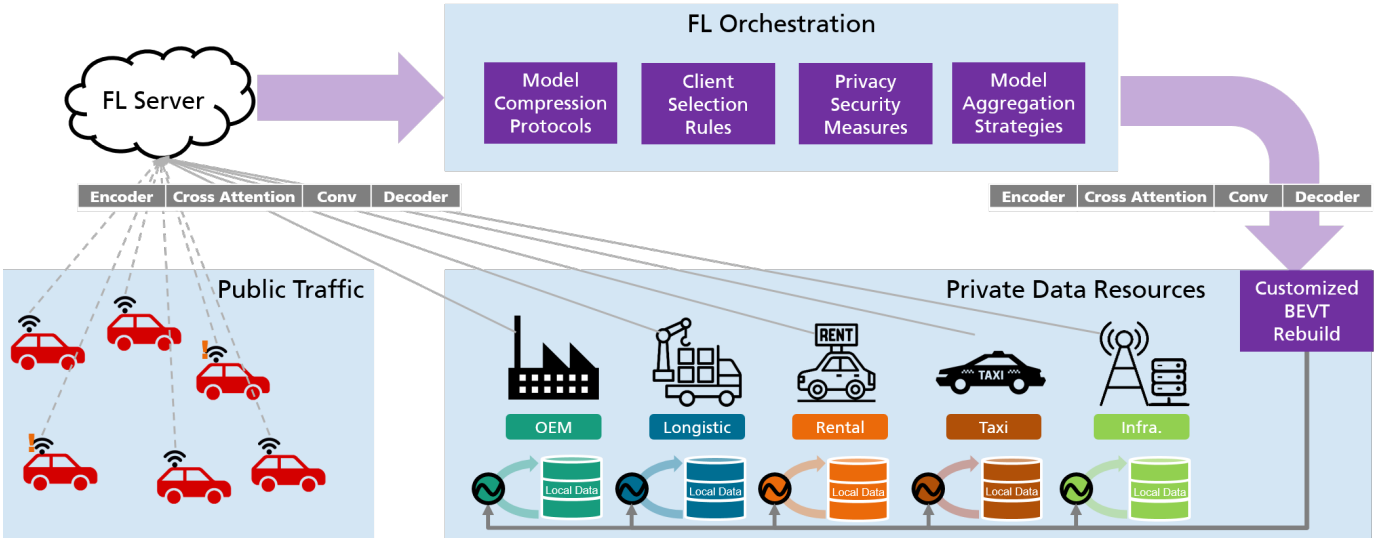


图 3. FedBEVT 系统概览展示了来自公共交通和私有数据资源的车辆客户端。联邦学习服务器通过模型压缩、客户端选择和安全聚合协议管理训练过程，这些协议与典型深度学习模型联邦学习框架中的协议类似。请注意，智能道路交通系统中的不同客户端会导致数据异质性。

这些客户端可能来自各行各业，包括原始设备制造商（OEM）、物流公司和出租车公司、汽车租赁公司，或交通管理部门等基础设施所有者。

在一般的联邦学习中，每个客户端从服务器接收全局模型，并在本地训练若干训练轮次，然后将训练后的模型返回服务器，融合为新的全局模型。此过程会持续一定数量的通信轮次，直至获得训练好的模型。联邦学习的目标是最小化以下内容：

$$f(w) = \frac{1}{N}\sum_{k=1}^{K}\sum_{i=1}^{N_k} f_i(w) = \mathbb{E}_{X_i \in X}[f_i(w)], \qquad (1)$$

其中 $f_i(w) = l(BevT(X_i)\ Y_i;\ w)$，$l$ 表示损失函数，$BevT$ 表示一种基于Transformer的BEV模型，包含三个主要组件：图像编码器、BEV Transformer和BEV解码器。图像特征编码器利用CNN网络对多摄像头图像进行编码。BEV Transformer的核心通常是一种多层注意力结构 [48]。在本文中，我们采用Fax-Attention方案 [3]，该方案同时涉及图像特征与BEV查询之间的稀疏跨视角注意力和自注意力。此外，BEV查询和位置嵌入需要作为BEVT的输入。BEV查询是一种用于表示网格世界的可学习嵌入，其最终目标是在该查询中定位其他交通目标的位置。位置嵌入包含每个Transformer图像特征的位置信息，包括用于将摄像头视角转换为BEV视角的坐标系转换相关信息。这提高了生成鸟瞰图特征时输入的直观性。我们采用 [11]中的方法，并将位置嵌入的输出表示为

如下所示，假设每个摄像头的外参 j为$R_{i,\ j,\ k} \in \mathbb{R}^{3\times3}$ （旋转）和$t_{i,\ j,\ k} \in \mathbb{R}^3$:

$$z_{i,k} = \mathbf{\Phi}^{(v_k)}(\mathbf{T}(R_{i,j,k}, t_{i,j,k}, p_{i,j,k})_{j=1}^{L_k}). \qquad (2)$$

这里，p 中的图像坐标系位置特征首先使用$\mathbf{T}$转换到三维车辆坐标系。接下来，通过 $\mathbf{\Phi}^{(v_k)}$对所得位置进行映射，生成位置嵌入 $z_{i,k}$。我们将 $\mathbf{\Phi}$ 中的参数记为 v_k，以区别于第 III-B 节所用 Transformer 模型 u中的其他参数。

鸟瞰图特征经过多个 Fax-Attention层处理后，再由解码器网络处理，以生成鸟瞰视角下的语义分割结果，从而能够直接确定运动平面上周围汽车的位置。我们的方法解决了两个关键挑战：(*i*) 为独特的传感器系统开发定制模型，以及 (*ii*) 利用摄像头数量各异的数据有效训练模型。

B. 摄像头关注个性化

不同客户端的摄像机位置存在差异，这可能导致传统联邦学习的结果偏离本地数据的最优值。这是因为在构建 Transformer 模型时，摄像机位置差异被编码到位置嵌入中，进而直接影响参数训练。

为应对这一挑战，我们提出了基于摄像头关注个性化的联邦学习（FedCaP）。该方法将如公式 2 所示，用于位置嵌入的多层感知机中的本地训练参数 v_k 与 Transformer分离，并将其作为给定客户端 k的私有参数。除 v_k外，我们共享 BEVT 中 w_k 的其余参数，并将这些公共参数称为客户端k的 u_k 。因此，每个通信轮次中的全局更新表示为来自所有

客户端的聚合后的公共参数，我们将其记为 u。随后，本地训练的目标是最小化：

$$F_k(u, v_k) = \frac{1}{N_k} \sum_{i=1}^{N_k} f_i(u, v_k), \tag{3}$$

全局目标由公式1改写为：

$$f(u, v_1, .., v_K) = \sum_{k=1}^{K} \frac{N_k}{N} F_k(u, v_k). \tag{4}$$

每个客户端都会上传 BEVT 中的编码器、解码器及其他 Transformer 部件。服务器聚合这些参数后，将全局参数与私有参数结合起来，以训练下一轮数据。

具体而言，每个客户端的本地更新 k 包含两个阶段：

$$\begin{aligned} v_k &\leftarrow v_k - \eta_v \nabla_v F(BevT; \mathcal{B}), \\ u_k &\leftarrow u_k - \eta_u \nabla_u F(BevT; \mathcal{B}), \end{aligned} \tag{5}$$

其中，η_v 和 η_u 分别是用于 v_k 和 u_k 的学习率，[2]；$\mathcal{B}$ 是本地数据集 $\mathcal{D}_k$ 的一个批次，批量大小为 $\mathbb{B}$。本地训练轮数记为 E，每个训练轮次由索引 e 表示。共享参数 u 通过聚合进行全局更新，具体如下：

$$u \leftarrow \sum_{k=1}^{K} \frac{N_k}{N} u_k. \tag{6}$$

如果考虑从所有客户端中仅选择一个子集 $\mathcal{S}^t$（其中 $M = |\mathcal{S}^t|$）参与聚合，则公式6可改写为：

$$u \leftarrow \sum_{k \in \mathcal{S}^t} \frac{N_k}{N} u_k. \tag{7}$$

客户端 k 的损失函数表示为：$F(u, v_k) = \mathbb{E}_{\mathcal{B} \subseteq \mathcal{D}_k} F(u, v_k; \mathcal{B})$。FedCaP 的伪代码见算法1。

C. 自适应多摄像头掩码

对于每个客户端，多视角数据点可能来自不同数量的摄像头。为促进这些客户端之间的联邦学习，我们引入自适应多摄像头掩码（AMCM），根据综合视场（FoV）在鸟瞰图查询上自适应地叠加掩码。本质上，所有客户端都以统一的尺寸和空间维度初始化鸟瞰图查询。

图2以两个客户端为例展示了这一过程：第一个使用四个摄像头，第二个仅使用两个摄像头，即前视图和后视图。根据适用于每种客户端类型的总视场，对鸟瞰图查询进行掩码处理，从而确保仅在视场区域内激活。每个摄像头与鸟瞰图查询之间的对应关系，使每个客户端的数据点都投影到尺寸一致的鸟瞰图查询上。这有助于在联邦学习中更有效地进行聚合。

[2]为统一超参数，我们将这些学习率设为相等，即 $\eta = \eta_v = \eta_u$。

算法1 FedCaP：基于摄像头注意力个性化的联邦学习**服务器输入：** 通信轮次数量 T**客户端输入：** 学习率 η_v, η_u**客户端输入：** 本地训练轮数 E_k**输出：** $BevT$，包含公共参数 u 和私有参数v_k 对每个客户端 1：服务器向所有客户端初始化 u^0 2：**对于**通信轮次 $t = 1, 2, ..., T$**执行**3：$\mathcal{S}^t \leftarrow ClientSelection(\mathcal{S})$ ▷ 联邦学习编排中的可选客户端选择操作 4：**对于**每个客户端 $k \in \mathcal{S}^t$ **并行执行**5：**如果**客户端 k 首次被选中**则**6：客户端 k 根据公式2将摄像头标定参数编码到 v_k 中 7：**结束如果**_8：$\Delta u_k \leftarrow ClientUpdate(u)$ _9：$u_k \leftarrow u + \Delta u_k$10：**结束循环** 11：$u \leftarrow Aggregate(\{u_k | \forall k \in \mathcal{S}^t\})$ ▷ 在服务器端进行安全聚合

12: **结束循环**

13: 14: 客户端更新(u**) 15:** $u_k \leftarrow u$**16:** $BevT \leftarrow Rebuild$**(**u_k**,** v_k**) 17: 对于训练轮次** $e = 1$**, 2, ...,** E**，执行 18: 对于每个批次** $\mathcal{B} \subseteq \mathcal{D}_k$**，执行 19:** $v_k \leftarrow v_k - \eta_v \nabla_v F(BevT; \mathcal{B})$ ▷ **在本地更新私有模型权重 20:** $u_k \leftarrow u_k - \eta_u \nabla_u F(BevT; \mathcal{B})$ ▷ **在本地更新公共模型权重 21: 结束循环 22: 结束循环 23:** $\Delta u_k \leftarrow u_k - u$_**24:** $\Delta u_k \leftarrow Compression(\Delta u_k)$ ▷ **联邦学习编排中的可选压缩操作**_**25: 返回** Δu_k

D. 系统框架

我们的联邦学习框架的系统概览如图 3 所示。该框架将公共交通和私有数据资源都视为客户端。由于网联自动驾驶车辆的无线网络连接可能不稳定，我们会在联邦学习过程中选择客户端，以避免掉队者效应。另一方面，私有数据资源中的客户端通常是联邦学习结果的使用者；由于网络质量较高，它们的本地训练模型会在所有通信轮次中参与聚合。

为减少每轮通信的数据量，客户端可在通信前采用稀疏化和量化等压缩技术 [54], [55]。为增强模型在网络中的暴露度，我们在联邦学习中采用了结合秘密共享的安全聚合协议 [56]。

我们的个性化联邦学习方法要求客户端仅共享模型的一部分，同时将

位置嵌入（摄像头和图像嵌入）保留在本地。安全聚合后，每个客户端通过拼接全局和本地模型分区来重新构建定制化 BEVT。随后，在本地数据集上训练拼接后的模型，并再次将其解耦，以便进一步进行模型聚合。

IV. 数据集

我们的联邦学习方法需要一个多样化的数据集，其中包含配备不同传感器安装位置和数量的各类车辆，并在不同城市中配有对应的鸟瞰视角真值。然而，由于目前没有符合我们要求的真实世界数据集，我们转而使用高保真模拟器 CARLA [57] 和全栈自动驾驶仿真框架 OpenCDA [58]来收集所需数据。为了模拟不同的摄像头传感器安装位置，我们采用三种不同类型的数据采集车辆：紧凑型轿车、皮卡和公交车。每辆数据采集车辆都会在天气条件一致的不同城市中行驶，并配备四个摄像头，以提供 360 度环视视野，与 [3]类似。我们采用后处理技术控制每帧中的摄像头数量，确保收集的数据符合质量标准。不过，这些摄像头的安装姿态因车辆型号不同而存在显著差异。汽车、公交车和卡车数据集分别包含 8352、1796 和 1800 帧，其中分别涵盖 52、14 和 9 个不同场景。每个场景都包含多种交通状况和道路类型，并遵循与 OPV2V [59]，类似的数据采集流程，以丰富场景复杂度。

表 I 展示了在多种车辆客户端（包括汽车、公交车和卡车）中用于数据采集的传感器配置参数。图 4 定性展示了来自公交车、卡车和汽车客户端的多摄像头输入数据之间的差异。

表 I 用于汽车、公交车和卡车客户端数据采集的传感器配置参数。

车辆类型	摄像头	高度（米）	横滚角（度）	俯仰角（度）	偏航角（度）
Car	前部	1.8	0	0	0
	Left	1.8	0	0	100
	右	1.8	0	0	-100
	Rear	1.8	0	0	180
Bus	前部	3.2	0	-5	0
	Left	3.2	0	-5	100
	右	3.2	0	-5	-100
	Rear	3.2	0	-5	180
卡车	前部	4.8	0	-5	0
	Left	4.8	0	-5	100
	右	4.8	0	-5	-100
	Rear	4.8	0	-5	-80

V. 实验

A. 实验设置

用例。在道路交通系统场景中，联邦学习客户端通常包括工业企业（例如汽车原始设备制造商）以及公共交通中的单个网联车辆。一般而言，这些网联车辆存储的数据集规模相对较小，但车辆数量庞大。相比之下，汽车企业的数据集通常大得多，尽管其数量远少于车辆。我们在实验设计中考虑了这两个因素。

同时，每个客户端的数据均来自各自的传感器系统，因此客户端之间存在数据异质性。我们的实验设计通过在不同类型的车辆（汽车、公交车和卡车）中采用不同的传感器系统配置，将这种数据方差来源纳入考量。实验中使用的具体传感器系统配置列于表 I。

为了全面评估各方法在道路交通系统中潜在联邦学习应用场景下的性能，我们定义了以下四种用例（UC）：

- 用例1考虑两家工业企业共同训练 BEVT。每家企业的本地数据集均采集自卡车和公交车。为进一步扩大训练数据集的规模，联邦学习中采用了一个开放数据集，并将其视为虚拟客户端。该虚拟客户端的数据集存储在服务器上，并作为第三个数据孤岛。
- 用例2涉及四家工业企业。其目标是利用其他客户端的数据集升级各自的模型，但不直接访问其他客户端的原始数据。
- 用例3展示了在车联网中使用采集到的公共交通数据进行训练的一种典型方式，这通常涉及较多客户端（联邦学习中共24个客户端）。每个客户端都拥有少量数据集，这些数据集采集自一两个特定驾驶场景，例如晴天在道路交通拥挤的城市道路上行驶。
- 用例4针对摄像头数量不同的客户端开展联邦学习，这种情况可能出现在制造商使用不同传感器系统采集数据集时。

我们在表 II 中总结了四种 UC 的动机和描述，并在图 5 中展示了客户端之间的数据分布。

BEVT架构。我们首先将输入图像$X_{i,\ k} \in \mathbb{R}^{L_k \times H \times W \times 3}$ 输入经过3层ResNet34编码器。为确保输入之间的一致性，我们使用AMCM将所有输入调整为 $L'_k = 4$。随后，图像特征在不同空间分辨率下进行编码，得到形状为$\mathbb{R}^{L'_k \times 64 \times 64 \times 128}$、$\mathbb{R}^{L'_k \times 32 \times 32 \times 256}$、 $\mathbb{R}^{L'_k \times 16 \times 16 \times 512}$的张量。接下来，我们在 $\mathbb{R}^{128 \times 128 \times 128}$ 中的BEV嵌入（查询）与编码后的图像特征（键和值）之间执行基于FAX交叉注意力的Transformer操作。此步骤会在 $\mathbb{R}^{32 \times 32 \times 128}$中生成鸟瞰图特征。为了将这些鸟瞰图特征转换为 $\{0,1\}^{256 \times 256}$中的最终鸟瞰图结果，我们使用一个带有3层双线性上采样模块的解码器。



图 4. 可视化 FedBEVT 数据集中汽车、公交车和卡车客户端的四视角相机数据点及鸟瞰视图真值（GT）。

表 II 四种用例的动机和描述概述。

ID	动机	描述
UC 1	企业利用其他工业领域的数据改进模型训练 公司和开放数据集。	2 个客户端分别拥有来自公交车和卡车的数据； 另有 1 个虚拟客户端拥有来自汽车的数据。
UC 2	各公司利用彼此的数据改进模型训练。	2 个客户端分别拥有来自公交车和卡车的数据；2 个客户端拥有来自汽车的数据。
UC 3	一个模型使用来自多辆网联车辆的数据进行训练 仅包含在一两个特定场景中收集的数据。	3 个客户端拥有来自公交车的数据 ，4 个客户端拥有来自卡车的数据 ，以及 17 个客户端拥有来自汽车的数据。
UC 4	各公司利用彼此的数据改进模型训练， 每个数据点的传感器数量不同。	3 个客户端的数据来自汽车，这些汽车上的摄像头数量分别为 这些客户端的汽车分别配有 1、3、4 个摄像头。

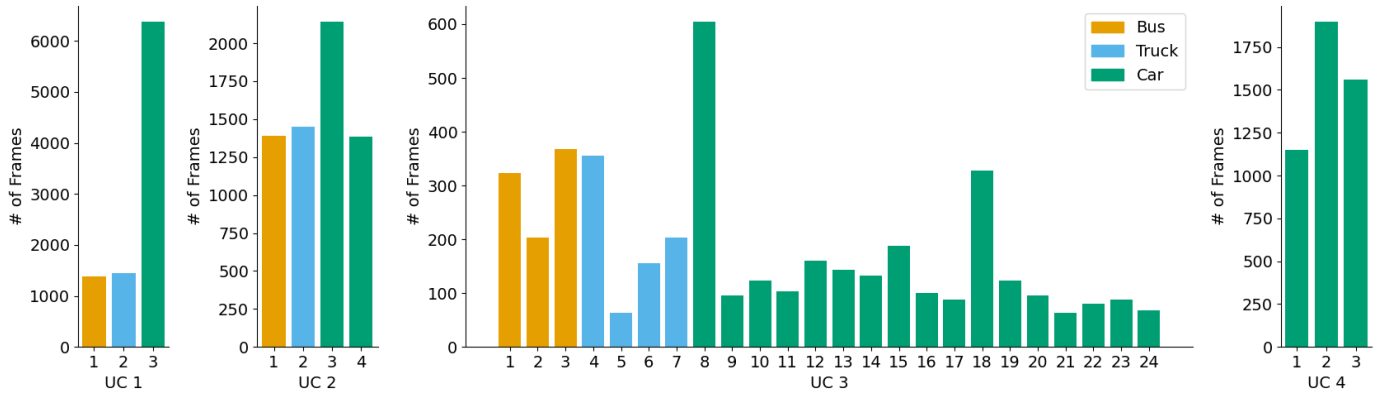


图 5. 用于评估 UC 1-4 中 FedBEVT 的联邦数据集分布。每种车辆类型的具体摄像头配置记录于表 I 中。这种配置多样性导致了数据异质性。

超参数。我们的评估在一个集群上进行，该计算机集群配备 4× 块 NVIDIA-A100-PCIE-40GB GPU和 4× 32核AMD EPYC 7513 CPU。该环境-

运行环境为 Linux 系统，使用 PyTorch 1.8.1 和 CUDA 11.1。在训练过程中，我们将初始学习率设为 2e−5 ，并在包含 20 个通信

轮次的预热阶段保持学习率不变。之后，我们采用余弦退火学习率调度器 [60]。每个客户端在本地训练一个训练轮次，批量大小为 4，并使用 AdamW 优化器 [61]。

基线方法。 鉴于目前关于 BEVT 联邦学习的研究有限，我们首次尝试在该平台上训练 FedAvg。为验证 FedCaP 的有效性，我们还将近期关于联邦 Transformer 学习的研究成果作为 BEVT 上的基线方法。除了展示各客户端本地训练的结果外，我们还将 FedCaP 与以下基线方法进行比较：

- FedAvg [36] 是联邦学习的原始算法，旨在为所有客户端训练一个共同的全局模型。
- FedRep [45] 在客户端之间共享数据表示，并为每个客户端学习独有的本地头部。请注意，我们只允许客户端将其图像编码器层个性化为本地头部。
- FedTP [62] 为每个客户端使用个性化注意力，同时在客户端之间聚合其他参数。

B. 性能

为比较 FedCaP 与其他基线方法的性能，我们比较了在四个 UC 中训练得到的模型所达到的平均交并比（IoU）。

UC 1. 如表 III 所示，FedBEVT 的交并比相比本地训练提升了 50% 以上，这显然得益于其间接利用了其他客户端的本地训练数据。此外，FedCaP 的整体性能优于基础方法 FedAvg 和另外两种个性化联邦学习方法。尽管 FedCaP 在公交车客户端上训练出的模型略优于 FedCaP，但所需的通信轮次显著更多。

UC 2. 表 IV 展示了 UC 2 的结果。在该场景中，各客户端的数据量更加均衡，但结果仍与 UC 1 类似。在汽车客户端 A 上，FedRep 的训练结果与 FedCaP 相当；而在其他客户端上，FedCaP 的表现更优。

UC 3.我们针对仅有一个或两个场景数据点的 24 个客户端，对 FedCaP 和 FedAvg 的性能进行了比较。由于 UC 3 面向网络环境较差的场景，我们将每种方法的训练限制为 100 个通信轮次。图 6 显示，超过 80%（24 个客户端中的 20 个）通过 FedCaP 获得了性能更优的个性化模型。

用例4。 尽管 AMCM 允许摄像头数量不同的客户端共同训练模型，并丰富数据资源，但也可能导致客户端之间的数据异质性。因此，如果训练数据充足，AMCM 可能会使训练结果变差。如表 V 所示，拥有单摄像头数据的客户端在不使用 AMCM 时能够取得更好的结果，因为另外两个客户端都拥有前置摄像头，可以仅基于这类数据训练模型。然而，在为拥有三摄像头的客户端训练模型时，AMCM 可以让更多客户端参与联邦学习，并提升训练性能。使用 FedCaP 后，性能还可以进一步

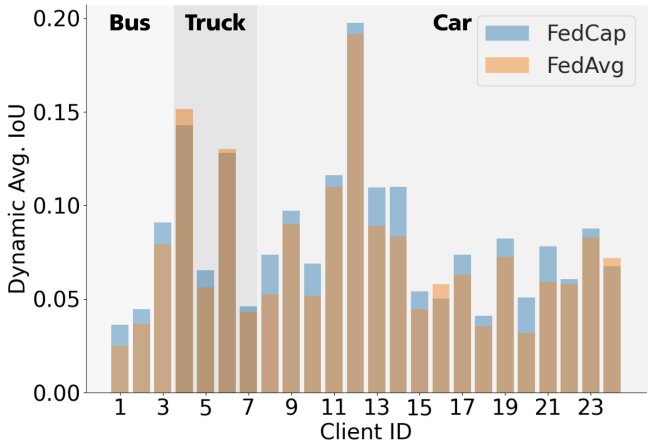


图 6. UC 3 中 FedCaP 与 FedAvg 的比较，其中联邦学习由包含公交车、卡车和汽车的 24 个客户端组成。每个客户端的本地数据仅来自一个或两个场景。

提升，因为 FedCaP 能够减轻 AMCM 的影响。对于拥有四摄像头的客户端而言，由于其数据与单摄像头客户端的数据差异显著，数据异质性成为更突出的问题。尽管如此，FedCaP 仍能缓解这些影响，并在三种方法中取得最佳模型。

C. 有效性

为验证个性化 FedBEVT 在解决客户端本地数据集之间差异方面的有效性，我们以 UC 1 为代表性示例开展消融实验，为不同客户端训练模型。

我们首先仅使用本地数据训练模型，并在不同客户端上对其进行测试。具体而言，我们使用包含 OPV2V 数据的虚拟车辆客户端训练模型。尽管该模型在汽车客户端数据上达到了 30.48% 的交并比，但其在公交车和卡车测试集上的性能显著下降（分别为 7.39% 和 2.01%），甚至不如使用本地卡车数据训练的模型在卡车测试集上的表现。这促使我们采用 FedBEVT。

随后，我们采用不同的个性化 FedBEVT，即 FedTP、FedRep 和 FedCaP，为不同客户端训练模型，并在各自的测试集以及其他客户端的测试集上进行评估。表 VI、表 VII 和表 VIII 所示的结果一致表明，只有本地个性化模型最适合本地数据。由于汽车的摄像头位姿和数量通常在生产后不会发生变化，因此本地收集的训练数据集与未来用于推理的数据具有很强的相似性。这与我们的实验设计相一致。因此，这些结果进一步凸显了个性化 FedBEVT 对于鸟瞰图道路交通感知的重要性。

D. 视觉分析

在图7中，我们直观地展示了，与其他方法相比，FedCaP在三种车辆类型（即公交车、卡车和汽车）的测试数据集上生成的鸟瞰图地图更加准确且完整。我们在第一列展示了每个多视角摄像头系统的前置摄像头视图。显然，不同客户端的摄像头系统在高度上存在显著差异

表 III UC 1 中 FedCaP 与其他基线方法的比较。

方法	公交车客户端 $N_1 = 1388$			卡车客户端 $N_2 = 1448$			虚拟车辆客户端[1] $N_3 = 6372$		
	训练损失	IoU↑	通信↓	训练损失	IoU↑	通信↓	训练损失	IoU↑	通信↓
本地训练	0.0716	5.42%	-	0.1740	4.16%	-	-	-	-
FedAvg [36]	0.0343	16.29%	315	0.0448	11.10%	290	0.0096	34.86%	410
FedRep [45]	0.0176	**19.40%**	350	0.0303	14.20%	325	0.0073	34.37%	380
FedTP [62]	0.0281	10.72%	195	0.0457	8.82%	260	0.0122	33.50%	390
FedCaP（我们的方法）	0.0528	**19.32%**	**240**	0.04332	**15.01%**	295	0.0223	**35.44%**	350

[1] 我们将虚拟车辆客户端设置为使用 OPV2V [59] 数据集。

表 IV UC 2 中 FedCaP 与其他基线方法的比较。

方法	公交车客户端 $N_1 = 1388$			卡车客户端 $N_2 = 1448$			汽车客户端 A $N_3 = 2140$			汽车客户端 B $N_4 = 1384$		
	训练损失	IoU↑	通信↓	训练损失	IoU↑	通信↓	训练损失	IoU↑	通信↓	训练损失	IoU↑	通信↓
本地训练	0.0716	5.42%	-	0.1740	4.16%	-	0.0288	13.15%	-	0.0298	9.41%	-
FedAvg [36]	0.2202	6.92%	105	0.0483	5.89%	180	0.0518	14.09%	175	0.0899	14.61%	175
FedRep [45]	0.0170	7.45%	320	0.1319	7.03%	105	0.0352	**18.94%**	345	0.0394	15.41%	330
FedTP [62]	0.0765	6.73%	175	0.1015	5.42%	140	0.0183	17.29%	335	0.0389	15.56%	300
FedCaP（我们的方法）	0.0061	**10.32%**	340	0.1294	**7.33%**	65	0.0162	**18.40%**	360	0.0268	**16.16%**	370

表 V 用例4中自适应多摄像头掩码（AMCM）有效性评估。

方法	单摄像头 $N_1 = 1152$ [1][1]	三摄像头 $N_2 = 1896$ [1,2,3][1]	四摄像头 $N_3 = 1560$ [1,2,3,4][1]
FedAvg[2]	**34.14%**	15.50%	17.87%
FedAvg + AMCM	30.17%	19.20%	16.95%
FedCaP + AMCM	31.89%	**20.89%**	**20.59%**

[1] 括号内的数字表示感知系统中摄像头的索引。具体而言，前置摄像头记为 1，左侧摄像头记为 2，右侧摄像头记为 3，后置摄像头记为 4。[2] 不使用 AMCM 的 FedAvg：每个客户端中配置相同的摄像头用于联邦学习。例如，要为单摄像头客户端训练模型，会使用其他客户端的所有前置摄像头数据进行联邦学习。不使用 AMCM 的 FedAvg 有一个显著局限：它只能为采用特定摄像头系统的客户端训练模型。

表 VI 使用 FedTP 训练的个性化模型在各客户端上的鸟瞰视图检测平均交并比。

本地测试集 \ 个性化模型	公交车客户端	卡车客户端	汽车客户端
公交车客户端	**10.72%**	4.61%	9.12%
卡车客户端	6.92%	**8.82%**	2.10%
汽车客户端	12.45%	6.76%	**33.50%**

表 VII 使用 FedRep 训练的个性化模型在各客户端上的鸟瞰视图检测平均交并比。

本地测试集 \ 个性化模型	公交车客户端	卡车客户端	汽车客户端
公交车客户端	**19.40%**	17.73%	9.39%
卡车客户端	9.59%	**14.20%**	4.48%
汽车客户端	26.04%	20.39%	**34.37%**

表 VIII 使用 FedCaP 训练的个性化模型在各客户端上的鸟瞰视图检测平均交并比。

本地测试集 \ 个性化模型	公交车客户端	卡车客户端	汽车客户端
公交车客户端	**19.32%**	17.48%	13.71%
卡车客户端	13.01%	**15.01%**	8.89
汽车客户端	26.94%	25.26%	**35.44%**

。尽管较高的摄像头可能提供更好的BEV视图，但也会减少每个目标所占据的像素数量，从而使目标检测更加困难。然而，BEV视图的真实标注保持不变——见第二列——因为z轴上的高度信息已不再相关。

FedRep。尽管 FedRep 的目标识别能力相对较强，

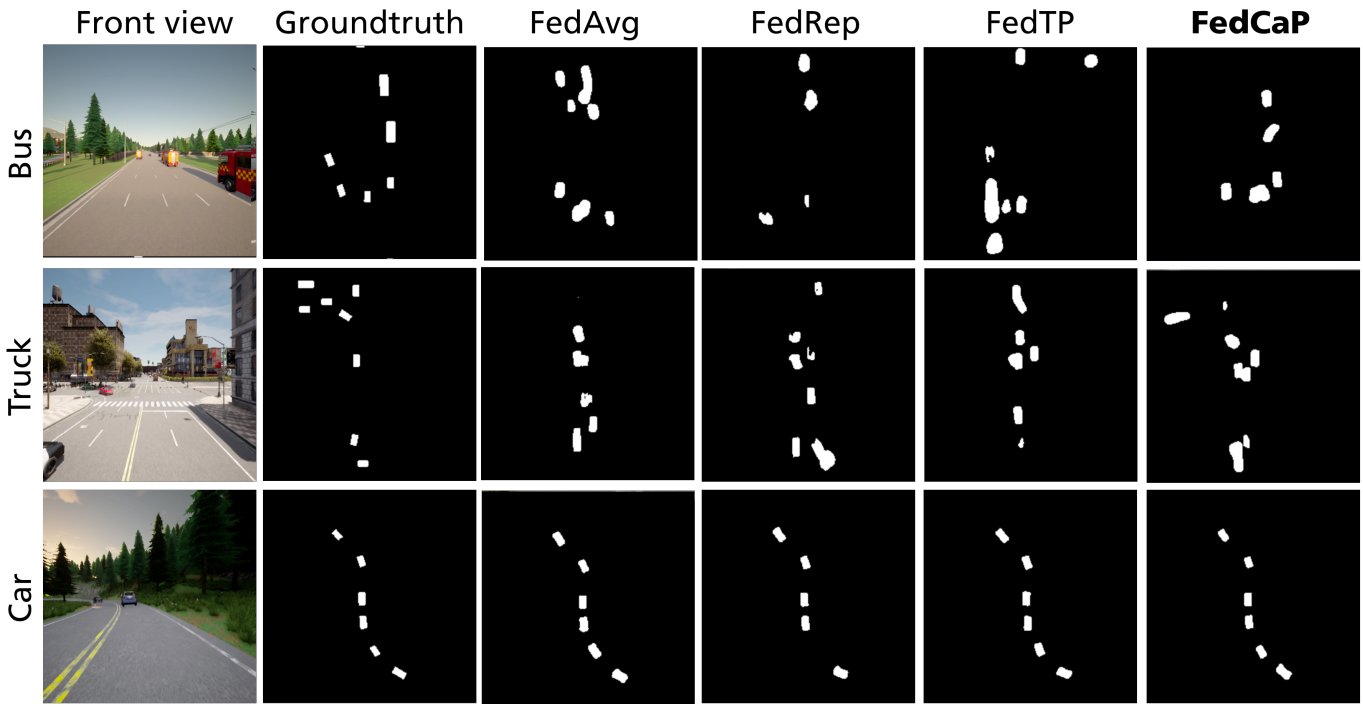


图 7. 不同联邦学习方法的鸟瞰图结果对比。

但仍可能漏检一些目标车辆。它在一定程度上训练编码器的初始层，以适应图像像素的差异。然而，FedRep 并未明确考虑传感器视角的差异。当由不同传感器高度造成的数据异质性产生影响时，目标识别能力便会下降。

FedTP。我们观察到，FedTP 训练的模型更容易将路边树木和建筑物识别为目标车辆。FedTP 直接解耦跨注意力的参数，以考虑数据差异，但模型重新耦合时，这可能导致注意力和原始注意力头的优化方向出现偏差。

FedCaP。通过对与外参和内参相关的嵌入进行隐私保护，FedCaP 直接考虑了客户端之间摄像头系统配置的差异。该方法取得了颇具前景的整体准确率和目标识别能力。如图 7 所示，在公交车数据中，与其他方法相比，只有该方法能够识别所有目标。在卡车数据中，只有该方法能够识别十字路口左侧的目标。在汽车数据中，该方法能够识别所有目标，并准确估计其尺寸。

VI. 结论

本研究探讨了联邦学习在训练用于道路交通鸟瞰图感知的基于Transformer的模型时的效率。我们的分析发现了两个可能妨碍联邦学习方法性能的数据异质性问题。为应对这些挑战，我们提出了两种新颖技术，即 FedCaP 和 AMCM。我们通过收集一个新数据集并将其分发给客户端来评估所提方法的有效性，从而构建典型用例在联邦设置中。实验结果表明，通过个性化位置嵌入并增加可用于训练的数据资源，所提方法显著提升了联邦学习的整体性能。总之，本研究彰显了联邦学习应用于鸟瞰图感知模型的潜力，并提出了有效解决数据异质性挑战的方法。

VII. 致谢

本项目广泛使用了 OpenCDA 生态系统中的工具链 [63]，包括 OpenCOOD [59] 和 OpenCDA 仿真工具 [58]。

参考文献

[1] X. Xia 等，“自动驾驶系统数据采集与分析平台”，《交通研究*C*辑：新兴技术》，第151卷，第104–120页，2023年，DOI：10.1016/j.trc.2023.104120。[2] Z. Meng、X. Xia、R. Xu、W. Liu 和 J. Ma，“HYDRO-3D：利用三维激光雷达进行协同感知的混合目标检测与跟踪”，*IEEE* 智能车辆汇刊，2023年，DOI：10.1109/TIV.2023.3282567。[3]R. Xu 等，“CoBEVT：利用稀疏Transformer进行协同鸟瞰图语义分割”，载于机器人学习会议（*CoRL*）。PMLR，2022年12月。 [在线]。可访问：https: //proceedings.mlr.press/v205/xu23a/xu23a.pdf [4] Y. Ma 等，“以视觉为中心的鸟瞰图感知：综述”，*arXiv*预印本 *arXiv:2208.02797*，2022年。[5] W. Liu 等，“网联自动驾驶汽车中的控制技术及其应用：系统综述”，*IEEE*《物联网期刊》，2023年。[6] Z. Liu 等，“BEVFusion：基于统一鸟瞰图表示的多任务多传感器融合”，*arXiv*预印本 *arXiv:2205.13542*，2022年。[7] W. Liu、X. Xia、L. Xiong、Y. Lu、L. Gao 和 Z. Yu，“考虑信号测量特性的自动驾驶车辆侧滑角估计”，《*IEEE*传感器期刊》，第21卷，第19期，第21 675–21687页，2021年，DOI：10.1109/JSEN.2021.3059050。

[8]X. Xia、N. P. Bhatt、A. Khajepour 和 E. Hashemi，“适用于不同环境的集成式基于惯性传感器和激光雷达的地图匹配定位”，IEEE 智能车辆汇刊，2023年，DOI：10.1109/TIV.2023.3298892。[9] Z. Li 等，“BEVformer：通过时空 Transformer 从多摄像头图像中学习鸟瞰图表示”，载于《计算机视觉—ECCV 2022》。沙姆：Springer，2022年10月，第1–18页，DOI：10.1007/978-3-031-20077-9 1。[10] J. Huang、G. Huang、Z. Zhu、Y. Yun 和 D. Du，“BEVDet：鸟瞰图中的高性能多摄像头三维目标检测”，arXiv预印本 arXiv:2112.11790，2021年。[11] B. Zhou 和 P. Kr¨ahenb¨uhl，“用于实时地图视图语义分割的跨视角 Transformer”，载于2022年 IEEE/CVF 计算机视觉与模式识别会议（CVPR），2022年，第13 750–13 759页，DOI：10.1109/CVPR52688.2022.01339。[12] Y. Liu、T. Wang、X. Zhang 和 J. Sun，“PETR：用于多视角三维目标检测的位置嵌入变换”，载于《计算机视觉—ECCV 2022》。Springer，2022年，第531–548页，DOI：10.1007/978-3-031-19812-0 31。[13] A. Swerdlow、R. Xu 和 B. Zhou，“从鸟瞰视图布局生成街景图像”，arXiv预印本 arXiv:2301.04634，2023年。[14] A. Hu 等，“FIERY：利用环视单目摄像头从鸟瞰图中进行未来实例分割”，载于 IEEE/CVF 国际计算机视觉会议（ICCV），2021年，URL：10.1109/ICCV48922.2021.01499。[15] Y. Liu 等，“PETRv2：从多摄像头图像进行三维感知的统一框架”，arXiv预印本 arXiv:2206.01256，2022年。[16] Y. Zhang 等，“BEVerse：面向以视觉为中心的自动驾驶的鸟瞰图统一感知与预测”，arXiv预印本 arXiv:2205.09743，2022年。[17] H. Caesar 等，“nuScenes：用于自动驾驶的多模态数据集”，载于 IEEE/CVF计算机视觉与模式识别会议论文集，2020年，第11 621–11 631页，DOI：10.1109/CVPR42600.2020.01164。[18] J. Philion 和 S. Fidler，“Lift, splat, shoot：通过隐式反投影到三维空间对来自任意相机组的图像进行编码”，载于《计算机视觉—ECCV 2020》。Springer，2020年，第194–210页，DOI：10.1007/978-3-030-58568-6 12。[19] H. Zhang、J. Bosch 和 H. H. Olsson，“实时端到端联邦学习：汽车领域案例研究”，载于2021年 IEEE 第45届计算机、软件与应用会议（COMPSAC）。IEEE，2021年，第459–468页，DOI：10.1109/COMPSAC51774.2021.00070。[20] A. M. Elbir、B. Soner、S. Ç¨oleri、D. G¨und¨uz 和 M. Bennis，“车联网中的联邦学习”，载于2022年 IEEE 地中海通信与网络国际会议（MeditCom）。IEEE，2022年，第72–77页，DOI：10.1109/MeditCom55741.2022.9928621。[21] A. Nguyen 等，“用于自动驾驶的深度联邦学习”，载于2022年 IEEE 智能车辆研讨会（IV），2022年，第1824–1830页，DOI：10.1109/IV51971.2022.9827020。[22] Y. Li、X. Tao、X. Zhang、J. Liu 和 J. Xu，“面向自动驾驶的隐私保护联邦学习”，IEEE 智能交通系统汇刊，第23卷，第7期，第8423–8434页，2022年，DOI：10.1109/TITS.2021.3081560。[23] D. Jallepalli、N. C. Ravikumar、P. V. Badarinath、S. Uchil 和 M. A. Suresh，“自动驾驶车辆目标检测的联邦学习”，载于2021年 IEEE 第七届大数据计算服务与应用国际会议（BigDataService），2021年，第107–114页，DOI：10.1109/BigDataService52369.2021.00018。[24] K. Xie 等，“利用脉冲神经网络进行高效联邦学习以实现交通标志识别”，IEEE 车辆技术汇刊，第71卷，第9期，第9980–9992页，2022年，DOI：10.1109/TVT.2022.3178808。[25] X. Kong、K. Wang、M. Hou、X. Hao、G. Shen、X. Chen 和 F. Xia，“面向5G车联网的基于联邦学习的车牌识别方案”，IEEE 工业信息学汇刊，第17卷，第12期，第8523–8530页，2021年，DOI：10.1109/TII.2021.3067324。[26] M. P. Aparna、R. Gandhiraj 和 M. Panda，“利用联邦学习预测自动驾驶的转向角：车联网通信的影响”，载于2021年第12届计算、通信与网络技术国际会议（ICCCNT），2021年，第1–7页，DOI：10.1109/ICCCNT51525.2021.9580097。[27] X. Kong、H. Gao、G. Shen、G. Duan 和 S. K. Das，“FedVCP：面向社交车联网的基于联邦学习的协作定位方案”，IEEE 计算社会系统汇刊，第9卷，第1期，第197–206页，2022年，DOI：10.1109/TCSS.2021.3062053。[28] H. Zhang 等，“利用联邦学习辅助的可重构全息表面进行自动驾驶协同无线 SLAM”

1–17，2023年智能车辆汇刊，DOI：10.1109/TIV.2023.3285592。[29] S. R. Pokhrel 和 J. Choi，“基于区块链的自动驾驶车辆联邦学习：分析与设计挑战”，IEEE《通信汇刊》，第68卷，第8期，第4734–4746页，2020年，DOI：10.1109/TCOMM.2020.2990686。[30] C. Zhu、X. Zhu、J. Ren 和 T. Qin，“面向无人机边缘计算网络的区块链赋能联邦学习：问题与解决方案”，IEEE Access，第10卷，第56 591–56 610页，2022年，DOI：10.1109/AC-CESS.2022.3174865。[31] Y. He、K. Huang、G. Zhang、F. R. Yu、J. Chen 和 J. Li，“Bift：面向互联与自动驾驶车辆的基于区块链的联邦学习系统”，IEEE《物联网期刊》，第9卷，第14期，第12 311–12 322页，2022年，DOI：10.1109/JIOT.2021.3135342。[32] W. Zhang 等，“面向工业物联网设备故障检测的基于区块链的联邦学习”，IEEE《物联网期刊》，第8卷，第7期，第5926–5937页，2020年，DOI：10.1109/JIOT.2020.3032544。[33] L. Fantauzzo 等，“FedDrive：将联邦学习推广至自动驾驶中的语义分割”，载于2022年IEEE/RSJ智能机器人与系统国际会议（IROS），2022年，第11 504–11 511页，DOI：10.1109/IROS47612.2022.9981098。[34] R. Song、L. Zhou、V. Lakshminarasimhan、A. Festag 和 A. Knoll，“应对协同智能交通系统层级异质性的联邦学习框架”，载于2022年IEEE第25届智能交通系统国际会议（ITSC），2022年，第3502–3508页，DOI：10.1109/ITSC55140.2022.9922064。[35] Z. Du 等，“车载物联网联邦学习：最新进展与开放问题”，IEEE《计算机学会开放期刊》，第1卷，第45–61页，2020年，DOI：10.1109/OJCS.2020.2992630。[36] B. McMahan、E. Moore、D. Ramage、S. Hampson 和 B. A. y Arcas，“去中心化数据上的深度网络通信高效学习”，载于人工智能与统计学。PMLR，2017年，第1273–1282页。 [在线]。可访问：https://proceedings.mlr.press/v205/xu23a/xu23a.pdf [37] M. G. Arivazhagan、V. Aggarwal、A. K. Singh 和 S. Choud-hary，“带个性化层的联邦学习”，arXiv预印本 arXiv:1912.00818，2019年。[38] F. Hanzely、B. Zhao 和 M. Kolar，“个性化联邦学习：统一框架与通用优化技术”，arXiv预印本 arXiv:2102.09743，2021年。[39] K. Pillutla 等，“带部分模型个性化的联邦学习”，载于国际机器学习会议，第162卷。PMLR，2022年7月，第17 716–17 758页。 [在线]。可访问：https://proceedings.mlr.press/v162/pillutla22a/pillutla22a.pdf [40] A. Z. Tan、H. Yu、L. Cui 和 Q. Yang，“迈向个性化联邦学习”，IEEE《神经网络与学习系统汇刊》，2022年，DOI：10.1109/TNNLS.2022.3160699。[41] J. Zhang、S. Guo、X. Ma、H. Wang、W. Xu 和 F. Wu，“面向个性化联邦学习的参数化知识迁移”，神经信息处理系统进展（NIPS 2021），第34卷，第10 092–10 104页，2021年。 [在线]。可访问：https://proceedings.neurips.cc/paper/2021/hash/5383c7318a3158b9bc261d0b6996f7c2-Abstract.html [42] A. Shamsian、A. Navon、E. Fetaya 和 G. Chechik，“使用超网络的个性化联邦学习”，载于国际机器学习会议。PMLR，2021年，第9489–9502页。 [在线]。可访问：http://proceedings.mlr.press/v139/shamsian21a/shamsian21a.pdf [43] P. P. Liang、T. Liu、L. Ziyin、R. Salakhutdinov 和 L.-P. Morency，“立足局部、着眼全局：具有局部与全局表征的联邦学习”，NeurIPS联邦学习研讨会，2019年9月。 [在线]。可访问：https://proceedings.neurips.cc/paper files/paper/2019/file/3a0844cee4fcf57de0c71e9ad3035478-Paper.pdf – [44] L. Yi、G. Wang、X. Liu、Z. Shi 和 H. Yu，“Fedgh：具有广义全局头部的异质联邦学习”，arXiv预印本 arXiv:2303.13137，2023年。[45] L. Collins、H. Hassani、A. Mokhtari 和 S. Shakkottai，“利用共享表征实现个性化联邦学习”，载于国际机器学习会议。PMLR，2021年7月，第2089–2099页。[在线]。可访问：https://proceedings.mlr.press/v139/collins21a.html [46] Y. Huang、L. Chu、Z. Zhou、L. Wang、J. Liu、J. Pei 和 Y. Zhang，“非独立同分布数据上的个性化跨机构联邦学习”，载于AAAI人工智能会议论文集，第35卷，第9期，2021年，第7865–7873页，DOI：10.1609/aaai.v35i9.16960。[47] D. Bui 等，“联邦用户表征学习”，arXiv预印本 arXiv:1909.12535，2019年。[48] A. Vaswani 等，“注意力就是你所需要的一切”，神经信息处理系统进展（NIPS 2017），第30卷，第5998–6008页，

2017年。 [在线]。可访问：https://papers.nips.cc/paper files/paper/2017/ file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf [49] A. Dosovitskiy 等，“一张图像胜过16×16个词：大规模图像识别的Transformer”，载于国际学习表征会议，2021年。 [在线]。可访问：https://openreview.net/ forum?id=YicbFdNTTy [50] R. Xu 等，“V2X-ViT：基于视觉Transformer的车对万物协同感知”，载于欧洲计算机视觉会议（ECCV），2022年，第107–124页，DOI：10.1007/978-3-031-19842-7 7。[51] ——，“V2V4Real：面向车车协同感知的真实世界大规模数据集”，arXiv预印本 arXiv:2303.07601，2023年。[52] W. Liu、K. Quijano 和 M. M. Crawford，“YOLOv5-Tassel：基于迁移学习改进YOLOv5的RGB无人机影像穗部检测”，IEEE《应用地球观测与遥感精选主题期刊》，第15卷，第8085–8094页，2022年，DOI：10.1109/JSTARS.2022.3206399。[53] L. Qu 等，“重新思考用于应对联邦学习数据异质性的架构设计”，载于IEEE/CVF计算机视觉与模式识别会议论文集，2022年，第10 061–10 071页，DOI：10.1109/CVPR52688.2022.00982。[54] Y. Lin、S. Han、H. Mao、Y. Wang 和 B. Dally，“深度梯度压缩：降低分布式训练的通信带宽”，载于国际学习表征会议，2018年。[在线]。可访问：https://openreview.net/forum?id=SkhQHMW0W [55] R. Song、L. Zhou、L. Lyu、A. Festag 和 A. Knoll，“ResFed：通过传输深度压缩残差实现通信高效的联邦学习”，arXiv预印本 arXiv:2212.05602，2022年。[56] K. Bonawitz 等，“面向隐私保护机器学习的实用安全聚合”，载于ACM SIGSAC计算机与通信安全会议，2017年，第1175–1191页，DOI：10.1145/3133956.3133982。[57]A. Dosovitskiy、G. Ros、F. Codevilla、A. Lopez 和 V. Koltun，“CARLA：开放的城市驾驶模拟器”，载于机器人学习会议。PMLR，2017年，第1–16页。 [在线]。可访问：http://proceedings.mlr.press/v78/dosovitskiy17a/dosovitskiy17a.pdf [58] R. Xu 等，“OpenCDA：集成协同仿真的开放协同驾驶自动化框架”，载于IEEE智能交通系统国际会议（ITSC），2021年，第1155–1162页，DOI：10.1109/ITSC48978.2021.9564825。[59] ——，“OPV2V：用于车车通信感知的开放基准数据集与融合流水线”，载于2022年国际机器人与自动化会议（ICRA）。IEEE，2022年，第2583–2589页，DOI：10.1109/ICRA46639.2022.9812038。[60] I. Loshchilov 和 F. Hutter，“SGDR：带热重启的随机梯度下降”，载于国际学习表征会议，2017年。 [在线]。可访问：https://openreview.net/ forum?id=Skq89Scxx[61] ——，“解耦权重衰减正则化”，载于国际学习表征会议，2019年。 [在线]。可访问：https://openreview.net/forum?id = 6 7 Bkg RiCqY [62] H. Li 等，“FedTP：通过Transformer个性化实现联邦学习”，IEEE《神经网络与学习系统汇刊》，第1 15页，2023年，10 1109 2023 3269062 – DOI：. /TNNLS. . . [63] R. Xu 等，“OpenCDA协同驾驶自动化研究开源生态系统”，IEEE 智能车辆汇刊，第8卷，第4期，第2698–2711页，2023年，DOI：10.1109/TIV.2023.3244948。

附录 A 收敛性分析

基于第 III-B 节对 FedCaP 的描述，我们首先提出以下假设，这些假设与先前关于 [37]–[39], [43] 的研究一致：

假设 1。（L-光滑性）。 k号客户端的局部模型梯度 $\nabla_u F_k(u, v_k)$，$\nabla_v F_k(u, v_k)$，$\nabla_{uv_k} F_k(u, v_k)$ 以及 $\nabla_{v_k u} F_k(u, v_k)$ 均满足 *L-Lipschitz* 条件，即：

$$\begin{aligned}
||\nabla_u F_k(u^{t+1}, v_k) - \nabla_u F_k(u^t, v_k)|| &\le L_u ||u^{t+1} - u^t||, \\
||\nabla_v F_k(u^{t+1}, v_k) - \nabla_v F_k(u^t, v_k)|| &\le L_{vu} ||u^{t+1} - u^t||, \\
||\nabla_v F_k(u, v_k^{t+1}) - \nabla_v F_k(u, v_k^t)|| &\le L_v ||v_k^{t+1} - v_k^t||, \\
||\nabla_u F_k(u, v_k^{t+1}) - \nabla_u F_k(u, v_k^t)|| &\le L_{uv} ||v_k^{t+1} - v_k^t||,
\end{aligned} \tag{8}$$

且对于某个 $\mathcal{X}$，

$$\mathcal{X} \ge \max\left(\frac{L_{uv}}{\sqrt{L_u L_v}}, \frac{L_{vu}}{\sqrt{L_u L_v}}\right). \tag{9}$$

假设 2。（有界梯度）。 每个客户端的局部随机梯度方差有界，即：

$$\begin{aligned}
\mathbb{E}_{B \subseteq \mathcal{D}_k} ||\nabla_u F(u, v_k; B) - \nabla_u F(u, v_k)||^2 &\le \sigma_u^2, \\
\mathbb{E}_{B \subseteq \mathcal{D}_k} ||\nabla_v F(u, v_k; B) - \nabla_v F(u, v_k)||^2 &\le \sigma_v^2.
\end{aligned} \tag{10}$$

假设 3。（有界噪声）。 存在 $\zeta \ge 0$ 和 $P \ge 0$，适用于所有 u 和 V，使得：

$$\begin{aligned}
&||\frac{1}{K} \sum_{k=1}^{k} \nabla_u F(u, v_k) - \nabla_u F(u, V)||^2 \\
&\qquad \le \zeta^2 + P^2 ||\nabla_u F(u, V)||^2.
\end{aligned} \tag{11}$$

在 F 不包含正则化项的情况下，P等于 *0*。

定理 1.（收敛性）。 *FedCaP* 在整个训练过程中均可收敛。

证明。假设假设 1 成立，我们可以为第 k 个客户端的变量 u 和 v_k 的更新建立如下界限：

$$\begin{aligned}
F_k(u^{t+1}, v_k^{t+1}) - F_k(u^t, v_k^{t+1}) &\le \langle \nabla_u F_k(u^t, v_k^{t+1}), (u^{t+1} - v^t) \rangle \\
&\quad + \frac{L_u}{2} ||u^{t+1} - u^t||^2 \qquad (12) \\
F_k(u^t, v_k^{t+1}) - F_k(u^t, v_k^t) &\le \langle \nabla_u F_k(u^t, v_k^t), (u^{t+1} - v^t) \rangle \\
&\quad + \frac{L_v}{2} ||v_k^{t+1} - v_k^t||^2, \qquad (13)
\end{aligned}$$

其中

$$\begin{aligned}
\langle \nabla_u F_k(u^t, v_k^{t+1}), (u^{t+1} - v^t) \rangle &\le \langle \nabla_u F_i(u^t, v_k^t), (u^{t+1} - u^t) \rangle \\
&\quad + \mathcal{X}^2 \frac{L_v}{2} ||v_k^{t+1} - v_k^t||^2 \\
&\quad + \mathcal{X}^2 \frac{L_u}{2} ||u^{t+1} - u^t||^2.
\end{aligned} \tag{14}$$

通过整合式（12）、（13）和（14）所表示的不等式，我们可以为 FedCaP 中单轮通信后的完整更新建立一个界限。可表示为：

$$\begin{aligned}
&F_k(u^{t+1}, v_k^{t+1}) - F_k(u^t, v_k^t) \le \langle \nabla_u F(u^t, V^t), (u^{t+1} - u^t) \rangle \\
&+ \frac{1}{K} \sum_{k=1}^{K} \langle \nabla_u F(u^t, v_k^t), (v_k^{t+1} - v_k^t) \rangle \\
&+ \frac{L_u}{2} (1 + \mathcal{X}^2) ||u^{t+1} - u^t||^2 \\
&+ \frac{L_v}{2} (1 + \mathcal{X}^2) ||v_k^{t+1} - v_k^t||^2.
\end{aligned} \tag{15}$$

对于不等式 15，我们可以分别考察各项的期望：

$$\begin{aligned}
&\mathbb{E}[\langle \nabla_u F(u^t, V^t), (u^{t+1} - u^t) \rangle] \le -\frac{E \eta_u}{2} ||\nabla_u F(u^t, V^t)||^2 \\
&+ \frac{\eta_u}{2} \sum_{e=0}^{E-1} \mathbb{E} ||\frac{1}{K} \sum_{k=1}^{K} \nabla_u F(u_{k,e}^t, v_{k,e}^t) - \nabla_u F(u^t, V^t)||^2 \\
&\le -\frac{E \eta_u}{2} ||\nabla_u F(u^t, V^t)||^2 \\
&+ \frac{\eta_u}{2K} \sum_{e=0}^{E-1} \sum_{k=1}^{K} \mathbb{E} ||\nabla_u F(u_{k,e}^t, v_{k,e}^t) - \nabla_u F(u^t, V^t)||^2.
\end{aligned} \tag{16}$$

$$\begin{aligned}
&\mathbb{E}[\frac{1}{K} \sum_{k=1}^{K} \langle \nabla_u F(u^t, v_k^t), (v_k^{t+1} - v_k^t) \rangle] \\
&= -\eta_v \frac{m}{K^2} \sum_{k=1}^{K} \sum_{e=0}^{E-1} \mathbb{E} (\nabla_v F(u^t, v_k^t) \nabla_v F(u_k^t - v_k^t)) \\
&\le -\frac{M \eta_v e}{2K^2} \sum_{k=1}^{K} ||\nabla_v F(u^t, v_k^t)||^2 \\
&+ \frac{M \eta_v}{2K^2} \sum_{k=1}^{K} \sum_{e=0}^{E-1} \\
&\mathbb{E} (2\mathcal{X}^2 L_u L_v ||u_{k,e}^t - u^t||^2 + 2L_v^2 ||v_{k,e}^t - v_k^t||^2).
\end{aligned} \tag{17}$$

$$\begin{aligned}
&\mathbb{E}[\frac{L_u}{2} (1 + \mathcal{X}^2) ||u^{t+1} - u^t||^2] \le \frac{L_u (1 + \mathcal{X}^2) \eta_u^2 E^2 \sigma_u^2}{2M} \\
&+ \frac{L_u (1 + \mathcal{X}^2) \eta_u^2 E}{2} \sum_{e=0}^{E-1} \mathbb{E} ||\frac{1}{M} \sum_{k \in \mathcal{S}^t} \nabla_u F(u_{k,e}^t, v_{k,e}^t)||^2 \\
&\le \frac{L_u (1 + \mathcal{X}^2) \eta_u^2 E^2}{2M} (\sigma_u^2 + 12\zeta^2 (1 - \frac{M}{K})) \\
&+ 3L_u (1 + \mathcal{X}^2) \eta_u^2 E^2 (1 + P^2) ||\nabla F(u^t, V^t)||^2 \\
&+ \frac{3L_u (1 + \mathcal{X}^2) \eta_u E^2}{2K} \sum_{k=1}^{K} \sum_{e=0}^{E-1} \\
&\mathbb{E} (L_u^2 ||u_{k,e}^t - u^t||^2 + \mathcal{X}^2 L_u L_v ||v_{k,e}^t - v_k^t||^2).
\end{aligned} \tag{18}$$

$$
\begin{aligned}
&\mathbb{E}[\frac{L_v}{2}(1+\mathcal{X}^2)||v_k^{t+1}-v_k^t||^2] \\
&\leq \frac{L_v(1+\mathcal{X}^2)M}{2K^2}\sum_{k=1}^{K}\mathbb{E}(||v_{k,E}^t-v_k^t||^2) \\
&\leq \frac{3L_v(1+\mathcal{X}^2)\eta_v^2E^2M}{2K^2}\sum_{k=1}^{K}(||\nabla_v F(u^t,v_k^t)||^2 \\
&+\sum_{e=0}^{}E-1\mathbb{E}(L_v^2||v_{k,e}^t-v_k^t||^2+\mathcal{X}^2L_uL_v||u_{k,e}^t-u^t||^2)) \\
&+\frac{L_v(1+\mathcal{X}^2)\eta_v^2E^2M\sigma_v^2}{2K}.
\end{aligned}
\tag{19}
$$

将 17、18、19 和 20 代入 16，可得 FedCaP 中一个通信轮次整体更新的如下不等式：

$$
\begin{aligned}
&\mathbb{E}[F(u^{t+1},V^{t+1})-F(u^t,V^t)] \leq -\frac{\eta_u*E}{4}||\nabla_u F(u^t,V^t)||^2 \\
&-\frac{\eta_vEM}{4K^2}\sum_{k=1}^{K}||\nabla_v F(u^t,v_k^t)||^2 \\
&+\frac{L_u(1+\mathcal{X}^2)\eta_u^2E^2}{2}(\sigma_u^2+\frac{12\zeta^2}{M}(1-M/K)) \\
&+\frac{L_v(1+\mathcal{X}^2)\eta_v^2E^2\sigma_v^2M}{2K} \\
&+\frac{2}{K}\sum_{k=1}^{K}\sum_{e=0}^{E-1}\mathbb{E}||u_{k,e}^t-u^t||^2(L_u^2\eta_u+\frac{M}{K}\mathcal{X}^2L_uL_v\eta_v) \\
&+\frac{2}{K}\sum_{k=1}^{K}\sum_{e=0}^{E-1}\mathbb{E}||v_{k,e}^t-v_k^t||^2(\frac{M}{K}L_v^2\eta_v+\mathcal{X}^2L_uL_v\eta_u).
\end{aligned}
\tag{20}
$$

接下来，我们评估不等式 20 中最后两项的界限。假设假设 2 和 3 均成立，并对各项中的常数进行化简后，可得：

$$
\begin{aligned}
&\mathbb{E}[F(u^{t+1},V^{t+1})-F(u^t,V^t)] \leq -\frac{C||\nabla_u F(u^t,V^t)||^2}{8L_u} \\
&-\frac{CM\frac{1}{K}\sum_{k=1}^{K}||\nabla_v F(u^t,v_k^t)||^2}{8L_vK} \\
&+C^2(1+\mathcal{X}^2)(\frac{\sigma_u^2}{2L_u}+\frac{M\sigma_v^2}{nL_v}+\frac{6\zeta^2}{L_vM}(1-\frac{M}{K})) \\
&+C^3(1+\mathcal{X}^2)(1-\frac{1}{E})(\frac{24\zeta^2}{L_u}+\frac{4\sigma_u^2}{L_u}+\frac{4\sigma_v^2}{L_u}),
\end{aligned}
\tag{21}
$$

学习率为 $\eta_u \leq (12EL_u(1+\mathcal{X}^2)(1+P^2))^{-1}$ 且 $\eta_v \leq (6EL_v(1+\mathcal{X}^2))^{-1}$，其中 $C>0$ 表示某个正常数。最后，FedCaP 的收敛速率可表示如下：

$$
\begin{aligned}
&\frac{1}{T}\sum_{t=0}^{T-1}(\frac{1}{L_u}\mathbb{E}||\nabla_u F(u^t,V^t)||^2+\frac{M}{L_vK^2}\sum_{k=1}^{K}\mathbb{E}||\nabla_v F(u^t,v_k^t)||^2) \\
&\leq\sqrt{\frac{(F(u^0,V^0)-F^*)(1+\mathcal{X}^2)(\frac{\zeta^2}{L_u}(1-\frac{M}{K})+\frac{\sigma_u^2}{L_u}+\frac{M\sigma_v^2}{KL_v})}{T}} \\
&+\sqrt[3]{\frac{(F(u^0,V^0)-F^*)^2(1+\mathcal{X}^2)(\frac{\zeta^2}{L_u}+\frac{\sigma_u^2}{L_u}+\frac{\sigma_v^2}{L_v})(1-\frac{1}{E})}{T^2}} \\
&+\frac{(F(u^0,V^0)-F^*)(1+\mathcal{X}^2)(1+P^2)}{T} \\
&+\frac{(F(u^0,V^0)-F^*)\sqrt{\frac{K}{M}(1-\frac{1}{E})(1+\mathcal{X}^2)(1+P^2)}}{T} \\
&\leq \mathcal{O}(\frac{1}{\sqrt{T}})+\mathcal{O}(\frac{1}{\sqrt[3]{T^2}})+\mathcal{O}(\frac{1}{T}),
\end{aligned}
\tag{22}
$$

$F(u^0,\ V^0)$ 表示初始值为 u^0 和 V^0 的函数 F，而 F^* 则包含 F 的所有其他实例。显然，FedCaP 的收敛速率近似为 $1/\sqrt{T}$，与原始随机梯度下降（SGD）一致，但这仅在以下条件下成立：

$$
T \geq \frac{KL_vL_u(1+P^2)(F(u^0,V^0)-F^*)}{\zeta^2L_v(K-M)+\sigma_u^2L_v+ML_u\sigma_v^2}. \tag{23}
$$

在每轮都选中所有客户端且局部损失函数中不存在正则化项的情况下，我们可以将条件 23 改写如下：

$$
T \geq \frac{KL_vL_u(F(u^0,V^0)-F^*)}{\sigma_u^2L_v+KL_u\sigma_v^2}. \tag{24}
$$