arXiv:2305.04205v3 [cs.CV] 6 Sep 2023

Bi-Mapper：面向自动驾驶的整体式鸟瞰视图语义映射

李思宇[1], 杨凯伦[1], 石昊[2], 张嘉铭[3,4],林嘉诚[5],滕志峰[3],和李志勇[1,5]

摘要—覆盖道路基本元素的道路场景语义地图，是自动驾驶系统不可或缺的组成部分。将其呈现为鸟瞰视图（BEV）时，可为定位和规划提供重要的感知基础。目前，假设深度的先验知识可在标定参数的辅助下，引导模型直接学习将前视图转换为BEV。然而，这种方法在表示远处物体时会受到几何畸变的影响。此外，另一类不依赖先验知识的方法则能从全局视角隐式学习前视图与BEV之间的变换。考虑到融合不同学习方法可能带来意想不到的益处，我们提出了用于俯视道路场景语义理解的Bi-Mapper框架，将全局视角与局部先验知识相结合。为增强两者之间可靠的交互，我们提出了一种异步互学习策略。同时，我们设计了跨空间损失（*ASL*），以减轻几何畸变带来的负面影响。在nuScenes和Cam2BEV数据集上的大量实验结果验证了所提出的Bi-Mapper框架中各模块的稳定有效性。与现有的道路映射网络相比，所提出的Bi-Mapper在nuScenes数据集上的交并比提高了2.1% 。此外，我们还验证了Bi-Mapper在真实驾驶场景中的泛化性能。源代码已公开发布于BiMapper。

索引词—用于视觉感知、制图和智能交通系统的深度学习

I. 引言

IN 对于自动驾驶系统而言，语义地图是一项重要的基础要素，会影响定位和规划等下游工作。近年来，鸟瞰视图（BEV）地图展现出卓越的性能 [1]。它能够以简洁的范式构建高精地图（HD-map），并可据此轻松生成路径规划 [2]。与高精地图的表示方式类似，BEV道路地图可以表示道路分隔线等基本道路元素，

稿件收到日期：2023年5月6日；修订日期：2023年7月4日；录用日期：8月2023年28日。本文经副编辑和审稿人的意见评估后，由编辑 Javier Civera 推荐发表。本研究部分得到国家自然科学基金（项目编号：U21A20518和61976086）以及杭州思锐影像科技有限公司的资助。
（通讯作者：*Kailun Yang* 和 *Zhiyong Li*。）[1]S. Li、K. Yang 和 Z. Li 就职于中国长沙湖南大学机器人学院，邮编410082（电子邮箱：kailun.yang@hnu.edu.cn；zhiy- ong.li@hnu.edu.cn）。[2]H. Shi 就职于中国杭州浙江大学现代光学仪器国家重点实验室，邮编310027。[3] J. Zhang 和 Z. Teng 就职于德国卡尔斯鲁厄理工学院拟人化与机器人研究所，卡尔斯鲁厄76131。[4]J. Zhang 还就职于英国牛津大学工程科学系，牛津OX1 3PJ。[5]J. Lin 和 Z. Li 就职于中国长沙湖南大学计算机科学与电子工程学院，邮编410082。数字对象标识符（DOI）：见本页顶部。

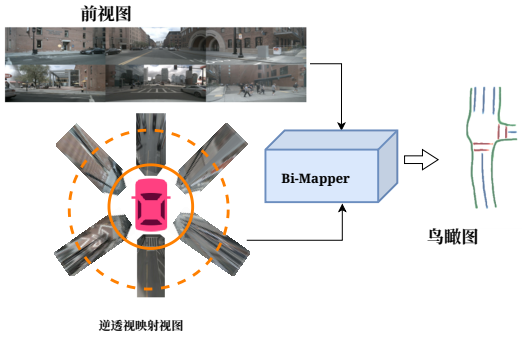

图1. Bi-Mapper 从前视图和逆透视映射视图构建鸟瞰图道路地图。它们从两个不同视角描述物体之间的关系。逆透视映射视图由假设深度生成，对近处物体的表征具有鲁棒性，但对远处物体存在畸变，如橙色圆圈所示。

人行横道和边界，从而实现全面的驾驶场景理解 [3]。

如今，鸟瞰图道路地图通常由前视图图像转换得到 [4]。根据内参和外参，可以利用深度信息直接构建鸟瞰图道路地图。然而，不具备深度信息的消费级单目摄像头是一种经济高效的选择。因此，BEV映射的核心在于有效学习高质量特征，并从前视图场景中预测俯视语义。

目前，实现BEV语义映射主要有两个方向。采用显式深度估计的BEV映射是其中一个方向 [5]，，它根据标定参数将二维像素投影到三维点。逆透视映射（IPM） [6] 是其中一种特殊情况，它为每个像素赋予一个假设深度。这种具有重要指导作用的先验知识可能会带来一些问题，例如在远处物体的表征中产生几何畸变 [7]。如图1所示，在逆透视映射视图图像中，远处物体显得模糊，而近处物体则相对清晰。另一个方向依靠深度学习，通过前视图变换器隐式学习深度信息和标定参数 [8], [9]。尽管在训练初期学习较慢，但经过充分训练后，其表现较好。鉴于这两种策略各有优势，我们提出一个值得探讨的问题：二者之间的交互能否产生互补效应，从而提升BEV场景理解能力?

为此，我们提出了 Bi-Mapper，这是一种以并行方式兼顾先验知识和视图变换器学习的框架。据我们所知，这是首次提出一种同时利用透视视图和逆透视映射视图进行鸟瞰视图语义映射的双流框架。具体而言，该框架结合了两条流的能力，即局部自视图流（LV）和全局交叉视图流（GV）。LV 在假设深度的先验知识指导下运行，而 GV 则利用了 ..... 的自学习能力。

使模型能够隐式推断语义。一方面，先验知识具有内在的指导作用，这意味着它能在训练初期指明学习方向。为应对几何畸变带来的负面影响，我们提出一种跨空间损失（*ASL*）来缓解这些影响。它在另一个空间，即相机系统坐标系中监督学习。需要注意的是，鸟瞰视图的真实值位于自车坐标系中。另一方面，在学习初期，来自 GV 学习的信息相对匮乏。因此，我们提出一种异步互学习（*AML*）策略，直到两个流地位平等时才开始互学习。具体而言，在初始阶段，只有 LV 充当教师。所提出的异步策略有助于各流学习有效知识。

我们在 nuScenes [10] 和 Cam2BEV [11]数据集上进行了大量实验。Bi-Mapper在nuScenes数据集上的交并比达到37。9%，比现有方法的最佳表现高2。1%。在Cam2BEV数据集上也取得了出色的结果，交并比达到86。7%，比当前方法高4。1%。此外，所提出的方法还能持续有效地为HDMapNet等最先进方法带来互补增益 [8]和 LSS [5]。其交并比分别提升了4。0% 和6。6%。

总而言之，我们研究的主要贡献如下：

- 我们提出了一种端到端的Bi-Mapper框架，可同时从不同视角学习鸟瞰图地图。其中一条流侧重于先验知识，另一条则利用自学习能力，并通过跨流交互获取互补知识。
- 鉴于先验知识能够引导学习方向，我们设计了一种异步互学习策略，以平衡两个流，并缩小BEV语义映射中先验知识的差距。
- 为提高网络的准确性，我们提出了跨空间损失，该损失也能缓解几何畸变问题。
- 大量实验表明，所提出算法的性能优越，且各关键组件均有效。

II. 相关工作

BEV感知研究已受到学者关注 [2], [12]。本节简要概述BEV感知的研究进展。当前研究主要分为三类：一类基于逆透视映射（IPM）变换器，第二类通过深度估计实现，另一类则利用深度神经网络直接学习该参数模型。

基于IPM的方法： IPM是最早、最简单的实现鸟瞰视图感知的解决方案 [13], [14]。许多研究工作都在此基础上学习语义知识。该工作 [15]首先通过IPM将前视图中的特征转换为俯视图，然后使用网络学习语义特征。IPM存在一个明显的问题：物体的几何形状在远处可能会发生扭曲。

Zou 等人 [16] 设计了两个分支，用于学习逆透视映射与几何特征之间的差异，并通过互学习机制实现这一目标。出于类似的动机，我们提出了一种异步学习方案。考虑到学习初期学习水平不一致的问题，我们的方法最初仅设置一条流作为教师。当两条流的学习水平相当时，便开始互学习，从而确保学习到的特征具有较高的可靠性。此外，Can 等人 [17]考虑了时序信息的聚合，这可以解决动态障碍物问题。除了我们研究的语义地图外，矢量地图也是一项重要的感知任务，专注于矢量实例。作为矢量地图领域的先驱工作，VectorMapNet [18],也利用逆透视映射进行视角变换。

基于深度估计的方法： 第二种范式通过深度估计实现BEV感知。这一方向的经典研究工作是 [5]。Philion 等人通过深度离散化估计每个像素的深度概率。随后，利用估计出的深度生成视锥特征图，以表示三维网格特征。基于此，Xie 等人 [19]改进了该方法，以降低运行内存占用。他们假设深度沿光线均匀分布，即沿相机光线的网格具有相同的特征。近期，Huang 等人 [20]通过优化视角变换来加快推理速度，并采用数据增强策略 [19]，进一步改进了该方法。Hu 等人 [21]提出了一种用于感知、预测和规划的端到端网络。他们借鉴 [5]中的方法进行BEV映射。然而，所有这些方法都基于隐式学习的深度。 [22], 表明，将深度真实值用作显式监督能够取得出色效果。在动态环境中，时序信息十分重要。因此，Hu 等人 [23] 基于估计深度融合时序特征，以预测当前状态和未来状态。相较于深度源相机坐标系，高度对应于车辆坐标系。因此，研究三维空间的高度对于构建鸟瞰视角地图具有重要意义。Li 等人 [1] 考虑了高度与时间的并行性，并引入了时序特征。近期，这项改进也已应用于矢量地图网络MapTR [24]。

基于学习的方法： Lu 等人 [25] 使用变分自编码器学习前视图与俯视图数据之间的转换。输入为图像，输出为语义占据栅格地图。在 [5], 中，深度被划分为 $1m$ 个单元，而 Roddick等人 [26] 则按照由近及远的原则将其划分为四个部分。他们利用金字塔中的深层特征，通过将其压缩至瓶颈层并展平为特征图，来学习近距离特征图。这样做的原因是，远处的小物体通常无法在深层中清晰呈现。在 [27], 中，提出了垂直变换器和平面变换器，以验证其准确性。Pan 等人 [28] 提出了另一种方法，使用多层感知机（MLP）学习这种转换。同时，他们还采用域适应，以便将该方法部署于真实环境中。

在上述工作基础上，研究者提出了改进方法。在 [8], 中，多视图图像被转换

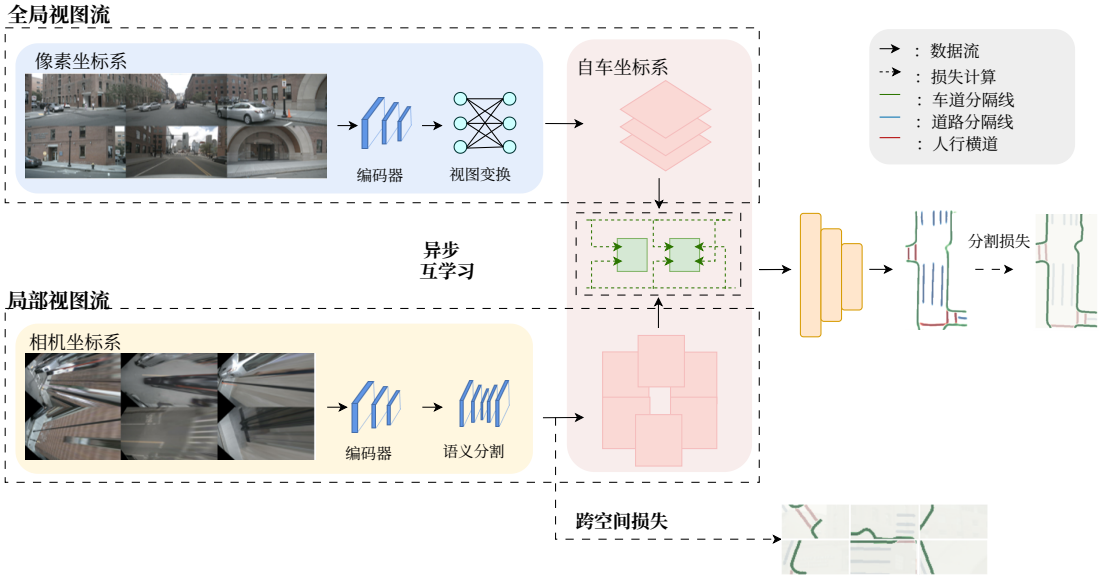

图 2. 所提出的 Bi-Mapper 网络框架。该框架由全局交叉视图流、局部自视图流、异步互学习和解码器模块组成。视图变换 将每个视角下的像素层特征转换到相机坐标系中。语义分割 学习每个逆透视映射视图的语义特征。除分割损失外，还引入了跨空间损失，以缓解几何畸变问题。

到相机坐标系中，所采用的是多个多层感知机。随后，通过外参在自车坐标系中构建鸟瞰图地图。然而，我们利用多层感知机直接学习内参和外参。由这些参数生成的逆透视映射可以引导多层感知机的方向。为提高准确性，汽车和道路的对抗学习在 [9]中得到了研究。同时，他们引入了一种循环自监督方案。具体而言，他们将鸟瞰视图投影到前视图，并计算循环损失。在此基础上，Yang [29] 添加了一个双重注意力模块。随着视觉变换器展现出卓越性能，研究人员发现，它适合弥合前视图与鸟瞰视图 [30], [31], [32]之间的差距。然而，视觉变换器的计算复杂度很高。实际上，图像中的每条垂直线都与鸟瞰视图中对应的射线相关。因此，通过 [33]中的注意力机制学习了这种对应关系。同时，加入了鸟瞰视图中射线之间的上下文信息，以增强空间关系。Gong等人 [34] 也通过添加几何先验知识关注了这种对应关系。

深度互学习作为模型蒸馏的一个分支，由 [35]提出。其核心在于学习多流深层特征之间的关系。Kim等 [36] 证明，互知识蒸馏不仅在融合分类器中表现良好，也明显有利于子网络。因此，它在多模态网络中得到了广泛应用。文献 [37] 和 [38]都关注相机与激光雷达数据流之间的特征蒸馏。前者的互学习旨在对来自目标框的特征施加加权损失。后者通过互损失，使用相机和激光雷达数据流之间BEV特征图的均方误差。不同于这些针对最终目标的工作，我们提出了一种

针对特征层的异步互学习策略会考虑两条流的学习程度。两种特征

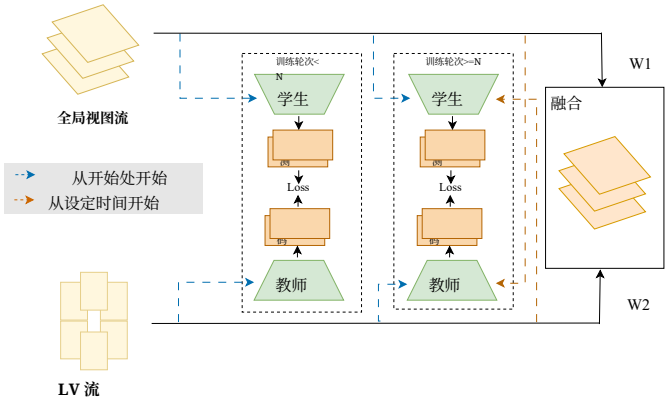

图3. 异步互学习与融合模块。融合各层以学习鸟瞰图布局。

III. 方法

我们首先介绍三种坐标系（第III-A节），并概述所提出的 Bi-Mapper框架（第III-B节）。然后，我们分别介绍全局跨视角流（第III-C节）和局部自视图流（第III-D节）的细节、跨空间损失（第III-E节）以及异步互学习模块（第III-F节）。

A. 问题表述

该变换涉及三个坐标系：像素坐标系、相机坐标系和自车坐标系。这三个坐标系之间有两个变换矩阵：一个内参矩阵T_{in} 和一个外参矩阵 T_{ex}。

$$T_{in} = \begin{bmatrix} f_x & 0 & c_x \\ 0 & f_y & c_y \\ 0 & 0 & 1 \end{bmatrix}, \qquad T_{ex} = [R|T],$$

其中 $f_x = \alpha f$, $f_y = \beta f$。 f为焦距。像素坐标系中的物体在 u 轴上缩放了 α 倍

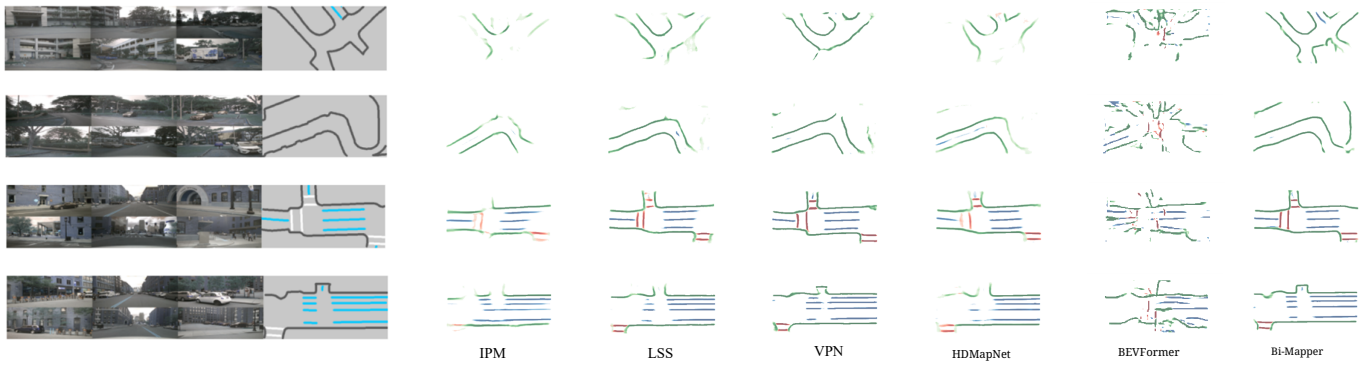

图4. nuScenes验证集上的映射结果。从左到右依次为 IPM [6], LSS [5], VPN [28], HDMapNet [8], BEVFormer [1], 以及我们提出的Bi-Mapper。

表I nuScenes数据集上的比较结果。∗ 表示数据来自相应的参考文献（交并比（%））（倒角距离（米））。

方法	分隔线			人行横道			边界			所有类别	
	IoU	CD_p	倒角距离 L_ CD	IoU	CD_p	倒角距离 L_ CD	IoU	CD_p	倒角距离 L_ CD	IoU	
逆透视映射 [6]	38.6*	1.045	0.812	0.941	19.3*	1.085	1.420	1.232	39.3*	0.523	1.494	0.968	32.4*
LSS [5]	38.3*	1.054	0.468	0.782	14.9*	0.657	0.454	0.573	39.3*	0.453	0.515	0.478	30.8*
VPN [28]	36.5*	0.994	0.246	0.644	15.8*	0.474	0.138	0.329	35.6*	0.620	0.551	0.588	29.3*
HDMapNet [8]	40.6*	0.874	0.312	0.699	18.7*	0.358	1.055	0.946	39.5*	0.285	0.346	0.323	32.9*
BEVFormer [1]	42.1	0.487	**0.006**	0.212	23.8	0.201	0.027	0.118	41.6	0.034	**0.001**	0.019	35.8
Bi-Mapper	**43.8**	**0.147**	0.010	**0.084**	**25.7**	**0.045**	**0.009**	**0.026**	**44.2**	**0.004**	0.006	**0.005**	**37.9**

相对于相机坐标系，在 v 轴上缩放了 β 倍。u 和 v 是像素坐标系。同时，原点分别偏移了 c_x 和 c_y。R 为旋转矩阵，T 为平移矩阵，它们是通过相机与自车之间的标定得出的。

三个坐标系之间的变换由公式计算得出。1。该工作 [8] 利用多层感知机学习 T_{in} 和 Z_c。随后 T_{ex} 用于在自车坐标系中获取鸟瞰视图特征图。不同于 [8], 所提出的方法学习 Z_c, T_{in} 以及 T_{ex} 。

$$Z_c * \begin{bmatrix} u \\ v \\ 1 \end{bmatrix} = T_{in} * \begin{bmatrix} X_c \\ Y_c \\ Z_c \end{bmatrix} \begin{bmatrix} X_c \\ Y_c \\ Z_c \\ 1 \end{bmatrix} = T_{ex} * \begin{bmatrix} X_w \\ Y_w \\ Z_w \\ 1 \end{bmatrix}, \quad (1)$$

其中，X_c、Y_c和 Z_c 位于相机坐标系中，而 X_w、Y_w和 Z_w 位于自车坐标系中。

*B. Bi-Mapper*框架

如图2所示，Bi-Mapper框架包括两条流，即全局交叉视图流（GV）和局部自视图流（LV）、跨空间损失以及异步互学习模块。GV流以完全依赖学习的方式，在自车坐标系中融合不同视角的特征，从而能够实现多视角交叉学习的结果。我们注意到，在以往基于学习的方法中，几乎没有探索过将像素坐标转换到鸟瞰视图坐标的物理过程。具体而言，BEV道路地图位于自车坐标系中；在简单假设的基础上，可以根据内参获得像素坐标系与相机坐标系之间的转换，而另一种转换则来自外参。考虑到成像的物理模型是一项重要的先验知识，其中涉及三个坐标系——像素坐标系、相机坐标系和自车坐标系，LV流分别遵循该模型，从多视角中学习局部特征。然而，这一简单假设会导致几何畸变。因此，我们进一步设计了跨空间损失，以减小这种副作用。两条流采用不同的学习方式，但其目标是一致的。为此，我们设计了异步互学习模块，以提升分割能力。随后，将GV流和LV流融合为全局特征图。最后，使用分割头获得BEV道路地图。

C. 全局跨视角

先前的工作 [8] 已证明，多层感知机能够构建像素与相机坐标系之间的映射模型。相机坐标系与自车坐标系之间的变换遵循外参。然而，本节对此提出了新的见解：我们使用多层感知机直接在像素坐标系与自车坐标系之间进行变换。这是因为内参和外参之间的关联涉及深度。例如，深度决定了二者之间的关联，而这种关联可以在像素级别上学习。因此，我们使用多层感知机直接学习三个坐标系之间的视角变换。不同视角的特征图被融合为一张全局地图。

具体而言，来自各个视角的图像经编码器嵌入后，可得到像素坐标系中的特征图。随后，使用多层感知机将像素层中的特征图转换到自车坐标系中。不同视角采用共享的两层多层感知机架构，但参数各不相同。随后，将不同视角的特征图累加到自车坐标系中的全局特征图上，并通过独立的多层感知机层自适应地调整自车视角下跨视角特征的权重：

$$F_g = \sum_{n=0}^{5} \varphi_2^n(ReLU(\varphi_1^n(F_p^n))), \quad (2)$$

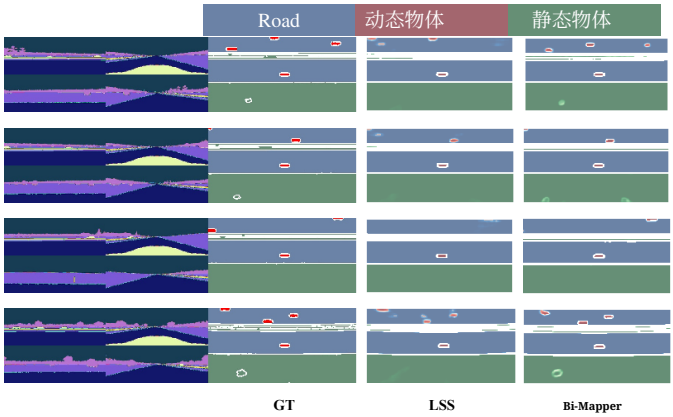

图 5. Cam2BEV 数据集上的映射结果。从左到右依次为 LSS [5]和 Bi-Mapper 的结果。其中， $F_g \in R^{h\times w\times c}$ 为全局特征图， $F_p^n \in R^{H\times W\times C}$ 为像素层中第 n 个视图的特征图， $\varphi_i(\cdot)$ 为第 i 层多层感知机。 $ReLU$ 表示修正线性单元激活。

D. 局部自视图

由于全局视图流充分利用网络来学习变换模型，我们的 LV 流则基于先验参数，并将先验知识融入神经网络架构。上述元素，包括道路分界线、车道分界线和人行横道，构成了本研究的研究范围。这些元素位于车道上，因此可以推断，逆透视映射 [6],适用于这些元素。因此，利用内参可以获得相机坐标系中距地面约一米高的平面图像。需要注意的是，该图像存在几何畸变，下一节（第 III-E节）将对此进行分析。

为获得各相机坐标系中的语义地图，首先将各视角的图像转换到各自对应的相机坐标系中。随后，采用基础语义分割网络 U-Net，生成各相机坐标系中的语义地图。最后，利用外参将各相机坐标系中得到的语义地图合并为自车坐标系中的全局语义地图。

E. 跨空间损失

在生成 BEV道路地图的过程中，会用到三个坐标系。通常，真实值位于自车坐标系或像素坐标系中。较少有 BEV 研究 [8], [15] 在相机坐标系中考虑真实值。但对于局部自视角流，如果不加以处理，几何畸变就会被带入融合过程，最终妨碍另一条流的学习。为缓解这些问题，我们提出使用跨空间损失。具体而言，全局地图中的自车位姿和外参可推导出每帧各视角对应的相机坐标系下的真实值。随后，使用

表 II Cam2BEV 数据集上的对比结果。

方法	Road	动态物体	静态物体	所有类别
逆透视映射 [6]	39.1	33.7	56.1	43.0
LSS [5]	97.1	56.5	94.3	82.6
VPN [28]	52.4	31.0	67.6	50.3
HDMapNet [8]	39.6	1.0	66.1	35.6
Bi-Mapper	**97.7**	**67.5**	**95.0**	**86.7**

交叉熵损失计算各相机坐标系下的结果与真实值之间的差异，以确保生成的 BEV道路地图的准确性，并减少可能出现的几何畸变带来的影响：

$$Loss_{ASL} = \sum_{j=0}^{5} L_{ij}(F_{gt}, F_{pre}), \tag{3}$$

其中 j 是视图数量， L_{ij} 是交叉熵损失。

F. 异步互学习与融合

虽然两条流通过不同的方法学习BEV道路地图，但它们具有相似的语义表达。此外，在互学习设置下，它们可以相互利用对方的信息。默认情况下，深度互学习（DML）是适用于此场景的解决方案。通常，只要两条流都由真实值监督，DML就会同时应用于这两条流之间。然而，在本案例中，全局视图流和LV 流并未直接受到标签监督，因此需要异步互学习与融合模块，如图3所示。

起初，LV 流凭借其先验知识以及在相机坐标系中接受跨空间损失监督的优势，充当全局视图流的教师。相反，全局视图流在训练初期的学习程度较低。随着全局视图流学习变换，它可以通过异步互学习同时指导LV 流，而该过程受交叉熵损失约束：

$$Loss_{mutual} = \left\{ \begin{array}{ll} Loss_{LV} & n < N, \\ Loss_{LV} + Loss_{GV} & else. \end{array} \right\} \tag{4}$$

其中， $Loss_{LV}$ 表示教师是 LV 流。$Loss_{GV}$ 表示教师是全局视图流。 N 是训练轮次。本质上，两条流通过预测特征图的前景和背景，相互学习彼此的特征学习成果，其中教师充当掩码。经过互学习后，两条流通过加权求和操作融合为全局特征图。

G. 训练损失

总体而言，损失可通过以下方式计算：

$$Loss = Loss_{BCE} + Loss_{ASL} + \alpha \cdot Loss_{mutual} \tag{5}$$

其中， $Loss_{BCE}$ 是二元交叉熵损失，用于缩小自车坐标系中预测与真实值之间的差距。 $Loss_{ASL}$ 是跨空间损失。$Loss_{mutual}$ 是互学习的结果。 α 是一个平衡权重，根据经验设为 0.1。

表 III 不同模块的消融结果。

方法	分隔线	人行横道	边界	所有类别
基线	40.1	22.2	39.7	34.0
基线 +CSL	43.0	24.7	42.9	36.9
基线 + ASL+AML	**43.8**	**25.7**	**44.2**	**37.9**

表 IV 不同模块的消融结果。

方法	分隔线	人行横道	边界	所有类别
LV 教师	42.8	24.0	42.6	36.5
GV教师	42.7	23.6	43.1	36.5
同步	42.9	23.9	43.9	36.5
异步	**43.8**	**25.7**	**44.2**	**37.9**

IV. 实验

为验证我们提出的方法，我们开展了一系列实验，分为三个部分：首先，将其与最先进的鸟瞰视图方法进行比较；其次，通过消融研究分析 Bi-Mapper 的各个组成部分；最后，在实际机器人导航场景中评估 Bi-Mapper。

A. 实验数据集

nuScenes [10] 是一个全面、通用的自动驾驶数据集。每帧包含六个相机视角，适用于BEV映射任务。此外，该数据集还提供详细的内参和外参，这些是该任务的核心。根据数据集中道路元素列表，我们选取了三种道路元素，即 道路分隔线、人行横道和车道分隔线。

Cam2BEV [11] 包含两个由语义分割道路场景图像构成的合成数据集，共有四个相机视角。此外，它还提供鸟瞰图中的真实值。在本研究中，道路场景包含三个语义类别：道路、动态物体和静态物体。

B. 评估指标

交并比（IoU）：预测结果与真实值之间的交并比如下所示：

$$IoU(M_1, M_2) = \frac{|M_1 \cap M_2|}{|M_1 \cup M_2|}, \tag{6}$$

其中， M_1 和 M_2 分别表示语义预测和真实值。

倒角距离（CD）：这是 [8] 中定义的评判标准。参考文献证明，它可以用于比较结构化输出。它表示预测与真实值中每条曲线上各点之间的距离：

$$\begin{aligned} CD_p &= \frac{1}{C_1} \sum_{x \in C_1} \min_{y \in C_2} ||x, y||, \\ CD_L &= \frac{1}{C_2} \sum_{x \in C_2} \min_{y \in C_1} ||x, y||, \end{aligned} \tag{7}$$

其中， C_1 和 C_2 分别是预测和真实值中的点集。 x 和 y 是像素坐标。

表 V 异步互学习开始时间的实验结果。

Time	分隔线	人行横道	边界	所有类别
第 5 次	43.8	25.7	44.2	37.9
第 8 次	42.6	24.4	42.9	36.6
第 12 次	40.7	21.9	40.5	34.4

表 VI 异步互学习启动时间的结果。“S”表示同步。“A”表示异步。

变体		分隔线	人行横道	边界	所有类别
L2 损失	S	40.4	24.8	42.0	35.7
KL 散度损失	S	43.2	24.2	43.7	37.0
	A	**43.6**	**24.6**	**43.6**	**37.3**
交叉熵损失	S	42.9	23.9	43.9	36.9
	A	**43.8**	**25.7**	**44.2**	**37.9**

C. 实现细节

两条流的编码器均为 EfficientNet-B0 [39]。鸟瞰图解码器如参考文献 [8] 所示。使用 Adam 优化网络，训练 30 轮次。初始学习率为 0.001，并在第 10 轮次衰减至原来的 0.1 倍。BEV道路地图的宽度和高度为 200×400 ，对应的范围分别为 $(-15m, 15m)$ 和 $(-30m, 30m)$，分辨率为 0.15。对于局部自视图流，相机坐标系中的图像尺寸为 400×800，通过逆透视映射（IPM）获得。该过程假设 y 在相机坐标系中为 $1m$ 。 x 和 z 的范围分别为 $(-5m, 5m)$ 和 $(3m, 29m)$。这是因为学习到的特征图尺寸为 32×88。

D. 与其他方法的比较

我们将 Bi-Mapper 与四种方法进行了比较。经典方法是 IPM [6] ，它是我们方法的基本组成部分。首先在像素坐标系中对图像进行学习，然后通过内参与外参将其投影到自车坐标系中。此外，LSS [5] 利用深度估计。VPN [28] 和 HDMapNet [8]使用多层感知机学习变换。BEVFormer [1]在当前鸟瞰图感知研究中具有相对较好的性能，因此将其加入比较。需要说明的是，在所有实验中，视图变换并不一致。

nuScenes 上的结果： 表 I展示了 nuScenes数据集上的结果比较。显然，Bi-Mapper 的性能优于此前所有方法。此外，Bi-Mapper 的运行时间与其他方法相当，具体为 $2.4s$。所有类别的交并比（37。9%）高于最佳性能（35。8%），提升了 2。1%。这一增益得益于所提出的 LV 流和全局视图流的融合，能够同时考虑局部学习和全局学习。其他方法只关注其中一个方面。随后，我们使用倒角距离比较不同方法的性能。倒角距离的对象是语义像素。 CD_p 表示精确率， CD_L表示召回率。可以清楚地观察到，我们的方法具有出色的性能。

Cam2BEV 上的结果： 表 II展示了在 Cam2BEV 数据集上的比较结果。Bi-Mapper 和 LSS 表现出色。Bi-Mapper 比 LSS 高 3.9% 在

表 VII 不同融合权重的消融结果

变体 GV 和 LV 的权重	分隔线	人行横道	边界	所有类别
拼接	40.5	25.1	41.4	35.7
1.0 1.0	37.7	20.7	40.0	32.8
0.1 1	39.9	21.0	41.1	34.0
1 0.1	**43.8**	**25.7**	**44.2**	**37.9**

表 VIII 激光雷达视图流（包含 CSL 和 AMU）的增益效果。

方法	分隔线	人行横道	边界	所有类别
LSS	38.3	14.9	39.3	30.8
LSS +	**43.5(+5.2)**	**24.9(+10.0)**	**43.7(+4.4)**	**37.4(+6.6)**
HDMapNet	40.6	18.7	39.5	32.9
HDMapNet +		**24.4(+5.7)**	**43.2(+3.7) 3**	

交并比。这表明所提出的 Bi-Mapper 能够无缝适应各种道路场景。

nuScenes 上的可视化：可视化结果如图 4所示。与其他方法相比，Bi-Mapper 能够在各种场景中构建精度极高的鸟瞰视图道路地图。然而，一些较远处的线条难以看清。

Cam2BEV 上的可视化： 图 5 展示了 LSS 和 Bi-Mapper 在 Cam2BEV 数据集上的可视化结果，二者表现相当。它们都具备较强的能力，能够学习有关道路和静态物体的语义信息。但 Bi-Mapper 在识别动态物体方面更具优势。

E. 消融研究

核心模块： 鸟瞰视图道路地图网络的核心包括两部分：一是跨空间（ASL），二是异步互学习（AML）。我们首先验证了它们的有效性。基线包含两条流（LV 和 GV），二者以加权方式融合，并采用分割损失。随后加入 ASL。最后一次实验中加入 AML。结果如表 III所示。这两个模块分别使交并比提高了 2.9%和 1.9% 。这表明它们有助于实现稳健的语义映射。同时，我们补充展示了 ASL 消融实验的可视化结果，如图 6 所示。我们比较了三种方法：逆透视映射、LV+GV，以及 LV+GV+ASL。从结果可以看出，逆透视映射对远处目标的识别较为模糊。我们提出的双流结构和 ASL 模块表现更佳。值得注意的是，这些模块对处理逆透视映射的几何畸变有积极作用。

异步互学习： 接下来，如表 IV 所示，我们验证了异步性在互学习中的有效性。我们设计了四组实验。首先，分别从两条流进行单独学习。然后进行同步和异步互学习。实验结果证明了互学习的必要性。此外，异步互学习优于同步互学习，这表明提出异步性的意义。

开始时间： AML 的开始时间是一个值得探索的课题，它体现了互学习的影响。结果如表 V 所示。请注意，LV 教师从训练开始时便启动。这表明将开始时间设为第五个训练轮次更好。更晚的异步介入时间会拉低

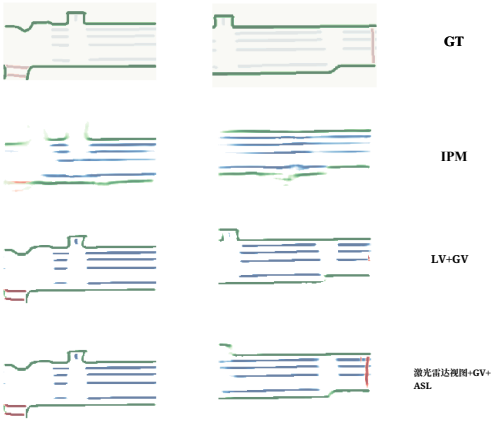

图 6. ASL 消融实验的可视化结果。从上到下分别为真实值、逆透视映射、LV+GV 和 LV+GV+ASL。

准确率。出现这种情况的原因可能是，长期运行的 LV 教师使网络陷入单一的学习方向，互学习的影响很小。同时，我们尝试了三种损失选择，结果列于表 VI。结果 37。9% 的交并比表明，所提出的异步互学习策略是有效的。

融合：有趣的是，我们发现不同的融合权重也会影响语义映射的准确率。因此，表 VII 展示了融合权重差异的消融实验。当 GV 和 LV 的权重分别为 1 和 0。1 时，效果最佳。这证实了 LV 流仍然存在几何畸变，而较小的权重可以减小其影响。

关于 Bi-Mapper 的有效性：最后，我们通过 LSS [5] 和 HDMapNet [8]测试了结合 ASL 和 AML 的 LV 流所带来的增益效果。这两个网络将作为全局视图流的替代模块。如表IV-D所示，该模块始终能显著提升它们的准确率。因此，将我们提出的模块应用于其他方法，有望获得互补的性能提升。

F. 实际机器人导航场景中的应用

为验证模型的泛化能力，我们将模型应用于实际机器人导航场景。我们使用一台四轮运输机器人，在工业园区和校园两个地点采集了真实世界数据集。该机器人配备了一台双目摄像头和三台单目摄像头。

虽然在这些场景中没有真实值可用于测量准确性，但我们以视觉方式展示了一组结果。如图7所示。尽管某些区域的预测并不正确，但与其他方法相比，我们的方法能够输出更准确的BEV道路地图。出现这种情况的原因是，面对不同场景时，标定参数并不相同。其他方法只能从数据集中学习相应的参数，这可能无法适应新的环境。然而，我们的方法将标定参数作为输入属性，而不是直接学习的对象。具体而言，经过标定参数变换后的逆透视映射视图会在激光雷达视图中进一步学习。

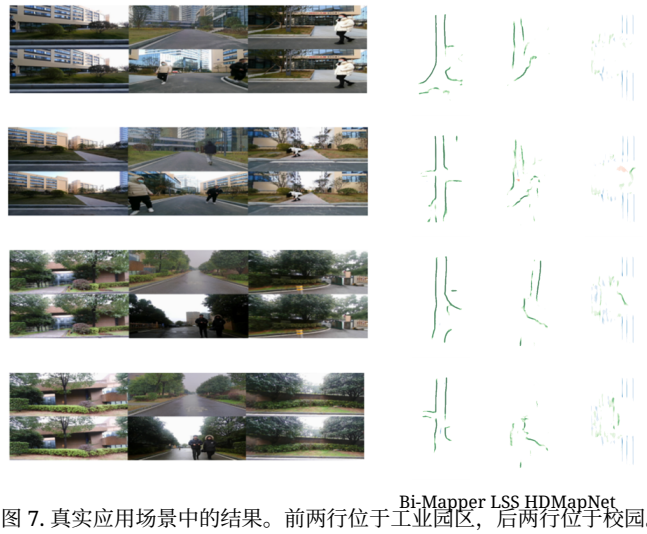

图 7. 真实应用场景中的结果。前两行位于工业园区，后两行位于校园。由于只有四个视角，左右图像会重复输入。V. 结论

本文提出了 Bi-Mapper，用于自动驾驶中的整体式 BEV语义映射。Bi-Mapper 是一种双流网络，可从前视图和逆透视映射视图构建 BEV道路地图，并采用跨空间损失和异步互学习来增强自上而下的语义映射。在 nuScenes 和 Cam2BEV 数据集上进行的大量实验表明，该方法性能出色，并有望持续提升各种映射方法的准确性。此外，它在实际应用场景中展现出稳健的语义映射效果。未来，我们计划利用时序信息构建覆盖范围更广、精度更高的 BEV道路地图。

参考文献

[1] Z. Li 等人，“BEVFormer：通过时空变换器从多相机图像中学习鸟瞰图表示”，载于 ECCV 论文集，2022 年，第 1–18 页。[2] H. Li 等人，“深入探究鸟瞰图感知的难题：综述、评估与实践指南”，arXiv 预印本 arXiv:2209.05324，2022 年。[3] K. Peng 等人，“MASS：面向稠密俯视理解的激光雷达数据多注意力语义分割”，IEEE 智能交通系统汇刊，第 23 卷，第 9 期，第 15 824–15 840 页，2022 年。[4] Y. B. Can、A. Liniger、O. Unal、D. Paudel 和 L. Van Gool，“利用车载摄像头理解道路语义的鸟瞰图”，IEEE 机器人与自动化快报，第 7 卷，第 2 期，第 3302–3309 页，2022 年。[5] J. Philion 和 S. Fidler，“Lift, splat, shoot：通过隐式反投影至 3D 对任意摄像头配置的图像进行编码”，载于 ECCV 论文集，2020 年，第 194–210 页。[6] H. A. Mallot、H. H. Bülthoff、J. Little 和 S. Bohrer，“逆透视映射简化了光流计算和障碍物检测”，《生物控制论》，第 64 卷，第 3 期，第 177–185 页，1991 年。[7] A. Bar-Hillel、R. Lerner、D. Levi 和 G. Raz，“道路和车道检测的最新进展：综述”，《机器视觉与应用》，第 25 卷，第 3 期，第 727–745 页，2014 年。[8] Q. Li、Y. Wang、Y. Wang 和 H. Zhao，“HDMapNet：在线高清地图构建与评估框架”，载于 ICRA 论文集，2022 年，第 4628–4634 页。[9] W. Yang 等人，“专注投射视角：通过跨视角变换进行单目道路场景布局估计”，载于 CVPR 论文集，2021 年，第 15 531–15 540 页。[10] H. Caesar 等人，“nuScenes：自动驾驶多模态数据集”，载于 CVPR 论文集，2020 年，第 11 618–11 628 页。[11] L. Reiher 等人，“Cam2BEV”，2020 年。 [在线]。网址：https://github.com/ika-rwth-aachen/Cam2BEV [12] Y. Ma 等人，“以视觉为中心的鸟瞰图感知：综述”，arXiv 预印本 arXiv:2208.02797，2022 年。[13] S. Sengupta、P. Sturgess、L. Ladick`y 和 P. H. Torr，“从街景图像自动生成稠密视觉语义地图”，载于 IROS 论文集，2012 年，第 857–862 页。[14] S. Ammar Abbas 和 A. Zisserman，“从图像中获取鸟瞰图的几何方法”，载于 CVPR 论文集，2019 年，第 4095–4104 页。[15] L. Reiher、B. Lampe 和 L. Eckstein，“一种将多台车载摄像头图像转换为鸟瞰视图语义分割图像的仿真到现实深度学习方法”，载于 ITSC 论文集，2020 年，第 1–7 页。[16] J. Zou 等人，“HFT：通过混合特征变换提升透视表示”，arXiv 预印本 arXiv:2204.05068，2022 年。[17] Y. B. Can、A. Liniger、O. Unal、D. Paudel 和 L. Van Gool，“利用车载摄像头理解道路语义的鸟瞰图”，IEEE 机器人与自动化快报，第 7 卷，第 2 期，第 3302–3309 页，2022 年。[18] Y. Liu、T. Yuan、Y. Wang、Y. Wang 和 H. Zhao，“VectorMapNet：端到端矢量化高清地图学习”，ICML 论文集，2023 年。[19] E. Xie 等人，“M^2BEV：采用统一鸟瞰视图表示的多相机联合 3D 检测与分割”，arXiv 预印本 arXiv:2204.05088，2022 年。[20] B. Huang 等人，“Fast-BEV：迈向实时车载鸟瞰感知”，arXiv 预印本 arXiv:2301.07870，2023 年。[21] S. Hu、L. Chen、P. Wu、H. Li、J. Yan 和 D. Tao，“ST-P3：通过时空特征学习实现端到端视觉自动驾驶”，载于 ECCV 论文集，2022 年，第 533–549 页。[22] Y. Li 等人，“BEVDepth：为多视角三维目标检测获取可靠深度”，载于 AAAI 论文集，2023 年。[23] A. Hu 等人，“FIERY：从环视单目摄像头预测鸟瞰图中的未来实例”，载于 ICCV 论文集，2021 年，第 15 253–15 262 页。[24] B. Liao、S. Chen、X. Wang、T. Cheng、Q. Zhang、W. Liu 和 C. Huang，“Maptr：面向在线矢量化高清地图构建的结构化建模与学习”，ICLR 论文集，2023 年。[25] C. Lu、M. J. G. van de Molengraft 和 G. Dubbelman，“采用卷积变分编码器—解码器网络的单目语义占据栅格映射”，IEEE 机器人与自动化快报，第 4 卷，第 2 期，第 445–452 页，2019 年。[26] T. Roddick 和 R. Cipolla，“利用金字塔占用网络从图像预测语义地图表示”，载于 CVPR 论文集，2020 年，第 11 135–11 144 页。[27] N. Gosala 和 A. Valada，“利用单目前视图像进行鸟瞰全景分割”，IEEE 机器人与自动化快报，第 7 卷，第 2 期，第 1968–1975 页，2022 年。[28] B. Pan、J. Sun、H. Y. T. Leung、A. Andonian 和 B. Zhou，“用于感知周围环境的跨视角语义分割”，IEEE 机器人与自动化快报，第 5 卷，第 3 期，第 4867–4873 页，2020 年。[29] S. Gao、Q. Wang 和 Y. Sun，“S2G2：用于自动驾驶的单目摄像头半监督语义鸟瞰栅格地图生成”，IEEE 机器人与自动化快报，第 7 卷，11 981，2022 4，11 974 期，第 – 页。[30] B. Zhou 和 P. Krähenbühl，“用于实时地图视角语义分割的跨视角变换器”，载于 CVPR 论文集，2022 年，第 13 750–13 759 页。[31] S. Chen、T. Cheng、X. Wang、W. Meng、Q. Zhang 和 W. Liu，“通过几何引导核变换器实现高效稳健的二维到鸟瞰视角表示学习”，arXiv 预印本 arXiv:2206.04584，2022 年。[32] L. Peng、Z. Chen、Z. Fu、P. Liang 和 E. Cheng，“BEVSegFormer：从任意摄像头配置实现鸟瞰图语义分割”，载于 WACV 论文集，2023 年，第 5935–5943 页。[33] A. Saha、O. Mendez、C. Russell 和 R. Bowden，“将图像转换为地图”，载于 ICRA 论文集，2022 年，第 9200–9206 页。[34] S. Gong 等人，“GitNet：用于鸟瞰分割的基于几何先验的变换”，载于 ECCV 论文集，2022 年，第 396–411 页。[35] Y. Zhang、T. Xiang、T. M. Hospedales 和 H. Lu，“深度互学习”，载于 CVPR 论文集，2018 年，第 4320–4328 页。[36] J. Kim、M. Hyun、I. Chung 和 N. Kwak，“用于在线互知识蒸馏的特征融合”，载于 ICPR 论文集，2021 年，第 4619–4625 页。[37] Z. Chen、Z. Li、S. Zhang、L. Fang、Q. Jiang 和 F. Zhao，“BEVDistill：用于多视角三维目标检测的跨模态知识蒸馏”，arXiv 预印本 arXiv:2211.09386，2022 年。[38] Y. Hong、H. Dai 和 Y. Ding，“用于单目三维目标检测的跨模态知识蒸馏网络”，载于 ECCV 论文集，2022 年，第 87–104 页。[39] M. Tan 和 Q. Le，“EfficientNet：重新思考卷积神经网络的模型缩放”，载于 ICML 论文集，2019 年，第 6105–6114 页。