为了在复杂和具有挑战性的驾驶场景中安全运行，自动驾驶车辆（AVs）需要对其环境有全面的感知。在实际交通中，仅依赖本地传感器数据无法实现这一点，因为存在视线限制、有限的感知范围或恶劣的天气条件。为了克服这些限制，集体感知（CP）是一种有前景的方法，其中关于环境的信息在联网和自动驾驶车辆（CAVs）之间共享。CP 的主要挑战是如何将集体共享的信息与本地感知的信息融合。CP 可以分为三个类别：早期、中期和晚期融合。晚期融合中仅分享检测到的物体状态，这在带宽上效率高且不受传感器域差距影响；然而，由于信息损失，性能通常低于其他融合方法。在早期融合中，传感器的原始数据如 LiDAR 点云被交换，这不会损失信息并能达到高性能；然而，带宽需求非常高。在中期融合中，交换神经网络特征，它既在带宽上高效又能达到非常高的性能；然而，如果不是所有的 CAVs 运行相同的模型，它会受到特征域差距的影响。但在中期融合中，不仅不同的模型会导致域差距，CAVs 之间使用不同传感器也会导致传感器域差距。特别是对 LiDAR 传感器而言，分辨率和视场的高多样性使得传感器之间的域差距可能非常大。早期融合和中期融合都容易受到传感器域变化的影响。

在实际应用中，制造商将部署多种不同的传感器系统，而目前最先进的 CP 方法仅考虑所有 CAVs 之间的同质传感器设置。这主要是由于以往的 CP 数据集仅包含一种类型的传感器数据；然而，最近发布的 SCOPE [1] 数据集包含每个 CAV 的三种不同激光雷达传感器的数据，这允许对 CP 中 Sensor2Sensor 领域差距进行广泛研究。

因此，我们研究了 CP 中的 Sensor2Sensor 领域差异，并进一步提出了一种称为 S2S-NET 的传感器领域鲁棒融合架构。

我们的主要贡献是：

在第 I 节中，我们介绍了与我们方法相关的工作。之后，在第 **??** 节中，我们提出了 S2S-Net，这是一种针对基于激光雷达的集体感知的领域鲁棒融合架构。第 II 节描述了所进行的实验和评估。结果在第 **??** 节中提供。最后，我们给出了结论并展望了未来的研究。

I. 相关工作

A. 数据集

为了研究 CP 中由于不同传感器域导致的域间差异，需要包含来自不同传感器类型数据的数据集。Axmann 等人 [2] 发布了 LUCOOP 真实世界的 V2V 数据集。该数据集由装备了不同 LiDAR 传感器的三个 CAV 的录音构成。第一个 CAV 配备了两个 LiDAR 传感器。其中一个是水平安装的 Hesai Pandar 64 层 LiDAR，另一个是垂直安装在 CAV 后部的 Velodyne VLP16 16 层 LiDAR。第二个 CAV 配备了一个 Hesai Pandar-XT32 32 层 LiDAR，最后一个 CAV 配备了 Velodyne VLP16 LiDAR。所有 LiDAR 的记录频率为 10 Hz。每个 CAV 记录的帧数如下：CAV1 和 CAV2 各 15,000 帧，CAV3 记录了 7,000 帧。需要注意的是，三个 CAV 为数据集记录目的作为车队多次行驶相同市区内环路线。这导致它们的相对位置总是相似的，并且场景多样性非常低。因此，该数据集不适合作为基准。

由 Xiang 等人提出的 V2X-Real 是一个真实世界的 V2X 数据集。为了记录数据，两辆 CAV 配备了 Robosense Ruby Plus 128 通道 360° LiDAR，测程为 200 m ，同时两台 RSU 配备了 Ouster OS1-128/64 通道 360° LiDAR 传感器，测程为 40 m 。此外，CAV 和 RSU 还配备了 2-4 个 1920 × 1080 px 的 RGB 摄像头。总体而言，该数据集包含 33,000 帧，涵盖来自 10 个不同类别的物体，包括各种车辆、行人和骑车者。然而，数据集中仅有两辆 CAV，它们都配备了相同的 128 层 LiDAR 传感器。因此，该数据集不适合用于 V2V CP 中的传感器域适应研究。此外，由于数据采集仅在一个交叉路口进行，场景多样性受到严重限制。

在 TUMTraf V2X 集体感知真实世界数据集 [3] 中，该数据集使用一个 RSU 和一个 CAV 在一个大型交叉路口记录，共包含 1000 帧。一个安装在标志桥上的 Ouster LiDAR OS1-64 64 层 360° LiDAR，具有 120 m 范围，和四个 1920 × 1200 px RGB 摄像头作为 RSU。CAV 配备了一个 Robosense RS-LiDAR-32 360° LiDAR，具有 32 层和 200 m 范围。此外，CAV 配备了一个 1920 × 1200 px RGB 摄像头。然而，由于帧数有限且没有额外的 CAV，该数据集不适合用于 V2V 集体感知中的 Sensor2Sensor 领域自适应。2022 年，Yu 等人 [4] 提出了 DAIR-V2X 数据集，它是一个不同子数据集的家族，其中 DAIR-V2X-C 是用于车与基础设施之间的 CP。这是第一个现实世界的大规模集体感知数据集。28 个不同的交叉路口配备了两个 300 层 100° LiDAR 传感器，具有 280 m 感知范围，以及两个 1920 × 1080 px RGB 摄像头，它们以 10 Hz 记录。此外，一个 CAV 配备了一个 40 层 360° LiDAR 和一个 1920 × 1080 px RGB 摄像头。该数据集包含 13k 帧，涵盖了包括行人和骑自行车者在内的不同对象类别。然而，由于 DAIR-V2X-C 仅包含一辆车辆，因此不适合评估 V2V 集体感知中的 Sensor2Sensor 领域自适应。Gamerdinger 等人发布了合成的 SCOPE 数据集，该数据集是使用 CARLA 模拟器生成的。该数据集具有多种环境特征，包括城市交叉路口、乡村道路和高速公路，包含 44 个场景，总共 17,600 帧，描绘了不同的对象类别，如汽车、货车、（摩托）骑自行车的人和行人。CAV 的数量在不同场景之间变化，从 3 到 21 不等，作为合理的交通密度选择了约为 50 % 的 CAV 比率。这些 CAV 配备了 5 个 1920 × 1080 px 的摄像头和三种不同的激光雷达传感器，一个 Velodyne HDL64 64 层 360° ，一个 Velodyne VLP32 32 层 360° ，以及一个具有 52 条线和 70° × 的视场的 Blickfeld Cube 固态激光雷达。CARLA 激光雷达模型缺乏真实的物理特性，如光束发散。对于 SCOPE，使用了一个改进的激光雷达传感器模型，其中包括物理特性。除了 CAV，SCOPE 在合适的场景中使用了 RSU，比如城市交叉路口。这些 RSU 配备了与 CAV 相同的传感器。由于增强的传感器模型、每个传感器单元的三种不同激光雷达传感器、广泛的场景多样性以及各种不同的道路使用者，SCOPE 数据集非常适合研究传感器到传感器的领域适应性。

到目前为止，CP 中的传感器域差距大多未被研究；然而，最近王等人发布了一个名为 V2X-DGPE 的框架，他们在其中解决了基于车辆到基础设施（V2I）感知的域差距问题。他们使用了 DAIR V2X [4] 数据集进行评估。

V2X-DGPE 利用一个共享特征提取主干网络，用于车辆传感器数据以及基础设施数据。由于共享主干在两个传感器域上进行训练，因此没有对未见传感器域的领域适应能力进行评估，这构成了未来部署 CAVs 的最大挑战，因为模型无法在所有可能的传感器域上进行训练。

Zhi 等人通过分别在车辆数据和基础设施数据上训练他们的 DCGNN 框架，解决了 DAIR V2X 数据集上的 V2I 传感器域间隙问题。在测试过程中，来自两个领域的数据都被用作输入。他们在 DAIR V2X 数据集上的结果显示，基础设施模型相比于在协同数据上评估的车辆模型表现有大幅下降。他们的结果突显了传感器域间隙对 CP 性能的强烈影响。然而，由于缺少仅车辆感知和仅基础设施感知的基线结果，因此无法量化这种影响。此外，他们还在一个自生成的基于 CARLA [5] 的数据集上评估了他们的方法，该数据集具有 128 层和 32 层 LiDAR 传感器，且他们的方法表现得更好；但由于该数据集的关键信息没有提供，同时也缺少本地检测的基准，因此不清楚改善的性能是由什么引起的。

为了降低计算复杂度和内存消耗，我们基于 MR3D-Net [6] 提出了一种轻量级且传感器领域鲁棒的融合架构，我们省略了 MR3D-Net 中的多分辨率输入流，仅使用单一分辨率输入流来共享稀疏体素网格。正如在 MR3D-Net 中所示，稀疏体素网格的细节级别和数据大小之间存在权衡。由于细节的变化可能也会影响 S2S-Net 的领域自适应能力，我们在统一的主干网络上使用高输入分辨率来评估 S2S-Net，因为这预期是最容易受到 Sensor2Sensor 领域差距影响的，因为它不像低分辨率那样统一不同的传感器分辨率。图 **??** 给出了所提出的 S2S-Net 架构的概览。我们在 OpenPCDet [7] 工具箱中实现了 S2S-Net 用于基于 LiDAR 的 3D 目标检测。

B. 环境表示

与其他现有 CP 方法相比，S2S-Net 利用稀疏体素网格作为环境表示，而不是原始传感器数据、特征图或对象状态。正如 MR3D-Net 所证明的，稀疏体素网格是一个有效且紧凑的 CP 环境表示。稀疏体素网格是通过将点云的空间划分为一个均匀的网格，然后仅存储包含点的体素坐标来构建的。稀疏体素网格相比于其他环境表示有一些重要的优势。稀疏体素网格比点云更加紧凑，可以实现高达 94 % 的数据压缩，同时仍然保留高细节水平 [6] 。此外，稀疏体素网格是一个统一的表示方法，可以与各种模型融合，而不会像中间融合方法那样遭受特征域差距的影响。在 S2S-Net 中，稀疏体素网格不包含任何特征，只交换体素网格的坐标以减少传输的数据量。然后作为模型的输入，体素的中心点被用作特征，这可以从坐标、体素大小和体素网格的原点计算得到。

C. 架构

为了融合集体数据与本地传感器数据，我们利用两条输入流：集体主干网络和本地主干网络。集体主干网络处理集体共享的信息，而本地主干网络处理自我传感器数据。每个输入流由四个连续的卷积块组成，其中一个卷积块由一个稀疏卷积层和两个子流形卷积组成。在这些层之后，应用批量规范化和 ReLU 激活函数。为了减少体素网格的空间维度，稀疏卷积层可能具有步幅。集体输入流和本地输入流不一定具有相同的分辨率。如果输入分辨率不匹配，我们在集体和本地输入流中使用不同的步幅将空间分辨率降低到一个通用分辨率。为了融合不同分辨率流之间的信息，我们使用散射操作。

由于体素网格中的稀疏性，两个体素网格的特征通道无法进行拼接，因为有些体素仅存在于两个体素网格中的一个，而其他体素则同时存在于两个体素网格中。因此，特征通道的拼接会导致每个体素的特征数量不平衡，使得卷积操作无法应用。为了融合稀疏的体素网格，我们应用了散射操作，对于那些同时出现在两个体素网格中的体素，称为重复体素，应用元素级的 max 函数，保持特征通道的数量不变。其他仅出现在一个体素网格中的体素，称为单体素，则保持不变。最终的体素网格就是来自两个网格的散射重复体素和单体素的并集。

在 S2S-Net 中，散射操作应用于每个卷积块之后，其中集合骨干网和局部骨干网共享相同的空间分辨率。散射的稀疏体素网格随后用作局部骨干网中下一个卷积块的输入，将集合信息以不同的分辨率传播到局部骨干网。这样，输入流可以从不同的输入中学习，使神经网络能够以不同的方式从多个来源提取特征。在四个卷积块之后，结果由最终的稀疏卷积层处理，然后通过沿 z 轴连接体素特征映射到鸟瞰视图，形成一个 2D 特征图，然后用作物体检测器的输入。

对于两个输入流，我们选择体素大小为 $5\,\mathrm{cm} \times 5\,\mathrm{cm} \times 10\,\mathrm{cm}$ 。对于训练和评估，我们使用最大网格大小为 $280\,\mathrm{m} \times 80\,\mathrm{m} \times 4\,\mathrm{m}$ ，产生输入分辨率为 $5600 \times 1600 \times 40$ 体素。在局部骨干网络和集体骨干网络中，第二、第三和第四稀疏卷积使用步长为二来减少空间维度。两个输入流在所有稀疏卷积块中共享相同数量的输出通道，即 $[16, 32, 64, 64]$ 。所有层的核大小为 $3 \times 3 \times 3$ 。

II. 评估

A. 数据集

对于评估，我们使用 SCOPE 数据集 [1] ，如在章节 I 中讨论的。由于每个 CAV 使用了三种不同的激光雷达传感器，SCOPE 数据集非常适合用于评估 V2V 集体感知中的领域适应能力。此外，由于不同的场景、各种道路使用者和多变的环境条件，SCOPE 数据集使得在 CP 中进行 3D 物体检测的评估更加具有表现力。

为了评估泛化性能，我们分别在每个传感器上进行训练，其中自我传感器和其他 CAV（协作自动驾驶车辆）的传感器是相同的。然后在测试期间，我们保持与训练阶段相同的自我传感器，并分别评估每个其他 CAV 传感器的模型。与自我传感器相同的其他 CAV 传感器的评估作为基准，因为这是训练域，而其他两个传感器则是未见过的域。此外，由于在实际场景中 CAV 也不一定使用相同的传感器，我们还评估一个将 CAV 随机分配到可用传感器域的情况，该情况包括来自训练域和两个未见过的传感器域的数据。我们不评估在训练和测试中更换自我传感器时的域适应能力，因为自我传感器是已知的，并且模型可以在该传感器域上进行训练，仅其他 CAV 的传感器域是未知的。作为基准，我们还在仅使用自我传感器数据而不融合任何协作信息的情况下训练和测试 S2S-Net。为此，我们使用自我传感器数据作为集体和局部骨干网的输入。

我们在 SCOPE 训练集上训练 S2S-Net。我们使用 S2S-Net 作为主干网，并与 PV-RCNN [8] 一起用于目标检测。为了评估，我们报告了在 SCOPE 测试集上的平均精度（AP）结果。作为交并比（IoU）阈值，我们为汽车使用 0.7 ，针对行人和骑自行车的人使用 0.5 。我们在 x 方向范围为 $[-140, 140]$ m ，y 方向范围为 $[-40, +40]$ m ，z 方向范围为 $[-4, 1]$ m 内评估检测结果，因为这是 SCOPE 的官方评估范围。训练和测试期间，我们不限制 CAV 之间的通信范围，即评估范围内的所有可用信

TABLE I: 在 SCOPE 测试集上，对于逐个以及每个 CAV 随机选择传感器的情况下，S2S-Net 和仅本地感知的平均精度结果。对于汽车使用 0.7 的 IoU 阈值，对于行人和骑行者使用 0.5 的 IoU 阈值。

		HDL64 Test			VLP32 Test			Blickfeld Cube Test			Random Sensor Test		
Method	Sensors Train	Car	Pedestrian	Cyclist	Car	Pedestrian	Cyclist	Car	Pedestrian	Cyclist	Car	Pedestrian	Cyclist
No Fusion	HDL64	34.15	20.51	20.27	-	-	-	-	-	-	-	-	-
	VLP32	-	-	-	38.32	29.81	22.19	-	-	-	-	-	-
	Blickfeld Cube	-	-	-	-	-	-	8.30	1.72	2.08	-	-	-
S2S-Net	HDL64	50.04	32.12	26.30	52.89	30.26	30.01	46.07	30.06	23.40	50.53	31.33	27.77
	VLP32	41.04	32.16	22.17	42.38	35.53	23.83	38.44	28.13	17.01	40.08	34.15	21.89
	Blickfeld Cube	20.36	7.05	7.19	24.48	7.29	8.91	12.18	4.70	3.31	19.69	7.38	6.26

息都会被融合。为了使传感器之间的比较公平，我们在测试期间不对任何边界框进行过滤，即使它们完全被遮挡且内部没有点或体素。这与大多数其他集体感知基准不同；然而，过滤遮挡的边界框会偏向于低分辨率和较小视场的传感器，因为例如盲点不会影响感知性能。这以及较大的评估范围会导致结果相比其他集体感知基准，如 OPV2V [9] ，较低。

从无融合基线结果来看，VLP32 显然在所有类别中都取得了最佳结果，对汽车、行人和骑自行车者的平均精度分别为 38.32 、29.81 和 22.19 。HDL64 表现略差，AP 分别减少了 4.17 、9.3 和 1.92 百分点（p.p.）。Blickfeld Cube 获得的结果最低，AP 仅为 8.3 、1.72 和 2.08 。Blickfeld Cube 性能较低的原因仅在于其仅有 70° 的狭窄水平视场，导致了巨大的盲点。相比之下，VLP32 和 HDL64 是 360° 传感器，因此盲点只能由遮挡或感测范围的限制引起，从而导致显著更高的性能。

与 S2S-Net 的融合显然可以改进所有传感器领域中的仅局部感知的结果。对于使用 HDL64 训练和测试的 S2S-Net，汽车、行人和骑行者的 AP 分别是 50.04 、32.12 和 26.30 ，这相比分别提高了 15.89 p.p.、11.61 p.p. 和 6.03 p.p.。对于领域外的测试，使用 HDL64 训练的 S2S-Net 也表现出非常高的领域鲁棒性。对于 VLP32 作为未见领域，性能甚至可以通过 2.85 p.p. 和 3.71 p.p. 分别在汽车和骑行者方面提高，只有行人的 AP 降低了 1.86 p.p.。这表明 S2S-Net 对传感器分辨率的变化非常鲁棒。此外，通过 Blickfeld Cube，使用 HDL64 训练的 S2S-Net 对于汽车、行人与骑自行车者能够维持较高的平均精度，分别为 46.07 、30.06 和 23.30 。尽管这导致了平均精度下降了 3.97 p.p.、2.06 p.p. 和 2.9 p.p.，但相比只使用 HDL64 进行本地感知，汽车、行人与骑自行车者的感知仍然提高了 11.92 p.p.、9.55 p.p. 和 3.13 p.p.。考虑到 Blickfeld Cube 的本地感知性能较低，感知性能的下降很可能并非由域差异引起，而是由 Blickfeld Cube 的视野角度过小造成的，这导致即使在集体感知中也存在盲区。在测试过程中，为每个 CAV 随机选择传感器的结果与单独传感器测试的结果相似，平均精度介于个别传感器结果之间。这进一步强调了 S2S-Net 的强传感器域鲁棒性。在其他传感器上训练的 S2S-Net 也实现了类似的结果，检测性能没有显著下降，可以追溯到传感器域差距。

III. 结论与未来工作

在这项工作中，我们提出了 S2S-Net，一种传感器域稳健融合架构，使用稀疏体素网格作为紧凑和统一的环境表示用于集体感知。这是首次针对 V2V 集体感知中的 Sensor2Sensor 域间隙的工作。

我们证明了 S2S-Net 对于传感器域的变化具有很强的鲁棒性，并且无论共享给自车的数据来自哪个传感器，所有类别的局部感知都可以显著提高。即使在传感器差异很大的情况下改变测试域或为每个 CAV 选择不同的传感器，感知性能也没有出现明显的下降，这种下降可以归因于传感器域的差异。

对于未来的研究，我们将扩展我们的评估到不同的体素网格分辨率。此外，我们希望将其他融合方法纳入我们的研究，以进一步描述在 V2V 集体感知中的传感器领域差异。此外，我们还计划进行实验，结合基础设施传感器，以研究由传感器不同放置引起的领域差异。

References

[1] J. Gamerdinger, S. Teufel, P. Schulz, S. Amann, J.-P. Kirchner, and O. Bringmann, "Scope: A synthetic multi-modal dataset for collective perception including physical-correct weather conditions," in *2024 IEEE 27th International Conference on Intelligent Transportation Systems (ITSC)*, 2024, pp. 1–8.

[2] J. Axmann, R. Moftizadeh, J. Su, B. Tennstedt, Q. Zou, Y. Yuan, D. Ernst, H. Alkhatib, C. Brenner, and S. Schön, "Lucoop: Leibniz university cooperative perception and urban navigation dataset," in *2023 IEEE Intelligent Vehicles Symposium (IV)*, 2023.

[3] W. Zimmer, G. A. Wardana, S. Sritharan, X. Zhou, R. Song, and A. C. Knoll, "Tumtraf v2x cooperative perception dataset," *arXiv preprint arXiv:2403.01316*, 2024.

[4] H. Yu, Y. Luo, M. Shu, Y. Huo, Z. Yang, Y. Shi, Z. Guo, H. Li, X. Hu, J. Yuan, *et al.*, "Dair-v2x: A large-scale dataset for vehicle-infrastructure cooperative 3d object detection," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 21 361–21 370.

[5] A. Dosovitskiy, G. Ros, F. Codevilla, A. Lopez, and V. Koltun, "CARLA: An Open Urban Driving Simulator," in *Conference on robot learning*. PMLR, 2017, pp. 1–16.

[6] S. Teufel, J. Gamerdinger, G. Volk, and O. Bringmann, "Mr3d-net: Dynamic multi-resolution 3d sparse voxel grid fusion for lidar-based collective perception," *arXiv preprint arXiv:2408.06137*, 2024.

[7] O. D. Team, "OpenPCDet: An open-source toolbox for 3d object detection from point clouds," https://github.com/open-mmlab/OpenPCDet, 2020.

[8] S. Shi, L. Jiang, J. Deng, Z. Wang, C. Guo, J. Shi, X. Wang, and H. Li, "PV-RCNN++: Point-voxel feature set abstraction with local vector representation for 3d object detection," *International Journal of Computer Vision*, vol. 131, no. 2, pp. 531–551, 2023.

[9] R. Xu, H. Xiang, X. Xia, X. Han, J. Li, and J. Ma, "OPV2V: An open benchmark dataset and fusion pipeline for perception with vehicle-to-vehicle communication," in *2022 International Conference on Robotics and Automation (ICRA)*. IEEE, 2022, pp. 2583–2589.