中图法分类号:TP391　　**文献标识码**: A　　**文章编号**: 1006-8961(2025)11-3651-14

论文引用格式:Liu X F, Liu Y, Li H R, Zhang Y and Guo Y L. 2025. Large-model driven test-time adaptation for multi-modal point cloud semantic segmentation. Journal of Image and Graphics, 30(11):3651-3664(刘雪帆, 刘砚, 李浩然, 张晔, 郭裕兰. 2025. 大模型驱动的多模态点云语义分割测试时自适应方法. 中国图象图形学报, 30(11):3651-3664)[DOI:10.11834/jig.240762]

大模型驱动的多模态点云语义分割测试时自适应方法

刘雪帆,刘砚,李浩然,张晔,郭裕兰*

中山大学深圳校区电子与通信工程学院,深圳　518107

摘　要:**目的**　点云语义分割在面对跨域分布差异时常出现性能下降,测试时自适应(test-time adaptation,TTA)可以通过在测试阶段利用目标域的无标签数据对源域训练的模型进行在线微调,从而缓解域偏移问题。然而,传统方法往往难以精确处理点云的空间连续性与局部结构约束,适应效果有限。为增强模型泛化能力,部分方法引入二维图像利用跨模态信息以增强模型的适应性,但跨模态对齐误差易导致语义碎片化的问题,影响语义分割性能。针对上述挑战,本文提出一种结合视觉大模型知识的测试时自适应点云语义分割方法。**方法**　首先,利用CLIP(contrastive language-image pre-training)文本编码器生成类别对应的文本嵌入,将视觉—文本先验知识融入逐点特征的预测过程,为点云提供泛化能力更强的语义补充信息;其次,通过SAM(segment-anything-model)生成的区域掩码对点云特征进行局部的一致性约束,有效缓解因对齐误差导致的特征不连续及进而产生的语义碎片化问题,提升模型的语义分割性能。**结果**　本文方法在3个数据集划分的3个真实场景(数据集—数据集、地点—地点、时间—时间)中,与现有的测试时自适应和无监督域自适应方法进行了对比。实验结果表明,本文方法在数据集—数据集场景中的性能提升尤为显著。在地点—地点和时间—时间场景中,本文方法也优于当前先进模型。此外,本文的测试时自适应方法在无法获取源域数据的条件下,仍能超越部分无监督域自适应方法,展现出较高的实用价值。**结论**　本文提出的利用视觉大模型知识引导测试时自适应方法,通过融合视觉—文本信息和局部特征一致性约束,显著提升了点云语义分割在多种场景中的泛化性能。

关键词:点云;语义分割;测试时自适应(TTA);视觉基础模型;多模态

Large-model driven test-time adaptation for multi-modal point cloud semantic segmentation

Liu Xuefan, Liu Yan, Li Haoran, Zhang Ye, Guo Yulan*

School of Electronic and Communication Engineering, Sun Yat-sen University, Shenzhen　518107, China

Abstract: **Objective** Point cloud semantic segmentation aims at assigning semantic labels to individual points in 3D space. However, the performance of segmentation models often deteriorates when they are applied to unseen domains due to domain shifts——differences in the distribution of data between the source and target domains. To address this problem, domain adaptation techniques, particularly test-time adaptation (TTA), have gained attention. Unlike traditional unsuper-

收稿日期:2024-12-27;**修回日期**:2025-04-08;**预印本日期**:2025-04-15

*通信作者:郭裕兰　guoyulan@sysu.edu.cn

基金项目:国家自然科学基金项目(U20A20185, 62372491);广东省基础与应用基础研究基金项目(2022B1515020103, 2023B1515120087);深圳市科技计划资助(RCYX20200714114641140)

Supported by: National Natural Science Foundation of China(U20A20185, 62372491); Guangdong Basic and Applied Basic Research Foundation(2022B1515020103, 2023B1515120087); Shenzhen Science and Technology Program(RCYX20200714114641140)

vised domain adaptation (UDA), TTA does not require access to labeled source-domain data, which makes it a privacy-preserving and practical solution. TTA aims to fine-tune a model on unlabeled target-domain data, which enables it to generalize better to the target domain without compromising privacy. Recent advancements have incorporated multi-modal data, such as pairing point cloud data with images, to further improve domain adaptation performance. These methods enhance feature representations and provide richer statistical estimations for pseudo-label generation. However, in the case of domain shift, alignment deviations may occur between modalities, which misleads the model to make incorrect predictions. In addition, the domain differences of the modalities may further lead to inconsistent model understanding and fragmented semantic predictions, which affects the performance of point cloud semantic segmentation tasks. To this end, the use of pre-trained foundation models, such as contrastive language-image pretraining (CLIP) and segment anything model (SAM), with their rich intrinsic prior knowledge, has shown promising results in improving semantic segmentation performance in domain adaptation. **Method** This study proposes a foundation model-driven TTA method that incorporates knowledge from pre-trained models to improve multi-modal domain adaptation. The approach consists of two main components. One is the CLIP-based text-embedded segmentation layer. Based on the CLIP model, this layer uses text embeddings tailored to the categories of the target domain to replace the original segmentation layer in the model. This integration enhances the semantic generalization capability of the model, which allows it to adapt better to the target domain without requiring additional labeled data. The capability of CLIP to align text and image feature spaces through contrastive learning aids in transferring semantic knowledge from large-scale datasets to domain-specific tasks. The other is the SAM mask-based feature consistency constraint module. Built upon SAM, this module generates object masks with consistent semantics across the point cloud and image modalities. By enforcing feature consistency within these masks, the module improves feature extraction and helps the model maintain semantic coherence between different modalities. This way ensures that features remain aligned and consistent throughout the adaptation process, which enhances segmentation performance, particularly in scenarios involving complex domain shifts. **Result** The method is evaluated on three multi-modal domain adaptation scenarios using datasets: nuScenes, A2D2, and SemanticKITTI. These datasets offer diverse challenges, such as geographic and lighting domain shifts, as well as cross-dataset adaptation. Our method is compared with state-of-the-art TTA and UDA approaches to assess its effectiveness. Experimental results demonstrate the effectiveness of the proposed method across various domain adaptation scenarios. In the A2D2-SemanticKITTI adaptation setting, the method achieves a mean intersection over union (mIoU) of 48.3% in the 2D adaptation scenario, which surpasses the state-of-the-art method by 4.6%. In the 3D adaptation setting, the method achieves an mIoU of 42.9%. These results highlight the capability of the proposed method to improve 2D and 3D semantic segmentation performance by leveraging two proposed modules. Compared with UDA methods, which rely on source and target domain data, the proposed method performs competitively despite only utilizing unlabeled target-domain data. This performance demonstrates the practicality of the method in real-world scenarios where access to source data is limited due to privacy concerns. Furthermore, qualitative results illustrate substantial improvements in segmentation quality, with the method producing highly accurate and coherent results in complex environments. Ablation studies further show the validation of the proposed modules. The text-embedded segmentation layer significantly enhanced 2D performance, with the mIoU increased from 45.8% to 48.1%. The mask-constrained feature consistency module had a greater impact on 3D performance, which improved the mIoU from 40.2% to 41.8%. When both modules were combined, the method achieved the highest performance across 2D and 3D tasks, with an mIoU of 48.3% for 2D and 42.9% for 3D. These experiments confirm that the combination of both modules is crucial for achieving optimal performance in multi-modal domain adaptation. **Conclusion** In this study, we propose a TTA method that uses the knowledge of a large visual model to guide the test. By integrating visual-text information and local feature consistency constraints, it substantially improves the generalization performance of point cloud semantic segmentation in various scenarios. Extensive experiments across three real-world scenarios demonstrate the superior performance of our method.

Key words: point cloud; semantic segmentation; test-time adaptation(TTA); vision foundation model; multi-modality

0 引 言

点云语义分割任务旨在为包含丰富空间信息的三维点云中的每个点分配语义标签，从而实现对场景的精确理解，帮助系统感知和识别环境中的物体和结构，在自动驾驶、机器人导航和虚拟现实等领域具有重要应用。近年来，深度学习的快速发展显著推动了点云语义分割的进步（Qi 等，2017a；孙刘杰等，2024），得益于大规模标注数据训练的模型，分割性能取得了显著提升。然而，现实世界中的点云数据来源多样，传感器参数差异导致点云密度不同；而光照、天气等环境因素则会影响点云的反射系数，进而使得数据分布产生差异。这些设备和环境差异导致基于标准数据集训练的模型在面对未见过的目标域数据时，常常出现性能大幅下降的现象。

为了解决域迁移的问题，域自适应（domain adaptation，DA）技术被提出以解决训练数据和测试数据分布不同的问题，使得模型能够将在源域上训练学到的知识应用到目标域，并提高模型的性能。然而，传统的域自适应方法通常假设目标域中存在标注数据，这在实际应用中往往难以实现。无监督域自适应（unsupervised domain adaptation，UDA）作为一种更通用的解决方案，仅利用源域的标注数据和目标域的无标签数据来提高模型在目标域上的表现。其核心思想是通过对抗训练（张桂梅 等，2020）、特征对齐（Yi 等，2021）等技术，将源域和目标域的特征分布尽可能对齐。然而，在某些实际应用场景中，特别是在涉及隐私保护或安全性要求较高的情况下，源域数据可能无法获得，这使得传统的域自适应方法变得不可行。因此测试时自适应（test-time adaptation，TTA）（Wang 等，2020；Zhang 等，2023）逐渐成为研究热点。与无监督域自适应不同，测试时自适应不依赖源域数据，而是仅利用未标注的目标域数据对源域模型进行在线微调。这使得测试时自适应在无法获取源域数据的场景中展现出更高的实用性，更适配真实需求。

现有的一些测试时适应方法通过利用目标域的统计信息（Vu 等，2024）或生成伪标签（Ruan 和 Tang，2024）来微调源域模型，从而提高模型在目标域测试数据上的表现。然而，点云数据由于其固有的无序性和更高的空间复杂性，现有方法无法有效解决域偏移问题。为此，部分方法（Shin 等，2022）引入了二维图像数据，通过图像与点云之间的对应关系和特征一致性，增强了模型的跨域稳定性。具体而言，图像数据提供了丰富的二维信息，如颜色和纹理，而点云数据则精确描述了三维空间结构。通过结合这两种模态，模型能够从图像和点云的互补特征中获得更全面的上下文，增强了对几何和语义特征的理解，为点云在目标域上的适应过程提供更充分的引导，显著提升了在目标域上的表现。

尽管现有方法通过利用目标域的统计信息或引入额外的模态等方式提升了模型在目标域测试数据上的表现，但仍面临诸多挑战。首先，在多模态方法中，二维图像的颜色、纹理和细节通常能为三维点云提供有价值的语义信息，从而帮助模型理解场景。然而，二维数据与三维数据之间的域间差异表现出不同的特征：在二维数据中，域间差异通常表现为图像之间的统计特性差异；而在三维点云中，域间差异则主要体现在点的密度差异。由于模态间的域差异，源域上已对齐的特征空间在目标域中常常出现对齐偏差。在这种情况下，二维语义信息由于域间差异的不稳定性和特征对齐偏差，可能会对三维点云的语义预测产生误导。此外，各模态内部存在的域偏移以及模态间对齐偏差进一步导致模型难以有效整合来自不同模态的特征信息，从而产生碎片化的语义预测，如图 1（c）所示，具体表现在模型无法对目标域中的对象保持一致的理解，而是将对象理解切割成多个局部区域，导致语义理解呈现断裂和不连贯的状态。

近年来出现的预训练大模型以其出色的泛化能力，能够有效解决点云语义分割在测试时域自适应阶段面临的上述挑战。CLIP（contrastive language-image pretraining）（Radford 等，2021）通过对比学习对齐文本和图像的特征空间，使其能够提取丰富的视觉—文本知识，通过将 CLIP 引入到多模态方法中，可以为二维模型提供稳定的泛化性语义信息；SAM（segment anything model）（Kirillov 等，2023）则能够高效地进行图像目标区域分割，实现在各种场景下的通用分割功能，帮助模型提取连续的特征。这些大模型所蕴含的先验知识有助于应对多模态数据在测试时适应过程中的挑战，从而提升模型在领域偏移下的鲁棒性和适应能力。

本文提出一种大模型驱动的测试时自适应方

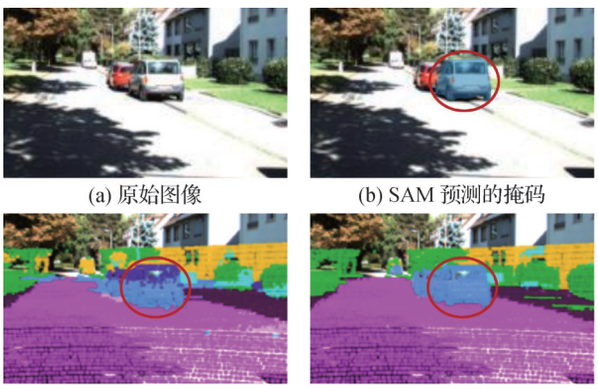
(a) 原始图像　(b) SAM 预测的掩码
(c) 碎片化语义现象　(d) 特征一致性约束后的预测结果

图1　碎片化语义现象与特征一致性约束后的校正结果

Fig. 1　Fragmented semantic phenomenon and the corrected results after feature consistency constraint

((a) original image; (b) mask predicted by SAM; (c) fragmented semantic phenomenon; (d) prediction result after feature consistency constraint)

法,利用预训练模型的知识提升点云语义分割模型在域迁移中的适应能力。具体而言,本文提出两个模块:一个基于CLIP的文本嵌入分割层和一个基于SAM掩码的特征一致性约束模块。通过将分割层替换为针对目标领域类别定制的CLIP文本嵌入,整合大模型所提供的丰富语义先验,为模型的分类预测提供了语义信息的显式指导,从而提高模型的语义泛化性和跨域稳定性。此外,利用SAM生成一致语义的对象掩码,约束掩码区域内特征的一致性,有效提升了特征表达的稳定性,解决了局部区域特征不连续导致的语义预测碎片化问题,如图1(d)所示。通过这两个模块,模型能够在目标域上获得更好的分割性能。

本文的主要贡献总结如下:1)提出一种大模型驱动的测试时自适应点云语义分割方法,针对点云数据的无序性和空间复杂性,弥补了传统方法在细粒度几何信息捕捉上的不足;2)提出基于CLIP的文本嵌入分割层,通过为模型的分类预测提供显式语义指导提高模型的语义泛化性;3)设计了基于SAM掩码的特征一致性约束模块,利用掩码中的一致性语义增强模型在跨模态对齐下提取特征的连续性和稳定性,降低了语义分割中语义碎片化风险;4)在多种场景下的实验验证了本文方法在实际任务中的优越性,消融实验进一步证明了各模块在测试时自适应场景下提升语义分割性能方面的有效性。

1　相关工作

1.1　点云语义分割

点云语义分割是三维场景理解中的一个重要任务,旨在将每个点云中的点分类到预定义的语义类别。早期工作在预处理阶段将点云转化为规则数据,例如划分体素网格(Maturana和Scherer,2015)或进行二维投影(Su等,2015)以便使用标准的卷积网络进行处理,但这类方法在处理稀疏和不规则点云时容易丢失空间信息。PointNet(Qi等,2017a)首次提出一种直接处理3D点云的方法,该方法使用共享多层感知器(multilayer perceptron,MLP)学习每个点的特征。PointNet++(Qi等,2017b)引入了层次化结构,从而进一步学习点云数据的局部几何特征。MinkowskiNet(Choy等,2019)和RandLA-Net(Hu等,2020)等方法提出高效的稀疏卷积和局部特征学习策略,有效地降低了大规模点云的计算复杂度,并能够在大规模环境中取得良好的分割精度。

近年来,多模态融合方法被用于进一步提高点云语义分割的性能。在点云分割任务中,一些工作(Jaritz等2019;Zhuang等,2021)通过将RGB图像和深度图等其他模态数据与点云数据对齐,为分割任务提供更多信息,这些基于多模态的方法使得点云处理在复杂场景中的表现得到了显著提升。

1.2　域自适应

域自适应(DA)是迁移学习的一个重要研究方向,旨在将源域中训练得到的知识迁移到目标域,从而解决源域和目标域之间分布的差异。在域自适应的研究中,无监督目标域自适应(UDA)是人们广泛关注的任务之一。无监督域自适应旨在通过源域带标签的数据和目标域无标签的数据之间的映射,缓解域间的分布差异。无监督域自适应方法广泛应用于二维和三维数据(Bian等,2022;Saleh等,2019;Achituve等,2021;Wu等,2019;Yi等,2021),近年来也有一些方法将其扩展到多模态数据(Jaritz等,2020;Peng等,2021;Xing等,2023)。例如,xMUDA(cross-modal unsupervised domain adaptation)(Jaritz等,2020)通过在源域和目标域的两种模态之间进行一致性学习,而DsCML(dynamic sparse-to-dense cross modal learning)(Peng等,2021)则采用对抗性学习,结合动态稀疏到密集的跨模态学习。这些方

法在目标域上进行适应时通常需要访问源域的数据，由于这一局限性，无监督域自适应某些应用中，尤其是因为隐私原因无法获取到源域数据时并不总是可行的。

测试时自适应（TTA）（Wang等，2020；Zhang等，2023；Ni等，2024）被提出以解决无法访问源域数据的问题，从而在无需访问源域数据的情况下，使现有的模型能够快速适应新的目标数据。TENT（test entropy minimization）（Wang等，2020）提出一种基于熵最小化的简单而有效的方法，该方法通过优化测试时的批归一化（batch normalization，BN）参数调整模型，从而减少测试时的域间差异。MM-TTA（multi-modal test-time adaptation）（Shin等，2022）的提出旨在处理复杂的多模态3D语义分割任务，其通过结合模态内和模态间的模块，使得这些模块协同工作，从而生成更可靠的自学习信号。越来越多基于多模态方法（Cao等，2023；Weijler等，2024）进一步充分利用多模态数据的互补性，在动态域转移的情况下，快速调整模型并保持较高的鲁棒性，实现更好的分割性能。

总体来说，测试时自适应方法相较于传统的无监督域自适应方法，具有更高的灵活性，能够在无需访问源数据的情况下实现快速的模型适应。这使得测试时自适应方法在现实应用中，特别是在数据受限或环境不断变化的场景中，表现出更大的潜力。

1.3 视觉大模型

近年来，随着计算能力的提升和大规模数据集的可用性，视觉大模型在计算机视觉领域取得了显著进展。其核心优势在于能够从海量多模态数据中学习抽象表示，并具备强大的迁移学习能力。通过在大规模数据集上的预训练，视觉大模型能够捕捉更丰富的视觉特征，提升在多种视觉任务中的表现，并减少对标注数据的依赖。

CLIP（Radford等，2021）是跨模态学习大模型的代表之一，采用对比学习方法联合训练大量图像—文本对，实现了图像与文本的深度融合。CLIP在零样本任务中表现出色，展现了卓越的泛化能力和跨模态融合优势。类似的模型，如ALIGN（a large-scale image and noisy-text embedding）（Jia等，2021），也通过噪声文本与图像的对比学习，利用海量数据提升了跨模态理解能力。此外，BLIP（bootstrapping language-image pre-training）（Li等，2022）和Florence（Yuan等，2021）等模型在视觉—语言任务中也取得了显著进展，推动了跨模态理解和推理的发展。

在图像分割领域，SAM（Kirillov等，2023）代表了一种创新的思路，专注于实现高效且精确的图像分割。不同于传统方法，SAM通过大规模训练数据，使模型能够自动识别并分割多种物体。同时，SAM结合了交互式分割功能，用户通过简单输入即可调整分割结果，极大提升了任务的灵活性和效率。随着相关研究的不断深入，许多后续工作（Ravi等，2024；Ren等，2024）进一步推动了通用分割模型性能提升，探索了更多技术创新，在实例分割、动态场景分割等任务中取得了显著成效。

总体而言，视觉大模型正在推动计算机视觉技术的快速发展。其强大的泛化能力使得模型能够在各种下游任务中表现出色，无论是零样本学习、少样本学习，还是跨领域迁移任务，都展现出显著的优势。通过在大规模多模态数据集上预训练，这些模型能够有效获取丰富的特征和语义表示，支持各种任务的实现。

2 本文方法

测试时自适应的目标是在源域使用带标签的成对点云和图像进行训练，得到多模态的源域模型后，再在目标域使用无标签的点云和图像对源域模型进行微调，从而获得在目标域上表现良好的模型。

本文所使用的模型框架如图2所示。在源域的训练阶段，二维模型和三维模型均使用真实标签进行全监督训练。在目标域适应阶段，受到MM-TTA（Shin等，2022）的启发，本文使用一种异步更新的模型架构进行训练。具体而言，两种模态分别设置了快模型和慢模型，快模型使用损失函数快速适应目标域的分布；而为了尽可能保留来自源域的信息，慢模型使用指数加权移动平均（exponential moving average，EMA）（Tarvainen和Valpola，2017）策略进行更新。通过慢模型对快模型的预测结果进行约束，减少在目标域适应阶段中可能出现的噪声或不稳定性。在所有模型中，网络的分割层被替换为基于CLIP的文本嵌入，以整合丰富的语义先验信息。此外，在快模型快速适应目标域的过程中利用SAM生成的掩码来约束局部的特征一致性，从而使提取的特征保持稳定性和连续性。

2.1 源域模型的训练

源域模型通过成对的源域图像和点云，以及真实标签进行全监督训练。在训练过程中，源域点云 $\boldsymbol{x}_S^{3D} \in \mathbf{R}^{N\times 3}$ 输入到三维模型 M_S^{3D} 中以生成三维特征图 $\boldsymbol{F}^{3D} \in \mathbf{R}^{N\times C}$，其中 N 为点的个数，C 为特征维度。类似地，使用二维模型 M_S^{2D} 对源域图像 $\boldsymbol{x}_S^{2D} \in \mathbf{R}^{H\times W\times 3}$ 进行特征提取。然而，由于训练过程使用的标签仅为三维点标签 $\boldsymbol{y}_S^{3D}$，需要将提取到的二维特征反投影到点云中，提取点云中 N 个点对应的二维特征，以生成二维特征图 $\boldsymbol{F}^{2D} \in \mathbf{R}^{N\times C}$，然后将获取到的特征图 $\boldsymbol{F}^{2D}$ 和 $\boldsymbol{F}^{3D}$ 分别通过各自的文本嵌入分割层，从而获取每个像素与每个点的语义概率预测 $\boldsymbol{P}^{2D} \in \mathbf{R}^{N\times K}$ 和 $\boldsymbol{P}^{3D} \in \mathbf{R}^{N\times K}$，其中，$K$ 表示类别的总数。语义概率预测与真实标签使用交叉熵损失进行分割损失，具体为

$$L_{\text{seg}}(\boldsymbol{x}_S, \boldsymbol{y}_S^{3D}) = -\frac{1}{N}\sum_{n=1}^{N}\sum_{k=1}^{K} y_S^{(i,k)} \log(\boldsymbol{P}^{(i,k)}) \tag{1}$$

式中，$\boldsymbol{x}_S$ 为 $\boldsymbol{x}_S^{3D}$ 或 $\boldsymbol{x}_S^{2D}$。源域二维模型和三维模型分别根据各自模态的损失进行反向传播训练。

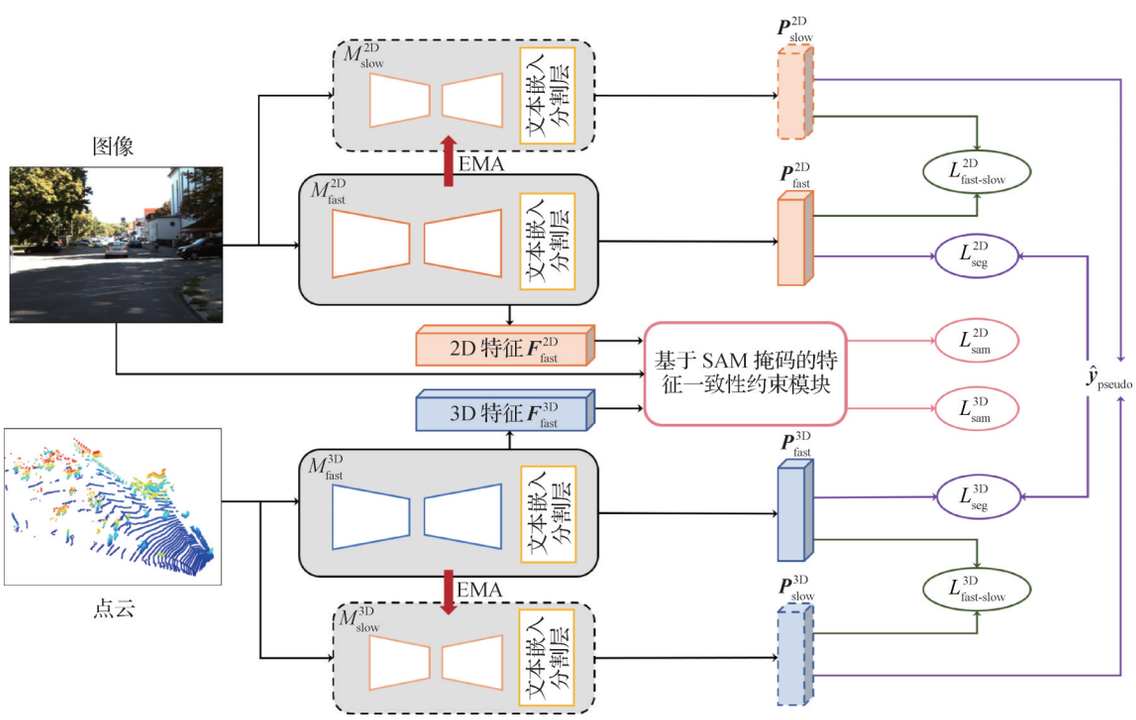

图2 本文方法的模型框架

Fig. 2 Framework of our method

2.2 目标域模型训练中的异步更新架构

为了减少在测试时训练阶段域间迁移的影响，本文使用了一种异步更新模型框架用于在目标域上对模型进行更新。二维模态和三维模态均设置了快模型（$M_{T,\text{fast}}^{2D}$ 和 $M_{T,\text{fast}}^{3D}$）和慢模型（$M_{T,\text{slow}}^{2D}$ 和 $M_{T,\text{slow}}^{3D}$），并且使用对应模态的源域模型参数进行初始化。在这两种模型中，只有与批归一化相关的参数会被更新。具体而言，快模型的参数使用损失函数通过反向传播进行更新；而慢模型使用指数加权移动平均更新，具体为

$$M_{T,\text{slow}}^{t} = \lambda M_{T,\text{slow}}^{t-1} + (1-\lambda) M_{T,\text{fast}}^{t} \tag{2}$$

式中，t 表示第 t 次迭代；M 表示 M^{2D} 或 M^{3D}；λ 是一个超参数，用于控制慢模型更新的速度。

为了在更新过程中保持快模型和慢模型之间的一致性，最大程度保留源域信息，同时降低目标域噪声对模型参数的负面影响，基于快慢模型预测结果之间的 KL（Kullback-Leibler）散度计算损失函数，具体定义为

$$L_{\text{fast-slow}} = D_{\text{KL}}(\boldsymbol{P}_{\text{fast}} \| \boldsymbol{P}_{\text{slow}}) \tag{3}$$

式中，D_{KL} 表示 KL 散度，用于衡量两个概率分布之间差异的非对称性度量；$\boldsymbol{P}$ 表示 $\boldsymbol{P}^{2D}$ 或 $\boldsymbol{P}^{3D}$，表示模型的预测输出。

此外，在每次迭代中成对的目标样本会被输入到对应模态的慢模型中以生成伪标签。具体而言，

两个模态慢模型的预测结果通过平均值进行融合，最终将预测概率最高的类别作为伪标签，记做 $\hat{\boldsymbol{y}}_{\text{pseudo}}$。生成的伪标签随后用于计算与预测结果的分割损失，从而更新快速模型中的BN层参数，具体为

$$L_{\text{seg}}(\boldsymbol{x}_T, \hat{\boldsymbol{y}}_{\text{pseudo}}) = -\frac{1}{N}\sum_{n=1}^{N}\sum_{k=1}^{K}\hat{y}_{\text{pseudo}}^{(i,k)}\log\left(\boldsymbol{P}_{\text{fast}}^{(i,k)}\right) \quad (4)$$

式中，$\boldsymbol{x}_T$ 为 $\boldsymbol{x}_T^{\text{3D}}$ 或 $\boldsymbol{x}_T^{\text{2D}}$；$\boldsymbol{P}_{\text{fast}}$ 为对应的 $\boldsymbol{P}_{\text{fast}}^{\text{2D}}$ 或 $\boldsymbol{P}_{\text{fast}}^{\text{3D}}$，表示各模态快模型的预测输出。

2.3　基于CLIP的文本嵌入分割层

在语义分割任务中，分割层通常为一个线性层，用于基于提取的特征预测每个点或像素的语义类别。其过程可以表示为 $\boldsymbol{y} = \boldsymbol{W}\boldsymbol{x} + \boldsymbol{b}$，其中，$\boldsymbol{x}$ 是输入的特征，$\boldsymbol{y}$ 表示对每个类别的预测概率，$\boldsymbol{W}$ 是线性层的权重，$\boldsymbol{b}$ 是偏置项，在本文中，偏置项设置为0。

在本方法中，通过使用CLIP的文本嵌入替代传统的权重矩阵，整合了大模型中具有强泛化性的先验信息，从而更有效地对齐提取的像素特征和点云特征与语义类别的表示。对于每个类别，构造形式为“A point of [class]$_k$”的文本提示 T_k，其中 $k \in 1, 2, \cdots, K$。这些提示通过CLIP文本编码器处理，以产生每个类别的对应文本嵌入 $\boldsymbol{E}_k$，具体为

$$\boldsymbol{E}_k = \text{CLIP}(T_k) \quad (5)$$

这些文本嵌入被拼接为权重矩阵 $\boldsymbol{W}_{\text{text}} \in \mathbf{R}^{K\times C}$，$C$ 表示分割层输入特征的维度。随后，分割层中的传统权重被替换为基于文本嵌入的权重，从而使得分割层的表示形式为 $\boldsymbol{y} = \boldsymbol{W}_{\text{text}}\boldsymbol{x}$，如图3所示。

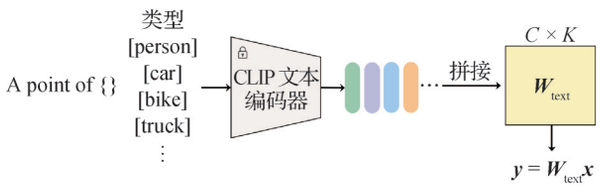

图3　基于CLIP的文本嵌入分割层

Fig. 3　CLIP-based text-embedded segmentation layer

2.4　基于SAM掩码的特征一致性约束模块

在SAM生成的掩码内保持语义一致性对提升本文方法的性能至关重要。为了充分利用掩码中的一致性信息，通过约束每个掩码内的2D特征和3D特征之间的一致性，从而改善特征提取效果，如图4所示。

首先，使用SAM对二维图像生成一组原始掩码 $\boldsymbol{M} = \{\boldsymbol{M}_1, \boldsymbol{M}_2, \cdots, \boldsymbol{M}_G\}$，其中 G 表示掩码数量；第 i 个掩码记为 $\boldsymbol{M}_i \in \{0,1\}^{H\times W}$，其中1表示该位置的像素属于该掩码。然后将这些掩码反投影回点云，即将像素的掩码值赋给点云中对应的点，从而获取三维有效掩码 $\tilde{\boldsymbol{M}}^{\text{3D}} = \left\{\tilde{\boldsymbol{M}}^{\text{3D}}{}_1, \tilde{\boldsymbol{M}}^{\text{3D}}{}_2, \cdots, \tilde{\boldsymbol{M}}_G^{\text{3D}}\right\}$。具体而言，

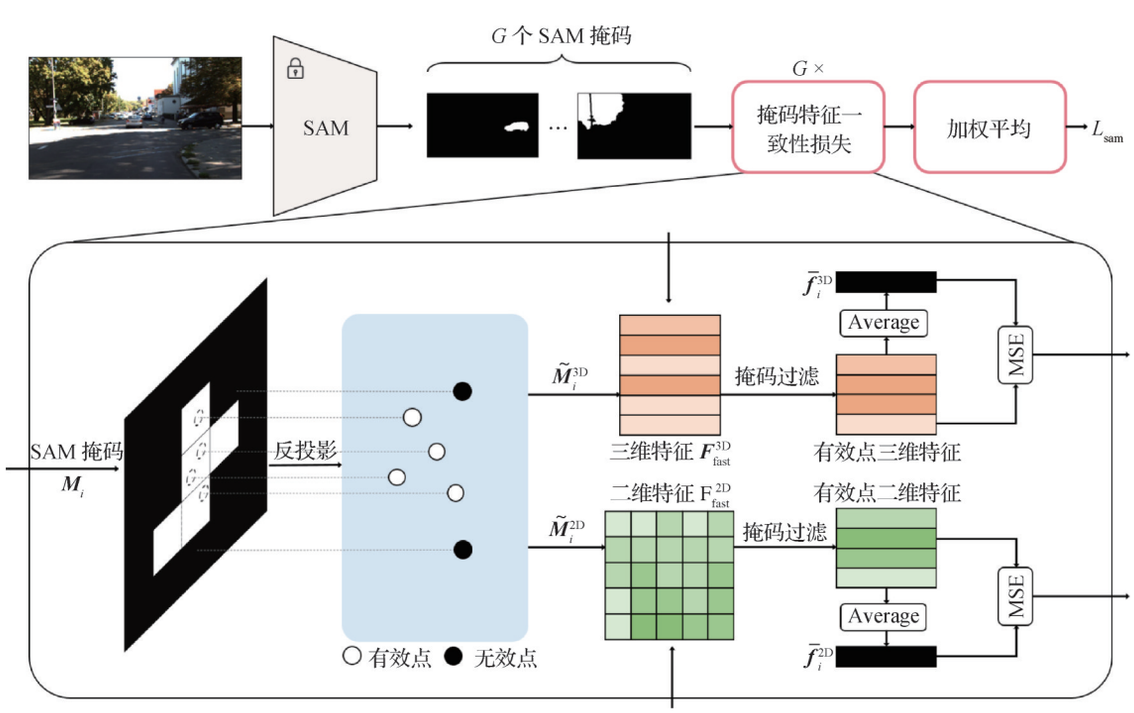

图4　基于SAM的特征一致性约束模块

Fig. 4　SAM-based feature consistency constraint module

对于一个三维有效掩码$\tilde{\boldsymbol{M}}_i^{3D} \in \{0,1\}^{\tilde{N}_i}$，其中，$\tilde{N}_i$表示第$i$个掩码内有效点的数量，有效掩码中的每个值表示该三维点对应的像素是否属于原始掩码$\boldsymbol{M}_i$。然后使用三维有效掩码从三维特征图中提取出三维有效特征$\tilde{\boldsymbol{F}}_i^{3D} \in \mathbf{R}^{\tilde{N}_i \times C}$。此外，提取出每个有效点对应的像素位置，从而获取二维有效掩码$\tilde{\boldsymbol{M}}^{2D} = \left\{\tilde{\boldsymbol{M}}_1^{2D}, \tilde{\boldsymbol{M}}_2^{2D}, \cdots, \tilde{\boldsymbol{M}}_G^{2D}\right\}$，然后根据有效掩码过滤提取出有效点对应的二维有效特征$\tilde{\boldsymbol{F}}_i^{2D} \in \mathbf{R}^{\tilde{N}_i \times C}$。

然后，对于掩码内的有效特征进行平均，计算得到掩码的局部特征$\bar{\boldsymbol{f}}_i$，从而获得局部视觉表示，具体为

$$\bar{\boldsymbol{f}}_i = \frac{1}{\tilde{N}_i} \sum_{\tilde{\boldsymbol{f}} \in \tilde{\boldsymbol{F}}_i} \tilde{\boldsymbol{f}} \tag{6}$$

式中，$\tilde{\boldsymbol{F}}$为$\tilde{\boldsymbol{F}}^{2D}$或$\tilde{\boldsymbol{F}}^{3D}$。为了确保掩码内的语义一致性，使用均方误差损失来计算每个点或像素的特征与对应的掩码局部特征之间的差异，得到单个掩码的特征一致性损失，具体为

$$L_{mi} = \frac{1}{\tilde{N}_i} \sum_{\tilde{\boldsymbol{f}} \in \tilde{\boldsymbol{F}}_i} \left\| \tilde{\boldsymbol{f}} - \bar{\boldsymbol{f}}_i \right\|^2 \tag{7}$$

式中，$\|\cdot\|^2$表示平方欧几里得范数。最后，考虑到不同掩码中有效点和像素的数量不同，简单地对所有掩码设置相同的权重可能会干扰训练信号。为解决这一问题，本文提出一种基于几何信息的加权策略，用于掩码级损失的组合。通常，较大的掩码在训练过程中会贡献更多的信号。最终，得到基于掩码的特征一致性约束损失，具体为

$$L_{\text{sam}} = \frac{\sum_{i=1}^{G} L_{mi} \cdot \tilde{N}_i}{\sum_{i=1}^{G} \tilde{N}_i} \tag{8}$$

通过在所有掩码上应用这种加权损失，模型能够在同一掩码内约束二维和三维特征保持一致，从而提高语义连贯性，并在测试时适应过程中增强两种模态的特征表达稳定性。

2.5 测试时自适应阶段的损失函数

在测试时自适应阶段，快模型使用损失函数反向传播对BN层相关参数进行更新，两个模态模型的总损失可以表示为

$$L^M = \beta_1 L_{\text{seg}}^M + \beta_2 L_{\text{fast-slow}}^M + \beta_3 L_{\text{sam}}^M \tag{9}$$

式中，$M \in \{2D, 3D\}$表示模态，$\beta_1, \beta_2, \beta_3$分别表示各部分损失的权重。

3 实验结果与分析

3.1 数据集与实验设置

为了全面评估本文提出的方法有效性，在3个不同的领域适应场景中进行测试，在这些设置中，使用到了nuScenes（Caesar等，2020）、A2D2（Geyer等，2020）和SemanticKITTI（Behley等，2019）数据集。

1）nuScenes是一个大规模的自动驾驶数据集，专注于复杂城市驾驶环境，涵盖美国和新加坡的高速公路、城市街道和交叉路口等场景。数据来源包括5个摄像头、1个32通道LiDAR和多个高精度雷达，且覆盖不同天气和光照条件。该数据集包含1 000个20 s的场景，支持3D目标检测、跟踪、语义分割和行为预测等任务。

2）A2D2数据集聚焦于复杂的城市场景。它结合了230万像素的高分辨率相机和16通道LiDAR，生成稀疏的点云数据，适用于城市环境中的数据处理研究。该数据集包含多种交通场景，如行人、骑行者和其他车辆的标注，提供精确的3D位置信息和语义标签，供自动驾驶技术进行高质量的训练和验证。

3）SemanticKITTI是基于KITTI数据集扩展的语义分割数据集，专注于高精度点云数据标注。它采用64通道激光雷达和0.7百万像素相机，生成密集的点云数据，适用于目标检测和场景理解任务。该数据集涵盖城市和乡村环境，以及不同天气条件下的场景，提供丰富的空间结构信息和语义标签，如道路、行人、车辆和建筑物。

为了确保公平比较，本文严格遵循之前的研究（Jaritz等，2020；Shin等，2022），在3个域适应场景中评估了本文方法。具体而言，A2D2-SemanticKITTI用于评估跨数据集的适应能力，重点考察两个数据集在传感器配置、点云密度以及采集精度上的差异，模型需要能够跨越这些差异进行有效的迁移学习。A2D2中的点云较为稀疏，而SemanticKITTI的点云则密集且具有高精度，因此，模型不仅需要适应点云稀疏与密集之间的差异，还需要有效处理传感器视角和分辨率带来的变化。nuScenes数据集被划分为两个设置：美国—新加坡，用于测试不同区域之间的地理适应能力，在这种场景下模型需要适应不同的城市布局、道路结构；白天—夜晚，用于评估模型在

光照变化下的适应性，在这种设置下，光照变化对图像的质量影响较大，白天与夜晚的光照强度、阴影以及反射差异都会导致图像信息的变化。

3.2 评估指标

本文使用平均交并比（mean intersection over union，mIoU）评估模型性能。mIoU 通过综合考虑每个类别的交集与并集的比值，反映了模型在每个类别上的预测精度。mIoU 计算为

$$f_{\mathrm{mIoU}}=\frac{1}{K}\sum_{k=1}^{K}\frac{TP_k}{TP_k+FP_k+FN_k} \tag{10}$$

式中，K 表示类别总数，TP_k（true positive）、FP_k（false positive）和 FN_k（false negative）分别为类别 k 的真阳性、假阳性和假阴性的数量。mIoU 能够全面衡量模型在多个类别上的预测能力，且通过平均各类别的 IoU 值，有效避免了类别不平衡对评估结果的偏差。本文在评估过程中，分别计算了二维模型和三维模型的 mIoU，并融合二维模型和三维模型的预测结果计算综合 mIoU（Avg）。

3.3 实验细节

本文采用基于 ResNet34（residual network）（He 等，2016）的 U-Net（Ronneberger 等，2015）和 Minkowski U-Net 分别作为二维和三维分割的骨干网络。这些模型能够提取特征，并输出逐像素或逐点对每个类别的预测得分。在 3 个设置上，利用真实标签全监督训练得到多模态的源域模型。然后，使用源域模型的参数初始化目标域对应模态的快慢模型。在测试时适应过程中，使用损失函数对快模型的 BN 层进行更新，然后使用 EMA 策略对慢模型进行更新。

在源域训练和目标域适应阶段，采用了相同的优化器设置。具体而言，二维模型使用 Adam 优化器，基础学习率设置为 0.001，β 配置为（0.99，0.999）。三维模型使用随机梯度下降法（stochastic gradient descent，SGD）优化器，基础学习率为 0.24，动量值为 0.9。此外，在训练过程中使用 MultiStepLR 调度器动态调整学习率，在每个指定的里程碑时将学习率降低至原本的 10%，不同实验设置下的里程碑有所不同。所有实验都使用一张 NVIDIA GeForce RTX 3090 GPU 上进行。

3.4 实验结果分析

本文在 3.1 节所提出的 3 个实验设置上与现有的测试时域自适应方法（TTA）进行了对比，包括 TENT（Wang 等，2020）和 MM-TTA（Shin 等，2022）。此外也与无监督域自适应方法（UDA）进行了对比，包括 xMUDA（cross-modal unsupervised domain adaptation）（Jaritz 等，2020）、DsCML（dynamic sparse-to-dense cross modal learning）（Peng 等，2021）和 CMCL（cross-modal contrastive learning）（Xing 等，2023）。实验结果如表 1 所示。

1）A2D2-SemanticKITTI。该设置主要考察的是在不同数据集之间，尤其是传感器配置不同的情况下模型的适应能力。实验结果表明，本文提出的方法在数据集之间适应性效果取得了显著提升。相较于最先进的 MM-TTA 方法，在二维上的 mIoU 提升了 4.6%，达到了 48.3%；在三维性能上达到了 42.9%；综合三维和二维预测结果后，mIoU 达到了 51.0%。此外，本文提出的测试时自适应方法超过了部分无监督域自适应方法。这对于实际场景中在不同数据集之间进行模型适应具有重要意义。得益于大模型中丰富的先验信息，模型能够在点云间的域差异较大的情况下，利用稳定的二维信息引导模型在目标域上展现出更好的分割效果。

2）美国—新加坡。此设置主要考察不同地区之间道路分布等差异下模型适应的能力。结果表明，本文方法在各项指标上均取得了最佳性能，二维上的 mIoU 达到了 51.5%，三维相比 TENT 提高 6.2%，达到了 52.9%，综合三维和二维的预测结果，mIoU 达到了 60.7%。与无监督域自适应方法相比，本文方法提升效果有限，这是因为在同一数据集中地区不同情况下，域间差异较小，使用源域数据相较于大模型中的信息能够更有效地帮助模型提取出域间不变的特征，从而提升模型在目标域上的性能。尽管如此，本文方法效果也能接近早期的无监督域自适应方法，在无法获取源域数据的情况下依然有较高的实用价值。

3）白天—夜晚。此设置主要考察在不同光照条件下模型的适应能力。实验结果表明，本文方法二维上的 mIoU 达到了 45.1%，三维上的 mIoU 达到了 43.3%，综合二维与三维的预测结果 mIoU 达到了 51.9%。与先进的 MM-TTA 相比，本文方法在两项指标上取得了更优的性能。此外，从表 1 中还可以观察到 TENT 方法在二维任务上的表现甚至不如仅源域的结果。这表明在昼夜变化等光照差异较大的情况下，图像的视觉表征会发生显著变化，而 TENT

表1　在3种不同的域自适应设置下本文方法与其他UDA和TTA方法的实验结果对比(mIoU)

Table 1　Comparison of the our method with other UDA and TTA methods under three different domain adaptation settings(mIoU)

/%

方法	类型	A2D2-SemanticKITTI			美国—新加坡			白天—夜晚		
		2D	3D	综合	2D	3D	综合	2D	3D	综合
仅源域	–	32.5	40.4	42.0	49.7	48.0	56.7	40.0	37.6	48.2
xMDUA(Jaritz等,2020)	UDA	36.8	43.3	42.9	59.3	52.0	56.7	46.2	44.2	50.0
DsCML(Peng等,2021)	UDA	39.6	45.1	44.5	61.3	53.3	63.6	48.0	45.7	51.0
DsCML + CMAL + PL(Peng等,2021)	UDA	46.8	51.8	52.4	63.9	56.3	65.1	50.1	48.7	53.0
CMCL(Xing等,2023)	UDA	42.8	47.7	48.0	62.0	54.2	64.5	49.0	46.6	51.6
CMCL + PL(Xing等,2023)	UDA	47.6	51.3	52.8	63.3	57.1	66.7	50.6	49.9	54.5
TENT(Wang等,2020)	TTA	38.9	35.6	39.5	50.4	46.7	58.9	36.5	42.2	42.0
MM-TTA(Hard)(Shin等,2022)	TTA	43.3	42.4	47.0	–	–	–	42.6	43.6	51.1
MM-TTA(Soft)(Shin等,2022)	TTA	43.7	42.5	47.1	–	–	–	44.2	**43.7**	51.8
本文	TTA	**48.3**	**42.9**	**51.0**	**51.5**	**52.9**	**60.7**	**45.1**	43.3	**51.9**

注:加粗、下划线字体分别表示各列TTA方法的最优、次优结果。“–”表示对应文献中未对该项指标提供相关信息。

方法主要依赖统计信息进行领域适应,但未能有效缓解光照差异带来的挑战。相比之下,本文方法在二维任务上的mIoU得到了显著提升,这表明本文方法通过模态融合和所提出的两个模块能够充分利用大模型的信息,帮助模型更好地适应不同光照条件下的视觉变化,显著提升了二维任务的表现。然而,在三维任务中,由于点云数据的表示主要是通过雷达反射而形成,对光照变化的依赖较小,因此三维的表现并未因为模型对光照变化的适应能力变强而有显著改善。

图5展示了在3组实验设置中的可视化结果。从图中可以看出,本文提出的方法能够有效纠正错误的语义预测,显著提升模型的准确性,表现出更强的语义泛化能力。此外,对于一些语义不连续的区域,本文方法也能够有效改善这一现象,表明了本文方法在增强特征的语义连续性方面的优势,显著提高了分割效果。

3.5　消融实验

3.5.1　各模块有效性验证

为了验证本文提出的基于CLIP的文本嵌入分割层和基于SAM掩码的特征一致性约束模块的有效性,本文在3个实验设置下(即A2D2-SemanticKITTI、美国—新加坡、白天—黑夜)进行了消融研究。消融实验的结果如表2所示。在3组实验中可以明显看到,文本嵌入分割层可以显著提高二维语义分割性能,在3个实验设置下分别从45.8%、49.7%和43.2%提升到48.1%、52.6%和44.8%,这表明基于CLIP的文本嵌入分割层在图像上可以显著增强模型的语义泛化能力,使模型对二维信息做出更准确的预测。当模型仅使用特征一致性约束模块时,三维分割性能较二维有更显著的提升,分别从40.2%、48.0%和38.8%提升到41.8%、49.8%和42.8%,尤其是在白天—夜晚最为显著,这表明基于SAM的特征一致性约束可以增强局部语义信息的连续性,使模型在进行三维预测时能够做出更准确的判断。当两个模块结合时,在A2D2-SemanticKITTI和白天—夜晚场景下都取得了最佳的性能,而在美国—新加坡这个设置下三维预测结果取得了最佳性能,这说明两种方法能够利用各自的优势对最终的性能各有贡献。这些实验验证了本文提出的两个模块的有效性。

3.5.2　EMA更新动量参数分析

为了验证本文提出的模块在异步更新架构下进行目标域适应时的稳定性,本文研究了不同动量更新参数对模型性能的影响,如图6所示。从结果可以看出,当动量因子为0时,模型的性能有明显的下

图 5　本文方法在 3 种域自适应设置下的实验结果可视化

Fig. 5　Experimental result visualization of our method under three domain adaptation setting

表 2　文本嵌入分割层和特征一致性约束模块的有效性消融实验结果(mIoU)

Table 2　Ablation study results of text-embedded segmentation layer and feature consistency constraint module(mIoU)

/%

文本嵌入分割层	特征一致性约束模块	A2D2-SemanticKITTI			美国—新加坡			白天—夜晚		
		2D	3D	综合	2D	3D	综合	2D	3D	综合
×	×	45.8	40.2	46.1	49.7	48.0	56.7	43.2	38.8	49.4
√	×	48.1	42.2	50.9	**52.6**	49.2	60.4	44.8	39.1	49.8
×	√	46.2	41.8	49.0	51.8	49.8	**61.1**	43.7	42.8	51.7
√	√	**48.3**	**42.9**	**51.0**	51.5	**52.9**	60.7	**45.1**	**43.3**	**51.9**

注：加粗字体表示各列最优结果。“√”和“×”分别表示使用和未使用。

降，这体现了 EMA 更新策略的有效性。此外，当动量因子不为 0 时，模型性能在小范围内波动，表明本文所提出的模块能够在异步更新架构中保持鲁棒性。

3.5.3　分割层文本提示分析

为了研究不同文本提示对模型性能的影响，本文使用了两种不同的文本提示设置：一种是三维和二维模型均用“A point of $[\text{class}]_k$”作为文本提示；另一种是三维和二维模型分别使用“A point of $[\text{class}]_k$”和“A pixel of $[\text{class}]_k$”作为文本提示。本文使用上述两种文本提示，在 3 个实验设置下进行语义预测，实验结果如表 3 所示。从表中的实验结果可以看到，使用不同的文本提示组合对模型的性能影响并不大，这可能是因为 CLIP 模型本身能够较好地聚焦

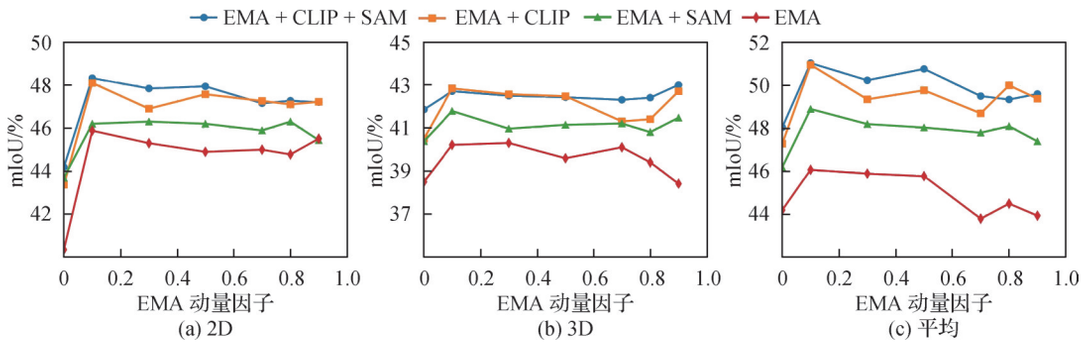

图6　快慢模型架构中EMA更新动量参数分析

Fig. 6　Analysis of EMA momentum parameter in fast-slow model architecture ((a) 2D; (b) 3D; (c) average)

表3　不同分割层文本提示分析(mIoU)

Table 3　Analysis of different text prompts in segmentation layers (mIoU)

/%

二维文本提示	三维文本提示	A2D2-SemanticKITTI			美国—新加坡			白天—夜晚		
		2D	3D	平均	2D	3D	平均	2D	3D	平均
A point of $[class]_k$	A point of $[class]_k$	48.3	42.9	51.0	51.5	52.9	60.7	45.1	43.3	51.9
A pixel of $[class]_k$	A point of $[class]_k$	48.2	42.9	51.0	51.6	53.0	60.7	44.8	43.3	51.8

于文本中真正的语义类别词汇信息。此外,在二维视觉表达上“point”和“pixel”的差距较小,所以在CLIP模型的特征空间中,这两种文本提示所产生的文本特征差异也较小,因此在这两种文本提示下的分割性能并无显著差异。

4　结　论

本文提出一种大模型驱动的测试时自适应方法,用于增强跨领域的语义分割能力。该方法包括一个基于CLIP的文本嵌入分割层,为模型提供显式的语义信息,增强模型在面对域偏移时的语义泛化能力;此外,该方法还包括一个基于SAM掩码的特征一致性约束模块,利用大模型生成的高质量掩码约束局部特征一致性,解决模型在域偏移情况下出现的语义碎片化问题,从而提高特征的泛化能力。这两个模块有效地解决了域偏移带来的挑战。通过在多个基准数据集上进行实验,验证了该方法的优越性。此外,本文还进行了消融研究,以验证每个模块对整体性能提升的贡献。在未来的工作中,将使用更先进的大模型,并进一步研究提取大模型信息的策略,从而帮助模型能够在有域偏移的场景下获得更好的分割性能。

参考文献(References)

Achituve I, Maron H and Chechik G. 2021. Self-supervised learning for domain adaptation on point clouds//Proceedings of 2021 IEEE Winter Conference on Applications of Computer Vision. Waikoloa, USA: IEEE: 123-133 [DOI: 10.1109/WACV48630.2021.00017]

Behley J, Garbade M, Milioto A, Quenzel J, Behnke S, Stachniss C and Gall J. 2019. SemanticKITTI: a dataset for semantic scene understanding of LiDAR sequences//Proceedings of 2019 IEEE/CVF International Conference on Computer Vision. Seoul Korea (South): IEEE: 9296-9306 [DOI: 10.1109/ICCV.2019.00939]

Bian Y K, Hui L, Qian J J and Xie J. 2022. Unsupervised domain adaptation for point cloud semantic segmentation via graph matching//Proceedings of 2022 IEEE/RSJ International Conference on Intelligent Robots and Systems. Kyoto, Japan: IEEE: 9899-9904 [DOI: 10.1109/IROS47612.2022.9981603]

Caesar H, Bankiti V, Lang A H, Vora S, Liong V E, Xu Q, Krishnan A, Pan Y, Baldan G and Beijbom O. 2020. nuScenes: a multimodal dataset for autonomous driving//Proceedings of 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle, USA: IEEE: 11618-11628 [DOI: 10.1109/CVPR42600.2020.01164]

Cao H Z, Xu Y C, Yang J F, Yin P Y, Yuan S H and Xie L H. 2023. Multi-modal continual test-time adaptation for 3D semantic segmentation//Proceedings of 2023 IEEE/CVF International Conference on Computer Vision. Paris, France: IEEE: 18763-18773 [DOI: 10.

1109/ICCV51070.2023.01724］

Choy C，Gwak J and Savarese S. 2019. 4D spatio-temporal ConvNets：minkowski convolutional neural networks//Proceedings of 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Long Beach，USA：IEEE：3070-3079［DOI：10.1109/CVPR.2019.00319］

Geyer J，Kassahun Y，Mahmudi M，Ricou X，Durgesh R，Chung A S，Hauswald L，Pham V H，Mühlegg M，Dorn S，Fernandez T，Jänicke M，Mirashi S，Savani C，Sturm M，Vorobiov O，Oelker M，Garreis S and Schuberth P. 2020. A2D2：Audi autonomous driving dataset［EB/OL］.［2024-12-27］. https://arxiv.org/pdf/2004.06320.pdf

He K M，Zhang X Y，Ren S Q and Sun J. 2016. Deep residual learning for image recognition//Proceedings of 2016 IEEE Conference on Computer Vision and Pattern Recognition. Las Vegas，USA：IEEE：770-778［DOI：10.1109/CVPR.2016.90］

Hu Q Y，Yang B，Xie L H，Rosa S，Guo Y L，Wang Z H，Trigoni N and Markham A. 2020. RandLA-Net：efficient semantic segmentation of large-scale point clouds//Proceedings of 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle，USA：IEEE：11105-11114［DOI：10.1109/CVPR42600.2020.01112］

Jaritz M，Gu J Y and Su H. 2019. Multi-view PointNet for 3D scene understanding//Proceedings of 2019 IEEE/CVF International Conference on Computer Vision Workshops. Seoul，Korea（South）：IEEE：3995-4003［DOI：10.1109/ICCVW.2019.00494］

Jaritz M，Vu T H，De Charette R，Wirbel E and Pérez P. 2020. xMUDA：cross-modal unsupervised domain adaptation for 3D semantic segmentation//Proceedings of 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle，USA：IEEE：12602-12611［DOI：10.1109/CVPR42600.2020.01262］

Jia C，Yang Y F，Xia Y，Chen Y T，Parekh Z，Pham H，Le Q，Sung Y H，Li Z and Duerig T. 2021. Scaling up visual and vision-language representation learning with noisy text supervision//Proceedings of the 38th International Conference on Machine Learning. Vienna，Austria：PMLR：4904-4916

Kirillov A，Mintun E，Ravi N，Mao H Z，Rolland C，Gustafson L，Xiao T T，Whitehead S，Berg A C，Lo W Y，Dollár P and Girshick R. 2023. Segment anything//Proceedings of 2023 IEEE/CVF International Conference on Computer Vision. Paris，France：IEEE：3992-4003［DOI：10.1109/ICCV51070.2023.00371］

Li J N，Li D X，Xiong C M and Hoi S. 2022. BLIP：bootstrapping language-image pre-training for unified vision-language understanding and generation//Proceedings of the 39th International Conference on Machine Learning. Baltimore，USA：PMLR：12888-12900

Maturana D and Scherer S. 2015. VoxNet：a 3D convolutional neural network for real-time object recognition//Proceedings of 2015 IEEE/RSJ International Conference on Intelligent Robots and Systems. Hamburg，Germany：IEEE：922-928［DOI：10.1109/IROS.2015.7353481］

Ni J Y，Yang S Q，Xu R，Liu J M，Li X Q，Jiao W Y，Chen Z H，Liu Y and Zhang S H. 2024. Distribution-aware continual test-time adaptation for semantic segmentation//Proceedings of 2024 IEEE International Conference on Robotics and Automation. Yokohama，Japan：IEEE：3044-3050［DOI：10.1109/ICRA57147.2024.10610045］

Peng D，Lei Y J，Li W，Zhang P P and Guo Y L. 2021. Sparse-to-dense feature matching：intra and inter domain cross-modal learning in domain adaptation for 3D semantic segmentation//Proceedings of 2021 IEEE/CVF International Conference on Computer Vision. Montreal，Canada：IEEE：7088-7097［DOI：10.1109/ICCV48922.2021.00702］

Qi C R，Su H，Kaichun M and Guibas L J. 2017a. PointNet：deep learning on point sets for 3D classification and segmentation//Proceedings of 2017 IEEE Conference on Computer Vision and Pattern Recognition. Honolulu，USA：IEEE：77-85［DOI：10.1109/CVPR.2017.16］

Qi C R，Yi L，Su H and Guibas L J. 2017b. PointNet++：deep hierarchical feature learning on point sets in a metric space//Proceedings of the 31st International Conference on Neural Information Processing Systems. Long Beach，USA：Curran Associates Inc.：5105-5114

Radford A，Kim J W，Hallacy C，Ramesh A，Goh G，Agarwal S，Sastry G，Askell A，Mishkin P，Clark J，Krueger G and Sutskever I. 2021. Learning transferable visual models from natural language supervision//Proceedings of the 38th International Conference on Machine Learning.［s.l.］：PMLR：8748-8763

Ravi N，Gabeur V，Hu Y T，Hu R H，Ryali C，Ma T，Khedr H，Rädle R，Rolland C，Gustafson L，Mintun E，Pan J，Alwala K V，Carion N，Wu C Y，Girshick R，Dollár P and Feichtenhofer C. 2024. SAM 2：segment anything in images and videos［EB/OL］.［2024-12-27］. https://arxiv.org/pdf/2408.00714.pdf

Ren T H，Liu S L，Zeng A L，Lin J，Li K C，Cao H，Chen J Y，Huang X Y，Chen Y K，Yan F，Zeng Z，Zhang H，Li F，Yang J，Li H，Jiang Q and Zhang L. 2024. Grounded SAM：assembling open-world models for diverse visual tasks［EB/OL］.［2024-12-27］. https://arxiv.org/pdf/2401.14159.pdf

Ronneberger O，Fischer P and Brox T. 2015. U-Net：convolutional networks for biomedical image segmentation//Proceedings of the 18th International Conference on Medical Image Computing and Computer-Assisted Intervention. Munich，Germany：Springer：234-241［DOI：10.1007/978-3-319-24574-4_28］

Ruan X Q and Tang W. 2024. Fully test-time adaptation for object detection//Proceedings of 2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition. New Orleans，USA：IEEE：1038-1047［DOI：10.1109/CVPRW63382.2024.00110］

Saleh K，Abobakr A，Attia M，Iskander J，Nahavandi D，Hossny M and Nahvandi S. 2019. Domain adaptation for vehicle detection from bird's eye view LiDAR point cloud data//Proceedings of 2019

IEEE/CVF International Conference on Computer Vision Workshops. Seoul, Korea(South): IEEE: 3235-3242 [DOI: 10.1109/ICCVW.2019.00404]

Shin I, Tsai Y H, Zhuang B B, Schulter S, Liu B Y, Garg S, Kweon I S and Yoon K J. 2022. MM-TTA: multi-modal test-time adaptation for 3D semantic segmentation//Proceedings of 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition. New Orleans, USA: IEEE: 16907-16916 [DOI: 10.1109/CVPR52688.2022.01642]

Su H, Maji S, Kalogerakis E and Learned-Miller E. 2015. Multi-view convolutional neural networks for 3D shape recognition//Proceedings of 2015 IEEE International Conference on Computer Vision. Santiago, Chile: IEEE: 945-953 [DOI: 10.1109/ICCV.2015.114]

Sun L J, Zeng T F, Fan J X and Wang W J. 2024. Double-view feature fusion network for LiDAR semantic segmentation. Journal of Image and Graphics, 29(1): 205-217 (孙刘杰, 曾腾飞, 樊景星, 王文举. 2024. 大场景双视角点云特征融合语义分割方法. 中国图象图形学报, 29(1): 205-217) [DOI: 10.11834/jig.220943]

Tarvainen A and Valpola H. 2017. Mean teachers are better role models: weight-averaged consistency targets improve semi-supervised deep learning results//Proceedings of the 31st International Conference on Neural Information Processing Systems. Long Beach, USA: Curran Associates Inc.: 1195-1204

Vu T A, Sarkar S, Zhang Z Y, Hua B S and Yeung S K. 2024. Test-time augmentation for 3D point cloud classification and segmentation//Proceedings of 2024 International Conference on 3D Vision. Davos, Switzerland: IEEE: 1543-1553 [DOI: 10.1109/3DV62453.2024.00153]

Wang D Q, Shelhamer E, Liu S T, Olshausen B A and Darrell T. 2020. Tent: fully test-time adaptation by entropy minimization//Proceedings of the 9th International Conference on Learning Representations. Vienna, Austria: OpenReview.net

Weijler L, Mirza M J, Sick L, Ekkazan C and Hermosilla P. 2024. TTT-KD: test-time training for 3D semantic segmentation through knowledge distillation from foundation models [EB/OL]. [2024-12-27]. https://arxiv.org/pdf/2403.11691.pdf

Wu B C, Zhou X Y, Zhao S C, Yue X Y and Keutzer K. 2019. SqueezeSegV2: improved model structure and unsupervised domain adaptation for road-object segmentation from a LiDAR point cloud//Proceedings of 2019 IEEE International Conference on Robotics and Automation. Montreal, Canada: IEEE: 4376-4382 [DOI: 10.1109/ICRA.2019.8793495]

Xing B W, Ying X H, Wang R B, Yang J F and Chen T Y. 2023. Cross-modal contrastive learning for domain adaptation in 3D semantic segmentation//Proceedings of the 37th AAAI Conference on Artificial Intelligence. Washington, USA: AAAI Press: 2974-2982 [DOI: 10.1609/aaai.v37i3.25400]

Yi L, Gong B Q and Funkhouser T. 2021. Complete and label: a domain adaptation approach to semantic segmentation of LiDAR point clouds//Proceedings of 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Nashville, USA: IEEE: 15358-15368 [DOI: 10.1109/CVPR46437.2021.01511]

Yuan L, Chen D D, Chen Y L, Codella N, Dai X, Gao J, Hu H, Huang X, Li B, Li C Y, Liu C, Liu M C, Liu Z C, Lu Y M, Shi Y, Wang L J, Xiao B, Xiao Z, Yang J W, Zeng M, Zhou L W and Zhang P C. 2021. Florence: a new foundation model for computer vision [EB/OL]. [2024-12-27]. https://arxiv.org/pdf/2111.11432.pdf

Zhang G M, Pan G F and Liu J X. 2020. Domain adaptation for semantic segmentation based on adaption learning rate. Journal of Image and Graphics, 25(5): 913-925 (张桂梅, 潘国峰, 刘建新. 2020. 域自适应城市场景语义分割. 中国图象图形学报, 25(5): 913-925) [DOI: 10.11834/jig.190424]

Zhang Y F, Wang X, Jin K X, Yuan K, Zhang Z, Wang L, Jin R and Tan T N. 2023. AdaNPC: exploring non-parametric classifier for test-time adaptation//Proceedings of the 40th International Conference on Machine Learning. Honolulu, USA: PMLR: 41647-41676

Zhuang Z W, Li R, Jia K, Wang Q C, Li Y Q and Tan M K. 2021. Perception-aware multi-sensor fusion for 3D LiDAR semantic segmentation//Proceedings of 2021 IEEE/CVF International Conference on Computer Vision. Montreal, Canada: IEEE: 16260-16270 [DOI: 10.1109/ICCV48922.2021.01597]

作者简介

刘雪帆,女,硕士研究生,主要研究方向为点云场景理解。
E-mail:liuxf83@mial2.sysu.edu.cn
郭裕兰,通信作者,男,教授,主要研究方向为三维视觉与模式识别。E-mail:guoyulan@sysu.edu.cn
刘砚,男,博士研究生,主要研究方向为点云场景理解。
E-mail:liuy2297@mail2.sysu.edu.cn
李浩然,男,博士研究生,主要研究方向为点云压缩。
E-mail:lihr67@mail2.sysu.edu.cn
张晔,男,博士后,主要研究方向为人机交互。
E-mail:zhangy2658@sysu.edu.cn