This CVPR Findings paper is the Open Access version, provided by the Computer Vision Foundation. Except for this watermark, it is identical to the accepted version; the final published version of the proceedings is available on IEEE Xplore.

UniD-Shift: Towards Unified Semantic Segmentation via Interpretable Shared–Private Multimodal Decomposition

Shuai Zhang[1], Zhecheng Shi[1], Zhuoxiao Li[1], Jing Ou[1], Tengxi Wang[1], Yuan Liu[2], Wufan Zhao[1*]
[1]HKUST(GZ), [2]HKUST
{szhang240, jou719, twang744}@connect.hkust-gz.edu.cn, yuanly@ust.hk
{zhechengs, zhuoxiaoli, wufanzhao}@hkust-gz.edu.cn

Abstract

Semantic segmentation of large-scale 3D point clouds is crucial for applications such as autonomous driving and urban digital twins. However, the sparse sampling pattern of LiDAR and the view-dependent geometric distortion in image observations complicate cross-modal alignment and hinder stable fusion. Inspired by the fact that 2D images captured by cameras are representations of the 3D world, we recognize that the features learned from 2D and 3D segmentation share some common semantics, while other aspects remain modality-specific. This insight motivates a unified multimodal framework for joint 2D-3D semantic segmentation. We combine a SAM-based vision encoder with a SPTNet-based geometric encoder to extract complementary semantic and geometric representations. The resulting features from both modalities are explicitly decomposed into shared and private subspaces, where the shared components summarize semantic factors common to both domains, and the private components preserve properties that are unique to each modality. A lightweight attention-based fusion module aggregates the shared features into a consistent cross-modal representation, and a regularized training objective ensures both semantic alignment and subspace independence. Experiments on the SemanticKITTI and nuScenes benchmarks demonstrate consistent improvements in segmentation accuracy over representative multimodal baselines, accompanied by competitive computational efficiency. Cross-domain evaluation on nuScenes USA-Singapore shows stable performance under distribution shifts, demonstrating strong generalization. The implementation code is publicly available at: https://github.com/shuaizhang69/UniD-Shift.

1. Introduction

Semantic segmentation of large-scale 3D point clouds is a central problem in intelligent perception and supports key applications in autonomous driving, urban reconstruction, and digital-twin modeling [15, 23, 44, 70, 74]. Assigning semantic labels to individual points produces a structured representation of urban environments and enables reliable analysis of complex outdoor scenes. Current sensing systems often combine LiDAR measurements with RGB images to capture both geometry and appearance. LiDAR provides accurate spatial structure, whereas images offer dense visual information that reflects surface color and texture. Although these observations originate from different sensing principles, they describe the same physical world and share consistent semantic content such as object categories and spatial relations. Their complementary characteristics indicate that integrating 2D and 3D observations is an effec-

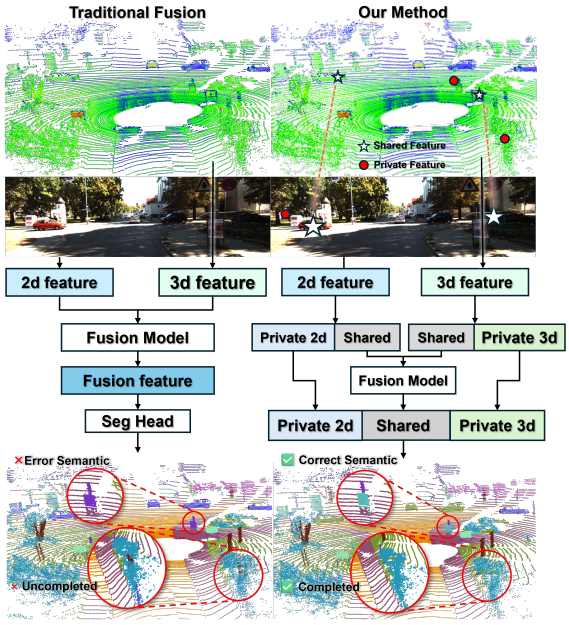


Figure 1. Existing multimodal fusion strategies (left) often mix 2D and 3D features without explicit semantic disentanglement, leading to redundant and ambiguous representations. Our proposed shared-private fusion framework (right) separates modality-invariant and modality-specific features, enabling efficient and interpretable 2D-3D interaction.

tive strategy for improving the robustness and completeness of scene understanding.

Recent progress in multimodal scene understanding has significantly advanced 3D semantic segmentation within single-domain scenarios. As illustrated in Fig. 1, existing 2D-3D fusion methods consistently demonstrate that combining geometric structure from LiDAR with appearance information from images improves feature quality and produces clearer semantic boundaries [20, 36, 37, 60]. Fusion architectures that incorporate cross-modal alignment and joint reasoning further enhance the consistency of geometric and visual representations.Despite these advances, many approaches rely on heavy feature concatenation or large attention modules, increasing computational cost and limiting scalability in large outdoor scenes. Their reliance on full-feature fusion often introduces redundant representations and leads to unstable optimization during training. These limitations reveal a persistent gap between accurate multimodal integration and stable model behavior, indicating the need for a fusion strategy that maintains efficiency while preserving interpretability.

To address these limitations, we introduce a shared-private feature decomposition framework called **UniD-Shift**. The central assumption is that 2D and 3D modalities contain both common attributes reflecting consistent semantics and unique attributes encoding modality-dependent characteristics. Effective multimodal learning therefore requires integrating shared attributes while preserving distinctive modality-specific information. To achieve this, the proposed decomposition separates modality-invariant representations that describe stable semantic concepts from modality-specific representations that retain appearance variation in images and geometric detail in point clouds. This separation aligns with the inherent distinction between image-based appearance and LiDAR geometry. By modeling shared and modality-specific information separately, the framework promotes consistent semantic correspondence while reducing interference from irrelevant modality-dependent variations. The resulting representation is more compact and stable, improving training stability and cross-domain generalization.

Overall, our contributions are summarized as follows:

- A unified multimodal framework is proposed to achieve both high single-domain accuracy and strong cross-domain generalization through feature interaction.
- A shared-private decomposition mechanism enables explicit disentanglement of modality-invariant and modality-specific representations, improving semantic alignment and interpretability.
- A dual-branch design integrating a SAM-based vision encoder with a sparse convolution-transformer backbone effectively combines semantic richness and geometric precision.
- Extensive experiments on nuScenes and SemanticKITTI demonstrate superior segmentation accuracy and robustness compared with existing 2D-3D fusion and domain adaptation methods.

2. Related work

2.1. LiDAR-based 3D Semantic Segmentation

Existing methods for LiDAR-based 3D semantic segmentation are generally categorized into point-based [42, 43, 49], projection-based [12, 16, 32, 46, 63], and voxel-based frameworks [10, 14, 17, 21, 26, 38, 45, 62, 77] . Point-based models directly process irregular point sets to preserve fine geometric structures, yet their computational overhead limits scalability in large outdoor scenes. Projection-based approaches transform point clouds into structured range representations, enabling efficient 2D convolutional or transformer-based processing but introducing spatial distortions and information loss in occluded regions. Voxel-based methods discretize the 3D space into regular grids and use sparse convolutions to model volumetric structure. This strategy offers a balanced compromise between accuracy and efficiency, but the fixed resolution restricts the representation of small objects. More recently, transformer-based architectures [24, 53, 55, 64, 65, 73] have improved long-range dependency modeling, but their quadratic complexity constrains large-scale deployment. Overall, the inherent sparsity and limited semantic abstraction of LiDAR data motivate the integration of complementary visual information to enhance contextual understanding.

2.2. Multimodal 2D-3D Fusion for 3D Semantic Segmentation

Fusing LiDAR and RGB improves 3D segmentation by coupling geometric precision with visual semantics [29, 31, 40, 50, 60, 61]. Early dual-stream systems project points to the image plane and merge features via residual or distillation modules, whereas later approaches jointly optimize intra-modal learning and inter-modal interaction to enhance complementarity [1, 2, 79]. However, most existing methods implicitly assume that features from different modalities can be directly aligned without explicitly modeling their shared and modality-specific components. This often leads to suboptimal fusion, where complementary information is not fully utilized and modality-dependent noise may be introduced. Hierarchical and semantic-guided strategies refine structural coherence through adaptive weighting, and vision-language encoders provide contextual priors and open-vocabulary capability [35, 57]. Despite these advances, many pipelines still depend on dense attention or direct concatenation, which raises computational cost and weakens interpretability. Performance degradation under illumination, weather, or sensor shifts further underscores the

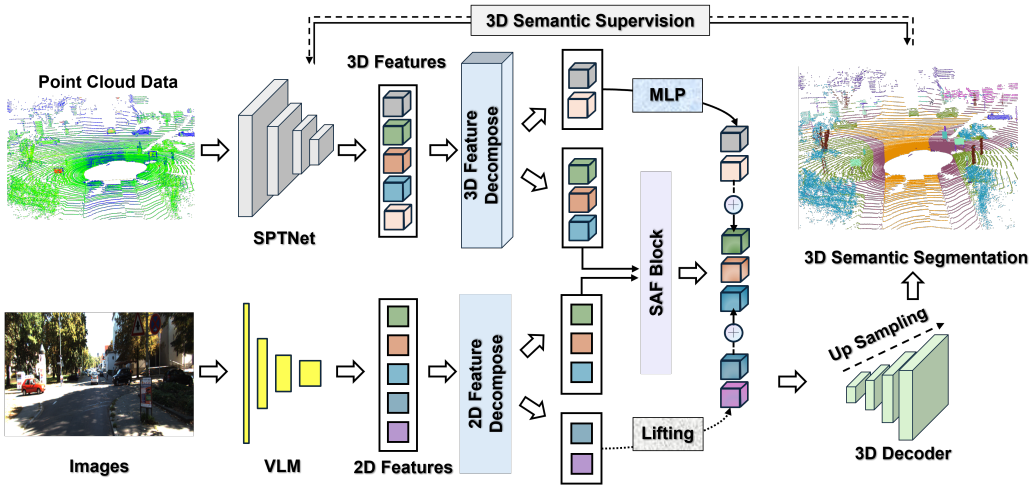


Figure 2. Overall architecture of the proposed **UniD-Shift**. The model takes synchronized 2D images and 3D point clouds as inputs. The 3D branch employs a SPTNet backbone to extract hierarchical geometric features, while the 2D branch utilizes a SAM-based encoder to obtain semantically enriched visual representations. Both modalities are decomposed into shared and private components, then feed into the Shared Attention Fusion (SAF) block into a unified latent representation.

need for domain-adaptive multimodal fusion.

2.3. Domain Adaptation and Generalization in 3D Semantic Segmentation

Domain adaptation and generalization address distribution shifts in multimodal 3D perception by transferring from labeled sources to unlabeled or unseen targets. Representative methods promote cross-modal consistency via bidirectional learning or align class prototypes through adversarial objectives, while fusion-distillation and style-transfer strategies aim to improve target robustness [6, 7, 19, 20, 30, 33, 34, 41, 56–58, 68, 72]. Recent unified representations reduce modality and domain gaps with structured quantization and latent regularization. Nevertheless, many frameworks still treat fusion and adaptation separately, which limits semantic consistency across modalities and stability across domains; a unified formulation remains necessary.

3. Methodology

3.1. Framework Overview

The proposed framework is outlined in Fig. 2. It adopts a dual-branch structure for 2D images and 3D point clouds, enabling reliable fusion and consistent generalization. The 3D branch uses sparse convolution and transformer layers [69] to extract geometric structure, while the 2D branch applies a SAM-based encoder [22] to produce semantically coherent visual features. The resulting representations are projected into a shared space and decomposed into shared and private components. Shared features represent the semantic content that is consistent across modalities and are fused through the Shared Attention Fusion block. Private features retain the modality-dependent information needed to complement the shared representation. The fused output is decoded into detailed segmentation predictions, and the training objectives guide the network toward better alignment, clearer feature separation, and improved accuracy.

3.2. Dual-Branch Feature Extraction

To capture both fine-grained geometric structures and high-level semantic information, we design a dual-branch feature extraction architecture that jointly leverages 2D imagery and 3D point clouds.The two modalities are inherently complementary, as LiDAR provides precise spatial geometry whereas RGB imagery contributes dense contextual semantics. To achieve a robust multimodal representation, each branch is carefully designed to preserve its intrinsic advantages while maintaining consistent cross-modal alignment.

For the image branch, we adopt the Segment Anything Model (SAM) as the vision encoder due to its strong generalization ability and proven effectiveness in segmentation-oriented visual tasks. SAM has demonstrated exceptional capability in capturing object boundaries and semantic coherence across diverse datasets [27, 52, 57]. By integrating the hierarchical visual tokens produced by the Segment Anything Model (SAM), the framework extracts multi-scale features that encode both local structural details and global semantic relationships. The resulting features are projected into a shared embedding space through a lightweight transformation layer, ensuring semantic consistency with the geometric features extracted from the 3D branch.

For the point cloud branch, we employ a Sparse Convolution-Transformer hybrid network (SPTNet), which efficiently captures fine-grained geometric variations through sparse convolutions and models long-range dependencies via transformer-based self-attention. This design

allows the network to encode both local and contextual spatial relations, generating structurally complete and semantically aligned 3D representations. The features produced by both branches are normalized into a common latent space, forming the basis for subsequent shared-private feature decomposition and semantic fusion.

3.3. Shared-Private Feature Decomposition and Semantic Fusion

To structure the multimodal interaction and reduce redundancy during fusion, each modality is designed to be decomposed into a shared component that captures modality-invariant semantics and a private component that preserves modality-specific characteristics. This design is formalized by representing each encoded feature as the sum of its shared and private components,

$$\mathbf{F}^m = \mathbf{S}^m + \mathbf{R}^m, \quad m \in \{2\mathrm{D}, 3\mathrm{D}\}, \tag{1}$$

where $\mathbf{S}^m$ encodes semantic information shared across modalities, and $\mathbf{R}^m$ preserves modality-specific attributes rooted in either geometric structure or visual appearance. The shared and private components are obtained through learnable linear projections and normalization layers, providing a structured latent space for subsequent fusion.

To ensure that the decomposition is meaningfully achieved during training, two regularization constraints are introduced to guide the separation of shared and private subspaces. The first constraint encourages the shared components to encode consistent semantic structure across modalities by aligning their second-order statistics:

$$\mathcal{L}_{\text{gram}} = \frac{1}{C_s^2} \left\| \mathbf{S}_{512}^{2\mathrm{D}\,\top} \mathbf{S}_{512}^{2\mathrm{D}} - \mathbf{S}_{512}^{3\mathrm{D}\,\top} \mathbf{S}_{512}^{3\mathrm{D}} \right\|_F^2 . \tag{2}$$

The second constraint promotes independence between private components by penalizing their cross-correlation,

$$\mathcal{L}_{\text{diff}} = \frac{1}{C_p^2} \left\| \mathbf{R}_{512}^{2\mathrm{D}\,\top} \mathbf{R}_{512}^{3\mathrm{D}} \right\|_F^2 , \tag{3}$$

thereby reducing redundancy and preventing modality-specific variations from interfering with shared semantics.

With the decomposed representations obtained, the shared features $\mathbf{S}_{512}^{2\mathrm{D}}$ and $\mathbf{S}_{512}^{3\mathrm{D}}$ are integrated by a Shared Attention Fusion (SAF) module, which aggregates modality-invariant information according to their semantic relevance. The attention weights $\mathbf{A}$ are computed as

$$\mathbf{A} = \mathrm{Softmax}\left(\frac{\mathbf{Q}(\mathbf{S}_{512}^{3\mathrm{D}}) \mathbf{K}(\mathbf{S}_{512}^{2\mathrm{D}})^\top}{\sqrt{d}} \right), \tag{4}$$

where $\mathbf{Q}(\cdot)$ and $\mathbf{K}(\cdot)$ denote linear transformations and d is the feature dimension. The 3D features are assigned as queries to guide the point-level prediction process, whereas the 2D features serve as keys to deliver supplementary semantic information obtained from the image branch. The fused shared representation is then computed as

$$\mathbf{S}_{1024} = \mathbf{A}\mathbf{V}(\mathbf{S}_{512}^{2\mathrm{D}}) + (1 - \mathbf{A})\mathbf{V}(\mathbf{S}_{512}^{3\mathrm{D}}), \tag{5}$$

where $\mathbf{V}(\cdot)$ denotes the value transformation. This aggregation balances the contribution of both modalities and maintains stable semantic correspondence in the shared space. For clarity, all feature dimensions are explicitly indicated in Eqs. (2)-(5) and the SAF formulation.

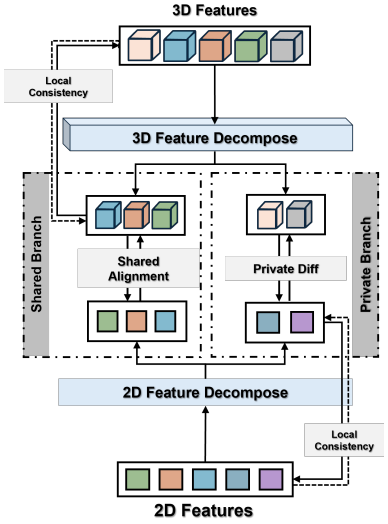


Figure 3. Architecture of the proposed shared-private feature decomposition.

Finally, the fused shared representation is concatenated with the 3D private features to produce the multimodal representation used for semantic prediction:

$$\mathbf{F}^{\text{fused}} = [\mathbf{S}; \mathbf{R}^{3\mathrm{D}}], \tag{6}$$

which is subsequently decoded into per-point semantic labels. By combining structured feature decomposition with attention-based fusion, the framework preserves modality-specific information while aligning semantic concepts across modalities, resulting in a compact and semantically coherent multimodal embedding.

3.4. Training Objectives

The training strategy is designed to jointly optimize semantic prediction accuracy, cross-modal consistency, and feature disentanglement. Accordingly, a set of complementary loss functions is introduced, each addressing a distinct learning objective. These components include semantic supervision for both modalities, cross-modal knowledge alignment, and regularization terms that stabilize the shared-private decomposition process.

Table 1. Quantitative results on the nuScenes validation set. Numbers for some baselines are taken from their original papers.

Method	Modality	barrier	bicycle	bus	car	construction	motorcycle	pedestrian	traffic cone	trailer	truck	driveable	other flat	sidewalk	terrain	manmade	vegetation	mIoU(%)
RangeNet++ [39]	L	66.0	21.3	77.2	80.9	30.2	66.8	69.6	52.1	54.2	72.3	94.1	66.6	63.5	70.1	83.1	79.8	65.5
PolarNet [71]	L	74.7	28.2	85.3	90.9	35.1	77.5	71.3	58.8	57.4	76.1	96.5	71.1	74.7	74.0	87.3	85.7	71.0
SalsaNext [11]	L	74.8	34.1	85.9	88.4	42.2	72.4	72.2	63.1	61.3	76.5	96.0	70.8	71.2	71.5	86.7	84.4	72.2
Cylinder3D [80]	L	76.4	40.3	91.3	93.8	51.3	78.0	78.9	64.9	62.1	84.4	96.8	71.6	76.4	75.4	90.5	87.4	76.1
SFCNet [75]	L	76.7	40.4	89.5	91.3	46.7	82.0	78.1	65.8	69.4	80.6	96.6	71.6	74.5	74.9	89.0	87.5	75.9
RangeVit [3]	L	75.5	40.7	88.3	90.1	49.3	79.3	77.2	66.3	65.2	80.0	96.4	71.4	73.8	73.8	89.9	87.2	75.2
MemorySeg [28]	L	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	76.7
PMF [81]	LC	74.1	46.6	89.8	92.1	57.0	77.7	80.9	**70.9**	64.6	82.9	95.5	73.3	73.6	74.8	89.4	87.7	76.9
2DPASS [61]	LC	78.4	52.5	95.4	93.4	62.5	88.9	83.1	68.1	75.6	**89.1**	96.9	75.5	76.3	75.7	89.1	86.9	80.5
PTv3 [55]	L	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	80.2
PTv3+PPT [55]	L	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	81.2
CSFNet (large) [36]	LC	77.8	52.6	**96.6**	94.4	**64.1**	**89.3**	82.9	67.9	**77.0**	87.7	96.8	76.6	76.2	76.1	89.3	86.7	80.8
U2MKD [47]	L	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	**83.1**
UniD-shift(ours)	LC	**78.9**	**52.7**	96.5	93.8	63.6	89.2	**84.5**	68.6	72.9	87.6	**97.1**	**77.9**	**77.8**	**77.7**	**91.3**	**88.9**	81.0

Notes. L: LiDAR-only. **LC**: LiDAR-Camera fusion.

Table 2. Quantitative results on the nuScenes test set.

LiDAR-only		LiDAR-Camera	
Method	mIoU (%)	Method	mIoU (%)
PolarNet [71]	69.4	PMF [81]	77.0
Cylinder3D [76]	77.2	2D3DNet [13]	80.0
CMDFusion [8]	80.8	MSeg3D[29]	81.1
LidarMultiNet [66]	81.4	U2MKD[47]	84.2
LiDARFormer [78]	81.5	TASeg [54]	**84.6**
SphereFormer [25]	**81.9**	UniD-Shift (Ours)	81.2

Shared-private regularization. Two regularization terms are introduced to stabilize feature decomposition and maintain disentanglement between subspaces. The Gram alignment $\mathcal{L}\text{gram}$ (Eq. 2) enforces consistency between the shared representations of both modalities, ensuring that they capture coherent semantic structures. Meanwhile, the decorrelation $\mathcal{L}\text{diff}$ (Eq. 3) reduces redundancy across the private subspaces, promoting independence between modality-specific components.

Segmentation loss. Semantic supervision is provided for both the 2D and 3D branches to ensure accurate category prediction within each modality. Each segmentation loss integrates a weighted cross-entropy term with a Lovász-Softmax loss, enhancing class balance and boundary precision:

$$\mathcal{L}_{\text{seg}}^{m} = \mathcal{L}_{\text{CE}}^{m} + \mathcal{L}_{\text{Lovasz}}^{m}, \quad m \in \{2\text{D}, 3\text{D}\}. \tag{7}$$

The 3D segmentation branch provides point-wise geometric supervision, while the 2D branch refines the fused prediction by leveraging dense pixel-level semantics.

Cross-modal knowledge distillation. To maintain semantic consistency between modalities, a bidirectional Kullback-Leibler divergence is adopted to align the predictive distributions of the image and point cloud branches:

$$\mathcal{L}_{\text{xm}} = \text{KL}\big(\log p(\hat{\mathbf{y}}^{3\text{D}}),\, p(\hat{\mathbf{y}}^{2\text{D}})\big)\,, \tag{8}$$

where $\hat{\mathbf{y}}^{2\text{D}}$ and $\hat{\mathbf{y}}^{3\text{D}}$ denote the fused and point-level predictions, respectively. This objective promotes bidirectional semantic adaptation and mutual feature calibration.

Overall optimization. All objectives are jointly optimized in an end-to-end manner. The total training loss is expressed as

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{seg}}^{3\text{D}} + \lambda_{\text{seg2D}}\mathcal{L}_{\text{seg}}^{2\text{D}} + \lambda_{\text{xm}}\mathcal{L}_{\text{xm}} + \lambda_{\text{gram}}\mathcal{L}_{\text{gram}} + \lambda_{\text{diff}}\mathcal{L}_{\text{diff}}, \tag{9}$$

where the weighting coefficients λ_{seg2D}, λ_{xm}, λ_{gram}, and λ_{diff} determine the relative contribution of each component. This joint optimization encourages balanced learning of semantic accuracy, modality alignment, and disentangled representation structure.

4. Experiments and Analysis

4.1. Experimental Setup

Benchmarks. Experiments are performed on the nuScenes [5] and SemanticKITTI [4] benchmarks, both providing synchronized LiDAR-camera data with dense semantic labels. nuScenes offers multimodal driving sequences with 16 categories, while SemanticKITTI provides annotated LiDAR sequences with 19 categories. These datasets serve as the basis for evaluating the proposed framework in single-domain settings, enabling a comprehensive analysis of segmentation accuracy and multimodal fusion behavior.

Model and Training Configuration. All experiments are implemented in PyTorch and conducted on a distributed setup with eight NVIDIA RTX 4090 GPUs. The network is trained for 100 epochs using stochastic gradient descent with an initial learning rate of 0.24, momentum of 0.9, weight decay of 10^{-4}, and a cosine annealing schedule. Random rotation, flipping, and uniform scaling are applied during data augmentation to improve generalization. The image branch adopts a SAM-based vision-language encoder, and the point cloud branch employs an SPTNet backbone. Both branches are jointly optimized through shared-private feature decomposition and semantic fusion to enhance cross-modal complementarity and maintain modality-specific distinctiveness.



Figure 4. Visualization of object segmentation performance on the nuScenes validation set.

4.2. Single-Domain Results

We evaluate UniD-Shift under the single-domain setting on the nuScenes [5] and SemanticKITTI [4] benchmarks. The results in Tables 1–3 indicate consistent improvements over existing multimodal segmentation frameworks. UniD-Shift achieves an mIoU of 81.0% on nuScenes validation set (Table 1), 81.2% on nuScenes test set (Tables 2,8), and 71.8% on SemanticKITTI (Tables 3), exceeding the performance of 2DPASS[61], CSFNet[36], and other strong multimodal fusion baselines. The improvements extend across categories with different scales, densities, and surface characteristics, indicating that the fusion strategy preserves geometric detail while maintaining stable semantic structure. UniD-Shift integrates information from dense imagery and sparse LiDAR in a consistent manner, producing predictions that remain well organized rather than conflicting or overly smoothed. Results across both datasets further show that the model establishes a coherent multimodal representation that remains reliable in regions affected by occlusion, irregular sampling, or strong geometric variation.

The visual comparisons in Figs. 4 and 5 include both single-modality baselines (SPVCNN [48], Point Transformer V2 [53]) and multimodal baselines (2DPASS [61], CSFNet [36]). The single-modality models retain strong geometric detail but often lose continuity when the point distribution becomes sparse. The multimodal baselines improve category coverage but still produce scattered or incomplete regions in scenes with complex depth changes or dense structural layouts.UniD-Shift delivers clearer object contours and more continuous semantic regions across both datasets. The predictions preserve structural integrity in areas where point density becomes irregular, while maintaining stable category transitions across adjacent regions. In particular, the fused representation reduces fragmented patches and isolated artifacts that frequently appear in baseline results. Moreover, the spatial layout remains coherent across the scene, even under strong depth variation or complex urban geometry. These observations indicate that UniD-Shift learns a resilient multimodal representation and maintains reliable performance in large-scale environments. The consistent advantage across diverse conditions suggests stable fusion across varying sensing conditions.

4.3. Ablation Study

As shown in Table 4, we perform an ablation study on the nuScenes to investigate how each design choice influences the overall performance. Replacing SPVCNN with SPTNet results in a clear improvement in geometric feature quality, which reflects the benefit of transformer-based modeling in capturing long-range structural dependencies. The substitution of the 2D encoder with SAM brings an additional performance gain, as the foundation model supplies richer semantic signals together with more stable pixel-level representations. Although these upgrades enhance the representational capacity of each modality, their effect remains limited when the interaction between branches relies on KL-based alignment. This observation underscores the need for a dedicated fusion mechanism that can fully exploit the complementary nature of 2D and 3D information.

Introducing the shared-private fusion module produces a more substantial improvement, since this component reorganizes multimodal features by separating modality-invariant semantics from modality-specific information. This separation reduces interference between branches and yields a more coherent correspondence across modalities. The configuration that integrates SAM, SPTNet, and the shared-private fusion achieves the best performance in the ablation study, reaching an mIoU of **74.8%** in Table 4. The comparison indicates that the performance improvement is primarily attributable to the structured fusion mechanism. This design regulates the interaction between heterogeneous modalities and strengthens the integration of geometric information with semantic content. For more ablation details, see Supplementary Section A, Table 7.

4.4. Cross-Domain Extension and Generalization

We conducted a representative cross-domain adaptation experiment on the nuScenes dataset under the **USA** $\rightarrow$ **Singapore** setting to examine the robustness of UniD-Shift un-

Table 3. Quantitative results on the SemanticKITTI validation set. Numbers for some baselines are taken from their original papers.

Method	Modality	car	bicycle	motorcycle	truck	other vehicle	person	bicyclist	motorcyclist	road	parking	sidewalk	other ground	building	fence	vegetation	trunk	terrain	pole	traffic sign	mIoU (%)
RangeNet++ [39]	L	89.4	26.5	48.4	33.9	26.7	54.8	69.4	0.0	92.9	37.0	69.9	0.0	83.4	51.0	83.3	54.0	68.1	49.8	34.0	51.2
SqueezeSegV2 [51]	L	82.7	15.1	22.7	25.6	26.9	22.9	44.5	0.0	92.7	39.7	70.7	0.1	71.6	37.0	74.6	35.8	68.1	21.8	22.2	40.8
SqueezeSegV3 [59]	L	87.1	34.3	48.6	47.5	47.1	58.1	53.8	0.0	95.3	43.1	78.2	0.3	78.9	53.2	82.3	55.5	70.4	46.3	33.2	53.3
RandLA-Net [18]	L	92.0	8.0	12.8	74.8	46.7	52.3	46.0	0.0	93.4	32.7	73.4	0.1	84.0	43.5	83.7	57.3	73.1	48.0	27.3	50.0
SalsaNext [11]	L	90.5	44.6	49.6	86.3	54.6	74.0	81.4	0.0	93.4	40.6	69.1	0.0	84.6	53.0	83.6	64.3	64.2	54.4	39.8	59.4
MinkowskiNet [10]	L	95.0	23.9	50.4	55.3	45.9	65.6	82.2	0.0	94.3	43.7	76.4	0.0	87.9	57.4	87.4	67.7	71.5	63.5	43.6	58.5
SPVCNN [48]	L	96.5	44.8	63.1	59.9	64.3	12.0	86.0	0.0	93.9	42.4	75.9	0.0	88.8	59.1	88.0	67.5	73.0	63.5	44.3	62.3
Cylinder3D [80]	L	96.4	61.5	78.2	66.3	69.8	80.8	93.3	0.0	94.9	41.5	78.0	1.4	87.5	50.0	86.7	72.2	68.8	63.0	42.1	64.9
FA-ResNet [67]	L	93.7	30.6	33.2	33.9	27.2	51.6	30.6	**18.7**	91.3	**65.2**	75.5	**26.0**	90.6	62.7	83.1	63.5	66.3	51.0	48.4	55.7
PTv3 [55]	L	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	70.8
PTv3+PPT [55]	L	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	**72.3**
PMF [81]	LC	95.4	47.8	62.9	68.4	75.2	78.9	71.6	0.0	**96.4**	43.5	80.5	0.1	88.7	60.1	88.6	72.7	75.3	65.5	43.0	63.9
2DPASS [61]	LC	97.2	55.3	78.3	90.9	70.7	80.7	93.5	0.1	94.3	53.8	82.0	12.1	93.0	71.8	**89.1**	72.3	75.7	66.0	54.2	70.0
MSeg3D [29]	LC	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	69.0
U2MKD [47]	LC	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	69.6
UniSeg [35]	LC	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	71.3
CSFNet (large) [36]	LC	**97.6**	**61.5**	88.5	**96.5**	**82.8**	**82.4**	**93.6**	2.2	94.8	52.1	**83.0**	0.0	**93.4**	**73.7**	**89.1**	**72.8**	75.6	**66.3**	52.3	71.5
UniD-shift(ours)	LC	96.8	59.4	**89.2**	92.3	80.1	**82.4**	**93.6**	13.1	94.7	61.5	78.3	14.2	93.3	60.2	83.2	71.7	**76.0**	66.1	**57.8**	71.8

Notes. L: LiDAR-only. **LC**: LiDAR-Camera fusion.



Figure 5. Visualization of object segmentation performance on the SemanticKitti validation set.

der clear changes in urban layout and sensing conditions. The experiment follows a standard cross-domain protocol in which the network is optimized with annotated data from the USA domain and evaluated on the Singapore domain. Both domains share the same sensing configuration and pipeline, ensuring that variations arise from environmental differences rather than data preparation.

Table 4. Comprehensive ablation study on backbone, fusion strategy, SAF query direction, and loss contributions.More detailed ablation studies are provided in Supplementary Section A (Table 7).

Exp	2D Module	3D Module	Fusion / Setting	mIoU (%)
Backbone + Fusion Ablation				
1 (Baseline)	ResNet	SPVCNN	KL	68.2(↓6.6)
2	ResNet	SPTNet	KL	70.1(↓4.7)
3	SAM	SPTNet	KL	73.4(↓1.4)
4 (Ours)	SAM	SPTNet	Shared-Private Fusion	**74.8**
SAF Query Direction Ablation				
5	SAM	SPTNet	2D→3D	70.9(↓3.9)
6	SAM	SPTNet	2D↔3D	71.5(↓3.3)
7 (Ours)	SAM	SPTNet	3D→2D	**74.8**
Loss Ablation				
8	SAM	SPTNet	w/o Gram	73.7(↓1.1)
9	SAM	SPTNet	w/o Decorr	74.4(↓0.4)
10	SAM	SPTNet	w/o Gram+Decorr	73.4(↓1.4)

As shown in Table 5, UniD-Shift reaches an mIoU of **74.5%** on the USA to Singapore benchmark and outperforms xMUDA (69.4%), UniDSeg (72.9%), and UniDxMD (74.3%). The advantage appears consistently across both 2D and 3D branches, suggesting that the shared-private formulation improves cross-modal alignment stability under domain shifts. The margin over fusion-based adaptation models indicates robust representations under variations in layout, objects, and background.

The overall trend in Table 5 shows that UniD-Shift delivers higher scores not only in the combined multimodal prediction but also in the unimodal components, revealing

Table 5. Cross-domain results on **nuScenes (USA/Sing.)**. "Uni." denotes uni-modal UDA, and "Cross." denotes cross-modal UDA. Comparison highlights the effectiveness of multimodal adaptation.

Mod.	Method	2D	3D	xM
Uni.	MinEnt [50]	57.6	61.5	66.0
	logCORAL [40]	64.4	63.2	69.4
	PL [31]	62.0	64.8	70.4
Cross.	xMUDA [20]	64.4	63.2	69.4
	AUDA [34]	63.2	61.9	67.5
	DsCML [41]	62.7	61.8	67.7
	Dual-Cross [30]	60.9	61.2	68.4
	SSE-xMUDA [72]	66.7	65.4	70.8
	BFtD [56]	63.7	62.2	69.4
	FtD++ [58]	69.7	64.6	69.8
	UniDSeg [57]	67.2	67.6	72.9
	MM2D3D [7]	71.7	66.8	72.4
	UniDxMD [33]	73.2	68.5	74.3
	UniD-Shift (Ours)	**73.3**	**68.8**	**74.5**

that the fusion strategy strengthens each branch instead of amplifying domain bias. The results point to a representation that absorbs complementary information from the two modalities in a balanced manner and maintains semantic consistency under geographic and structural variation. This behavior suggests that UniD-Shift forms a more transferable feature space, enabling reliable segmentation performance when applied to unseen urban environments.

5. Discussion

5.1. Interpretability and Fusion Efficiency

We further evaluate the balance between accuracy and efficiency to examine the practical value of the framework. The shared-private decomposition organizes multimodal features into distinct semantic roles and reduces interference during fusion, leading to a stable optimization process. Table 6 shows that UniD-Shift reaches an mIoU of **71.8%** on SemanticKITTI with a latency of **240 ms**. The results indicate that the frozen SAM encoder increases the overall parameter count while keeping the trainable portion compact, and the inference time remains lower than other SAM-based designs. This pattern reflects a fusion mechanism that limits computation and restricts cross-modal exchange to essential information. The evidence suggests UniD-Shift achieves a balanced trade-off between accuracy and computational cost, yielding an efficient and stable multimodal solution.

5.2. Cross-Domain Robustness and Generalization

The proposed framework exhibits strong robustness in cross-domain evaluation. In the USA→Singapore adaptation setting of the nuScenes dataset, UniD-Shift attains the highest performance among all compared methods, reaching an mIoU of **74.5%** as shown in Table 5. The improvement reflects the ability of the shared-private decomposition to separate stable semantic information from domain-dependent variations, thereby reducing the influence of differences in spatial layout, road structure, and environmental appearance across cities. The model forms a unified latent representation that supports reliable transfer from labeled source data to the unlabeled target domain and preserves semantic consistency throughout the prediction process. The overall trend indicates that UniD-Shift sustains high segmentation accuracy while maintaining stable behavior across urban environments, demonstrating its effectiveness for large-scale multimodal domain adaptation.

Table 6. Comparisons between efficiency (run-time) and accuracy on the SemanticKITTI validation set. A clear trade-off between latency and segmentation performance can be observed.

Method	#Params (M)	Latency (ms)	mIoU (%) ↑
MinkowskiNet [10]	21.7	48	61.1
Cylinder3D [80]	56.3	75	65.9
SPVCNN [48]	21.8	52	63.8
UniSeg [35]	147.6	145	71.3
CSFNet [36]	45.4	122	71.5
UniDSeg (CLIP) [57]	350.96 (38.69)	214	N/A
UniDSeg (SAM) [57]	353.44 (38.69)	296	N/A
UniD-Shift (SAM)	359.80 (35.20)	240	71.8

5.3. Limitations

Despite its strong performance, the framework depends on reliable geometric correspondence between LiDAR and camera data, which may constrain its use in scenes where the two modalities are not precisely synchronized. In addition, employing a large visual backbone increases memory demand during training and reduces scalability on devices with limited computational resources. Future research will investigate more compact fusion structures and adaptive alignment mechanisms to lower overhead and enhance suitability for real-time and large-scale applications.

6. Conclusion

In this paper, we propose UniD-Shift, a unified multimodal framework for 3D semantic segmentation that integrates image and point cloud information through shared-private feature decomposition. The method integrates SAM-driven visual priors with SPTNet-based geometric encoding to form a coherent latent space capable of capturing both semantic organization and spatial structure. Extensive experiments on nuScenes and SemanticKITTI demonstrate that UniD-Shift delivers superior segmentation accuracy while maintaining computational efficiency and training stability throughout the entire process. Cross-domain evaluations further confirm that the learned representations effectively under diverse environmental and sensor conditions, establishing the framework as a reliable and interpretable solution for the large-scale multimodal 3D scene understanding.

Acknowledgments

This work is supported by the Young Scientists Fund of the National Natural Science Foundation of China (Grant No. 42401567), Guangdong Provincial Project (Grant No. 2024QN11G095), AI Research and Learning Base of Urban Culture, Guangdong Provincial Department of Education (Grant No. 2023WZJD008), and the Open Fund of the Technology Innovation Center for 3D Real Scene Construction and Urban Refined Governance, Ministry of Natural Resources (Grant No. 2024PF-1)

References

[1] Zhaochong An, Guolei Sun, Yun Liu, Runjia Li, Min Wu, Ming-Ming Cheng, Ender Konukoglu, and Serge Belongie. Multimodality helps few-shot 3d point cloud semantic segmentation. *arXiv preprint arXiv:2410.22489*, 2024. 2

[2] Zhaochong An, Guolei Sun, Yun Liu, Runjia Li, Junlin Han, Ender Konukoglu, and Serge Belongie. Generalized few-shot 3d point cloud segmentation with vision-language model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16997–17007, 2025. 2

[3] Angelika Ando, Spyros Gidaris, Andrei Bursuc, Gilles Puy, Alexandre Boulch, and Renaud Marlet. Rangevit: Towards vision transformers for 3d semantic segmentation in autonomous driving. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5240–5250, 2023. 5

[4] Jens Behley, Martin Garbade, Andres Milioto, Jan Quenzel, Sven Behnke, Cyrill Stachniss, and Jurgen Gall. Semantickitti: A dataset for semantic scene understanding of lidar sequences. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 9297–9307, 2019. 5, 6

[5] Holger Caesar, Varun Bankiti, Alex H Lang, Sourabh Vora, Venice Erin Liong, Qiang Xu, Anush Krishnan, Yu Pan, Giancarlo Baldan, and Oscar Beijbom. nuscenes: A multimodal dataset for autonomous driving. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 11621–11631, 2020. 5, 6

[6] Haozhi Cao, Yuecong Xu, Jianfei Yang, Pengyu Yin, Shenghai Yuan, and Lihua Xie. Mopa: Multi-modal prior aided domain adaptation for 3d semantic segmentation. In *2024 IEEE International Conference on Robotics and Automation (ICRA)*, pages 9463–9470. IEEE, 2024. 3

[7] Adriano Cardace, Pierluigi Zama Ramirez, Samuele Salti, and Luigi Di Stefano. Exploiting the complementarity of 2d and 3d networks to address domain-shift in 3d semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 98–109, 2023. 3, 8

[8] Jun Cen, Shiwei Zhang, Yixuan Pei, Kun Li, Hang Zheng, Maochun Luo, Yingya Zhang, and Qifeng Chen. Cmdfusion: Bidirectional fusion network with cross-modality knowledge distillation for lidar semantic segmentation. *IEEE Robotics and Automation Letters*, 9(1):771–778, 2023. 5, 3

[9] Xuechao Chen, Shuangjie Xu, Xiaoyi Zou, Tongyi Cao, Dit-Yan Yeung, and Lu Fang. Svqnet: Sparse voxel-adjacent query network for 4d spatio-temporal lidar semantic segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 8569–8578, 2023. 3

[10] Christopher Choy, JunYoung Gwak, and Silvio Savarese. 4d spatio-temporal convnets: Minkowski convolutional neural networks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 3075–3084, 2019. 2, 7, 8

[11] Tiago Cortinhal, George Tzelepis, and Eren Erdal Aksoy. Salsanext: Fast, uncertainty-aware semantic segmentation of lidar point clouds. In *International Symposium on Visual Computing*, pages 207–222. Springer, 2020. 5, 7

[12] Jiajun Deng, Wengang Zhou, Yanyong Zhang, and Houqiang Li. From multi-view to hollow-3d: Hallucinated hollow-3d r-cnn for 3d object detection. *IEEE Transactions on Circuits and Systems for Video Technology*, 31(12):4722–4734, 2021. 2

[13] Kyle Genova, Xiaoqi Yin, Abhijit Kundu, Caroline Pantofaru, Forrester Cole, Avneesh Sud, Brian Brewington, Brian Shucker, and Thomas Funkhouser. Learning 3d semantic segmentation with only 2d image supervision. In *2021 International Conference on 3D Vision (3DV)*, pages 361–372. IEEE, 2021. 5, 3

[14] Benjamin Graham, Martin Engelcke, and Laurens Van Der Maaten. 3d semantic segmentation with submanifold sparse convolutional networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 9224–9232, 2018. 2

[15] Yulan Guo, Hanyun Wang, Qingyong Hu, Hao Liu, Li Liu, and Mohammed Bennamoun. Deep learning for 3d point clouds: A survey. *IEEE transactions on pattern analysis and machine intelligence*, 43(12):4338–4364, 2020. 1

[16] Abdullah Hamdi, Silvio Giancola, and Bernard Ghanem. Mvtn: Multi-view transformation network for 3d shape recognition supplementary material. 2

[17] Ji Hou, Angela Dai, and Matthias Nießner. 3d-sis: 3d semantic instance segmentation of rgb-d scans. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4421–4430, 2019. 2

[18] Qingyong Hu, Bo Yang, Linhai Xie, Stefano Rosa, Yulan Guo, Zhihua Wang, Niki Trigoni, and Andrew Markham. Randla-net: Efficient semantic segmentation of large-scale point clouds. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 11108–11117, 2020. 7

[19] Maximilian Jaritz, Tuan-Hung Vu, Raoul de Charette, Emilie Wirbel, and Patrick Pérez. xmuda: Cross-modal unsupervised domain adaptation for 3d semantic segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 12605–12614, 2020. 3

[20] Maximilian Jaritz, Tuan-Hung Vu, Raoul De Charette, Émilie Wirbel, and Patrick Pérez. Cross-modal learning for domain adaptation in 3d semantic segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(2):1533–1544, 2022. 2, 3, 8

[21] Li Jiang, Hengshuang Zhao, Shaoshuai Shi, Shu Liu, Chi-Wing Fu, and Jiaya Jia. Pointgroup: Dual-set point grouping for 3d instance segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and Pattern recognition*, pages 4867–4876, 2020. 2

[22] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C Berg, Wan-Yen Lo, et al. Segment anything. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 4015–4026, 2023. 3

[23] Maxim Kolodiazhnyi, Anna Vorontsova, Anton Konushin, and Danila Rukhovich. Oneformer3d: One transformer for unified point cloud segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 20943–20953, 2024. 1

[24] Xin Lai, Jianhui Liu, Li Jiang, Liwei Wang, Hengshuang Zhao, Shu Liu, Xiaojuan Qi, and Jiaya Jia. Stratified transformer for 3d point cloud segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 8500–8509, 2022. 2

[25] Xin Lai, Yukang Chen, Fanbin Lu, Jianhui Liu, and Jiaya Jia. Spherical transformer for lidar-based 3d recognition. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 17545–17555, 2023. 5, 3

[26] Alex H Lang, Sourabh Vora, Holger Caesar, Lubing Zhou, Jiong Yang, and Oscar Beijbom. Pointpillars: Fast encoders for object detection from point clouds. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 12697–12705, 2019. 2

[27] Minhyeok Lee, Suhwan Cho, Jungho Lee, Sunghun Yang, Heeseung Choi, Ig-Jae Kim, and Sangyoun Lee. Effective sam combination for open-vocabulary semantic segmentation. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 26081–26090, 2025. 3

[28] Enxu Li, Sergio Casas, and Raquel Urtasun. Memoryseg: Online lidar semantic segmentation with a latent memory. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 745–754, 2023. 5

[29] Jiale Li, Hang Dai, Hao Han, and Yong Ding. Mseg3d: Multi-modal 3d semantic segmentation for autonomous driving. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 21694–21704, 2023. 2, 5, 7, 3

[30] Miaoyu Li, Yachao Zhang, Yuan Xie, Zuodong Gao, Cuihua Li, Zhizhong Zhang, and Yanyun Qu. Cross-domain and cross-modal knowledge distillation in domain adaptation for 3d semantic segmentation. In *Proceedings of the 30th ACM International Conference on Multimedia*, pages 3829–3837, 2022. 3, 8

[31] Yunsheng Li, Lu Yuan, and Nuno Vasconcelos. Bidirectional learning for domain adaptation of semantic segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 6936–6945, 2019. 2, 8

[32] Ming Liang, Bin Yang, Shenlong Wang, and Raquel Urtasun. Deep continuous fusion for multi-sensor 3d object detection. In *Proceedings of the European conference on computer vision (ECCV)*, pages 641–656, 2018. 2

[33] Zhengyin Liang, Hui Yin, Min Liang, Qianqian Du, Ying Yang, and Hua Huang. Unidxmd: Towards unified representation for cross-modal unsupervised domain adaptation in 3d semantic segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 20346–20356, 2025. 3, 8

[34] Wei Liu, Zhiming Luo, Yuanzheng Cai, Ying Yu, Yang Ke, José Marcato Junior, Wesley Nunes Gonçalves, and Jonathan Li. Adversarial unsupervised domain adaptation for 3d semantic segmentation with multi-modal learning. *ISPRS Journal of Photogrammetry and Remote Sensing*, 176:211–221, 2021. 3, 8

[35] Youquan Liu, Runnan Chen, Xin Li, Lingdong Kong, Yuchen Yang, Zhaoyang Xia, Yeqi Bai, Xinge Zhu, Yuexin Ma, Yikang Li, et al. Uniseg: A unified multi-modal lidar segmentation network and the openpcseg codebase. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 21662–21673, 2023. 2, 7, 8

[36] Yang Luo, Ting Han, Yujun Liu, Jinhe Su, Yiping Chen, Jinyuan Li, Yundong Wu, and Guorong Cai. Csfnet: Cross-modal semantic focus network for sematic segmentation of large-scale point clouds. *IEEE Transactions on Geoscience and Remote Sensing*, 2025. 2, 5, 6, 7, 8

[37] Yang Luo, Ting Han, Xiaorong Zhang, Yujun Liu, Duxin Zhu, Jinyuan Li, Yiping Chen, Yundong Wu, Guorong Cai, Yingchao Piao, et al. Paseg: positional-guided segmenter with multimodal semantic alignment for enhancing urban scene 3d semantic segmentation. *International Journal of Digital Earth*, 18(1):2528811, 2025. 2

[38] Daniel Maturana and Sebastian Scherer. Voxnet: A 3d convolutional neural network for real-time object recognition. In *2015 IEEE/RSJ international conference on intelligent robots and systems (IROS)*, pages 922–928. IEEE, 2015. 2

[39] Andres Milioto, Ignacio Vizzo, Jens Behley, and Cyrill Stachniss. Rangenet++: Fast and accurate lidar semantic segmentation. In *2019 IEEE/RSJ international conference on intelligent robots and systems (IROS)*, pages 4213–4220. IEEE, 2019. 5, 7

[40] Pietro Morerio, Jacopo Cavazza, and Vittorio Murino. Minimal-entropy correlation alignment for unsupervised deep domain adaptation. *arXiv preprint arXiv:1711.10288*, 2017. 2, 8

[41] Duo Peng, Yinjie Lei, Wen Li, Pingping Zhang, and Yulan Guo. Sparse-to-dense feature matching: Intra and inter domain cross-modal learning in domain adaptation for 3d semantic segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 7108–7117, 2021. 3, 8

[42] Charles R Qi, Hao Su, Kaichun Mo, and Leonidas J Guibas. Pointnet: Deep learning on point sets for 3d classification and segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 652–660, 2017. 2

[43] Charles Ruizhongtai Qi, Li Yi, Hao Su, and Leonidas J Guibas. Pointnet++: Deep hierarchical feature learning on point sets in a metric space. *Advances in neural information processing systems*, 30, 2017. 2

[44] Wentao Qu, Jing Wang, YongShun Gong, Xiaoshui Huang, and Liang Xiao. An end-to-end robust point cloud semantic segmentation network with single-step conditional diffusion models. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 27325–27335, 2025. 1

[45] Shuran Song and Jianxiong Xiao. Deep sliding shapes for amodal 3d object detection in rgb-d images. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 808–816, 2016. 2

[46] Hang Su, Subhransu Maji, Evangelos Kalogerakis, and Erik Learned-Miller. Multi-view convolutional neural networks for 3d shape recognition. In *Proceedings of the IEEE international conference on computer vision*, pages 945–953, 2015. 2

[47] Tianfang Sun, Zhizhong Zhang, Xin Tan, Yong Peng, Yanyun Qu, and Yuan Xie. Uni-to-multi modal knowledge distillation for bidirectional lidar-camera semantic segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 46(12):11059–11072, 2024. 5, 7, 3

[48] Haotian Tang, Zhijian Liu, Shengyu Zhao, Yujun Lin, Ji Lin, Hanrui Wang, and Song Han. Searching efficient 3d architectures with sparse point-voxel convolution. In *European conference on computer vision*, pages 685–702. Springer, 2020. 6, 7, 8, 3

[49] Hugues Thomas, Charles R Qi, Jean-Emmanuel Deschaud, Beatriz Marcotegui, François Goulette, and Leonidas J Guibas. Kpconv: Flexible and deformable convolution for point clouds. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 6411–6420, 2019. 2

[50] Tuan-Hung Vu, Himalaya Jain, Maxime Bucher, Matthieu Cord, and Patrick Pérez. Advent: Adversarial entropy minimization for domain adaptation in semantic segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 2517–2526, 2019. 2, 8

[51] Bichen Wu, Xuanyu Zhou, Sicheng Zhao, Xiangyu Yue, and Kurt Keutzer. Squeezesegv2: Improved model structure and unsupervised domain adaptation for road-object segmentation from a lidar point cloud. In *2019 international conference on robotics and automation (ICRA)*, pages 4376–4382. IEEE, 2019. 7

[52] Guanyao Wu, Haoyu Liu, Hongming Fu, Yichuan Peng, Jinyuan Liu, Xin Fan, and Risheng Liu. Every sam drop counts: Embracing semantic priors for multi-modality image fusion and beyond. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 17882–17891, 2025. 3

[53] Xiaoyang Wu, Yixing Lao, Li Jiang, Xihui Liu, and Hengshuang Zhao. Point transformer v2: Grouped vector attention and partition-based pooling. *Advances in Neural Information Processing Systems*, 35:33330–33342, 2022. 2, 6

[54] Xiaopei Wu, Yuenan Hou, Xiaoshui Huang, Binbin Lin, Tong He, Xinge Zhu, Yuexin Ma, Boxi Wu, Haifeng Liu, Deng Cai, et al. Taseg: Temporal aggregation network for lidar semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15311–15320, 2024. 5, 3

[55] Xiaoyang Wu, Li Jiang, Peng-Shuai Wang, Zhijian Liu, Xihui Liu, Yu Qiao, Wanli Ouyang, Tong He, and Hengshuang Zhao. Point transformer v3: Simpler faster stronger. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4840–4851, 2024. 2, 5, 7

[56] Yao Wu, Mingwei Xing, Yachao Zhang, Yuan Xie, Jianping Fan, Zhongchao Shi, and Yanyun Qu. Cross-modal unsupervised domain adaptation for 3d semantic segmentation via bidirectional fusion-then-distillation. In *Proceedings of the 31st ACM International Conference on Multimedia*, pages 490–498, 2023. 3, 8

[57] Yao Wu, Mingwei Xing, Yachao Zhang, Xiaotong Luo, Yuan Xie, and Yanyun Qu. Unidseg: Unified cross-domain 3d semantic segmentation via visual foundation models prior. *Advances in Neural Information Processing Systems*, 37: 101223–101249, 2024. 2, 3, 8

[58] Yao Wu, Mingwei Xing, Yachao Zhang, Yuan Xie, Kaibei Peng, and Yanyun Qu. Fusion-then-distillation: Toward cross-modal positive distillation for domain adaptive 3d semantic segmentation. *IEEE Transactions on Circuits and Systems for Video Technology*, 2025. 3, 8

[59] Chenfeng Xu, Bichen Wu, Zining Wang, Wei Zhan, Peter Vajda, Kurt Keutzer, and Masayoshi Tomizuka. Squeezesegv3: Spatially-adaptive convolution for efficient point-cloud segmentation. In *European Conference on Computer Vision*, pages 1–19. Springer, 2020. 7

[60] Jianyun Xu, Ruixiang Zhang, Jian Dou, Yushi Zhu, Jie Sun, and Shiliang Pu. Rpvnet: A deep and efficient range-point-voxel fusion network for lidar point cloud segmentation. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 16024–16033, 2021. 2

[61] Xu Yan, Jiantao Gao, Chaoda Zheng, Chao Zheng, Ruimao Zhang, Shuguang Cui, and Zhen Li. 2dpass: 2d priors assisted semantic segmentation on lidar point clouds. In *European conference on computer vision*, pages 677–695. Springer, 2022. 2, 5, 6, 7, 3

[62] Yan Yan, Yuxing Mao, and Bo Li. Second: Sparsely embedded convolutional detection. *Sensors*, 18(10):3337, 2018. 2

[63] Bin Yang, Wenjie Luo, and Raquel Urtasun. Pixor: Real-time 3d object detection from point clouds. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, pages 7652–7660, 2018. 2

[64] Yu-Qi Yang, Yu-Xiao Guo, Jian-Yu Xiong, Yang Liu, Hao Pan, Peng-Shuai Wang, Xin Tong, and Baining Guo. Swin3d: A pretrained transformer backbone for 3d indoor scene understanding. *arXiv preprint arXiv:2304.06906*, 2023. 2

[65] Yu-Qi Yang, Yu-Xiao Guo, and Yang Liu. Swin3d++: Effective multi-source pretraining for 3d indoor scene understanding. *arXiv preprint arXiv:2402.14215*, 2024. 2

[66] Dongqiangzi Ye, Zixiang Zhou, Weijia Chen, Yufei Xie, Yu Wang, Panqu Wang, and Hassan Foroosh. Lidarmultinet: Towards a unified multi-task network for lidar perception. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 3231–3240, 2023. 5, 3

[67] Lixin Zhan, Wei Li, and Weidong Min. Fa-resnet: Feature affine residual network for large-scale point cloud segmentation. *International Journal of Applied Earth Observation and Geoinformation*, 118:103259, 2023. 7

[68] Boxiang Zhang, Zunran Wang, Yonggen Ling, Yuanyuan Guan, Shenghao Zhang, and Wenhui Li. Mx2m: masked cross-modality modeling in domain adaptation for 3d semantic segmentation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 3401–3409, 2023. 3

[69] Shuai Zhang, Yiping Chen, Biao Wang, Dong Pan, Wuming Zhang, and Aiguang Li. Sptnet: Sparse convolution and transformer network for woody and foliage components separation from point clouds. *IEEE Transactions on Geoscience and Remote Sensing*, 62:1–18, 2024. 3

[70] Shuai Zhang, Biao Wang, Yiping Chen, Shuhang Zhang, and Wuming Zhang. Point and voxel cross perception with lightweight cosformer for large-scale point cloud semantic segmentation. *International Journal of Applied Earth Observation and Geoinformation*, 131:103951, 2024. 1

[71] Yang Zhang, Zixiang Zhou, Philip David, Xiangyu Yue, Zerong Xi, Boqing Gong, and Hassan Foroosh. Polarnet: An improved grid representation for online lidar point clouds semantic segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9601–9610, 2020. 5, 3

[72] Yachao Zhang, Miaoyu Li, Yuan Xie, Cuihua Li, Cong Wang, Zhizhong Zhang, and Yanyun Qu. Self-supervised exclusive learning for 3d segmentation with cross-modal unsupervised domain adaptation. In *Proceedings of the 30th ACM International Conference on Multimedia*, pages 3338–3346, 2022. 3, 8

[73] Hengshuang Zhao, Li Jiang, Jiaya Jia, Philip HS Torr, and Vladlen Koltun. Point transformer. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 16259–16268, 2021. 2

[74] Weiguang Zhao, Rui Zhang, Qiufeng Wang, Guangliang Cheng, and Kaizhu Huang. Bfanet: Revisiting 3d semantic segmentation with boundary feature analysis. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 29395–29405, 2025. 1

[75] Yu Zheng, Guangming Wang, Jiuming Liu, Marc Pollefeys, and Hesheng Wang. Spherical frustum sparse convolution network for lidar point cloud semantic segmentation. *Advances in Neural Information Processing Systems*, 37:121827–121858, 2024. 5

[76] Hui Zhou, Xinge Zhu, Xiao Song, Yuexin Ma, Zhe Wang, Hongsheng Li, and Dahua Lin. Cylinder3d: An effective 3d framework for driving-scene lidar semantic segmentation. *arXiv preprint arXiv:2008.01550*, 2020. 5, 3

[77] Zixiang Zhou, Yang Zhang, and Hassan Foroosh. Panoptic-polarnet: Proposal-free lidar point cloud panoptic segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 13194–13203, 2021. 2

[78] Zixiang Zhou, Dongqiangzi Ye, Weijia Chen, Yufei Xie, Yu Wang, Panqu Wang, and Hassan Foroosh. Lidarformer: A unified transformer-based multi-task network for lidar perception. In *2024 IEEE International Conference on Robotics and Automation (ICRA)*, pages 14740–14747. IEEE, 2024. 5, 3

[79] Runsong Zhu, Shi Qiu, Zhengzhe Liu, Ka-Hei Hui, Qianyi Wu, Pheng-Ann Heng, and Chi-Wing Fu. Rethinking end-to-end 2d to 3d scene segmentation in gaussian splatting. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3656–3665, 2025. 2

[80] Xinge Zhu, Hui Zhou, Tai Wang, Fangzhou Hong, Yuexin Ma, Wei Li, Hongsheng Li, and Dahua Lin. Cylindrical and asymmetrical 3d convolution networks for lidar segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9939–9948, 2021. 5, 7, 8

[81] Zhuangwei Zhuang, Rong Li, Kui Jia, Qicheng Wang, Yuanqing Li, and Mingkui Tan. Perception-aware multi-sensor fusion for 3d lidar semantic segmentation. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 16280–16290, 2021. 5, 7, 3