DUAL-DIFFUSION SPATIAL INTERACTION REPRESENTATION GENERALIZATION FOR 3D SEMANTIC SEGMENTATION

Anonymous authors
Paper under double-blind review

ABSTRACT

3D semantic segmentation in unknown environments is crucial. Current methods utilize cross-modal learning and introduce 2D semantic priors to improve domain-generalized 3D segmentation. However, the heterogeneous representations of different modalities hinder cross-modal interaction, limiting the adaptation of 2D semantics to 3D space. In this paper, Dual-diffusion Spatial Interaction Representation (D^2SIR) is proposed to explore how cross-modal diffusion representations enhance 3D spatial semantics to improve the generalization of 3D semantic segmentation. Specifically, Spatial Representation Alignment (SRA) is first proposed to align the cross-modal diffusion representations, which designs manifold alignment of attributes and spatial coordinates to promote discriminativeness and spatial consistency before modalities interact. Secondly, Bidirectional Interaction Representation (BIR) models dual-diffusion interactions, which designs spatial joint adaptive representations to enable 2D and 3D bidirectional interactions. Finally, Spatial Semantic Modulation (SSM) is proposed to enhance 3D spatial semantics through effective interaction representations, which modulates local semantics in 3D space while suppressing spatially uncorrelated interaction noise, thereby improving the generalization of 3D semantic segmentation. Extensive experiments demonstrate that the proposed method has state-of-the-art performance. The code will be available.

1 INTRODUCTION

3D semantic segmentation aims to assign semantic categories to point clouds in 3D space, with wide applications in autonomous driving Li et al. (2021) and robotics Zhu & Chen (2024); He et al. (2026). Deep learning has made significant progress in this field in recent years Guo et al. (2021). However, acquiring large-scale datasets with fine-grained 3D point-level annotations remains costly and time-consuming Xiao et al. (2023a). Furthermore, the distribution of data in real-world environments varies significantly Li et al. (2026). Models trained on a single source domain often struggle to generalize to unseen target domains. These challenges hinder the reliable deployment of 3D segmentation systems in real-world environments.

Domain-generalized 3D semantic segmentation (DG3SS) aims to learn from labeled source domains and generalize to unseen target domains. Existing methods can be categorized into LiDAR-only methods and cross-modal methods Dong et al. (2026). LiDAR-only methods improve the generalization of 3D segmentation domains through point cloud data augmentation Park et al. (2025; 2024); He et al. (2024) or geometrically invariant representation learning Kim et al. (2026); Yang et al. (2026); He et al. (2025). However, 3D point cloud data collected by a single LiDAR sensor tends to characterize geometric structures, and the amount of information represented by a single modality is limited. Recently, cross-modal methods have utilized rich semantic and geometric cues from images and point clouds, which can learn diverse features and improve domain generalization. UniDSeg Wu et al. (2024) pioneered the integration of prior knowledge from 2D vision models into a cross-modal framework, significantly improving the generalization of 3D segmentation. XDiff3D Li et al. (2025b) utilizes a frozen diffusion model to extract 2D image features, capturing the potential contextual distribution and spatial layout between objects, thereby improving the generalization potential of 3D segmentation.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

However, the heterogeneous representations of different modalities make cross-modal interaction difficult, thus limiting the generalization potential of cross-modal representations. Although 2D vision foundation models or 2D diffusion models can be used to introduce rich semantic representations, simply using tokens for query tokens hinders deep interaction and learning between modal representations. Specifically, there are two reasons related to representation: **1) Disparate model representations.** 3D and 2D models use different mathematical operators to compute features. 3D segmentation models use sparse tensors to compute feature representations, while 2D segmentation models use operators such as 2D convolutions. This architectural heterogeneity means that cross-modal learning is often restricted to relying on feature queries or indirect supervision signals, making deep fusion and interaction difficult, thereby limiting the full realization of multimodal advantages. **2) Heterogeneous modal representations.** 2D images and 3D point clouds have different physical properties. Therefore, 2D features represent visual information based on dense pixels, while 3D features represent spatial information based on voxels. This physical heterogeneity poses a challenge to effective interaction between 2D and 3D representations, while spatially misaligned 2D semantics affect the fine-grained understanding of 3D space.

In this paper, a Dual-diffusion Spatial Interaction Representation (D^2SIR) method is proposed to explore how cross-modal diffusion representations enhance 3D spatial semantics to improve the generalization of 3D segmentation. Specifically, for the first problem, 3D diffusion model with similar operators is adopted to provide a foundational representation for direct and deep interaction between 2D and 3D representations. For the second problem, the proposed modules employ an *"Alignment-Interaction-Enhancement"* process to progressively leverage diffusion representations for enhancing 3D semantic generalization. *1) Alignment:* Spatial Representation Alignment (SRA) aligns the cross-modal diffusion representations, which designs manifold alignment of attributes and spatial coordinates to promote discriminativeness and spatial consistency before modalities interact. *2) Interaction:* Bidirectional Interaction Representation (BIR) models dual-diffusion interactions, which designs spatial joint adaptive representation to enable 2D and 3D bidirectional interactions. *3) Enhancement:* Spatial Semantic Modulation (SSM) is proposed to enhance 3D spatial semantics through effective interaction representations, which modulate local semantics in 3D space while suppressing spatially uncorrelated interaction noise, thereby improving the generalization of 3D semantic segmentation. Extensive experiments validate the superior performance of the proposed method. Our main contributions can be summarized as follows:

- The proposed D^2SIR explores the ability of cross-modal diffusion representations to enhance 3D spatial semantics, providing a new perspective on domain generalization in 3D segmentation.
- The proposed SRA, BIR, and SSM modules employ an "Alignment-Interaction-Enhancement" process that effectively interacts cross-modal diffusion representations within an aligned 3D space and enhances semantics based on valid spatial interaction relationships, thereby enhancing 3D semantic generalization.
- Superior performance across extensive experiments demonstrates the effectiveness of the proposed method.

2 Related Work

Domain-generalized 3D semantic segmentation aims to perform point-level recognition on point clouds of unknown distributions in a target domain Dong et al. (2026). Some methods utilize multiple source domain data for learning, but this requires acquiring multiple aligned datasets with class labels simultaneously Li et al. (2023); Saltori et al. (2023). Therefore, single-source domain generalization becomes a more challenging approach Xiao et al. (2023b), which can be divided into LiDAR-only methods and cross-modal methods Dong et al. (2026). LiDAR-only methods focus on point cloud data augmentation Park et al. (2025; 2024); He et al. (2024); Zhao et al. (2024) or learning geometrically invariant representations Kim et al. (2026); Yang et al. (2026); He et al. (2025); Wu et al. (2026). However, LiDAR sensors are primarily focused on characterizing geometry, and their information content is limited Jaritz et al. (2020). In recent years, cross-modal methods have leveraged 2D visual foundational models Wu et al. (2024) or 2D diffusion models Li et al. (2025b) to extract contextual semantics from images, thereby improving generalization. However, relying solely on 2D prior information to assist 3D segmentation prevents 2D semantics from effectively adapting to

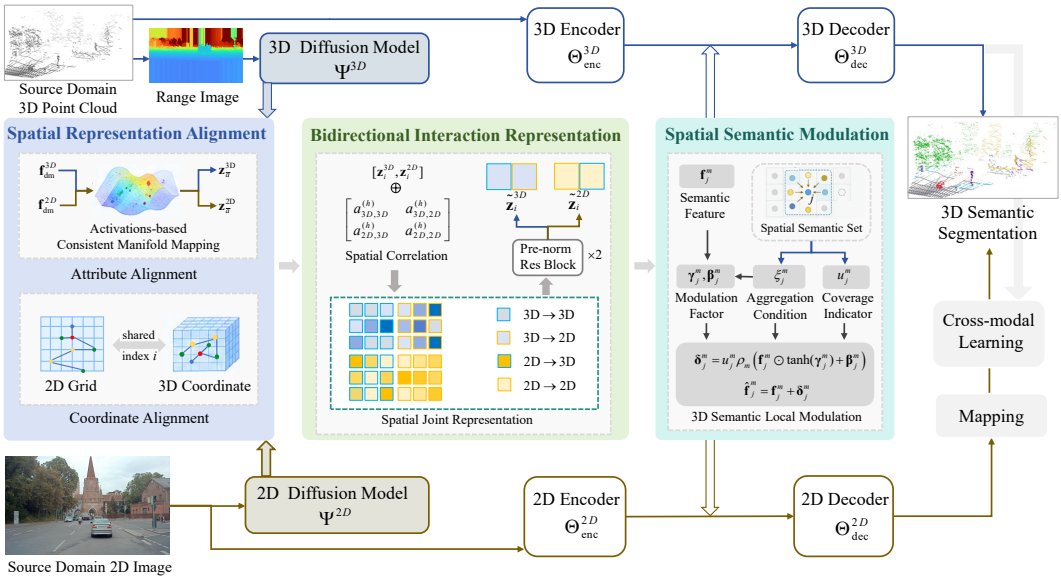


Figure 1: **Overview of the D^{2}SIR framework**. SRA aligns the cross-modal diffusion representation, promoting discriminativeness and spatial consistency before modalities interact. BIR explicitly models dual-diffusion interactions, facilitating effective bidirectional interactions between 2D and 3D. SSM enhances 3D spatial semantics through effective interaction representations to improve segmentation generalization.

3D space, as heterogeneous representations constrain spatial cross-modal learning. Therefore, the proposed method explores how cross-modal diffusion representations enhance 3D spatial semantics to improve the generalization of 3D segmentation.

Diffusion models enhance domain generalization through data augmentation and representation transfer Tang et al. (2023); Meng et al. (2024). Consequently, methods such as DIDEX Niemeijer et al. (2024), DGInStyle Jia et al. (2024), and CLOUDS Benigmim et al. (2024) leverage 2D diffusion models to generate variations in appearance and content beyond the source domain, thereby improving domain generalization for 2D images. QueryDiff utilizes diffusion-based semantics to guide 2D segmentation queries Li et al. (2025a). Recently, 2D diffusion representations have also been applied to domain generalization in 3D segmentation. XDiff3D extracts contextual representations from a frozen image diffusion model and transfers them to a 3D branch through query aggregation and cross-modal learning Li et al. (2025b). However, while these studies demonstrate the value of image diffusion priors for cross-domain semantic prediction, no existing work has investigated the generalization capability of 3D diffusion representations for 3D semantic recognition. Therefore, our method explores the synergistic representation of diffusion and semantics in 3D space, offering a new perspective on cross-domain generalization in 3D recognition.

LiDAR diffusion models provide 3D generative priors that are independent of image texture Fuest et al. (2026); Cao et al. (2025). LiDM learns the distribution of outdoor scenes within the latent space of compressed range maps Ran et al. (2024), while SG-LDM Xiang et al. (2025) employs semantic guidance for LiDAR generation. Additionally, CDSegNet Qu et al. (2025) and DOODLE Oh et al. (2026) utilize the diffusion process for closed-set semantic segmentation and open-set recognition, respectively. These works indicate that LiDAR diffusion representations can encode scene structure, geometric integrity, and semantic distributions. Distinct from these methods, we further explore how to harness the advantages of such 3D diffusion representations for 3D spatial semantic recognition, aiming to enhance the generalization capability of the recognition process.

3 D^{2}SIR METHOD

3.1 PROBLEM DEFINITION AND FRAMEWORK

Problem Definition. The goal of DG3SS is to learn a robust model Θ on a labeled source domain dataset $\mathcal{D}_s = \{(I, X, Y)\}$ that can perform pointwise semantic predictions without accessing the target domain $\mathcal{D}_t$ during training. Here, $\mathcal{D}_s$ and $\mathcal{D}_t$ have different domain distributions, I represents a 2D image, $X = \{x_i\}_{i=1}^{N}$ represents a 3D point cloud, $Y \in \{1, ..., C\}^N$ are 3D pointwise labels

with C classes, and N is the number of points. Due to the domain shift between the source and target data distributions, cross-modal domain generalization methods Jaritz et al. (2020); Wu et al. (2024) often employ a dual-branch architecture to process 2D and 3D modalities separately, and utilize cross-modal learning mechanisms to mitigate domain shift. Specifically, the frozen VFMs is used as a 2D encoder Θ^{2D}_{enc}, with a convolution-based 2D decoder Θ^{2D}_{dec} to generate a 2D prediction P^{2D}. The sparse encoder Θ^{3D}_{enc} and sparse decoder Θ^{3D}_{dec} are used to predict 3D point cloud segmentation P^{3D}. Furthermore, an auxiliary classifier is used to predict P^{2D}_{aux} and P^{3D}_{aux}, and cross-modal learning loss $\mathcal{L}_{CL}$ is used to minimize the Kullback-Leibler (KL) divergence between the primary predictions and the imitation predictions:

$$\mathcal{L}_{\text{CL}} = \mathcal{L}^{2D\leftarrow 3D}_{\text{CL}} + \mathcal{L}^{3D\leftarrow 2D}_{\text{CL}} = \frac{1}{N}\sum_{i=1}^{N}\sum_{c=1}^{C} P^{3D}_{i,c} \log \frac{P^{3D}_{i,c}}{P^{2D}_{\text{aux},i,c}} + \frac{1}{N}\sum_{i=1}^{N}\sum_{c=1}^{C} P^{2D}_{i,c} \log \frac{P^{2D}_{i,c}}{P^{3D}_{\text{aux},i,c}}. \quad (1)$$

Consequently, the model's overall optimization objective is to learn a robust segmentation model Θ by minimizing the source domain segmentation loss $\mathcal{L}_{seg}$ and the cross-modal learning loss:

$$\min_{\Theta}(\mathbb{E}_{(I,X,Y)\sim\mathcal{D}_s}[\mathcal{L}_{\text{seg}}(\Theta(I,X),Y)] + \mathbb{E}_{(I,X)\sim\mathcal{D}_s}[\mathcal{L}_{\text{CL}}]). \quad (2)$$

This cross-modal approach learns prior knowledge from 2D models that enhances 3D segmentation. However, relying solely on 2D prior information to assist 3D segmentation cannot effectively adapt 2D semantics to 3D space. Therefore, it is necessary to further explore cross-modal deep interaction and representation learning.

Dual-diffusion Framework. Our framework is illustrated in Figure 1. Firstly, a 2D diffusion model Ψ^{2D} is used to generate 2D diffusion features $\mathbf{f}^{2D}_{\text{dm}} = \Psi^{2D}(I)$, which perceives rich object texture details and semantic information Tang et al. (2023). However, the different model representations limit cross-modal interaction and fail to fully leverage the advantages of multimodal processing. Therefore, a 3D diffusion model Ψ^{3D} with a similar architecture is used to generate 3D diffusion features $\mathbf{f}^{3D}_{\text{dm}} = \Psi^{3D}(\hat{X})$, which represent different geometries and captures rich physical spatial patterns in 3D scenes. Here, $\hat{X}$ is a range image obtained through spherical projection of the point cloud X, where each pixel represents the reflection distance of the laser beam emitted by the LiDAR sensor. Both diffusion models are based on the U-Net architecture, but use different operators to compute features.

However, these two diffusion processes exhibit heterogeneous modal representations. 2D images and 3D point clouds possess different physical properties, and their features are represented at points in different physical spaces. If cross-modal learning is performed directly, ineffective cross-modal spatial relationships introduce noise into the semantic information, hindering the generalization capability of segmentation. Therefore, D^{2}SIR enhances semantic information on effective cross-modal spatial relationships to improve 3D segmentation generalization. Specifically, Spatial Representation Alignment (SRA) aligns the attributes and coordinates in 2D and 3D spaces. Secondly, Bidirectional Interaction Representation (BIR) constructs a joint representation of bidirectional interactions. Finally, Spatial Semantic Modulation (SSM) constructs effective cross-modal spatial relationships and modulates semantics with interaction representations in 3D space, enhancing 3D segmentation generalization. Details are presented below.

3.2 Spatial Representation Alignment

2D and 3D diffusion models employ different architectures and activation mechanisms, resulting in drastically different properties of diffusion features. Furthermore, due to the different organization of the image grid and the distance view grid, these two representations lack a direct coordinate correspondence. Therefore, it is necessary to align attributes and coordinates separately.

Attribute-aligned Manifold Representation. Due to the differences in modal representations, attribute alignment cannot be achieved based solely on the number of module layers. Activations Meng et al. (2024) represent the functional roles of diffusion features, facilitating the compression of high-dimensional diffusion features and enabling discriminative representation within a low-dimensional manifold space. Therefore, based on a combination of discriminative activations, the high-dimensional feature representations of 2D and 3D diffusion models can be mapped to a low-dimensional manifold space. Specifically, for Ψ^{2D} and Ψ^{3D}, we synchronously extract diffusion

activations at timestep t^{2D} and t^{3D}, including cross-attention queries, residual block outputs, second cross-attention queries, and self-attention keys. Note that Ψ^{3D} is an unconditional diffusion model and does not incorporate a cross-attention mechanism. Consequently, the two query roles are realized via self-attention queries. For modality m, the discriminative screening $\mathcal{H}^m$ can be formulated as $\mathbf{z}^m_{\bar{a},k} = \mathcal{H}^m_{\ell^m_k}[\Psi^m, t^m]$, where ℓ^m_k denotes the selected layer, $k \in \mathcal{K} = \{1,2,3,4\}$ represents the k-th activation role. Collecting the activations compatible with these four roles yields an attribute-aligned manifold representation:

$$\mathbf{z}^m_{\bar{a}} = \left(\mathbf{z}^m_{\bar{a},1}, \mathbf{z}^m_{\bar{a},2}, \mathbf{z}^m_{\bar{a},3}, \mathbf{z}^m_{\bar{a},4}\right), \qquad m \in \{2D, 3D\}, \tag{3}$$

where $\mathbf{z}^m_{\bar{a}}$ represents a sequence of activations with ordered roles, enforcing attribute alignment between the 2D and 3D manifold representations. The specific options are provided in the appendix.

Coordinate-aligned Manifold Representation. The attribute-aligned manifold representations are organized on the image grid and the range-view grids respectively, which means that the spatial locations of the represented points are not aligned, making spatial interaction impossible. Therefore, it is necessary to establish a cross-modal spatial correspondence to align the coordinates.

Specifically, for a point $\mathbf{x}_i$ in the point cloud, image pixels aligned with the point cloud are selected based on the camera coordinates, and the projected coordinates are denoted as π^{2D}_i. Simultaneously, the range image coordinates π^{3D}_i are recorded according to the spherical projection. Although these coordinates are located in different views, they all correspond to the same 3D point. Furthermore, since the scales of $\mathbf{z}^m_{\bar{a},k}$ vary, for modality m, the operator $\mathcal{T}^m_k$ is used to maintain a consistent scale between π^m_i and the k-th activation $\mathbf{z}^m_{\bar{a},k}$. Therefore, a cross-modal spatial relation query is:

$$a^m_i = \mathbb{I}\left[\pi^m_i \in \Omega^m_{\text{In}} \ \wedge \ \mathcal{T}^m_k(\pi^m_i) \in \Omega^m_{\bar{a},k}, \quad \forall\, k \in \mathcal{K}\right], \tag{4}$$

where $a^m_i \in \{0,1\}$, Ω^m_{In} and $\Omega^m_{\bar{a},k}$ denote the valid ranges of the input view and the k-th activation, respectively. This query method avoids multiple transformations between grid-level features and point cloud-level features. Furthermore, we concatenate all pointwise activations according a^m_i to obtain pointwise manifold representation:

$$\mathbf{p}^m_i = \begin{cases} \Big\Vert_{k\in\mathcal{K}} \ \mathcal{G}\left(\mathbf{z}^m_{\bar{a},k} | \mathcal{T}^m_k(\pi^m_i)\right), & a^m_i = 1, \\ \mathbf{0}, & a^m_i = 0, \end{cases} \tag{5}$$

where $\mathcal{G}$ represents the nearest-neighbor search on the manifold representation based on the effective coordinates. $\Vert$ denotes concatenation. Let $C^m = \sum_{k\in\mathcal{K}} C^m_k$ denote the concatenated channel dimension, thus $\mathbf{p}^{2D}_i \in \mathbb{R}^{C^{2D}}$ and $\mathbf{p}^{3D}_i \in \mathbb{R}^{C^{3D}}$. The two pointwise features inherit the same point order, without cross-point reordering or dense matching between the image and range-view grids. Finally, for a point cloud with N points, the aligned representations are

$$\mathbf{z}^{2D}_{\text{a}} = \begin{bmatrix} (\mathbf{p}^{2D}_1)^\mathsf{T} \\ \vdots \\ (\mathbf{p}^{2D}_N)^\mathsf{T} \end{bmatrix} \in \mathbb{R}^{N\times C^{2D}}, \qquad \mathbf{z}^{3D}_{\text{a}} = \begin{bmatrix} (\mathbf{p}^{3D}_1)^\mathsf{T} \\ \vdots \\ (\mathbf{p}^{3D}_N)^\mathsf{T} \end{bmatrix} \in \mathbb{R}^{N\times C^{3D}}, \tag{6}$$

where the i-th rows of the two outputs correspond to the same point $\mathbf{x}_i$, while retaining the independent validity of a^{2D}_i and a^{3D}_i. Therefore, the correspondence between different coordinates can be determined via the shared point index i, thereby achieving a coordinate-aligned manifold representation. This approach paves the way for subsequent manifold interaction computations based on 3D point cloud coordinates.

3.3 Bidirectional Interaction Representation

To achieve effective cross-modal information exchange between aligned representations, BIR organizes the modality-specific manifolds into a joint manifold representation and explicitly models the bidirectional dependencies between the two modalities at spatially corresponding points. Although cross-attention can model inter-modal correlations, calculating the correlation once is biased. Mathematically, cross-attention approximates a conditional distribution, representing an asymmetric, unidirectional injection of information. Therefore, inspired by joint self-attention that models joint distributions in a simple way, we compute a joint representation capturing bidirectional correlations from the 2D and 3D aligned manifold representations, explicitly modeling cross-modal bidirectional interactions at spatial points.

Normalization Mapping. For point i, modality-specific mapping and normalization are employed to project the aligned representations into a shared d-dimensional space:

$$\mathbf{z}_i^{2D} = \mathrm{LN}^{2D}\big(\mathbf{W}^{2D}(\mathbf{z}_{\mathrm{a}}^{2D})_i^{\mathsf{T}}\big) + \mathbf{e}^{2D}, \qquad \mathbf{z}_i^{3D} = \mathrm{LN}^{3D}\big(\mathbf{W}^{3D}(\mathbf{z}_{\mathrm{a}}^{3D})_i^{\mathsf{T}}\big) + \mathbf{e}^{3D}, \tag{7}$$

where $\mathbf{W}^{2D} \in \mathbb{R}^{d\times C^{2D}}$ and $\mathbf{W}^{3D} \in \mathbb{R}^{d\times C^{3D}}$ are branch-specific learnable projections, LN^{2D} and LN^{3D} denote LayerNorm operations with independent parameters, and $\mathbf{e}^{2D}$ and $\mathbf{e}^{3D}$ are the corresponding modality embeddings.

Spatial Joint Representation. To model a joint representation for spatial points, point-level correlations are calculated based on the spatial queries a_i^{2D} and a_i^{3D}. Specifically, let the ordered modality sequence be $\mathcal{M}_i = [\mathbf{z}_i^{2D}, \mathbf{z}_i^{3D}]$. Let $\kappa_i = a_i^{2D} + a_i^{3D} \in \{0,1,2\}$, then the local cross-modal interaction unit $\mathbf{Z}_i$ for point i can be expressed as:

$$\mathbf{Z}_i = \begin{cases} \mathrm{Stack}(\mathbf{z}_i^{2D}, \mathbf{z}_i^{3D}), & (a_i^{2D}, a_i^{3D}) = (1,1), \\ \mathrm{Stack}(\mathbf{z}_i^{2D}), & (a_i^{2D}, a_i^{3D}) = (1,0), \\ \mathrm{Stack}(\mathbf{z}_i^{3D}), & (a_i^{2D}, a_i^{3D}) = (0,1), \\ \varnothing, & (a_i^{2D}, a_i^{3D}) = (0,0). \end{cases} \tag{8}$$

where the length of $\mathbf{Z}_i$ is determined by the number of valid modalities at that point. Each non-empty sequence undergoes interaction using multi-head self-attention. Let $\mathbf{Z}_i' = \mathrm{LN}(\mathbf{Z}_i)$, the dimension of each head h is $d_h = d/H$, and the learnable matrices for each head are $W_Q^{(h)}$, $W_K^{(h)}$, $W_V^{(h)}$, and the multi-head output matrix is W_O. The attention weights $A_i^{(h)}$ and the multi-head output are:

$$\mathbf{A}_i^{(h)} = \mathrm{Softmax}\left(\frac{(\mathbf{Z}_i'\mathbf{W}_Q^{(h)})(\mathbf{Z}_i'\mathbf{W}_K^{(h)})^{\mathsf{T}}}{\sqrt{d_h}}\right), \ \mathrm{MHA}(\mathbf{Z}_i') = \left(\Big\Vert_{h=1}^{H} \mathbf{A}_i^{(h)}\Big(\mathbf{Z}_i'\mathbf{W}_V^{(h)}\Big)\right)\mathbf{W}_O. \tag{9}$$

In general, $\mathbf{A}_i^{(h)} \in \mathbb{R}^{\kappa_i\times\kappa_i}$. When the 2D tokens z_i^{2D} and 3D tokens z_i^{3D} are both valid, the matrix can be written as

$$\mathbf{A}_i^{(h)} = \begin{bmatrix} a_{2D,2D}^{(h)} & a_{2D,3D}^{(h)} \\ a_{3D,2D}^{(h)} & a_{3D,3D}^{(h)} \end{bmatrix}, \tag{10}$$

where $a_{2D,3D}^{(h)}$ denotes the 2D token reading information from the 3D token, and $a_{3D,2D}^{(h)}$ denotes the 3D token reading information from the 2D token. The two off-diagonal terms establish the bidirectional cross-modal exchange at the same point, while different points remain independent.

Adaptive Interaction and Separation. To facilitate cross-modal bidirectional interaction, two parameter-unshared pre-norm residual blocks are employed to update each non-empty token. Let $\mathbf{Z}_i^{(0)} = \mathbf{Z}_i$, the feature $\mathbf{Z}_i^{(\ell)}$ of the ℓ-th layer ($\ell \in \{1,2\}$) can be calculated as:

$$\begin{aligned} \mathbf{H}_i^{(\ell)} &= \mathbf{Z}_i^{(\ell-1)} + \mathrm{MHA}^{(\ell)}\Big(\mathrm{LN}^{(\ell)}(\mathbf{Z}_i^{(\ell-1)})\Big), \\ \mathbf{Z}_i^{(\ell)} &= \mathbf{H}_i^{(\ell)} + \mathrm{FFN}^{(\ell)}\Big(\mathrm{LN}^{(\ell)}(\mathbf{H}_i^{(\ell)})\Big), \qquad \widetilde{\mathbf{Z}}_i = \mathbf{Z}_i^{(2)}. \end{aligned} \tag{11}$$

When $\kappa_i = 2$, the two modalities interact bidirectionally inside these residual blocks. When $\kappa_i = 1$, the single valid token only passes through a single-token residual transform. When $\kappa_i = 0$, the point bypasses the interaction module. All attention computations are confined to the token sequence of one point. After interaction, the results are restored to the branches according to the modality positions (pos) recorded in $\mathcal{M}_i$:

$$[\widetilde{\mathbf{z}}_i^{2D}, \widetilde{\mathbf{z}}_i^{3D}] = \widetilde{\mathbf{z}}_i = \begin{cases} \widetilde{\mathbf{Z}}_i[\mathrm{pos}(\mathcal{M}_i|m)], & a_i^m = 1, \\ \mathbf{0}, & a_i^m = 0. \end{cases} \tag{12}$$

The resulting $\widetilde{\mathbf{z}}_i^{2D}$ and $\widetilde{\mathbf{z}}_i^{3D}$ retain their modality identity, and both have completed same-point bidirectional interaction when both sides are valid. This approach enables the modeling of cross-modal bidirectional correlations at each spatially corresponding point while maintaining modal separation, thereby achieving effective cross-modal bidirectional interaction.

3.4 SPATIAL SEMANTIC MODULATION

Interactive representations require effective enhancement of the segmented semantic space to improve generalization. Although affine transformations Perez et al. (2018) from diffusion features to segmentation features can modulate information, they may incorrectly transmit spatially unrelated local semantics, affecting the generalization of fine-grained recognition. There is a physical spatial discrepancy between the external conditions in this 3D space and the low-dimensional manifold representations, preventing direct modulation. Therefore, SSM enhances semantic information on effective cross-modal spatial relationships. Firstly, effective cross-modal spatial semantics are constructed based on the spatial arrangements of different dimensions. Secondly, interactive representations are used to modulate spatial semantics in 3D space, which suppresses spurious cross-modal correlations and enhances relevant local 3D semantics, thereby improving the generalization ability of point cloud segmentation.

Cross-modal Spatial Semantic Representation. Let $g_m(i)$ be the mapping from a manifold interaction point i to a spatial semantic entity j in modality m, where $j \in \Omega_m$, and Ω_m denotes the 3D spatial semantic set. Note that Ω_{2D} is obtained by filtering the deepest-layer patch features of Θ_{enc}^{2D} via sparse projection from point clouds to images, while Ω_{3D} is obtained by coarsening the input sparse coordinates based on the bottleneck stride. Only entities satisfying $a_i^m = 1$ and capable of being mapped via g_m participate in aggregation, thus the spatial semantic set can be computed as:

$$\mathcal{I}_j^m = \{i \mid a_i^m = 1,\ g_m(i) = j\} \tag{13}$$

Therefore, valid points falling into j are used to compute the aggregated value ξ_j^m via an averaging operation, and a coverage indicator u_j^m is used to mark whether the entity has received at least one valid point:

$$\xi_j^m = \frac{1}{\max\big(|\mathcal{I}_j^m|, 1\big)} \sum_{i \in \mathcal{I}_j^m} \widetilde{\mathbf{z}}_i^m, \qquad u_j^m = \mathbb{I}\big[|\mathcal{I}_j^m| > 0\big] \tag{14}$$

where $\mathbb{I}$ is the Boolean indicator function, and an empty set corresponds to a value of $\xi_j^m = \mathbf{0}$ and $u_j^m = 0$. ξ_j^m and u_j^m can correct the manifold geometry based on the spatial arrangement of different modalities, enabling subsequent decoupled modulation to be performed in 3D space with a specified coverage area.

3D Semantic Local Modulation To effectively map manifold interaction representations back into the 3D semantic feature space, 3D semantic local modulation utilizes the aggregation condition ξ_j^m to predict scaling and shifting parameters along the channel dimension, while the coverage indicator u_j^m controls whether an enhancement operation is permitted at a given location, thereby preventing interference with the semantic features of uncovered 3D regions. Specifically, the condition ξ_j^m is first mapped using a conditional projection ϕ^m to jointly predict the modulation factor:

$$[\boldsymbol{\gamma}_j^m, \boldsymbol{\beta}_j^m] = \mathrm{MLP}^m\big(\mathbf{f}_j^m \parallel \bar{\xi}_j^m\big), \quad \bar{\xi}_j^m = \phi^m(\xi_j^m). \tag{15}$$

where $\mathbf{F}^m = \{\mathbf{f}_j^m \mid j \in \Omega_m\}$ represents the semantic features of the branch. The symbol $\parallel$ denotes concatenation along the channel dimension, and ϕ^m consists of a linear layer, layer normalization, and a ReLU activation function. Subsequently, the result is written back to the semantic features based on u_j^m:

$$\boldsymbol{\delta}_j^m = u_j^m \rho_m \left(\mathbf{f}_j^m \odot \tanh(\boldsymbol{\gamma}_j^m) + \boldsymbol{\beta}_j^m\right), \qquad \widehat{\mathbf{f}}_j^m = \mathbf{f}_j^m + \boldsymbol{\delta}_j^m. \tag{16}$$

where ρ_m is a branch-specific learnable residual scaling coefficient. When $u_j^m = 0$, $\boldsymbol{\delta}_j^m = \mathbf{0}$, ensuring that uncovered locations remain unchanged. Furthermore, the final layer of the MLP employs zero initialization to ensure that, at the initialization stage, the modulated semantic features align with the state where the coverage indicator is disabled. By leveraging aggregation conditions and coverage indicators, 3D semantic local modulation effectively maps interaction features back into the 3D semantic space. This method modulates local semantics in 3D space while suppressing spatially uncorrelated interaction noise, thereby enhancing the generalization capability of 3D segmentation.

4 EXPERIMENTS AND ANALYSIS

4.1 EXPERIMENTAL SETUP

Dataset Description. Four autonomous driving datasets are used to evaluate domain generalization for 3D semantic segmentation. **nuScenes** Caesar et al. (2020) is a real-world dataset collected in

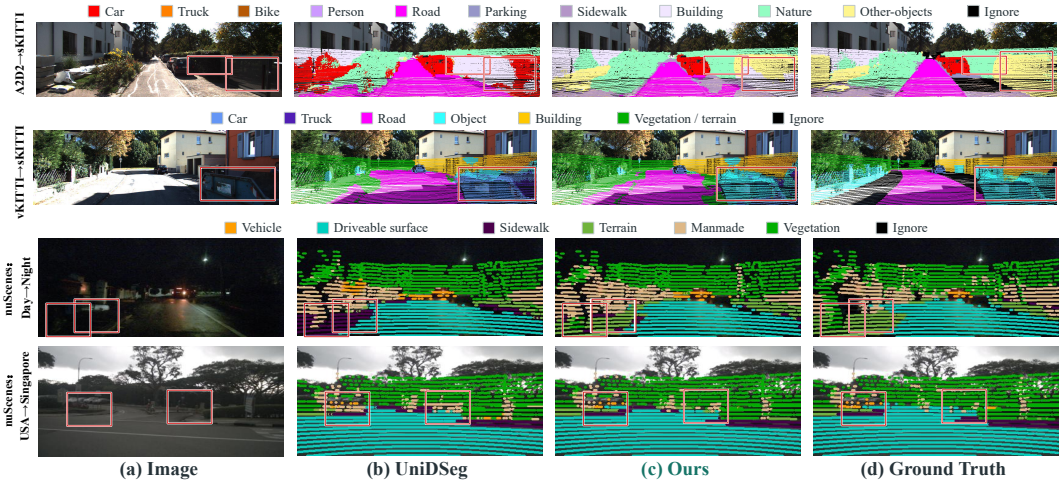


Figure 2: **Qualitative comparison** with current methods across four benchmarks.

Table 1: **Quantitative comparison** of cross-modal domain adaptation and domain generalization methods for 3D semantic segmentation. The results are the mIoU of the 2D branch, the 3D branch, and their ensemble (xM) on the target test set.

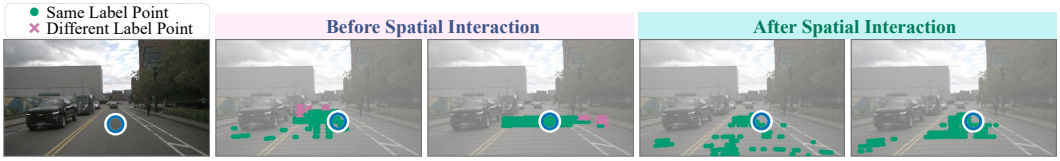
	Source/Target	nuScenes:USA/Sing			nuScenes:Day/Night			vKITTI/sKITTI			A2D2/sKITTI		
Task	Method	2D	3D	xM	2D	3D	xM	2D	3D	xM	2D	3D	xM
DA	xMUDA Jaritz et al. (2020)	64.4	63.2	69.4	55.5	69.2	67.4	42.1	46.7	48.2	38.3	46.0	44.0
	AUDA Liu et al. (2021)	64.0	64.0	69.2	55.6	69.8	64.8	35.8	37.8	41.3	43.0	43.6	46.8
	DsCML Peng et al. (2021)	65.6	56.2	66.1	50.9	49.3	53.2	38.4	38.4	45.5	39.6	45.1	44.5
	Dual-Cross Li et al. (2022)	64.7	58.1	66.5	58.5	69.7	68.0	40.7	35.1	44.2	45.9	40.0	48.6
	SSE Zhang et al. (2022)	64.9	63.9	69.2	62.8	69.0	68.9	45.9	40.0	49.6	44.5	46.8	48.4
	BFtD Wu et al. (2023)	63.7	62.2	69.4	57.1	70.4	68.3	41.5	45.5	51.5	40.5	44.4	48.7
	MM2D3D Cardace et al. (2023)	**71.7**	66.8	72.4	**70.5**	70.2	72.1	53.4	50.3	56.5	42.3	46.1	46.2
	VFMSeg Xu et al. (2025)	70.0	65.6	72.3	60.6	70.5	66.5	57.2	52.0	61.0	45.0	52.3	50.0
	UniDSeg Wu et al. (2024)	67.2	67.6	72.9	63.2	71.2	71.2	60.5	50.9	62.0	50.7	55.4	57.5
	XDiff3D Li et al. (2025b)	–	**69.5**	74.1	–	**73.5**	72.3	–	53.1	63.3	–	**57.6**	58.8
	Ours	69.0	68.9	**74.8**	61.7	71.6	**72.6**	**63.9**	**68.0**	**69.8**	**55.6**	49.8	**60.3**
DG	xMUDA Jaritz et al. (2020)	58.7	62.3	68.6	43.0	68.9	59.6	25.7	37.4	39.0	34.9	36.7	41.6
	MM2D3D Cardace et al. (2023)	69.7	62.3	70.9	**65.3**	63.2	68.3	37.7	40.2	44.2	39.6	35.9	43.6
	UniDSeg Wu et al. (2024)	66.5	64.5	72.3	57.0	70.5	70.0	57.6	44.7	60.0	48.8	46.3	54.4
	XDiff3D Li et al. (2025b)	–	66.7	73.5	–	**72.4**	71.6	–	46.9	61.3	–	49.1	56.2
	Ours	**70.9**	**70.8**	**75.0**	59.0	70.9	**72.1**	**60.1**	**59.7**	**65.4**	**50.0**	**53.2**	**57.6**

Boston and Singapore using a 32-beam LiDAR. **sKITTI** Behley et al. (2019) is a dataset collected in Karlsruhe, Germany. **A2D2** Geyer et al. (2020) is an outdoor dataset collected in Munich, Gaimersheim, and Ingolstadt, Germany. **vKITTI** Gaidon et al. (2016) is a synthetic dataset collected in a virtual environment.

Implementation Details. Following the settings of baseline Wu et al. (2024), we use Adam with an iteration-based MultiStepLR schedule. The initial learning rate is 10^{-3} for all trainable components, and is divided by 10 at each scheduled milestone. The voxel size is set to 5 cm. The model architecture is the same as the baseline, where Θ^{2D}_{enc} is ViT-L/14 Dosovitskiy et al. (2021) and Θ^{3D} is SCN U-Net Graham et al. (2018). To avoid domain leakage, Ψ^{2D} use the frozen Stable Diffusion v1.5 Rombach et al. (2022) pretrain on LAION-2B Schuhmann et al. (2022), Ψ^{3D} use the frozen unconditional LiDM Ran et al. (2024) pretrain on kitti360 Liao et al. (2023). Further details are provided in the appendix.

4.2 Compared with the State-of-the-arts

Quantitative Comparison. Table 1 compares domain adaptation and domain generalization methods across four cross-domain settings. Our method achieves state-of-the-art performance in all xM results. While UniDSeg Wu et al. (2024) utilizes 2D VFM and XDiff3D Li et al. (2025b) leverages 2D diffusion model to introduce rich semantic representations, using only tokens for queries hinders deep interaction and learning between modal representations. Our method explores cross-modal diffusion representations to enhance 3D spatial semantics, thus improving the generalization of 3D segmentation. Table 2 presents the classwise comparison with state-of-the-art methods, demonstrating the general superiority of our method across categories. To further demonstrate the advantages of our proposed method across different diffusion models, Table 4 compares XDiff3D and our method in the 3D results, proving the superiority of our proposed method.



(a) Scene and Query Point (b) 2D Responses (c) 3D Responses (d) 2D Responses (e) 3D Responses

Figure 3: **Visualization** of point responses before and after spatial interaction.

Table 2: **Comparison of classwise IoU** of the 3D results on vKITTI→sKITTI.

Class	DA			DG	
	UniDSeg	XDiff3D	Ours	UniDSeg	Ours
Veg	60.66	61.18	**85.64**	61.45	**83.94**
Buil	53.37	58.04	**69.80**	36.82	**45.94**
Road	82.24	83.26	**94.81**	66.96	**94.89**
Obj	4.64	5.09	**45.28**	13.09	**34.56**
Tru	24.53	30.02	**34.92**	2.44	**16.53**
Car	80.13	**81.64**	77.68	71.93	**81.99**
mIoU	50.9	53.1	**68.02**	42.12	**59.64**

Table 3: **Component ablation** on DG. 2DDM and 3DDM denote the 2D and 3D diffusion models, respectively.

3DDM	2DDM	SRA	BIR	SSM	vKITTI/sKITTI			nuScenes:USA/Sing		
					2D	3D	xM	2D	3D	xM
					57.6	44.7	60.0	66.5	64.5	72.3
✓					57.9	33.0	60.9	66.3	64.4	71.8
	✓				57.9	43.2	62.0	67.1	64.7	72.6
✓	✓	✓			57.4	46.0	61.3	66.4	64.9	72.4
✓	✓	✓	✓		58.0	58.7	63.8	67.8	69.4	73.5
✓	✓	✓	✓	✓	**60.1**	**59.6**	**65.4**	**70.9**	**70.8**	**75.0**

Table 4: Comparison of different diffusion models (DM) on vKITTI → sKITTI.

DM	XDiff3D	Ours
SD 1.4	46.1	**58.4**
SD 1.5	46.6	**59.6**
SD 2.1	46.9	**58.5**

Table 5: Comparison of different fusion approaches.

Method	2D	3D	xM
Cross-Attention	59.3	54.1	60.5
Concat-MLP	58.2	56.9	63.2
Add	**60.9**	55.9	63.3
Gated Residual	58.0	58.7	63.8
SSM	60.1	**59.6**	**65.4**

Table 6: Ablation study on different diffusion timesteps t^{2D} and t^{3D}.

Timestep	t^{2D}	t^{3D}
50	**65.4**	64.8
100	64.5	64.4
150	64.8	**65.4**
250	64.0	63.9
500	62.1	64.3

Table 7: Ablation study on different d.

d	mIoU (%)
32	60.7
64	61.7
128	**65.4**
256	63.9
512	64.8
1024	65.0

Qualitative Comparison. Figure 2 is the qualitative comparison in four benchmarks. Our method demonstrates significant advantages, maintaining more accurate segmentation in spatial regions with challenging domain shifts, which shows superior generalization in 3D scene segmentation.

4.3 Ablation Study and Analysis

Component Ablation. Table 3 evaluates the contribution of each component under two benchmark settings. Using UniDSeg as the baseline, 2DDM yielded performance gains due to its rich image semantic priors. While 3DDM is rich in geometric information, using it in isolation can disrupt the semantic information within the backbone. The performance gains observed upon adding BIR highlight the advantage of interaction, going beyond mere representation alignment. Finally, the inclusion of SSM leads to significant performance improvements, indicating that effectively enhancing semantic information through cross-modal spatial relationships can substantially boost the generalization capability of 3D segmentation.

Visual Verification. Figure 3 visualizes the impact of 2D and 3D feature interactions, which compares the top 10% of response points similar to the query point. After the interaction, both the 2D and 3D branches more strongly focus on high-affinity responses on the true semantic categories, demonstrating the effectiveness of spatial interactions.

Impact of Different Fusion Approaches. Table 5 compares the impact of SSM with other fusion approaches on DG3SS. The proposed SSM achieves the highest mIoU for both 3D and xM. Although direct addition yielded higher 2D scores, the lack of spatial information limited its effectiveness in 3D and xM results. The proposed method better preserves cross-modal advantages during the enhancement process to improve 3D segmentation.

Parameter Ablation. Table 6 studies the impact of timestep t^{2D} and t^{3D}. Under fixed one to validate the other experiments, $t^{3D} = 150$ and $t^{2D} = 50$ yielded the best performance. Table 7 studies the shared dimension d in BIR. Experimental results show that $d = 128$ provides the best performance in a relatively compact interaction space.

5 Conclusions

In this paper, a dual-diffusion spatial interaction representation (D^{2}SIR) is proposed to explore cross-modal diffusion representations to enhance 3D spatial semantics, which improves the generalization of 3D segmentation. The proposed SRA, BIR, and SSM modules employ an alignment-interaction-enhancement process that effectively interacts dual-diffusion representations within an aligned 3D space and enhances semantics based on valid spatial interaction, thereby enhancing 3D semantic generalization. Nevertheless, the proposed method only explores how diffusion features enhance 3D spatial semantic representations. Future work could incorporate physical world principles and explore world model priors to reinforce spatial semantics and further enhance generalization.

Reproducibility statement

The model formulation and training objective are described in Section 3. Section 4.1 specifies the datasets, segmentation backbones, and pretrained diffusion models. The supplementary material provides the dataset splits, shared label mappings, image–point correspondence, task-specific training schedules, detailed diffusion activation and model configurations. Together, these descriptions specify the method and experimental setup used in this work. We will release the source code to facilitate reproduction of our results.

Ethics statement

This work uses existing synthetic and real-world driving benchmarks for offline evaluation of 3D semantic segmentation. It does not collect new data or release raw sensor recordings. Access to the original datasets remains governed by their providers' terms. The reported results evaluate benchmark segmentation accuracy and do not constitute a safety assessment of an autonomous driving system.

AI use statement

In this work, we used generative AI tools to assist in discussing the methodology and experiment, revising related code, and analyzing and interpreting selected experimental results. We also used it to prepare scientific visualizations from experimental data, and polish the manuscript and assist in the latex preparation of mathematical formulas and tables. The authors checked the adopted method descriptions against the implementation, reviewed the relevant code and execution outputs, compared reported numerical results with experimental records, and verified the provenance of figures, cited sources, and final text. AI outputs were not treated as independent evidence for experimental measurements or bibliographic facts. The authors take full responsibility for the final submission.

References

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*, pp. 9297–9307, 2019.

Yasser Benigmim, Subhankar Roy, Slim Essid, Vicky Kalogeiton, and Stéphane Lathuilière. Collaborating foundation models for domain generalized semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 3108–3119, 2024.

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*, pp. 11621–11631, 2020.

Pu Cao, Feng Zhou, Qing Song, and Lu Yang. Controllable generation with text-to-image diffusion models: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2025.

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 Workshops*, pp. 98–109, 2023. doi: 10.1109/CVPRW59228.2023.00015.

Hao Dong, Moru Liu, Kaiyang Zhou, Eleni Chatzi, Juho Kannala, Cyrill Stachniss, and Olga Fink. Advances in multimodal adaptation and generalization: From traditional approaches to foundation models. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2026.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations*, 2021.

Michael Fuest, Pingchuan Ma, Ming Gui, Johannes Schusterbauer, Vincent Tao Hu, and Björn Ommer. Diffusion models and representation learning: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2026.

Adrien Gaidon, Qiao Wang, Yohann Cabon, and Eleonora Vig. Virtual worlds as proxy for multi-object tracking analysis. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 4340–4349, 2016.

Jakob Geyer, Yohannes Kassahun, Mentar Mahmudi, Xavier Ricou, Rupesh Durgesh, Andrew S Chung, Lorenz Hauswald, Viet Hoang Pham, Maximilian Mühlegg, Sebastian Dorn, et al. A2d2: Audi autonomous driving dataset. *arXiv preprint arXiv:2004.06320*, 2020.

Benjamin Graham, Martin Engelcke, and Laurens van der Maaten. 3D semantic segmentation with submanifold sparse convolutional networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 9224–9232, 2018.

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, 2021.

Pei He, Licheng Jiao, Lingling Li, Xu Liu, Fang Liu, Wenping Ma, Shuyuan Yang, and Ronghua Shang. Domain generalization-aware uncertainty introspective learning for 3d point clouds segmentation. In *Proceedings of the 32nd ACM International Conference on Multimedia*, pp. 651–660, 2024.

Pei He, Lingling Li, Licheng Jiao, Ronghua Shang, Fang Liu, Shuang Wang, Xu Liu, and Wenping Ma. Domain-aware category-level geometry learning segmentation for 3d point clouds. In *2025 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 28324–28333. IEEE, 2025.

Shuting He, Peilin Ji, Yitong Yang, Changshuo Wang, Jiayi Ji, Yinglin Wang, and Henghui Ding. A survey on 3d gaussian splatting applications: Segmentation, editing, and generation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2026.

Maximilian Jaritz, Tuan-Hung Vu, Raoul de Charette, Émilie 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*, pp. 12602–12611, 2020. doi: 10.1109/CVPR42600.2020.01262.

Yuru Jia, Lukas Hoyer, Shengyu Huang, Tianfu Wang, Luc Van Gool, Konrad Schindler, and Anton Obukhov. DGInStyle: Domain-generalizable semantic segmentation with image diffusion models and stylized semantic control. In *Computer Vision – ECCV 2024*, volume 15097 of *Lecture Notes in Computer Science*, pp. 91–109. Springer, 2024. doi: 10.1007/978-3-031-72933-1_6.

Hyeonseong Kim, Yoonsu Kang, Changgyoon Oh, and Kuk-Jin Yoon. Towards enhanced representation learning for single-source domain generalization in lidar semantic segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2026.

Danyang Li, Tianhao Wu, Bin Lin, Zhenyuan Chen, Yang Zhang, Yuxuan Li, Ming-Ming Cheng, and Xiang Li. Wow-seg: A word-free open world segmentation model. In *International Conference on Learning Representations*, volume 2026, pp. 155933–155958, 2026.

Fan Li, Xuan Wang, Min Qi, Zhaoxiang Zhang, and Yuelei Xu. Better to teach than to give: Domain generalized semantic segmentation via agent queries with diffusion model guidance. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 36129–36139. PMLR, 2025a.

Fan Li, Xuan Wang, Xuanbin Wang, Zhaoxiang Zhang, and Yuelei Xu. No object is an island: Enhancing 3D semantic segmentation generalization with diffusion models. In *Advances in Neural Information Processing Systems*, volume 38, pp. 75836–75857, 2025b.

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*, pp. 3829–3837, 2022. doi: 10.1145/3503161.3547990.

Miaoyu Li, Yachao Zhang, Xu Ma, Yanyun Qu, and Yun Fu. BEV-DG: Cross-modal learning under bird's-eye view for domain generalization of 3D semantic segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 11632–11642, 2023.

Ying Li, Lingfei Ma, Zilong Zhong, Fei Liu, Michael A. Chapman, Dongpu Cao, and Jonathan Li. Deep learning for lidar point clouds in autonomous driving: A review. *IEEE Transactions on Neural Networks and Learning Systems*, 32(8):3412–3432, 2021.

Yiyi Liao, Jun Xie, and Andreas Geiger. KITTI-360: A novel dataset and benchmarks for urban scene understanding in 2D and 3D. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(3):3292–3310, 2023.

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. doi: 10.1016/j.isprsjprs.2021.04.012.

Benyuan Meng, Qianqian Xu, Zitai Wang, Xiaochun Cao, and Qingming Huang. Not all diffusion model activations have been evaluated as discriminative features. In *Advances in Neural Information Processing Systems*, volume 37, pp. 55141–55177, 2024. doi: 10.52202/079017-1750.

Joshua Niemeijer, Manuel Schwonberg, Jan-Aike Termöhlen, Nico M. Schmidt, and Tim Fingscheidt. Generalization by adaptation: Diffusion-based domain extension for domain-generalized semantic segmentation. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pp. 2830–2840, 2024.

Changgyoon Oh, Hyeonseong Kim, Daehyun We, Jongoh Jeong, Yujeong Chae, and Kuk-Jin Yoon. DOODLE: Diffusion-based out-of-distribution learning for open-set LiDAR semantic segmentation. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pp. 2273–2283, 2026. doi: 10.1109/WACV61042.2026.00225.

Junsung Park, Kyungmin Kim, and Hyunjung Shim. Rethinking data augmentation for robust lidar semantic segmentation in adverse weather. *European Conference on Computer Vision*, 2024.

Junsung Park, Hwijeong Lee, Inha Kang, and Hyunjung Shim. No thing, nothing: Highlighting safety-critical classes for robust lidar semantic segmentation in adverse weather. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 6690–6699. IEEE, 2025.

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*, pp. 7108–7117, 2021.

Ethan Perez, Florian Strub, Harm de Vries, Vincent Dumoulin, and Aaron Courville. FiLM: Visual reasoning with a general conditioning layer. *Proceedings of the AAAI Conference on Artificial Intelligence*, 32(1):3942–3951, 2018. doi: 10.1609/aaai.v32i1.11671.

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 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 27325–27335, 2025.

Haoxi Ran, Vitor Guizilini, and Yue Wang. Towards realistic scene generation with LiDAR diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 14738–14748, 2024.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 10684–10695, 2022.

Cristiano Saltori, Aljosa Osep, Elisa Ricci, and Laura Leal-Taixè. Walking your LiDOG: A journey through multiple domains for LiDAR semantic segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 196–206, 2023.

Christoph Schuhmann, Romain Beaumont, Richard Vencu, Cade Gordon, Ross Wightman, Mehdi Cherti, Theo Coombes, Aarush Katta, Clayton Mullis, Mitchell Wortsman, Patrick Schramowski, Srivatsa Kundurthy, Katherine Crowson, Ludwig Schmidt, Robert Kaczmarczyk, and Jenia Jitsev. LAION-5B: An open large-scale dataset for training next generation image-text models. In *Advances in Neural Information Processing Systems*, volume 35, pp. 25278–25294, 2022.

Luming Tang, Menglin Jia, Qianqian Wang, Cheng Perng Phoo, and Bharath Hariharan. Emergent correspondence from image diffusion. In *Advances in Neural Information Processing Systems*, volume 36, pp. 1363–1389, 2023.

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*, pp. 490–498, 2023. doi: 10.1145/3581783.3612013.

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. In *Advances in Neural Information Processing Systems*, volume 37, pp. 101223–101249, 2024. doi: 10.52202/079017-3210.

Yao Wu, Mingwei Xing, Yachao Zhang, Fangyong Wang, Xiaopei Zhang, and Yanyun Qu. Beyondsparse: Facilitating mamba to enhance cross-domain 3d semantic segmentation in adverse weather. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 10826–10834, 2026.

Zhengkang Xiang, Zizhao Li, Amir Khodabandeh, and Kourosh Khoshelham. SG-LDM: Semantic-guided LiDAR generation via latent-aligned diffusion. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 24965–24976, 2025.

Aoran Xiao, Jiaxing Huang, Dayan Guan, Xiaoqin Zhang, Shijian Lu, and Ling Shao. Unsupervised point cloud representation learning with deep neural networks: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(9):11321–11339, 2023a.

Aoran Xiao, Jiaxing Huang, Weihao Xuan, Ruijie Ren, Kangcheng Liu, Dayan Guan, Abdulmotaleb El Saddik, Shijian Lu, and Eric P Xing. 3d semantic segmentation in the wild: Learning generalized models for adverse-condition point clouds. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 9382–9392, 2023b.

Jingyi Xu, Weidong Yang, Lingdong Kong, Youquan Liu, Qingyuan Zhou, Rui Zhang, Zhijun Li, Wen-Ming Chen, and Ben Fei. Visual foundation models boost cross-modal unsupervised domain adaptation for 3D semantic segmentation. *IEEE Transactions on Intelligent Transportation Systems*, 26(11):20287–20301, 2025. doi: 10.1109/TITS.2025.3587430.

Longyu Yang, Shangbo Yuan, Lu Zhang, Jun Liu, Heng tao Shen, Xiaofeng Zhu, and Ping Hu. Explicit geometry-reflectance domain shift modeling for robust lidar segmentation in adverse weather. *International Journal of Computer Vision*, 134(8):363, 2026.

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*, pp. 3338–3346, 2022. doi: 10.1145/3503161.3547987.

Haimei Zhao, Jing Zhang, Zhuo Chen, Shanshan Zhao, and Dacheng Tao. Unimix: Towards domain adaptive and generalizable lidar semantic segmentation in adverse weather. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 14781–14791, 2024.

Chaoyang Zhu and Long Chen. A survey on open-vocabulary detection and segmentation: Past, present, and future. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 46(12): 8954–8975, 2024.