BeyondSparse: Facilitating Mamba to Enhance Cross-Domain 3D Semantic Segmentation in Adverse Weather

Yao Wu[1], **Mingwei Xing**[2], **Yachao Zhang**[3*], **FangyongWang**[4], **Xiaopei Zhang**[5], **Yanyun Qu**[2,3*]

[1]College of Computer and Data Science, Fuzhou University
[2]Key Laboratory of Multimedia Trusted Perception and Efficient Computing, Ministry of Education of China, Institute of Artificial Intelligence, Xiamen University
[3]School of Informatics, Xiamen University
[4]Hanjiang National Laboratory
[5]University of California
wuyao@fzu.edu.cn, yyqu@xmu.edu.cn

Abstract

Domain generalization (DG) and domain adaptation (DA) for 3D semantic segmentation enable the model to maintain high performance while avoiding labor-intensive and time-consuming annotation of target-domain data. However, under adverse weather conditions, the injection of spatial noise will affect the reflectivity of LiDAR point clouds, exacerbate domain distribution discrepancies, and degrade the generalization ability of the model. Current methods mainly rely on sparse convolution-based architecture. Due to its limited receptive field, the model captures varying local geometric information when dealing with point clouds of different sparsities, thereby limiting its transferability. To this end, we propose BeyondSparse, a novel cross-domain 3D semantic segmentation method under adverse weather that incorporates a state-space model into a 3D sparse convolution-based architecture, sequentially modeling all features to learn domain-invariant representations. This method consists of two main components: domain feature decoupling and Mamba-based encoder. The former performs feature disentanglement before sequential modeling, while the latter performs global modeling on voxelized point cloud data. In addition, we introduce a token-style augmentation to capture the intrinsic properties of input data. Extensive experimental results demonstrate that our method outperforms SOTA competitors in both DG and DA tasks, for instance, achieving +4.6% and +0.8% mIoU on "SynLiDAR to SemanticSTF".

Code — https://github.com/Barcaaaa/BeyondSparse

Introduction

3D semantic segmentation interprets the distribution of objects in 3D space and is widely applied in fields such as autonomous driving and robotics (Zhang et al. 2021, 2024, 2025). Despite existing sparse convolution-based models achieving high-precision segmentation under specific conditions (Tang et al. 2020; Choy, Gwak, and Savarese 2019), they lack adaptability to complex and unseen scenarios. Especially in adverse weather conditions (*e.g.*, fog, snow, and rain), which can affect the spatial positioning, reflectivity, and semantic information of LiDAR point clouds (Xiao et al. 2023) (See Figure 1). Therefore, it is crucial to develop a reliable 3D model that can handle various adverse weather.

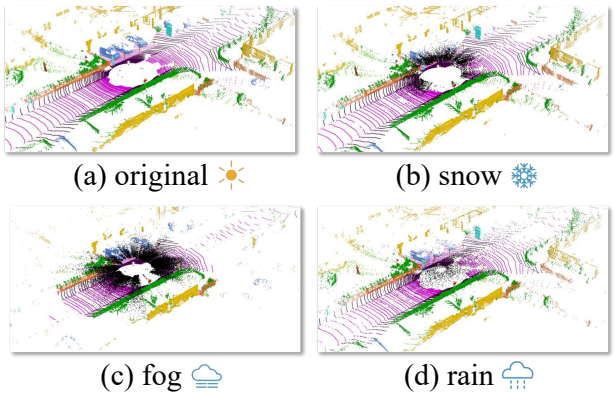

Figure 1: Visualization of LiDAR point cloud under three simulated adverse weather conditions. The black region indicates noisy points affected by weather.

Domain generalization and domain adaptation are two effective transfer learning techniques that address domain distribution discrepancies caused by the complex structural properties of point cloud data. Current methods primarily focus on designing a mix-based paradigm within specific domains. For instance, PointDR (Xiao et al. 2023) employed random geometric transformations on point clouds and integrated them with aggregation embeddings, while UniMix (Zhao et al. 2024) simulated weather conditions to construct a bridge domain between the source domain and the target domain. However, these methods mainly rely on sparse convolution-based architectures, whose limited receptive fields hinder their ability to capture global information. Global contextual information can boost model recognition of common features across domains and improve its generalization ability for unknown environments.

Recently, Mamba (Gu and Dao 2023) effectively captures long-term dependencies and global information with-

*Corresponding Author.

out quadratic complexity. Inspired by its success in image and point cloud classification (Zhu et al. 2024; Liu et al. 2024a), we aim to use Mamba for mitigating domain shifts in 3D semantic segmentation. However, directly integrating the global modeling of Mamba into existing 3D models poses a significant challenge. Besides, within the Mamba, the hidden state plays a crucial role in storing historical information from sequential data, which facilitates the modeling of long-range associations and enhances the global receptive field of the model (Long et al. 2024). However, when faced with data from unknown domains, hidden states may introduce domain-specific feature representations to the sequential data, leading to noise accumulation and thus weakening the generalization ability of Mamba to handle distributional shifts. Consequently, how to extract domain-invariant representation for Mamba modeling to mitigate the accumulation of noise is another challenge.

For the first challenge, we consider integrating the Mamba block into existing 3D segmentation models via a series of lightweight adapters. Our objective is to maintain the original inference efficiency while enhancing its global modeling abilities. Regarding the second challenge, we aim to decompose the features into domain-invariant and domain-specific components. By leveraging only domain-invariant features for global modeling, noise accumulation is mitigated.

In this work, we propose **BeyondSparse**, which leverages Mamba to construct structural priors **beyond** the abilities of existing 3D **sparse**-based architecture while simultaneously exploiting long-range dependency modeling to transcend limited voxel receptive fields for enhanced domain-invariant representation learning. To tackle error accumulation from domain-specific information in unknown domains, we introduce a Domain Feature Decoupling (DFD) module that separates features into domain-invariant and domain-specific components, using only the domain-invariant representation to improve model generalization ability. Then, we introduce a Mamba-based Encoder (ME) to perform global modeling on voxelized point cloud data, integrating Mamba layers with 3D sparse convolution blocks to enhance the understanding of global structure. Additionally, we introduce Token Style Augmentation (TSA) that stylizes sequential data input to the Mamba layer, enriching feature representation. This pre-exposes target-relevant cues to the model during training, further enhancing its generalization ability.

Our main contributions can be summarized as follows:

- We propose BeyondSparse, a novel cross-domain 3D semantic segmentation method under adverse weather that incorporates a state-space model into a 3D sparse-based architecture, sequentially modeling all features to learn domain-invariant representation.
- DFD performs feature disentanglement before sequential modeling, reducing noise accumulation during sequential modeling. ME performs global modeling on voxelized point cloud data to enhance global structure understanding. TSA captures the intrinsic characteristics of input data to enrich feature representation.
- Extensive experiments demonstrate that our method outperforms SOTA competitors in both DG and DA tasks.

Related Work

3D Domain Generalization and Adaptation

To address performance degradation across domains, 3D DG has been extensively studied (Wu et al. 2024a). This learning paradigm aims to learn a robust model from one or multiple source domains, enabling it to perform well on unseen target domains. Kim et al. (Kim et al. 2023b) employed random subsampling of point cloud data to simulate unseen domains. 3DLabelProp (Sanchez, Deschaud, and Goulette 2023) propagated labels from past sequences to newly registered scans. PointDR (Xiao et al. 2023) randomized geometric styles and aggregated embeddings in adverse weather conditions, while UniMix (Zhao et al. 2024) simulated weather conditions to construct a bridge domain between the source and target domains. Unlike 3D DG, 3D DA can access target-domain data but lacks the corresponding labels (Wu et al. 2024b, 2025). Mix3D (Nekrasov et al. 2021) directly concatenated two samples, while Cuboid Mixing (Ding et al. 2022) divided the scenes into several cuboids for mixing across domains. CosMix (Saltori et al. 2023) and ConDA (Kong, Quader, and Liong 2023) constructed an intermediate domain by utilizing joint supervision signals from both the source and target domains. Our endeavor is tailored to learn 3D domain-invariant representations using global modeling of Mamba.

State Space Models

Transformer (Vaswani et al. 2017) has revolutionized the underlying architecture for deep learning but suffers from quadratic computational complexity. To address this, more efficient operators are proposed, such as linear attention (Wang et al. 2020), Flash Attention (Dao et al. 2022), and State Space Model (SSM) series (Gu and Dao 2023; Gu, Goel, and Ré 2022; Nguyen et al. 2022). Especially for Mamba (Gu and Dao 2023), it stands out for its integration of selective mechanisms, which effectively capture long-range dependencies and process large-scale data in linear time. This innovation had extended into the visual domain through variants like VMamba (Liu et al. 2024b), which incorporated a cross-scan module. In point cloud processing, PointMamba (Liang et al. 2024) enhanced global modeling of point clouds by rearranging input patches based on a 3D Hilbert curve, while Point Mamba (Liu et al. 2024a) utilized an octree-based ordering to efficiently capture spatial relationships. These advances highlight the proficiency of Mamba in handling large-scale point cloud data.

Method

Problem Definition

We tackle the problem of 3D DG and DA, in which we have access to labeled source data $\mathcal{S} = \{x_i, y_i\}_{i=1}^{N_s}$ and unlabeled target data $\mathcal{T} = \{x_i\}_{i=1}^{N_t}$ used solely for DA. x signifies a LiDAR point cloud scan, y represents the corresponding point-wise semantic annotations, N_s and N_t represent the size of the source and target domain data, respectively. For DG, the objective is to develop a segmentation model $\mathcal{F}$ that effectively generalize to point clouds originating from $\mathcal{T}$ by

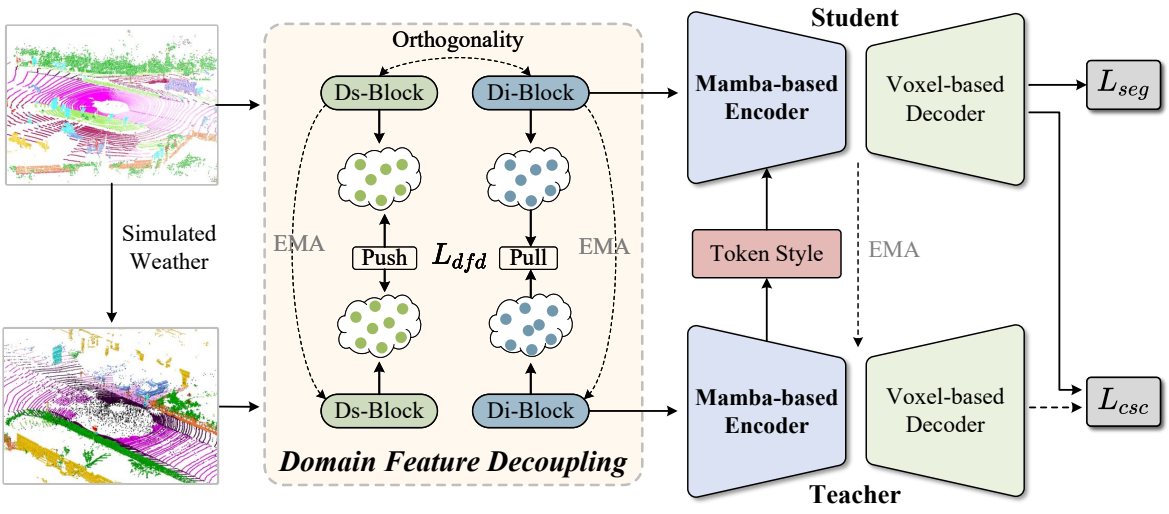

Figure 2: The overall framework of BeyondSparse primarily consists of three components: Domain Feature Decoupling, Mamba-based Encoder, and Token Style Augmentation. Only the Mamba-based Encoder is employed during the inference.

exclusively utilizing data from $\mathcal{S}$, *i.e.*, $\mathcal{F} : \mathcal{S} \rightarrow \hat{y}_T$. For DA, the model $\mathcal{F}$ leverages both source and target domain data to predict 3D target labels, represented as $\mathcal{F} : \mathcal{S} \cup \mathcal{T} \rightarrow \hat{y}_T$.

Overall Framework

As illustrated in Figure 2, we first apply specific data augmentations to the point clouds from the source domain to simulate challenging weather conditions, such as rain (Rasshofer, Spies, and Spies 2011; Teufel et al. 2022), snow (Hahner et al. 2022), and fog (Hahner et al. 2021), to closely mimic real-world adverse scenarios. Then, we design a dual-branch network, where the student branch processes the original source-domain data, while the teacher branch processes the augmented source-domain data. The teacher branch updates its weights using an Exponential Moving Average (EMA) mechanism. Through Domain Feature Decoupling, the point cloud features are decomposed into domain-invariant feature f_i and domain-specific feature f_s. Only f_i is fed into a Mamba-based Encoder for feature extraction, ensuring the domain-invariant property of the learning features. Additionally, a token-style augmentation is employed to incorporate features generated by the teacher branch, thereby enhancing those from the student branch. Finally, a voxel-based decoder is adopted to perform 3D semantic segmentation.

Domain Feature Decoupling

For the student branch, we design two structurally identical 3D sparse convolutional blocks: Ds-Block and Di-Block. These two blocks are responsible for extracting 3D domain-invariant features f_i and 3D domain-specific features f_s, respectively. Similarly, within the teacher branch, we obtain the EMA-updated domain-invariant features f_i^{ema} and domain-specific features f_s^{ema} from point clouds under simulated adverse weather conditions. Notably, Ds-Block and Di-Block should possess distinct parameters, thereby providing two different views of feature representation. Specifically, we enforce the orthogonality of the weights (w_i and w_s) of two blocks by minimizing their cosine similarity. This constraint is intended to facilitate the learning process of internal representations within the model, allowing models with similar architectures to develop the capability for capturing domain-specific and domain-invariant representations. Hence, the weight discrepancy loss $\mathcal{L}_{wd}$ is as follows:

$$\mathcal{L}_{wd} = \frac{\overrightarrow{w_i} \cdot \overrightarrow{w_s}}{\|\overrightarrow{w_i}\| \, \|\overrightarrow{w_s}\|}, \tag{1}$$

Furthermore, we minimize the Manhattan distance between f_i and f_i^{ema} to boost their cross-domain consistency. Conversely, we maximize the Manhattan distance between f_s and f_s^{ema} to ensure that they can be distinguished. The constraint formula is as follows:

$$\mathcal{L}_i = -\frac{1}{K} \sum_{k=1}^{K} \left| f_{i,k} - f_{i,k}^{ema} \right|, \tag{2}$$

$$\mathcal{L}_s = \frac{1}{K} \sum_{k=1}^{K} \left| f_{s,k} - f_{s,k}^{ema} \right|, \tag{3}$$

where K is the length of the token embedding sequence. Finally, the domain feature decoupling loss is formulated as:

$$\mathcal{L}_{dfd} = \lambda_1 \mathcal{L}_{wd} + \lambda_2 \mathcal{L}_i + \lambda_3 \mathcal{L}_s, \tag{4}$$

where $\{\lambda_i\}_{i=1}^{3}$ represent the loss balance parameters.

Mamba-based Encoder

Mamba (Gu and Dao 2023) has been demonstrated to effectively capture long-range dependencies and global information while avoiding quadratic complexity (Zhu et al. 2024). To this end, we consider integrating Mamba with 3D sparse convolution for global modeling on voxelized point cloud data. As depicted in Figure 3, within each encoder layer, we

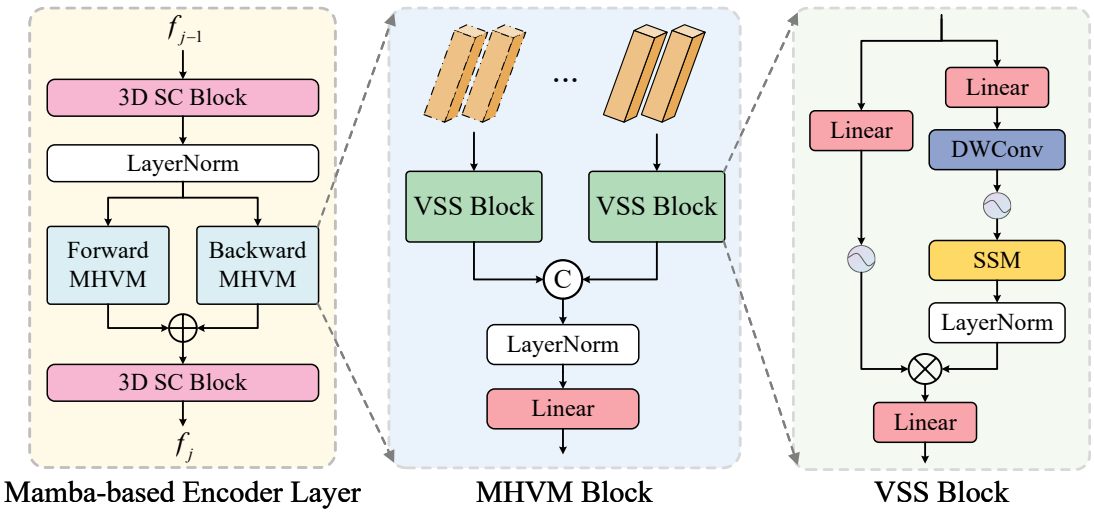

Figure 3: The architecture of Mamba-based Encoder, which mainly includes Multi-head Visual Mamba (MHVM) block and 3D sparse convolution (SC) block.

attach Multi-head Visual Mamba (MHVM) blocks $\mathcal{V}(\cdot)$ between two 3D sparse convolution (SC) blocks $\mathcal{SC}(\cdot)$. During the encoding process, we adopt a sequential scanning method along the 'XYZ' axes to capture local features. By performing forward and backward scanning of MHVM, we can extract geometric structure information from different directions, thereby improving the understanding of spatial context. This implementation is written as:

$$f^{'}_{j-1} = LN(\mathcal{SC}(f_{j-1})), \tag{5}$$

$$f_j = \mathcal{SC}(\mathcal{M}(f^{'}_{j-1}) + \mathcal{M}(reverse(f^{'}_{j-1}))), \tag{6}$$

where f_j denotes the token embedding at the j-th layer, $reverse(\cdot)$ denotes the backward scanning of input sequences, and $LN(\cdot)$ denotes Layer Normalization.

Concretely, the design of MHVM draws inspiration from the multi-head self-attention (Vaswani et al. 2017). It initially divides the input token sequence into independent subsets along the feature dimension. Each subset is then processed individually by the Visual State Space (VSS) block (Zhu et al. 2024), allowing the model to focus on information across various subspaces. After that, the outputs from all groups are concatenated along the feature dimension and subsequently undergo Layer Normalization and Linear Transformation to produce the final output representation at that layer. This design not only enhances the learning of fine-grained features but also maintains sensitivity to the global context, enabling the encoder to achieve superior performance while maintaining computational efficiency.

Token Style Augmentation

Adaptive Instance Normalization (AdaIN) (Huang and Belongie 2017), as a normalized feature stylization technique, has been widely applied in image processing and has been proven in enhancing model generalization across diverse datasets (Noori et al. 2024). Inspired by this, we treat each point-level feature as a token and use the statistics from the teacher branch to stylize the token in the student branch. Specifically, we select the corresponding samples from the teacher branch and synthesize the stylized features. This method adjusts the statistical properties of features, allowing the model to better capture the intrinsic characteristics of input, thus improving its robustness under varying conditions. The token style augmentation is formulated as:

$$\phi(f_j) = \gamma_{mix}\frac{f_j - \mu(f_j)}{\sigma(f_j)} + \beta_{mix}, \tag{7}$$

$$\gamma_{mix} = \alpha\sigma(f_j) + (1-\alpha)\sigma\left(f_j^{ema}\right), \tag{8}$$

$$\beta_{mix} = \alpha\mu(f_j) + (1-\alpha)\mu\left(f_j^{ema}\right), \tag{9}$$

where α denotes the mixing coefficient, which is obtained from the interval $[0, 1]$ using uniform sampling, $\mu(\cdot)$ and $\sigma(\cdot)$ are two functions returning channel-wise mean and standard deviation of input feature.

Overall Loss

The point-wise supervised segmentation loss of the source domain is formulated as follows:

$$\mathcal{L}_{seg} = -\frac{1}{N}\sum_{n=1}^{N}\sum_{c=1}^{C} y(n,c)\log P(n,c), \tag{10}$$

where C is the number of classes and P is the segmentation prediction. Then, we impose a consistency loss (He et al. 2020) to constrain the distribution discrepancy across branches, as follows:

$$\mathcal{L}_{csc} = -\frac{1}{N}\sum_{k=1}^{N}\log\frac{\exp(f_k\mathcal{B}^{+}/\tau)}{\sum_{j=1}^{C}\exp(f_k\mathcal{B}_j/\tau)}, \tag{11}$$

where τ is a temperature hyperparameter. The memory bank, denoted as $\mathcal{B}$, is updated via f^{ema} from the teacher branch. Samples within $\mathcal{B}$ that belong to the same class as f_k are considered positive samples and are indicated as $\mathcal{B}^{+}$.

Finally, the overall loss is as follows:

$$\mathcal{L} = \mathcal{L}_{seg} + \lambda_{csc}\mathcal{L}_{csc} + \mathcal{L}_{dfd}. \tag{12}$$

Method	car	bi.cle	mt.cle	truck	oth-v.	pers.	bi.clst	mt.clst	road	parki.	sidew.	oth-g.	build.	fence	veget.	trunk	terra.	pole	traf.	mIoU
Oracle	89.4	42.1	0.0	59.9	61.2	69.6	39.0	0.0	82.2	21.5	58.2	45.6	86.1	63.6	80.2	52.0	77.6	50.1	61.7	54.7
SemanticKITTI → SemanticSTF																				
Source-only	55.9	0.0	0.2	1.9	10.9	10.3	6.0	0.0	61.2	10.9	32.0	0.0	67.9	41.6	49.8	27.9	40.8	29.6	17.5	24.4
Dropout (Srivastava et al. 2014)	62.1	0.0	15.5	3.0	11.5	5.4	2.0	0.0	58.4	12.8	26.7	1.1	72.1	43.6	52.9	34.2	43.5	28.4	15.5	25.7
Perturbation (Xiao et al. 2023)	74.4	0.0	0.0	23.3	0.6	19.7	0.0	0.0	60.3	10.8	33.9	0.7	72.0	45.2	58.7	17.5	42.4	22.1	9.7	25.9
PolarMix (Xiao et al. 2022a)	57.8	1.8	3.8	16.7	3.7	26.5	0.0	2.0	65.7	2.9	32.5	0.3	71.0	48.7	53.8	20.5	45.4	25.9	15.8	26.0
MMD (Li et al. 2018)	63.6	0.0	2.6	0.1	11.4	28.1	0.0	0.0	67.0	14.1	37.9	0.3	67.3	41.2	57.1	27.4	47.9	28.2	16.2	26.9
PCL (Yao et al. 2022)	65.9	0.0	0.0	17.7	0.4	8.4	0.0	0.0	59.6	12.0	35.0	1.6	74.0	47.5	60.7	15.8	48.9	26.1	27.5	26.4
PointDR (Xiao et al. 2023)	67.3	0.0	4.5	19.6	9.0	18.8	2.7	0.0	62.6	12.9	38.1	0.6	73.3	43.8	56.4	32.2	45.7	28.7	27.4	28.6
DGLSS (Kim et al. 2023a)	72.6	0.1	11.7	29.4	13.7	**48.3**	0.5	21.2	65.0	**20.2**	38.3	3.8	**78.9**	51.8	57.0	**36.4**	47.0	26.9	34.9	34.6
UniMix (Zhao et al. 2024)	82.7	6.6	8.6	4.5	**15.1**	35.5	**15.5**	37.7	55.8	10.2	36.2	1.3	72.8	40.1	49.1	33.4	34.9	23.5	33.5	31.4
DGUIL (He et al. 2024)	77.9	**10.6**	19.1	26.0	9.7	46.3	6.0	9.3	**69.1**	18.0	38.6	**9.4**	73.3	51.2	60.8	30.9	50.8	**31.8**	22.3	35.5
BeyondSparse (Ours)	**83.8**	0.9	**29.8**	**32.9**	10.9	38.0	5.0	**67.6**	67.2	16.4	**42.9**	5.8	77.7	**51.8**	**60.8**	31.5	**51.2**	30.2	**42.3**	**39.4**
SynLiDAR → SemanticSTF																				
Source-only	27.1	3.0	0.6	15.8	0.1	25.2	1.8	5.6	23.9	0.3	14.6	0.6	36.3	19.9	37.9	17.9	41.8	9.5	2.3	15.0
Dropout (Srivastava et al. 2014)	28.0	3.0	1.4	9.6	0.0	17.1	0.8	0.7	34.2	6.8	19.1	0.1	35.5	19.1	42.3	17.6	36.0	14.0	2.8	15.2
Perturbation (Xiao et al. 2023)	27.1	2.3	2.3	16.0	0.1	23.7	1.2	4.0	27.0	3.6	16.2	0.8	29.2	16.7	35.3	22.7	38.3	17.9	5.1	15.2
PolarMix (Xiao et al. 2022a)	39.2	1.1	1.2	8.3	1.5	17.8	0.8	0.7	23.3	1.3	17.5	0.4	45.2	24.8	46.2	20.1	38.7	7.6	1.9	15.7
MMD (Li et al. 2018)	25.5	2.3	2.1	13.2	0.7	22.1	1.4	7.5	30.8	0.4	17.6	0.2	30.9	19.7	37.6	19.3	43.5	9.9	2.6	15.1
PCL (Yao et al. 2022)	30.9	0.8	1.4	10.0	0.4	23.3	**4.0**	7.9	28.5	1.3	17.7	1.2	39.4	18.5	40.0	16.0	38.6	12.1	2.3	15.5
PointDR (Xiao et al. 2023)	37.8	2.5	2.4	23.6	0.1	26.3	2.2	3.3	27.9	7.7	17.5	0.5	47.6	25.3	45.7	21.0	37.5	17.9	5.5	18.5
DGLSS (Kim et al. 2023a)	47.9	2.9	3.4	17.4	1.1	28.0	2.4	7.3	28.8	10.2	18.1	0.2	48.9	25.3	46.5	21.4	**45.2**	17.9	4.5	19.8
UniMix (Zhao et al. 2024)	**65.4**	0.1	**3.9**	16.9	**5.3**	**32.3**	2.0	**19.3**	52.1	5.0	**27.3**	**3.0**	49.4	20.3	58.5	22.7	23.2	26.9	10.4	23.4
DGUIL (He et al. 2024)	43.3	2.8	2.6	**23.2**	3.2	31.3	2.5	4.4	34.3	**9.2**	17.9	0.3	57.1	**27.6**	50.0	24.2	41.5	19.0	6.1	21.1
BeyondSparse (Ours)	46.6	**4.9**	2.9	19.6	4.9	30.4	1.7	3.2	**59.4**	7.4	22.0	1.7	**66.3**	22.5	**58.6**	**25.1**	28.5	**28.5**	**13.2**	**23.6**

Table 1: Quantitative results (mIoU, %) of two domain generalization scenarios. **Bold** indicates the best value.

Method	S.KITTI → STF.				SynLiDAR → STF.			
	D-fog	L-fog	Rain	Snow	D-fog	L-fog	Rain	Snow
Source-only	29.5	26.0	28.4	21.4	16.9	17.2	17.2	11.9
Dropout	29.3	25.6	29.4	24.8	15.3	16.6	20.4	14.0
Perturbation	26.3	27.8	30.0	24.5	16.3	16.7	19.3	13.4
PolarMix	29.7	25.0	28.6	25.6	16.1	15.5	19.2	15.6
MMD	30.4	28.1	32.8	25.2	17.3	16.3	20.0	12.7
PCL	28.9	27.6	30.1	24.6	17.8	16.7	19.3	14.1
PointDR	31.3	29.7	31.9	26.2	19.5	19.9	21.1	16.9
DGLSS	34.2	34.8	36.2	32.1	-	-	-	-
UniMix	33.5	34.8	30.2	**34.9**	**24.3**	22.9	26.1	20.9
DGUIL	36.3	34.5	35.5	33.3	-	-	-	-
BeyondSparse	**41.1**	**35.6**	**38.9**	34.3	21.6	**23.6**	**28.6**	**25.2**

Table 2: Quantitative results (mIoU, %) of two domain generalization scenarios in each weather-specific subset.

Experiment

Datasets

Similar to the testing benchmark employed by PointDR (Xiao et al. 2023), we conducted a series of experiments on two prevalent settings: SemanticKITTI→SemanticSTF and SynLiDAR→SemanticSTF. SemanticKITTI (Behley et al. 2019) collected LiDAR point clouds of urban landscapes under normal weather conditions. Our analysis is predicated on the training subset, which encompasses 19 semantic classes as the source domain dataset. SynLiDAR (Xiao et al. 2022b) presents synthetic point clouds derived from a variety of simulated environments. Characterized by its diversity in scene structures and configurations, this dataset comprises over 19 billion points. For our study, we chose the training subset, which is annotated with 19 semantic classes, as the source domain dataset. SemanticSTF (Xiao et al. 2023) collected LiDAR point clouds of urban landscapes under adverse weather conditions, including but not limited to fog (dense fog and light fog), rain, and snow. This dataset serves as the target domain for the evaluation.

Implementation Details

Following previous work (Xiao et al. 2023), we adopt the widely recognized MinkowskiNet (Choy, Gwak, and Savarese 2019) with sparse convolution (Tang et al. 2020) as the backbone. For optimization, we employ stochastic gradient descent (SGD) with a momentum of 0.9. The batch size is set to 4, and the initial learning rate is 0.24, with a decay factor of 0.0001. We utilize Intersection over Union (IoU) as the evaluation metric for each class, along with the mean IoU (mIoU) across all classes. All experiments are conducted using one NVIDIA RTX 3090 GPU. For hyperparameters, λ_{csc} and λ_1 are set to 0.1, λ_2 and λ_3 are set to 0.001.

Quantitative and Qualitative Comparison

Domain Generalization. Table 1 presents the quantitative results of two domain generalization scenarios. 'Oracle' refers to a model trained solely on the target-domain data, serving as an upper bound. 'Source-only' refers to a model trained solely on the source-domain data without any generalization techniques, serving as a lower bound. Compared to the baseline model, PointDR, our method increases mIoU by 10.8% and 5.1%. Compared to the latest method, DGUIL, our method still achieves remarkable results, exhibiting 3.9% and 2.5% mIoU improvement. This is because our method effectively models long-range dependencies to transcend limited voxel receptive fields. Besides, we

Method	car	bi.cle	mt.cle	truck	oth-v.	pers.	bi.clst	mt.clst	road	parki.	sidew.	oth-g.	build.	fence	veget.	trunk	terra.	pole	traf.	mIoU
Oracle	89.4	42.1	0.0	59.9	61.2	69.6	39.0	0.0	82.2	21.5	58.2	45.6	86.1	63.6	80.2	52.0	77.6	50.1	61.7	54.7
SemanticKITTI → SemanticSTF																				
Source-only	55.9	0.0	0.2	1.9	10.9	10.3	6.0	0.0	61.2	10.9	32.0	0.0	67.9	41.6	49.8	27.9	40.8	**29.6**	17.5	24.4
ADDA (Tzeng et al. 2017)	65.6	0.0	0.0	21.0	1.3	2.8	1.3	16.7	64.7	1.2	35.4	0.0	66.5	41.8	57.2	32.6	42.2	23.3	26.4	26.3
Ent-Min (Vu et al. 2019)	69.2	0.0	10.1	31.0	5.3	2.8	2.6	0.0	65.9	2.6	35.7	0.0	72.5	42.8	52.4	32.5	44.7	24.7	21.1	27.2
Self-training (Zou et al. 2019)	71.5	0.0	10.3	33.1	7.4	5.9	1.3	0.0	65.1	6.5	36.6	0.0	67.8	41.3	51.7	32.9	42.9	25.1	25.0	27.6
CoSMix (Saltori et al. 2023)	65.0	1.7	22.1	25.2	7.7	33.2	0.0	0.0	64.7	11.5	31.1	0.9	62.5	37.8	44.6	30.5	41.1	30.9	28.6	28.4
UniMix (Zhao et al. 2024)	75.3	0.9	44.9	11.7	**13.6**	38.2	**50.3**	**31.9**	71.1	15.0	**46.4**	6.5	74.3	**51.0**	49.8	36.8	34.4	25.5	28.9	37.2
BeyondSparse (Ours)	**87.4**	**6.2**	**50.4**	**45.1**	6.1	**56.7**	0.3	8.8	**74.4**	**15.9**	45.3	**29.4**	**75.2**	49.8	**66.0**	**39.2**	**64.5**	25.1	**47.7**	**41.8**
SynLiDAR → SemanticSTF																				
Source-only	27.1	3.0	0.6	15.8	0.1	25.2	1.8	5.6	23.9	0.3	14.6	0.6	36.3	19.9	37.9	17.9	**41.8**	9.5	2.3	15.0
ADDA (Tzeng et al. 2017)	55.8	0.0	3.6	26.1	1.3	25.2	7.5	9.9	17.2	**23.4**	4.4	0.9	43.9	18.4	45.2	21.8	33.6	28.0	19.7	20.3
Ent-Min (Vu et al. 2019)	48.3	0.1	5.6	**28.7**	0.1	23.3	2.5	19.8	19.3	6.7	22.6	1.4	46.9	20.7	43.2	25.2	34.1	26.0	22.2	20.9
Self-training (Zou et al. 2019)	50.6	0.0	6.1	31.0	0.5	26.0	4.8	12.0	20.7	4.6	23.5	1.5	45.3	19.5	44.6	25.0	35.1	29.2	20.8	21.1
CoSMix (Saltori et al. 2023)	51.5	0.2	5.0	28.1	0.0	26.5	**17.0**	9.9	20.2	3.6	24.6	2.2	52.6	20.6	47.5	24.3	34.6	28.2	24.1	22.1
UniMix (Zhao et al. 2024)	**73.6**	0.0	**7.9**	26.9	2.9	29.1	13.7	**21.8**	38.0	8.0	26.3	3.4	56.0	21.2	56.1	**29.6**	38.0	28.2	26.5	26.7
BeyondSparse (Ours)	70.7	**5.1**	4.9	7.8	**8.3**	**37.9**	0.2	0.7	**65.2**	1.3	**31.8**	**11.5**	**72.2**	**27.7**	**64.4**	27.9	25.5	**30.5**	**29.7**	**27.5**

Table 3: Quantitative results (mIoU, %) of two domain adaptation scenarios. **Bold** indicates the best value.

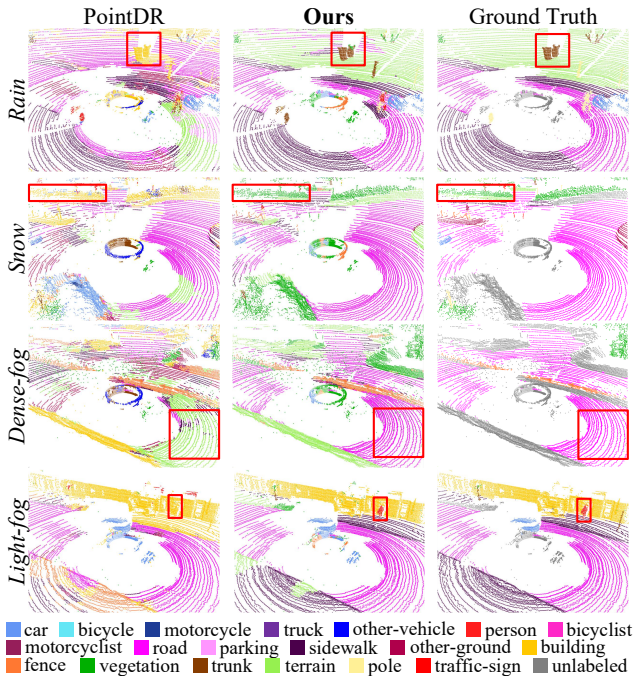

Figure 4: Qualitative results in adverse weather conditions.

evaluate the results in each weather-specific subset, including dense fog, light fog, rain, and snow. As shown in Table 2, it is observed that our method consistently achieves the best performance across multiple adverse weather conditions. For instance, in SemanticKITTI→SemanticSTF, our method surpasses 'DGUIL' with a gain of 4.8% mIoU under dense fog conditions.

Domain Adaptation. Under the DA setting, we have access to target-domain data. Hence, we employ a mixture of target-domain data and source-domain data augmented with weather simulation in the teacher branch. As shown in Table 3, compared with the SOTA method 'UniMix', our method achieves gains of 4.6% and 0.8% in mIoU, reaching top rankings of 41.8% and 27.5%.

	DFD	ME	TSA	mIoU (%)
PointDR				28.6
#1	✓			33.2
#2		✓		32.4
#3			✓	30.8
#4	✓	✓		36.5
#5		✓	✓	35.4
#6	✓		✓	37.2
#7	✓	✓	✓	39.4

Table 4: Ablation study on the effectiveness of each component in BeyondSparse for domain generalization.

Visualization Figure 4 validates our Table 2 results, showing BeyondSparse's superiority over DG methods. The red box indicates the differences in predictions. By the merit of global modeling from Mamba, our method segments detailed objects well. From top to bottom, the focus is the trunk of a tree, vegetation near the road, the profile of the road, and, most importantly, people walking safely on the road.

Ablation Studies

Different Components. As shown in Table 4, we conduct a series of experiments to evaluate the improvement of each component in BeyondSparse. When DFD is used alone (#1), the mIoU is increased by 4.6%, which highlights the role of feature decoupling. A 3.8% mIoU improvement is achieved using ME (#2), indicating that global modeling promotes the learning of spatial-invariant features of the model across domains. Performing TSA only on the Mamba layer can also enrich feature representation, yielding a 2.2% mIoU gain. The combined application of both modules (#4, #5, #6) yields a more overall improvement, reaching 36.5%, 35.4%, and 37.2% mIoU. It is observed that these three components

	Position	Cos. Sim.↓	mIoU (%)↑
#1	After Input	**0.139**	**33.2**
#2	Before Bottleneck	0.215	31.4
#3	After Bottleneck	0.236	31.0
#4	Before Output	0.322	30.2

Table 5: Impact of DFD at different positions.

Scenario	Module	mIoU (%)
SemanticKITTI→SemanticSTF	Di-Block	39.4
	Ds-Block	30.6
SynLiDAR→SemanticSTF	Di-Block	23.6
	Ds-Block	20.3

Table 6: Impact of decoupled features generated by DFD.

can work synergistically to not only distinguish domain-invariant and domain-specific features but also learn global invariant features, significantly improving the generalization ability of the model. Finally, combining the three proposed components to reach peak value (39.4% mIoU).

Role of Decoupling. As shown in Table 5, we evaluate the impact of DFD at different positions within the segmentation network. #1 involves performing feature disentanglement immediately after data input, positioning it ahead of the segmentation network. It is observed that this position-setting achieves the best performance, with the lowest inter-feature similarity between f_i and f_s, reaching a value of 33.2% mIoU. Besides, #2 and #3 examine the effects of incorporating DFD before and after the Bottleneck, which is situated between the encoder and the decoder. #4 add DFD before the output, *i.e.,* after the decoder but before the segmentation head. These results suggest that feature adjustment at an early stage better facilitates the learning of generalizable representations by the model. Moreover, as shown in Table 6, we input f_i and f_s generated by Di-Block and Ds-Block into ME for global modeling, respectively. It is observed that using f_i is higher than f_s. For instance, on SemanticKITTI→SemanticSTF, using Di-Block surpasses Ds-Block by 8.8% mIoU. It means that separating domain-specific features before the segmentation network enhances the ability of the model to discriminate information.

Error Accumulation Analysis. Domain-specific information tends to be accumulated or even amplified by the hidden states, leading to a decline in the generalization ability of the 3D model. The Kullback-Leibler (KL) divergence can be used to quantify the difference between two feature distributions. We employ an unsupervised domain adaptation framework to compute the KL divergence between the feature distributions of the source and target domains during training. This serves as an indicator of the discrepancy between the two distributions, thereby indirectly reflecting the accumulation of errors. As illustrated in Figure 5, when DFD is not utilized, the KL divergence between the source and target domains exhibits a significant upward trend after 100k iterations. This suggests that during the sequence modeling, error accumulation introduced by domain-specific features

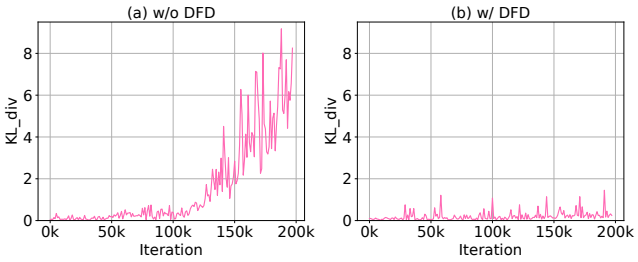

Figure 5: Error accumulation analysis during the global modeling of Mamba. 'KL_div' means KL divergence.

	L.5	L.4	L.3	L.2	L.1	mIoU (%)
	-	-	-	-	-	28.6
#1	✓	-	-	-	-	30.2
#2	✓	✓	-	-	-	31.8
#3	✓	✓	✓	-	-	**32.4**
#4	✓	✓	✓	✓	-	31.7
#5	✓	✓	✓	✓	✓	30.6

Table 7: Impact of MHVM blocks at different layers of ME.

leads to a gradual increase in the discrepancy between the two feature distributions. Conversely, when DFD is applied, and only domain-invariant features are used for sequence modeling, the discrepancy between the two feature distributions is confined within a certain range and does not show an increasing trend. This indicates that DFD not only effectively disentangles features but also addresses the issue of error accumulation in Mamba for cross-domain tasks.

Design of Mamba-based Encoder. As shown in Table 7, we evaluate the impact of integrating MHVM blocks at different layers 'L.x' of ME. Concretely, 'L.1' means the shallowest layer of ME, whereas 'L.5' corresponds to the deepest layer. It is observed that integrating MHVM at relatively deep layers (#3) is beneficial for improving performance (32.4% mIoU). This is because the global modeling of 3D deep features can be effectively learned by subsequent feature layers. Conversely, when MHVM is applied exclusively to the deepest layer (#1), performance declines. Moreover, an attempt to uniformly integrate MHVM across all layers of ME (#5) also results in a performance decrease. The introduction of MHVM at 'L.1', which is closer to the output, however, disrupts the encoding of basic visual representation, thereby degrading the quality of the final prediction.

Conclusion

In this work, we delve into cross-domain 3D semantic segmentation under adverse weather conditions, named BeyondSparse. For this purpose, we introduce Mamba to construct structural priors beyond the abilities of existing 3D sparse-based architecture while simultaneously exploiting long-range dependency modeling to transcend limited voxel receptive fields for domain-invariant representation learning. Our method achieves SOTA performance in both DA and DG, as proved by extensive experiments on two scenarios.

Acknowledgments

This work was supported in part by the National Natural Science Foundation of China under Grant 62176224, Grant 62176092, Grant 62222602, and Grant 62306165; in part by the Science and Technology on Sonar Laboratory under Grant 2024-JCJQ-LB-32/07.

References

Behley, J.; Garbade, M.; Milioto, A.; Quenzel, J.; Behnke, S.; Stachniss, C.; and Gall, J. 2019. Semantickitti: A dataset for semantic scene understanding of lidar sequences. In *ICCV*, 9297–9307.

Choy, C. B.; Gwak, J.; and Savarese, S. 2019. 4D Spatio-Temporal ConvNets: Minkowski Convolutional Neural Networks. In *CVPR*, 3075–3084.

Dao, T.; Fu, D. Y.; Ermon, S.; Rudra, A.; and Ré, C. 2022. FlashAttention: Fast and Memory-Efficient Exact Attention with IO-Awareness. In *NeurIPS*, 16344–16359.

Ding, R.; Yang, J.; Jiang, L.; and Qi, X. 2022. DODA: Data-Oriented Sim-to-Real Domain Adaptation for 3D Semantic Segmentation. In *ECCV*, 284–303.

Gu, A.; and Dao, T. 2023. Mamba: Linear-time sequence modeling with selective state spaces. *arXiv preprint arXiv:2312.00752*.

Gu, A.; Goel, K.; and Ré, C. 2022. Efficiently Modeling Long Sequences with Structured State Spaces. In *ICLR*, 1–27.

Hahner, M.; Sakaridis, C.; Bijelic, M.; Heide, F.; Yu, F.; Dai, D.; and Gool, L. V. 2022. LiDAR Snowfall Simulation for Robust 3D Object Detection. In *CVPR*, 16343–16353.

Hahner, M.; Sakaridis, C.; Dai, D.; and Van Gool, L. 2021. Fog simulation on real LiDAR point clouds for 3D object detection in adverse weather. In *ICCV*, 15283–15292.

He, K.; Fan, H.; Wu, Y.; Xie, S.; and Girshick, R. 2020. Momentum Contrast for Unsupervised Visual Representation Learning. In *CVPR*, 9726–9735.

He, P.; Jiao, L.; Li, L.; Liu, X.; Liu, F.; Ma, W.; Yang, S.; and Shang, R. 2024. Domain Generalization-Aware Uncertainty Introspective Learning for 3D Point Clouds Segmentation. In *ACM MM*, 651–660.

Huang, X.; and Belongie, S. J. 2017. Arbitrary Style Transfer in Real-Time with Adaptive Instance Normalization. In *ICCV*, 1510–1519.

Kim, H.; Kang, Y.; Oh, C.; and Yoon, K. 2023a. Single Domain Generalization for LiDAR Semantic Segmentation. In *CVPR*, 17587–17598.

Kim, H.; Kang, Y.; Oh, C.; and Yoon, K.-J. 2023b. Single domain generalization for lidar semantic segmentation. In *CVPR*, 17587–17598.

Kong, L.; Quader, N.; and Liong, V. E. 2023. ConDA: Unsupervised Domain Adaptation for LiDAR Segmentation via Regularized Domain Concatenation. In *ICRA*, 9338–9345.

Li, H.; Pan, S. J.; Wang, S.; and Kot, A. C. 2018. Domain Generalization With Adversarial Feature Learning. In *CVPR*, 5400–5409.

Liang, D.; Zhou, X.; Xu, W.; Zhu, X.; Zou, Z.; Ye, X.; Tan, X.; and Bai, X. 2024. PointMamba: A Simple State Space Model for Point Cloud Analysis. In *NeurIPS*, 32653–32677.

Liu, J.; Yu, R.; Wang, Y.; Zheng, Y.; Deng, T.; Ye, W.; and Wang, H. 2024a. Point mamba: A novel point cloud backbone based on state space model with octree-based ordering strategy. *arXiv preprint arXiv:2403.06467*.

Liu, Y.; Tian, Y.; Zhao, Y.; Yu, H.; Xie, L.; Wang, Y.; Ye, Q.; Jiao, J.; and Liu, Y. 2024b. VMamba: Visual State Space Model. In *NeurIPS*, 103031–103063.

Long, S.; Zhou, Q.; Li, X.; Lu, X.; Ying, C.; Luo, Y.; Ma, L.; and Yan, S. 2024. DGMamba: Domain Generalization via Generalized State Space Model. In *ACM MM*, 3607–3616.

Nekrasov, A.; Schult, J.; Litany, O.; Leibe, B.; and Engelmann, F. 2021. Mix3D: Out-of-Context Data Augmentation for 3D Scenes. In *3DV*, 116–125.

Nguyen, E.; Goel, K.; Gu, A.; Downs, G. W.; Shah, P.; Dao, T.; Baccus, S.; and Ré, C. 2022. S4ND: Modeling Images and Videos as Multidimensional Signals with State Spaces. In *NeurIPS*, 2846–2861.

Noori, M.; Cheraghalikhani, M.; Bahri, A.; Hakim, G. A. V.; Osowiechi, D.; Ayed, I. B.; and Desrosiers, C. 2024. TFS-ViT: Token-level feature stylization for domain generalization. *PR*, 149: 110213.

Rasshofer, R. H.; Spies, M.; and Spies, H. 2011. Influences of weather phenomena on automotive laser radar systems. *ARS*, 9: 49–60.

Saltori, C.; Galasso, F.; Fiameni, G.; Sebe, N.; Poiesi, F.; and Ricci, E. 2023. Compositional Semantic Mix for Domain Adaptation in Point Cloud Segmentation. *TPAMI*, 45(12): 14234–14247.

Sanchez, J.; Deschaud, J.-E.; and Goulette, F. 2023. Domain generalization of 3d semantic segmentation in autonomous driving. In *ICCV*, 18077–18087.

Srivastava, N.; Hinton, G.; Krizhevsky, A.; Sutskever, I.; and Salakhutdinov, R. 2014. Dropout: A Simple Way to Prevent Neural Networks from Overfitting. *JMLR*, 15(1): 1929–1958.

Tang, H.; Liu, Z.; Zhao, S.; Lin, Y.; Lin, J.; Wang, H.; and Han, S. 2020. Searching Efficient 3D Architectures with Sparse Point-Voxel Convolution. In *ECCV*, 685–702.

Teufel, S.; Volk, G.; Von Bernuth, A.; and Bringmann, O. 2022. Simulating realistic rain, snow, and fog variations for comprehensive performance characterization of lidar perception. In *VTC*, 1–7.

Tzeng, E.; Hoffman, J.; Saenko, K.; and Darrell, T. 2017. Adversarial Discriminative Domain Adaptation. In *CVPR*, 2962–2971.

Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A. N.; Kaiser, L.; and Polosukhin, I. 2017. Attention is All you Need. In *NeurIPS*, 5998–6008.

Vu, T. H.; Jain, H.; Bucher, M.; Cord, M.; and Perez, P. 2019. ADVENT: Adversarial Entropy Minimization for Domain Adaptation in Semantic Segmentation. In *CVPR*, 2517–2526.

Wang, S.; Li, B. Z.; Khabsa, M.; Fang, H.; and Ma, H. 2020. Linformer: Self-attention with linear complexity. *arXiv preprint arXiv:2006.04768*.

Wu, Y.; Xing, M.; Zhang, Y.; Luo, X.; Xie, Y.; and Qu, Y. 2024a. UniDSeg: Unified Cross-Domain 3D Semantic Segmentation via Visual Foundation Models Prior. In *NeurIPS*, 101223–101249.

Wu, Y.; Xing, M.; Zhang, Y.; Xie, Y.; and Qu, Y. 2024b. Clip2uda: Making frozen clip reward unsupervised domain adaptation in 3d semantic segmentation. In *ACMMM*, 8662–8671.

Wu, Y.; Xing, M.; Zhang, Y.; Xie, Y.; and Qu, Y. 2025. Fusion-Then-Distillation: Toward Cross-Modal Positive Distillation for Domain Adaptive 3D Semantic Segmentation. *TCSVT*, 35(9): 9030–9045.

Xiao, A.; Huang, J.; Guan, D.; Cui, K.; Lu, S.; and Shao, L. 2022a. PolarMix: A General Data Augmentation Technique for LiDAR Point Clouds. In *NeurIPS*, 11035–11048.

Xiao, A.; Huang, J.; Guan, D.; Zhan, F.; and Lu, S. 2022b. Transfer learning from synthetic to real lidar point cloud for semantic segmentation. In *AAAI*, 2795–2803.

Xiao, A.; Huang, J.; Xuan, W.; Ren, R.; Liu, K.; Guan, D.; El Saddik, A.; Lu, S.; and Xing, E. P. 2023. 3d semantic segmentation in the wild: Learning generalized models for adverse-condition point clouds. In *CVPR*, 9382–9392.

Yao, X.; Bai, Y.; Zhang, X.; Zhang, Y.; Sun, Q.; Chen, R.; Li, R.; and Yu, B. 2022. PCL: Proxy-based Contrastive Learning for Domain Generalization. In *CVPR*, 7087–7097.

Zhang, Y.; Hu, R.; Li, R.; Qu, Y.; Xie, Y.; and Li, X. 2024. Cross-modal match for language conditioned 3d object grounding. In *AAAI*, 7359–7367.

Zhang, Y.; Lan, Y.; Xie, Y.; Li, C.; and Qu, Y. 2025. Cross-cloud consistency for weakly supervised point cloud semantic segmentation. *TNNLS*, 36(8): 14452–14463.

Zhang, Y.; Qu, Y.; Xie, Y.; Li, Z.; Zheng, S.; and Li, C. 2021. Perturbed self-distillation: Weakly supervised large-scale point cloud semantic segmentation. In *ICCV*, 15520–15528.

Zhao, H.; Zhang, J.; Chen, Z.; Zhao, S.; and Tao, D. 2024. UniMix: Towards Domain Adaptive and Generalizable LiDAR Semantic Segmentation in Adverse Weather. In *CVPR*, 14781–14791.

Zhu, L.; Liao, B.; Zhang, Q.; Wang, X.; Liu, W.; and Wang, X. 2024. Vision Mamba: Efficient Visual Representation Learning with Bidirectional State Space Model. In *ICML*, 62429–62442.

Zou, Y.; Yu, Z.; Liu, X.; Kumar, B. V. K. V.; and Wang, J. 2019. Confidence Regularized Self-Training. In *ICCV*, 5981–5990.