UAWTrack: Universal 3D Single Object Tracking in Adverse Weather

Yuxiang Yang[1], Hongjie Gu[1], Yingqi Deng[1], Zhekang Dong[1], Zhiwei He[1], Jing Zhang[2*]

[1]School of Electronics and Information, Hangzhou Dianzi University, China
[2]School of Computer Science, Wuhan University, China
{yyx, hongjiegu, den, englishp, zwhe}@hdu.edu.cn, jingzhang.cv@gmail.com

Abstract

3D single object tracking (3D SOT) in LiDAR point clouds is essential for autonomous driving. Most existing 3D SOT methods focus on clear weather, where point clouds are more defined. However, adverse weather conditions lead to sparser and noisier point clouds, significantly degrading tracking performance and posing safety risks. In this study, we introduce **UAWTrack**, a universal 3D SOT model designed to perform effectively across diverse real-world weather conditions. UAWTrack comprises three key modules: 1) Voxel Feature Extraction, which mitigates the perturbations in point clouds caused by adverse weather; 2) Motion-centric Spatial-temporal Aggregation and Motion-guided Feature Fusion, capturing motion clues and sampling dense BEV motion features to address the issue of sparsity; and 3) Weather-Specific Tracker, which efficiently handles tracking in various weather conditions. To fill the gap of lacking benchmarks for 3D SOT in adverse weather, we simulate physically valid adverse weather conditions on the KITTI and NuScenes datasets, creating two benchmarks: KITTI-A and NuScenes-A. Extensive experiments demonstrate that UAWTrack achieves state-of-the-art performance under all weather conditions.

Code — https://github.com/HDU-VRLab/UAWTrack

Introduction

3D single object tracking (3D SOT) in LiDAR point clouds is a crucial task for autonomous robots, such as self-driving cars and drones. For the 3D SOT models to be safely used in autonomous driving, they must maintain reliable performance under diverse real-world weather conditions. However, adverse weather poses significant challenges due to sensor distortion. Specifically, the pulsed laser light from LiDAR struggles to penetrate water particles, leading to two key issues: 1) Attenuation reduces signal strength from solid objects, causing sparse 3D point clouds; and 2) Backscattering introduces false peaks in the received signal, leading to perturbations in 3D point clouds. As illustrated in Fig. 1, there is a noticeable disparity between point clouds under different weather conditions. Particularly, point clouds near the LiDAR are more susceptible to perturbation and noise, and the same tracked object may exhibit varying levels of sparsity depending on the weather.

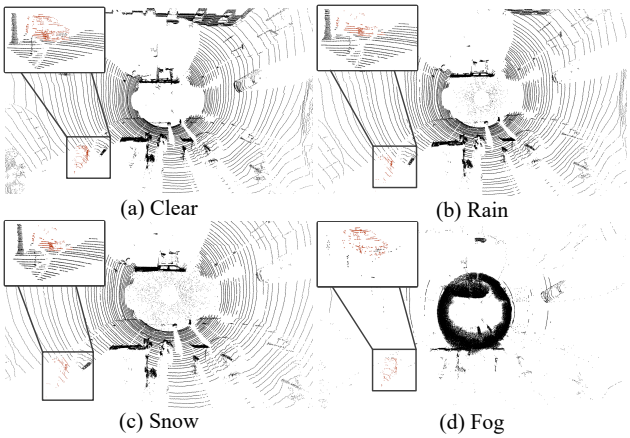

Figure 1: Visualization of point clouds from the same scene under varying weather conditions in our KITTI-A dataset.

While current 3D SOT models (Qi et al. 2020; Hui et al. 2021; Zheng et al. 2022; Yang et al. 2023) perform well in clear weather, they struggle to address the challenges posed by adverse conditions. For instance, methods like P2B (Qi et al. 2020), 3D-SiamRPN (Fang et al. 2020), BAT (Zheng et al. 2021), and PTT (Shan et al. 2021) follow a point-based pipeline, extracting point-wise discriminative features and calculating point-wise similarity through feature aggregation. However, given the inherent sparsity and irregularity of point clouds, which are further exacerbated by weather-induced issues like missing data and perturbations, point-wise feature encoding is not well-suited for adverse conditions. An intuitive solution is to use domain adaptation techniques to transfer knowledge to specific weather conditions, thereby mitigating the domain shift between the source (clear weather) and target domains (adverse weather). However, this strategy is less efficient when dealing with multiple weather types simultaneously, limiting its practical applicability.

To address the challenges mentioned above, we introduce UAWTrack, a versatile approach inspired by Mixture of Experts (MoE) (Shazeer et al. 2017) that learns weather-

*Corresponding author.
Copyright © 2025, Association for the Advancement of Artificial Intelligence (www.aaai.org). All rights reserved.

robust representations, enabling reliable target tracking under diverse weather conditions in a single model. First, we depart from the point-based pipelines used in previous 3D SOT methods to minimize information loss. Instead, we propose Voxel Feature Extraction (VFE) to capture dense BEV features from two consecutive frames, which are universal across weather conditions, mitigating the performance degradation caused by point clouds perturbations. Next, we employ the Motion-centric Spatial-temporal Aggregation (MSA) and Motion-Guided Feature Fusion (MFF) to capture motion information between sequential frames effectively. This approach overcomes the limitations of previous appearance-matching methods, which often fail due to significant appearance changes in adverse weather, as demonstrated by M^2Track (Zheng et al. 2022). By integrating dense BEV features at multiple scales and assigning multiple sampling points for global feature aggregation, we address the issues of sparsity and perturbations. Additionally, to enhance efficiency and precision, we designed a Weather Router (WR) that uses a learned gating function to selectively route inputs to the most relevant expert. This router works in conjunction with the Weather-Specific Tracker (WS-Tracker), composed of four experts, to handle specific input subspaces effectively.

Currently, there is a lack of real adverse weather datasets for 3D SOT tasks. To address this, we use simulation techniques similar to those in other fields (Hao et al. 2024; Kong et al. 2024) to generate datasets. Specifically, by applying physically-based adverse weather simulation algorithms (Hahner et al. 2021, 2022; Dong et al. 2023) to two classic autonomous driving datasets, KITTI (Geiger, Lenz, and Urtasun 2012) and NuScenes (Caesar et al. 2020), we create two synthetic datasets: KITTI-A and NuScenes-A. These datasets include three adverse weather conditions (rain, snow, and fog) and one clear weather condition. Our experiments on the datasets demonstrate that UAWTrack achieves state-of-the-art performance in both clear and adverse weather conditions. In summary, our contributions are as follows:

- We propose UAWTrack, the first 3D SOT model that offers universal adaptability and robustness across various weather conditions.
- We introduce VFE, MSA, and MFF to handle point cloud missing and perturbation issues in adverse weather through voxel representation and motion modeling.
- We present WS-Tracker and WR to address the domain gap between different adverse weather types, enhancing the performance of our model.

Related Work

3D Single Object Tracking. Current methods (Qi et al. 2020; Zhou et al. 2022; Ma et al. 2023; Xu et al. 2023b) for 3D single object tracking typically assume clear weather conditions and can be divided into two paradigms: Siamese and motion-centric. The Siamese paradigm, which includes SC3D (Giancola, Zarzar, and Ghanem 2019), P2B (Qi et al. 2020), BAT (Zheng et al. 2021), employs Siamese networks to extract and match features between the target template and the search region. SC3D, a pioneer in 3D SOT, compares the target template with multiple candidate patches in the current frame. P2B uses a Siamese network to encode features from both the template and the search region point clouds, enhancing target-specific features by incorporating clues from the template into the search area seed. BAT introduces a box-aware feature fusion module to enhance geometric features through point-to-box relations, while CXTrack (Xu et al. 2023a) uses a transformer-based network to capture and utilize contextual information. In contrast, the motion-centric paradigm, exemplified by M^2Track (Zheng et al. 2022), first segments the target point clouds and then explicitly models target motion. However, these traditional approaches are often vulnerable in adverse weather conditions.

3D Perception in Adverse Weather. In real-world applications like autonomous driving, adverse weather poses a major challenge to 3D perception. Significant research has been dedicated to tasks such as 3D object detection and LiDAR semantic segmentation. For example, STF (Bijelic et al. 2020) provides a multi-modal dataset for 3D object detection in adverse conditions and a fusion model that adaptively combines different modal features. SRKD (Huang et al. 2024) simulates the Waymo Open Dataset (WOD) using rain simulation and employs knowledge distillation to enhance the robustness of 3D detectors under rainy conditions. SemanticSTF (Xiao et al. 2023) extends STF by incorporating semantic labels. Unimix (Zhao et al. 2024) creates a bridge domain within a teacher-student framework to address domain gaps. However, research on adverse weather conditions in the 3D SOT field remains scarce.

Mixture-of-Experts (MoE). MoE employs multiple submodels (experts), each tailored to specific data types, with gating networks to determine the appropriate model for each data type, thereby reducing interference between different data types (Xu et al. 2023c). For instance, Sparsely-Gated MoE (Shazeer et al. 2017) integrates MoE with Long Short-Term Memory (LSTM) models, introducing Sparse MoE. GShard (Lepikhin et al. 2020) is the first to apply MoE to Transformers, replacing Feed Forward Network (FFN) layers with MoE structures, where each expert is an FFN. MoFME (Zhang et al. 2024), with its uncertainty-aware router and feature modulation expert, excels in multi-task deweathering settings. Inspired by these advancements, our proposed method leverages the MoE structure to handle various weather conditions in 3D SOT.

Proposed Method

In this paper, we propose a novel 3D SOT model named UAWTrack that offers universal adaptability and robustness across various weather conditions. As shown in Fig. 2, UAWTrack comprises (a) VFE, (b) MSA, and (c) MFF to extract effective features across different weather conditions. It further utilizes (e) WS-Tracker, which consists of four parallel experts, to enhance the adaptability of the model to diverse weather conditions. Specifically, the aggregated features are routed to the appropriate expert via (d) WR to get the final prediction.

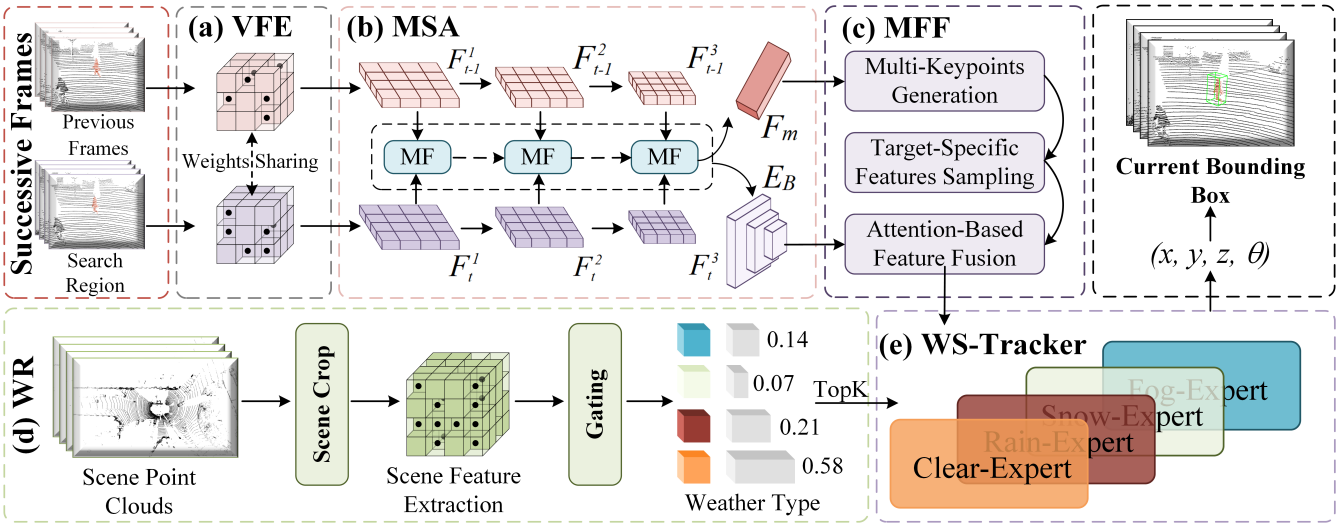

Figure 2: Overview of the UAWTrack architecture: (a) Voxel Feature Extraction (VFE), (b) Motion-centric Spatial-temporal Aggregation (MSA), (c) Motion-guided Feature Fusion (MFF), (d) Weather Router (WR), and (e) Weather-Specific Tracker (WS-Tracker).

Voxel Feature Extraction

LiDAR-scanned point clouds are inherently sparse and particularly prone to perturbation due to backscattering in adverse weather conditions. To address this, we adopt a voxel-based representation (Zhou and Tuzel 2018; Yang et al. 2023) instead of the point-based approach used in previous works, enabling more effective learning of point cloud features. Specifically, we voxelize two consecutive frames of point clouds, P_{t-1} and P_t, under varying weather conditions. Given a point cloud set $P = \{p_1, p_2, p_3, \ldots, p_n\}$, where n denotes the number of points, V represents voxel size, and each point $p_i = \{x_i, y_i, z_i\}$ provides 3D coordinate, the voxel indices $\{i, j, k\}$ corresponding to p_i are calculated as follows:

$$i = [\frac{x_i}{V}], \quad j = [\frac{y_i}{V}], \quad k = [\frac{z_i}{V}]. \tag{1}$$

By applying feature extraction to point clouds within the same voxel grid, the effects of perturbations and sparsity are mitigated to some extent. Using 3D sparse convolution (Graham 2015), we extract sparse 3D features from ordered voxel grids in both the previous frame and the search region. These sparse 3D features are then flattened along the Z-axis and downsampled to obtain multi-scale BEV features of the previous frame and current frame, which are denoted as $\{F^1_{t-1}, F^2_{t-1}, F^3_{t-1}\}$ and $\{F^1_t, F^2_t, F^3_t\}$, respectively.

Motion-centric Spatial-temporal Aggregation

In adverse weather, point clouds suffer from both backscatter and attenuation, leading to increased sparsity. Traditional methods using Siamese appearance matching search for targets within a predefined region, but performance degrades due to the incomplete nature of the point clouds. To address this, as shown in Fig. 2 (b), we introduce a Motion-centric Spatial-temporal Aggregation module. In this module, BEV feature maps at the s^{th} scale from the previous and current frames (F^s_{t-1} and F^s_t) are fed into a sub-module called MF which is depicted in Fig. 3, where spatial alignment, element-wise summation, and convolution are used to extract motion feature M^s for that scale. Specifically, except for the first MF, we also use the motion feature M^s from the previous scale as the input of the next MF, as shown in Fig. 3. This process encodes feature correlations between frames, capturing the motion cues across all matches. Additionally, by continuously downsampling the feature maps, we improve the receptive field, allowing the multi-scale motion features $\{M^s, s \in 1, 2, 3\}$ to encode both global and local motion information. Finally, a global max-pooling is applied on M^3 to derive the initial motion feature F_m.

Motion-guided Feature Fusion

Target-specific features provide critical prior knowledge for object tracking. To this end, we introduce the motion-guided feature fusion module to propagate the target information into the final motion feature with multi-keypoints sampling and attention-based feature fusion.

Multi-Keypoints Generation. We generate N sampling keypoints $S \in \mathbb{R}^{N\times 2}$ (i.e., represented as their coordinates), comprising N_F fixed keypoints (S_F), N_{t-1} learnable keypoints (S_P), and N_t learnable keypoints (S_C). Specifically, S_F and S_P are responsible for sampling target-specific features from the previous frame's BEV features, while S_C samples them from the current frame's BEV features. As illustrated in Fig. 4, we place the fixed keypoints S_F directly at the four corners of the target's 2D bounding box (centered on the BEV plane) in the previous frame. Unlike fixed keypoints, the learnable keypoints adjust according to varying motion features, enabling the neural network to capture the most representative object features under different weather

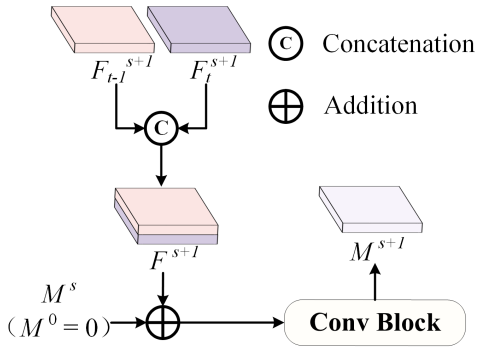

Figure 3: Diagram of Motion Fusion (MF).

conditions. Specifically, given the initial motion feature F_m from the Motion-centric Spatial-temporal Aggregation module, S_P and S_C are generated using a linear network Φ as follows:

$$D = \textbf{sigmoid}(\Phi(F_m)) - 0.5 \in \mathbb{R}^{(N_{t-1}+N_t)\times 2}, \quad (2)$$

$$[D_{t-1}, D_t] = \textbf{split}(D), \quad (3)$$

$$S_P = D_{t-1} \times [l, w], \quad S_C = D_t \times [L, W]. \quad (4)$$

Here, $D_{t-1} \in \mathbb{R}^{N_{t-1}\times 2}$ and $D_t \in \mathbb{R}^{N_t \times 2}$. l and w represent the length and width of the target's bounding box in the point cloud, while L and W denote the point cloud's range along the X and Y axes, respectively.

Target-Specific Features Sampling. The fixed keypoints S_F and learnable keypoints S_P are derived from the target's bounding box, making them aware of the target's location and useful for extracting target-specific features from the previous frame. Meanwhile, the learnable keypoints S_C, informed by the motion feature F_m, focus on capturing the moving target's features in the current frame. Utilizing these keypoints and BEV features from timestamps $t-1$ and t, we efficiently sample target-specific features. Technically, the keypoints are first projected onto the BEV feature maps (F^1_{t-1} and F^1_t), followed by bilinear interpolation for feature sampling. This process yields multi-sampling target-specific features $F_S \in \mathbb{R}^{(N_F+N_{t-1}+N_t)\times C}$, where C denotes the number of feature channels.

Attention-Based Feature Fusion. This module efficiently maps the multi-sampling target-specific features and multi-scale motion features to a final motion feature. This module is illustrated in Fig. 4. Given the flattened token sequence $E_T = \{F_S, F_m\}$ and $E_B = \{M_s, s \in 1, 2, 3\}$, each attention layer performs (1) self-attention within E_T, and (2) cross-attention to E_B, i.e.,

$$F^{sa}_{motion} = \textbf{SelfAttn}(E_T), \quad (5)$$

$$F^{ca}_{motion} = \textbf{CrossAttn}(\{E_B + PE\}, F^{sa}_{motion}), \quad (6)$$

where PE represents positional encodings.

Each self/cross-attention is followed by a residual connection and a layer normalization. The next attention layer takes the updated tokens from the previous layer. Note that

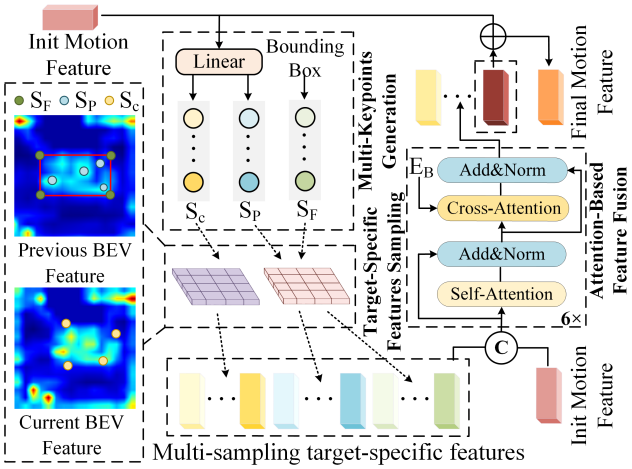

Figure 4: Diagram of Motion-guided Feature Fusion.

the positional encodings are added to the multi-scale motion features whenever they participate in a cross-attention. Finally, we add the init motion feature F_m to the updated motion feature F^{ca}_{motion} output from the last attention layer as the final motion feature F^{final}_{motion}.

Weather-Specific Tracker

There are differences in the distribution of point cloud data between clear weather and adverse weather. A standard dense model which performs well in adverse weather, may forget the knowledge acquired in clear weather. We expect to construct a unified model that can cover all clear and adverse weather types. To address this, inspired by the idea of MoE (Lepikhin et al. 2020), we design WS-Tracker, which comprises four weather-specific experts, i.e., Clear-Expert, Rain-Expert, Snow-Expert, and Fog-Expert, solving problems that new information overwrites previously acquired knowledge. These sub-modules share the same network structure but have unshared weights, allowing them to adaptively process motion features under various weather conditions. Each weather-specific expert directly predicts the target's position and angle $\{x, y, z, \theta\}$ based on the final motion feature F^{final}_{motion}:

$$x, y, z, \theta = \textbf{MLP}(F^{final}_{motion}), \quad (7)$$

where MLP represents a multi-layer perception.

Weather Router. We introduce WR, a lightweight auxiliary module designed to determine the appropriate expert for processing motion features, as shown in Fig. 1. Given the significant geometrical variations in point clouds near the LiDAR sensor under different weather conditions (see Fig. 1 and more examples in Appendix), we first crop the point clouds P_{crop} centered on the LiDAR sensor within the range of $[-7, -7, -1.5; 7, 7, 1.5]$ along the X, Y, and Z axes to use as input.

Traditional point-based methods, such as PointNet (Qi et al. 2017a), struggle with classifying large-scale scene point clouds, while PointNet++ (Qi et al. 2017b) is hindered

Weather Type	Tracker	Source	Car	Pedestrian	Van	Cyclist	Mean	Mean by Category
Clear	P2B	CVPR'20	54.4 / 68.7	36.9 / 59.6	42.2 / 50.8	27.9 / 37.7	45.2 / 62.5	40.4 / 54.2
	M^2Track	CVPR'22	62.7 / 76.0	49.9 / 80.4	61.4 / 77.2	73.1 / 93.4	57.3 / 78.4	61.8 / 81.8
	GLT-T	AAAI'23	66.3 / 79.3	44.8 / 71.5	44.8 / 52.7	58.3 / 87.8	54.9 / 73.8	53.5 / 72.9
	MBPTrack	ICCV'23	70.6 / 81.3	58.3 / 85.0	**67.0/ 78.4**	69.6 / 93.0	64.9 /82.9	66.4 / 84.4
	UAWTrack	Ours	**71.8 / 84.0**	**62.2 / 88.9**	64.5 / 77.7	**74.8 / 94.1**	**67.1 / 85.8**	**68.3 / 86.1**
Rain	P2B	CVPR'20	35.0 / 42.1	36.7 / 58.8	20.6 / 23.2	27.5 / 38.0	34.3 / 47.6	30.0 / 40.5
	M^2Track	CVPR'22	42.9 / 52.1	50.5 / 81.8	26.3 / 32.5	73.4 / 93.5	45.4 / 64.1	48.3 / 65.0
	GLT-T	AAAI'23	38.5 / 45.8	44.9 / 71.0	21.7 / 23.3	55.9 / 84.2	40.2 / 55.6	40.3 / 56.1
	MBPTrack	ICCV'23	55.0 / 66.9	**58.5 / 86.1**	25.6 / 28.5	68.9 / 92.8	54.2 / 72.4	52.0 / 68.6
	UAWTrack	Ours	**58.3 / 70.2**	58.0 / 85.8	**41.0 / 49.5**	**74.5 / 94.1**	**57.0 / 75.6**	**60.3 / 74.9**
Snow	P2B	CVPR'20	34.3 / 41.4	36.3 / 58.5	20.6 / 23.2	28.7 / 39.4	33.8 / 47.1	30.0 / 40.6
	M^2Track	CVPR'22	41.1 / 49.6	51.2 / 82.6	27.3 / 33.7	72.5 / 93.2	44.9 / 63.4	48.0 / 64.8
	GLT-T	AAAI'23	34.8 / 41.1	42.3 / 67.8	21.1 / 22.6	57.4 / 87.4	37.3 / 52.0	38.9 / 54.7
	MBPTrack	ICCV'23	53.7 / 65.5	56.7 / 82.7	23.1 / 24.8	69.1 / 92.9	52.6 / 69.6	50.7 / 66.5
	UAWTrack	Ours	**57.2 / 68.9**	**56.9 / 84.6**	**39.6 / 46.9**	**74.3 / 94.0**	**55.9 / 74.3**	**57.0 / 73.6**
Fog	P2B	CVPR'20	46.7 / 57.7	15.0 / 25.5	40.2 / 47.7	19.9 / 24.8	31.8 / 42.2	30.5 / 38.9
	M^2Track	CVPR'22	60.2 / 72.7	22.1 / 40.1	59.3 / **75.1**	53.3 / 83.4	43.5 / 59.0	48.7 / 67.8
	GLT-T	AAAI'23	58.4 / 69.1	21.5 / 38.0	47.0 / 56.8	32.7 / 50.4	40.9 / 54.1	53.7 / 53.6
	MBPTrack	ICCV'23	66.5 / 77.6	23.6 / 39.4	59.6 / 69.1	59.7 / 84.0	47.2 / 60.5	52.3 / 67.5
	UAWTrack	Ours	**72.8 / 85.6**	**29.1 / 51.0**	**62.6** / 74.7	**66.4 / 90.1**	**52.8 / 69.8**	**57.7 / 75.3**

Table 1: Comparison with state-of-the-art methods on KITTI-A dataset. Bold and underline denote the best and the second-best scores, respectively. Success / Precision are used for evaluation.

by slower inference speeds. To address these challenges, we utilize voxel-based representations (Yang et al. 2023, 2024), reducing the voxel grid size to capture geometric information more effectively. This allows us to extract scene geometrical features F_{scene}^{crop}, which are then input into the *Gating* network. The network predicts the probability (G) of activating each expert, guiding the selection of the appropriate expert to process the motion features. This process can be formulated as follows:

$$G = \textbf{Softmax}(\textbf{TopK}(\textbf{MLP}(F_{scene}^{crop}))), \tag{8}$$

$$\textbf{TopK}(x) = \begin{cases} x, & if\ x\ is\ in\ the\ top\ K\ elements \\ 0, & otherwise \end{cases}. \tag{9}$$

In order to decrease computational expenses, the model activates experts sparsely by using the TopK($\cdot$) function, which sets all but the top K elements of the router weight to zero.

Experiments

We evaluated our proposed UAWTrack on two simulated benchmarks (i.e., KITTI-A and NuScenes-A as detailed below), comparing its performance under adverse weather conditions with that of representative 3D SOT methods.

Datasets. Previous research (Hahner et al. 2022; Dong et al. 2023; Zhao et al. 2024; Rasshofer, Spies, and Spies 2011) indicates that the LiDAR system's received power can be modeled linearly. In non-elastic scattering, range-dependent received power P_R is the convolution of transmitted power P_T and the optical system's impulse response H. In adverse weather conditions, scattering caused by rain, fog, and snowflakes can affect the laser pulse H. These scatterers are represented as distributions of spherical particles with varying sizes, which diminish the initial reflectance of solid objects. Following them, we create two benchmarks for 3D SOT in adverse weather using physically valid adverse weather simulations (Hahner et al. 2022; Dong et al. 2023; Hahner et al. 2021): KITTI-A and NuScenes-A. These datasets extend the clear-weather conditions of KITTI and NuScenes by incorporating three types of adverse weather - rain, snow, and fog - each with three severity levels. Specifically, KITTI-A contains 500 sequences, which are split into training (210 sequences) and testing sets (290 sequences), following the settings in previous works (Yang et al. 2023; Xu et al. 2023b). Compared to KITTI-A, NuScenes-A is a more challenging dataset which contains 7,000 and 1,500 scenes for training and testing, respectively. We show some examples from the KITTI-A dataset in Fig. 1.

Based on these datasets, we evaluate the generalization ability of existing 3D SOT models. Specifically, in Tables 1 and 2, we report the performance of P2B (Qi et al. 2020), M^2Track (Zheng et al. 2022), GLT-T (Nie et al. 2023), and MBPTrack (Xu et al. 2023b) on the KITTI-A and NuScenes-A benchmarks. All models exhibit varying degrees of performance degradation under adverse weather conditions compared to clear weather. This degradation highlights the challenges posed by adverse weather.

Implementation Details. In this work, we use the residual log-likelihood estimation (Li et al. 2021; Yang et al. 2023) for precise target regression, and cross-entropy loss to update the learnable WR weights based on weather type labels. Notably, We first train WR until convergence, then train the remaining components of UAWTrack with WR fixed. Our model is trained with a batch size of 128 and an initial learning rate of 1×10^{-4} using the AdamW optimizer on two

Weather Type	Tracker	Car	Pedestrian	Truck	Trailer	Bus	Mean	Mean by Category
Clear	P2B	34.43 / 36.70	16.18 / 29.52	39.61 / 35.24	48.60 / 35.62	36.54 / 29.02	30.32 / 34.27	35.07 / 33.22
	M^2Track	47.59 / 55.65	24.77 / 50.16	42.78 / 41.93	44.42 / 34.40	42.78 / 41.93	40.21 / 51.27	39.38 / 42.60
	GLT-T	36.93 / 40.40	23.05 / 44.71	44.68 / 42.92	46.43 / 37.92	39.99 / 33.46	34.24 / 41.67	38.22 / 39.88
	MBPTrack	54.60 / 62.85	40.85 / 69.67	54.98 / 54.65	59.52 / 47.88	55.32 / 50.43	50.91 / 63.09	53.05 / 57.10
	UAWTrack	**62.82 / 69.21**	**44.77 / 76.06**	**65.33 / 64.71**	**74.07 / 70.60**	**59.40 / 52.30**	**58.23 / 70.24**	**61.26 / 66.58**
Rain	P2B	34.34 / 36.51	16.32 / 29.84	39.49 / 35.25	49.63 / 37.18	36.74 / 28.84	30.33 / 34.30	35.30 / 33.52
	M^2Track	47.35 / 55.48	24.82 / 50.26	42.63 / 41.84	44.53 / 34.76	42.63 / 41.84	40.09 / 51.22	39.38 / 42.78
	GLT-T	36.40 / 39.80	22.95 / 44.34	44.55 / 42.81	46.73 / 38.82	39.58 / 32.95	33.91 / 41.23	38.04 / 39.74
	MBPTrack	53.91 / 60.89	40.71 / 69.42	54.84 / 54.62	59.81 / 48.04	55.16 / 50.08	55.16 / 50.08	52.89 / 56.61
	UAWTrack	**62.80 / 69.19**	**44.72 / 76.06**	**65.32 / 64.62**	**73.90 / 70.41**	**59.14 / 52.10**	**58.19 / 70.21**	**61.18 / 66.48**
Snow	P2B	34.32 / 36.49	16.39 / 30.11	39.25 / 35.01	49.34 / 36.64	35.75 / 28.01	30.28 / 34.30	35.01 / 33.25
	M^2Track	47.33 / 55.45	24.76 / 50.22	42.58 / 41.72	44.59 / 34.84	42.58 / 41.72	40.05 / 51.17	39.30 / 42.62
	GLT-T	36.37 / 39.74	22.88 / 44.17	44.27 / 42.44	46.84 / 38.55	39.26 / 32.74	33.84 / 41.10	37.92 / 39.53
	MBPTrack	54.45 / 62.73	40.84 / 69.62	55.13 / 54.92	59.51 / 47.81	55.17 / 50.23	50.84 / 63.04	53.02 / 57.06
	UAWTrack	**62.74 / 69.06**	**44.68 / 76.09**	**65.26 / 64.59**	**74.07 / 70.62**	**59.00 / 51.98**	**58.14 / 70.15**	**61.15 / 66.47**
Fog	P2B	31.51 / 32.48	15.88 / 29.41	39.24 / 35.99	45.70 / 35.79	35.03 / 27.53	28.47 / 31.99	33.47 / 32.24
	M^2Track	42.56 / 50.28	22.77 / 45.98	36.97 / 36.21	42.10 / 32.61	36.97 / 36.21	36.08 / 46.36	35.76 / 38.56
	GLT-T	34.29 / 37.63	21.34 / 41.15	40.68 / 37.82	41.81 / 32.25	35.17 / 28.92	31.60 / 38.28	34.66 / 35.55
	MBPTrack	54.21 / 61.51	37.21 / 63.37	54.36 / 53.79	60.33 / 50.22	52.67 / 47.59	49.55 / 60.47	51.76 / 55.30
	UAWTrack	**62.03 / 68.20**	**40.78 / 70.04**	**63.90 / 63.71**	**73.87 / 70.49**	**56.03 / 49.65**	**55.94 / 67.80**	**59.32 / 64.42**

Table 2: Comparison with state-of-the-art methods on NuScenes-A dataset. Bold and underline denote the best and the second-best scores, respectively. Success / Precision are used for evaluation.

S_F	S_P	S_C	Clear	Rain	Snow	Fog
			70.2 / 82.5	57.0 / 68.8	55.7 / 67.5	71.9 / 84.4
✓			70.8 / 82.8	57.4 / 69.1	56.0 / 67.9	72.1 / 84.6
✓	✓		71.1 / 83.1	57.7 / 69.2	56.4 / 68.1	72.4 / 85.1
✓		✓	71.4 / 83.5	58.0 / 69.7	57.0 / 68.4	72.5 / 85.3
✓	✓	✓	71.8 / 84.0	58.3 / 70.2	57.2 / 68.9	72.8 / 85.6

Table 3: Influence of keypoint sampling strategy.

MFF	Clear	Rain	Snow	Fog
	64.3 / 77.6	50.8 / 62.1	49.0 / 60.1	65.8 / 77.6
✓	71.8 / 84.0	58.3 / 70.2	57.2 / 68.9	72.8 / 85.6

Table 4: Ablation of MFF module on KITTI-A benchmark.

NVIDIA GTX 4090 GPUs. The network is implemented in PyTorch with MMEngine (Contributors 2022).

Metrics. Following common practice, we utilize the metrics of Success and Precision from the One Pass Evaluation (OPE) (Wu, Lim, and Yang 2013) framework for performance assessment. Precision assesses the distance between the centers of the two corresponding bounding boxes, while success rate calculates the Intersection over Union (IoU) between the predicted bounding box and the ground truth bounding box.

Main Results and Analysis

Results on KITTI-A. We present a detailed comparison on the KITTI-A dataset between our proposed method, UAWTrack, and previous state-of-the-art methods, including Siamese trackers like P2B, GLT-T, MBPTrack, and the motion-based tracker M^2Track. As shown in Table 1, UAWTrack consistently outperforms others across various categories and weather conditions, achieving the highest average Success and Precision rates. Notably, UAWTrack significantly surpasses MBPTrack under all weather conditions, highlighting the advantages of voxel-based representation and motion modeling in adverse weather. Compared to M^2Track, UAWTrack shows remarkable improvement, with mean Success and Precision increases of 9.8% and 7.4% in clear weather, 11.6% and 11.5% in rain, 11.0% and 10.9% in snow, and 9.3% and 10.8% in fog. The classification results of WR for various weather conditions in KITTI-A are presented in the Appendix. Additionally, as depicted in Fig. 5, UAWTrack maintains strong performance across all severity levels compared to other 3D SOT trackers. We also visualize the tracking results for qualitative comparisons in Fig. 6.

Results on NuScenes-A. NuScenes-A presents a more challenging 3D SOT task than KITTI-A due to the inclusion of real-world adverse weather conditions. We compared UAWTrack against state-of-the-art methods (P2B, M^2Track, GLT-T, and MBPTrack) on the NuScenes-A dataset. As shown in Table 2, UAWTrack consistently outperforms all prior methods across every category and weather condition. Notably, UAWTrack surpasses MBPTrack by 8.22% and 7.15% in clear weather, 3.03% and 20.13% in rain, 7.30% and 7.11% in snow, and 6.39% and 7.33% in fog, for mean success rate and mean precision rate, respectively. These results demonstrate UAWTrack's exceptional tracking performance in adverse weather, even with extremely sparse point clouds, and its strong potential for handling diverse weather conditions within a unified model. Similarly, the classification results of WR for various weather conditions in NuScenes-A will also be presented in the Appendix.

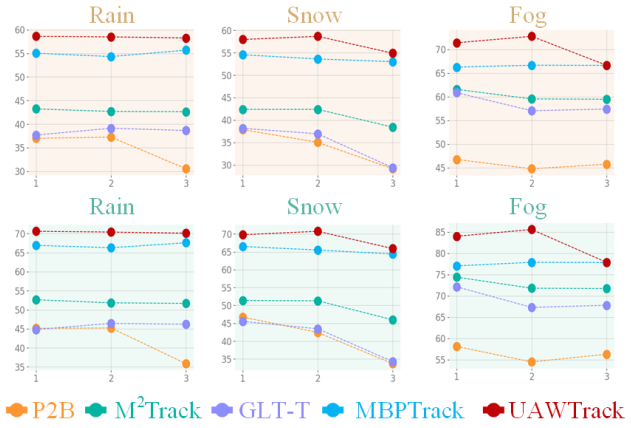

Figure 5: The success (upper) and precision (lower) metrics of state-of-the-art 3D SOT trackers under each of the three severity levels on KITTI-A.

Weather Type	WS-Tracker	Car	Pedestrian	Van	Cyclist
Clear		70.1 / 83.2	58.1 / 84.4	60.3 / 72.3	71.2 / 93.0
	✓	71.8 / 84.0	59.9 / 86.1	66.6 / 79.0	73.8 / 94.1
Rain		55.8 / 66.5	57.8 / 83.9	33.0 / 36.1	72.2 / 92.7
	✓	58.3 / 70.2	59.3 / 85.8	33.8 / 38.1	73.5 / 94.0
Snow		54.8 / 65.6	58.1 / 84.7	31.1 / 32.8	72.2 / 92.7
	✓	57.2 / 68.9	59.5 / 86.2	33.4 / 37.7	73.2 / 93.9
Fog		71.2 / 83.7	26.0/ 46.4	60.0 / 72.3	64.1 / 90.7
	✓	72.8 / 85.6	27.5 / 48.1	65.6 / 78.3	65.8 / 91.2

Table 5: Effectiveness of WS-Tracker. Without WS-Tracker, a single expert is used to handle motion features across all weather conditions.

Ablation Study

Influence of keypoint sampling strategy. To assess the impact of different keypoint sampling strategies on tracking performance, we conducted an ablation study on UAWTrack using the KITTI-A benchmark. As shown in Table 3, relying solely on "fixed keypoints" (S_F) leads to a notable performance drop across all weather conditions for the car category compared to using a combination of fixed and learnable keypoints (S_F + S_P + S_C). Additionally, employing either S_F + S_P or S_F + S_C also results in performance declines under all weather conditions, highlighting the complementary role of both types of learnable keypoints in sampling effective features from previous and current frames.

Effectiveness of the MFF module. To investigate the impact of the Motion-guided Feature Fusion module, we directly take the motion feature F_m as the query and the multi-scale features $\{M^s, s \in 1, 2, 3\}$ as the key and value, and then execute the cross-attention mechanism. After that, we send F_m to the weather-specific tracker module. Table 4 shows that the MFF module can provide the target-specific features required by 3D single object tracking, thus achieving a significant performance boost.

Effectiveness of WS-Tracker. To evaluate the effectiveness of WS-Tracker, we trained a single expert to handle motion features across all weather conditions. Since our UAWTrack utilizes the output of WR to provide weather information, disabling WS-Tracker necessarily means disabling WR as well. Without WS-Tracker and WR, this architecture lacks the capability to handle weather-specific information, leading to a performance gap compared to the architecture that incorporates these components. As demonstrated in Table 5, using the car category as an example, WS-Tracker and WR enhance the tracking performance by 1.7% Success and 0.9% Precision in clear weather, 1.5% and 2.1% in rain, 1.3% and 1.4% in snow, and 1.3% and 1.4% in fog. This notable improvement across all categories and weather conditions underscores the importance of using individual experts to process motion features under corresponding weather conditions.

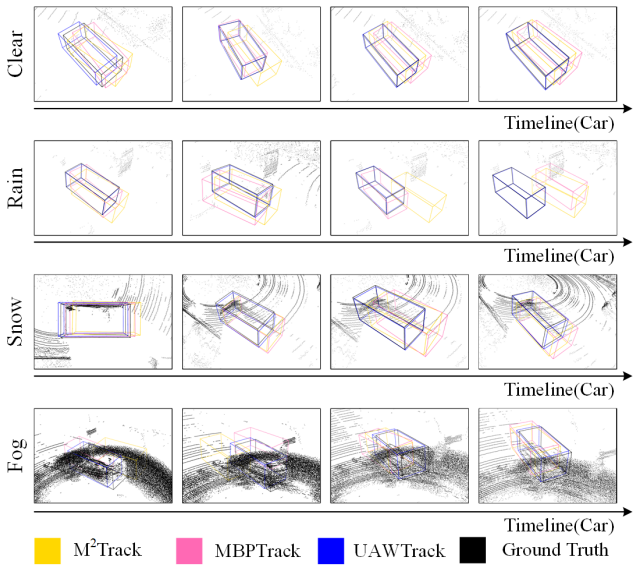

Figure 6: Visualization of tracking results by our UAWTrack and state-of-the-art methods.

Conclusion

In this paper, we present UAWTrack, a universal 3D Single Object Tracking (SOT) model that effectively addresses the challenges of tracking in adverse weather conditions. Our model introduces several critical components: a Voxel Feature Extraction module to mitigate perturbation and sparsity in point clouds, a Motion-centric Spatial-temporal Aggregation and Motion-guided Feature Fusion module that leverages motion clues to enhance feature representations, and a Mixture-of-Experts based Weather-Specific Tracker that adapts to various weather conditions. In addition, we fill the gap of lacking benchmarks for 3D SOT in adverse weather by simulating realistic adverse weather and creating the KITTI-A and NuScenes-A datasets. Extensive experiments show UAWTrack beats other top trackers in all weather conditions, improving the reliability and safety of 3D SOT in autonomous driving. We hope it inspires more research on 3D SOT in adverse weather.

Acknowledgments

This work was supported by the National Natural Science Foundation of China under Grant 62376080, the Zhejiang Provincial Major Research and Development Project of China under Grant 2024C01143, and the Zhejiang Provincial Natural Science Foundation Key Fund of China under Grant LZ23F030003.

References

Bijelic, M.; Gruber, T.; Mannan, F.; Kraus, F.; Ritter, W.; Dietmayer, K.; and Heide, F. 2020. Seeing Through Fog Without Seeing Fog: Deep Multimodal Sensor Fusion in Unseen Adverse Weather. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

Caesar, H.; Bankiti, V.; Lang, A. H.; Vora, S.; Liong, V. E.; Xu, Q.; Krishnan, A.; Pan, Y.; Baldan, G.; and Beijbom, O. 2020. nuscenes: A multimodal dataset for autonomous driving. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 11621–11631.

Contributors, M. 2022. MMEngine: OpenMMLab Foundational Library for Training Deep Learning Models.

Dong, Y.; Kang, C.; Zhang, J.; Zhu, Z.; Wang, Y.; Yang, X.; Su, H.; Wei, X.; and Zhu, J. 2023. Benchmarking robustness of 3d object detection to common corruptions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 1022–1032.

Fang, Z.; Zhou, S.; Cui, Y.; and Scherer, S. 2020. 3d-siamrpn: An end-to-end learning method for real-time 3d single object tracking using raw point cloud. *IEEE Sensors Journal*, 21(4): 4995–5011.

Geiger, A.; Lenz, P.; and Urtasun, R. 2012. Are we ready for autonomous driving? the kitti vision benchmark suite. In *2012 IEEE conference on computer vision and pattern recognition*, 3354–3361. IEEE.

Giancola, S.; Zarzar, J.; and Ghanem, B. 2019. Leveraging shape completion for 3d siamese tracking. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 1359–1368.

Graham, B. 2015. Sparse 3D convolutional neural networks. In *Procedings of the British Machine Vision Conference 2015*.

Hahner, M.; Sakaridis, C.; Bijelic, M.; Heide, F.; Yu, F.; Dai, D.; and Van Gool, L. 2022. Lidar snowfall simulation for robust 3d object detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 16364–16374.

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 *Proceedings of the IEEE/CVF international conference on computer vision*, 15283–15292.

Hao, X.; Wei, M.; Yang, Y.; Zhao, H.; Zhang, H.; Zhou, Y.; Wang, Q.; Li, W.; Kong, L.; and Zhang, J. 2024. Is Your HD Map Constructor Reliable under Sensor Corruptions? *arXiv preprint arXiv:2406.12214*.

Huang, X.; Wu, H.; Li, X.; Fan, X.; Wen, C.; and Wang, C. 2024. Sunshine to rainstorm: Cross-weather knowledge distillation for robust 3d object detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, 2409–2416.

Hui, L.; Wang, L.; Cheng, M.; Xie, J.; and Yang, J. 2021. 3d siamese voxel-to-bev tracker for sparse point clouds. *Advances in Neural Information Processing Systems*, 34: 28714–28727.

Kong, L.; Xie, S.; Hu, H.; Niu, Y.; Ooi, W. T.; Cottereau, B. R.; Ng, L. X.; Ma, Y.; Zhang, W.; Pan, L.; et al. 2024. The robodrive challenge: Drive anytime anywhere in any condition. *arXiv preprint arXiv:2405.08816*.

Lepikhin, D.; Lee, H.; Xu, Y.; Chen, D.; Firat, O.; Huang, Y.; Krikun, M.; Shazeer, N.; and Chen, Z. 2020. Gshard: Scaling giant models with conditional computation and automatic sharding. *arXiv preprint arXiv:2006.16668*.

Li, J.; Bian, S.; Zeng, A.; Wang, C.; Pang, B.; Liu, W.; and Lu, C. 2021. Human Pose Regression with Residual Log-likelihood Estimation. In *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*.

Ma, T.; Wang, M.; Xiao, J.; Wu, H.; and Liu, Y. 2023. Synchronize feature extracting and matching: A single branch framework for 3d object tracking. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 9953–9963.

Nie, J.; He, Z.; Yang, Y.; Gao, M.; and Zhang, J. 2023. Glt-t: Global-local transformer voting for 3d single object tracking in point clouds. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, 1957–1965.

Qi, C. R.; Su, H.; Mo, K.; and Guibas, L. J. 2017a. Pointnet: Deep learning on point sets for 3d classification and segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 652–660.

Qi, C. R.; Yi, L.; Su, H.; and Guibas, L. J. 2017b. Pointnet++: Deep hierarchical feature learning on point sets in a metric space. *Advances in neural information processing systems*, 30.

Qi, H.; Feng, C.; Cao, Z.; Zhao, F.; and Xiao, Y. 2020. P2b: Point-to-box network for 3d object tracking in point clouds. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 6329–6338.

Rasshofer, R. H.; Spies, M.; and Spies, H. 2011. Influences of weather phenomena on automotive laser radar systems. *Advances in Radio Science*, 49–60.

Shan, J.; Zhou, S.; Fang, Z.; and Cui, Y. 2021. PTT: Point-track-transformer module for 3D single object tracking in point clouds. In *2021 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 1310–1316. IEEE.

Shazeer, N.; Mirhoseini, A.; Maziarz, K.; Davis, A.; Le, Q.; Hinton, G.; and Dean, J. 2017. Outrageously Large Neural Networks: The Sparsely-Gated Mixture-of-Experts Layer. *arXiv: Learning,arXiv: Learning*.

Wu, Y.; Lim, J.; and Yang, M.-H. 2013. Online object tracking: A benchmark. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2411–2418.

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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 9382–9392.

Xu, T.-X.; Guo, Y.-C.; Lai, Y.-K.; and Zhang, S.-H. 2023a. CXTrack: Improving 3D point cloud tracking with contextual information. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 1084–1093.

Xu, T.-X.; Guo, Y.-C.; Lai, Y.-K.; and Zhang, S.-H. 2023b. Mbptrack: Improving 3d point cloud tracking with memory networks and box priors. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 9911–9920.

Xu, Y.; Zhang, J.; Zhang, Q.; and Tao, D. 2023c. Vitpose++: Vision transformer for generic body pose estimation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.

Yang, Y.; Deng, Y.; Nie, J.; and Zhang, J. 2023. BEVTrack: A Simple Baseline for Point Cloud Tracking in Bird's-Eye-View. *arXiv preprint arXiv:2309.02185*.

Yang, Y.; Deng, Y.; Zhang, J.; Gu, H.; and Dong, Z. 2024. SiamMo: Siamese Motion-Centric 3D Object Tracking. *arXiv preprint arXiv:2408.01688*.

Zhang, R.; Luo, Y.; Liu, J.; Yang, H.; Dong, Z.; Gudovskiy, D.; Okuno, T.; Nakata, Y.; Keutzer, K.; Du, Y.; et al. 2024. Efficient Deweahter Mixture-of-Experts with Uncertainty-Aware Feature-Wise Linear Modulation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, 16812–16820.

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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 14781–14791.

Zheng, C.; Yan, X.; Gao, J.; Zhao, W.; Zhang, W.; Li, Z.; and Cui, S. 2021. Box-aware feature enhancement for single object tracking on point clouds. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 13199–13208.

Zheng, C.; Yan, X.; Zhang, H.; Wang, B.; Cheng, S.; Cui, S.; and Li, Z. 2022. Beyond 3d siamese tracking: A motion-centric paradigm for 3d single object tracking in point clouds. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 8111–8120.

Zhou, C.; Luo, Z.; Luo, Y.; Liu, T.; Pan, L.; Cai, Z.; Zhao, H.; and Lu, S. 2022. Pttr: Relational 3d point cloud object tracking with transformer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 8531–8540.

Zhou, Y.; and Tuzel, O. 2018. VoxelNet: End-to-End Learning for Point Cloud Based 3D Object Detection. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*.