2026 IEEE International Conference on Robotics and Automation (ICRA 2026)
June 1-5, 2026. Vienna, Austria

4D Radar Diffusion with Adaptive Visual-Aided Condition for Point Cloud Enhancement

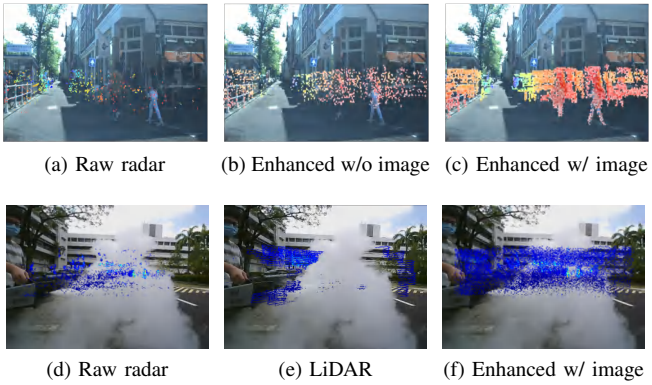
Renxiang Xiao[1], Yuanfan Zhang[2], Wei Liu[1], Guangzhong Dong, Yunjiang Lou[1] and Liang Hu[1*]

***Abstract*— Despite its resilience in adverse weather, millimeter-wave (mmWave) radar yields sparse and noisy point clouds that limit its perception and localization performance. Diffusion models have recently gained attention for enhancing millimeter-wave radar in perception tasks due to their strong denoising and generative capabilities. Yet, the enhanced radar point cloud is still far from expected due to a lack of texture information and errors caused by inherent sensor–model mismatch between LiDAR and radar. In this paper, we propose an adaptive vision-aided radar data enhancement method based on a conditional diffusion model for denoising and densifying radar point clouds. The pipeline decomposes mmWave radar into depth and BEV views, fuses the depth view with synchronized images, and uses the fused features together with BEV tokens to condition the diffusion model. LiDAR is used only for training supervision, but not for inference. Extensive experiments demonstrate that our proposed method produces dense and geometrically consistent radar point clouds, validating the effectiveness of the introduced vision-aid for radar enhancement. Notably, our method even works well in scenarios under visual occlusions. The accurate odometry and high-fidelity map reconstruction using enhanced radar point cloud highlights the great potential of our method for other downstream tasks in robotics and autonomous driving.**

I. INTRODUCTION

In recent years, millimeter wave radar combined with vision has been gaining increasing attention as a promising perception solution for autonomous driving in the automotive industry. As an alternative to optical ranging sensors such as LiDAR, millimeter-wave radar can perform perception [1]–[3] and localization [4]–[8] tasks reliably even under adverse weather conditions (rain, fog, snow). However, the inherent noise and sparsity problems of the radar, especially the newer type of 4D radar, impair its performance in downstream tasks. Therefore, enhancing the raw radar data to obtain denser and higher-quality point clouds is crucial to expand the potential of millimeter-wave radar.

The efforts to enhance 4D millimeter-wave radar span both the raw signal in the frequency domain and the point cloud in the data domain. Although pure signal processing enhancement and convolutional networks [9]–[11] can remove ghost points caused by noise, they intrinsically lack the generative capacity to recover the continuous geometry. Data-domain enhancement implemented through aggregation or deterministic regression [12]–[18] fails to capture alternative plausible structures and provides no principled representation of uncertainty. By contrast, diffusion probabilistic models learn a generative prior over explicit point cloud distributions and perform stochastic stepwise denoising, allowing them to naturally encode cross-modal uncertainty and reconstruct complete geometry from a single noisy observation.



(a) Raw radar (b) Enhanced w/o image (c) Enhanced w/ image

(d) Raw radar (e) LiDAR (f) Enhanced w/ image

Fig. 1. **Radar enhancement of the proposed method in different conditions (All projected onto the corresponding images for better visualization). Top row** (clear environments): (a) raw radar; (b) radar-enhanced results by our method with radar-only input; (c) radar-enhanced results by our method. **Bottom row** (smoke environments): (d) raw radar; (e) camera/LiDAR degraded by smoke; (f) Our method uses reliable image texture and reverts to radar-only in occluded regions, preserving a dense and geometrically accurate representation in smoke.

However, current diffusion-based approaches for 4D radar enhancement [19], [20] rely on LiDAR for supervision during training but only use sparse radar input at inference. As shown in Fig. 1d and 1e, the differences in sensor physics models of LiDAR and radar produce inconsistent observations, leading to incomplete or discontinuous geometry inference, particularly under occlusions or complex object appearances as shown in Fig. 1b. To overcome it, we introduce synchronized visual information with an adaptive self-attention at inference. Visual observations provide dense texture cues that complement radar measurements, effectively bridging the inconsistency and guiding the denoising process toward high-density, geometrically consistent radar point clouds even in adverse conditions, as shown in Fig. 1f.

To this end, we propose a unified radar–vision framework that integrates synchronized 4D radar sweeps and RGB images with a conditional latent diffusion model (LDM) for radar data enhancement. Each radar frame is orthogonally

* Corresponding author. Email: l.hu@hit.edu.cn.

This work was supported in part by the National Natural Science Foundation of China under Grant 62573157, the Science Center Program of National Natural Science Foundation of China under Grant 62188101, Shenzhen Science and Technology Program under Grant SYSPG20241211173609005, and under Grant JCYJ20241202123714019.

[1]R. Xiao, W. Liu, G. Dong, Y. Lou and L. Hu are with School of Mechanical Engineering and Automation, Harbin Institute of Technology, Shenzhen, China.

[2]Y. Zhang is with School of Computer Science and Technology, Harbin Institute of Technology, Shenzhen, China.

projected and fused with its corresponding image into the shared bird's-eye-view (BEV) token grid, which serves as the condition for the diffusion model. The LDM produces LiDAR-aligned latent features that are subsequently decoded by a lightweight network into a dense point cloud. LiDAR is used only during training, via a frozen encoder that defines the latent space and provides supervision; at inference, the model relies solely on radar and image inputs.

The main contributions of our paper are summarized as follows:

- We propose a conditional latent-diffusion model that fuses synchronized radar and image features to generate denoised, dense point clouds. The introduced visual features guide the denoising process toward high-density, geometrically consistent radar point clouds.
- We propose a forward depth–BEV decomposition for radar point clouds, which facilitates fusing depth with image tokens and hence stable and efficient sampling by the conditional diffusion model.
- Extensive comparative experiments verify the effectiveness of our method in radar point-cloud denoising and generation. We further demonstrate its potential through the downstream task of high-fidelity reconstruction using the enhanced radar point cloud, which was infeasible with the raw radar data.

II. RELATED WORK

A. Radar Data Enhancement

Existing radar data enhancement methods are mainly divided into signal domain enhancement [9], [10] and data domain enhancement [12]–[18]. Signal domain enhancement methods directly apply deep learning or signal processing methods to the original or intermediate frequency data of the radar to improve the angular resolution and virtual aperture before forming the point cloud. [9] improves mmWave radar point-cloud quality from a point-cloud processing perspective, but it is still limited by observation sparsity and has difficulty extrapolating surface structures that completely lack support. [10] uses a residual encoding-decoding network to recover details suppressed by the sampling rate limit, thereby generating continuous angle-range texture under extremely sparse input. [11] embeds the micro-Doppler spectrum into U-Net and uses the dual-base virtual array of antennas to infer dense corner points.

These methods are still limited by the inherent observation sparsity. Therefore, data domain enhancement methods are focused on predicting the point cloud distribution to guide point cloud densification. [12] sends the voxelized sparse 4D point cloud to the 3D CNN network to regress the voxel occupancy probability. [13] uses a conditional GAN network to generate dense points supervised by synchronized LiDAR. [14] uses an autoregressive Transformer to generate a complete point cloud by radar point tokens. DenserRadar [15] predicts 3D offsets through an MLP and regresses sparse radar points to corresponding dense LiDAR points. [16] further introduces temporal memory and recursively fuses features of consecutive frames to reduce drift. [17] and [18] use visual-inertial odometry and monocular depth prior as weak supervision, respectively, and use cross-modal consistency to repair occluded areas. Nonetheless, all the above regressors only produce a single best guess completion and are unable to infer surfaces that lie entirely outside the sparse measurement support. In contrast, Diffusion models equipped with stochastic generative priors synthesize such missing structure faithfully, yielding denser and more complete point clouds.

B. Radar Diffusion

Recent research on diffusion models for radar data enhancement can be roughly divided into two categories: methods that directly transfer image-domain DDPM priors to radar representations [21]–[23], and methods that construct explicit conditions to guide the diffusion process [20], [24]–[26]. Early works primarily adopt 2D representations, Radar-Diffusion [20] and RadarINV [24] formulate radar point cloud enhancement as cross-modal denoising and super-resolution in BEV space without pair LiDAR-radar supervision, while Diffradar [21] and Tuel et al. [22] reconstruct high-quality 2D scene distributions or remote sensing imagery through diffusion-based decoding. These methods, however, largely neglect the vertical structure of radar point clouds. More recent studies extend diffusion models to 3D recovery. Luan et al. [19] achieve radar point cloud super-resolution toward LiDAR-like density, Wu et al. [23] demonstrate the advantage of range-image priors over BEV or range-azimuth heatmaps, R2LDM [25] encodes LiDAR and 4D radar jointly into latent voxels for diffusion and reconstruction, and Zhang et al. [26] propose a diffusion-based framework that suppresses sidelobes and generates LiDAR-like semantic point clouds through sparse 3D semantic networks.

Despite these advances, existing diffusion-based methods commonly rely on LiDAR-anchored training assumptions while being deployed under radar-only conditions at inference, leading to a sensor-physics mismatch that degrades generalization. To overcome this limitation, we introduce a vision-aided fusion module that conditions the denoising process with synchronized camera–radar tokens to mitigate this mismatch and enhance generalization.

III. METHODOLOGY

A. Overview

The proposed framework enhances a single mmWave radar sweep under visual guidance, delivering a densified, low-noise point cloud, as shown in Fig. 2. The pipeline projects the 4D radar data into a BEV image, and a forward view depth image, embedding them as radar tokens T_B and T_D. The depth image tokens are fused with synchronised image tokens through the Visual-Aided Fusion (VAF) module to produce fused tokens T. T_B and T are fed to the Cross-View Fusion (CVF) module to produce the conditional stack C_t. A conditional LDM then operates in the latent space, which

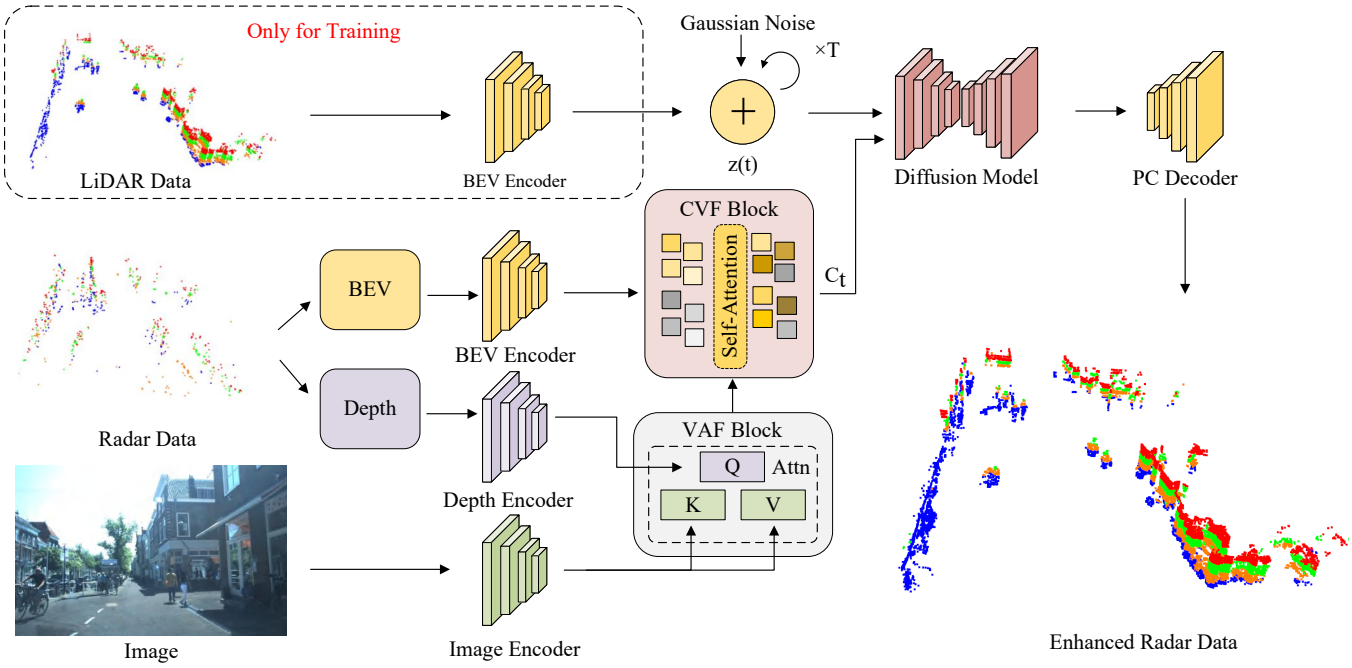


Fig. 2. **Overall system architecture.** The raw radar sweep is rasterised into two orthogonal views. The BEV branch is patch-embedded into radar tokens, while the depth branch forms query tokens that cross-attend to image features extracted by an image encoder. The fusion tokens are connected with BEV tokens and refined by self-attention to form the conditional token stack C_t. The LiDAR tokens are merely input as supervision during training. At inference, the diffusion model denoises iteratively conditioned on C_t without requiring LiDAR, and a task-specific point-cloud decoder maps the recovered latent to a dense radar point cloud.

is decoded by a point-cloud decoder into a dense LiDAR-like point cloud. Training relies solely on accumulated multi-frame LiDAR as supervision, while the inference process needs only radar and image inputs.

B. Tokenization and Condition Construction

Data Processing: Raw LiDAR scans are first de-grounded using Patchwork++ [27] which combines adaptive ground likelihood estimation with region-wise vertical plane fitting. To mitigate sparsity, five consecutive radar sweeps are motion-compensated and stacked using ground-truth poses. Cross-modal calibration is then applied, the 4D radar points are reprojected into the camera frame to obtain both an image-aligned forward-view depth map and a BEV raster. Let $p_R = [x_r, y_r, z_r]^\top \in \mathbb{R}^3$ denote a single radar point expressed in the radar frame, $p_C = [x_c, y_c, z_c]^\top \in \mathbb{R}^3$ denote the same point expressed in the camera frame. It is transformed to the camera frame by the radar-to-camera extrinsics:

$$p_C = R_R^C\, p_R + t_R^C. \tag{1}$$

where $p_R \in \mathbb{R}^3$ and $p_C \in \mathbb{R}^3$ denote a single 3D point in the radar and camera frames, respectively; $R_R^C \in SO(3)$ and $t_R^C \in \mathbb{R}^3$ denote the rotation and translation from the radar frame to the camera frame. The pixel coordinate (u, v) is then obtained by homogeneous projection with the camera intrinsics, and the corresponding depth is z_c.

Independently, both LiDAR and stacked radar point clouds are voxelized on the ground plane to form BEV raster images $B_l \in \mathbb{R}^{H_B \times W_B}$ and $B_r \in \mathbb{R}^{H_B \times W_B}$, respectively, where the spatial resolution H_B, W_B determined by the in-plane voxel size.

Tokenized: Given the LiDAR and Radar BEV images, a shared Transformer encoder ψ reduces dimensionality and extracts features:

$$F_B^l = \psi(B_l) \in \mathbb{R}^{H \times W \times C}, F_B^r = \psi(B_r) \in \mathbb{R}^{H \times W \times C}. \tag{2}$$

For the synchronised depth image $D \in \mathbb{R}^{H_D \times W_D}$ and the RGB image $I \in \mathbb{R}^{H_I \times W_I \times 3}$, their latent representations are also extracted by the depth encoder ϕ and the image encoder π:

$$F_D = \phi(D) \in \mathbb{R}^{H \times W \times C}, F_I = \pi(I) \in \mathbb{R}^{H \times W \times C}. \tag{3}$$

Visual-Aided Fusion: VAF performs asymmetric, geometrically anchored image and depth fusion. Although the field of view (FOV) of the image and depth map is the same, radar depth maps have holes in the pixel grid due to data sparsity, and the image is prone to failure under lighting changes, reflections, or motion blur. To retain the depth grid as a geometric anchor while selectively injecting visual evidence, we use depth as a query to preserve its geometric coordinates and use the image as a key-value memory. Specifically, we first divide each feature into N patches, then we flatten each patch into a vector of length L. The concatenated vectors produce the depth and image token matrices $T_D, T_I \in \mathbb{R}^{N \times L}$. Cross-attention is then calculated between them:

$$Q = T_D W_Q, \qquad K = T_I W_K, \qquad V = T_I W_V, \tag{4}$$

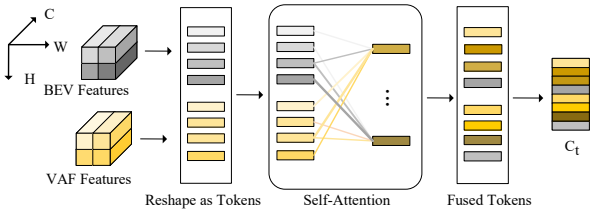


Fig. 3. **The process of CVF.** BEV features and fused depth features from VAF are reshaped and concatenated into a joint token set, over which global self-attention performs symmetric, bidirectional fusion. The fused tokens C_t combine scene-level structure with fine boundaries and are used to condition the diffusion model.

$$T = \text{softmax}\left(\frac{QK^{\top}}{\sqrt{d_k}} + B\right)V, \tag{5}$$

where W_Q, W_K, W_V are the weight matrices, B is the bias, d_k represents the dimensionality of the feature vector K, and the output T is dimensionally consistent with the depth token matrix T_D. This asymmetric attention mechanism design can be understood as kernel regression of the conditional expectation of depth geometry given known image semantics. Specifically, the depth features provide a fundamental geometric prior, which is further integrated with image semantics to estimate the spatial structure of the unknown part. The cross-attention design constrains weights to within the valid depth occupancy range, effectively supporting radar with sparse or uniform depth distribution and avoiding the introduction of invalid information when the image is perturbed.

Cross-View Fusion: CVF performs global feature fusion between the radar BEV features and fused features output by VAF. Since multi-view fusion requires symmetric bidirectional coordination to enforce scene-level consistency, we concatenate the output T from VAF with the BEV token T_B obtained from F_B^r to form a new token set $T' = [T; T_B] \in \mathbb{R}^{2N\times L}$. Then self-attention is applied to the joint set:

$$Q = T'W_Q,\ K = T'W_K,\ V = T'W_V, \tag{6}$$

$$C_t = \text{softmax}\left(\frac{QK^{\top}}{\sqrt{d_K}} + B\right)V. \tag{7}$$

CVF implements a cross-view information fusion scheme through the attention mechanism, achieving non-local regularization and cross-scale alignment in one step. The BEV view propagates global structure to the forward view, while the forward view feeds back the fine boundary and semantic information fused in the previous step to the BEV information. As shown in Fig. 3, the mutual information exchange between the two views generates C_t for the diffusion network, which serves as a condition for geometric consistency and information completeness.

C. Denoising and Generation

We adopt latent-space diffusion with $z_t \in \mathbb{R}^{H\times W\times C}$ as the diffusion variable. In the training stage, z_t is obtained by adding noise to extracted features from LiDAR BEV images:

$$z_t = \sqrt{\alpha_t}\, z_{t-1} + \sqrt{1-\alpha_t}\,\epsilon, \quad \epsilon \sim \mathcal{N}(0, I) \tag{8}$$

$$z_0 = F_B^l, \tag{9}$$

where α_t is the stepwise adjusted variance parameter and I represents the identity matrix. While in the inference stage, we start from $z_t \sim \mathcal{N}(0, I)$. Given the fixed condition $C_t \in \mathbb{R}^{2N\times L}$, each reverse step is a Gaussian posterior [28]:

$$p_\theta(z_{t-1} \mid z_t, C_t) = \mathcal{N}\big(z_{t-1};\, \mu_\theta(z_t, t, C_t),\, \tilde{\beta}_t I\big), \tag{10}$$

$$\tilde{\beta}_t = \frac{1-\bar{\alpha}_{t-1}}{1-\bar{\alpha}_t}\beta_t,\ \beta_t = 1-\alpha_t,\ \bar{\alpha}_t = \textstyle\prod_{s=1}^{t}\alpha_s. \tag{11}$$

Using the ε-parameterisation, the mean is

$$\mu_\theta(z_t, t, C_t) = \frac{1}{\sqrt{\alpha_t}}\Big(z_t - \frac{\beta_t}{\sqrt{1-\bar{\alpha}_t}}\,\varepsilon_\theta(z_t, t, C_t)\Big), \tag{12}$$

where ε_θ is a single conditional noise predictor. We integrate the probability–flow ODE with DPM-Solver++ [29] under log-SNR stepping to deterministically obtain $\hat{z}_0 \in \mathbb{R}^{H\times W\times C}$.

Decoder Architecture. The latent geometry decoder $\mathcal{D}_\phi$ is implemented using a Transformer structure symmetric to the BEV encoder. It receives the latent space features $\hat{z}_0$ generated by the diffusion network, decodes it to a high-resolution output, and finally parses it to obtain an enhanced point cloud P_{e}.

D. Training Details

Training Protocol: We use a two-stage scheme followed by short joint fine-tuning. First, a symmetric encoder–decoder is pre-trained on ground-truth point clouds with a visibility-weighted reconstruction objective; the encoder is then frozen and used to compute latent codes z_0. Second, the conditional LDM is trained on these fixed z_0 while supervision in explicit space through the frozen decoder. Finally, we unfreeze the decoder and perform a brief joint fine-tuning with a smaller learning rate for the decoder.

Objective Function: We optimize a single objective that couples the latent diffusion loss with explicit geometric terms. In the diffusion domain we use the standard noise-prediction loss [28]:

$$L_{\text{diffusion}} = \mathbb{E}_{t, z_0, \varepsilon}\left\|\varepsilon - \varepsilon_\theta(z_t, t, C_t)\right\|_2^2. \tag{13}$$

After decoding, we obtain the enhanced point cloud P_{e} and use the LiDAR point cloud as ground-truth P_{GT}. A Chamfer Distance Loss in the explicit space is introduced:

$$L_{\text{CD}} = \sum_{p\in P_{\text{e}}}\min_{q\in P_{\text{GT}}}\|p-q\|_2^2 + \sum_{q\in P_{\text{GT}}}\min_{p\in P_{\text{e}}}\|q-p\|_2^2. \tag{14}$$

To capture higher-order geometric details, a perceptual loss based on a frozen PointNet++ feature extractor Φ is added:

$$L_p = \left\|\Phi\big(P_{\text{e}}\big) - \Phi\big(P_{\text{GT}}\big)\right\|_2^2. \tag{15}$$

Overall, the final objective is

$$L_{\text{total}} = L_{\text{diffusion}} + \lambda_{\text{CD}}\, L_{\text{CD}} + \lambda_p\, L_p. \tag{16}$$

where λ_{CD} and λ_p are hyper-parameters.

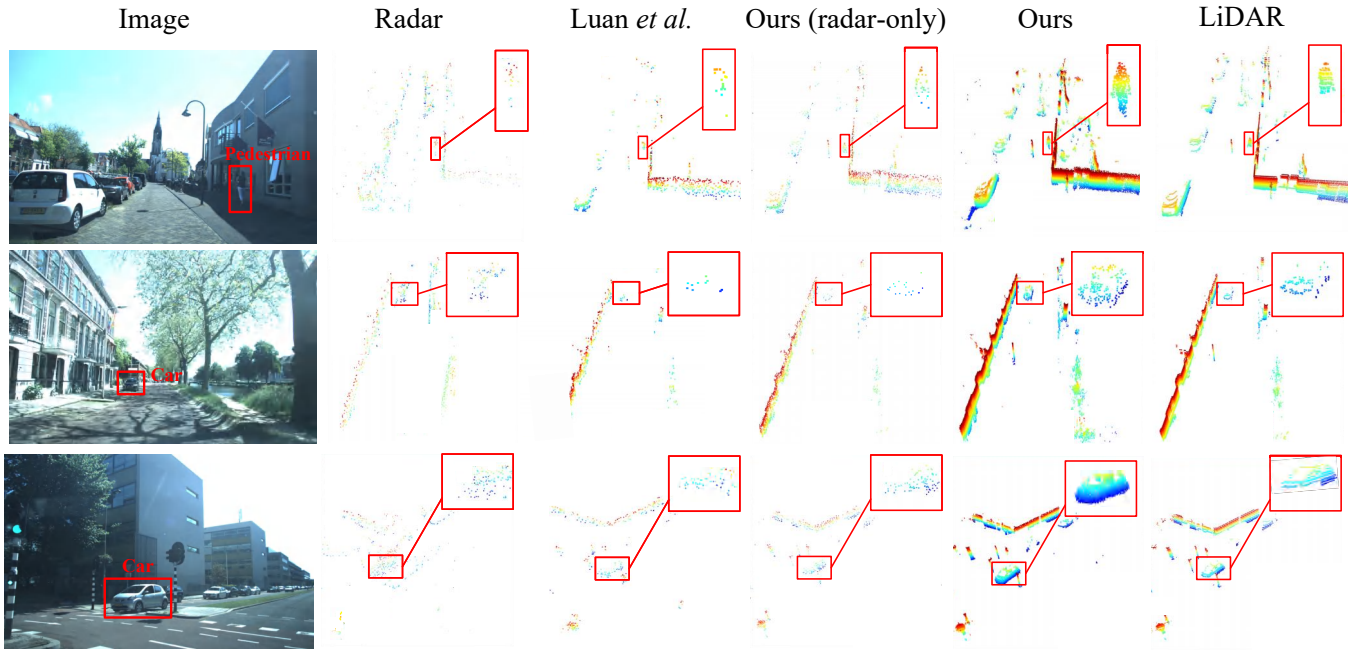


Fig. 4. **Qualitative performance comparisons of radar enhancement on the VoD dataset.** Compared to previous work, our method not only densifies the point clouds but also recovers the geometric outlines of pedestrians and vehicles in the image view, thanks to the incorporation of visual aids, as demonstrated by the magnified results in the small window.

IV. EXPERIMENT

A. Dataset

The View-of-Delft (VoD) dataset [30] contains 8600 frames of synchronized and calibrated 64-layer LiDAR, camera, and 4D radar data collected in complex urban traffic, and provides pose ground truth. Sequences 01, 02, 04, 08, 14, and 19 are allocated as the test set, with the remaining sequences for training.

NTU4DRadLM [31] provides measurements from a Livox Horizon LiDAR, a 640×480 monocular camera, an Oculii G7 4D millimetre-wave radar and the ground-truth of robot pose. Sequence Cp is allocated as the test set and sequence Garden, and Nyl (recorded at approximately $3.6\,\mathrm{km\,h^{-1}}$) for training.

B. Implementation Details

The conditional diffusion model and the joint fine-tuning process are trained on eight Nvidia RTX A6000 GPUs, and the visual-aid fusion network is trained on an Nvidia RTX 4090 GPU. The whole network is implemented on the Pytorch 2.3.1 framework. The diffusion model requires 50 steps to achieve the denoising results and is trained for 350k steps with a learning rate of 1e-5. All input images are resized to 640×320. All radar and LiDAR data superimposition over 5 consecutive frames. During inference, the decoder parameters ϕ remain fixed, all evaluations are performed in the explicit point-cloud domain.

C. Evaluation Metrics

We benchmark our method against several methods on the VoD dataset: the learning-based methods RadarHD [11] and RPDNet [9], and recent diffusion-model-based methods by Luan [19] and Zhang *et al.* [20].

For quantitative analysis, we employ the same metrics as [20] for evaluation, specifically Chamfer distance (CD), Hausdorff distance (HD), and F-Score [32]. Since [9], [11], [20] can only process 2D radar, for a fair comparison, we conduct 2D but not 3D evaluations for these methods and our own method. In experiments, we remove the Z-axis information from the radar point cloud.

D. Qualitative Analysis

The radar enhancement by different methods alongside corresponding LiDAR collected from the same location are visualized for comparison, as illustrated in Fig. 4. Compared with the raw radar input, both variants of our method reduce ghost returns and fill large BEV empty space, approaching the single-frame LiDAR point cloud in contour completeness. Compared with [19], our radar-only variant yields fewer outliers and better edge continuity, while the variant with visual fusion further improves structural fidelity and point-cloud precision in cluttered regions. The increased structural density leads to cleaner reconstructions and more accurate geometry around vegetation and specular facades, and even recovers complete geometric contours of objects that are only partially visible in LiDAR (*e.g.,* pedestrians and vehicles), as highlighted in the magnified windows.

E. Quantitative Comparison and Enhanced Results

a) Analysis in 2D Metirc: As shown in Tab. I, the proposed radar–camera method achieves the best performance on Chamfer Distance (CD), Hausdorff Distance (HD), and F-Score across all VoD sequences. Compared with the

TABLE I

2D QUANTITATIVE COMPARISONS ON THE VOD DATASET.

	01			02			04			08			14			19		
	CD ↓	HD ↓	F-Score ↑	CD ↓	HD ↓	F-Score ↑	CD ↓	HD ↓	F-Score ↑	CD ↓	HD ↓	F-Score ↑	CD ↓	HD ↓	F-Score ↑	CD ↓	HD ↓	F-Score ↑
Original	1.623	8.903	0.273	1.654	9.198	0.262	1.568	8.804	0.279	1.603	9.097	0.272	1.592	8.703	0.293	1.632	9.004	0.268
RPDNet [9]	1.983	10.804	0.169	2.012	10.903	0.158	1.922	10.402	0.189	1.957	10.703	0.179	1.902	10.103	0.201	1.993	10.597	0.178
Zhang [20]	1.283	7.603	0.332	1.227	7.402	0.352	1.213	7.198	0.361	1.257	7.597	0.341	1.242	7.298	0.358	1.272	7.502	0.351
RadarHD [11]	1.692	9.402	0.251	1.714	9.503	0.241	1.642	9.102	0.258	1.682	9.297	0.252	1.662	8.903	0.268	1.703	9.398	0.249
Luan *et al.* [19]	1.021	6.103	0.412	1.063	6.002	0.398	1.002	6.297	0.423	0.983	5.997	0.438	1.029	6.198	0.432	1.038	6.102	0.419
Ours (Radar-only)	1.041	5.903	0.421	1.028	5.802	0.409	1.013	6.098	0.432	1.002	5.698	0.448	1.043	5.998	0.441	1.048	5.902	0.438
Ours (Radar-camera)	**0.861**	**5.102**	**0.502**	**0.852**	**5.003**	**0.509**	**0.843**	**5.198**	**0.522**	**0.832**	**4.902**	**0.548**	**0.853**	**5.098**	**0.531**	**0.858**	**5.001**	**0.539**

TABLE II

3D QUANTITATIVE COMPARISONS ON THE CP SEQUENCE.

Methods	CD ↓	HD ↓	F-Score ↑
Original	1.744	7.260	0.267
Luan *et al.* [19]	1.334	5.585	0.314
Ours (Radar-only)	1.179	4.717	0.399
Ours	**0.329**	**2.027**	**0.598**

second-best baseline, it improves CD↓ by at least 15.4%, HD↓ by at least 13.6%, and F-Score↑ by at least 19.2%. Incorporating visual information into radar enhancement further yields at least 16.8% CD↓, 13.6% HD↓, and 19.2% F-Score↑, demonstrating that cross-modal fusion is critical for suppressing clutter and reconstructing occluded structures. These results confirm the necessity of the two-stage design, where denoising provides a clean prior and vision-guided densification enforces structural consistency.

b) Analysis in 3D Metirc: As shown in Tab II, our method outperforms the second-best baseline by at least 24.4% in CD↓, 24.2% in HD↓, and 90.5% in F-Score↑ in the CP sequence of NTU4DRadLM. Unlike the 2D case where the improvements are dominated by distance-based metrics, the most significant gain in 3D lies in F-Score, reflecting a +90.5% increase compared with only +19.2% in 2D. This highlights that volumetric completion along the z axis, enabled by orthogonal-view feature extraction and attention alignment, contributes substantially more than the planar BEV-based methods to establishing correct correspondences.

F. Evaluation of Visual Occlusion

To verify the robustness of our method in extreme environments, we evaluate it on the NTU4DRadLM-CP-Smoke sequence [31] with smoke interference and the same trajectory NTU4DRadLM-CP sequence without smoke interference. As shown in Fig. 1e, in a smoky environment, the LiDAR can only detect minimal noise, and the camera is also obscured by smoke, whereas our method remains effective even during the sudden appearance and dissipation of smoke in normal environments, as shown in Fig. 1f. In smoky areas where visual input is highly unreliable, our point cloud enhancement methods degenerate to reasoning based solely on the mmWave radar point cloud. The robust enhancement performance highlights the adaptive fusion capability of our VAF module based on the quality of the visual images.

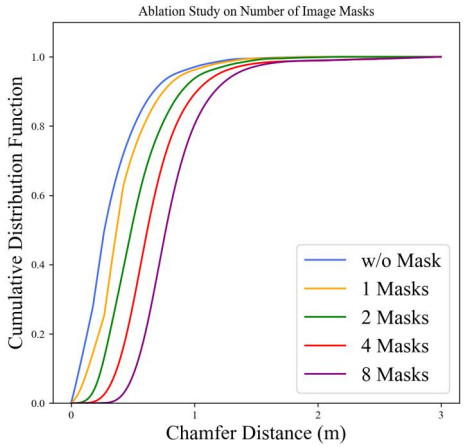


Fig. 5. The effects of the restricted visual inputs on the proposed method are shown in CDF curves. Proximity to the vertical axis indicates more favorable enhancement results.

To quantify the effectiveness of cross-modal densification under restricted visual input, we divide the image of the NTU4DRadLM-CP sequence into 32 even blocks and randomly mask some blocks before conducting the tests. Fig. 5 shows that the densification effect remains stable when not all image blocks are obscured. The radar-only variant, corresponding to masking all 32 blocks, records CD 1.179 (as shown in Tab. II). When one quarter of the image is masked (8 of 32 blocks), CD reaches 0.856, reducing CD by 27.4% relative to the radar-only variant, showing the robustness of radar enhancement under restricted visual evidence.

G. Qualitative Analysis on Downstream Tasks

Intuitively, geometrically reliable radar point clouds can open new paradigms for downstream tasks. To validate it, we integrate the proposed method into the latest publicly available LiDAR-vision SLAM baseline LiV-GS [33] that adopts a 3D Gaussian Splatting (3DGS) map representation. To the best of the authors' knowledge, no scene–reconstruction system currently couples radar measurements with 3DGS because raw 4D mmWave radar measurements have angular resolution orders of magnitude coarser and stronger noises than LiDAR, preventing the generation of stable Gaussian ellipsoids. This is why LiV-GS with raw radar (and vision) input fails to complete the entire trajectory, as shown in

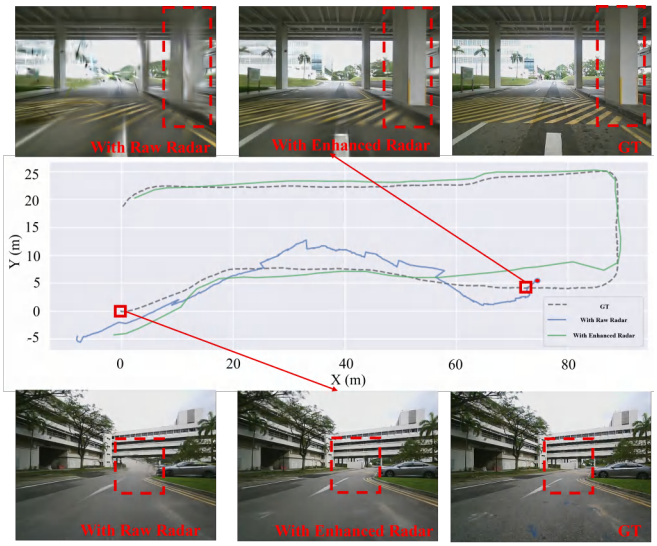


Fig. 6. **Downstream validation of our radar enhancement method.** Top and bottom: rendering results at the positions indicated by the red squares along the trajectory. Middle: odometries of LiV-GS with raw radar and with enhanced radar. The LiV-GS input with radar point clouds enhanced by our methods yields comparable localization accuracy and novel view synthesis fidelity to that using LiDAR and vision.

(a) Enhanced radar data depth (b) LiDAR data depth

Fig. 7. **Failure-case analysis.** (a) The enhanced radar point cloud projected onto the image plane. (b) The corresponding depth projection of LiDAR data.

Fig. 6. On the contrary, LiV-GS with enhanced radar by our method (and vision) not only completes accurate odometry along the entire trajectory, but also supplies the missing geometric prior required for renderable mapping.

H. Limitation

In experiments, we encountered a failure case happening in the CP sequence in the NTU4DRadLM dataset [31]. This is because millimetre-wave radar echoes can penetrate transparent surfaces such as glass curtain walls and return points from objects behind the glass. By contrast, LiDAR, which serves as the supervisory signal for data enhancement, lacks this capability. As shown in Fig. 7, this mismatch in the sensing model results in an erroneous dense point cloud distribution.

V. Conclusion

In this work, we propose a conditional diffusion model that leverages synchronized 4D radar sweeps and RGB images for radar data enhancement. The method projects radar point clouds into orthogonal views, fusing the depth view with synchronized images while encoding the BEV representation independently. The diffusion model is then conditioned on the fused visual–depth features and the BEV tokens to generate dense, geometrically continuous point clouds. Extensive experiments confirm that our method not only densifies the radar point clouds but also recover the vertical structure with distinct visual features, even in areas only partially visible in LiDAR. Furthermore, the augmented point clouds enable successful renderable reconstruction, which is infeasible with raw radar data, highlighting the advantage introduced by multimodal data fusion in our method. In future work, we plan to integrate physical models of heterogeneous sensors into the framework, aiming to build unified cross-modal enhancement representations that mitigate the sensing-model mismatch observed in challenging cases.

References

[1] K. Harlow, H. Jang, T. D. Barfoot, A. Kim, and C. Heckman, "A new wave in robotics: Survey on recent mmwave radar applications in robotics," *IEEE Transactions on Robotics*, vol. 40, pp. 4544–4560, 2024.

[2] S. Yao, R. Guan, Z. Peng, C. Xu, Y. Shi, W. Ding, E. G. Lim, Y. Yue, H. Seo, K. L. Man, J. Ma, X. Zhu, and Y. Yue, "Exploring radar data representations in autonomous driving: A comprehensive review," *IEEE Transactions on Intelligent Transportation Systems*, vol. 26, no. 6, pp. 7401–7425, 2025.

[3] S. Li, Z. Hong, Y. Chen, L. Hu, and J. Qin, "Get it for free: Radar segmentation without expert labels and its application in odometry and localization," *IEEE Robotics and Automation Letters*, vol. 10, no. 3, pp. 2678–2685, 2025.

[4] J. Zhang, H. Zhuge, Z. Wu, G. Peng, M. Wen, Y. Liu, and D. Wang, "4DRadarSLAM: A 4D imaging radar SLAM system for large-scale environments based on pose graph optimization," in *2023 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2023, pp. 8333–8340.

[5] Y. Zhang, R. Xiao, Z. Hong, L. Hu, and J. Liu, "Adaptive Visual-Aided 4D Radar Odometry Through Transformer-Based Feature Fusion," in *2024 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2024, pp. 12 529–12 535.

[6] Y. Cheng, M. Jiang, and Y. Liu, "Ms-vro: A multistage visual–millimeter-wave radar fusion odometry," *IEEE Transactions on Robotics*, vol. 40, pp. 3004–3023, 2024.

[7] K. Burnett, A. P. Schoellig, and T. D. Barfoot, "Continuous-time radar-inertial and lidar-inertial odometry using a gaussian process motion prior," *IEEE Transactions on Robotics*, vol. 41, pp. 1059–1076, 2025.

[8] J. Zhang, R. Xiao, H. Li, Y. Liu, X. Suo, C. Hong, Z. Lin, and D. Wang, "4DRT-SLAM: Robust SLAM in Smoke Environments using 4D Radar and Thermal Camera based on Dense Deep Learnt Features," in *2023 IEEE International Conference on Cybernetics and Intelligent Systems (CIS) and IEEE Conference on Robotics, Automation and Mechatronics (RAM)*, 2023, pp. 19–24.

[9] Y. Cheng, J. Su, M. Jiang, and Y. Liu, "A Novel Radar Point Cloud Generation Method for Robot Environment Perception," *IEEE Transactions on Robotics*, vol. 38, no. 6, pp. 3754–3773, 2022.

[10] M. Jiang, G. Xu, H. Pei, Z. Feng, S. Ma, H. Zhang, and W. Hong, "4D High-Resolution Imagery of Point Clouds for Automotive mmWave Radar," *IEEE Transactions on Intelligent Transportation Systems*, vol. 25, no. 1, pp. 998–1012, 2024.

[11] A. Prabhakara, T. Jin, A. Das, G. Bhatt, L. Kumari, E. Soltanaghai, J. Bilmes, S. Kumar, and A. Rowe, "High Resolution Point Clouds from mmwave Radar," in *2023 IEEE International Conference on Robotics and Automation (ICRA)*, 2023, pp. 4135–4142.

[12] Y. Cheng, J. Su, H. Chen, and Y. Liu, "A New Automotive Radar 4D Point Clouds Detector by Using Deep Learning," in *ICASSP 2021 - 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2021, pp. 8398–8402.

[13] R. Geng, Y. Li, D. Zhang, J. Wu, Y. Gao, Y. Hu, and Y. Chen, "DREAM-PCD: Deep Reconstruction and Enhancement of mmWave Radar Pointcloud," *IEEE Transactions on Image Processing*, vol. 33, pp. 6774–6789, 2024.

[14] P. Cai and S. Sur, "MilliPCD: Beyond Traditional Vision Indoor Point Cloud Generation via Handheld Millimeter-Wave Devices," *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, vol. 6, no. 4, 2022.

[15] Z. Han, J. Jiang, X. Ding, J. Wang, Q. Meng, S. Xu, L. He, and J. Wang, "DenserRadar: A 4D Millimeter-Wave Radar Point Cloud Detector Based on Dense LiDAR Point Clouds," in *2024 IEEE 27th International Conference on Intelligent Transportation Systems (ITSC)*, 2024, pp. 930–936.

[16] Y. Cheng, J. Su, M. Jiang, and Y. Liu, "A Novel Radar Point Cloud Generation Method for Robot Environment Perception," *IEEE Transactions on Robotics*, vol. 38, no. 6, pp. 3754–3773, 2022.

[17] C. Fan, S. Zhang, K. Liu, S. Wang, Z. Yang, and W. Wang, "Enhancing mmWave Radar Point Cloud via Visual-inertial Supervision," in *2024 IEEE International Conference on Robotics and Automation (ICRA)*, 2024, pp. 9010–9017.

[18] Z. Li, F. Ai, Y. Song, W. Wu, C. Song, and Z. Xu, "PCGNet: Point Cloud Generation Network for 3D Perception using Monocular Images and Radar," *IEEE Transactions on Intelligent Vehicles*, pp. 1–13, 2024.

[19] K. Luan, C. Shi, N. Wang, Y. Cheng, H. Lu, and X. Chen, "Diffusion-Based Point Cloud Super-Resolution for mmWave Radar Data," in *2024 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2024, pp. 11 171–11 177.

[20] R. Zhang, D. Xue, Y. Wang, R. Geng, and F. Gao, "Towards dense and accurate radar perception via efficient cross-modal diffusion model," *IEEE Robotics and Automation Letters*, 2024.

[21] J. Wu, R. Geng, Y. Li, D. Zhang, Z. Lu, Y. Hu, and Y. Chen, "Diffradar: High-Quality Mmwave Radar Perception With Diffusion Probabilistic Model," in *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2024, pp. 8291–8295.

[22] A. Tuel, T. Kerdreux, C. Hulbert, and B. Rouet-Leduc, "Diffusion models for interferometric satellite aperture radar," *arXiv preprint arXiv:2308.16847*, 2023.

[23] R. Wu, Z. Li, J. Wang, X. Xu, H. Yu, Z. Zheng, K. Huang, and G. Lu, "Diffusion-Based mmWave Radar Point Cloud Enhancement Driven by Range Images," *arXiv preprint arXiv:2503.02300*, 2025.

[24] Y. Yang, J. Liu, G. Luo, H. Li, E. Ahn, M. R. Azghadi, and T. Huang, "Unsupervised Radar Point Cloud Enhancement via Arbitrary LiDAR Guided Diffusion Prior," *arXiv preprint arXiv:2505.09887*, 2025.

[25] B. Zheng, S. Lu, R. Huang, M. Huang, F. Lu, W. Tian, G. Zhuo, and L. Xiong, "R2LDM: An Efficient 4D Radar Super-Resolution Framework Leveraging Diffusion Model," *arXiv preprint arXiv:2503.17097*, 2025.

[26] R. Zhang and F. Gao, "Sem-RaDiff: Diffusion-Based 3D Radar Semantic Perception in Cluttered Agricultural Environments," *arXiv preprint arXiv:2509.02283*, 2025.

[27] S. Lee, H. Lim, and H. Myung, "Patchwork++: Fast and Robust Ground Segmentation Solving Partial Under-Segmentation Using 3D Point Cloud," in *Proc. IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 2022, pp. 13 276–13 283.

[28] J. Ho, A. Jain, and P. Abbeel, "Denoising diffusion probabilistic models," *Advances in neural information processing systems*, vol. 33, pp. 6840–6851, 2020.

[29] C. Lu, Y. Zhou, F. Bao, J. Chen, C. Li, and J. Zhu, "DPM-Solver++: Fast Solver for Guided Sampling of Diffusion Probabilistic Models," *arXiv preprint arXiv:2211.01095*, 2022. [Online]. Available: https://arxiv.org/abs/2211.01095

[30] A. Palffy, E. Pool, S. Baratam, J. F. P. Kooij, and D. M. Gavrila, "Multi-Class Road User Detection With 3+1D Radar in the View-of-Delft Dataset," *IEEE Robotics and Automation Letters*, vol. 7, no. 2, pp. 4961–4968, 2022.

[31] J. Zhang, H. Zhuge, Y. Liu, G. Peng, Z. Wu, H. Zhang, Q. Lyu, H. Li, C. Zhao, D. Kircali *et al.*, "NTU4DRaDLM: 4D radar-centric multi-modal dataset for localization and mapping," in *2023 IEEE 26th International Conference on Intelligent Transportation Systems (ITSC)*. IEEE, 2023, pp. 4291–4296.

[32] M. Tatarchenko, S. R. Richter, R. Ranftl, Z. Li, V. Koltun, and T. Brox, "What do single-view 3D reconstruction networks learn?" in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2019, pp. 3405–3414.

[33] R. Xiao, W. Liu, Y. Chen, and L. Hu, "LiV-GS: LiDAR-Vision Integration for 3D Gaussian Splatting SLAM in Outdoor Environments," *IEEE Robotics and Automation Letters*, vol. 10, no. 1, pp. 421–428, 2025.