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

PSKDNet: Position-Supervised Keypoints Diffusion Network for Online Vectorized HD Map Construction

Mingkun Jiang[1,2], Jun Dong[2,*], Junming He[2], Guangyu Hou[1], Fan Ma[2], Shuang Wu[2], and Yujing Zhang[2]

***Abstract*— Online high-definition map construction represents a critical challenge in autonomous driving systems. Existing approaches suffer from generalization degradation when confronted with domain shifts across different geographical regions, particularly when facing limited training data in novel scenarios. To address this issue, we propose PSKDNet, a Position-Supervised Keypoints Diffusion Network that applies keypoints diffusion models to vectorized map for the first time. Our approach introduces Spatio-Temporal Keypoints Diffusion, which provides additional supervisory information through the diffusion process, thereby enhancing model generalization under domain shifts.To ensure accurate supervision of spatial relationships between map elements, we propose the Progressive Position Relation Transformer, which employs pointset similarity networks to obtain learnable position masks for explicitly supervising spatial relationships between map elements. Extensive experiments on nuScenes and Argoverse datasets demonstrate that PSKDNet achieves superior performance over state-of-the-art methods, with significant improvements in detection accuracy and robustness to environmental variations. To the best of our knowledge, this work represents the first successful application of diffusion models to vectorized map construction, opening new research directions in autonomous driving perception systems.**

I. INTRODUCTION

High-definition (HD) maps constitute fundamental infrastructure for autonomous driving systems, serving essential roles in vehicle localization, navigation, and path planning [1], [2], [3]. Traditional HD map construction has relied upon offline methodologies including simultaneous localization and mapping (SLAM) [4], [5] and photogrammetry [6], which require substantial human intervention and temporal investment while showing limited adaptability to dynamic road environments. Contemporary research has shifted toward online map construction using vehicle-mounted sensors [7], [8], [9]. These methodologies employ conventional frameworks that generate invariant outputs for given inputs, avoiding stochastic modeling and reducing complex perception problems to direct feature-to-geometry mappings. Real-world deployment scenarios face a critical challenge, when autonomous vehicles encounter novel geographical regions or road structure, the available training data becomes severely limited, leading to substantial performance degradation. This limited training data problem is particularly acute for complex geometric structures in vectorized maps, where insufficient information fails to capture the intricate positional relationships inherent in map elements. Traditional approaches struggle with this data scarcity as they rely solely on direct supervision from limited ground truth annotations, lacking additional supervisory signals to guide learning in novel scenarios. Diffusion models offer a promising solution by providing rich supervisory information through their iterative denoising process [10]. As illustrated in Fig.1, this process transforms completely random noisy keypoints through multiple DDIM [11] sampling steps into structured map element predictions, providing additional supervision during training.

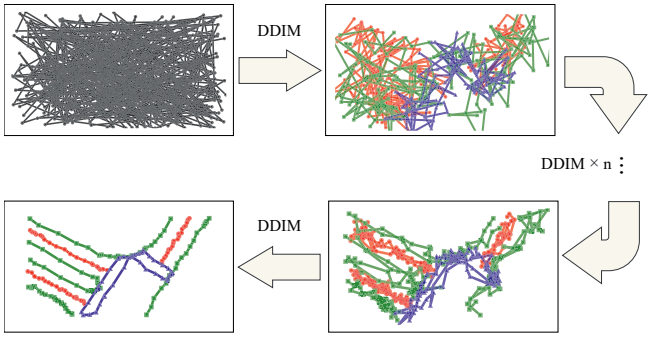

Fig. 1. Illustration of the DDIM denoising process for map element keypoint prediction. Starting from completely random noisy keypoints (top left), the model progressively denoises through multiple DDIM sampling steps to generate structured map element predictions (bottom left).

While diffusion models have been successfully applied to various domains, existing applications focus on different data modalities: DiffusionDet [12] applies diffusion to bounding boxes for object detection, LidarDM [13] targets point clouds for 3D generation, and PixelDiffusion [14] operates on pixel-level representations for image synthesis. However, none of these approaches address the unique challenges of vectorized map keypoints: first, the vectorized representation of map elements demands geometric precision substantially exceeding that of conventional object detection tasks, which requires careful design of the diffusion process for element keypoint-level supervision; second, the positional relationships between elements represent road structural information, which is also crucial for the credibility of scene construction.

In response to these challenges, we introduce PSKDNet, a novel framework that reformulates online HD map construction through keypoints diffusion modeling. To provide additional supervisory information for scenes in new geographical regions with differing distributions, we apply keypoints diffusion models to vectorized map, providing rich

[1]University of Science and Technology of China, Hefei, 230026
[2]Hefei Institutes of Physical Science, Chinese Academy of Sciences, Hefei, 230031.

ACKNOWLEDGMENT: This work was supported in part by The National Key Research and Development Program of China grant number 2022YFD2001400.

supervisory signals through the iterative denoising process. To ensure the accuracy of positional relationships between road elements, we design learnable position mask supervision that dynamically recovers structural information and adjusts the recovery intensity across different noise levels.

Our contributions can be summarized as follows:

- We propose PSKDNet, which reformulates online HD map construction through keypoints diffusion modeling via a Spatio-Temporal Keypoints Diffusion (STKD) that provides additional supervisory information for limited training data scenarios;
- We introduce the Progressive Position Relation Transformer (PPRT), which ensures the accuracy of positional relationships between road elements through learnable pointset similarity networks and position mask supervision, and propose three specialized evaluation metrics;
- Our method achieves superior performance over state-of-the-art approaches on both nuScenes and Argoverse2 datasets, with particularly significant improvements demonstrated on geographically-disjoint dataset splits where training data is limited;
- To our knowledge, this represents the first successful application of keypoints diffusion models to online HD map construction, opening new research directions in this domain.

II. Relation Works

A. Online Vectorized HD Map Construction

The landscape of online high-definition map construction in autonomous driving has undergone a paradigmatic shift from rasterized to vectorized representations. While HDMapNet [15] pioneered rasterized mapping via multi-view semantic segmentation, its computational inefficiency motivated vectorized alternatives. VectorMapNet [7] established keypoint-sequence representations achieving optimal storage-geometry balance. MapTR [16] integrated transformers for end-to-end vectorized learning, validated by benchmark supremacy. Subsequent refinements addressed specific challenges: MapTRv2 [17] enhanced multi-scale detection, StreamMapNet [8] incorporated temporal modeling, and BeMapNet [18] explored LiDAR-vision fusion. Emerging uncertainty research includes LaRa's [19] Bayesian frameworks and PivotNet's [20] keypoint confidence mechanisms. However, our method, within the overall framework, combines the previous conventional methods with the diffusion model of uncertainty, achieving stronger scene generalization.

B. Detection Transformers

Previous works for online high-definition map construction algorithms are inherently grounded in Detection Transformers. DETR [21] established object detection as a set prediction problem, eliminating complex hand-crafted components. While groundbreaking, its challenges like slow convergence spurred numerous refinements. Subsequent research focused heavily on computational efficiency, with Deformable DETR [22] reducing complexity via sparse attention, and Conditional DETR [23] accelerating convergence. Training stability improvements were achieved through methods like DN-DETR's [24] denoising training. Recent probabilistic approaches include DiffusionDet [12] integrating diffusion models and Particle-DETR [25] using particle filtering. However, an inherent conflict exists between the stochastic nature of diffusion processes and the stringent precision requirements for geometric representation. Our methodology reconciles this fundamental tension through synergistic integration of diffusion models with conventional object detection frameworks, establishing an enhanced-performance network architecture. This hybrid paradigm demonstrates significant efficacy in online high-definition map construction.

III. Methodology

A. Overall Architecture

Our proposed PSKDNet adopts a keypoints diffusion architecture that provides additional supervisory information for robust HD map construction in scenarios with limited training data. As illustrated in Fig. 2, the framework processes surrounding multi-view images through a shared encoder to extract bird's-eye-view (BEV) feature representations. Then, a specialized BEV feature generator projects the processed 2D features into BEV space, yielding a unified BEV feature that serves as the shared foundation for subsequent keypoints diffusion inference. Based on this unified BEV feature, the STKD employs its keypoints diffusion process to provide enriched queries to the decoder, while the PPRT module provides attention masks. The decoder then integrates these queries and masks to generate classification scores and geometric predictions for map elements.

B. Spatio-Temporal Keypoints Diffusion (STKD)

STKD applies diffusion models directly to vectorized map keypoints, providing additional supervisory information through the iterative denoising process. The architecture combines conventional detection queries with keypoints diffusion supervision to enhance model generalization with limited training data. The keypoints diffusion process comprises three components: Noise Injection Network, Spatio-Temporal Multi-Point Query Interpolation Module, and Denoising Prediction Network.

Noise Injection Network. The forward diffusion process employs designed noise injection mechanisms that progressively transform ground-truth target positions into Gaussian distributions, following Markovian dynamics to ensure controllable noise injection. Similar to conventional diffusion networks, this component is only incorporated during training to learn noise patterns, while during inference, randomly initialized keypoints are used instead. For vectorized map elements modeled through multiple keypoints, each map element is represented as a geometric structure containing K keypoints: $\mathbf{M} \in \mathbb{R}^{N \times K \times D}$, where N denotes the number of map elements per frame, K represents keypoint count, and D

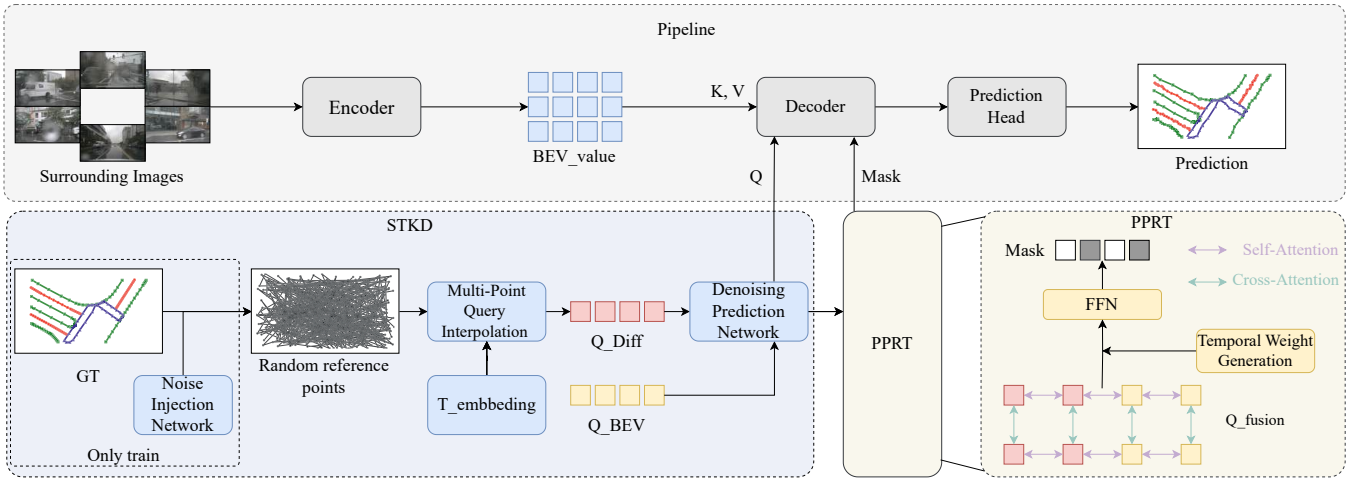

Fig. 2. Overview of the proposed PSKDNet architecture. The system processes surrounding multi-view images through an encoder to generate BEV features. The decoder and prediction head generate final map element predictions. Our proposed STKD provides queries while PPRT provides masks for the decoder.

indicates coordinate dimensionality. The diffusion noise injection process targets the complete geometric configuration of map elements:

$$\mathbf{M}_t = \sqrt{\bar{\alpha}_t}\mathbf{M}_0 + \sqrt{1-\bar{\alpha}_t}\boldsymbol{\eta} \in \mathbb{R}^{N\times K\times D} \tag{1}$$

where $\boldsymbol{\eta} \sim \mathcal{N}(0, \mathbf{I}_{N\times K\times D})$ represents a Gaussian noise tensor conforming to map element geometries, and the cumulative noise parameter sequence $\{\bar{\alpha}_t\}_{t=0}^{T}$ is defined with $\bar{\alpha}_t = \prod_{i=1}^{t}\alpha_i$ following a cosine scheduling strategy to ensure smooth noise injection.

Spatio-Temporal Multi-Point Query Interpolation Module. This module extends traditional query-based frameworks to multi-keypoint format: reference_points $\in \mathbb{R}^{N_{query}\times K\times D}$, where K denotes keypoint count and D represents coordinate dimensions, enabling precise modeling of vectorized map elements under noise conditions.

For noisy elements $\mathbf{M}_t \in \mathbb{R}^{N\times K\times D}$, the module employs spatial sampling and temporal modulation:

$$\mathbf{q}_i^{spatial} = \sum_{k=1}^{K}\omega_{i,k}\cdot \text{GridSample}(\mathbf{F}_{BEV}, 2\mathbf{M}_{t,i,k}-1) \tag{2}$$

$$\mathbf{q}_i^{modulated} = \text{LN}(\mathbf{q}_i^{spatial}\odot(1+\gamma_t)+\beta_t) \tag{3}$$

where $\omega_{i,k}$ represents keypoint importance weights, $\gamma_t = \text{MLP}_{scale}(\mathbf{e}_t)$, $\beta_t = \text{MLP}_{shift}(\mathbf{e}_t)$, $\mathbf{e}_t$ is time embedding, and LN denotes LayerNorm. The complete query interpolation implements feature transformations through decoder cascading: $\mathbf{h}_{l+1} = \text{Decoder}_l\left(\mathbf{q}^{modulated}, \mathbf{F}_{BEV}, \mathbf{e}_t\right)$.

Denoising Prediction Network. This network ϵ_θ predicts noise distributions at each timestep by implementing time-conditioned feature transformations through decoder layers. Within each decoder layer, the system utilizes noisy keypoint positions $\mathbf{x}_t$ generated by the diffusion process to extract corresponding spatial features $\mathbf{q}^{Diff}$ from BEV feature maps, subsequently performing feature concatenation with conventional queries $\mathbf{q}^{BEV}$. This design provides additional supervisory information through the keypoints diffusion process while maintaining spatial consistency. For computational efficiency optimization and detection accuracy enhancement, the system implements TopK selection mechanisms following each decoder layer, retaining only the highest-confidence K queries for subsequent processing:

$$\mathbf{q}_{fused} = \text{TopK}\left(\text{Concat}\left[\mathbf{q}^{BEV}, \mathbf{q}^{Diff}\right]\right) \tag{4}$$

C. Progressive Position Relation Transformer (PPRT)

The PPRT comprises two core components: a Learnable PointSet Similarity Network that extracts structured geometric descriptors from multi-keypoint map elements, and a Progressive Position Relation Supervision Module that transforms these features into spatial relationship weight matrices.

Learnable PointSet Similarity Network. The Learnable PointSet Similarity Network employs multi-head attention mechanisms to extract geometric descriptors from keypoint coordinates. The network embeds two-dimensional coordinates into high-dimensional features through linear transformations and learnable positional encodings. Unlike traditional geometric relationship formulas for rectangular bounding boxes with parameters $[x, y, w, h]$, this network designs specialized geometric descriptors for map elements with 20 keypoints that capture intrinsic structural characteristics. The computational pipeline includes point-level self-attention for intra-point-set dependencies, cross-attention for geometric similarities between point sets, and feedforward networks for descriptor generation. The PointSet Similarity Network computation is formulated as:

$$\mathbf{S}_{ij}^{\text{geo}} = \text{FFN}(\text{CrossAttn}(\text{SelfAttn}(\mathbf{P}_i), \text{SelfAttn}(\mathbf{P}_j))) \tag{5}$$

where $\mathbf{P}_i$ and $\mathbf{P}_j$ represent keypoint sets of the i-th and j-th map elements respectively, each containing K coordinate points.

TABLE I

COMPARISON WITH SOTA METHODS ON THE NUSCENES VALIDATION SET AT $60m \times 30m$ PERCEPTION RANGE.

Method	Backbone	Epochs	Temporal	AP_{ped}	AP_{div}	AP_{bou}	mAP
MapTR [16]	R50	24		46.3	51.5	53.1	50.3
MapVR [26]	R50	24		47.7	54.4	51.4	51.2
PivotNet [20]	R50	30		56.2	56.5	60.1	57.6
BeMapNet [18]	R50	30		57.7	62.3	59.4	59.8
MapTRv2 [17]	R50	24		59.8	62.4	62.4	61.5
StreamMapNet(Baseline)	R50	30	√	61.7	66.3	62.1	63.4
MGMap [27]	R50	24		61.8	65.0	67.5	64.8
SQD-MapNet [28]	R50	24	√	63.6	66.6	64.8	65.0
MapQR [29]	R50	24		63.4	68.0	67.7	66.4
HIMap [30]	R50	30		62.6	68.4	**69.1**	66.7
HRMapNet [31]	R50	24	√	65.8	67.4	68.5	67.2
PSKDNet(Ours)	R50	30	√	**67.5**	**70.4**	67.3	**68.4**(↑5.0)

Progressive Position Relation Supervision Module. This module transforms geometric similarity features from the PointSet Similarity Network into learnable position mask weights for decoder attention. We propose a temporal weight generation network that progressively modulates position relations according to diffusion timesteps. The network normalizes discrete diffusion timesteps t through $t/T_{\max}$ and processes them via fully-connected layers with ReLU activation to learn non-linear mappings from temporal information to weight vectors. The network outputs learnable position mask weights that dynamically modulate geometric position relationship embedding based on noise intensity:

$$\mathbf{M}_{\text{relation}}(t) = \text{Softmax}(\text{ReLU}(\mathbf{W} \cdot (t/T_{\max}) + \mathbf{b})) \quad (6)$$

This adaptive modulation provides noise-level-tailored embeddings. During early diffusion stages, higher weights preserve detailed geometric relationships, while weights progressively adjust as the denoising process advances to optimize the reconstruction of accurate positional relationships.

D. Loss Function

Our model employs a comprehensive loss function integrating conventional detection losses with keypoints diffusion losses:

$$\mathcal{L}_{total} = \lambda_{bev}\mathcal{L}_{bev} + \lambda_{diff}\mathcal{L}_{diff} \quad (7)$$

The BEV loss $\mathcal{L}_{BEV}$ follows StreamMapNet's design with classification, regression, and auxiliary components. The diffusion loss $\mathcal{L}_{diff}$ provides additional supervision through keypoints diffusion modeling across decoder layers. The diffusion process computes regression and classification losses independently at each layer:

$$\mathcal{L}^{(l)}_{diff\_reg} = \sum_{i=1}^{N^{(l)}} L_{SmoothL1}(\mathbf{x}^{(l,i)}_{pred}, \mathbf{x}^{(i)}_{gt}) \quad (8)$$

$$\mathcal{L}^{(l)}_{diff\_cls} = \sum_{i=0}^{N-1} L_{Focal}(\hat{p}_{\hat{\pi}(i)}, c_i) \quad (9)$$

TABLE II

COMPARISON WITH SOTA METHODS ON THE ARGOVERSE2 AT $60m \times 30m$ PERCEPTION RANGE.

Method	AP_{ped}	AP_{div}	AP_{bou}	mAP
StreamMapNet	62.0	59.5	63.0	61.5
SQD-MapNet [28]	64.9	60.2	64.9	63.3
MapTRv2 [17]	60.0	68.7	64.2	64.3
HRMapNet [31]	**65.1**	71.4	68.6	68.3
PSKDNet(Ours)	64.6	**72.2**	**68.8**	**68.5**(↑7.4)

where $\hat{\pi}$ represents instance-level optimal matching results. The complete diffusion loss aggregates across all L decoder layers:

$$\mathcal{L}_{diff} = \sum_{l=1}^{L} \left(\lambda_{cls}\mathcal{L}^{(l)}_{diff\_cls} + \lambda_{reg}\mathcal{L}^{(l)}_{diff\_reg} \right) \quad (10)$$

IV. EXPERIMENTAL

A. Datasets and Setup

We conduct experiments on two benchmark datasets: NuScenes [2] and Argoverse2 [32]. To address substantial location overlap in existing dataset partitions (84% in NuScenes, 54% in Argoverse2), we employ a dual validation enabling fair comparison with existing methods on original splits while evaluating generalization capabilities on geographically-disjoint partitions. Our re-partitioning follows StreamMapNet principles that minimizes spatial overlap while ensuring balanced geographical and meteorological distributions.

Our models are trained on eight GTX3090 GPUs with batch size 32, using AdamW optimizer with learning rate 5×10^{-4}. The network uses ResNet50 backbone and BEVFormer for BEV feature extraction. NuScenes uses 30 training epochs while Argoverse2 uses 24 epochs. Balanced loss weights $\lambda_{BEV} = 0.5$ and $\lambda_{diff} = 0.5$. The diffusion framework operates with $T = 1,000$ timesteps using cosine noise injection, while inference employs 8-step DDIM sampling.

B. Metrics

Following established protocols, we evaluate performance across three primary map element categories: pedestrian crossings, lane dividers, and road boundaries. mean Average

TABLE III

COMPARISON WITH SOTA METHODS ON THE NEW SPLIT DATASETS AT $60m \times 30m$ PERCEPTION RANGE.

Dataset	Method	Epoch	AP_{ped}	AP_{div}	AP_{bou}	mAP
nuScenes	VectorMapNet	120	15.8	17.0	21.2	18.0
	MapTR [16]	24	6.4	20.7	35.5	20.9
	StreamMapNet(Baseline)	24	29.6	30.1	41.9	33.9
	HRMapNet [31]	24	**36.9**	30.3	44.0	37.1
	PSKDNet(Ours)	24	36.6	**39.4**	**50.3**	**42.1**$^{(\uparrow 8.2)}$
Argoverse 2	VectorMapNet	120	35.6	34.9	37.8	36.1
	MapTR [16]	24	48.1	50.4	55.0	51.1
	StreamMapNet(Baseline)	24	56.9	55.9	61.4	58.1
	HRMapNet [31]	24	60.1	58.3	66.0	61.5
	PSKDNet(Ours)	24	**61.6**	**60.1**	**68.5**	**63.4**$^{(\uparrow 5.3)}$

Precision (mAP) serves as our primary evaluation metric. For validating our PPRT's effectiveness in preserving geometric structural information under diffusion noise, we propose three specialized metrics.

Position Dependency Metric. Measures spatial structural integrity by calculating the consistency of distance changes between adjacent points. Given point set $P = \{p_1, p_2, \ldots, p_n\}$, we compute distance sequence $d_i = \|p_{i+1} - p_i\|_2$ for $i = 1, 2, \ldots, n-1$, where the position dependency score is:

$$\text{PD} = \frac{1}{1 + \sigma(d)} \tag{11}$$

where $\sigma(d)$ is the standard deviation of the distance sequence.

Geometric Deviation Metric. Quantifies the average displacement between processed and original point sets. For original point set P_{orig} and processed point set P_{proc}:

$$\text{GD} = \frac{1}{n} \sum_{i=1}^{n} \|p_{\text{orig},i} - p_{\text{proc},i}\|_2 \tag{12}$$

Directional Consistency Metric. Evaluates orientation stability by computing the cosine similarity of direction vectors between adjacent line segments:

$$\text{DC} = \frac{1}{n-1} \sum_{i=1}^{n-1} \left| \frac{\vec{v}_{\text{orig},i} \cdot \vec{v}_{\text{proc},i}}{\|\vec{v}_{\text{orig},i}\|_2 \|\vec{v}_{\text{proc},i}\|_2} \right| \tag{13}$$

where $\vec{v}_{\text{orig},i} = p_{\text{orig},i+1} - p_{\text{orig},i}$ and $\vec{v}_{\text{proc},i} = p_{\text{proc},i+1} - p_{\text{proc},i}$ are direction vectors.

C. Comparison with state-of-the-art methods

Performance on nuScenes. Table I shows the comparative results on nuScenes validation set. PSKDNet achieves 68.4% mAP, improving 5.0 percentage points over baseline StreamMapNet with 63.4% mAP. Compared to the strongest temporal-based approach HRMapNet, PSKDNet delivers a 1.2 percentage point improvement, demonstrating the effectiveness of our approach.

Performance on Argoverse2. As shown in Table II, PSKDNet achieves 68.5% mAP on Argoverse2, improving 7.4 percentage points over StreamMapNet baseline with 61.5% mAP. Our approach excels in lane divider detection with 72.2% mAP, surpassing HRMapNet by 0.8 percentage points.

TABLE IV

ABLATION STUDY OF MAIN COMPONENTS ON NUSCENES DATASET.

STKD	PPRT	AP_{ped}	AP_{div}	AP_{bou}	mAP
		62.0	59.5	63.0	61.5
✓		63.7	68.6	67.8	66.7$^{(\uparrow 5.2)}$
	✓	62.9	66.7	64.2	64.6$^{(\uparrow 3.1)}$
✓	✓	**64.6**	**72.2**	**68.8**	**68.5**$^{(\uparrow 7.4)}$

TABLE V

QUANTITATIVE IMPROVEMENT RESULTS AT MAXIMUM NOISE LEVEL ($t = 1000$).

Method	PD ↑	GD ↓	DC ↑
Noise Only	0.210	9.340	0.708
With PPRT	0.469	3.203	0.763
Improvement	**+124.0%**	**−65.7%**	**+7.8%**

Performance on New Splits Table III shows performance on geographically-disjoint new data splits that represent domain shifts for evaluating generalization capabilities with limited training data. On nuScenes new split, PSKDNet achieves 42.1% mAP, surpassing StreamMapNet by 8.2 percentage points and HRMapNet by 4.0 percentage points. On Argoverse2's new split, our method attains 63.4% mAP, improving 5.3 percentage points over baseline. PSKDNet exhibits smaller performance degradation with 26.3pp on nuScenes and 5.1pp on Argoverse2 compared to baselines when transitioning from original to new splits under domain shifts. The differential impact correlates with data overlap ratios between Argoverse2's 54% overlap and nuScenes' 84% overlap. These results demonstrate PSKDNet's superior generalization when confronting domain shifts scenarios.

D. Ablation Study

Contributions of Main Components. Table IV demonstrates that STKD alone improves baseline by 5.2 mAP, demonstrating the effectiveness of keypoints diffusion modeling. PPRT alone contributes 3.1 mAP improvement, validating its spatial relationship supervision capability. The combination of both components achieves optimal performance with 7.4 mAP improvement, indicating synergistic integration of keypoints diffusion and position relation supervision .

Ablation Study of Position Relation Supervision. To validate PPRT effectiveness in preserving geometric structural

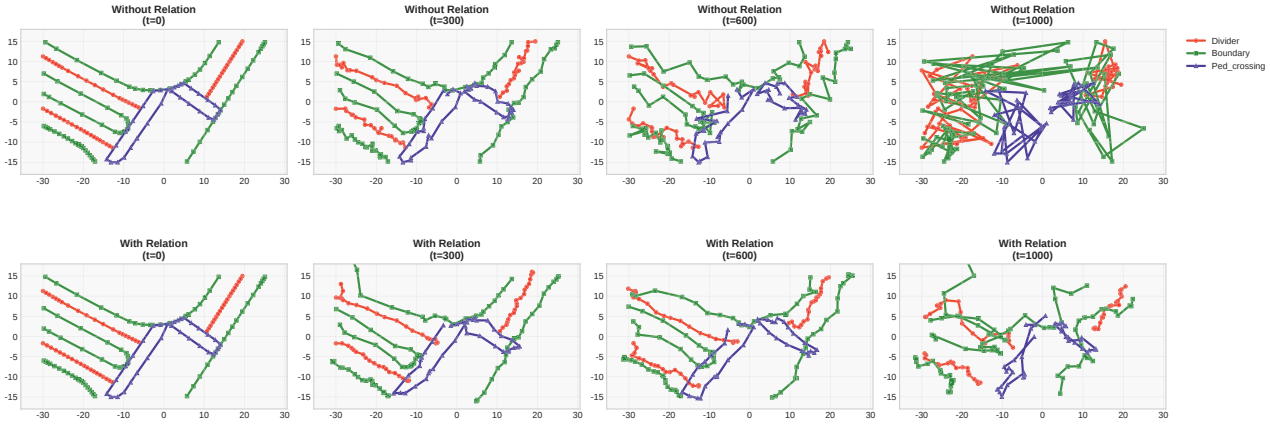

Fig. 3. Visual comparison of position relation supervision effects. The upper row shows the results without position relation supervision (Without Relation), while the lower row demonstrates the enhanced spatial relationship supervision achieved through PPRT (With Relation) across four critical timesteps.

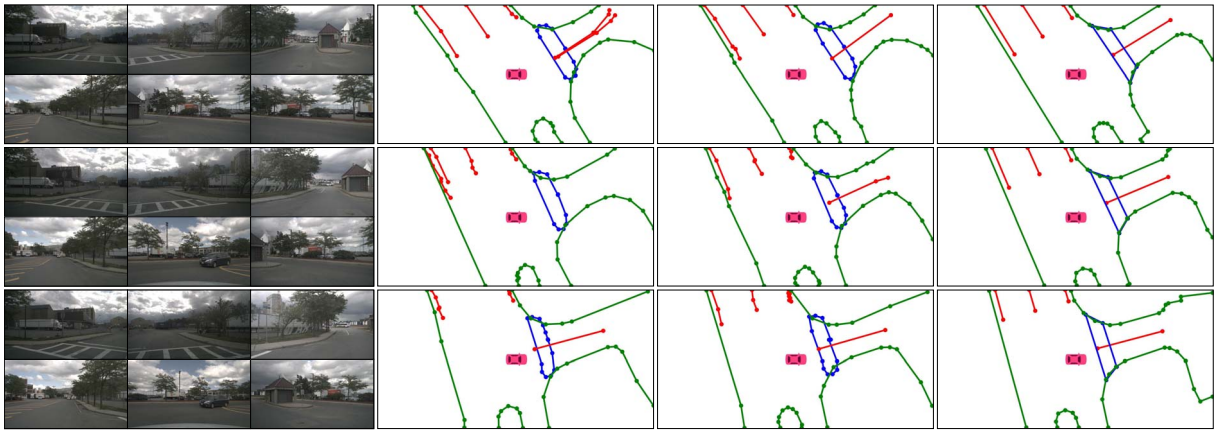

Fig. 4. Qualitative comparison between PSKDNet and StreamMapNet across three consecutive frames.

information under diffusion noise, we developed a comprehensive experimental framework using real map elements from the nuScenes dataset as ground truth.

Our qualitative analysis demonstrates PPRT's protective efficacy through visual comparison in Fig. 3. We evaluate four critical timesteps ($t = 0, 300, 600, 1000$) to simulate progressive noise intensification. Without position relation supervision, lane lines exhibit irregular distortions, road boundaries lose parallelism, and pedestrian crossings suffer severe deformation. PPRT maintains original geometric characteristics even at maximum noise levels ($t = 1000$). Table V shows quantitative improvements with 124.0% improvement in position dependency, 65.7% reduction in geometric deviation, and 7.8% enhancement in directional consistency, demonstrating that PPRT effectively restores the positional relationship between elements.

E. Qualitative Visualization

Fig. 4 presents qualitative comparisons between PSKDNet and baseline StreamMapNet across three consecutive frames, demonstrating superior detection consistency and accuracy. PSKDNet displays improved boundary delineation and reduced artifacts, maintaining spatio-temporal consistency and resulting in smoother map construction.

V. CONCLUSIONS

In this work, we introduced PSKDNet, which first introduced diffusion models into autonomous driving online HD map construction tasks to address the generalization degradation problem when networks encounter novel geographical regions with different distributions. The STKD applied keypoints diffusion models directly to vectorized map, providing additional supervisory information to enhance model generalization across diverse geographical environments. To ensure accurate modeling of positional relationships between road elements, we developed the PPRT, which employed learnable pointset similarity networks to enhance spatial relationships between map elements. Experimental validation on nuScenes and Argoverse2 datasets confirmed our approach's effectiveness, demonstrating that PSKDNet provided a robust framework for autonomous driving applications with enhanced generalization capabilities when facing geographical distribution shifts.

REFERENCES

[1] X. Chang, M. Xue, X. Liu, Z. Pan, and X. Wei, "Driving by the rules: A benchmark for integrating traffic sign regulations into vectorized hd map," in *Proceedings of the Computer Vision and Pattern Recognition Conference*, 2025, pp. 6823–6833.

[2] H. Caesar, V. Bankiti, A. H. Lang, S. Vora, V. E. Liong, Q. Xu, A. Krishnan, Y. Pan, G. Baldan, and O. Beijbom, "nuscenes: A multimodal dataset for autonomous driving," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2020, pp. 11 621–11 631.

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

[4] C. Cadena, L. Carlone, H. Carrillo, Y. Latif, D. Scaramuzza, J. Neira, I. Reid, and J. J. Leonard, "Past, present, and future of simultaneous localization and mapping: Toward the robust-perception age," *IEEE Transactions on robotics*, vol. 32, no. 6, pp. 1309–1332, 2017.

[5] K. Wang, J. Guo, K. Chen, and J. Lu, "An in-depth examination of slam methods: Challenges, advancements, and applications in complex scenes for autonomous driving," *IEEE Transactions on Intelligent Transportation Systems*, 2025.

[6] J. Levinson, J. Askeland, J. Becker, J. Dolson, D. Held, S. Kammel, J. Z. Kolter, D. Langer, O. Pink, V. Pratt *et al.*, "Towards fully autonomous driving: Systems and algorithms," in *2011 IEEE intelligent vehicles symposium (IV)*. IEEE, 2011, pp. 163–168.

[7] Y. Liu, T. Yuan, Y. Wang, Y. Wang, and H. Zhao, "Vectormapnet: End-to-end vectorized hd map learning," in *International Conference on Machine Learning*. PMLR, 2023, pp. 22 352–22 369.

[8] T. Yuan, Y. Liu, Y. Wang, Y. Wang, and H. Zhao, "Streammapnet: Streaming mapping network for vectorized online hd map construction," in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 2024, pp. 7356–7365.

[9] N. Peng, X. Zhou, M. Wang, X. Yang, S. Chen, and G. Chen, "Prevpredmap: Exploring temporal modeling with previous predictions for online vectorized hd map construction," in *2025 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*. IEEE, 2025, pp. 8134–8143.

[10] A. Q. Nichol and P. Dhariwal, "Improved denoising diffusion probabilistic models," in *International conference on machine learning*. PMLR, 2021, pp. 8162–8171.

[11] J. Song, C. Meng, and S. Ermon, "Denoising diffusion implicit models," *arXiv preprint arXiv:2010.02502*, 2020.

[12] S. Chen, P. Sun, Y. Song, and P. Luo, "Diffusiondet: Diffusion model for object detection," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2023, pp. 19 830–19 843.

[13] 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.

[14] T. Yang, R. Wu, P. Ren, X. Xie, and L. Zhang, "Pixel-aware stable diffusion for realistic image super-resolution and personalized stylization," in *European conference on computer vision*. Springer, 2024, pp. 74–91.

[15] Q. Li, Y. Wang, Y. Wang, and H. Zhao, "Hdmapnet: An online hd map construction and evaluation framework," in *2022 International Conference on Robotics and Automation (ICRA)*. IEEE, 2022, pp. 4628–4634.

[16] B. Liao, S. Chen, X. Wang, T. Cheng, Q. Zhang, W. Liu, and C. Huang, "Maptr: Structured modeling and learning for online vectorized hd map construction," in *International Conference on Learning Representations*, 2023.

[17] B. Liao, S. Chen, Y. Zhang, B. Jiang, Q. Zhang, W. Liu, C. Huang, and X. Wang, "Maptrv2: An end-to-end framework for online vectorized hd map construction," *International Journal of Computer Vision*, vol. 133, no. 3, pp. 1352–1374, 2025.

[18] L. Qiao, W. Ding, X. Qiu, and C. Zhang, "End-to-end vectorized hd-map construction with piecewise bezier curve," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 13 218–13 228.

[19] F. Bartoccioni, É. Zablocki, A. Bursuc, P. Pérez, M. Cord, and K. Alahari, "Lara: Latents and rays for multi-camera bird's-eye-view semantic segmentation," in *Conference on robot learning*. PMLR, 2023, pp. 1663–1672.

[20] W. Ding, L. Qiao, X. Qiu, and C. Zhang, "Pivotnet: Vectorized pivot learning for end-to-end hd map construction," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023, pp. 3672–3682.

[21] N. Carion, F. Massa, G. Synnaeve, N. Usunier, A. Kirillov, and S. Zagoruyko, "End-to-end object detection with transformers," in *European conference on computer vision*. Springer, 2020, pp. 213–229.

[22] X. Zhu, W. Su, L. Lu, B. Li, X. Wang, and J. Dai, "Deformable detr: Deformable transformers for end-to-end object detection," *arXiv preprint arXiv:2010.04159*, 2020.

[23] D. Meng, X. Chen, Z. Fan, G. Zeng, H. Li, Y. Yuan, L. Sun, and J. Wang, "Conditional detr for fast training convergence," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2021, pp. 3651–3660.

[24] F. Li, H. Zhang, S. Liu, J. Guo, L. M. Ni, and L. Zhang, "Dn-detr: Accelerate detr training by introducing query denoising," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2022, pp. 13 619–13 627.

[25] A. Nachkov, D. P. Paudel, M. Danelljan, and L. Van Gool, "Diffusion-based particle-detr for bev perception," in *2025 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*. IEEE, 2025, pp. 2725–2735.

[26] G. Zhang, J. Lin, S. Wu, Z. Luo, Y. Xue, S. Lu, Z. Wang *et al.*, "Online map vectorization for autonomous driving: A rasterization perspective," *Advances in Neural Information Processing Systems*, vol. 36, pp. 31 865–31 877, 2023.

[27] X. Liu, S. Wang, W. Li, R. Yang, J. Chen, and J. Zhu, "Mgmap: Mask-guided learning for online vectorized hd map construction," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 14 812–14 821.

[28] S. Wang, F. Jia, W. Mao, Y. Liu, Y. Zhao, Z. Chen, T. Wang, C. Zhang, X. Zhang, and F. Zhao, "Stream query denoising for vectorized hd-map construction," in *European Conference on Computer Vision*. Springer, 2024, pp. 203–220.

[29] Z. Liu, X. Zhang, G. Liu, J. Zhao, and N. Xu, "Leveraging enhanced queries of point sets for vectorized map construction," in *European Conference on Computer Vision*. Springer, 2024, pp. 461–477.

[30] Y. Zhou, H. Zhang, J. Yu, Y. Yang, S. Jung, S.-I. Park, and B. Yoo, "Himap: Hybrid representation learning for end-to-end vectorized hd map construction," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 15 396–15 406.

[31] X. Zhang, G. Liu, Z. Liu, N. Xu, Y. Liu, and J. Zhao, "Enhancing vectorized map perception with historical rasterized maps," in *European Conference on Computer Vision*. Springer, 2024, pp. 422–439.

[32] B. Wilson, W. Qi, T. Agarwal, J. Lambert, J. Singh, S. Khandelwal, B. Pan, R. Kumar, A. Hartnett, J. K. Pontes *et al.*, "Argoverse 2: Next generation datasets for self-driving perception and forecasting," *arXiv preprint arXiv:2301.00493*, 2023.