CDSegNet++: Enhancing Diffusion-Based Point Cloud Segmentation with Adaptive Scheduling and Consistency Fusion

Dongdong An[1], Mingyu Shou[1], Shuo Zhang[2*], Wenbing Tang[3]
[1]College of Information, Mechanical and Electrical Engineering, Shanghai Normal University, Shanghai, China
[2]School of Electronic Information Engineering, Shanghai DianJi University, Shanghai, China
[3]Northwest A&F University, Yangling, China
Email: {andongdong, 1000551217}@shnu.edu.cn, szhang@sdju.edu.cn

***Abstract*—Point cloud semantic segmentation is challenged by irregular sampling, sparsity, and ambiguous object boundaries caused by sparse point distributions and geometric discontinuities. Diffusion-based segmentors offer noise robustness, but existing designs suffer from (i) fixed noise schedules, (ii) single-stream boundary blurring, and (iii) indiscriminate feature fusion. We propose an enhanced Conditional-Noise Framework with three components. Adaptive Diffusion Scheduling Strategy (ADSS) adjusts noise level conditioned on scene complexity, improving denoising stability across easy and hard regions. Dual-Stream Semantic Consistency (DSSC) enforces auxiliary prediction consistency to sharpen object boundaries. Lightweight Adaptive Fusion Module (LAFM) employs gated fusion to suppress noisy cues. Experiments on ScanNet and SemanticKITTI show consistent gains: 67.50% mIoU (+4.23%) and 41.94% mIoU (+1.47%), while retaining single-step inference.**

***Index Terms*—Point Cloud Semantic Segmentation, Denoising Diffusion Probabilistic Models, Adaptive Learning, Consistency Training, Feature Fusion.**

I. Introduction

Point cloud semantic segmentation (PCSS) is fundamental for applications such as autonomous driving, robotic modeling, and augmented reality [1–3], yet it remains challenging due to noise, sparsity, and ambiguous object boundaries in point clouds. Deep learning methods have achieved impressive results [4], but most are sensitive to real-world perturbations.

Denoising Diffusion Probabilistic Models (DDPMs) offer inherent robustness through their noise modeling capability [5–7]. Recently, [8] introduced the Conditional-Noise Framework (CNF) to PCSS for the first time, constructing CDSegNet, achieving single-step inference and robustness to noise. However, we identify three critical limitations that hinder its performance in complex and varied real-world scenes:

(i) static noise schedules lack scene adaptability, causing over-perturbation in simple scenes and under-augmentation in complex ones; (ii) the single conditional network struggles to preserve fine details, as perturbed features easily obscure boundaries and small objects; (iii) the feature fusion mechanism is non-selective, relying on standard cross-attention that may introduce irrelevant noise and fail to prioritize semantically critical correspondences. To address aforementioned issues, we present an enhancement to the existing CDSegNet framework through introducing three novel modules:

Adaptive Diffusion Scheduling Strategy (ADSS), a scene-aware curriculum that adjusts noise strength and timesteps based on complexity and training phase; Dual-Stream Semantic Consistency (DSSC), a lightweight auxiliary head enforcing consistency at prediction and feature levels to improve boundary clarity and small-object recognition; Lightweight Adaptive Fusion Module (LAFM), a gated fusion mechanism with learnable residual weights for selective cross-modal interaction.

Extensive experiments on the indoor (ScanNet [9]) and outdoor (SemanticKITTI [10]) benchmarks demonstrate that our model outperforms the original CDSegNet and other methods, validating the effectiveness of our proposed modules.

*Corresponding author. This work was supported by the National Natural Science Foundation of China (NSFC) under Grant 62302308.

II. Related Work

A. Point Cloud Semantic Segmentation

Deep learning has advanced PCSS from point-based [11], voxel-based [12] to transformer-based architectures [13, 14]. Point Transformer V3 (PTv3) has achieved strong performance but overlooks real-world noise and sparsity. This motivates integrating robust diffusion models into segmentation.

B. Diffusion Models for 3D Vision

DDPMs excel in point cloud generation [15], completion [16], and upsampling [17]. Recently, CDSegNet [8] has enabled single-step segmentation via CNF, but it uses fixed noise schedules and standard cross-attention. Our work enhances CDSegNet with adaptive scheduling and selective fusion.

C. Adaptive Training Mechanism

Curriculum learning inspires dynamic adjustment of augmentation and modulate regularization parameters [18]. In 3D vision, scene-aware strategies leverage context to guide tasks like text-to-3D generation [19] and motion prediction [20]. However, adapting diffusion noise by scene complexity for segmentation is unexplored. Our ADSS fills this gap.

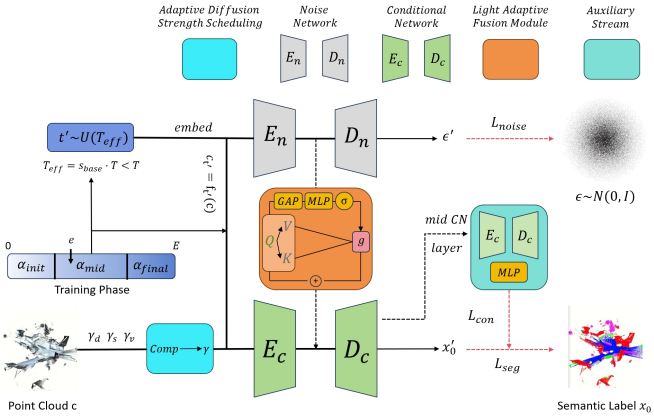

Fig. 1: Framework of CDSegNet++.

D. Consistency Learning and Knowledge Distillation

Dual-stream consistency improves multi-modal fusion, diffusion sampling and training [21, 22]. Yet enforcing consistency between conditional features and diffusion output for segmentation is under-explored. Our DSSC provides lightweight auxiliary supervision focusing on boundaries.

E. Efficient Attention Mechanisms

Standard attention inherently has prohibitive quadratic complexity [23]. Linear [24] and gated [25] variants reduce cost. In point clouds, managing attention is crucial. Our LAFM adopts adaptive gating to selectively emphasize relevant interactions, achieving a better balance between performance and efficiency.

III. Methodology

A. Framework Overview

Our approach builds upon the **Conditional-Noise Framework (CNF)** proposed by CDSegNet [8] and enhances it with three key modifications. As illustrated in Fig.1, the overall pipeline proceeds as follows:

First, The input point cloud $\mathbf{c}$ is fed simultaneously into the Noise Network (NN) and the Conditional Network (CN). Guided by ADSS, the NN performs adaptive noise addition and denoising learning on the conditional information, generating noise features rich in perturbation. The CN, serving as the segmentation backbone, extracts deep semantic features from the point cloud. At the bottleneck stage, an improved Feature Fusion Module (FFM)—specifically its core unit, the Lightweight Adaptive Fusion Module (LAFM)—selectively fuses the noise features from the NN into the semantic features of the CN. Finally, the CN outputs the primary semantic prediction $\mathbf{P}_{main}$. Additionally, a lightweight auxiliary head (DSSC) receives intermediate features from the CN and the original coordinates, generating an auxiliary prediction $\mathbf{P}_{aux}$, which is constrained for consistency with the main prediction during training and can be optionally fused during inference.

Our overall objective combines diffusion loss, segmentation loss, and consistency loss:

$$\mathcal{L}_{total} = \mathcal{L}_{noise} + \lambda_{seg}\mathcal{L}_{seg} + \lambda_{con}\mathcal{L}_{con} \tag{1}$$

where λ_{seg} and λ_{con} are both scalar balancing weights.

B. Conditional-Noise Framework Revisited

Following CDSegNet [8] , we employ CNF where NN serves as an auxiliary feature augmentation branch, while CN acts as a dominant segmentation backbone. NN predicts added noise ϵ from noisy condition $\mathbf{c_t}$ based on time step t:

$$\mathcal{L}_{noise}(\theta) = \mathbb{E}_{\boldsymbol{\epsilon}\sim\mathcal{N}(0,I)}\|\boldsymbol{\epsilon} - \epsilon_\theta(\boldsymbol{c_t}, t)\|^2 \tag{2}$$

where $\boldsymbol{c_t} = f_t(\boldsymbol{c}) = \sqrt{\bar{\alpha}_t}\boldsymbol{c} + \sqrt{1-\bar{\alpha}_t}\epsilon$, $\bar{\alpha}_t$ is determined by a predefined noise schedule $\{\beta_t\}$ through the cumulative product formula $\bar{\alpha}_t = \prod_{s=1}^{t}\alpha_s$ and $\alpha_s = 1 - \beta_s$ respectively.

The CN takes the original point cloud $\mathbf{c}$ as input to predict semantic labels $\mathbf{x_0}$. Its loss is a standard combination loss:

$$\mathcal{L}_{seg}(\psi) = \mathcal{L}_{CE}(\mathbf{x}_0, \mathbf{x}_0') + \mathcal{L}_{Lovasz}(\mathbf{x}_0, \mathbf{x}_0') \tag{3}$$

where $\mathcal{L}_{CE}$ is the cross-entropy loss and $\mathcal{L}_{Lovasz}$ is the Lovasz-Softmax loss [26] specifically for optimizing IoU.

C. Adaptive Diffusion Strength Scheduling

ADSS dynamically adjusts noise intensity based on scene complexity and training progress during the diffusion. It consists of two phases: offline estimation and online scheduling.

1) Scene Complexity Estimation: Given an input point cloud $\mathbf{c} \in \mathbb{R}^{N\times 3}$, we compute a scalar complexity score $\gamma \in [0.1, 1.0]$ that synthesizes three lightweight metrics:

Point Density Feature:

$$\gamma_d = min(\frac{N}{5000}, 1.0) \tag{4}$$

where N is the absolute number of points in the scene.

Spatial Distribution Feature:

$$\gamma_s = min(\frac{\|std(\mathbf{c})\|_2}{3.0}, 1.0) \tag{5}$$

where $std(\cdot)$ computes the standard deviation across spatial coordinates, reflecting point dispersion across the entire scene.

Volume Feature:

$$\gamma_v = min(\frac{\log(vol(\mathbf{c}) + 1)}{5.0}, 1.0) \tag{6}$$

where $vol(\mathbf{c})$ is the the axis-aligned bounding box volume, reflecting the spatial scale of the scene along all three axes.

The normalization constants (5000, 3.0, 5.0) are empirically determined from training set statistics. The final complexity score is:

$$\gamma = \frac{1}{3}(\gamma_d + \gamma_s + \gamma_v) \tag{7}$$

A higher value indicates a more complex scene and vice versa.

2) Adaptive Noise Scheduling: At training epoch e (out of total epochs E), we first determine a base strength factor s_{base}:

$$s_{base} = \begin{cases} \alpha_{init}, & if\ e/E < 0.3\ (Initial\ Phase) \\ \alpha_{mid}, & if\ 0.3 \leq e/E < 0.7\ (Middle\ Phase) \\ \alpha_{final}, & otherwise\ (Final\ Phase) \end{cases} \tag{8}$$

where α_{init}, α_{mid} and α_{final} are configurable parameters (e.g., 0.3, 0.6, 0.8), forming a curriculum learning plan.

Next, adjustment based on scene complexity: Simpler scenes receive increased diffusion strength for diverse perturbations; Complex scenes receive decreased strength to avoid corrupting the semantics. The adjusted strength factor s is:

$$s = s_{base} \cdot (1 + (1 - \gamma) \cdot \rho) \tag{9}$$

where ρ is a scaling factor controlling the influence of complexity. We clamp s to a reasonable range (e.g., $s \in [0.1, 1.5]$).

3) Parameter Recalculation and Sampling: Using s, we adjust noise schedule parameters: $\beta_t' = \beta_t \cdot s$ (for linear schedule, this scales the start and end points). We then recompute α_t' and $\bar{\alpha}_t'$. Concurrently, effective timesteps are adjusted altogether:

$$T_{eff} = \lfloor T \cdot min(1.0, s) \rfloor \tag{10}$$

During training, we sample timestep t from $[0, T_{eff}]$ (instead of $[0, T]$) and use the adjusted $\bar{\alpha}_t'$ for forward diffusion. This enables the NN to learn a denoising task, whose intensity dynamically matches both the scene and training phase.

D. Lightweight Adaptive Fusion Module

Original Transfer Module within FFM employs non-selective cross-attention, treating all feature interactions equally, and potentially propagating irrelevant noise cues. To address this, we propose LAFM that supersedes the Transfer Module while preserving its CrossBlock for bidirectional feature interaction (using either standard cross-attention or the efficient Serialized Cross Restomer). LAFM introduces three adaptive components for selective and stable fusion.

1) Feature Normalization for Stabilization: Before fusion, LayerNorm is applied separately and respectively to the semantic features $\mathbf{F}_s$ from CN and the noise features $\mathbf{F}_n$ from NN to ensure numerical stability and scale uniformity:

$$\tilde{\mathbf{F}}_s = LayerNorm(\mathbf{F}_s), \tilde{\mathbf{F}}_n = LayerNorm(\mathbf{F}_n) \tag{11}$$

2) Adaptive Gating: Not all noise features are equally beneficial—static fusion weights cannot adapt to varying scene complexities. We design a lightweight gating network that generates a scalar gating value $g \in (0, 1)$ based on global context of the CN features (obtained via global average pooling), adaptively modulating noise feature contribution:

$$\mathbf{g}_{context} = GAP(\tilde{\mathbf{F}}_s), g = \sigma(MLP_g(\mathbf{g}_{context})) \tag{12}$$

where σ is the Sigmoid function. This gate adaptively controls the amount of information flow from the noise features.

3) Feature Transformation and Residual Fusion: Direct addition of noise features may disrupt semantic representations. We first apply a lightweight nonlinear transformation to the normalized semantic features: $\mathbf{F}_s^{trans} = MLP_{trans}(\tilde{\mathbf{F}}_s)$.

Then, fuse it with projection-aligned noise features via a carefully controlled learnable residual weight λ_{res}:

$$\mathbf{F}_{enhanced} = \mathbf{F}_s + \lambda_{res} \cdot g \cdot (\mathbf{F}_s^{trans} + Proj(\tilde{\mathbf{F}}_n)) \tag{13}$$

where $Proj(\cdot)$ projects linearly for dimension matching, and λ_{res} is initialized to 0.1 and constrained to [0, 0.5].

This design allows the fusion process to self-adjust based on the input features, enhancing beneficial interactions while suppressing harmful noise, with low computational overhead.

E. Dual-Stream Semantic Consistency

The DSSC module introduces a parallel auxiliary semantic segmentation stream to provide additional supervision signals to the main network, focused on improving boundary regions.

1) Auxiliary Stream Structure: The auxiliary stream is a lightweight Multi-Layer Perceptron (MLP). It takes the concatenation of point cloud coordinates $\mathbf{c}$ and intermediate features $\mathbf{F}_s^{(mid)}$ from a middle layer of the CN as input:

$$\begin{aligned} \mathbf{F}_{aux} &= MLP_{aux}(concat(\mathbf{c}, \mathbf{F}_s^{(mid)})) \\ \mathbf{P}_{aux} &= Linear(\mathbf{F}_{aux}) \end{aligned} \tag{14}$$

2) Consistency Constraints Losses: During training, we impose three distinct and complementary losses in parallel to jointly constrain both the main and auxiliary branches:

Auxiliary Segmentation Loss, which directly supervises the auxiliary prediction with corresponding ground-truth labels:

$$\mathcal{L}_{aux} = \mathcal{L}_{CE}(\mathbf{P}_{aux}, \mathbf{x}_0) \tag{15}$$

Prediction Consistency Loss, which employs symmetric KL-divergence to bring output distributions of the main prediction $\mathbf{P}_{main}$ and the auxiliary prediction $\mathbf{P}_{aux}$ closer:

$$\mathcal{L}_{pred} = \frac{1}{2}(KL(\mathbf{P}_{main}||\mathbf{P}_{aux}) + KL(\mathbf{P}_{aux}||\mathbf{P}_{main})) \tag{16}$$

Feature Consistency Loss, which constrains the distance between the main and auxiliary branches in feature space:

$$\mathcal{L}_{feat} = \|\phi(\mathbf{F}_s) - \phi(\mathbf{F}_{aux})\|_2^2 \tag{17}$$

where ϕ is a shared projection layer across both branches.

The consistency loss sums three weighted terms:

$$\mathcal{L}_{con} = \lambda_{aux}\mathcal{L}_{aux} + \lambda_{pred}\mathcal{L}_{pred} + \lambda_{feat}\mathcal{L}_{feat} \tag{18}$$

3) Adaptive Fusion for Inference (Optional): During inference, we can optionally fuse the auxiliary prediction. If enabled, we apply temperature scaling to soften its distribution:

$$\mathbf{P}_{aux}^T = Softmax(\mathbf{P}_{aux}/\tau) \tag{19}$$

with $\tau = 2.0$, fused linearly with the main prediction:

$$\mathbf{P}_{final} = (1 - \lambda_{fuse}) \cdot \mathbf{P}_{main} + \lambda_{fuse} \cdot \mathbf{P}_{aux}^T \tag{20}$$

where $\lambda_{fuse} = 0.2$ controls the fusion, This operation can further improve the smoothness and accuracy of the prediction.

TABLE I: Best results on ScanNet and SemanticKITTI. Our method outperforms all competing approaches across all metric on ScanNet, and achieves the best mIoU among all compared methods on the more challenging SemanticKITTI.

Methods	ScanNet			SemanticKITTI		
	mIoU	mAcc	allAcc	mIoU	mAcc	allAcc
SparseUNet [28]	0.4880	0.5711	0.8285	0.4044	0.4996	0.8709
PTv1 [13]	0.4901	0.5771	0.8404	0.3257	0.3797	0.8344
OACNNs [29]	0.5385	0.6356	0.8468	0.3918	0.4849	0.8729
PTv2 [30]	0.5411	0.6354	0.8529	0.3991	**0.5105**	0.8605
PTv3 [4]	0.6000	0.7184	0.8540	0.4097	0.5026	0.8707
Baseline	0.6327	0.7428	0.8660	0.4047	0.4896	**0.8761**
Ours	**0.6750**	**0.7737**	**0.8862**	**0.4194**	0.5045	0.8749

TABLE II: Ablation study of CDSegNet with each our innovation module ADSS, DSSC and LAFM. Baseline is the original CDSegNet [8]. It shows validity of each single modules and their different combination using multiple modules.

Methods	ScanNet			SemanticKITTI		
	mIoU	mAcc	allAcc	mIoU	mAcc	allAcc
Baseline	0.6327	0.7395	0.8663	0.4047	0.4896	0.8761
ADSS	0.6667	0.7684	0.8816	0.4148	0.4955	0.8806
DSSC	0.6582	0.7581	0.8793	**0.4194**	**0.5045**	0.8749
ADSS+DSSC	0.6568	0.7641	0.8792	0.4115	0.4957	0.8793
LAFM	0.6330	0.7527	0.8650	0.4063	0.4894	0.8734
ADSS+LAFM	0.6635	0.7695	0.8812	0.4185	0.4933	**0.8821**
DSSC+LAFM	0.6626	0.7674	0.8797	0.4005	0.4828	0.8745
ADSS+DSSC+LAFM	**0.6750**	**0.7737**	**0.8862**	0.4005	0.4803	0.8775

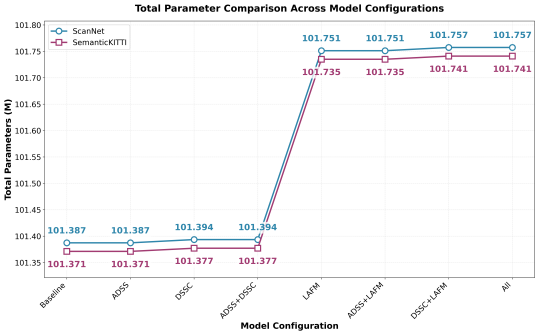

Fig. 2: Total parameters of combination of different modules

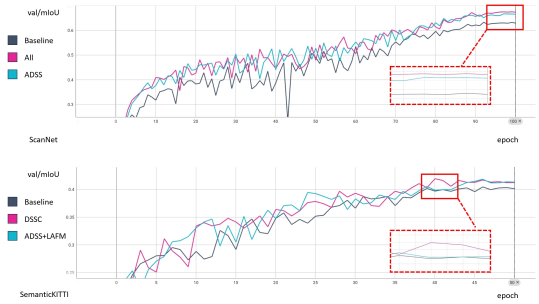

Fig. 3: mIoU of validation on ScanNet and SemanticKITTI.

F. Training and Inference

Training: Inspired by [27], we employ Geometric Loss Strategy (GLS) to balance gradient scales among three tasks: noise fitting, semantic segmentation, and consistency constraint. The GLS used for training is as follows:

$$\mathcal{L}_{GLS} = \sqrt[3]{\mathcal{L}_{noise} \cdot \mathcal{L}_{seg} \cdot \mathcal{L}_{con}} \tag{21}$$

This strategy helps stabilize the joint multi-task training.

Inference: Benefiting from CNF, our model requires only single-step deterministic forward propagation during inference. The process is as follows: Pure Gaussian noise $\mathbf{c}_T \sim \mathcal{N}(0, I)$ is fed into NN to obtain noise features. CN processes the original point cloud to obtain semantic features. LAFM fuse NN and CN. Finally, CN outputs the main prediction, which can be optionally fused with the auxiliary head's prediction to yield final segmentation result $\mathbf{x}'_0$.

IV. Experiment

A. Experiment Setup

Baseline: Original CDSegNet [8] with PTV3 backbone [4]. If our proposed modules (ADSS, DSSC, LAFM) are all disabled, our model reduces exactly to the baseline.

Dataset: ScanNet (indoor, 20 classes, 1201/312/100 train/val/test scenes, 0.04m voxel) and SemanticKITTI subset (outdoor, 19 classes, 2392/4071/1752 scenes, 0.08m voxel).

B. Segmentation Results and Comparison

1) Results of Indoor Scenes: Table I shows that our method achieves 67.50% mIoU, 77.37% mAcc, and 88.62% allAcc on ScanNet, surpassing all competitors. Compared to baseline, gains are +4.23% mIoU, +3.09% mAcc, and +2.02% allAcc improvements. Against PTv3, mIoU advantage reaches +7.50%. The improvements are consistent across all metrics, indicating that ADSS prevents over-perturbation in simple regions and under-augmentation in complex ones, while DSSC's auxiliary supervision refines fine-grained boundaries.

2) Results of Outdoor Scenes: On SemanticKITTI (Table I), our method achieves 41.94% mIoU (+1.47%), 50.45% mAcc (+1.49%), and 87.49% allAcc, outperforming all compared methods in mIoU. Notably, PTv2 has highest mAcc (51.05%), baseline leads in allAcc (87.61%), while our method leads in the class-balanced mIoU, benefiting from DSSC's auxiliary supervision that strengthens recognition of sparse small objects. ADSS's late-stage perturbation enhances generalization in sparse scenes, while DSSC's coordinate-aware stream preserves structural continuity of distant vegetation.

C. Ablation Study

1) Individual Module Contributions: Table II shows: ADSS yields largest gain on ScanNet (+3.40% mIoU), as indoor scenes have high complexity variation. DSSC gains more on SemanticKITTI (+1.47% mIoU) with sparse small objects. LAFM gives modest gains (ScanNet +0.03%, SemanticKITTI +0.16%) and serves as a stable fusion foundation.

2) Combination Effects and Parameter Efficiency: Fig.2 shows minimal parameter overhead: ADSS adds zero. DSSC adds 6K∼7K (on ScanNet and SemanticKITTI) and LAFM adds 0.364M. Full combination adds only 0.37M over baseline.

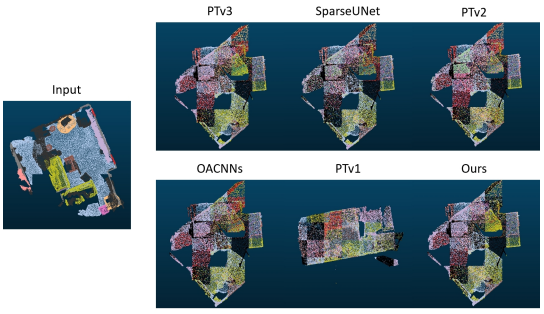

Fig. 4: Scene 0203-01 of visualization result of ScanNet

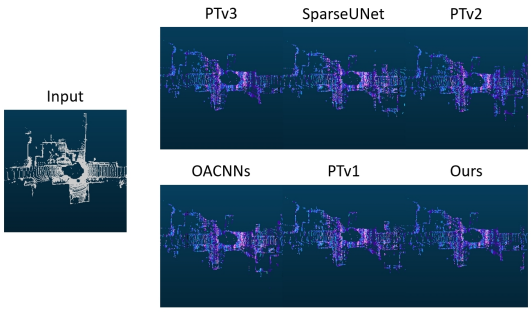

Fig. 5: Scene 11-300 of visualization result of SemanticKITTI

Notably, non-additive behavior appears: Full combination achieves best on ScanNet (67.50% mIoU) but underperforms DSSC alone on SemanticKITTI (40.05% mIoU vs 41.94%). This indicates tension between ADSS (which increases perturbation in late training) and DSSC (which stabilizes features) in sparse outdoor scenes. LAFM mitigates this: LAFM+ADSS outperforms ADSS alone, suggesting that the gating network successfully suppresses harmful dynamics.

3) Training Dynamics and Convergence Characteristics: Fig.3 reveals: (i) ADSS accelerates convergence—plateau at epoch 70 on ScanNet (80 for baseline), and reaches 40% mIoU within 40 epochs on SemanticKITTI. (ii) DSSC boosts late-stage performance—sustained improvements after 70 epochs on ScanNet and 35 epochs on SemanticKITTI. (iii) LAFM improves stability—combinations with LAFM exhibit smaller fluctuations during mid-training (epoch 30 to 70 on ScanNet, 15 to 35 on SemanticKITTI). The full combination stabilizes after epoch 90 on ScanNet but degrades after 40 epochs on SemanticKITTI (41.94% to 40.58%), consistent with Table II.

4) Qualitative Results: Fig.4 and Fig.5 show that our method reduces misclassified points. On ScanNet, boundaries become cleaner and objects are better segmented. On SemanticKITTI, structural continuity of sparse objects is preserved. It confirms synergistic gains in complex indoor scenes and the need for careful balance in sparse outdoor scenes.

D. Discussion

Our experiments reveal three key phenomena: (i) Scene dependency: ADSS gains +3.40% indoor with high complexity variation, while DSSC gains +1.47% outdoor. (ii) Non-additive combinations: Full combination is optimal on ScanNet but underperforms on SemanticKITTI, indicating tension between dynamic perturbation and stable supervision in sparse scene. (iii) Parameter efficiency: Core gains (ADSS+DSSC) cost only 7K additional parameters. LAFM adds 0.36M for stability.

These findings highlight a fundamental trade-off: ADSS reduces bias via adaptive augmentation, DSSC controls variance via regularization. Future work could design a learnable scheduler to dynamically balance perturbation strength and consistency weight based on real-time feature entropy, especially for outdoor scenes.

V. Conclusion

In this paper, we presented CDSegNet++, an enhanced diffusion-based point cloud semantic segmentation framework, addressing three limitations of the existing Conditional-Noise Framework: (i) fixed noise schedules lacking scene adaptability, (ii) single prediction streams prone to boundary blurring, and (iii) non-selective feature fusion. By introducing ADSS, DSSC and LAFM, our approach achieves strong performance on both indoor and outdoor benchmarks while preserving single-step inference. Ablation studies reveal non-additive module interactions, indicating tension between adaptive perturbation and consistency regularization in sparse outdoor scenes. Future work will explore dynamic balancing mechanisms to better coordinate these modules.

References

[1] S. Zhang *et al.*, "Multi-modal salient object detection via a unified diffusion model," in *ICASSP 2025-2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2025, pp. 1–5.

[2] S. Zhang, J. Huang *et al.*, "Dimsod: A diffusion-based framework for multi-modal salient object detection," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 39, no. 10, 2025, pp. 10 103–10 111.

[3] S. Zhang *et al.*, "Sod-diffusion: Salient object detection via diffusion-based image generators," in *Computer Graphics Forum*, vol. 43, no. 7. Wiley Online Library, 2024, p. e15251.

[4] X. Wu *et al.*, "Point transformer v3: Simpler faster stronger," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2024, pp. 4840–4851.

[5] S. Zhang *et al.*, "Tranbf: Deep transformer networks and bayesian filtering for time series anomalous signal detection in cyber-physical systems," in *2024 IEEE International Conference on Multimedia and Expo (ICME)*. IEEE, 2024, pp. 1–6.

[6] S. Zhang, J. Huang *et al.*, "Simplediffusion: A lightweight and efficient conditional diffusion model for multi-modal salient object detection," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 40, no. 15, 2026, pp. 12 753–12 761.

[7] S. Zhang *et al.*, "Seg-diffusion: Text-to-image diffusion model for open-vocabulary semantic segmentation,"

in *ICASSP 2025-2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2025, pp. 1–5.

[8] W. Qu *et al.*, "An end-to-end robust point cloud semantic segmentation network with single-step conditional diffusion models," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2025, pp. 27 325–27 335.

[9] A. Dai *et al.*, "Scannet: Richly-annotated 3d reconstructions of indoor scenes," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, July 2017.

[10] J. Behley *et al.*, "Semantickitti: A dataset for semantic scene understanding of lidar sequences," in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, October 2019.

[11] C. R. Qi *et al.*, "Pointnet: Deep learning on point sets for 3d classification and segmentation," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, July 2017.

[12] C. Choy *et al.*, "4d spatio-temporal convnets: Minkowski convolutional neural networks," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019.

[13] H. Zhao *et al.*, "Point transformer," in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, October 2021, pp. 16 259–16 268.

[14] J. Jiang *et al.*, "Deep vision transformer with association divergence for image anomaly detection and localization," in *International Conference on Neural Information Processing*. Springer, 2024, pp. 1–15.

[15] S. Luo *et al.*, "Diffusion probabilistic models for 3d point cloud generation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2021, pp. 2837–2845.

[16] Z. Lyu *et al.*, "A conditional point diffusion-refinement paradigm for 3d point cloud completion," 2022. [Online]. Available: https://arxiv.org/abs/2112.03530

[17] W. Qu *et al.*, "A conditional denoising diffusion probabilistic model for point cloud upsampling," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2024, pp. 20 786–20 795.

[18] K. Cheng *et al.*, "Diff-moe: Diffusion transformer with time-aware and space-adaptive experts," in *Forty-second International Conference on Machine Learning*, 2025. [Online]. Available: https://openreview.net/forum?id=JCUsWrwkKw

[19] Y. Yang *et al.*, "Prometheus: 3d-aware latent diffusion models for feed-forward text-to-3d scene generation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2025, pp. 2857–2869.

[20] J. Tang *et al.*, "A unified diffusion framework for scene-aware human motion estimation from sparse signals," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2024, pp. 21 251–21 262.

[21] S. Zhang *et al.*, "Anomalous signal detection for cyber-physical systems using interpretable causal neural network," in *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2023, pp. 1–5.

[22] S. Zhang, X. Hu *et al.*, "Causal fusion of convolutional neural network and vision transformer for image anomaly detection and localization," in *2024 IEEE International Conference on Multimedia and Expo (ICME)*. IEEE, 2024, pp. 1–6.

[23] S. Zhang *et al.*, "Feature-constrained and attention-conditioned distillation learning for visual anomaly detection," in *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2024, pp. 2945–2949.

[24] A. Katharopoulos *et al.*, "Transformers are RNNs: Fast autoregressive transformers with linear attention," in *Proceedings of the 37th International Conference on Machine Learning*, ser. Proceedings of Machine Learning Research, H. D. III and A. Singh, Eds., vol. 119. PMLR, 13–18 Jul 2020, pp. 5156–5165. [Online]. Available: https://proceedings.mlr.press/v119/katharopoulos20a.html

[25] B. Dhingra *et al.*, "Gated-attention readers for text comprehension," in *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, R. Barzilay and M.-Y. Kan, Eds. Vancouver, Canada: Association for Computational Linguistics, Jul. 2017, pp. 1832–1846. [Online]. Available: https://aclanthology.org/P17-1168/

[26] M. Berman *et al.*, "The lovasz-softmax loss: A tractable surrogate for the optimization of the intersection-over-union measure in neural networks," in *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2018, pp. 4413–4421.

[27] S. Chennupati *et al.*, "Multinet++: Multi-stream feature aggregation and geometric loss strategy for multi-task learning," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, June 2019.

[28] B. Graham *et al.*, "3d semantic segmentation with submanifold sparse convolutional networks," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018.

[29] B. Peng *et al.*, "Oa-cnns: Omni-adaptive sparse cnns for 3d semantic segmentation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2024, pp. 21 305–21 315.

[30] X. Wu *et al.*, "Point transformer v2: Grouped vector attention and partition-based pooling," in *Advances in Neural Information Processing Systems*, S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh, Eds., vol. 35. Curran Associates, Inc., 2022, pp. 33 330–33 342.