# SAM-GUIDED MULTI-VIEW FUSION FOR WEAKLY SUPERVISED 3D POINT CLOUD SEGMENTATION

*Yuena Qiao*[1], *Nanqing Liu*[2†], *Yongyi Su*[3], *Shijie Li*[3], *Xulei Yang*[3], *Bihan Wen*[4], *Nancy Chen*[3], *Tianrui Li*[1], *Xun Xu*[3]

[1] School of Computing and Artificial Intelligence, Southwest Jiaotong University, China
[2] Yunnan Normal University, China [3] Institute for Infocomm Research (I$^2$R), A*STAR, Singapore
[4] Nanyang Technological University, Singapore

## ABSTRACT

Point cloud semantic segmentation is fundamental for 3D scene understanding but requires costly dense annotations, limiting scalability in real-world applications. Weakly supervised point cloud semantic segmentation (WSPCSS) reduces labeling effort by using sparse annotations, yet existing methods struggle to learn comprehensive semantic representations under such constraints. In this paper, we propose a novel framework that augments weak 3D supervision with external cues from the 2D vision domain. Specifically, we leverage the Segment Anything Model (SAM) in a multi-view setting, where the point cloud is rendered into multiple 2D views, segmented by SAM, and back-projected into 3D space. To resolve inconsistencies across views, we introduce a hypergraph-based label propagation strategy that explicitly models high-order relationships between 3D points and 2D masks, enabling robust fusion of multi-view cues. A confidence-based filtering mechanism further enhances pseudo-label reliability. Extensive experiments on the S3DIS and ScanNet V2 benchmarks demonstrate that our approach outperforms existing weakly supervised methods, narrowing the gap to fully supervised performance under extremely sparse labels. Our method establishes a new paradigm for transferring 2D foundation model priors to 3D tasks, providing a scalable solution for large-scale 3D segmentation with minimal labeling cost.

***Index Terms*—** 3D Point Cloud Segmentation, Weakly Supervised Learning, Segment Anything Model, Label Propagation.

## 1. INTRODUCTION

Point cloud semantic segmentation is a fundamental task in 3D vision, with applications in robotics, autonomous driving, and augmented reality [1, 2]. Most existing methods rely on fully supervised learning with dense per-point annotations [3, 4, 5], which are costly and labor-intensive, especially for large-scale scenes such as indoor environments [6, 7, 8] or urban street views. To reduce annotation effort, weakly supervised point cloud semantic segmentation (WSPCSS) has emerged, using sparse or partial labels. Existing approaches mainly follow two strategies: (1) maximizing the utility of limited labels via point-level regularization and constraints [9, 10, 11, 12, 13]; and (2) leveraging unlabeled data through self-supervised or contrastive learning [14, 15, 16, 17]. While effective, these methods remain limited by weak supervision, often producing representations that lack sufficient semantic richness for real-world applications.

A natural question then arises: *can we enrich 3D learning with external cues beyond point-level annotations?* In this work, we explore the transfer of segmentation priors from the 2D vision domain to enhance weakly supervised 3D segmentation. Specifically, we leverage the Segment Anything Model (SAM) [18], a powerful foundation model trained on billions of masks, which exhibits strong generalization across diverse image domains[19, 20]. Our key idea is to harness SAM's high-quality 2D masks as auxiliary cues for 3D point clouds. To achieve this, we introduce a multi-view framework that projects a 3D point cloud into multiple rendered images and applies SAM to each view, generating diverse yet complementary segmentation proposals.

However, directly fusing multi-view 2D masks into 3D space is non-trivial. The masks from different views are not aligned in index or semantics, making naïve merging unreliable. To overcome this, we formulate the problem as hypergraph label propagation, where each 3D point is modeled as a vertex, and each 2D SAM mask serves as a hyperedge connecting points with shared 2D segmentation context. This hypergraph representation enables consistent propagation of semantic information across views, ultimately yielding robust and coherent 3D pseudo-labels. Our approach offers a new paradigm for WSPCSS: injecting rich 2D segmentation priors into 3D learning to compensate for limited supervision. Empirical results on challenging real-world datasets show that our method achieves significant improvements over prior WSPCSS baselines. Moreover, ablation studies confirm the effectiveness of both multi-view fusion and SAM-derived pseudo-labels in guiding 3D semantic understanding.

The main contributions of this paper are summarized as follows:

- We propose a novel paradigm for WSPCSS that transfers rich segmentation priors from the 2D Segment Anything Model (SAM) to the 3D domain.
- We introduce a hypergraph-based label propagation framework that effectively fuses multi-view SAM segmentations into high-quality 3D pseudo-labels, enabling robust supervision under sparse annotations.
- Extensive experiments on large-scale real-world datasets demonstrate that our method consistently outperforms state-of-the-art weakly supervised baselines.

## 2. METHODOLOGY

### 2.1. Overview

To formally define the WSPCSS task, we follow the settings proposed in [9]. In specific, a training dataset $D_{tr} = \{X_i, Y_i, M_i\}_{i=1\cdots N_{tr}}$ is provided, where $X_i \in \mathbb{R}^{D_i \times N}$ are the $N$ input points each with $D_i$ dimension feature, e.g. 3D coordinates with RGB color if available, $Y_i \in \{0,1\}^{K \times N}$ is the one-hot per-point segmentation label ($K$ categories) and $M_i \in \{0,1\}^N$ is a binary mask indicating whether ground-truth label is available. An encoder network

† Correspondence to: Nanqing Liu <lansing163@163.com>.

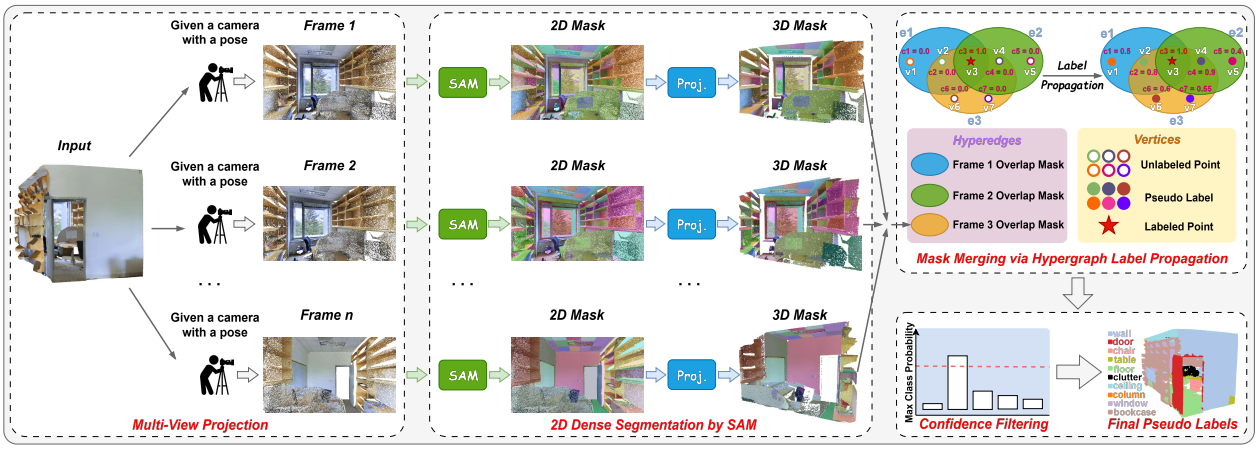


**Fig. 1**: Overview of the proposed SAM-based 3D hypergraph label propagation method.

$Z = f(X;\Theta)$ maps input points into a $D_o$ dimension feature space, $Z \in \mathbb{R}^{D_o \times N}$. A classifier $h(Z;\Omega) \in \mathbb{R}^{K \times N}$ maps encoded features into logits in segmentation category space.

Our goal is to enrich weak supervision with external 2D cues to supplement the limited weak labels. As illustrated in Fig. 1, the framework consists of four main steps: (1) **Multi-View Projection**: render the point cloud into multiple RGB views using virtual cameras; (2) **SAM Segmentation**: apply the Segment Anything Model (SAM) to obtain dense 2D masks; (3) **Back-Projection and Fusion**: map the 2D masks back to 3D points and fuse them using hypergraph label propagation; (4) **Pseudo-Label Training**: filter high-confidence predictions and use them as pseudo-labels to train a standard 3D segmentation network.

### 2.2. Multi-View Projection

To obtain diverse 2D observations from a sparse 3D point cloud $X \in \mathbb{R}^{3\times N}$, we represent each point in homogeneous coordinates as $\mathbf{p}_h = (x, y, z, 1)^\top$. The cloud is rendered from multiple virtual cameras placed uniformly on a sphere centered at the point cloud centroid to ensure broad coverage. Each camera is defined by intrinsic parameters $\mathbf{K} \in \mathbb{R}^{3\times3}$ (focal length and principal point) and extrinsic parameters $(\mathbf{R}, \mathbf{t})$, where $\mathbf{R} \in \mathrm{SO}(3)$ is the rotation matrix and $\mathbf{t} \in \mathbb{R}^3$ the translation vector. A 3D point $\mathbf{p}$ is projected onto the image plane by

$$\mathbf{u} = \mathbf{P}\mathbf{p}_h, \quad \mathbf{P} = \mathbf{K}\,[\mathbf{R} \mid \mathbf{t}] \in \mathbb{R}^{3\times4}. \tag{1}$$

where $\mathbf{u} = (u, v, w)^\top$ are homogeneous image coordinates. The final pixel location is obtained by perspective division $(u/w, v/w)$. The projection produces a set of rendered RGB images $\{I_i\}$ for subsequent SAM segmentation.

### 2.3. 2D Dense Segmentation by SAM

For each rendered image $I_i$, we apply the SAM in dense mode by uniformly sampling point prompts across the image. Each prompt generates one or more candidate masks $\{S_{ij}\}_{j=1}^{N_{Si}}$, resulting in dense and potentially overlapping coverage of the scene. To lift these masks back into 3D, we first obtain the depth $d(u, v)$ for each pixel $(u, v)$ in the rendered image. Since the images are generated from a known 3D point cloud, the depth at each pixel can be directly retrieved from the z-component of the corresponding 3D point in the camera coordinate frame, i.e. $d(u, v) = ([\mathbf{R}|\mathbf{t}]\mathbf{p}_h)_z$. Given the

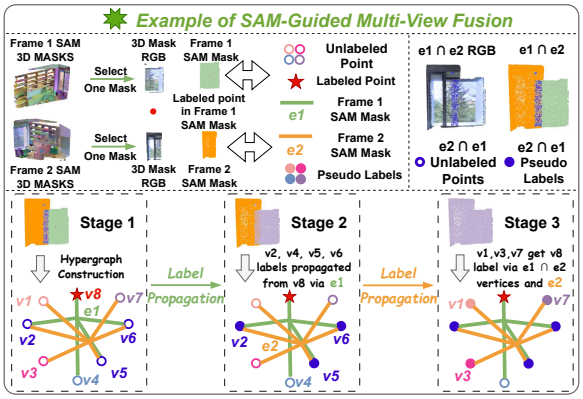


**Fig. 2**: Illustration of SAM-Guided Multi-View Fusion via Hypergraph Label Propagation.

pixel location $(u, v)$ and depth $d$, the 3D coordinate $\mathbf{p} = (x, y, z)^\top$ in the world frame is recovered by back-projection:

$$\mathbf{p} = \mathbf{R}^\top\Big(d \cdot \mathbf{K}^{-1}\begin{bmatrix} u & v & 1 \end{bmatrix}^\top - \mathbf{t}\Big), \tag{2}$$

### 2.4. Mask Merging via Hypergraph Label Propagation

Given multiple segmentation masks produced by SAM on different rendered views, our goal is to fuse these heterogeneous cues into a unified and consistent 3D label representation. This is challenging because masks from different views are not aligned in index or semantics, making direct merging unreliable. To address this, we formulate the problem as a hypergraph-based label propagation task.

**Hypergraph Construction**: Let $\mathbf{p}_k$ denote the $k$-th 3D point. Across multiple views, $\mathbf{p}_k$ may be associated with different 2D masks $\{S_{ij}\}$, but these associations vary by view and mask index. We construct a hypergraph $\mathcal{G} = (V, E)$, where each vertex $v_k \in V$ corresponds to a 3D point, and each hyperedge $e \in E$ represents a SAM-generated 2D mask that connects all 3D points projected inside it.

The hypergraph is encoded by an incidence matrix $\mathbf{H} \in \{0, 1\}^{N\times N_e}$, where $N$ is the number of points and $N_e$ the number of masks. Each entry is defined as

$$H_{k,e} = \begin{cases} 1, & \text{if point } k \in \text{mask } e, \\ 0, & \text{otherwise.} \end{cases} \tag{3}$$

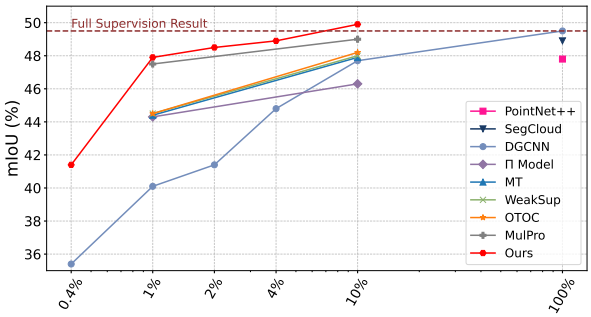


**Fig. 3**: Comparative results of several methods under different label ratios on the S3DIS Area-5.

This representation naturally embeds multi-view segmentation cues into a higher-order structure, allowing semantic information to propagate across points that co-occur in the same SAM masks.

**Label Propagation**: Following [27], we define the degree of a hyperedge $e$ as $d_e = \sum_k H_{k,e}$ and the degree of a vertex $k$ as $d_k = \sum_e H_{k,e}$, collected into diagonal matrices $\mathbf{D}_e = \mathrm{diag}(d_e)$ and $\mathbf{D}_v = \mathrm{diag}(d_k)$. The normalized affinity matrix is

$$\mathbf{S} = \mathbf{D}_v^{-\frac{1}{2}} \mathbf{H} \mathbf{D}_e^{-1} \mathbf{H}^\top \mathbf{D}_v^{-\frac{1}{2}}. \tag{4}$$

Let $\mathbf{Y} \in \{0,1\}^{N \times K}$ be the initial label matrix, where labeled points store one-hot ground-truth vectors and unlabeled points are zero. The iterative propagation rule is

$$\mathbf{F}^{(t+1)} = \alpha \mathbf{S} \mathbf{F}^{(t)} + (1-\alpha)\mathbf{Y}, \tag{5}$$

which converges to the closed-form solution:

$$\mathbf{F}^{\infty} = (1-\alpha)(\mathbf{I} - \alpha \mathbf{S})^{-1} \mathbf{Y}, \tag{6}$$

where $\mathbf{F}^{\infty} \in \mathbb{R}^{N \times K}$ contains the final propagated label distributions. Each row $\mathbf{F}_i$ represents the semantic confidence for point $\mathbf{p}_i$.

Fig. 2 illustrates the process using a window instance appearing in two consecutive frames. Even if 2D SAM-generated masks in different views are independent, their 3D projections overlap, allowing labels to propagate via shared hyperedges. Sparse annotations on a single point can thus spread to overlapping points, which in turn act as semantic bridges to transfer labels across frames. This staged propagation ensures that multi-view 2D semantic cues are fused into coherent 3D pseudo-labels.

**Confidence-based Filtering**: To improve pseudo-label reliability, we assign a label only if the prediction is confident. For each point, if the maximum probability exceeds a threshold $\delta$, the corresponding class is taken as its pseudo-label; otherwise, it is ignored during training:

$$\mathrm{PseudoLabel}(i) = \begin{cases} \arg\max(\mathbf{F}_k^{\infty}), & \text{if } \max(\mathbf{F}_k^{\infty}) \geq \delta, \\ -1, & \text{otherwise.} \end{cases} \tag{7}$$

## 3. EXPERIMENTS

### 3.1. Datasets and Experimental Settings

We evaluate our method on two widely used indoor scene segmentation benchmarks: ScanNet V2 [7] and S3DIS [6]. **ScanNet V2**: This dataset contains 20 semantic classes with 1,201 training scans, 312 validation scans, and 100 test scans. Following common practice, we evaluate our method on the validation set. **S3DIS**: This dataset consists of 272 rooms across six areas, with over 215 million points annotated with XYZ coordinates, RGB colors, and 13 semantic categories. Following prior work [9, 28], we use Area 5 as the test set and train on the remaining areas. For semantic enrichment, we render 46 images per room using Open3D: 36 horizontal views sampled every 10° and 10 vertical views sampled every 36°, ensuring diverse and uniform coverage of the scene. In hypergraph label propagation, we set $\alpha = 0.5$ and adopt DGCNN [23] as the backbone segmentation network for the S3DIS dataset, while for ScanNet V2 we adopt PTv2 [29] to better capture global context in large-scale indoor scenes. To reduce memory usage while maintaining proportional coverage, we uniformly downsample points per room: 8,000 points for rooms with fewer than 100,000 points, and 15,000 points otherwise.

### 3.2. Results on Weakly Supervised Segmentation

**S3DIS**: Tab. 1 summarizes quantitative results on S3DIS Area 5 across different supervision levels. Our approach outperforms most existing weakly supervised methods under all label budgets. In particular, with 10% labeled points, our method achieves 49.9% mIoU, surpassing all compared approaches, including those trained with 100% supervision. These results demonstrate that our framework serves as a powerful plug-in for segmentation networks such as DGCNN, delivering significant performance improvements with minimal labeling costs, and showing clear advantages, especially in highly sparse annotation scenarios. As shown in Fig. 3, our method surpasses all competing approaches under the highly constrained 1-pt supervision setting (one labelled point per category), and demonstrates a steady performance gain as the annotation ratio increases. Notably, with just 10% labeled data, our method approaches the fully supervised upper bound.

**ScanNet V2**: To further assess generalizability, we evaluate on ScanNet V2 under two challenging weak-label settings: 0.1% labeled points and 20 points per scene (20 pts) [13], as shown in Tab. 2. Despite the extreme sparsity, our method consistently outperforms prior state-of-the-art approaches. With only 20 points per scene, it surpasses CPCM by 2.6% mIoU, while under the 0.1% label setting, it achieves a 4.0% improvement. Remarkably, our approach even exceeds baselines trained with 1% labels, highlighting its robustness under severe label scarcity and its strong potential for practical deployment in scenarios with highly limited annotation budgets.

### 3.3. Ablation Study

As shown in Tab. 3, the baseline without any component achieves 40.1% mIoU. When we rely solely on the superpoints of [8] as geometric hyperedges (i.e., unsupervised segmentation of a point cloud into geometrically homogeneous clusters), the performance drops sharply to 31.4%, suggesting that such geometric grouping alone is ineffective under extremely sparse supervision (1pt). Replacing them with SAM-based hyperedges, constructed from masks rendered over 46 multi-view images, recovers the performance to 39.5%, highlighting the advantage of high-quality 2D segmentation priors for label propagation. When both geometric and SAM-based hyperedges are used, the mIoU rises to 41.4%, indicating that geometric consistency from 3D and texture-aware cues from 2D are complementary. Moreover, introducing confidence filtering and enhanced rendering strategies yields a substantial improvement, with the performance jumping from 41.4% to 47.5% (36 views) and further to 47.9% (46 views). This shows that confidence filtering is crucial for suppressing noisy pseudo-labels, while incorporating

**Table 1**: Comparison with different methods on the S3DIS Area-5.* denotes results based on our reimplementation.

| Setting | | Model | ceil. | floor | wall | beam | col. | win. | door | chair | table | book. | sofa | board | clutter | Avg. |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Ful. Sup. | | PointNet [3] | 88.8 | 97.3 | 69.8 | 0.1 | 3.9 | 46.3 | 10.8 | 52.6 | 58.9 | 40.3 | 5.9 | 26.4 | 33.2 | 41.1 |
| | | PointNet++ [21] | 90.3 | 95.6 | 69.3 | **0.1** | 13.8 | 26.7 | **44.1** | 64.3 | 70.0 | 27.8 | **47.8** | 30.8 | 38.1 | 47.8 |
| | | SegCloud [22] | 90.1 | 96.1 | 69.9 | 0.0 | **18.4** | 38.4 | 23.1 | **75.9** | **70.4** | **58.4** | 40.9 | 13.0 | **41.6** | 48.9 |
| | | DGCNN* [23] | **92.2** | **97.6** | **74.5** | 0.0 | 8.3 | **49.5** | 27.1 | 72.8 | 69.2 | 51.8 | 24.6 | **34.4** | 41.3 | **49.5** |
| Unsup. | | Kmeans | 59.8 | 63.3 | 34.9 | 21.5 | **24.6** | 34.2 | 29.3 | 35.7 | 33.1 | 45.0 | 45.6 | 41.7 | 30.4 | 38.4 |
| | | Ncut [24] | **63.5** | **63.8** | **37.2** | **23.4** | **24.6** | **35.5** | **29.9** | **38.9** | **34.3** | **47.1** | **46.3** | **44.1** | **31.5** | **40.0** |
| Weak Sup. | 1pt | Π Model [25] | 89.1 | 97.0 | 71.5 | 0.0 | 3.6 | 43.2 | 27.4 | 62.1 | 63.1 | 14.7 | 43.7 | 24.0 | 36.7 | 44.3 |
| | | MT [26] | 88.9 | 96.8 | 70.1 | 0.1 | 3.0 | 44.3 | 28.8 | 63.6 | 63.7 | 15.5 | **43.7** | 23.0 | 35.8 | 44.4 |
| | | WeakSup [9] | 90.1 | **97.1** | 71.9 | 0.0 | 1.9 | 47.2 | 29.3 | 62.9 | 64.0 | 15.9 | 42.2 | 18.9 | 37.5 | 44.5 |
| | | OTOC [16] | 89.0 | 96.6 | 69.0 | **0.2** | **7.6** | 43.6 | 34.4 | 59.4 | 59.7 | 16.1 | 43.2 | 36.9 | 37.1 | 45.6 |
| | | MulPro [12] | **90.1** | 96.3 | 71.8 | 0.0 | 6.7 | 46.7 | **39.2** | **67.2** | **67.4** | 21.8 | 39.2 | 33.0 | 38.0 | 47.5 |
| | | **Ours** | 89.2 | 95.5 | **74.7** | 0.0 | 6.9 | **51.0** | 35.7 | 60.3 | 60.3 | **54.0** | 17.8 | **38.8** | **38.4** | **47.9** |
| | 10% | Π Model [25] | 91.8 | 97.1 | 73.8 | 0.0 | 5.1 | 42.0 | 19.6 | 66.7 | 67.2 | 19.1 | 47.9 | 30.6 | 41.3 | 46.3 |
| | | MT [26] | **92.2** | 96.8 | 74.1 | 0.0 | 10.4 | 46.2 | 17.7 | 67.0 | 70.7 | 24.4 | 50.2 | 30.7 | 42.2 | 47.9 |
| | | WeakSup [9] | 90.9 | 97.3 | 74.8 | 0.0 | 8.4 | 49.3 | 27.3 | **69.0** | **71.7** | 16.5 | **53.2** | 23.3 | 42.8 | 48.0 |
| | | OTOC [16] | 91.2 | **97.7** | **78.0** | 0.0 | 6.3 | 46.3 | 31.6 | 65.7 | 64.4 | 8.2 | 52.5 | 41.6 | **43.1** | 48.2 |
| | | MulPro [12] | 89.7 | 96.9 | 75.5 | 0.0 | 14.0 | 45.7 | **40.7** | 68.5 | 66.8 | 13.9 | 49.4 | 34.4 | 41.2 | 49.0 |
| | | **Ours** | 85.3 | 94.1 | 71.7 | 0.0 | **17.7** | **51.9** | 38.4 | 67.1 | 65.4 | **49.1** | 23.0 | **47.0** | 38.4 | **49.9** |

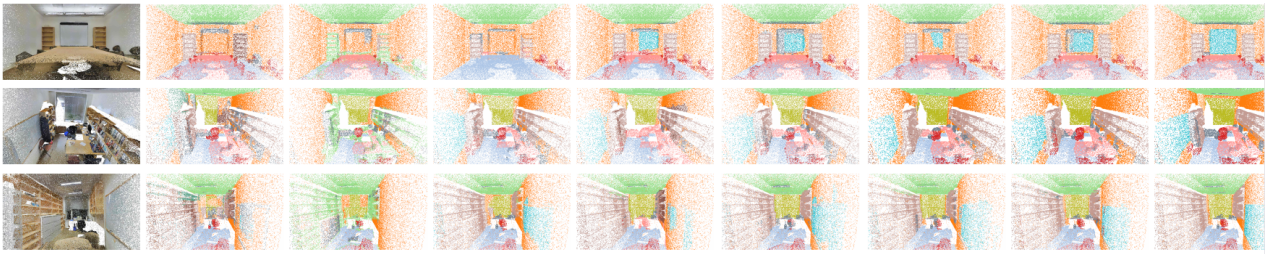


**Fig. 4**: Qualitative results of our method vs. DGCNN baseline (1pt) on the S3DIS dataset.

**Table 2**: Comparisons with state-of-the-art methods on ScanNet-v2 val. set.

| Method | Setting | mIoU |
|---|---|---|
| DGCNN [23] | Fully | 48.4 |
| MinkNet [30] | | 72.9 |
| PTv2 [29] | | 75.4 |
| HybridCR [31] | 1% | 56.9 |
| SQN [10] | 0.1% | 58.4 |
| CPCM [13] | | 63.8 |
| **Ours** | | **67.8** |
| WYPR [32] | 20 points/scene | 51.5 |
| CSC_LA_SEM [33] | | 53.1 |
| PointContrast_LA_SEM [34] | | 55.0 |
| MIL [11] | | 57.8 |
| OTOC [16] | | 59.4 |
| VIBUS [35] | | 58.6 |
| CPCM [13] | | 62.7 |
| **Ours** | | **65.3** |

more views continues to provide additional complementary cues that further refine the segmentation results.

Fig. 4 presents the qualitative results of our method on the S3DIS dataset, where the first six columns from left to right correspond to visualizations of different ablation settings. Under extremely limited supervision (1pt supervision), fully supervised methods such as DGCNN struggle to produce reasonable semantic structures, leading to fragmented predictions and confusion between semantic categories in some regions. In contrast, under the same supervision setting, our method yields more coherent segmentation results with clearer object boundaries. As the number of labeled points increases (e.g., 10% supervision), the overall segmentation quality is further

**Table 3**: Ablation study conducted on S3DIS Area-5, where the annotation cost is 1pt.

| #Views | Geometric Hyperedges | SAM-based Hyperedges | Confidence Filtering | mIoU (%) |
|---|---|---|---|---|
| - | | | | 40.1 |
| - | ✓ | | | 31.4 |
| 46 | | ✓ | | 39.5 |
| 46 | ✓ | ✓ | | 41.4 |
| 36 | ✓ | ✓ | ✓ | 47.5 |
| 46 | ✓ | ✓ | ✓ | 47.9 |

improved. And the ablation visualizations demonstrate that jointly incorporating geometric hyperedges and SAM-based hyperedges enhances region consistency, while confidence filtering effectively suppresses the propagation of noisy pseudo-labels, which is consistent with the quantitative results reported in Table 3. Furthermore, keeping other settings constant, increasing the number of views leads to qualitatively more coherent segmentation, while the corresponding quantitative gains are relatively minor.

## 4. CONCLUSION

We propose a weakly supervised 3D point cloud segmentation framework that exploits multi-view 2D cues and SAM-based dense masks. By modeling mask fusion as hypergraph label propagation, our method produces reliable pseudo-labels that effectively complement sparse supervision and guide 3D learning. Extensive experiments on S3DIS and ScanNet V2 demonstrate consistent improvements over existing methods. This work highlights the potential of transferring 2D foundation model priors to scalable 3D weak supervision, and future efforts will explore tighter end-to-end 2D–3D integration and extensions to instance-level and scene-level 3D understanding tasks

**Acknowledgement**: This research work is supported by the Agency for Science, Technology and Research (A*STAR) under its MTC Programmatic Funds (Grant No. M23L7b0021).

## 5. REFERENCES

[1] A. Nguyen and B. Le, "3d point cloud segmentation: A survey," in *RAM*, 2013.

[2] D. Zermas, I. Izzat, and N. Papanikolopoulos, "Fast segmentation of 3d point clouds: A paradigm on lidar data for autonomous vehicle applications," in *ICRA*, 2017.

[3] C. R. Qi et al., "Pointnet: Deep learning on point sets for 3d classification and segmentation," in *CVPR*, 2017.

[4] T. Le and Y. Duan, "Pointgrid: A deep network for 3d shape understanding," in *CVPR*, 2018.

[5] Q. Hu et al., "Randla-net: Efficient semantic segmentation of large-scale point clouds," in *CVPR*, 2020.

[6] I. et al. Armeni, "3d semantic parsing of large-scale indoor spaces," in *CVPR*, 2016.

[7] A. Dai et al., "Scannet: Richly-annotated 3d reconstructions of indoor scenes," in *CVPR*, 2017.

[8] L. Landrieu and M. Simonovsky, "Large-scale point cloud semantic segmentation with superpoint graphs," in *CVPR*, 2018.

[9] X. Xu and G. H. Lee, "Weakly supervised semantic point cloud segmentation: Towards 10x fewer labels," in *CVPR*, 2020.

[10] Q. Hu et al., "Sqn: Weakly-supervised semantic segmentation of large-scale 3d point clouds," in *ECCV*. Springer, 2022.

[11] C.-K. Yang et al., "An mil-derived transformer for weakly supervised point cloud segmentation," in *CVPR*, 2022.

[12] Y. Su, X., and K. Jia, "Weakly supervised 3d point cloud segmentation via multi-prototype learning," *IEEE Transactions on Circuits and Systems for Video Technology*, 2023.

[13] L. Liu et al., "Cpcm: Contextual point cloud modeling for weakly-supervised point cloud semantic segmentation," in *ICCV*, 2023.

[14] J. Wei et al., "Multi-path region mining for weakly supervised 3d semantic segmentation on point clouds," in *CVPR*, 2020.

[15] M. Cheng et al., "Sspc-net: Semi-supervised semantic 3d point cloud segmentation network," in *AAAI*, 2021.

[16] Z. Liu, X. Qi, and C.-W. Fu, "One thing one click: A self-training approach for weakly supervised 3d semantic segmentation," in *CVPR*, 2021.

[17] M. Liu et al., "Less: Label-efficient semantic segmentation for lidar point clouds," in *ECCV*, 2022.

[18] A. Kirillov et al., "Segment anything," in *ICCV*, 2023.

[19] Nanqing Liu, Xun Xu, Yongyi Su, Haojie Zhang, and Heng-Chao Li, "Pointsam: Pointly-supervised segment anything model for remote sensing images," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 63, pp. 1–15, 2025.

[20] Qiwei Lin, Nanqing Liu, Zhiyuan Tan, Yang Liu, and Qinghua Long, "Point-supervised oriented ship detection via segment anything model for sar images," *IEEE Geoscience and Remote Sensing Letters*, vol. 23, pp. 1–5, 2026.

[21] C. R. Qi et al., "Pointnet++: Deep hierarchical feature learning on point sets in a metric space," *NeurIPS*, 2017.

[22] L. Tchapmi et al., "Segcloud: Semantic segmentation of 3d point clouds," in *2017 international conference on 3D vision (3DV)*. IEEE, 2017.

[23] Y. Wang et al., "Dynamic graph cnn for learning on point clouds," *ACM Transactions on Graphics*, 2019.

[24] J. Shi and J. Malik, "Normalized cuts and image segmentation," *IEEE Transactions on pattern analysis and machine intelligence*, 2000.

[25] S. Laine and T. Aila, "Temporal ensembling for semi-supervised learning," in *International Conference on Learning Representations*, 2017.

[26] A. Tarvainen and H. Valpola, "Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results," *NeurIPS*, 2017.

[27] D. Zhou, J. Huang, and B. Schölkopf, "Learning with hypergraphs: Clustering, classification, and embedding," *NeurIPS*, 2006.

[28] Y. Zhang et al., "Perturbed self-distillation: Weakly supervised large-scale point cloud semantic segmentation," in *ICCV*, 2021.

[29] X. Wu et al., "Point transformer v2: Grouped vector attention and partition-based pooling," *NeurIPS*, 2022.

[30] C. Choy, J. Gwak, and S. Savarese, "4d spatio-temporal convnets: Minkowski convolutional neural networks," in *CVPR*, 2019.

[31] M. Li et al., "Hybridcr: Weakly-supervised 3d point cloud semantic segmentation via hybrid contrastive regularization," in *CVPR*, 2022.

[32] Z. Ren et al., "3d spatial recognition without spatially labeled 3d," in *CVPR*, 2021.

[33] J. Hou et al., "Exploring data-efficient 3d scene understanding with contrastive scene contexts," in *CVPR*, 2021.

[34] S. Xie et al., "Pointcontrast: Unsupervised pre-training for 3d point cloud understanding," in *ECCV*, 2020.

[35] B. Tian et al., "Vibus: Data-efficient 3d scene parsing with viewpoint bottleneck and uncertainty-spectrum modeling," *ISPRS Journal of Photogrammetry and Remote Sensing*, 2022.