Semi-supervised 3D Semantic Scene Completion with 2D Vision Foundation Model Guidance

Duc-Hai Pham[1], **Duc-Dung Nguyen**[2], **Anh Pham**[2], **Tuan Ho**[1], **Phong Nguyen**[1], **Khoi Nguyen**[1], **Rang Nguyen**[1]

[1]VinAI Research, Vietnam
[2]AITech Lab., Ho Chi Minh City University of Technology, VNU-HCM, Vietnam
{v.haipd13, v.tuanhl25, v.phongnh31, v.khoindm, v.rangnhm}@vinai.io, {nddung, anhpham}@hcmut.edu.vn

Abstract

Accurate prediction of 3D semantic occupancy from 2D visual images is crucial for enabling autonomous agents to understand their surroundings for planning and navigation. State-of-the-art methods typically rely on fully supervised approaches, requiring large labeled datasets obtained through expensive LiDAR sensors and meticulous voxel-wise annotation by human experts. The resource-intensive nature of this annotation process significantly limits the scalability and application of these methods. To address this challenge, we propose a novel semi-supervised framework that reduces reliance on densely annotated data. Our approach leverages 2D foundation models to extract essential 3D scene geometry and semantic cues, enabling a more efficient training process. The proposed framework has two key advantages: (1) **Generalizability**, as it is compatible with various 3D semantic scene completion methods, including 2D-3D lifting and 3D-2D transformer techniques; and (2) **Effectiveness**, as demonstrated by experiments on the SemanticKITTI and NYUv2 datasets, where our method achieves up to 85% of the fully supervised performance using only 10% of the labeled data. This approach not only reduces the cost of data annotation but also highlights its potential for broader adoption in vision-based systems for 3D semantic occupancy prediction.

Introduction

Vision-based 3D semantic scene completion (SSC) is a crucial vision task with numerous applications in both indoor and outdoor scenarios (Song et al. 2017; Liu et al. 2018; Roldao, de Charette, and Verroust-Blondet 2020; Wu et al. 2020; Li et al. 2020; Dourado et al. 2021; Huang et al. 2023b; Cao and de Charette 2022; Tong et al. 2023; Yao et al. 2023; Li et al. 2023; Zhang, Zhu, and Du 2023). Given an input image, the goal is to predict a complete 3D voxel representation of volumetric occupancy and semantic labels for a scene. This task is challenging due to the absence of 3D information in 2D images. Most SSC methods (Song et al. 2017; Li et al. 2023; Zhang, Zhu, and Du 2023) address this challenge through fully supervised learning, which relies on dense 3D SSC annotations. However, obtaining these annotations requires deploying vehicles equipped with LiDAR sensors and performing meticulous voxel-level human labeling, which severely limits the scalability and deployment of many SSC systems. An alternative approach to reduce the annotation costs is self-supervised learning (Cao and de Charette 2023; Huang et al. 2023a; Zhang et al. 2023a), which only requires unlabeled images for training. However, these methods often produce suboptimal results, with a significant performance gap compared to fully supervised approaches.

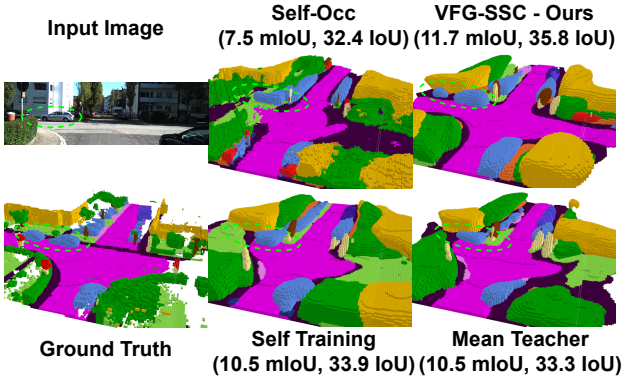

Figure 1: **Visual comparison on SemanticKITTI**: VFG-SSC surpasses self-supervised methods like SelfOcc by 4% mIoU and outperforms other semi-supervised methods: Self-Training and Mean Teacher by 1.5% mIoU.

To address these limitations, this paper introduces the problem of semi-supervised vision-based 3D semantic scene completion (Semi-SSC). In this setup, a small percentage of images are labeled with 3D semantic occupancy, while the majority remain unlabeled. To overcome the challenge of limited labeled data, semi-supervised learning approaches like *Self-Training* (Lee et al. 2013) and *Mean-Teacher* (Tarvainen and Valpola 2017) can be employed. In Self-Training (ST), a fully supervised SSC network first learns from labeled data and then predicts pseudo-labels for unlabeled data. These pseudo-labels are used to retrain the network, leveraging the unlabeled data. In contrast, the Mean-Teacher (MT) approach involves a more stable Teacher network, derived from the student network's Exponential Moving Average (EMA), which predicts pseudo-labels for unlabeled images. These pseudo-labels are combined with labeled data to continually train the student network. While there are ad-

vanced semi-supervised learning techniques for 3D, such as LaserMix (Kong et al. 2023), CutMix (Yun et al. 2019), and ClassMix (Olsson et al. 2021), these methods rely on a one-to-one correspondence between input and output or homogeneous data. This approach does not apply to Semi-SSC, where the input consists of 2D images and the output is in 3D voxels.

On the other hand, recent vision foundation models (VFMs) have shown the ability to robustly extract metric depth (Piccinelli et al. 2024) and semantics (Zhang et al. 2023b) from unlabeled images. However, these capabilities are not yet leveraged in SSC. To bridge this gap, we introduce VFG-SSC, a method that leverages 2D VFMs to generate 3D priors from unlabeled images, thereby improving the quality of pseudo-labels during training. Adapting 2D VFM knowledge to 3D presents challenges such as depth scale ambiguity and feature misalignment from segmentation models. To address these, we developed a novel attention-based enhancement module that (1) globally refines features and (2) maintains linear complexity, enabling efficient processing of large 3D datasets. Futhermore, we propose a technique to accumulate VFM information over image sequences, reducing noise and biases compared to using a single image. Our method can serve as a universal plugin for any vision-based SSC method, enabling a semi-supervised alternative for fully supervised approaches.

We evaluate our approach on the outdoor SemanticKITTI (Behley et al. 2019) and indoor NYUv2 (Silberman et al. 2012) datasets. Our results demonstrate strong performance across both settings, achieving 85% of the accuracy of fully supervised methods while utilizing only 10% of the 3D semantic occupancy annotations.

In summary, our contributions are as follows:

- We introduce a new semi-supervised 3D semantic scene completion (Semi-SSC) framework to reduce the high cost of 3D voxel-wise annotation.
- We propose VFG-SSC, a novel approach that leverages 3D cues from 2D vision foundation models for depth and segmentation.
- We design a new enhancement module to effectively combine 3D cues with 3D features, seamlessly integrating with existing SSC methods.
- Our pipeline achieves 85% of fully supervised performance using just 10% of the labeled data.

Related Work

This section reviews literature on fully, semi, and self-supervised 3D semantic scene completion, followed by an overview of the 2D depth and semantic segmentation foundation models used in our framework.

3D Semantic Scene Completion (SSC). The primary goal of 3D semantic scene completion (SSC) or 3D semantic occupancy prediction is to jointly infer the geometry and semantics of a scene. Initially introduced in SSCNet (Song et al. 2017) for indoor scenes like NYUv2 (Silberman et al. 2012), the task has since been extended to outdoor environments such as SemanticKITTI (Behley et al. 2019) (Roldao, de Charette, and Verroust-Blondet 2020; Li et al. 2020; Yan et al. 2021; Cheng et al. 2021; Mei et al. 2023; Xia et al. 2023; Jia et al. 2023). **Vision-based SSC** methods, which are cost-effective and easily deployable, have gained traction. MonoScene (Cao and de Charette 2022) was the first to use only RGB images for SSC by projecting 3D voxels onto 2D images and processing them with a 3D UNet to predict 3D semantic occupancy. VoxFormer (Li et al. 2023) improves upon this by leveraging depth estimation priors to focus cross-attention on voxels near object surfaces. OccFormer (Zhang, Zhu, and Du 2023) further enhances 3D feature encoding using windowed attention on local and global transformer pathways and employs a mask decoder, similar to Mask2Former (Cheng et al. 2022), to predict 3D SSC. However, these methods still rely on expensive voxel-wise annotations, limiting their practical application.

Self-supervised SSC. One way to reduce the annotation cost of SSC is through self-supervised learning, which leverages video sequences along with predicted depth and semantic segmentation as pseudo-labels. Methods like SceneRF (Cao and de Charette 2023), SelfOcc (Huang et al. 2023a), and OccNeRF (Zhang et al. 2023a) use volumetric rendering on predicted 3D semantic occupancy to generate RGB images, depth maps, and segmentation maps, allowing for loss computation with pseudo-labels. Although self-supervised SSC requires less supervision than our VFG-SSC approach, its performance is significantly lower than that of semi-supervised and fully-supervised methods, making it unsuitable for practical applications.

Semi-supervised learning for 2D and 3D LiDAR segmentation. Semi-supervised image segmentation methods (Chen et al. 2021; Liu et al. 2022; Kwon and Kwak 2022; Wang et al. 2022; Yang et al. 2023; Hoyer et al. 2023) have demonstrated strong performance on autonomous datasets by leveraging two key strategies: consistency regularization and entropy minimization. Consistency regularization employs augmentations to ensure stable predictions under perturbed inputs, while entropy minimization reduces uncertainty in the classification of unlabeled data, leading to more confident predictions. In 3D LiDAR segmentation, LaserMix (Kong et al. 2023) enforces consistency through mixing schemes, while methods like 3DIoUMatch (Wang et al. 2021) and DetMatch (Park et al. 2022) use multi-sensor data to generate high-quality pseudo-labels for object detection. However, these methods rely on a homogeneous input-output relationship. In contrast, Semi-SSCpresents a greater challenge as it takes 2D images as input and produces 3D semantic occupancy as output. Consequently, directly applying the above methods to Semi-SSC is non-trivial. Moreover, Semi-SSConly uses images for unlabeled data, avoiding the significant setup costs associated with collecting multi-sensor data.

2D foundation models for depth estimation and segmentation. Segmentation models like Segment Anything (Kirillov et al. 2023) have introduced class-agnostic segmentation, demonstrating robustness across diverse datasets. Following this, models such as X-Decoder (Zou et al. 2023), SEEM (Zou et al. 2024), and OpenSEED (Zhang et al.

Methods	Precision ↑	Recall ↑	mIoU ↑
Sup-only	47.53	47.09	9.16
Self-Training (ST)	42.83	**59.35**	9.48
3D clues as pseudo-labels	**48.60**	26.19	9.27
Our approach	48.01	58.36	**11.73**

Table 1: Comparison of Semi-SSC on 5% labeled SemanticKITTI validation data. The Sup-only baseline excludes pseudo-labels from unlabeled data during training.

2023b) have emerged as general-purpose panoptic segmentation tools, offering zero-shot application on autonomous datasets. For depth estimation, models like Metric3Dv2 (Yin et al. 2023) and UniDepth (Piccinelli et al. 2024) excel in robust metric depth estimation by leveraging large-scale datasets. In our work, we utilize these 2D VFMs to generate 3D clues, addressing the challenge of limited labeled data.

Semi-Supervised Semantic Scene Completion

Problem Setting: Given a sequence $\mathbf{Q}_t = \{I_{t-k}, I_{t-k+1} \ldots, I_{t-1}, I_t\}$ of k consecutive frames, the SSC model f_θ generates a semantic occupancy grid which is defined in the coordinate system of the ego vehicle at the timestamp t. Each voxel of the grid is categorized as either empty or occupied by a specific semantic class. The grid can be obtained as follows: $\hat{\mathbf{Y}}_t = f_\theta(\mathbf{Q_t})$ where $\hat{\mathbf{Y}}_t \in \mathbb{R}^{H \times W \times Z \times (C+1)}$. H, W, and Z denote the voxel grid's height, width, and depth, and C is the number of the semantic classes. In a semi-supervised setting (Semi-SSC), the dataset contains two non-overlapping subsets:

- Labeled data $\mathcal{D}^L = \{(\mathbf{Q}_t, \mathbf{Y}_t)\}_{t=1}^L$ where each image I_t has a corresponding 3D occupancy ground-truth $\mathbf{Y}_t$.
- Unlabeled data $\mathcal{D}^U = \{\mathbf{Q}_t\}_{t=1}^U$, which only contains images with **no LiDAR** and the amount of the unlabeled data is significantly larger than labeled data, i.e, $U \gg L$.

Our Customized Self-Training for Semi-SSC: A common strategy for tackling the semi-supervised problem is Self-Training. This involves first training a supervised model f_θ on labeled data $\mathcal{D}^L$ and then generating pseudo-labels for the unlabeled dataset $\mathcal{D}^U$. After that, the model f_θ is retrained using both the labeled and pseudo-labeled data. Based on this Self-Training approach, our key contribution is to enhance the quality of the pseudo-label by incorporating 3D priors extracted from 2D vision foundation models. We now outline our three-step process in detail.

- **Step 1 - Training on labeled data**: We first extract efficient 3D cues from 2D images using 2D foundation models. These cues are fused with 3D features derived from a 3D SSC network f_θ through our proposed enhancement module g_ϕ. The networks f_θ and g_ϕ are jointly trained on labeled data $\mathcal{D}^L$, enabling the proposed enhancement module to learn the efficient fusion between 3D clues and 3D features with the supervision of labeled data.
- **Step 2 - Generating pseudo-label for unlabeled data**: The trained f_θ and g_ϕ from Step 1 are used to predict unlabeled data $\mathcal{D}^U$, resulting in a pseudo-labeled dataset $\hat{\mathcal{D}}^L = \{(\mathbf{I}_t, \hat{\mathbf{Y}}_t)\}_{t=1}^U$.
- **Step 3 - Retraining with labeled and pseudo-labeled data**: Finally, we retrain the original SSC model f_θ using the combined dataset $\mathcal{D}^L \cup \hat{\mathcal{D}}^L$ with supervised segmentation losses. Only f_θ is utilized during test time; 3D cues and g_ϕ are not used at this stage.

The next part presents our contributions in Steps 1 and 2.

Our VFG-SSC Network

To highlight the importance of improving pseudo-label quality in Self-Training, we conducted a quick test by training f_θ with various setups, as shown in Tab. 1. As can be seen, using additional pseudo-labeled data from Self-Training (second row) boosts recall but lowers precision compared to the model trained on 5% labeled data only (first row). To overcome this limitation, we incorporate 3D cues from 2D VFMs as pseudo labels for the unlabeled data. These 3D cues offer higher precision but lower recall, as they provide accurate yet sparse estimates of depth and semantics. By combining these sources, we enhance both precision and recall, leading to an overall improvement in mIoU. This approach motivates our strategy to align 3D cues with Self-Training by integrating them as inputs during pseudo-label generation. We achieve this enhancement through an attention-based mechanism that effectively fuses the two feature sets.

Our VFG-SSC network, as shown in Fig. 2, includes a unique 3D clue generation module, a semantic scene completion network f_θ, and an enhancement module g_ϕ.

3D Clues Generation

As mentioned in the previous section, we exploit 2D VFMs of depth estimation and semantic segmentation to create 3D clues. A 3D clue is defined as a discrete set of voxels $\mathbf{Y}_{\text{clues}} \in \mathbb{R}^{H \times W \times Z \times (C+1)}$. A straightforward approach is to extract depth and semantic information from the current frame, and then use the depth data to project the segmentation into a 3D voxel representation. However, as shown in Fig. 4, 3D clues generated from a single frame often contain numerous irrelevant artifacts that do not align with the actual scene. To address this issue, we propose generating clues using temporal image sequences. Particularly, given a set of N input images of a scene, we extract depth maps and semantic segmentation map from VFMs. For each pixel $u = [u_x, u_y]$ and its corresponding depth d, we un-project the semantic map u to 3D point x in world coordinate by computing $x = E^{-1} \times K^{-1} \times [u_x, u_y, d]^T$, where K and E are the camera intrinsic and extrinsic matrices respectively. This results in a set of semantic point clouds $X \in \mathbb{R}^{N \times 4}$ (3D coordinates and semantic class). We convert these point clouds to voxels to obtain $\mathbf{Y}_{\text{clues}}$.

As shown in Fig. 4, accumulating 3D clues from temporal frames (Multi Frame) results in a more accurate scene representation compared to using a single frame. However, dynamic objects like vehicles become indistinguishable. To address this, we introduce a regularization pipeline that leverages semantic and geometric constraints. We begin by segmenting the semantic point cloud produced in the previous

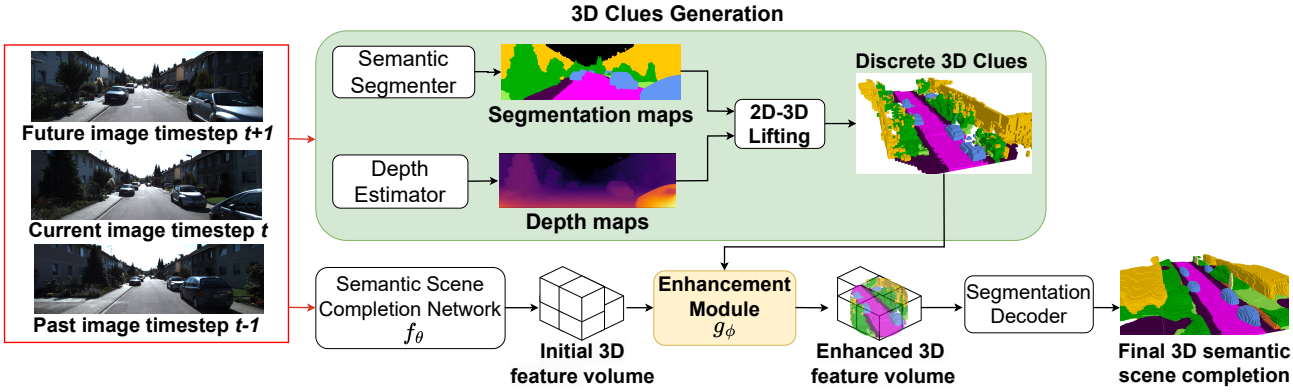

Figure 2: The overall architecture of the proposed VFG-SSC network. Our approach leverages 3D cues from 2D foundation models to enhance the inferred 3D feature volume and generate the final 3D semantic occupancy grid. The model is trained using all frames (solid red box) to produce pseudo-labels for the unlabeled data.

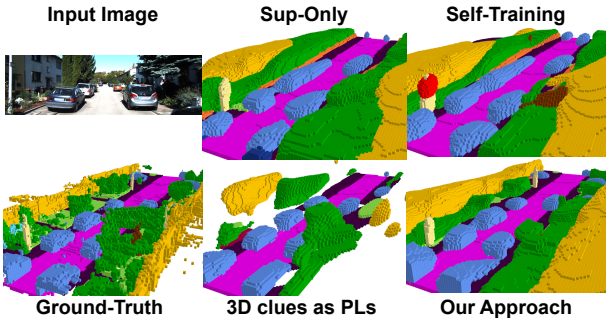

Figure 3: Qualitative of Semi-SSC approaches. With Sup-only and Self-Training, the scene layout is reasonably reconstructed, however, dynamic object prediction is incorrect due to many false positives. With 3D clues as pseudo-labels, objects are correctly predicted but occluded regions do not have a prediction (assigned as empty). Our method obtains reliable predictions (high precision) and reasonable reconstruction in occluded space (high recall).

step by class and applying heuristic filters: a radius filter to remove outliers based on the density of nearby points, and a statistical filter to eliminate points that deviate significantly from the average distance. The results in Tab. 8 show that this pipeline is crucial for effectively utilizing VFMs. A pseudo-code implementation is provided in the *Supplementary Material*.

3D Clue-Guided Enhancement Module

This module, g_ϕ, fuses the 3D clues from the previous module with the 3D features generated by the scene completion network f_θ to create an enhanced 3D feature volume. A naive approach is to concatenate these 3D clues with the features from f_θ to obtain a refined feature. However, our visualized results in Fig. 3 reveal that these features contain complementary information that can be better utilized. The 3D clues can reliably estimate the semantics and geometry

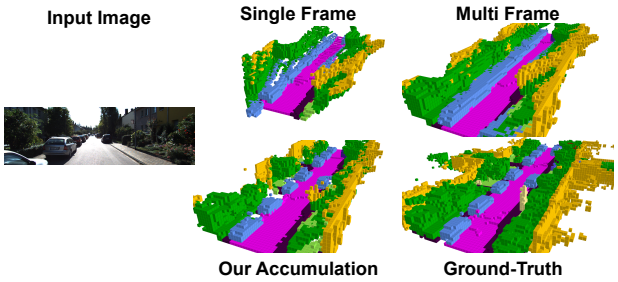

Figure 4: 3D clues visual comparisons on different settings: using only the current image and accumulated 3D clues from temporal frames with and without filtering.

of visible surfaces, but due to the limited camera FoV, many voxels remain without sufficient information. Concatenation alone may introduce ambiguities in these regions. An alternative approach involves cross-attention, which associates each feature in f_θ with all 3D clues, but this can lead to a significant increase in computational cost. To balance effectiveness and efficiency, we propose using a local attention mechanism to merge the two features, achieving both global context aggregation and linear computational complexity.

Our enhancement module uses a set of self-attention blocks to align the two features implicitly. Specifically, given 3D clues $\mathbf{Y}_{\text{clues}}$ and 3D features $\mathbf{Y}_{\text{feats}}$, we project the discretized 3D clues into continuous embeddings. We then concatenate the aligned features with the original 3D features and apply self-attention for enhancement. Note that we adopt Dilated Neighborhood Attention (Hassani et al. 2023), a recently proposed method with linear time and space complexity, to reduce the computational cost of self-attention.

Experiments

Datasets and metrics: We utilize two 3D semantic occupancy datasets, SemanticKITTI (Behley et al. 2019) and NYUv2 (Silberman et al. 2012), which represent outdoor

Backbone	Method	1%		5%		10%	
		mIoU↑	IoU↑	mIoU↑	IoU↑	mIoU↑	IoU↑
SceneRF	Self-supervised	N/A mIoU, 13.84 IoU					
SelfOcc	Self-supervised	5.02 mIoU, 21.97 IoU					
SelfOcc	Finetuned	5.55	29.57	7.5	32.43	8.3	33.66
OccFormer	Fully-Supervised	14.08 mIoU, 36.04 IoU					
	Supervised-only	6.59	26.31	9.16	32.73	10.45	34.25
	Mean-Teacher	6.76	26.54	9.41	31.67	10.53	33.28
	Self-Training	6.75	26.61	9.48	33.12	10.49	33.94
	VFG-SSC (Ours)	**9.40**	**30.40**	**11.73**	**35.76**	**12.38**	**35.25**
VoxFormer	Fully-Supervised	14.40 mIoU, 43.50 IoU					
	Supervised-only	6.35	34.04	9.56	39.68	10.52	40.28
	Mean-Teacher	7.20	**38.97**	9.58	37.70	10.56	38.66
	Self-Training	7.60	35.40	10.51	39.83	11.06	41.05
	VFG-SSC (Ours)	**9.32**	36.78	**11.15**	**39.96**	**12.19**	**41.57**

Table 2: Results on the val set of SemanticKITTI. Faded rows indicate reference-only results, not for direct comparison.

and indoor environments, respectively, to establish new semi-supervised benchmarks. For SemanticKITTI, we sample 40, 198, and 383 frames (corresponding to 1%, 5%, and 10% of the training set), consistent with existing setups (Wang et al. 2023; Behley et al. 2019). For NYUv2, we uniformly sample 40 and 80 frames (representing 5% and 10% of the training set). We evaluate performance using mIoU for semantic class and IoU for binary occupancy predictions.

Implementation details: For SemanticKITTI, we use OccFormer (Zhang, Zhu, and Du 2023) and VoxFormer (Li et al. 2023) as our primary SSC networks (f_θ). VoxFormer employs a two-stage training pipeline: we pre-train the query proposal with 1/5/10% labeled data in stage 1, then apply VFG-SSC in stage 2. For OccFormer, which uses additional depth supervision from LiDAR, we supplement the missing LiDAR data in unlabeled sets with pseudo-LiDAR generated from a depth model (Piccinelli et al. 2024). We maintain consistent settings across both models for fair comparison. For NYUv2, we use MonoScene (Cao and de Charette 2022) and OccDepth (Miao et al. 2023) as the SSC networks.

UniDepth (Piccinelli et al. 2024) is used for depth estimation, and SEEM (Zou et al. 2024) for semantics, serving as our foundational models. We employ 2 layers of Dilated Neighborhood Attention with 4 heads, a kernel size of 7, and 4 dilation rates (1, 2, 4, 8). Additional details on losses for training each SSC network and further implementation specifics are provided in the Supplementary Material.

Comparison with Related Methods

As the first work to address the 3D semantic scene completion task in a semi-supervised manner, we adapt two semi-supervised learning approaches as baselines: Mean Teacher (Tarvainen and Valpola 2017), which leverages consistency regularization, and Self-Training (Lee et al. 2013) with entropy minimization. In the Semi-SSC setting, we train our model using only 1%, 5%, and 10% of the labeled training data and compare its performance against other methods trained with the same data percentages. For reference, we also include results from fully supervised training using 100% labeled data, as well as two self-supervised methods, SceneRF (Cao and de Charette 2023) and SelfOcc (Huang et al. 2023a), which rely entirely on unlabeled data. We also evaluate a version of SelfOcc fine-tuned on the same labeled data percentages with an added segmentation head.

Quantitative results: For SemanticKITTI, we present IoU and mIoU in Tab. 2. Additionally, we validate our results by reporting VFG-SSC performance with a 10% labeled data setting on the SemanticKITTI hidden test set in Table Tab. 3. Our VFG-SSC significantly outperforms baseline methods. Remarkably, our approach achieves comparable results to fully supervised counterparts, with just a 15% performance gap using 10% labeled data. It is worth noting that on the 1% setting, Mean Teacher obtains a better IoU score on VoxFormer. Therefore, a robust Semi-SSC method should achieve excellent performance in both geometric and semantic completion. Moreover, on the SemanticKITTI hidden test set, our method compares favorably with some fully supervised methods like MonoScene, despite utilizing only 10% labeled occupancy annotation. This emphasizes the effectiveness of our 3D clues and enhancement module, which is effective in generating high-quality pseudo-labels for training any SSC backbones. For NYUv2, we report the results in Tab. 4. We consistently outperform other methods with both 5% and 10% of training data, underscoring the superiority of our approach over strong baselines. These results indicate that our VFG-SSC is generalizable to different architectures, can be applied to various labeled settings, and applies to both outdoor and indoor scenarios.

Qualitative results: We present the qualitative results of our method, VFG-SSC, compared to baseline models on the SemanticKITTI and NYUv2 datasets in Fig. 5 and Fig. 6, respectively. In outdoor scenarios, VFG-SSC outperforms the baselines, particularly in accurately predicting various instance classes, such as cars and people. In contrast, other baselines mis-segment people and produce irregular car shapes (row 1). Additionally, VFG-SSC excels in captur-

Method	mIoU↑	IoU↑	road	sidewalk	parking	other-grnd	building	car	truck	bicycle	motorcycle	other-vehi	vegetation	trunk	terrain	person	bicyclist	motorcyclist	fence	pole	traffic-sign
MonoScene	11.1	34.2	54.7	27.1	24.8	5.7	14.4	18.8	3.3	0.5	0.7	4.4	14.9	2.4	19.5	1.0	1.4	0.4	11.1	3.3	2.1
OccFormer	13.2	33.7	57.3	31.1	29.7	13.7	16.8	21.5	4.5	2.4	2.2	3.2	15.5	4.3	22.5	2.7	2.3	0.0	12.5	4.1	4.2
VoxFormer	14.3	42.0	56.5	30.0	25.8	12.3	23.7	23.4	5.1	3.4	1.6	2.7	24.5	8.6	24.4	1.6	1.3	0.0	15.4	6.3	6.0
VFG-SSC (Occ)	11.1	30.5	55.2	26.7	25.6	0.0	14.8	20.5	1.6	2.3	2.2	5.8	13.1	3.6	20.7	3.4	0.0	0.0	10.5	3.4	2.1
VFG-SSC (Vox)	11.8	40.6	55.5	23.8	16.3	0.0	20.9	21.5	2.2	1.4	1.8	3.5	23.7	8.2	20.0	1.5	0.0	0.0	12.5	5.8	4.9

Table 3: Quantitative comparisons between our method and other fully-supervised methods on the hidden testing set of SemanticKITTI (Behley et al. 2019) dataset. Faded rows denote results for reference purposes only, not direct comparison.

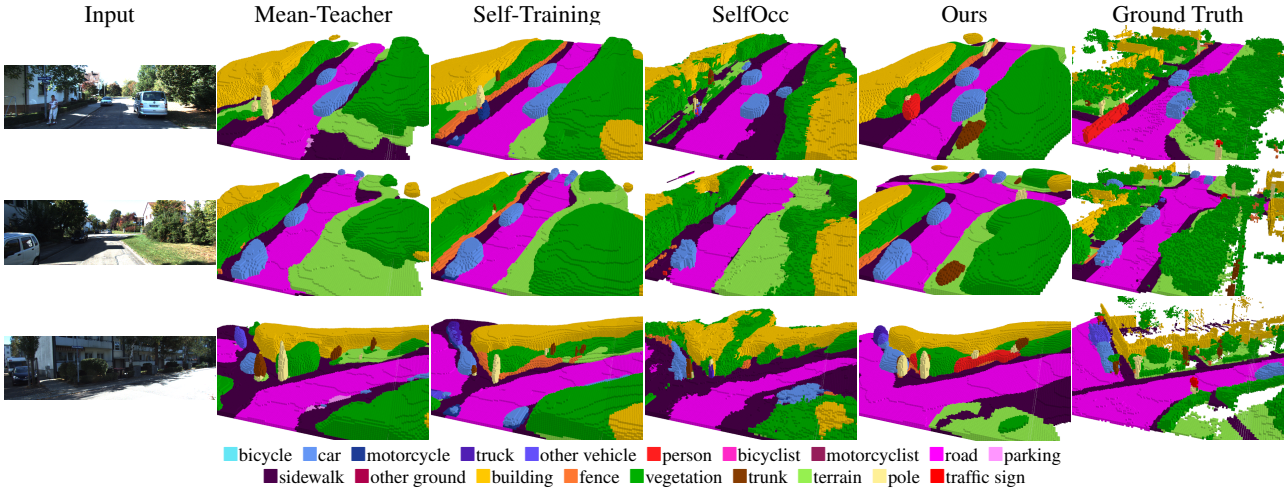

Figure 5: Qualitative results on the validation set of the SemanticKITTI (Behley et al. 2019) dataset.

SSC Network	Method	5%		10%	
		mIoU↑	IoU↑	mIoU↑	IoU↑
	Fully-Supervised	26.94 mIoU, 42.51 IoU			
MonoScene	Supervised-only	13.36	29.17	17.17	34.68
	Self-Training	14.00	30.64	17.32	35.20
	VFG-SSC	**18.23**	**43.35**	**22.20**	**44.37**
	Fully-Supervised	29.03 mIoU, 44.17 IoU			
OccDepth	Supervised-only	16.69	30.97	19.01	37.83
	Self-Training	16.72	30.49	19.42	38.07
	VFG-SSC	**22.54**	**43.61**	**23.92**	**44.85**

Table 4: Quantiative results on the testing set of the NYUv2 (Silberman et al. 2012) dataset. Faded rows denote results for reference purposes only, not direct comparison.

ing the broader scene layout, including long-range structures such as crossroads, sidewalks, and buildings (rows 2, 3). In indoor scenes, VFG-SSC effectively reconstructs the global scene layout, including both visible and occluded regions (row 2), while also recovering finer details of smaller objects, such as chairs, tables, and sofas (rows 1, 3).

Ablation Study

In this section, we ablate our approach on 5% labeled data of SemanticKITTI with OccFormer as the SSC network f_θ.

Enhancement techniques: In Tab. 5, we compare our attention-based alignment method against several enhancement techniques, including simple concatenation, BEVFusion (Liang et al. 2022), and OpenOccupancy's Adaptive Fusion (Wang et al. 2023). Additionally, we include an attention-only baseline where 3D features and 3D clues are concatenated along the feature dimension, followed by Deformable Self-Attention (Zhu et al. 2020). We evaluate the quality of the pseudo-labels generated using training ground truth (GT) labels (in step 2) and assess the final model's predictions against validation GT labels. The results demonstrate that our method consistently outperforms the other techniques on both the training and validation splits. Notably, even with the additional clues from VFMs, simply integrating them into an SSC network does not result in significant performance improvements compared to baselines without VFMs. We attribute the substantial performance gain to our enhancement module, which effectively leverages these additional clues.

Study on different depth estimation models: We conducted experiments using various depth foundation models

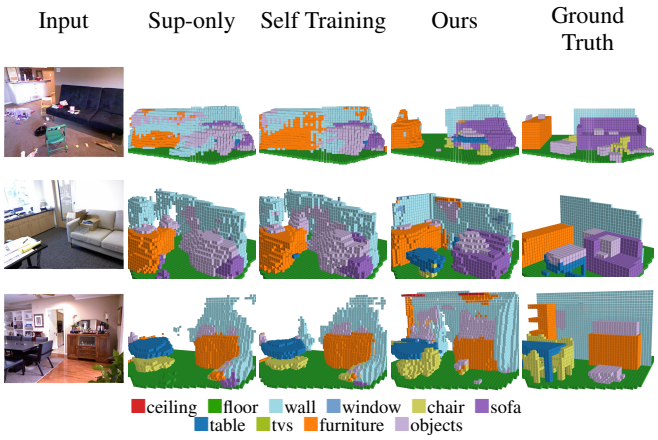

Figure 6: Qualitative results on the testing set of the NYUv2 (Silberman et al. 2012) dataset.

Enhancement Methods	Train Split mIoU ↑	Train Split IoU ↑	Val Split mIoU ↑	Val Split IoU ↑
None	8.37	31.60	9.48	33.12
Concat	13.03	42.29	10.01	34.05
BEVFusion	12.80	42.15	10.00	34.72
OpenOcc	12.42	40.08	10.37	34.37
Deformable	13.84	37.40	10.04	34.30
Ours	**15.59**	**42.73**	**11.73**	**35.76**

Table 5: Quantitative results of different fusion methods.

to generate 2D depth maps. Since the effectiveness of our method hinges on the accuracy of pseudo-labels, employing a more reliable depth predictor can potentially enhance the quality of the generated 3D clues. As shown in Tab. 6, our approach achieves the best results with UniDepth. It's worth mentioning that advancements in depth estimation will further enhance our method's performance. In this experiment, we used SEEM (Zou et al. 2024) as the segmentation model.

Study on different segmentation models: We present a study on various semantic foundation models like SEEM (Zou et al. 2024) (pretrained on COCO), OpenSEED (Zhang et al. 2023b) (pretrained on COCO, Object365), and SegFormer (Xie et al. 2021) (pretrained on CityScapes). The comparison results reported in Tab. 7 reveal marginal distinctions among these aforementioned models. SegFormer obtains the best results as the domain between training (CityScapes) and testing (SemanticKITTI) data is not substantial for 2D segmentation. This suggests that our method consistently produces high-quality 3D cues across diverse segmentation foundation models.

Impact analysis of temporal and regularization: Our analysis, detailed in Tab. 8, reveals several insights. In rows 1 and 2, we compare the performance of ST with and without the integration of a single frame of regularized 3D clues, noting a modest improvement of +0.09 mIoU. This enhancement can be attributed to the additional sparse information provided by the VFMs. In row 3, we evaluated the impact

Depth Model	Train Split mIoU ↑	Train Split IoU ↑	Val Split mIoU ↑	Val Split IoU ↑
Depth-Anything	12.63	37.66	10.38	35.62
Metric3Dv2	14.86	41.28	11.09	**36.20**
UniDepth	**15.59**	**42.73**	**11.73**	35.76

Table 6: Study on different depth foundation models.

Semantic Model	Train Split mIoU ↑	Train Split IoU ↑	Val Split mIoU ↑	Val Split IoU ↑
OpenSEED	14.80	40.90	11.40	36.68
SEEM	15.59	**42.73**	11.73	35.76
SegFormer	**16.18**	41.74	**11.94**	**36.78**

Table 7: Study on different semantic foundation models.

of incorporating temporal frames into 3D clue generation while omitting the regularization module, which resulted in a decrease of -0.62 mIoU. This drop indicates that the added temporal information from VFMs without regularization, introduces too much noise. However, when combining temporal accumulation with our regularization, we achieved a significant improvement of +2.25 mIoU. This demonstrates that while temporal information is valuable for VFG-SSC, regularization is essential for optimizing performance.

Regularize	Temporal	Train Split mIoU ↑	Train Split IoU ↑	Val Split mIoU ↑	Val Split IoU ↑
		8.37	31.6	9.48	33.12
✓		9.28	33.87	9.57	34.26
	✓	8.61	33.16	8.86	33.64
✓	✓	**15.59**	**42.73**	**11.73**	**35.76**

Table 8: Impact of VFG-SSC's temporal regularization.

Conclusion

In this paper, we have introduced a new approach for semi-supervised 3D semantic scene completion. Our method effectively utilizes 3D clues derived from 2D depth and segmentation foundation models and introduces an innovative fusion module to integrate these clues with 3D features, ensuring compatibility with most existing SSC techniques. Extensive experiments have demonstrated the effectiveness of our approach in both outdoor and indoor environments, achieving a minimal performance gap of just 15% compared to fully supervised methods while using only 10% of 3D semantic occupancy annotations. To the best of our knowledge, this is the first exploration of semi-supervised 3D SSC.

Limitations. Given the reliance on using 2D VFMs, our approach might not work well under challenging conditions (e.g., night-time, foggy, rainy) where VFMs fail to produce good results. Furthermore, large domain gaps between training and test data might harm our method, as our approach uses a small amount of labeled data to train.

Acknowledgements

We acknowledge Ho Chi Minh City University of Technology (HCMUT), VNU-HCM for supporting this study.

Most of the work was done at VinAI Research.

References

Behley, J.; Garbade, M.; Milioto, A.; Quenzel, J.; Behnke, S.; Stachniss, C.; and Gall, J. 2019. Semantickitti: A dataset for semantic scene understanding of lidar sequences. In *ICCV*.

Cao, A.-Q.; and de Charette, R. 2022. Monoscene: Monocular 3d semantic scene completion. In *CVPR*.

Cao, A.-Q.; and de Charette, R. 2023. Scenerf: Self-supervised monocular 3d scene reconstruction with radiance fields. In *ICCV*.

Chen, X.; Yuan, Y.; Zeng, G.; and Wang, J. 2021. Semi-supervised semantic segmentation with cross pseudo supervision. In *CVPR*.

Cheng, B.; Misra, I.; Schwing, A. G.; Kirillov, A.; and Girdhar, R. 2022. Masked-attention mask transformer for universal image segmentation. In *CVPR*.

Cheng, R.; Agia, C.; Ren, Y.; Li, X.; and Bingbing, L. 2021. S3cnet: A sparse semantic scene completion network for lidar point clouds. In *Conference on Robot Learning*.

Dourado, A.; De Campos, T. E.; Kim, H.; and Hilton, A. 2021. EdgeNet: Semantic scene completion from a single RGB-D image. In *ICPR*.

Hassani, A.; Walton, S.; Li, J.; Li, S.; and Shi, H. 2023. Neighborhood Attention Transformer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 6185–6194.

Hoyer, L.; Tan, D. J.; Naeem, M. F.; Van Gool, L.; and Tombari, F. 2023. SemiVL: Semi-Supervised Semantic Segmentation with Vision-Language Guidance. *arXiv preprint arXiv:2311.16241*.

Huang, Y.; Zheng, W.; Zhang, B.; Zhou, J.; and Lu, J. 2023a. Selfocc: Self-supervised vision-based 3d occupancy prediction. *arXiv preprint arXiv:2311.12754*.

Huang, Y.; Zheng, W.; Zhang, Y.; Zhou, J.; and Lu, J. 2023b. Tri-perspective view for vision-based 3d semantic occupancy prediction. In *CVPR*.

Jia, Y.; He, J.; Chen, R.; Zhao, F.; and Luo, H. 2023. OccupancyDETR: Making Semantic Scene Completion as Straightforward as Object Detection. *arXiv preprint arXiv:2309.08504*.

Kirillov, A.; Mintun, E.; Ravi, N.; Mao, H.; Rolland, C.; Gustafson, L.; Xiao, T.; Whitehead, S.; Berg, A. C.; Lo, W.-Y.; et al. 2023. Segment anything. *arXiv preprint arXiv:2304.02643*.

Kong, L.; Ren, J.; Pan, L.; and Liu, Z. 2023. Lasermix for semi-supervised lidar semantic segmentation. In *CVPR*.

Kwon, D.; and Kwak, S. 2022. Semi-supervised semantic segmentation with error localization network. In *CVPR*.

Lee, D.-H.; et al. 2013. Pseudo-label: The simple and efficient semi-supervised learning method for deep neural networks. In *Workshop on challenges in representation learning, ICML*, volume 3, 896. Atlanta.

Li, J.; Han, K.; Wang, P.; Liu, Y.; and Yuan, X. 2020. Anisotropic convolutional networks for 3d semantic scene completion. In *CVPR*.

Li, Y.; Yu, Z.; Choy, C.; Xiao, C.; Alvarez, J. M.; Fidler, S.; Feng, C.; and Anandkumar, A. 2023. Voxformer: Sparse voxel transformer for camera-based 3d semantic scene completion. In *CVPR*.

Liang, T.; Xie, H.; Yu, K.; Xia, Z.; Lin, Z.; Wang, Y.; Tang, T.; Wang, B.; and Tang, Z. 2022. Bevfusion: A simple and robust lidar-camera fusion framework. *Advances in Neural Information Processing Systems*, 35: 10421–10434.

Liu, S.; Hu, Y.; Zeng, Y.; Tang, Q.; Jin, B.; Han, Y.; and Li, X. 2018. See and think: Disentangling semantic scene completion. *NeurIPS*, 31.

Liu, Y.; Tian, Y.; Chen, Y.; Liu, F.; Belagiannis, V.; and Carneiro, G. 2022. Perturbed and strict mean teachers for semi-supervised semantic segmentation. In *CVPR*.

Mei, J.; Yang, Y.; Wang, M.; Huang, T.; Yang, X.; and Liu, Y. 2023. SSC-RS: Elevate LiDAR Semantic Scene Completion with Representation Separation and BEV Fusion. In *IROS*.

Miao, R.; Liu, W.; Chen, M.; Gong, Z.; Xu, W.; Hu, C.; and Zhou, S. 2023. Occdepth: A depth-aware method for 3d semantic scene completion. *arXiv preprint arXiv:2302.13540*.

Olsson, V.; Tranheden, W.; Pinto, J.; and Svensson, L. 2021. Classmix: Segmentation-based data augmentation for semi-supervised learning. In *WACV*.

Park, J.; Xu, C.; Zhou, Y.; Tomizuka, M.; and Zhan, W. 2022. Detmatch: Two teachers are better than one for joint 2d and 3d semi-supervised object detection. In *ECCV*.

Piccinelli, L.; Yang, Y.-H.; Sakaridis, C.; Segu, M.; Li, S.; Van Gool, L.; and Yu, F. 2024. UniDepth: Universal Monocular Metric Depth Estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 10106–10116.

Roldao, L.; de Charette, R.; and Verroust-Blondet, A. 2020. Lmscnet: Lightweight multiscale 3d semantic completion. In *3DV*.

Silberman, N.; Hoiem, D.; Kohli, P.; and Fergus, R. 2012. Indoor segmentation and support inference from rgbd images. In *ECCV*.

Song, S.; Yu, F.; Zeng, A.; Chang, A. X.; Savva, M.; and Funkhouser, T. 2017. Semantic scene completion from a single depth image. In *CVPR*.

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

Tong, W.; Sima, C.; Wang, T.; Chen, L.; Wu, S.; Deng, H.; Gu, Y.; Lu, L.; Luo, P.; Lin, D.; et al. 2023. Scene as occupancy. In *ICCV*.

Wang, H.; Cong, Y.; Litany, O.; Gao, Y.; and Guibas, L. J. 2021. 3dioumatch: Leveraging iou prediction for semi-supervised 3d object detection. In *CVPR*.

Wang, X.; Zhu, Z.; Xu, W.; Zhang, Y.; Wei, Y.; Chi, X.; Ye, Y.; Du, D.; Lu, J.; and Wang, X. 2023. OpenOccupancy: A Large Scale Benchmark for Surrounding Semantic Occupancy Perception. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 17850–17859.

Wang, Y.; Wang, H.; Shen, Y.; Fei, J.; Li, W.; Jin, G.; Wu, L.; Zhao, R.; and Le, X. 2022. Semi-supervised semantic segmentation using unreliable pseudo-labels. In *CVPR*.

Wu, S.-C.; Tateno, K.; Navab, N.; and Tombari, F. 2020. Scfusion: Real-time incremental scene reconstruction with semantic completion. In *3DV*.

Xia, Z.; Liu, Y.; Li, X.; Zhu, X.; Ma, Y.; Li, Y.; Hou, Y.; and Qiao, Y. 2023. SCPNet: Semantic Scene Completion on Point Cloud. In *CVPR*.

Xie, E.; Wang, W.; Yu, Z.; Anandkumar, A.; Alvarez, J. M.; and Luo, P. 2021. SegFormer: Simple and efficient design for semantic segmentation with transformers. *Advances in neural information processing systems*, 34: 12077–12090.

Yan, X.; Gao, J.; Li, J.; Zhang, R.; Li, Z.; Huang, R.; and Cui, S. 2021. Sparse single sweep lidar point cloud segmentation via learning contextual shape priors from scene completion. In *AAAI*.

Yang, L.; Qi, L.; Feng, L.; Zhang, W.; and Shi, Y. 2023. Revisiting weak-to-strong consistency in semi-supervised semantic segmentation. In *CVPR*.

Yao, J.; Li, C.; Sun, K.; Cai, Y.; Li, H.; Ouyang, W.; and Li, H. 2023. Ndc-scene: Boost monocular 3d semantic scene completion in normalized device coordinates space. In *ICCV*.

Yin, W.; Zhang, C.; Chen, H.; Cai, Z.; Yu, G.; Wang, K.; Chen, X.; and Shen, C. 2023. Metric3d: Towards zero-shot metric 3d prediction from a single image. In *ICCV*.

Yun, S.; Han, D.; Oh, S. J.; Chun, S.; Choe, J.; and Yoo, Y. 2019. Cutmix: Regularization strategy to train strong classifiers with localizable features. In *ICCV*.

Zhang, C.; Yan, J.; Wei, Y.; Li, J.; Liu, L.; Tang, Y.; Duan, Y.; and Lu, J. 2023a. Occnerf: Self-supervised multi-camera occupancy prediction with neural radiance fields. *arXiv preprint arXiv:2312.09243*.

Zhang, H.; Li, F.; Zou, X.; Liu, S.; Li, C.; Yang, J.; and Zhang, L. 2023b. A simple framework for open-vocabulary segmentation and detection. In *ICCV*.

Zhang, Y.; Zhu, Z.; and Du, D. 2023. OccFormer: Dual-path Transformer for Vision-based 3D Semantic Occupancy Prediction. *arXiv preprint arXiv:2304.05316*.

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

Zou, X.; Dou, Z.-Y.; Yang, J.; Gan, Z.; Li, L.; Li, C.; Dai, X.; Behl, H.; Wang, J.; Yuan, L.; et al. 2023. Generalized decoding for pixel, image, and language. In *CVPR*.

Zou, X.; Yang, J.; Zhang, H.; Li, F.; Li, L.; Wang, J.; Wang, L.; Gao, J.; and Lee, Y. J. 2024. Segment everything everywhere all at once. *Advances in Neural Information Processing Systems*, 36.