SparseFusion-Net: Sparse U-Net with Multi-Scale Fusion for 3D Panoptic Scene Completion

1st Thanh Phat Pham
Dept. of Electronics Engineering
Soongsil University
phatpham@soongsil.ac.kr

2nd Myungsik Yoo
Dept. of Electronics Engineering
Soongsil University
myoo@soongsil.ac.kr

***Abstract*—3D Panoptic Scene Completion (PSC) has recently emerged as a promising research direction in 3D computer vision, particularly for autonomous driving and robotic systems. Unlike conventional scene completion, which focuses solely on recovering the geometric structure of incomplete 3D environments, PSC not only reconstructs the full scene but also provides rich semantic and instance-level understanding. This enables a more comprehensive perception of the surrounding environment, which is critical for safe navigation and high-level decision-making. However, PSC remains an extremely challenging task due to the inherent sparsity and occlusion of real-world 3D data, as well as the need to effectively model both local geometric details and global context for accurate scene completion and object distinction. Inspired by that, we introduce a novel framework, named SparseFusion-Net, an encoder–decoder architecture over sparse voxels for large-scale 3D PSC. At the core of our model is Sparse Context Fusion block (SCF Block), which integrates information across multiple scales to aggregate global background context and local detailed information. This joint modeling not only enhances geometric completion but also improves semantic consistency and instance separation, leading to more accurate PSC. Experiments on the SemanticKITTI benchmark demonstrate that our framework outperforms state-of-the-art methods by a significant margin using only LiDAR input.**

***Index Terms*—3D panoptic scene completion, multi-scale fusion, autonomous driving**

I. Introduction

LiDAR scene completion plays a vital role in autonomous driving systems, aiming to reconstruct the full 3D geometry of real-world environments from sparse and occluded LiDAR point clouds. Due to the inherent incompleteness of LiDAR data, this task requires reasoning about missing structures while maintaining semantic consistency. Recently, PSC [1] has gained attention as it extends conventional scene completion by jointly predicting voxel-level semantics and instance-aware object masks, enabling the system to differentiate dynamic objects (e.g., vehicles, pedestrians) from static structures (e.g., roads, buildings) in the reconstructed scene.

In the field of PSC, the current state-of-the-art method, PaSCo [1], utilizes a 3D generative U-Net architecture enhanced with a panoptic-aware decoder. U-Net is built upon an encoder–decoder structures: the encoder progressively downsamples the input to extract hierarchical and multi-scale features, while the decoder upsamples these features to recover fine-grained spatial details and reconstruct the final scene output. However, standard skip connections in U-Net directly transfer low-level geometric details from encoder layers to decoder layers of the same spatial resolution without resolving the receptive field mismatch and semantic abstraction levels between them. This inconsistency gives rise to a semantic gap, referring to the difference in information representation between encoder and decoder features. Specifically, encoder features mainly encode local geometric cues, such as surface edges and object boundaries, while decoder features at the corresponding level represent high-level semantic and global contextual understanding derived from deeper layers, including object identity and spatial relationships within the scene. When these geometry-rich yet context-limited encoder features are directly fused into the decoder, the model receives information that is spatially precise but semantically shallow. This imbalance causes the decoder to overemphasize local details while losing global coherence, leading to incomplete reconstruction and inaccurate instance boundaries. For instance, when reconstructing two vehicles parked side by side, the model may blur or merge their contact boundaries, resulting in an inaccurate instance separation. At the same time, the overlapping body region is often reconstructed as a fragmented or incomplete shape, revealing the model's inability to recover coherent geometry under insufficient global context.

We observe that adjacent encoder layers in a U-Net architecture capture complementary aspects of the scene. The previous encoder layer focuses on local geometric structures and fine details, such as object boundaries and surface edges, due to its smaller receptive field. In contrast, the next encoder layer aggregates broader contextual and semantic information that describes the overall scene layout and object relationships. Leveraging both local and global cues is therefore essential to bridge the semantic gap, reconstruct missing regions, and maintain structural coherence in complex 3D environments.

To mitigate the limitation, we propose **SparseFusion-Net**, an efficient sparse 3D network for PSC task. To effectively process such sparse and irregular 3D data, our approach builds upon a sparse 3D generative U-Net backbone and incorporates a novel component called the **Sparse Context Fusion (SCF)** block, which explicitly fuses features from both previous (detail-preserving) and next (context-rich) encoder layers. This cross-scale interaction helps refine fine structures, strengthen semantic consistency, and improve overall panoptic completion performance.

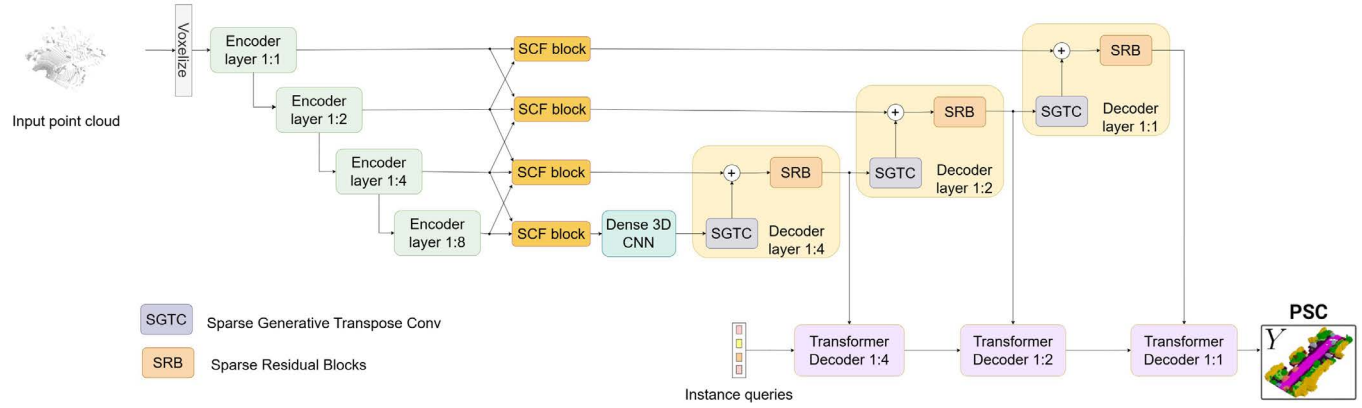

Fig. 1. Overall architecture of SparseFusion-Net Architecture. Best view in colors with zoom-in.

II. Methodology

A. Network Architecture

As illustrated in Fig. 1, SparseFusion-Net follows a sparse U-Net structure with four encoder stages and three decoder stages. The encoder extracts hierarchical geometric and semantic representations from the sparse input point cloud $X \in \mathbb{R}^{N\times 3}$, where N denotes the number of points.

To achieve this, the encoder begins by voxelizing the input point cloud into a sparse voxel grid, allowing efficient processing of large-scale 3D data without unnecessary computations on empty spaces. Each subsequent encoder stage applies sparse convolutions to downsample the features, progressively increasing the receptive field and capturing higher-level abstractions while preserving sparsity.

Each encoder layer is connected to an SCF block, which fuses the current-layer features with those from both the previous (detail-preserving, local detailed information) and next (context-rich, overall spatial layout of the entire scene) encoder layers. This design allows for richer cross-scale representation learning and enhances both reconstruction completeness and instance-level separation.

In the decoding stage, sparse generative transpose convolutions (SGTC) progressively recover spatial resolution, while sparse residual blocks (SRB) refine feature quality. Finally, the transformer-based decoder refines instance queries and produces a dense voxel representation $Y = (m_i, c_i)_{i=1}^{K}$, where each mask m_i corresponds to an instance or stuff region associated with a semantic label $c_i \in \{1, 2, \ldots, C\}$.

B. Sparse Context Fusion (SCF) block

The traditional Sparse U-Net structure [1], [3] directly fuses features between encoder and decoder layers of the same spatial resolution through skip connections. However, due to differences in the receptive fields across layers, this can cause semantic gaps in the fused features. To address this, we propose a novel Sparse Context Fusion block (SCF block) illustrated in Fig. 2, which incorporates multi-scale global context information to refine geometry and semantics, aiding in occlusion recovery and better object instance separation.

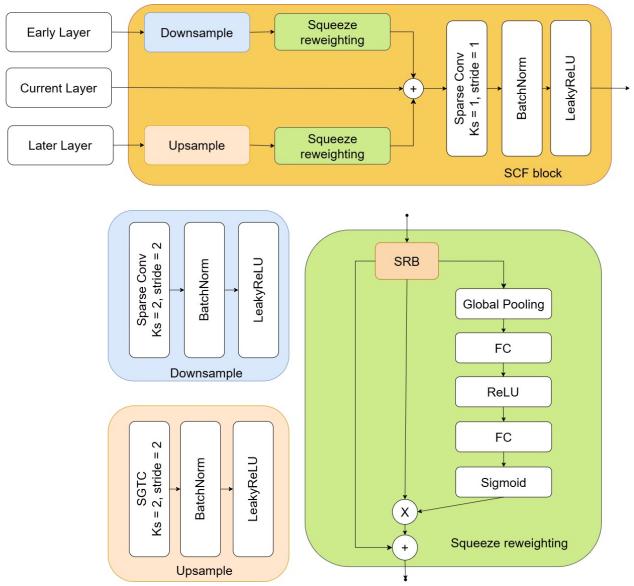

Fig. 2. Architecture of SCF Block. Best view in colors with zoom-in.

Before skip connection operation, we aggregate the feature from the current layer of the encoder denoted as F_c with the feature from the previous and next layer of the encoder denoted as F_p and F_n. Specifically, F_p is downsampled by Sparse Convolution (SC) layer with kernel size 2×2, followed by BatchNorm and LeakyReLU activations, formulated as $F'_p = \text{Down}(F_p)$, where Down() represents a Downsample block. F_n is upsampled by Sparse Generative Transpose Convolution (SGTC) layer with kernel size 2×2, followed by BatchNorm and LeakyReLu activations, formulated as $F'_n = \text{Up}(F_n)$, where Up() represents an Upsample block. Both F'_p and F'_n are refined using a squeeze re-weighting (SR) [7] for sparse tensor, designed to re-weight important voxel-wise features by modeling channel-wise dependencies in sparse voxels. The outputs are fused with the current layer feature F_c through summation:

$$F_{\text{fused}} = F_c + \text{SR}(F'_p) + \text{SR}(F'_n)$$

Finally, F_{fused} is processed via a 1×1 Sparse Convolution followed by BatchNorm and LeakyReLU activations to generate the output of the SCF block.

III. Experiment and Results

In this section, we present the experimental setup of the proposed SparseFusion-Net and compare it with state-of-the-art methods on the PSC benchmark derived from the SemanticKITTI dataset [4], following the evaluation introduced in PaSCo [1]. Although SemanticKITTI does not provide PSC labels by default, this benchmark extends its semantic labels with instance-level information for comprehensive panoptic evaluation.

A. Dataset

SemanticKITTI dataset has 11 sequences (sequences 00-07 and 09-10 for training, 08 for validation) with 19 semantic categories (8 thing classes and 11 stuff classes). To adapt the SemanticKITTI for the PSC task, we use DBSCAN [5], [6] for clustering object instances from ad-hoc classes with a distance threshold of $\epsilon = 1$ and a minimum points parameter $MinPts = 8$, extracting PSC labels following PaSCO [1].

B. Implementation details

We train SparseFusion-Net for 30 epochs on SemanticKITTI by a NVIDIA RTX-3090 GPU, using AdamW [2] optimizer, the learning rate is 10^{-4}, and batch size of 1. Also, for augmentation, we apply random rotations in [-30°, 30°] on SemanticKITTI dataset, random cropping to reduce the scene size to 80% along both the x and y axes, and random translations of ± 0.6m on the x/y axes and ± 0.4m on z axis.

For training loss, we use voxel-query semantic loss ($\mathcal{L}_{\text{voxel}}$), semantic loss ($\mathcal{L}_{\text{sem}}$), and masks matching loss ($\mathcal{L}_{\text{matched}}$) from PaSCo [1].

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{voxel}} + \mathcal{L}_{\text{sem}} + \mathcal{L}_{\text{matched}} \tag{1}$$

C. Results

We compared the performance of SparseFusion-Net with the existing method PaSCo on the SemanticKITTI validation set, as summarized in Table I. SparseFusion-Net consistently achieves superior results in all major panoptic metrics, including PQ†, PQ, SQ, and RQ. In particular, it improves upon the previous state-of-the-art [1] by +3.12/+1.6 on All-PQ†/PQ, together with a +3.41 gain in SQ, indicating more accurate segmentation, and a +2.36 increase in RQ, reflecting enhanced recognition quality. Furthermore, SparseFusion-Net attains a +1.27 gain on the auxiliary mIoU metric, indicating that it not only maintains semantic completeness but also achieves finer voxel-level discrimination. These results demonstrate that SparseFusion-Net enhances both geometric reconstruction and semantic consistency, resulting in more accurate and coherent 3D panoptic scene understanding.

TABLE I
Performance metric for PSC on SemanticKITTI Validation set.

Model	PQ†	PQ	SQ	RQ	mIoU
PaSCo*	23.08	12.12	49.8	19.29	26.47
SparseFusion-Net	**26.20**	**13.72**	**53.21**	**21.65**	**27.74**

* denotes our own re-implementation of PaSCo for a fair comparison

Acknowledgment

This work was supported by the National Research Foundation of Korea(NRF) grant funded by the Korea government(MSIT)(RS-2024-00335137)

References

[1] Cao, A.Q., Dai, A., de Charette, R., 2024. Pasco: Urban 3d panoptic scene completion with uncertainty awareness, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 14554–14564.

[2] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In ICLR, 2019.

[3] Luis Rolda o, Raoul de Charette, and Anne Verroust-Blondet. Lmscnet: Lightweight multiscale 3d semantic completion. In 3DV, 2020.

[4] J. Behley, M. Garbade, A. Milioto, J. Quenzel, S. Behnke, C. Stachniss, and J. Gall, "Semantickitti: A dataset for semantic scene understanding of lidar sequences," in Proceedings of the IEEE/CVF International Conference on Computer Vision, 2019, pp. 9297–9307.

[5] Hui Chen, Man Liang, Wanquan Liu, Weina Wang, and Peter Xiaoping Liu. An approach to boundary detection for 3d point clouds based on dbscan clustering. Pattern Recognition, 2022.

[6] Martin Ester, Hans-Peter Kriegel, Jörg Sander, Xiaowei Xu, et al. A density-based algorithm for discovering clusters in large spatial databases with noise. In KDD, 1996.

[7] Y. Wang, T. Shi, P. Yun, L. Tai, and M. Liu. Pointseg: Real-time semantic segmentation based on 3d lidar point cloud. arXiv preprint arXiv:1807.06288, 2018.

[8] Maxim Berman, Amal Rannen Triki, and Matthew B. Blaschko. The lov asz-softmax loss: A tractable surrogate for the optimization of the intersection-over-union measure in neural networks. In CVPR, 2018.