SegmentFusion: Hierarchical Context Fusion for Robust 3D Semantic Segmentation (Supplementary Material)

Anirud Thyagharajan[1], Benjamin Ummenhofer[2], Prashant Laddha[1], Om Ji Omer[1], Sreenivas Subramoney[1]
{[1]Processor Architecture Research Lab India, [2]Autonomous Agents Lab Germany}, Intel Labs

1. Network Architecture

Below is an example of the network architecture. The network architecture is composed of a stack ($N = 2$) of encoder blocks, followed by a final fully connected layer. Each encoder is composed of a scaled dot product attention block, (Figure 1 shows the first of such blocks). u and v capture segmentwise features in the projected space, and the attention $S(u, v)$ computes the interactions of features across segments. We constrain the interactions to flow only between spatially connected segments, which is performed by employing a Hadamard product with the adjacency matrix $\mathbf{A}$. This ensures that the error gradient flows through only relevant edges. To maintain regular gradient flow across the N blocks, we employ skip connections and propagate features onto the next encoder block. The number of hidden nodes K in each FC layer is 256, with group norm parameter as a group of 8 channels.

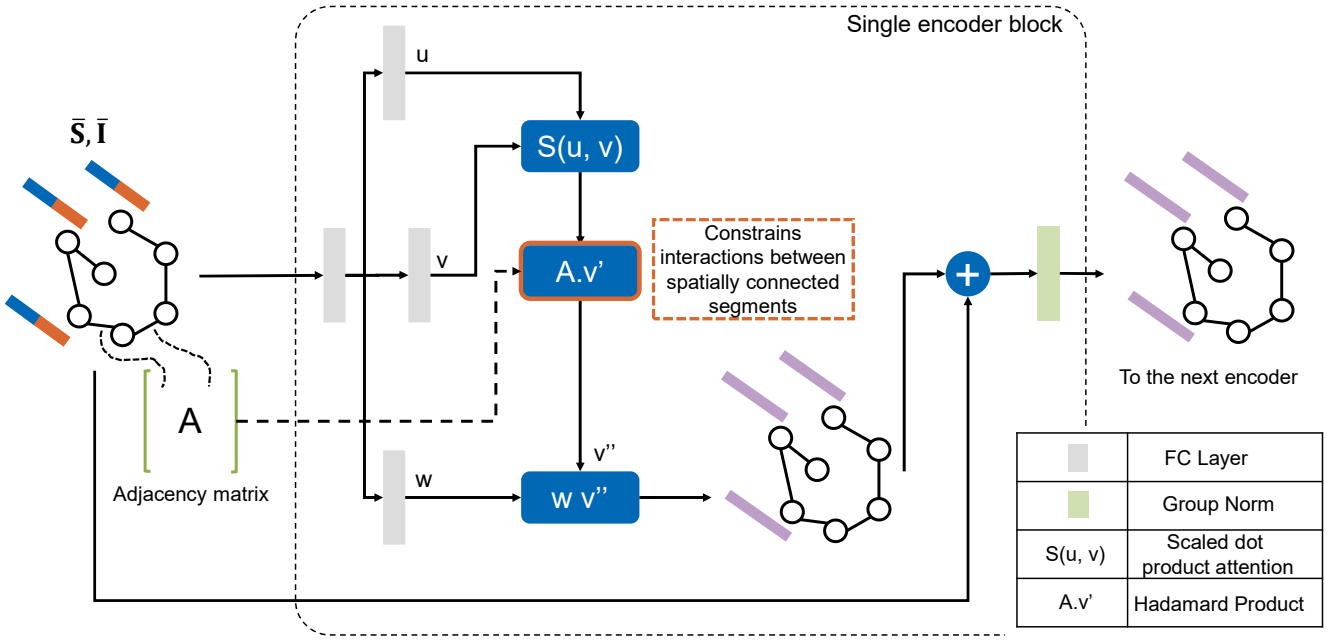

Figure 1. Description of the components of the first encoder block in the Segment-Fusion network architecture.

2. Training Setup

As part of data augmentation, we introduced random dropping of nodes in the graph, along with randomly resetting some spatial connections (by setting entries of $\mathbf{A}$ to 0) to make the fusion decisions more robust. The proposed Segment-Fusion network is extremely light-weight (consists of only about $0.5M$ parameters for the above $N = 2, K = 256$ configuration.) and can be trained on either a GPU or CPU. We trained the network on a Intel(R) i7-8700K CPU with 16 GB RAM.

3. Visualizations

Figures 2,3 and 4 illustrate additional visualizations of the impact of using Segment-Fusion on previous semantic estimators (MinkowskiNet [1], PointConv [4], SparseConvNet [2]) on the ScanNet dataset.

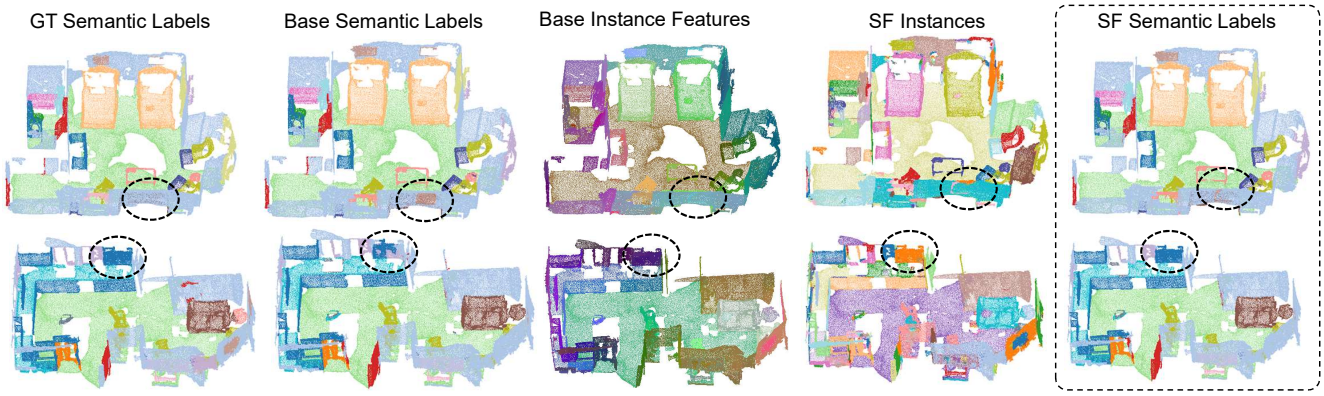

Figure 2. Qualitative results of Segment-Fusion on some sample point clouds of the ScanNet validation set using the MinkowskiNet semantic backbone.

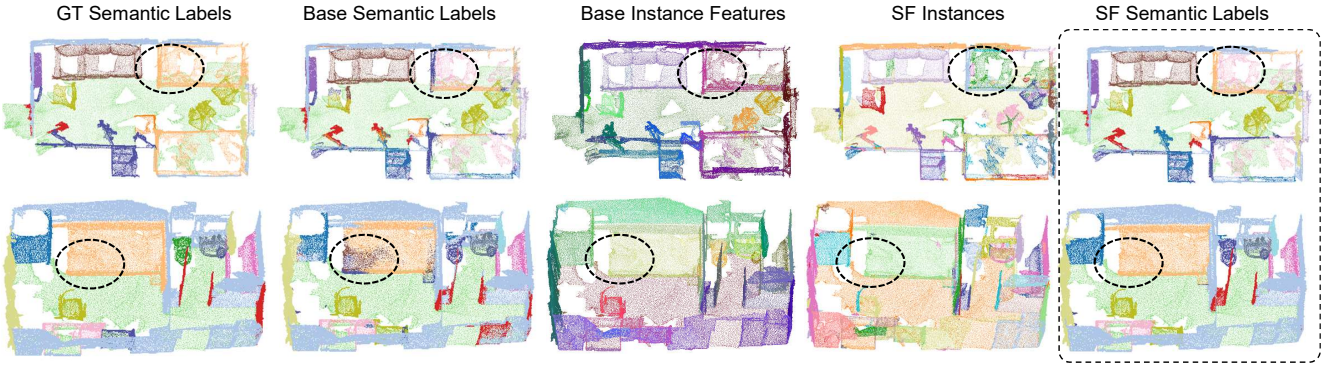

Figure 3. Qualitative results of Segment-Fusion on some sample point clouds of the ScanNet validation set using the PointConv semantic backbone.

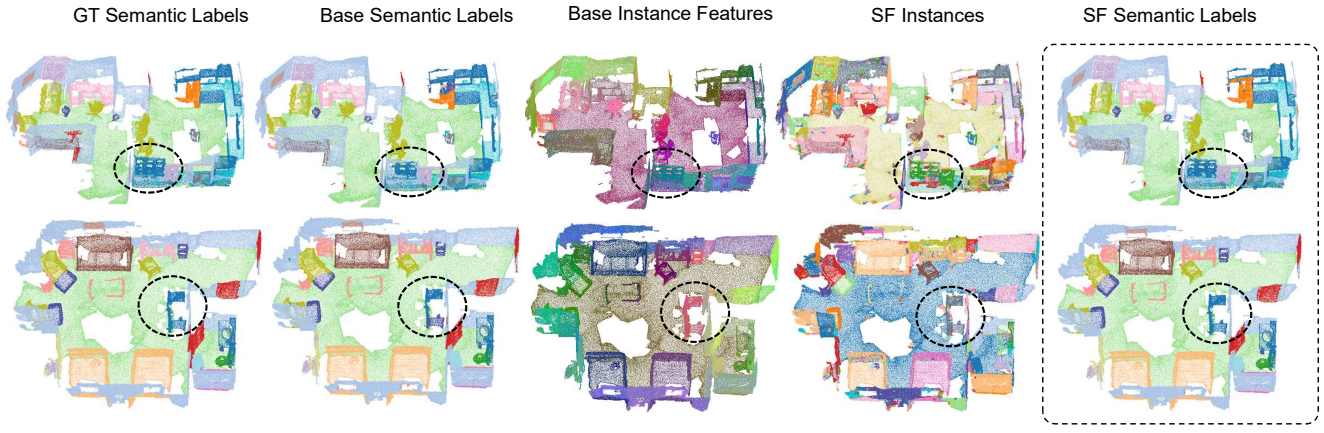

Figure 4. Qualitative results of Segment-Fusion on some sample point clouds of the ScanNet validation set using the SparseConvNet semantic backbone.

Table 1. Performance impact (mIoU) of Segment-Fusion on state-of-the-art semantic segmentation backbones on the ScanNet.

Model	Set	wall	floor	cabinet	bed	chair	sofa	table	door	window	bookshelf	picture	counter	desk	curtain	fridge	shower	toilet	sink	bathtub	furniture	**mean**
SparseConvNet	val	82.4	94.8	58.0	77.2	88.5	78.4	68.1	58.3	58.7	70.1	29.4	61.9	56.7	62.3	46.6	61.5	91.2	63.7	84.0	50.3	67.1
SparseConvNet + SF	val	85.3	97.1	61.9	79.5	91.6	82.1	72.3	61.4	61.0	71.6	36.1	71.0	61.7	65.8	47.1	69.9	95.5	73.8	92.3	54.0	**71.5**
PointConv	val	74.1	94.7	46.9	70.0	83.0	70.0	64.9	32.1	46.7	68.4	11.7	56.6	52.4	58.2	36.6	46.8	83.2	58.0	77.3	34.6	58.3
PointConv + SF	val	78.6	97.3	50.7	78.0	87.6	76.2	68.0	36.0	49.3	75.1	13.3	64.2	59.9	62.9	38.1	54.1	89.9	63.6	90.5	38.2	**63.6**
MinkNet42	val	84.3	95.1	63.3	78.9	91.5	87.7	74.4	60.2	65.0	80.1	24.9	65.1	65.8	78.1	55.4	69.9	92.2	69.1	86.4	61.0	72.4
MinkNet42 + SF	val	86.7	97.3	65.9	80.0	93.9	90.0	77.0	63.6	66.4	82.9	26.5	68.9	68.7	80.7	55.7	72.8	95.4	76.0	91.7	64.5	**75.2**

4. ScanNet Validation Set Evaluation

In Table 1, we present results of using Segment-Fusion on a variety of semantic backbones (MinkowskiNet [1], PointConv [4], SparseConvNet [2]) on the ScanNet validation set, supplemented with an instance backbone trained with the losses proposed in Occuseg [3]. Similar to the test set, we observe significant improvements of 4.4%, 5.1% and 2.8% respectively in mIoU scores (and gains in individual class scores). Thus, we observe that the improvements provided by Segment-Fusion do not depend specifically on the choice of the semantic segmentation backbone used.

References

[1] Christopher Choy, JunYoung Gwak, and Silvio Savarese. 4d spatio-temporal convnets: Minkowski convolutional neural networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3075–3084, 2019. 2, 3

[2] Benjamin Graham, Martin Engelcke, and Laurens Van Der Maaten. 3d semantic segmentation with submanifold sparse convolutional networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 9224–9232, 2018. 2, 3

[3] Lei Han, Tian Zheng, Lan Xu, and Lu Fang. Occuseg: Occupancy-aware 3d instance segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 2940–2949, 2020. 3

[4] Wenxuan Wu, Zhongang Qi, and Li Fuxin. Pointconv: Deep convolutional networks on 3d point clouds. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9621–9630, 2019. 2, 3