SemAffiNet: Semantic-Affine Transformation for Point Cloud Segmentation
Supplementary Material

In this supplementary material, we provide the detailed network architecture of SemAffiNet in Section A. Then in Section B, we show additional experiments information, including the datasets introduction, implementation details and class-wise 3D point cloud segmentation mIoU on the ScanNetV2 [4] dataset.

A. Network Architecture

A.1. The Backbone

We choose the BPNet [6] as our backbone, which consists of a 3D sparse convolution encoder-decoder branch and a 2D vanilla convolution encoder-decoder branch. The 3D branch implements the sparse convolution MinkowskiEngine [2] to build a ResNet-style architecture [5], while the 2D branch exploits the ResNet34 model as the encoder. The decoder blocks of these two branches are composed of residual blocks with interpolation operation for the 2D branch and transpose convolution for the 3D branch. Skip connections are added between the corresponding layers of encoders and decoders to pass low-level information. The bidirectional links between layers of the 2D branch and the 3D branch from the BPNet are kept to communicate knowledge between the two modalities. Please refer to the BPNet paper for further details about the bidirectional projections between 2D and 3D features.

A.2. The Implicit Semantic-Aware Module

The ISAM is a shared Transformer [10] encoder module that consists of $N_e = 6$ encoder blocks to fuse 2D and 3D multi-modal information. Each encoder block consists of a Multi-Head Attention layer with 8 heads, two LayerNorm layers and a Feed Forward layer. The input to the Transformer encoder is the concatenation of the high-level features from both 2D and 3D branches.

A.3. The Explicit Semantic-Aware Module

We design two similar explicit semantic-aware modules (ESAM) for 2D and 3D branches respectively. We wrap the learning parameters of ESAM into a Transformer [10] decoder module, predicting semantic-affine parameters and semantic class masks. Then semantic-affine transformations are then applied to multiple mid-level layers of the backbone decoder.

The **Transformer decoder** consists of $N_d = 6$ decoder blocks. Each decoder block consists of a Multi-Head Attention layer with 8 heads, three LayerNorm layers, a Multi-Head Cross Attention layer with 8 heads and a Feed Forward layer. The input to the Transformer decoder is the output of the ISAM in the corresponding modality and the learnable class queries of shape (N, d_h), where N is the number of classes and $d_h = 128$ is the hidden dimension. The position embeddings for the 2D branch are learnable embeddings with d_h dimensions. We obtain the position embeddings for the 3D branch by feeding the coordinates of 3D point clouds to an MLP block with 3 linear layers.

The output features h^u of shape (N, d_h) from the Transformer decoder layer u are prepared for predicting semantic-affine parameters and the output feature h^{-1} from the final layer of the Transformer decoder are used for predicting semantic class masks. Firstly, the output features of the Transformer decoder h^{-1} are fed through an MLP block with 3 linear layers to get the class masks M of shape (N, d_m), where $d_m = 128$ is the class mask dimension. Secondly, the intermediate output features h^u of the Transformer decoder layer u are fed through two separate MLP blocks with 5 linear layers to obtain class-specific affine parameters $\mathbf{s}^i, \mathbf{b}^i$ for the backbone decoder layer i, where the correspondence between i and u is $u = i + 2$.

The **Semantic-Affine Transformation** is implemented to 3 mid-level layers of the backbone decoders. Once obtaining the backbone decoder feature f^i at layer i, we can calculate the per-point or per-pixel classification prediction confidence as $A^i = Mf^i$. Then the semantic-affine transformation parameters for each point S^i, B^i can be calculated via Equation(1) in our main paper according to $\mathbf{s}^i, \mathbf{b}^i$ and A^i. Then we transform the normalized feature $\hat{f}^i_j$ of the backbone decoders with Equation(2) in our main paper to enhance semantic information of mid-level features.

B. Additional Experiments Information

B.1. Datasets Introduction

We evaluate our SemAffiNet with both point cloud semantic segmentation task on the ScanNetV2 dataset and RGB-D image segmentation task on the RGB-D NYUv2 dataset. We also verify the generalization ability of the proposed semantic-affine transformation on pure 3D S3DIS dataset and pure 2D Cityscapes dataset. In this subsection, we will introduce these datasets in detail.

ScanNetV2 [4] is one of the most commonly used indoor

Table 1. The class-wise segmentation results on the 3D point cloud segmentation task of the ScanNetV2 [4] dataset. We compare the proposed SemAffiNet with the BPNet [6] baseline under both 5cm and 2cm settings.

Method	mIoU	bath	bed	bkshf	cab	chair	cntr	curt	desk	door	floor	other	pic	fridge	shower	sink	sofa	table	toilet	wall	window
BPNet (5cm)	70.6	85.6	81.6	79.8	68.7	89.9	66.3	60.0	69.2	58.6	94.6	58.0	20.9	54.7	64.9	68.4	79.7	76.6	91.4	82.9	59.4
SemAffiNet (5cm)	72.1	85.9	80.8	82.7	69.9	90.7	65.6	66.3	71.8	61.7	94.5	56.9	26.0	52.0	71.4	66.0	82.3	76.8	92.6	83.8	65.0
BPNet (2cm)	72.5	86.7	79.5	80.1	66.9	90.8	62.3	74.9	69.3	63.3	95.0	56.3	34.1	55.6	71.5	65.9	83.1	73.7	92.6	84.9	63.5
SemAffiNet (2cm)	74.5	88.5	82.1	81.6	69.9	91.6	67.2	79.2	70.0	67.3	95.3	58.0	32.6	58.3	70.9	70.5	83.1	77.4	94.5	85.9	65.3

scenes datasets. It consists of over 1500 scans of indoor scenes annotated with 20 commonly seen semantic classes. The RGB-D video dataset provides both 2D image-level color data and 3D point-cloud geometry data, thus being robust and comprehensive for 3D scene understanding. We follow the official split to train on 1201 training scans and test on 312 validation scans.

S3DIS [1] is another indoor scene dataset. It contains dense 3D point clouds extracted from 6 large-scale areas scanned from 271 rooms in 3 buildings and is annotated by 13 semantic classes. We follow the common protocol to split Area 5 as the test set and use other Areas for training.

NYUv2 [7] is a popular RGB-D dataset that focuses on 2D image segmentation. It consists of 1,449 pairs of aligned RGB and depth images and we follow the official split to train on 795 samples and leave 654 for testing. We convert the depth image to pseudo point clouds according to camera pose, and the problem is transformed into 3D-2D multi-modality segmentation as ScanNetV2.

Cityscapes [3] is an outdoor 2D street-scene dataset that contains 2975 train images, 500 validation images and 1525 test images. The images have 1024*2048 resolutions and there are 19 classes. We do not use additional 20k images with coarse annotations.

B.2. Implementation Details

We implement our proposed SemAffiNet architecture with PyTorch [8] and utilize SGD optimizer [9] with base learning rate $2e^{-2}$ and weight decay $1e^{-4}$. The learning rate of parameters in ISAM and ESAM is reduced by a factor of 0.1 for more stable training. We implement a squared learning rate scheduler with a warming up process. We train the model for 100 epochs with batch size 16. The loss weights for vanilla 3D and 2D segmentation Cross Entropy loss are 1 and 0.1 respectively, following BPNet settings. The loss weights for the auxiliary mid-level Binary Cross Entropy loss for 3D and 2D branches are kept at 1 for our experiments.

B.3. Class-wise Segmentation Results

The class-wise segmentation results on the 3D point cloud segmentation task of ScanNetV2 [4] validation set are shown in Table 1. We compare the proposed SemAffiNet with the BPNet baseline under both 5cm and 2cm settings. From the class-wise segmentation results, we can conclude that our SemAffiNet achieves a higher IoU on most categories and a higher mIoU than the BPNet baseline.

References

[1] Iro Armeni, Ozan Sener, Amir R Zamir, Helen Jiang, Ioannis Brilakis, Martin Fischer, and Silvio Savarese. 3d semantic parsing of large-scale indoor spaces. In *CVPR*, pages 1534–1543, 2016. 2

[2] Christopher Choy, JunYoung Gwak, and Silvio Savarese. 4d spatio-temporal convnets: Minkowski convolutional neural networks. In *CVPR*, pages 3075–3084, 2019. 1

[3] Marius Cordts, Mohamed Omran, Sebastian Ramos, Timo Rehfeld, Markus Enzweiler, Rodrigo Benenson, Uwe Franke, Stefan Roth, and Bernt Schiele. The cityscapes dataset for semantic urban scene understanding. In *CVPR*, 2016. 2

[4] Angela Dai, Angel X Chang, Manolis Savva, Maciej Halber, Thomas Funkhouser, and Matthias Nießner. Scannet: Richly-annotated 3d reconstructions of indoor scenes. In *CVPR*, pages 5828–5839, 2017. 1, 2

[5] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, pages 770–778, 2016. 1

[6] Wenbo Hu, Hengshuang Zhao, Li Jiang, Jiaya Jia, and Tien-Tsin Wong. Bidirectional projection network for cross dimension scene understanding. In *CVPR*, pages 14373–14382, 2021. 1, 2

[7] Pushmeet Kohli Nathan Silberman, Derek Hoiem and Rob Fergus. Indoor segmentation and support inference from rgbd images. In *ECCV*, 2012. 2

[8] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. Pytorch: An imperative style, high-performance deep learning library. *NeurIPS*, 32:8026–8037, 2019. 2

[9] Sebastian Ruder. An overview of gradient descent optimization algorithms. *arXiv preprint arXiv:1609.04747*, 2016. 2

[10] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *NeurIPS*, pages 5998–6008, 2017. 1