# 3D INDOOR SCENE LONG TAIL SEGMENTATION

Chia-Wen Huang, Ying-Hua Huang,  Tsu-Hsien Shih, Po-Yu Chen

## ABSTRACT

Semantic segmentation of 3D point clouds is one of the most challenging problems in recent computer vision research. ScanNet200, a 200-class 3D indoor scene semantic segmentation benchmark, contains rare categories and a lot of imbalanced data which represent the diversity of the real-world environment compared to other datasets.  In this work, we aim to deal with the data imbalance problem with text features anchor learning during pre-training, and both data-sampling-based re-balance and loss-based re-balance during fine-tuning.

## PREDICTION RESULTS

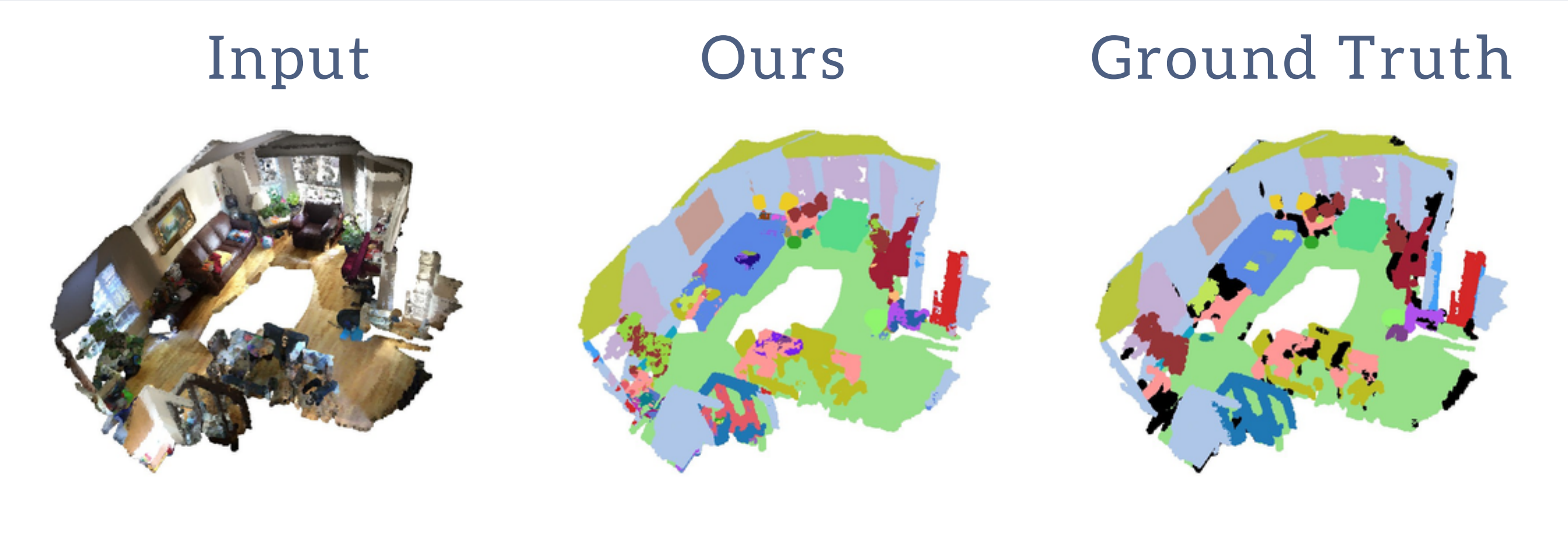

Fig. 1: The visualization of our results.

| | Test mIoU | Val mIoU | | | |
|---|---|---|---|---|---|
| | All | All | Head | Common | Tail |
| Cross Entropy | | | | | |
| Focal Loss | | | | | |
| Focal Loss + Augmentation | | | | | |

## CONCLUSION

In this work, we develop a two-stage training scheme to deal with the difficulties in the 3D semantic segmentation of ScanNet200 dataset. First, we leverage the state-of-art CLIP model in text-supervised contrastive learning, making the embedding learning for numerous classes more feasible. Then, we adopt several strategies to handle the data imbalance problem, such as instance augmentation and focal loss. The experiment result shows the effectiveness of these methods, and the visualization result reveals that the proposed method could generally reconstruct the outline of 3D instances.

## METHODOLOGY

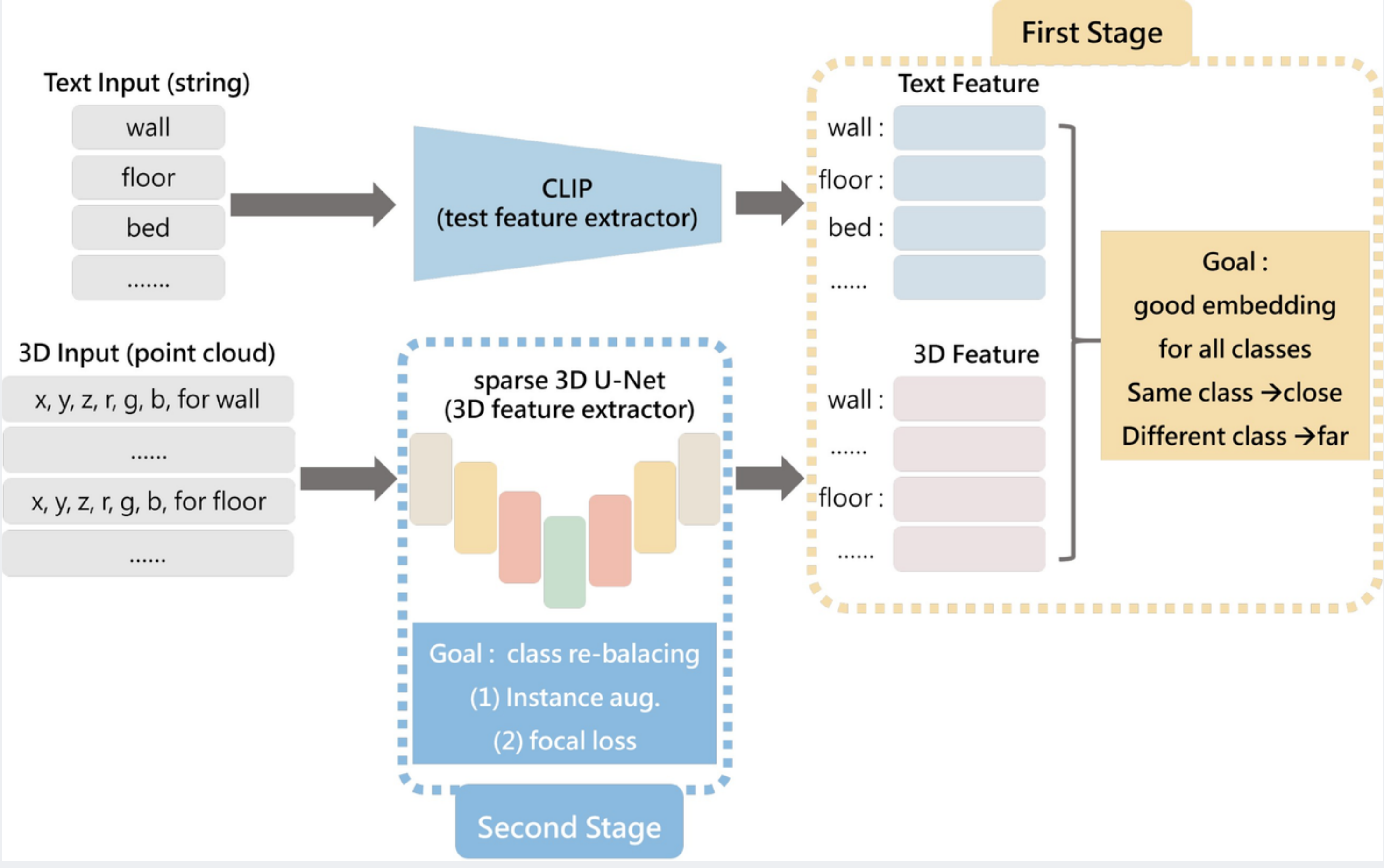

Fig. 2: Our framework

### ☃ Challenge 1 : Multi-class embedding – Text feature anchor learning

**(1) Text feature – CLIP**

We use pre-trained CLIP to get text features of the instance label. In the pre-training stage, we use contrastive loss to align geometric features and text features to get better initialization weights on the downstream semantic segmentation task.

**(2) 3D feature - sparse 3D U-Net**

We leverage a sparse 3D U-Net for 3D feature extraction by implementing the MinkowskiEngine, which benefits sparse convolutions in high dimensions. To be specific, the backbone is trained to map the output 3D feature into the text space during the pre-training phase, and then fine-tuned for our downstream semantic segmentation task.

### ☃ Challenge 2 : Class re-balancing

**(1) Focal loss and re-weighting factor**

Our re-weighting scheme is changing loss function from CE loss to focal loss, and applying weighting factor on focal loss. The weight factor is based on frequency of class, and therefore focusing on hard examples.

**(2) Instance augmentation**

We do instance augmentation on scene. First, we retrieve the tail instances for all scenes, adding them to scenes randomly to increase the occurrence of the tail instances.

Reference:
[1] Rozenberszki, D., Litany, O., & Dai, A. (2022). Language-Grounded Indoor 3D Semantic Segmentation in the Wild. 2022 Proceedings of the European Conference on Computer Vision (ECCV)
[2] Choy, C.B., Gwak, J., & Savarese, S. (2019). 4D Spatio-Temporal ConvNets: Minkowski Convolutional Neural Networks. 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 3070-3079.