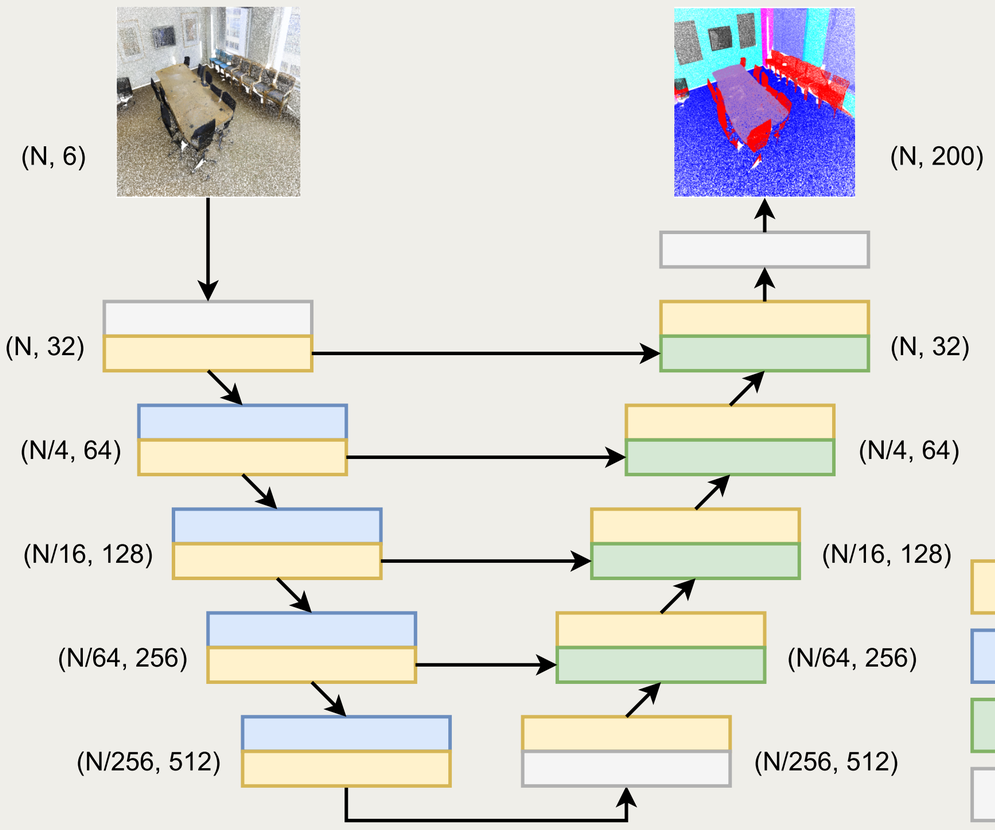
MIXFORMER3D

Group "TalkingtoMe":
李勝維 R11944004 張仲喆 R11922A15
魏湧致 R11944035 呂兆凱 R11922098

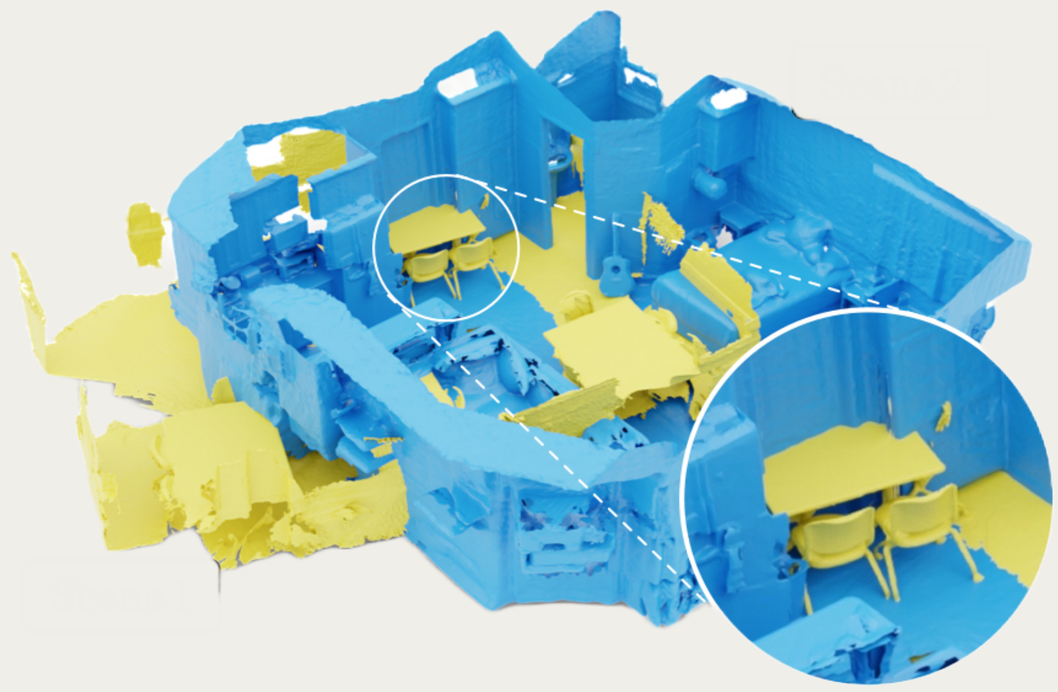
In this project, we adapt the well-known attention mechanism to 3D point cloud segmentation and demonstrate its effectiveness in combination with U-Net-inspired architecture and various long-tail recognition (LTR) techniques.

POINT TRANSFORMER



MIX3D: BREAKING CONTEXTUAL PRIORS

Mix3D is a data augmentation technique that increases generalization beyond the contextual priors of the training scenes and improves predictions for rare events not well captured by the training data.



DECOUPLING REPRESENTATION AND CLASSIFIER

The main idea:

- Data imbalance might not be an issue in learning high-quality representations.
- Strong long-tailed recognition ability can be achieved through **only** tuning the classifier.

Thus, we decouple the learning procedure into two stages:

1. **Representation learning**: Train the transformer (backbone) and classifier (linear head) using standard cross-entropy.
2. **Classification**: The classifier trained in previous stage is deprecated, and a new classifier is trained with frozen backbone, using class-balanced loss* and applying max-norm regularization.

※ Class-balanced loss: Any loss with class weight = $\beta^{\pi}, 0 < \beta < 1$, CE loss is provided as backbone in this instance.

LONG-TAIL LEARNING VIA LOGIT ADJUSTMENT

At inference time, the logit is adjusted using the following equation:

$$\arg\max_{y \in L} \exp(f_y(x)/\pi_y^{\tau}) = \arg\max_{y \in L} f'_y(x) - \tau \cdot \log(\pi_y)$$

where $\tau \cdot \log(\pi_y)$ is the class prior adjusted by the hyperparameter τ, which replaces regular weight normalization $(\text{view } \pi_y^{\tau} \text{ as } \|W_y\|_2)$

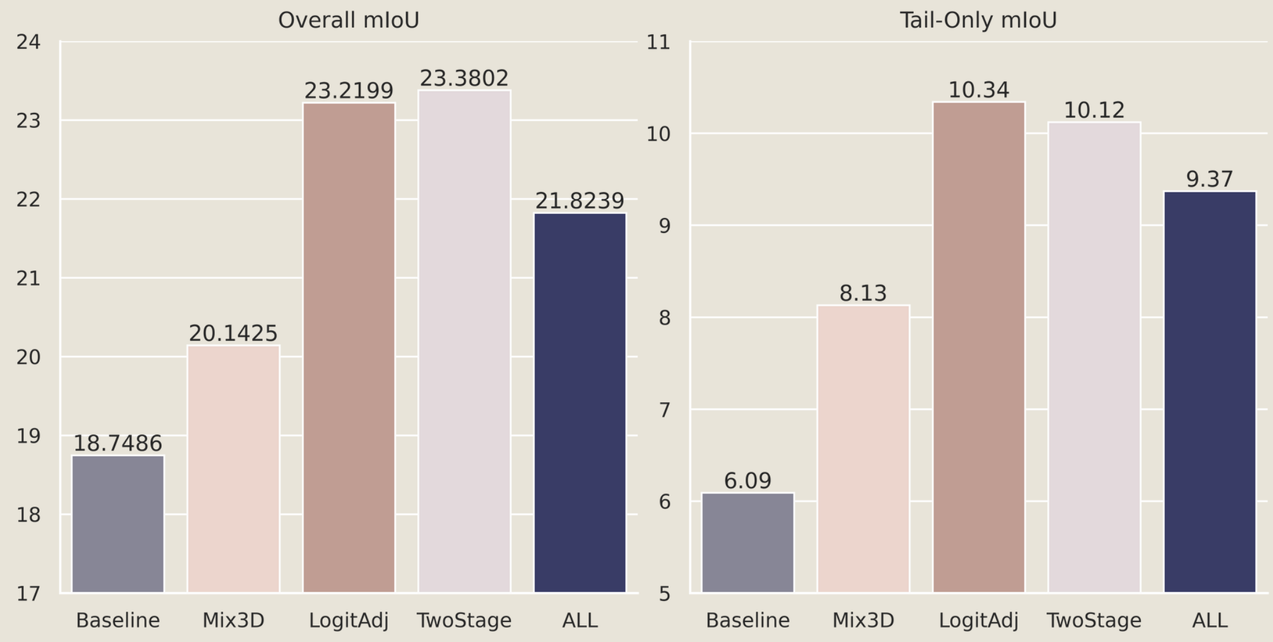
Logit adjustment is used to correct the probability distribution of the model's predictions and prevent an over-emphasis on the head classes at the expense of the tail classes.
In other words, Logit adjustment is virtually equivalent to weight normalization, but can result in more stable training as it is only applied during inference.

RESULTS AND ABLATION STUDY

Based on the Point Transformer architecture, we tested the following five settings on the testing set with their **overall** mIoU and **tail-only** mIoU:

- Baseline: The model was directly trained using cross-entropy.
- Mix3D: The Mix3D augmentation was applied to the baseline configuration.
- LogitAdj: The logit adjustment was applied to the Mix3D configuration.
- TwoStage: The decoupled two-stage training paradigm with Mix3D applied.
- ALL: Utilizes all the techniques aforementioned.

Finally, one model was trained under the LogitAdj configuration and two models were trained under the TwoStage configuration with different random seeds. These three models were then combined through voting ensemble. This resulted in mIoU score and a ranking of on the leaderboard.



CONCLUSION

In conclusion, the use of attention mechanisms in the point cloud semantic segmentation task has been shown to be effective. The Transformer model has demonstrated its ability to generalize well to other tasks.
Furthermore, we have applied various long-tail recognition methods that are typically used in the two-dimensional domain, such as decoupled training of representation and classification, class-balanced loss, and logit adjustment, to the three-dimensional point cloud domain and found them to be effective and transferable.

REFERENCE

1. H. Zhao, J. Li, J. Jia, P. Torr, and V. Koltun. Point Transformer. ICCV 2021.
2. A. Nekrasov, J. Schult, O. Litany, B. Leibe, and F. Engelmann. Mix3D: Out-of-Context Data Augmentation for 3D Scenes. 3DV 2021
3. Y. Cui, M. Jia, T. Lin, Y. Song, S. Belongie. Class-Balanced Loss Based on Effective Number of Samples. CVPR 2019
4. A. Menon, S. Jayasumana, A. Rawat, H. Jain, A. Veit, S. Kumar. Long-tail learning via logit adjustment. ICLR 2021.
5. S. Alshammari, Y. Wang, D. Ramanan, S. Kong. Long-Tailed Recognition via Weight Balancing. CVPR 2022