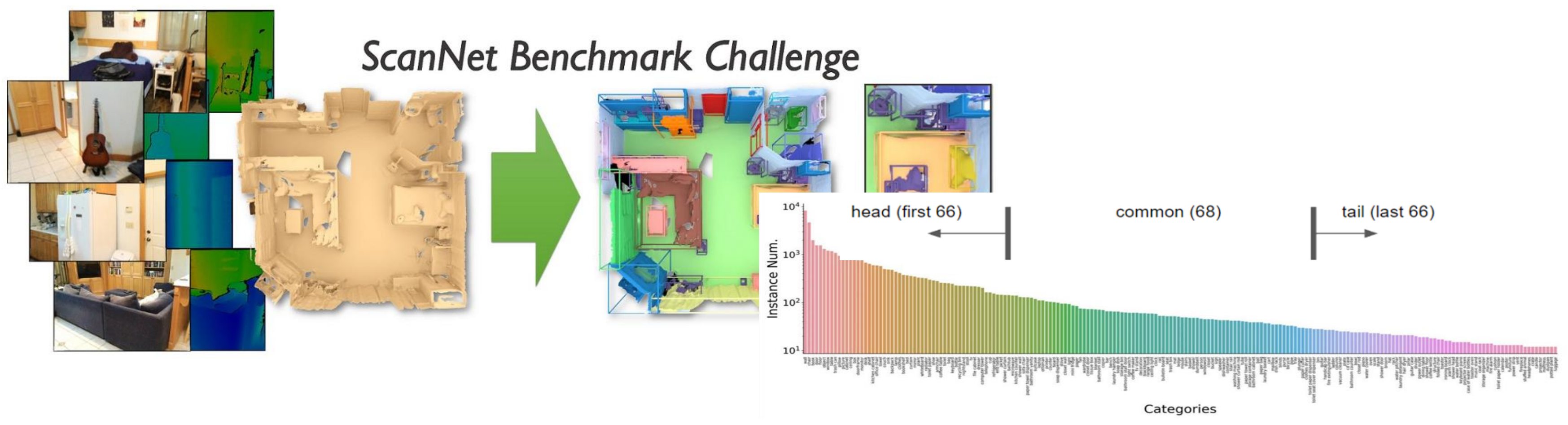
DLCV Fall 2022 Final Project

3D Indoor Scene Long Tail Segmentation

Team members: 陳烱濤、阮羿寧、羅恩至、溫威領、劉名凱

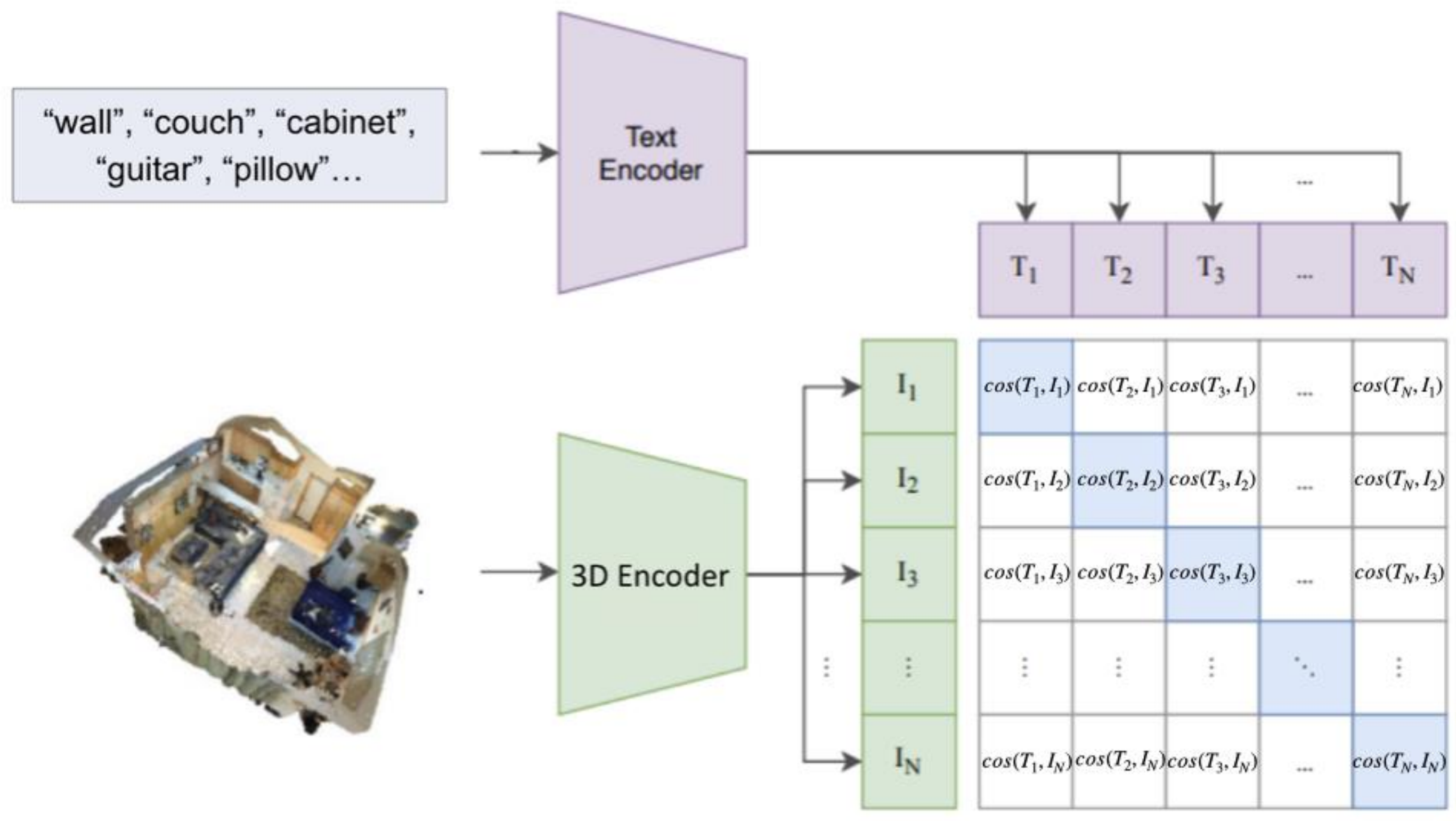
Problem

- We need to train a neural network to conduct **3D indoor scene semantic segmentation** on ScanNet200.
- ScanNet200 is a 3D point cloud scene dataset, including **XYZ position** and **RGB color** for each point. However, it's **imbalanced** for each class. We should try to deal with long-tailed class distributions.



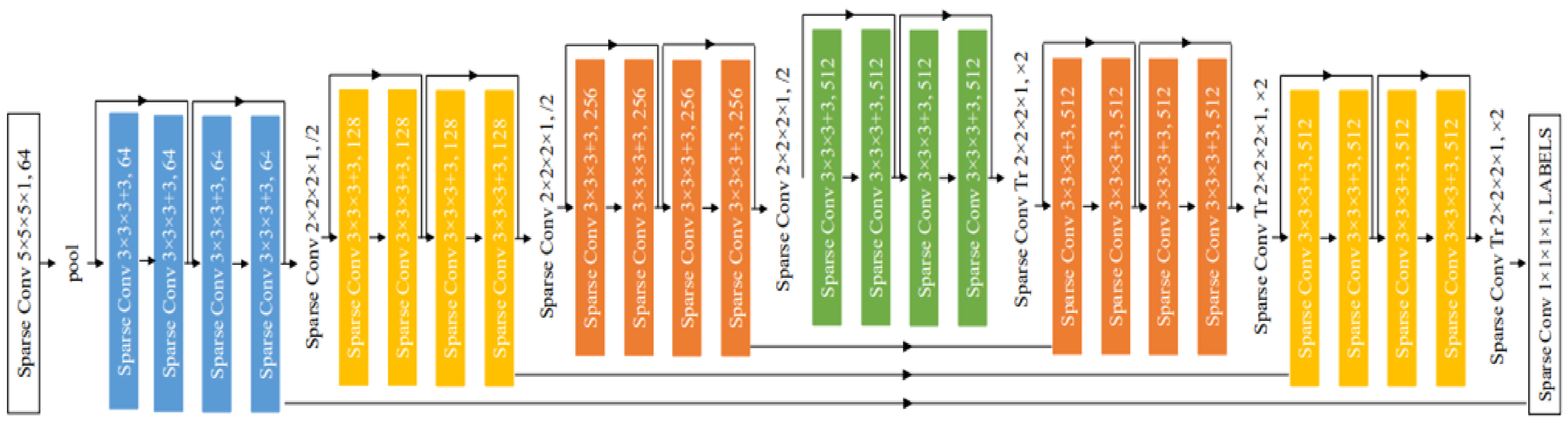
Language-Grounded 3D Feature Learning

- **Pipeline of 3D Feature Learning**



Ref: "Learning Transferable Visual Models From Natural Language Supervision" ICML 2021

- **3D Encoder**



MinkowskiUNet32

- **Text-supervised Contrastive Learning**

➢ For matching semantic text feature:

$$\mathcal{L}_{pos} = \sum_{i=1}^{N_p} max\left(0, \frac{f_i^s \cdot f_{h(i)}^t}{|f_i^s| \cdot |f_{h(i)}^t|} - t_{pos}\right),$$

➢ For non-matching semantic text feature:

$$\mathcal{L}_{neg} = \sum_{i=1}^{N_p} \frac{1}{|M|} \sum_{j \in M} max\left(0, t_{neg} - \frac{f_i^s \cdot f_j^t}{|f_i^s| \cdot |f_j^t|}\right),$$

➢ For final pre-training loss:

$$\mathcal{L} = \mathcal{L}_{pos} + \lambda \mathcal{L}_{neg}$$

Discussion

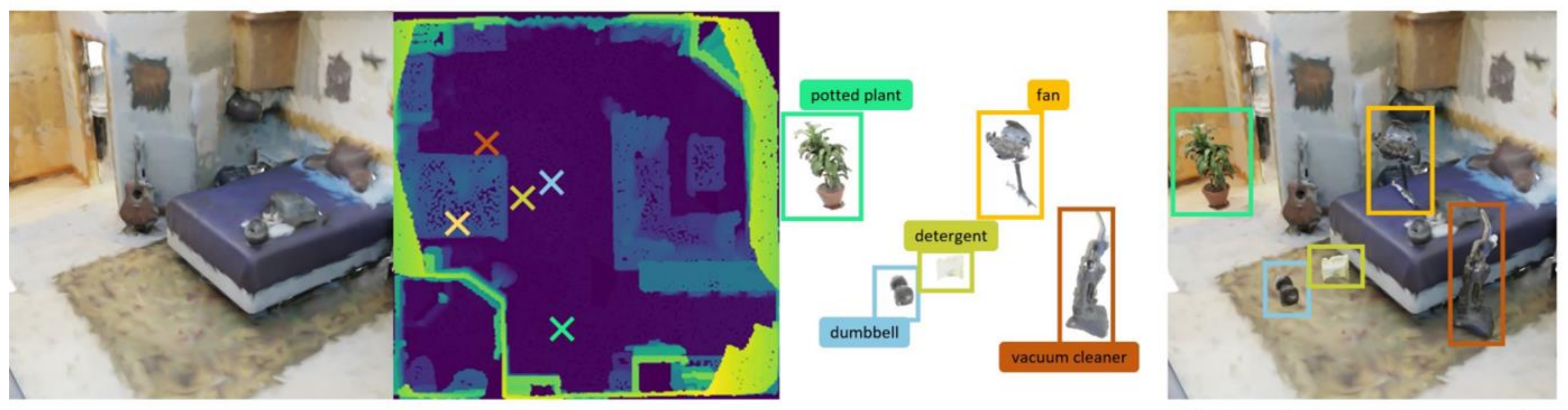
Implementation Summary

- Bring **text encoding** to leverage a **pre-trained CLIP** to map semantic labels to text features.
- Use **3D convolutional U-Net** as 3D encoder for 3D feature extraction.
- Implement **focal loss** to handle label imbalance problem.

Methods for dealing with imbalanced classes

- **Instance Sampling**

Placing infrequently seen instances in scenes, breaking context dependencies for recognition.



Original scan

Iterative instance placement

Sampled instances

Augmented scan with sampled instances

- **Class-Balanced Loss**

The focal loss proposes a modulating factor for a cross entropy loss:

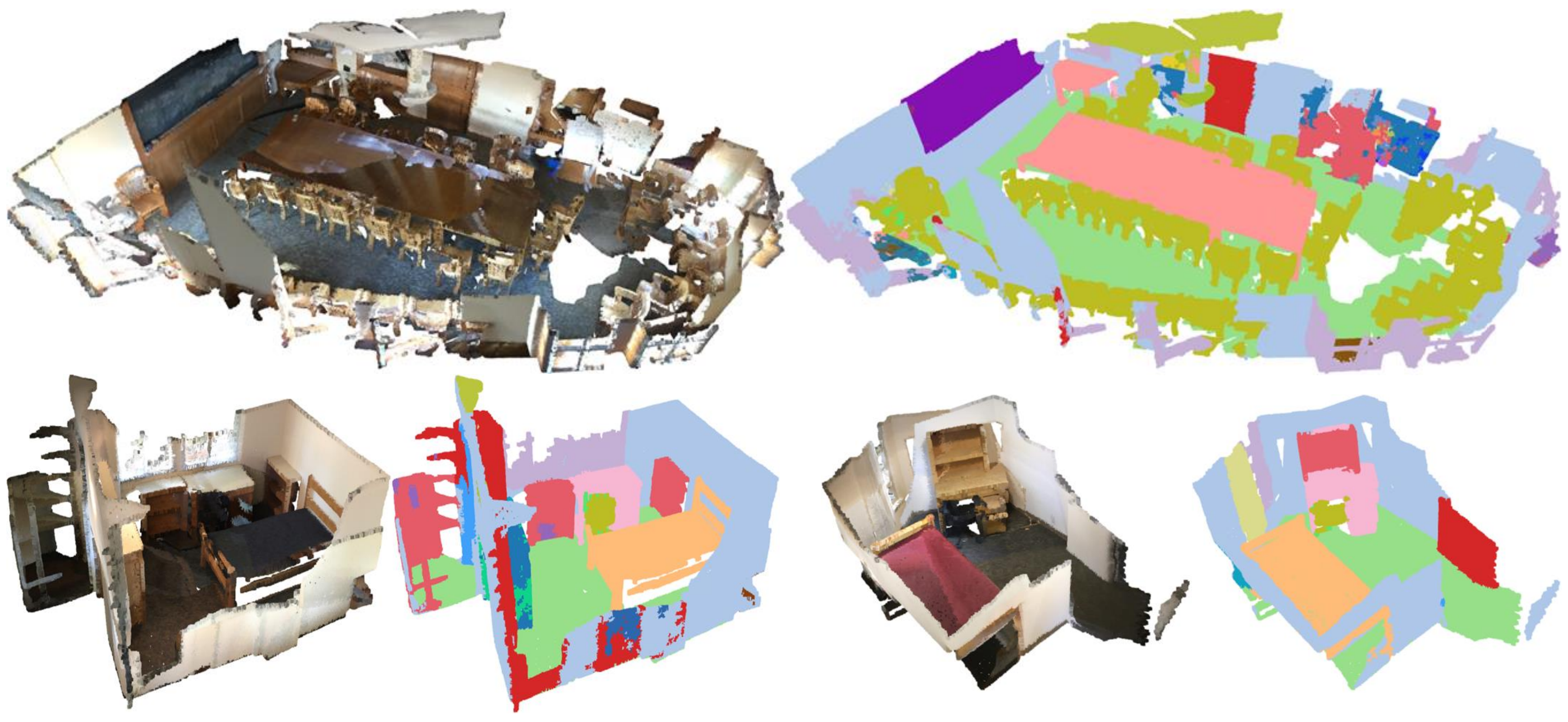
$$\mathcal{L}_{\text{focal}}(p_t) = -(1 - p_t)^{\gamma} log(p_t)$$

However, we did not see a direct improvement over cross entropy training by applying a focal loss, so we additionally re-balance the loss based on the class imbalance of the train set:

$$FL(p_t) = -\alpha(1 - p_t)^{\gamma} log(p_t),$$

$$\alpha_i = \frac{log(n_i)}{\sum_{j=1}^{N_{\text{class}}} log(n_j)}$$

Experiments Results



Method	mIoU
Res16UNet34C	
Fine-tune + focal loss	20.2
Res16UNet34D	
Fine-tune + CE loss	20.8
Fine-tune + focal loss	**22.8**
3D Feature Learning + Focal Fine-tune	

Ref: "Language-Grounded Indoor 3D Semantic Segmentation in the Wild" ECCV 2022