[SAM-Guided Masked Token Prediction for 3D Scene Understanding]

NeurIPS 2024

Zhimin Chen, Liang Yang, Yingwei Li, Longlong Jing, Bing Li

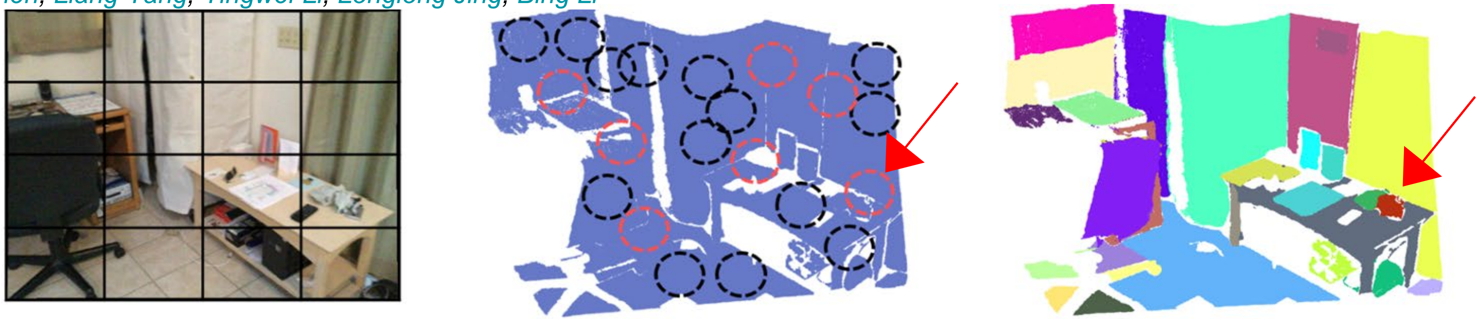
- Problem/Objective
  - 3D Scene understanding(Object detection, semantic segmentation)

- Contribution/Key Idea
  - Foundation model to improve 3D model
  - Distillation의 long-tail 문제 해결을 위한 group-balanced re-weighting method 제안
  - SOTA in Multi-task + Multi-dataset

김범준

SAM-Guided Masked Token Prediction for 3D Scene Understanding

Zhimin Chen, Liang Yang, Yingwei Li, Longlong Jing, Bing Li



(a) Patch based 2D tokenization method.

(b) KNN-based 3D tokenization method.

(c) Proposed SAM-guided 3D tokenization method.

Bridge3D 방식 → but) KNN tokenization으로 경계에 대한 정보 활용 bad

- 기존
  - CLIP2Scene, Seal, Bridge3D 등 Foundation model을 이용한 3D understanding 연구

- 한계점
  - 3D는 local info가 중요한데, scene단위 정보나 point단위 정보만 활용하여 학습하는 문제.
  - 전체 데이터를 같은 가중치로 다뤄서 long-tail 문제점 발생

[1]Chen, Runnan, et al. "Clip2scene: Towards label-efficient 3d scene understanding by clip." *CVPR2023*
[2] Liu, Youquan, et al. "Segment any point cloud sequences by distilling vision foundation models." *NerulPS2023*
[3]Chen, Zhimin, et al. "Bridging the domain gap: Self-supervised 3d scene understanding with foundation models." *NerulPS2023*

SAM-Guided Masked Token Prediction for 3D Scene Understanding

NeurIPS 2024

Zhimin Chen, Liang Yang, Yingwei Li, Longlong Jing, Bing Li

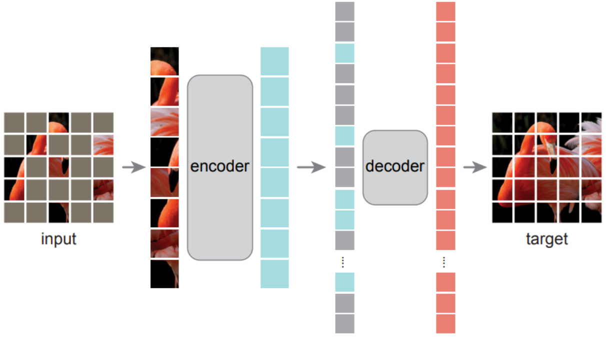
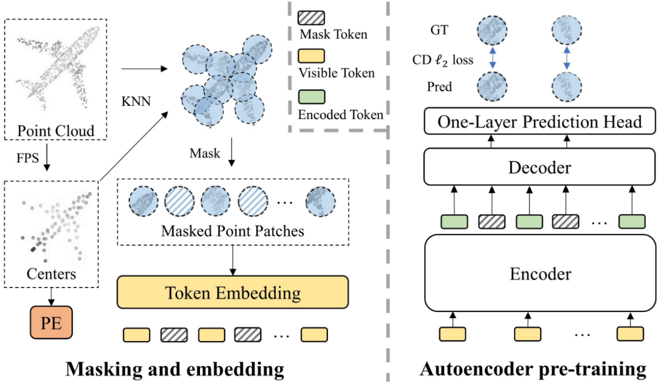


Figure 1. **Our MAE architecture**. During pre-training, a large



However, these 3D MAE applications have predominantly focused on masked point reconstruction. Recent studies [5, 69] have shown that masked feature prediction can be a more effective strategy for representation learning.

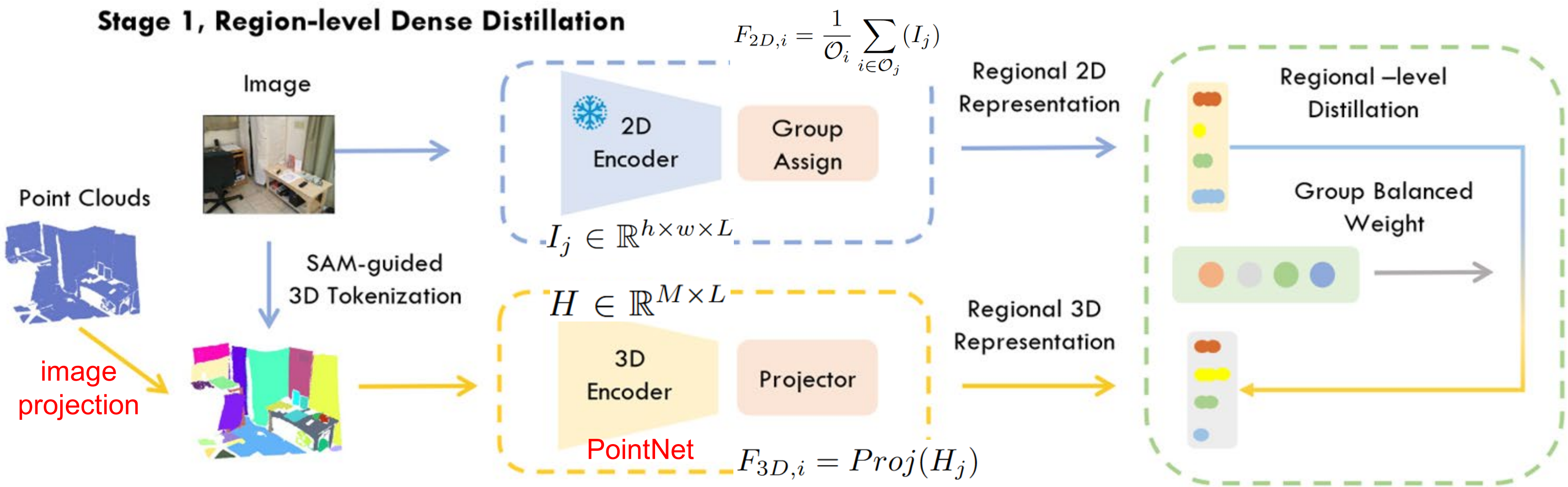
Masked feature prediction 기반으로한 2-stage framework 제시

[1]Pang, Yatian, et al. "Masked autoencoders for point cloud self-supervised learning." *ECCV2022*

김범준

SAM-Guided Masked Token Prediction for 3D Scene Understanding

Zhimin Chen, Liang Yang, Yingwei Li, Longlong Jing, Bing Li



To establish a precise correspondence between mask-level visual features and point tokens $\{x_i, p_i\}$, we align the point cloud tokens with the respective SAM masks, where x_i and p_i represent paired image and point features, respectively. This process is conducted offline, and the resulting labels are stored locally for easy access during the self-supervised training phase.

$$\mathcal{L}_{distill} = \frac{1}{M} \sum_{i}^{M} L_1(F_{2D,i}, F_{3D,i})$$

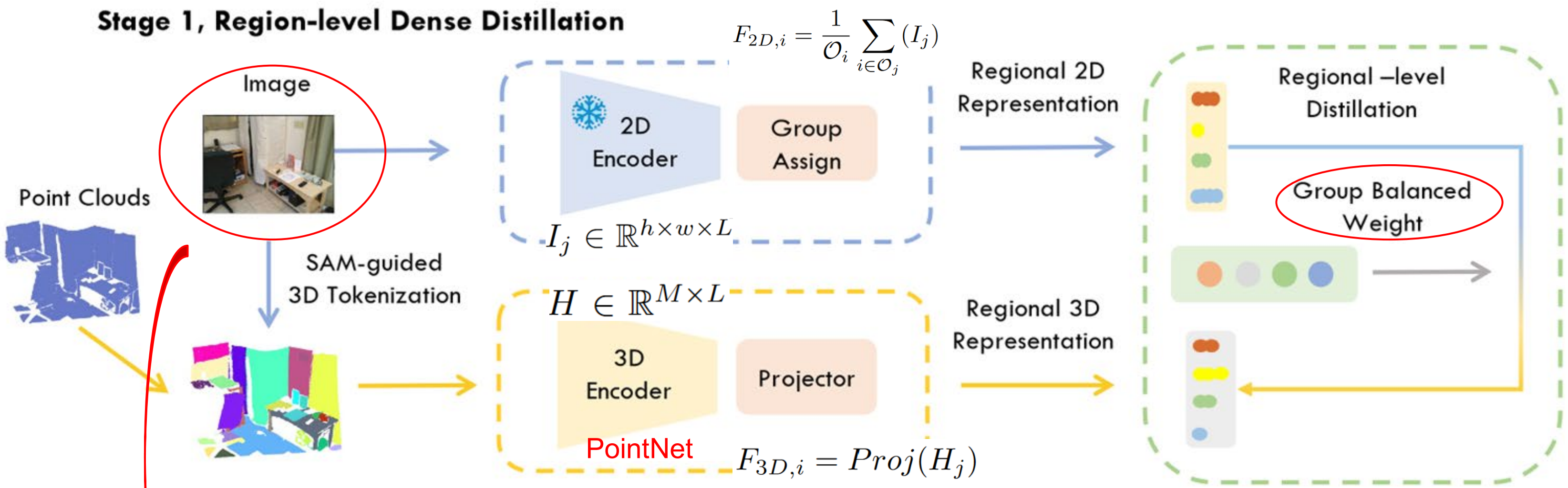
Then, we assign points to tokens based on their positions within the SAM-defined regions in the 2D images. Each patch's centroid is calculated as the average position of all points within that patch.

김범준

NeurIPS 2024

SAM-Guided Masked Token Prediction for 3D Scene Understanding

Zhimin Chen, Liang Yang, Yingwei Li, Longlong Jing, Bing Li



DINOv2

pixel feature extract → Average pooling SAM mask region → Cluster into distinct group

빈도 count

$$w_i = \frac{\tau_i}{\sum_{j=0}^{K} \tau_i}$$

$$k_i = \frac{n_i - n_{\min}}{n_{\max}}$$

$$\tau_i = 1.0 - k_i$$

$$\mathcal{L}_{distill} = \frac{1}{M} \sum_{i}^{M} L_1(F_{2D,i}, F_{3D,i})$$

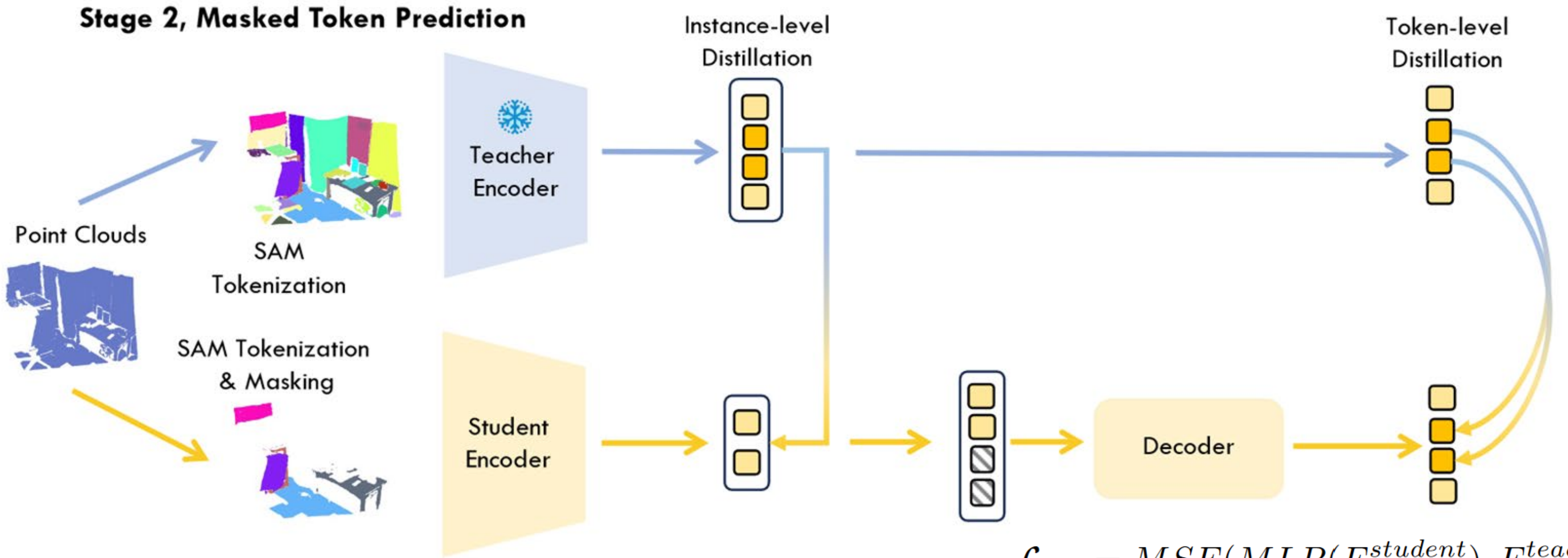
$$\mathcal{L}_{distill} = \frac{1}{M} \sum_{i}^{M} w_i L_1(F_{2D,i}, F_{3D,i})$$

김범준

[SAM-Guided Masked Token Prediction for 3D Scene Understanding](#)

NeurIPS 2024

Zhimin Chen, Liang Yang, Yingwei Li, Longlong Jing, Bing Li



For the instance-level knowledge prediction, we pool all point token features after the teacher encoder as $F_{ins}^{teacher}$ and after the student encoder as $F_{ins}^{student}$. The student model then predicts $F_{ins}^{teacher}$ using MLP layers. The instance prediction is formulated as follows:

$$\mathcal{L}_{ins} = MSE(MLP(F_{ins}^{student}), F_{ins}^{teacher})$$

We use the global features of the student model with only visible inputs to predict the global features of the teacher model with complete inputs. Additionally, we employ a token-level prediction loss to ensure that the student models can predict the masked tokens obtained from the teacher model's decoder.

$$\mathcal{L}_{token} = \frac{1}{N_m}\sum_{i=1}^{N_m} MSE(F_i^{student}, F_i^{teacher})$$

$$\mathcal{L}_{final} = \mathcal{L}_{ins} + \mathcal{L}_{token}$$

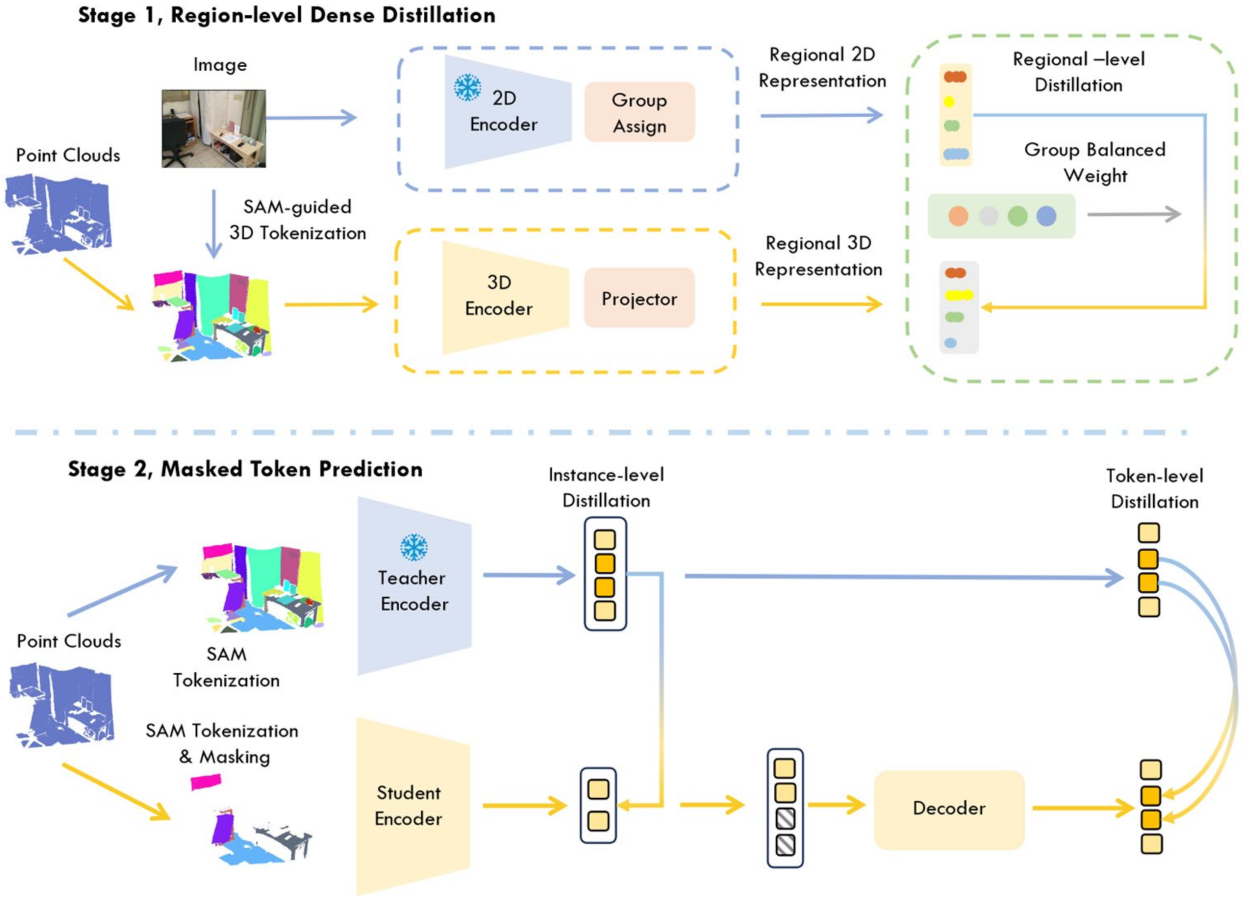
Where N_m is the number of masked tokens.

김범준

SAM-Guided Masked Token Prediction for 3D Scene Understanding

NeurIPS 2024

Zhimin Chen, Liang Yang, Yingwei Li, Longlong Jing, Bing Li



김범준

NeurIPS 2024

SAM-Guided Masked Token Prediction for 3D Scene Understanding

Zhimin Chen, Liang Yang, Yingwei Li, Longlong Jing, Bing Li

Methods	Pre-trained	SUN RGB-D AP_{25}	SUN RGB-D AP_{50}	ScanNetV2 AP_{25}	ScanNetV2 AP_{50}
VoteNet [45]	*None*	57.7	32.9	58.6	33.5
PointContrast [58]	✓	57.5	34.5	59.2	38.0
Hou et al. [29]	✓	-	36.4	-	39.3
4DContrast [9]	✓	-	38.2	-	40.0
DepthContrast [68]	✓	61.6	35.5	64.0	42.9
DPCo [35]	✓	60.2	35.5	64.2	41.5
3DETR [41]	*None*	58.0	30.3	62.1	37.9
+Plain Transformer	*None*	57.6	31.9	61.1	38.6
+Point-BERT[64]	-	-	-	61.0	38.3
+Point-MAE [43]	✓	-	-	63.4	40.6
+MaskPoint [37]	✓	-	-	63.4	40.6
+ACT [20]	✓	-	-	63.5	41.0
+PiMAE [7]	✓	59.9	33.7	63.0	40.2
+Bridge3D [10]	✓	61.8	37.1	65.3	44.2
+Ours	✓	**63.5(+1.7)**	**39.5(+2.4)**	**68.2 (+2.9)**	**48.4(+4.2)**
GroupFree3D [39]	*None*	63.0	45.2	67.3	48.9
+Plain Transformer	*None*	62.2	45.0	66.1	48.3
+Point-MAE [43]	✓	63.9	46.1	67.4	49.8
+PiMAE [7]	✓	65.0	46.8	67.9	50.5
+Bridge3D [10]	✓	67.9	48.5	69.1	51.9
+Ours	✓	**68.9(+1.0)**	**52.1(+3.6)**	**72.3(+3.2)**	**55.7(+3.8)**

Table 1: **3D object detection results on ScanNet and SUN RGB-D dataset.** We adopt the average precision with 3D IoU thresholds of 0.25 (AP_{25}) and 0.5 (AP_{50}) for the evaluation metrics.

김범준

SAM-Guided Masked Token Prediction for 3D Scene Understanding

Zhimin Chen, Liang Yang, Yingwei Li, Longlong Jing, Bing Li

Methods	Pre-trained	S3DIS mIoU	S3DIS mAcc	ScanNetV2 mIoU	ScanNetV2 mAcc
SR-UNet [58]	*None*	68.2	75.5	72.1	80.7
PointContrast [58]	✓	70.9	77.0	74.1	81.6
DepthContrast [68]	✓	70.6	-	73.1	-
Hou et al. [29]	✓	72.2	-	73.8	-
Standard Transformer [64]	*None*	60.0	68.6	-	-
PointBert [64]	✓	60.8	69.9	-	-
PViT [46]	*None*	64.4	69.9	-	-
PViT+Pix4Point [46]	✓	69.6	75.2	-	-
Plain Transformer	*None*	61.1	67.2	67.3	73.1
+Point-MAE [43]	✓	64.8	70.2	-	-
+Bridge3D [10]	✓	70.2	76.1	73.9	80.2
+Ours	✓	**71.8 (+1.6)**	**78.2(+2.1)**	**75.4(+1.5)**	**81.5(+1.3)**

Table 2: **3D semantic segmentation results on S3DIS and ScanNet dataset.** We adopt the mean accuracy (mAcc) and mean IoU (mIoU) for the evaluation metrics.

김범준

[SAM-Guided Masked Token Prediction for 3D Scene Understanding](#)

NeurIPS 2024

Zhimin Chen, Liang Yang, Yingwei Li, Longlong Jing, Bing Li

Dense Distillation	Masked Token Prediction	Balanced Re-weight	SAM-Guided Tokenzie	ScanNetV2 AP_{25}	ScanNetV2 AP_{50}	S3DIS $mIoU$	S3DIS $mAcc$
				61.1	38.6	61.1	67.2
✓				62.4	41.7	66.2	71.3
✓	✓			64.5	44.3	68.7	74.1
✓	✓	✓		66.0	46.1	69.7	75.9
✓	✓		✓	67.1	47.0	70.9	77.0
✓	✓	✓	✓	**68.2**	**48.4**	**71.8**	**78.2**

Table 3: **The effectiveness of each component.** Ablation study on the effectiveness of each component on 3D object detection and semantic segmentation tasks.

	ScanNetV2 AP_{25}	ScanNetV2 AP_{50}	S3DIS $mIoU$	S3DIS $mAcc$
Stage 1	65.2	45.1	69.1	75.3
Stage 1 + MTP in same stage	66.0	46.3	69.9	76.1
Stage 1 + Stage 2 (Ours)	**68.2**	**48.4**	**71.8**	**78.2**

Table 4: **The effectiveness of Stage.** Ablation study on the effectiveness of a two-stage framework on 3D object detection and semantic segmentation tasks. MTP here represents the masked token prediction

김범준