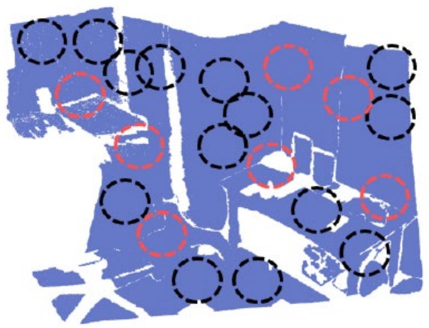
SAM-Guided Masked Token Prediction for 3D Scene Understanding

Zhimin Chen[1] , Liang Yang[2] , Yingwei Li[3] ,Longlong Jing[2], Bing Li[1]
Clemson University[1], The City University of New York[2], Johns Hopkins University[3]

- **Background**: Foundation models have advanced 3D task performance significantly.
- **Issues**: Limited alignment and long-tail distribution hinder 3D foundation model training.
- **Problematic**: How to align 2D and 3D features to improve 3D scene understanding.
- **Idea**: Use SAM-guided tokenization and a balanced re-weighting strategy for region-level distillation, combined with a two-stage masked token prediction framework.

Misalignment Problems

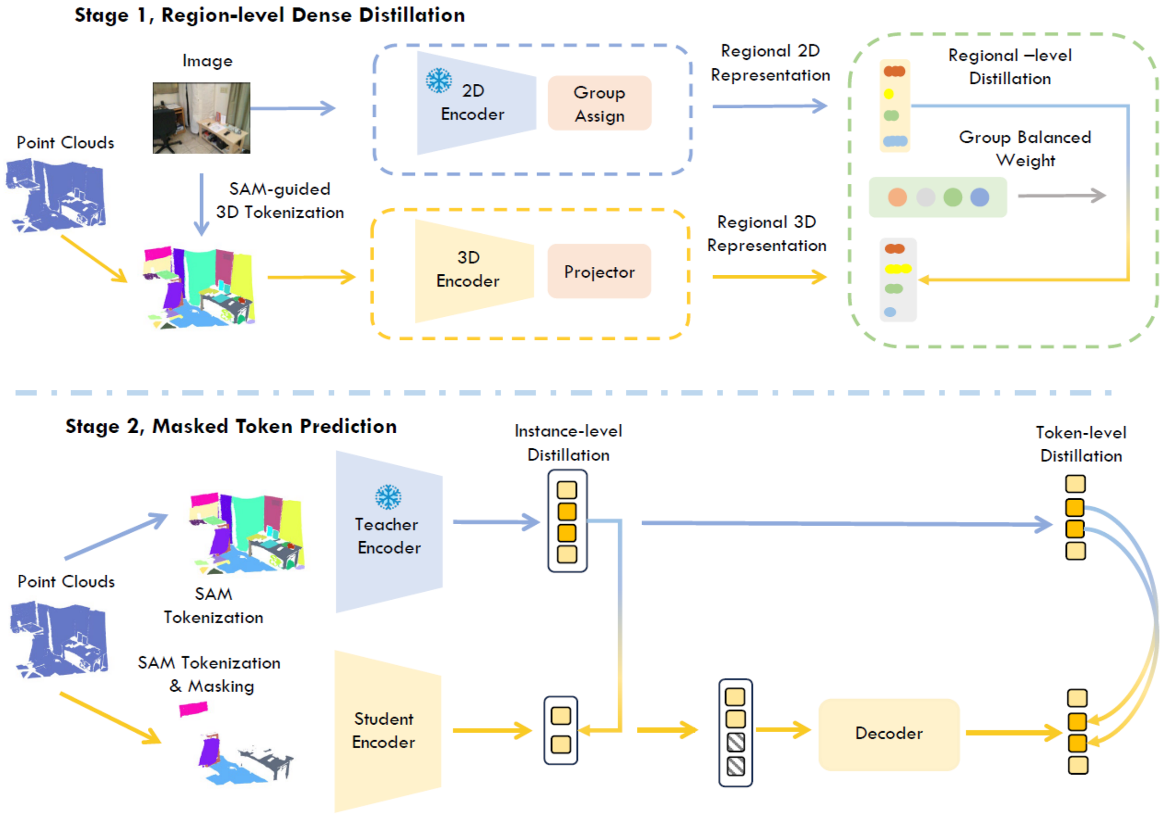


KNN-Based Tokenization

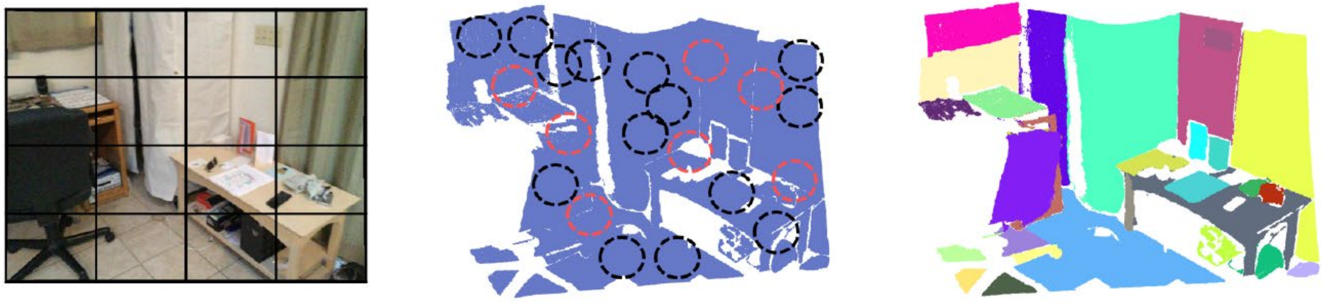
SAM Masks

As shown in the red circle, the KNN-based method may inadvertently group points from different SAM regions into the same tokens, leading to potential confusion within the 3D network.

Framework


Stage 1, Region-level Dense Distillation
Image
Point Clouds
SAM-guided
3D Tokenization
2D
Encoder
Group
Assign
3D
Encoder
Projector
Regional 2D
Representation
Regional 3D
Representation
Regional –level
Distillation
Group Balanced
Weight
Stage 2, Masked Token Prediction
Point Clouds
SAM
Tokenization
SAM Tokenization
& Masking
Teacher
Encoder
Student
Encoder
Instance-level
Distillation
Decoder
Token-level
Distillation

1, SAM-Guided Tokenization



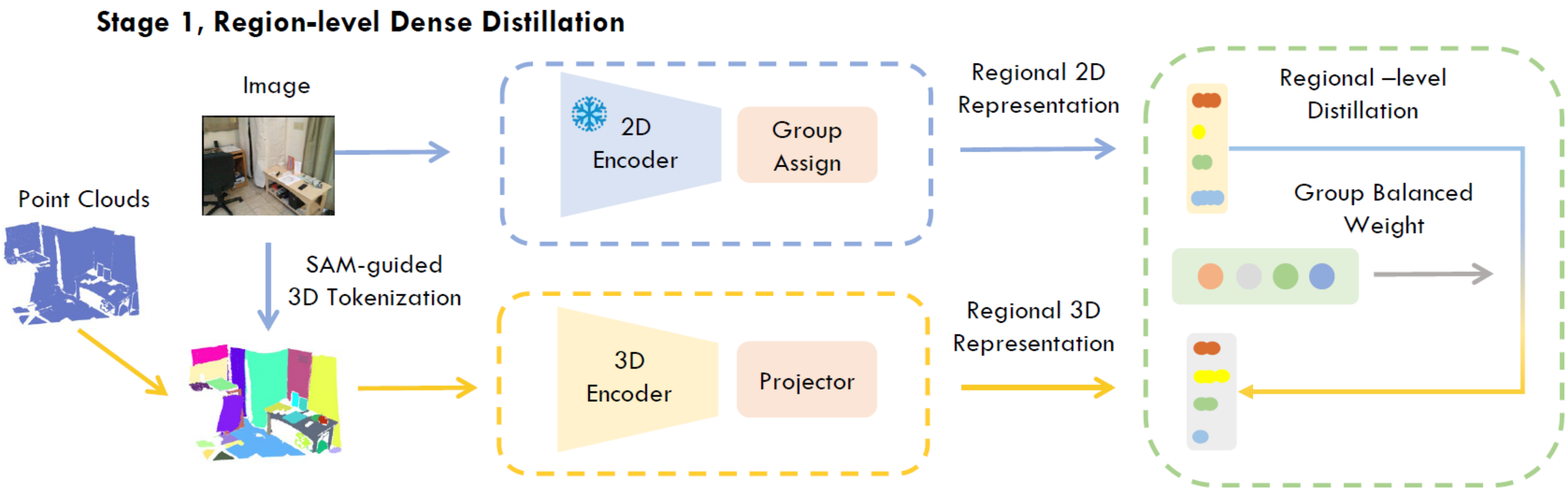
(a) Patch based 2D tokenization method.

(b) KNN-based 3D tokenization method.

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

Unlike previous methods leverage the KNN-based tokenization method, our method effectively employs SAM masks in tokenization to ensure seamless region-level knowledge distillation, thereby avoiding misalignment issues.

2, Stage 1



In the first stage, we input complete point clouds and leverage SAM masks to guide the point cloud tokenization, thereby seamlessly aligning the 2D and 3D region-level features for dense prediction. A group-balanced weight is applied during distillation to prevent bias towards the head representations.

2.1, Stage 1

Group Balanced Re-weighting

•**Challenge**: 3D datasets are inherently imbalanced, making traditional methods suboptimal for underrepresented (tail) classes.

Method:

1.**Region Grouping**: Apply SAM masks and use max pooling to obtain region-level features.

2.**Clustering**: Use K-means to assign features into K distinct groups (pseudo-labels).

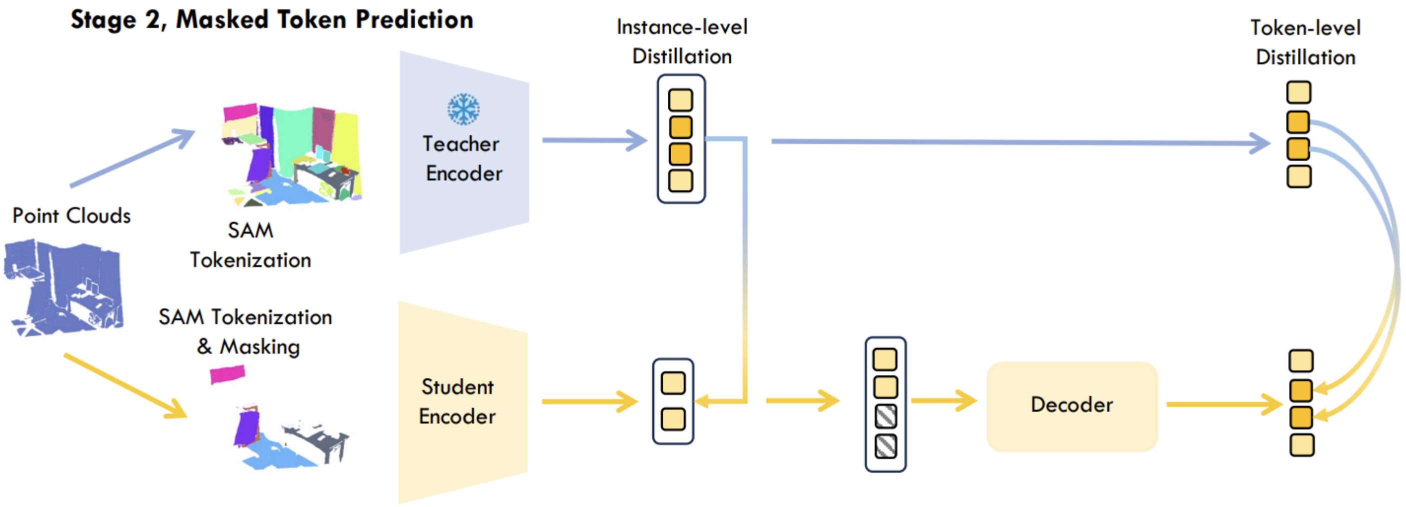
3.**Re-weighting Strategy**:

   1. Compute weights for each group to prioritize tail classes.
   2. Adjust the loss to balance learning across head and tail regions.

4.**Loss Function**: Use a weighted L1 loss to enhance representation learning.

5.**Outcome**: Balances representation learning, improving robustness on underrepresented regions.

3, Stage 2



In the second stage, we freeze the models trained in the first stage and have the student models predict instance-level features and masked tokens obtained from the teacher models.

Experiments

Methods	Pre-trained	SUN RGB-D		ScanNetV2	
		AP_{25}	AP_{50}	AP_{25}	AP_{50}
VoteNet [39]	*None*	57.7	32.9	58.6	33.5
PointContrast [49]	✓	57.5	34.5	59.2	38.0
Hou et al. [26]	✓	-	36.4	-	39.3
4DContrast [9]	✓	-	38.2	-	40.0
DepthContrast [55]	✓	61.6	35.5	64.0	42.9
DPCo [30]	✓	60.2	35.5	64.2	41.5
3DETR [35]	*None*	58.0	30.3	62.1	37.9
+Plain Transformer	*None*	57.6	31.9	61.1	38.6
+Point-BERT[51]	-	-	-	61.0	38.3
+Point-MAE [37]	✓	-	-	63.4	40.6
+MaskPoint [31]	✓	-	-	63.4	40.6
+ACT [15]	✓	-	-	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 [33]	*None*	63.0	45.2	67.3	48.9
+Plain Transformer	*None*	62.2	45.0	66.1	48.3
+Point-MAE [37]	✓	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)**

3D object detection results on ScanNet and SUN RGB-D dataset.

Experiments

Methods	Pre-trained	S3DIS $mIoU$	S3DIS $mAcc$	ScanNetV2 $mIoU$	ScanNetV2 $mAcc$
SR-UNet [49]	*None*	68.2	75.5	72.1	80.7
PointContrast [49]	✓	70.9	77.0	74.1	81.6
DepthContrast [55]	✓	70.6	-	73.1	-
Hou et al. [26]	✓	72.2	-	73.8	-
Standard Transformer [51]	*None*	60.0	68.6	-	-
PointBert [51]	✓	60.8	69.9	-	-
PViT [40]	*None*	64.4	69.9	-	-
PViT+Pix4Point [40]	✓	69.6	75.2	-	-
Plain Transformer	*None*	61.1	67.2	67.3	73.1
+Point-MAE [37]	✓	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)**

3D semantic segmentation results on S3DIS dataset and ScanNet

Ablation Study

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**

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

Conclusion

1, We introduce a novel two-stage SAM-guided masked token prediction framework that leverages foundation models for 3D scene understanding.

2, We present a group-balanced re-weighting method for long-tail representation distillation and a SAM-guided tokenization method to seamlessly align 2D and 3D region-level features.

3. Extensive experiments have been conducted to demonstrate the significance of our approach in various 3D downstream tasks.