西安交通大学
XI'AN JIAOTONG UNIVERSITY

清華大学
Tsinghua University

交叉信息研究院
Institute for Interdisciplinary Information Sciences

MEGVII 旷视

上海人工智能实验室
Shanghai Artificial Intelligence Laboratory

ICML | 2023
International Conference On Machine Learning

Contrast with Reconstruct

Contrastive 3D Representation Learning Guided by Generative Pretraining

Zekun Qi † Runpei Dong † Guofan Fan Zheng Ge Xiangyu Zhang Kaisheng Ma Li Yi

International Conference on Machine Learning (ICML)
July 23rd – 27th, 2023, Honolulu Hawaii

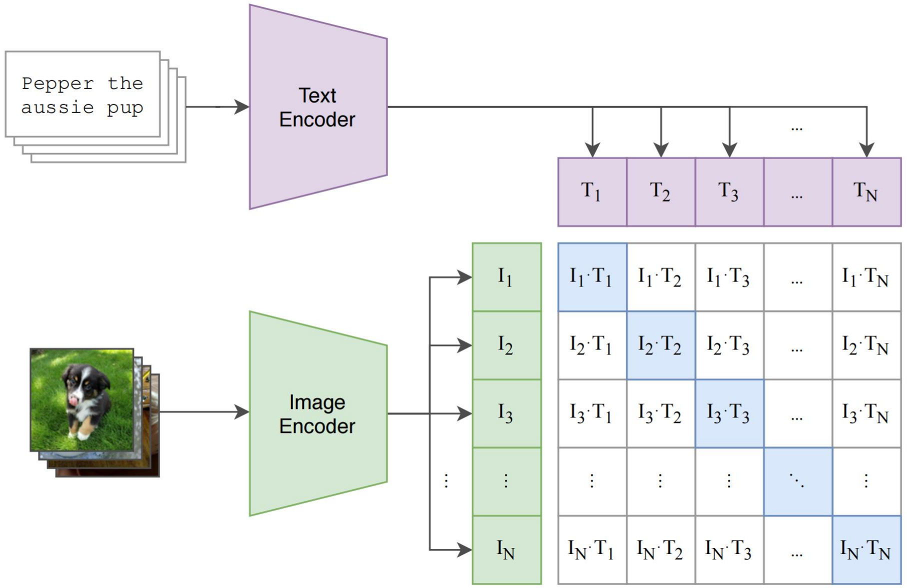
Zekun Qi

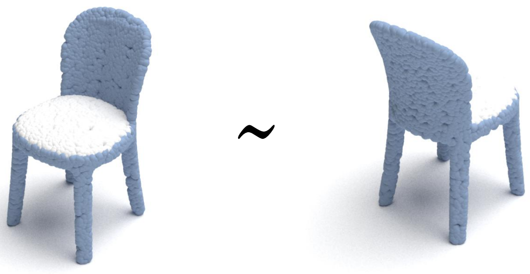
24 July 2023

Paradigm Comparison

Contrastive Learning

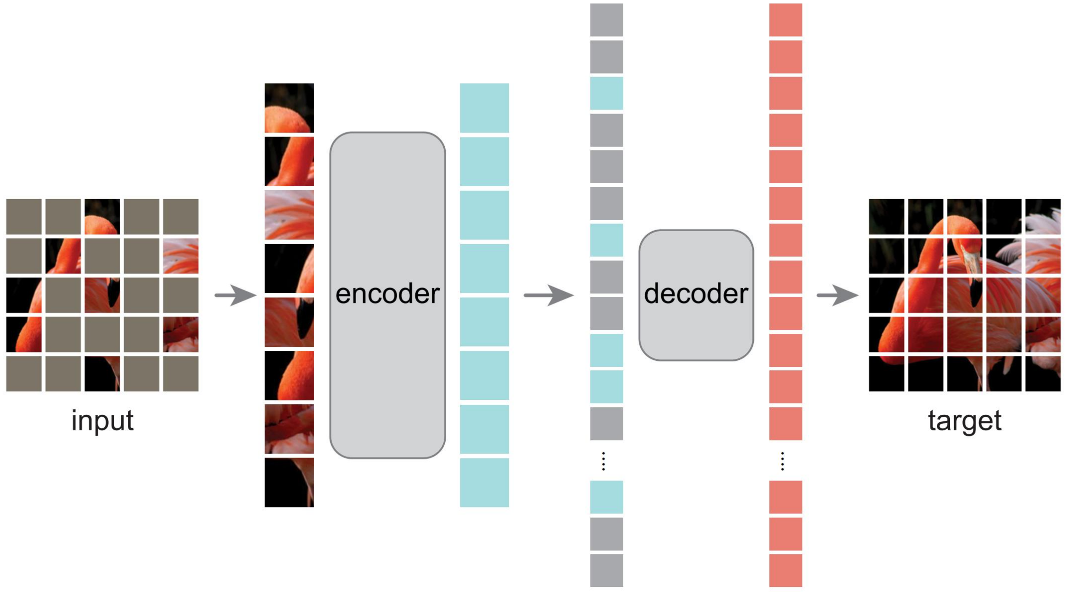
SimCLR, MoCo, PointContrast
CLIP, ALIGN, FLIP, GLIP, ULIP

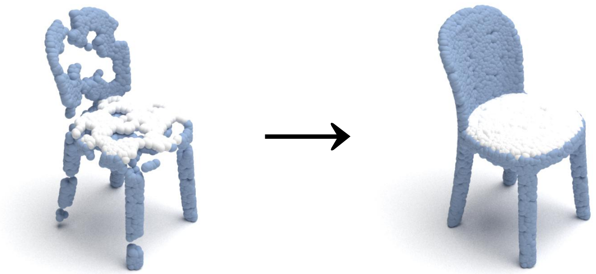




Generative Learning

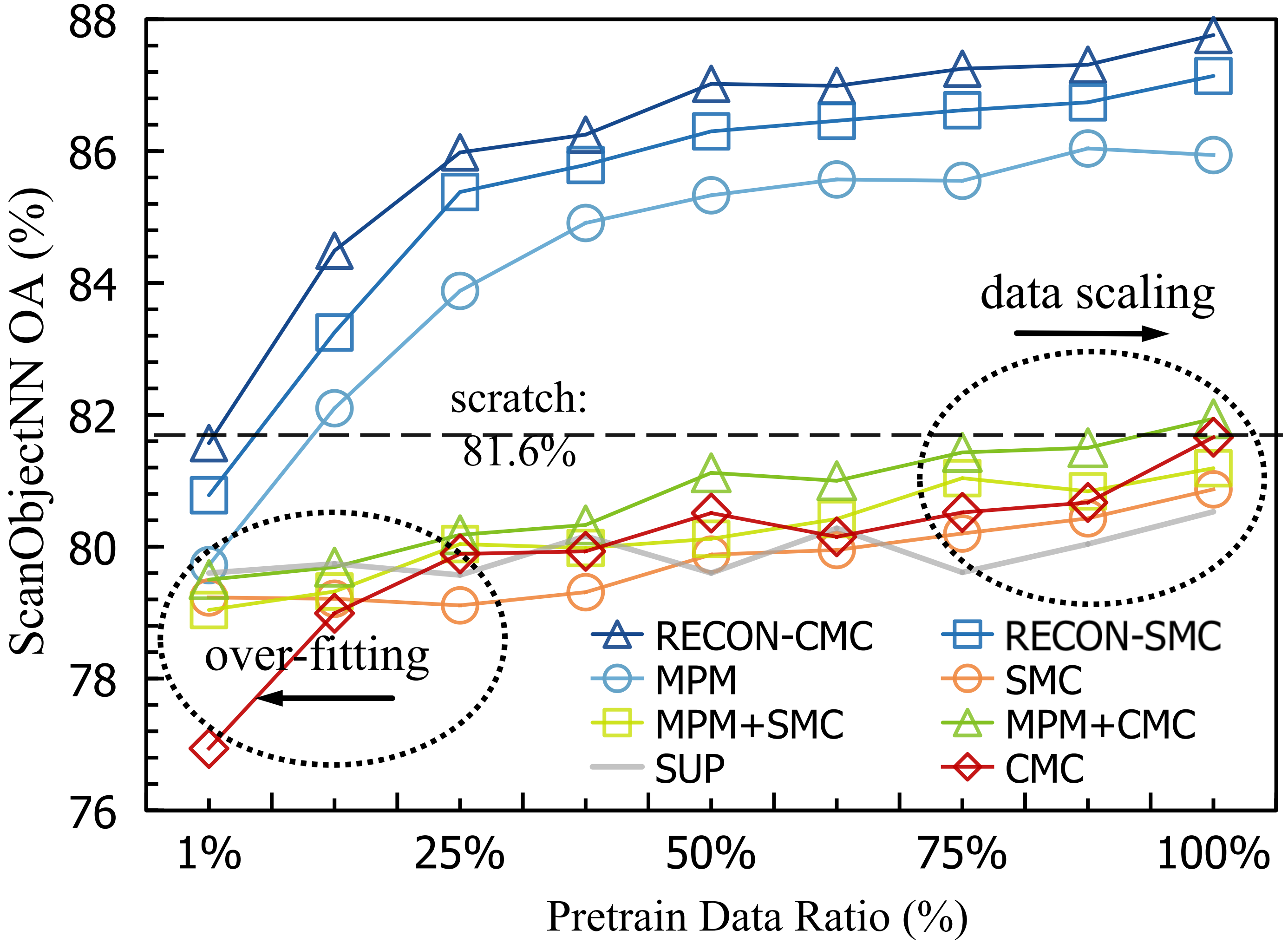
BERT, MAE, MaskFeat, Point-MAE
BEiT-3, M3AE, CoCa, VLBERT, MVP, EVA





Source: Learning Transferable Visual Models From Natural Language Supervision
Masked Autoencoders Are Scalable Vision Learners

Pattern differences



(a) Data Scaling Capacity Comparison

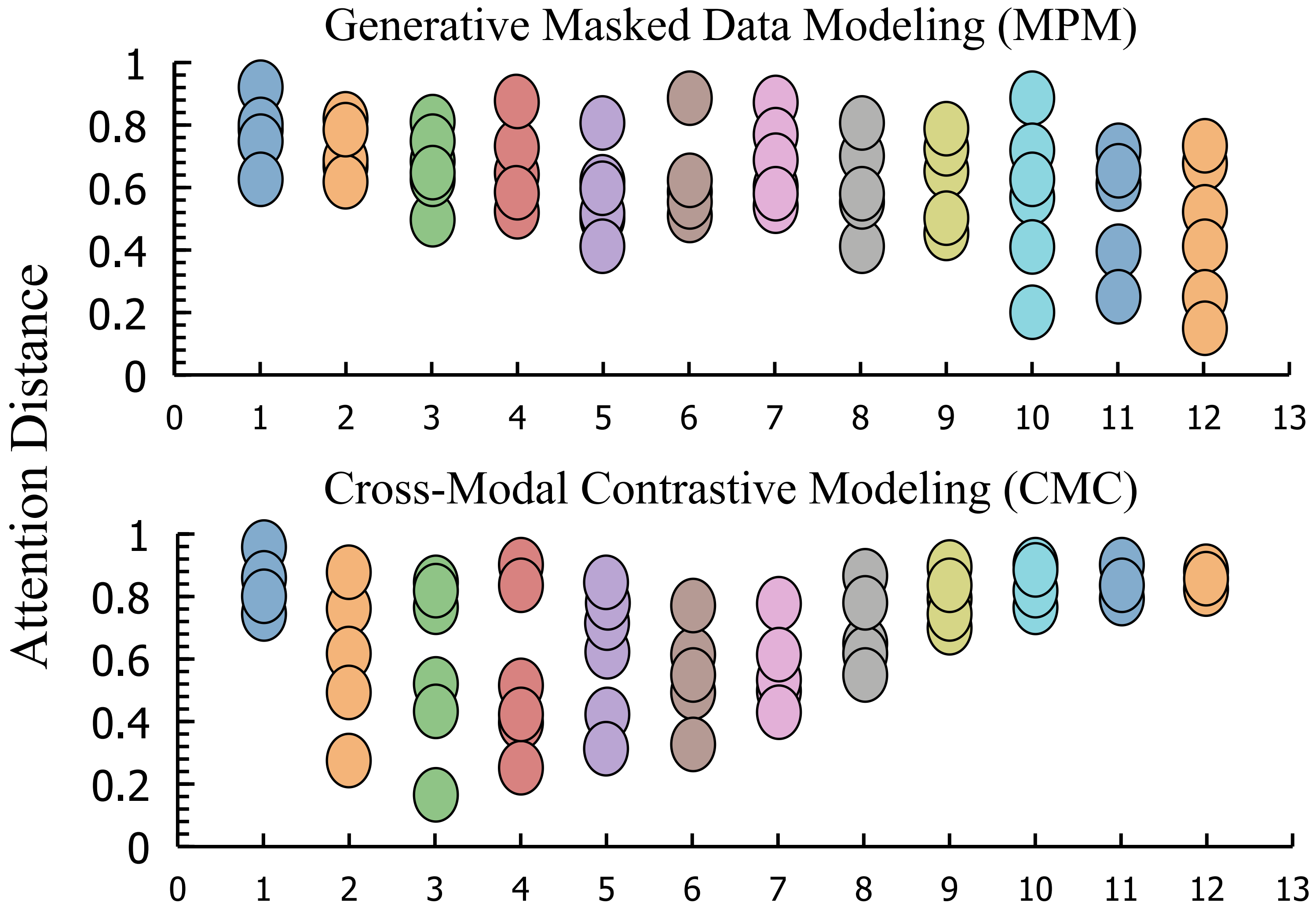
Representation over-fitting (Contrastive)

- contrastive models can easily find shortcuts with trivial representations

Data filling(Generative)

- generative models are less data-hungry that learn decent initialization with very few data

Pattern differences



(b) Average Attention Distance Comparison

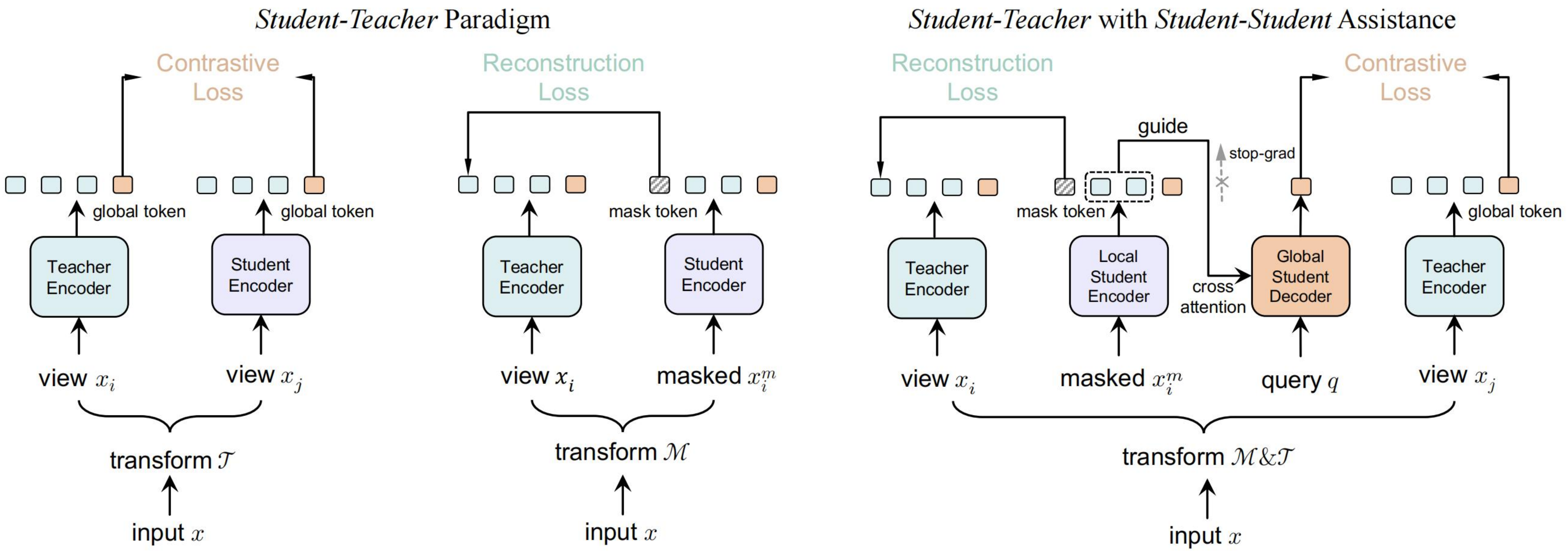
Global Representation (Contrastive)

- Pay more attention to long-range information

Local Representation (Generative)

- Pay more attention to short-range information

Unified View

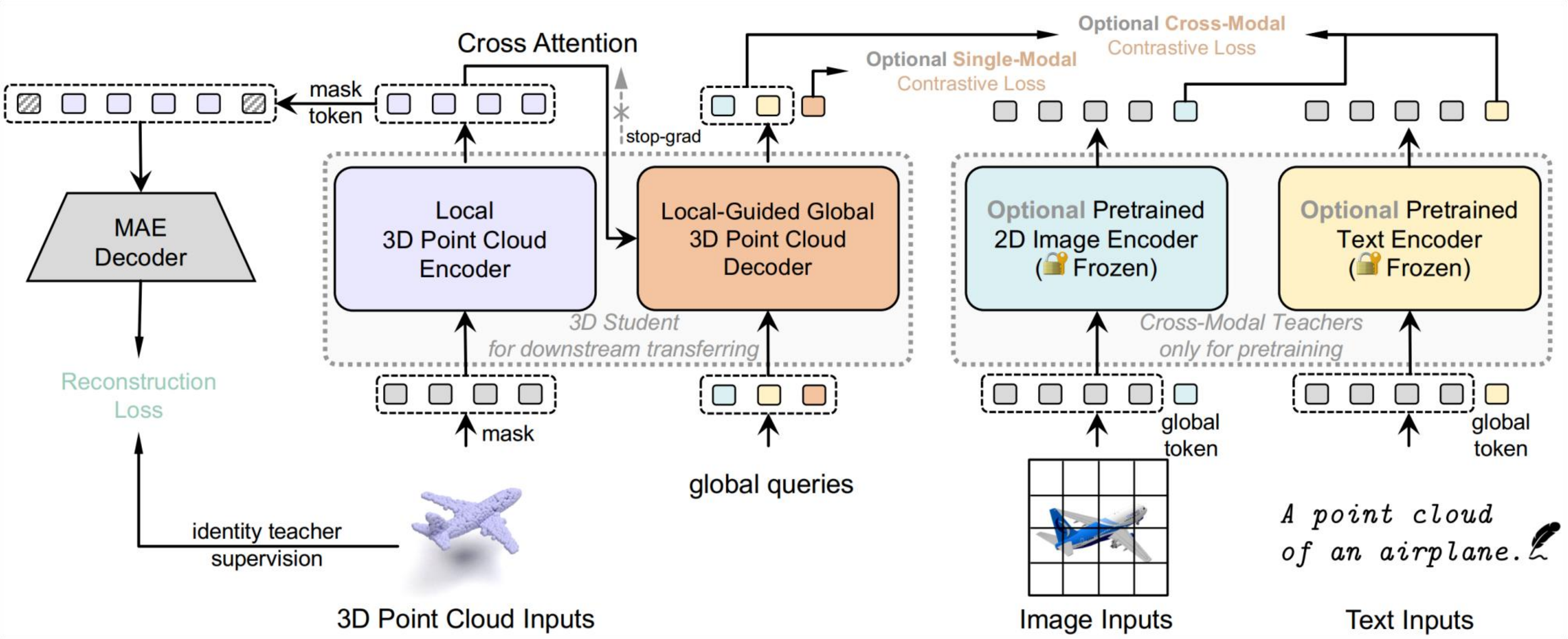


(a) contrastive modeling (b) generative masked modeling (c) ReCon

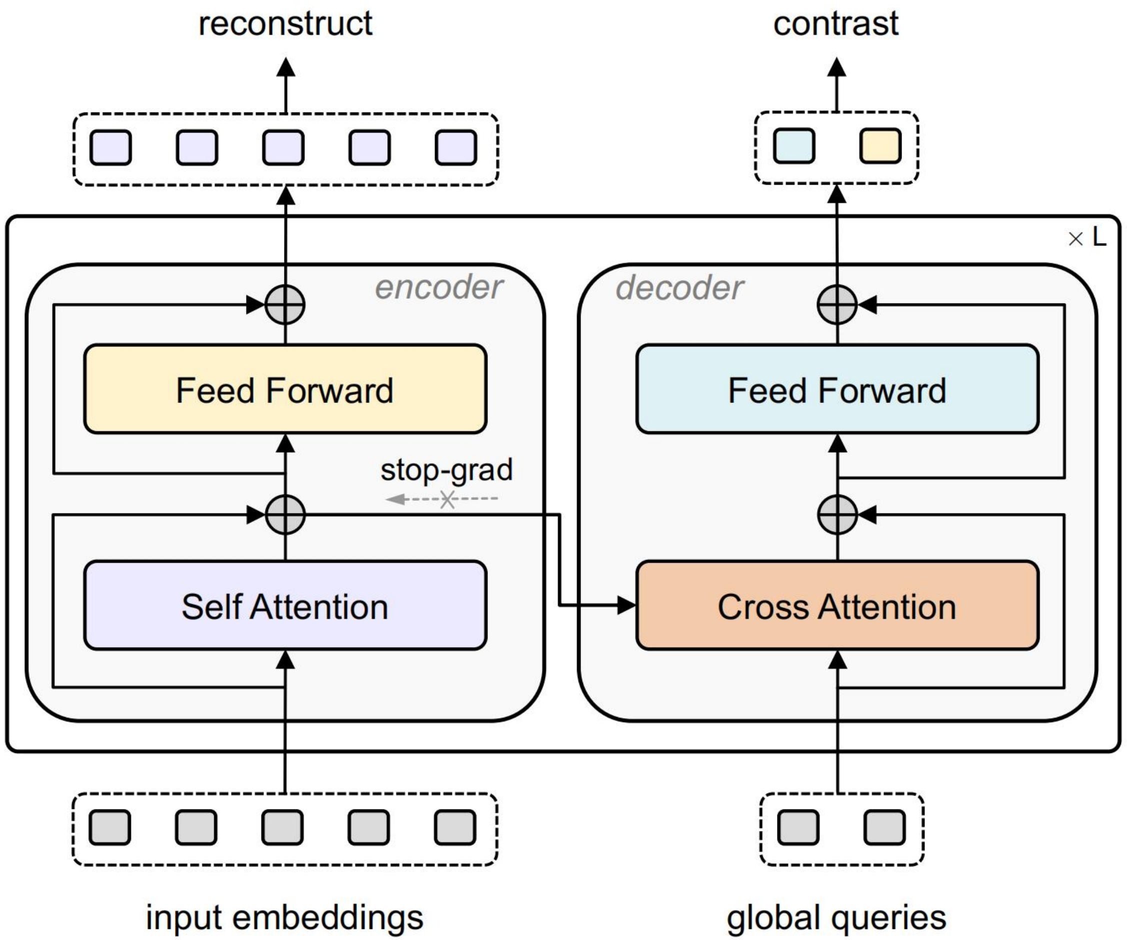
$$\min_{\theta} \mathop{\mathbb{E}}_{\substack{x\sim\mathcal{D},\\ \{\mathcal{T}_i,\mathcal{T}_j\}\in\mathcal{T}}} \mathcal{L}_{\mathbb{C}}^{\text{CON}}(z_i, z_j),$$
$$z_i = \mathcal{F}_\theta^S\left(\mathcal{T}_i(x)\right), z_j = \mathcal{F}_\phi^T\left(\mathcal{T}_j(x)\right),$$

$$\min_{\theta} \mathop{\mathbb{E}}_{\substack{x\sim\mathcal{D},\\ \{\mathcal{M}_i,\widetilde{\mathcal{M}}_i\}\in\mathcal{M}}} \mathcal{L}_{\mathbb{D}}^{\text{REC}}(z_i, z_j),$$
$$z_i = \mathcal{F}_\theta^S\left(\mathcal{M}_i(x)\right), z_j = \mathcal{F}_\phi^T\left(\widetilde{\mathcal{M}}_i(x)\right),$$

Contrast with Reconstruct



Block Design




```
# X: input embeddings (GxC)
# Q: global queries (N×C)

def block(X, Q):

    X = MHSA(norm(X)) + X
    Q = MHCA(norm(Q), X.detach()) + Q

    X = FFN(norm(X)) + X
    Q = FFN(norm(Q)) + Q

    return X, Q
```

Contrast with Reconstruct

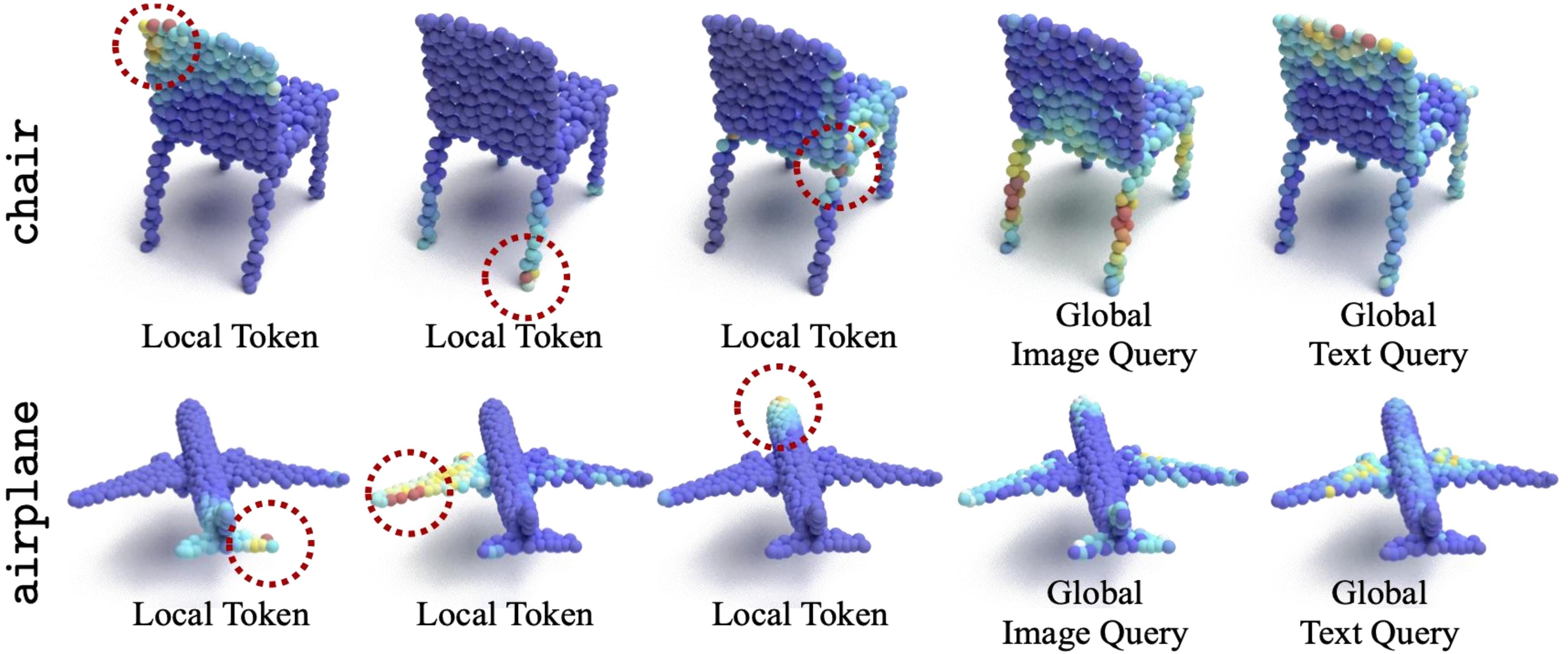
- Generative student can serve as a powerful regularization technique to alleviate ***over-fitting issue*** of the contrastive student.

- The ***data-filling issue*** of the generative student is alleviated due to the promising scaling capacity of the contrastive student.

- ReCon circumvents the discrepancies in ***attention patterns*** between generative learning and contrastive learning through a simple two-stream network architecture.

- Due to the equal number of global query tokens and contrastive teachers, the increase in FLOPs is very small compared to the single-stream network.

Attention Visualization



Downstream Tasks

State-of-the-art Performance on 3D Classification Tasks

Method	#P	#F	PT	MD	ScanObjectNN OBJ_BG	ScanObjectNN OBJ_ONLY	ScanObjectNN PB_T50_RS	ModelNet40 1k P	ModelNet40 8k P
Supervised Learning Only									
○ PointNet (Qi et al., 2017a)	3.5	0.5	×	×	73.3	79.2	68.0	89.2	90.8
○ PointNet++ (Qi et al., 2017b)	1.5	1.7	×	×	82.3	84.3	77.9	90.7	91.9
○ DGCNN (Wang et al., 2019)	1.8	2.4	×	×	82.8	86.2	78.1	92.9	-
○ PointCNN (Li et al., 2018)	0.6	-	×	×	86.1	85.5	78.5	92.2	-
○ SimpleView (Goyal et al., 2021)	-	-	×	×	-	-	80.5±0.3	93.9	-
○ MVTN (Hamdi et al., 2021)	11.2	43.7	×	×	92.6	92.3	82.8	93.8	-
○ PCT (Guo et al., 2021)	2.88	2.3	×	×	-	-	-	93.2	-
○ PointMLP (Ma et al., 2022)	12.6	31.4	×	×	-	-	85.4±0.3	94.5	-
○ PointNeXt (Qian et al., 2022)	1.4	3.6	×	×	-	-	87.7±0.4	94.0	-
○ P2P-HorNet (Wang et al., 2022)	-	34.6	✓	✓	-	-	89.3	94.0	-
with Single-Modal Self-Supervised Representation Learning (FULL)									
● Transformer (Vaswani et al., 2017)	22.1	4.8	×	×	83.04	84.06	79.11	91.4	91.8
● Transformer† (Vaswani et al., 2017)	43.6	5.3	×	×	84.90	86.12	81.64	91.6	92.0
● Point-BERT (Yu et al., 2022b)	22.1	4.8	×	×	87.43	88.12	83.07	93.2	93.8
● Point-MAE (Pang et al., 2022)	22.1	4.8	×	×	90.02	88.29	85.18	93.8	94.0
○ Point-M2AE (Zhang et al., 2022b)	15.3	3.6	×	×	91.22	88.81	86.43	94.0	-
● Point-MAE† (Pang et al., 2022)	43.6	5.3	×	×	92.60	91.91	88.42	93.8	94.0
● **RECON w/o vot.**	43.6	5.3	×	×	**94.15**	**93.12**	**89.73**	93.6	93.8
● **RECON w/ vot.**	43.6	5.3	×	×	**94.49**	**93.29**	**90.35**	**93.9**	**94.2**
with Cross-Modal Self-Supervised Representation Learning (FULL)									
● ACT (Dong et al., 2023)	22.1	4.8	✓	×	93.29	91.91	88.21	93.7	94.0
● **RECON-Tiny w/o vot.**	11.4	2.4	✓	✓	**93.80**	**92.94**	**89.10**	93.3	93.6
● **RECON-Small w/o vot.**	19.0	3.2	✓	✓	**94.15**	**93.12**	**89.52**	93.5	93.8
● **RECON w/o vot.**	43.6	5.3	×	✓	**94.66**	**93.29**	**90.32**	**94.0**	**94.2**
● **RECON w/o vot.**	43.6	5.3	✓	✓	**95.18**	**93.63**	**90.63**	**94.1**	**94.3**
● **RECON w/ vot.**	43.6	5.3	✓	✓	**95.35**	**93.80**	**91.26**	**94.5**	**94.7**

Downstream Tasks

Table 3. **Zero-shot 3D object classification domain transfer on ModelNet40 (MN-40) and ModelNet10 (MN-10).** Top-1 accuracy (%) is reported. Ensemb. denotes whether to use the ensemble strategy with multiple text inputs.

Method	Backbone	Ensemb.	MN-10	MN-40
○ PointCLIP (Zhang et al., 2022c)	ResNet-50	×	30.2	20.2
● CLIP2Point (Huang et al., 2022)	Transformer	✓	66.6	49.4
● RECON	Transformer	×	**74.2**	**60.6**
● RECON	Transformer	✓	**75.6**	**61.7**

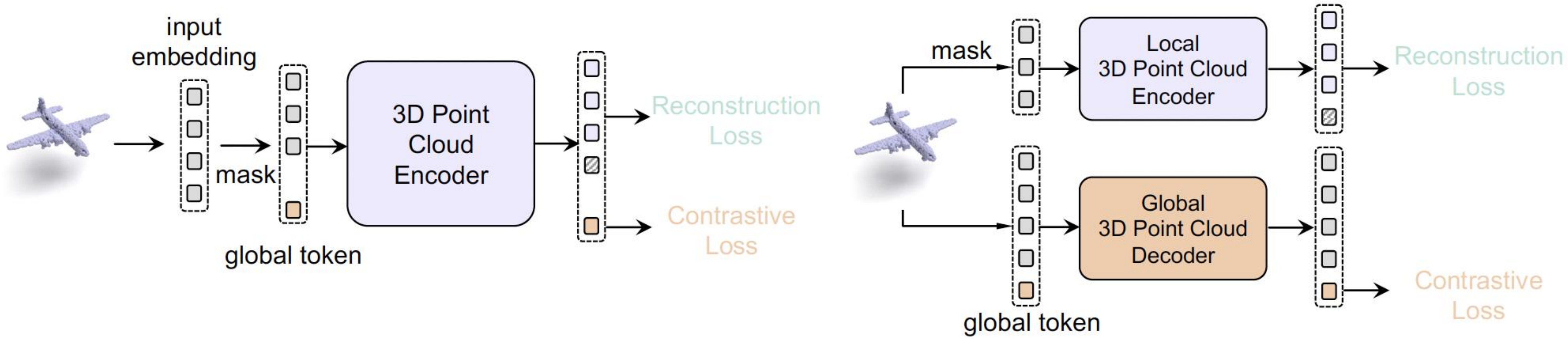
Table 2. **Few-shot classification results on ModelNet40.** † represent results of our proposed ● RECON-block built backbone architecture. Overall accuracy (%) without voting is reported.

Method	5-way		10-way	
	10-shot	20-shot	10-shot	20-shot
○ DGCNN	31.6 ± 2.8	40.8 ± 4.6	19.9 ± 2.1	16.9 ± 1.5
○ OcCo	90.6 ± 2.8	92.5 ± 1.9	82.9 ± 1.3	86.5 ± 2.2
with Self-Supervised Representation Learning (FULL)				
● Transformer	87.8 ± 5.2	93.3 ± 4.3	84.6 ± 5.5	89.4 ± 6.3
● Transformer†	90.2 ± 5.9	94.3 ± 4.4	85.2 ± 5.9	89.9 ± 6.1
○ OcCo	94.0 ± 3.6	95.9 ± 2.3	89.4 ± 5.1	92.4 ± 4.6
● Point-BERT	94.6 ± 3.1	96.3 ± 2.7	91.0 ± 5.4	92.7 ± 5.1
● MaskPoint	95.0 ± 3.7	97.2 ± 1.7	91.4 ± 4.0	93.4 ± 3.5
● Point-MAE	96.3 ± 2.5	97.8 ± 1.8	92.6 ± 4.1	95.0 ± 3.0
○ Point-M2AE	96.8 ± 1.8	98.3 ± 1.4	92.3 ± 4.5	95.0 ± 3.0
● Point-MAE†	96.4 ± 2.8	97.8 ± 2.0	92.5 ± 4.4	95.2 ± 3.9
● ACT	96.8 ± 2.3	98.0 ± 1.4	93.3 ± 4.0	95.6 ± 2.8
● RECON	$\mathbf{97.3} \pm \mathbf{1.9}$	$\mathbf{98.9} \pm \mathbf{1.2}$	$\mathbf{93.3} \pm \mathbf{3.9}$	$\mathbf{95.8} \pm \mathbf{3.0}$
with Self-Supervised Representation Learning (MLP-LINEAR)				
● Point-MAE†	91.1 ± 5.6	91.7 ± 4.0	83.5 ± 6.1	89.7 ± 4.1
● ACT	91.8 ± 4.7	93.1 ± 4.2	84.5 ± 6.4	90.7 ± 4.3
● RECON	$\mathbf{96.9} \pm \mathbf{2.6}$	$\mathbf{98.2} \pm \mathbf{1.4}$	$\mathbf{93.6} \pm \mathbf{4.7}$	$\mathbf{95.4} \pm \mathbf{2.6}$
with Self-Supervised Representation Learning (MLP-3)				
● Point-MAE†	95.0 ± 2.8	96.7 ± 2.4	90.6 ± 4.7	93.8 ± 5.0
● ACT	95.9 ± 2.2	97.7 ± 1.8	92.4 ± 5.0	94.7 ± 3.9
● RECON	$\mathbf{97.4} \pm \mathbf{2.2}$	$\mathbf{98.5} \pm \mathbf{1.4}$	$\mathbf{93.6} \pm \mathbf{4.7}$	$\mathbf{95.7} \pm \mathbf{2.7}$

Table 10. **Linear SVM classification on ModelNet40.** Overall accuracy (%) without voting is reported.

Method	Hierachical	ModelNet40
● Point-BERT (Yu et al., 2022b)	×	87.4
○ OcCo (Wang et al., 2021)	✓	89.2
○ CrossPoint (Afham et al., 2022)	✓	91.2
○ PointM2AE (Zhang et al., 2022b)	✓	92.9
● RECON	×	**93.4**

Additional Baselines



(a) vanilla multi-task learning (b) two-tower network

Table 9. **Study of the additional baseline**. Overall accuracy (%) without voting is reported.

Method	ScanObjectNN	ModelNet40
Vanilla Multi-task Learning	82.53	91.6
Two-Tower Network	85.05	92.1
ReCon	**90.63**	**94.1**

Due to the pattern difference issue, both simple combinations fail to yield satisfactory generalization performance.

西安交通大学 XI'AN JIAOTONG UNIVERSITY
清華大學 Tsinghua University
交叉信息研究院 Institute for Interdisciplinary Information Sciences
MEGVII 旷视
上海人工智能实验室 Shanghai Artificial Intelligence Laboratory

ICML | 2023
International Conference On Machine Learning

Contrast with Reconstruct

Contrastive 3D Representation Learning Guided by Generative Pretraining

Zekun Qi † Runpei Dong † Guofan Fan Zheng Ge Xiangyu Zhang Kaisheng Ma Li Yi

arXiv

GitHub

Thanks!