Autoencoders as Cross-Modal Teachers: Can Pretrained 2D Image Transformers Help 3D Representation Learning?

Runpei Dong · Zekun Qi · Linfeng Zhang · Junbo Zhang · Jianjian Sun · Zheng Ge
Li Yi · Kaisheng Ma

International Conference on Learning Representations (ICLR)
May 1st – 5th, 2023, Kigali Rwanda

Foundation Models in 2D vision and NLP

Google
BERT

OpenAI
ChatGPT

Meta

Source: (left) web; (middle) web; (right) Rob Mulla YouTube

Question

What makes 3D representation learning more challenging than 2D vision or NLP?

Motivation: Open Challenges & Issues

- Data Dessert & Pattern Difference

(a) Natural Languages

A: There is a plane on the runway.

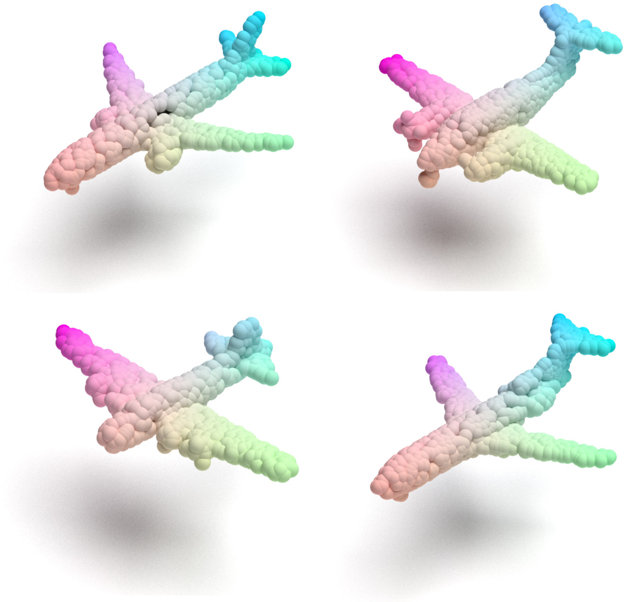
Q: How to describe the right pictures?

A: Several planes at the airport.

(b) 2D RGB Images

RGB Semantic

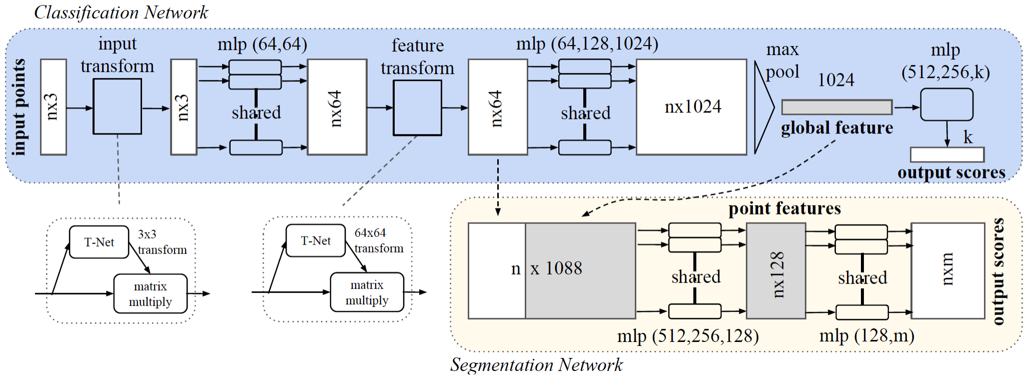
(c) 3D Point Clouds



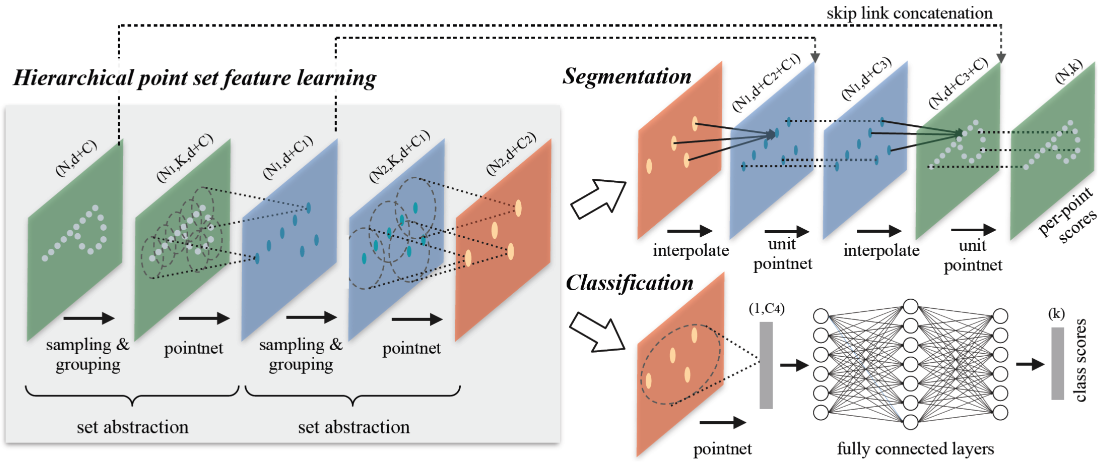
	(a) Natural Languages	(b) 2D RGB Images	(c) 3D Point Clouds
Format	Free Form Words	Regular Pixels	Cartesian Coordinates
Scale	Broad	Large	Moderate
Semantics	Dense & Structured	Sparse & Unstructured	Sparse & Unstructured

Motivation: Open Challenges & Issues

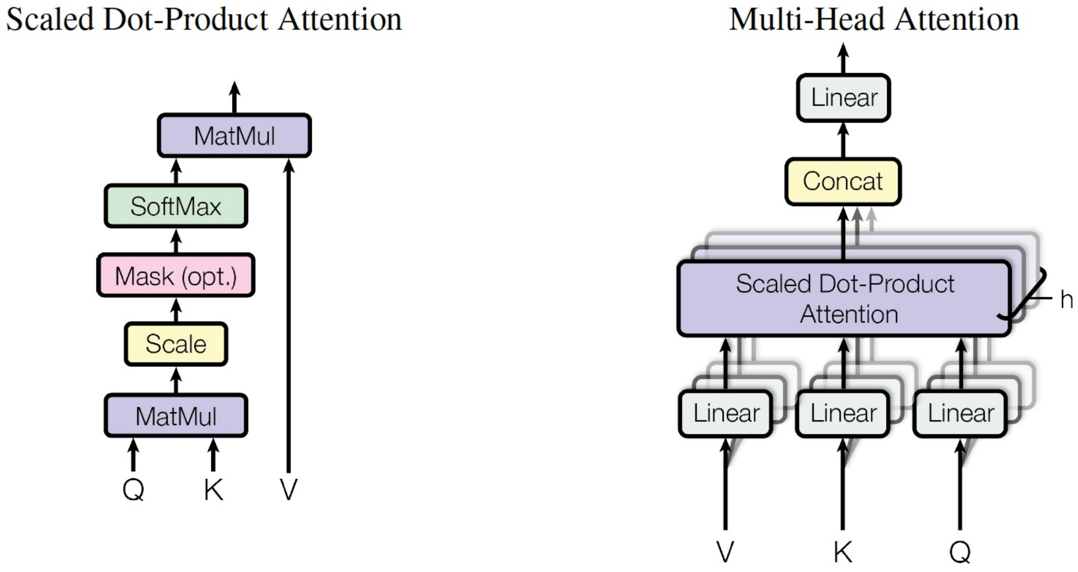
- Architectural Disunity



PointNet, 2017



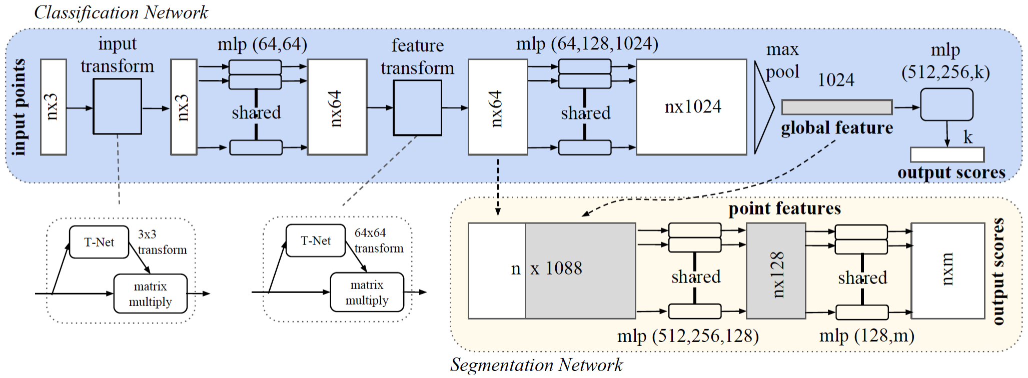
PointNet++, 2017



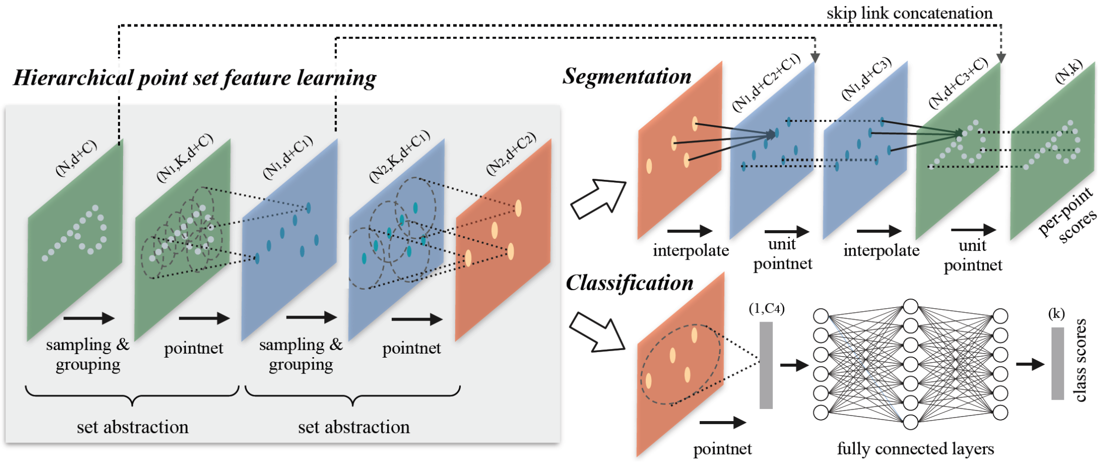
Atten is All You Need, 2017

Motivation: Open Challenges & Issues

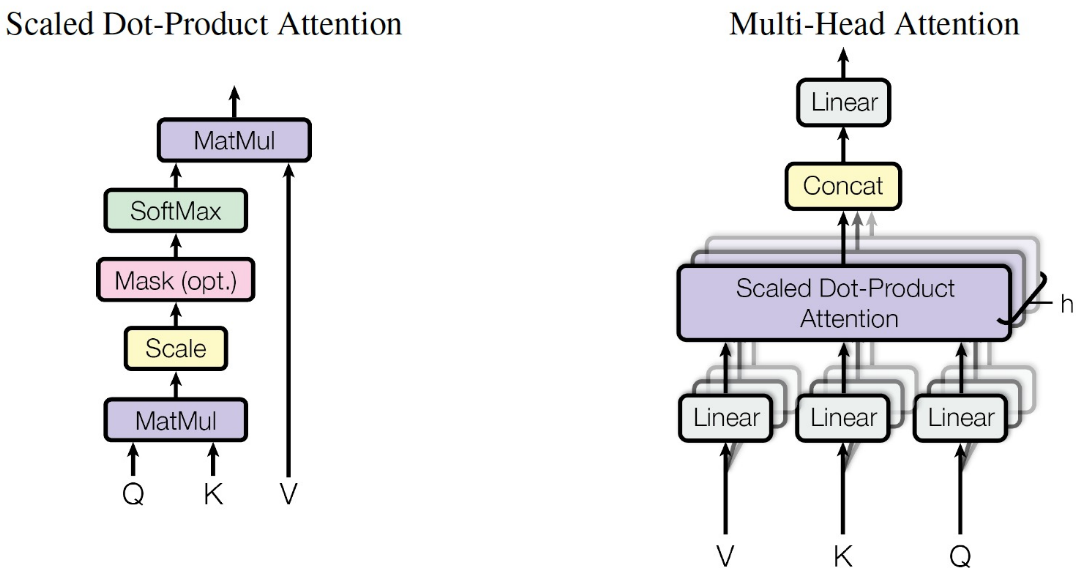
- Architectural Disunity



PointNet, 2017



PointNet++, 2017



Atten is All You Need, 2017

ACT: Autoencoders as Cross-Modal Teachers

- Preliminary: A Unified View of Masked Modeling with KD

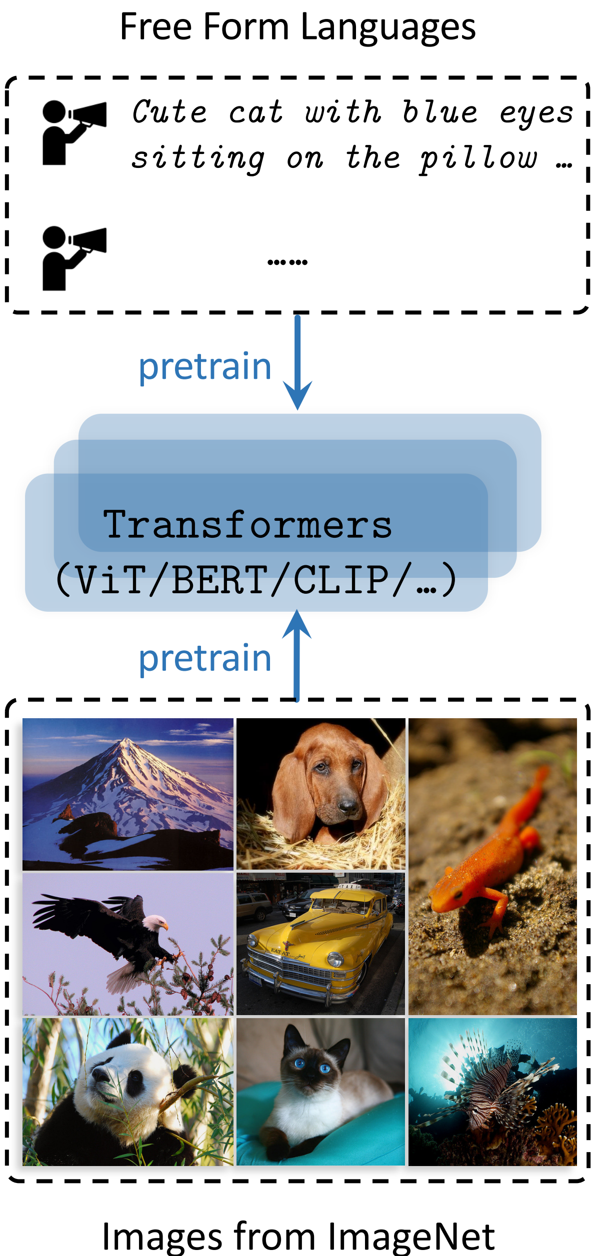
Data Definition

- Student network: $f_{\mathcal{S}}$; Teacher network: $f_{\mathcal{T}}$
- A sequence of N_t tokens: $\mathbf{T} = \{\mathbf{t}_i | i = 1, 2, \ldots, N_t\}$
- A set of masks: $\mathcal{M} = \{m_i | i = 1, 2, \ldots, N_t\} \in \{0, 1\}^{N_t}$
- A learnable corruption embedding: $\boldsymbol{e}^{[\mathrm{M}]}$
- Corrupted input: $\mathbf{Z}^{\mathcal{M}} = \mathbb{1}(\mathcal{M}) \odot \boldsymbol{e}^{[\mathrm{M}]} + \mathbb{1}(1 - \mathcal{M}) \odot \mathbf{T}$
- Distance function defined in some metric space $\mathbb{D}$: $\mathcal{L}_{\mathbb{D}}(\cdot, \cdot)$
- Decoders: $h_{\mathcal{S}}$ and $h_{\mathcal{T}}$

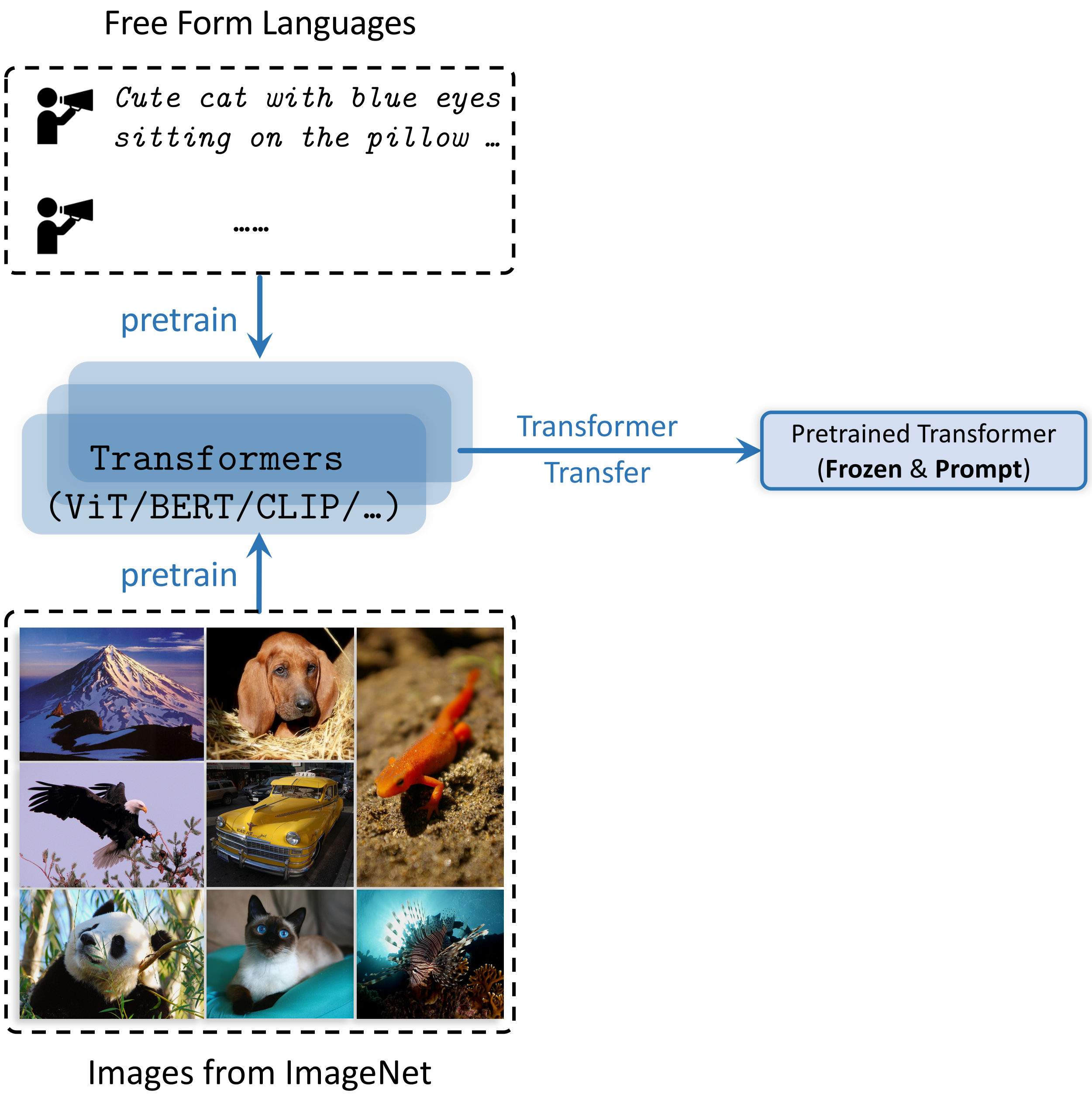
Unified Objective (minimize)

$$-\sum_{i=1}^{N_t} m_i \cdot \mathcal{L}_{\mathbb{D}}\left(h_{\mathcal{S}} \circ f_{\mathcal{S}}(\mathbf{Z}^{\mathcal{M}}), h_{\mathcal{T}} \circ f_{\mathcal{T}}(\mathbf{T})\right)$$

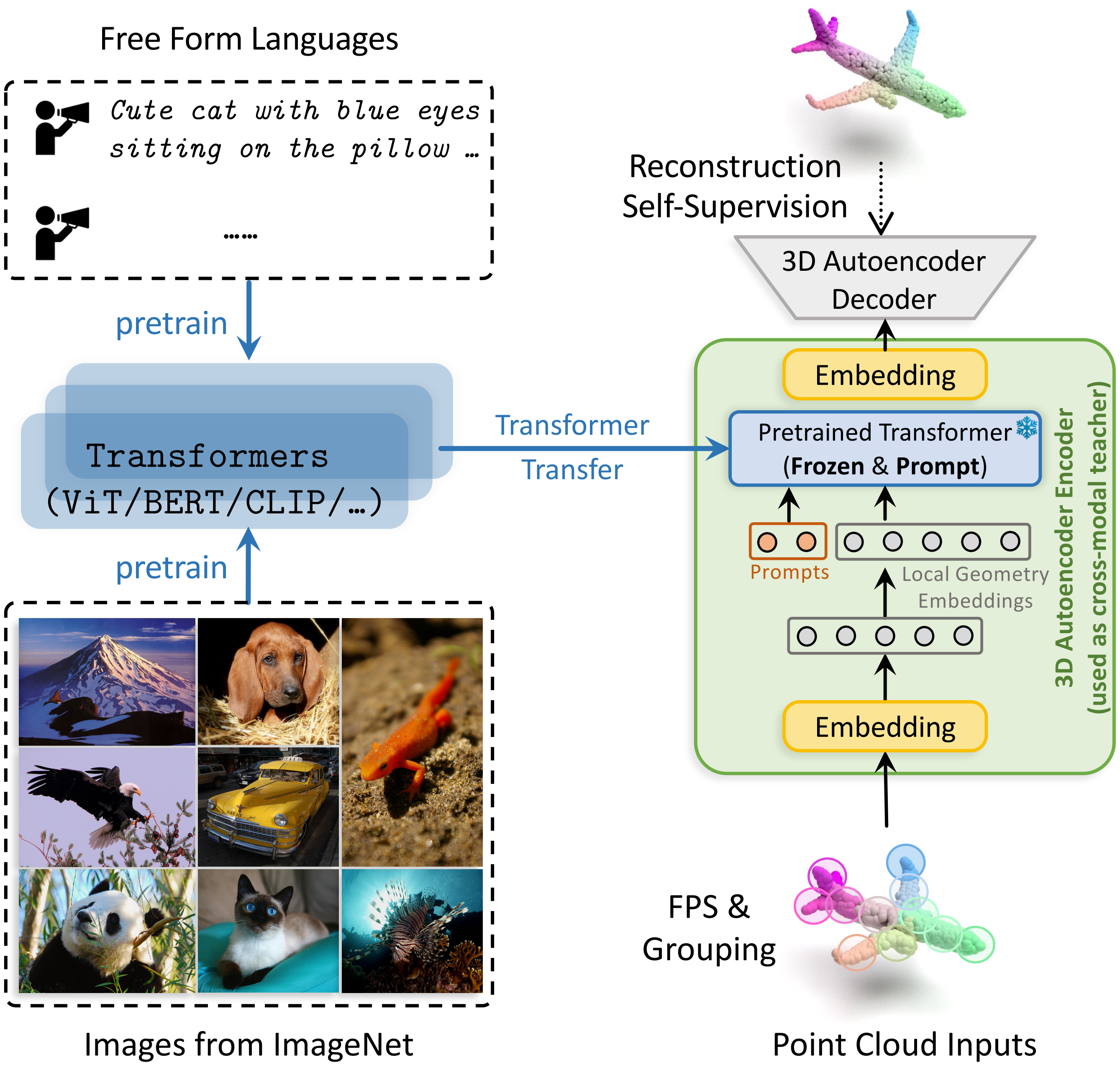
ACT: Autoencoders as Cross-Modal Teachers



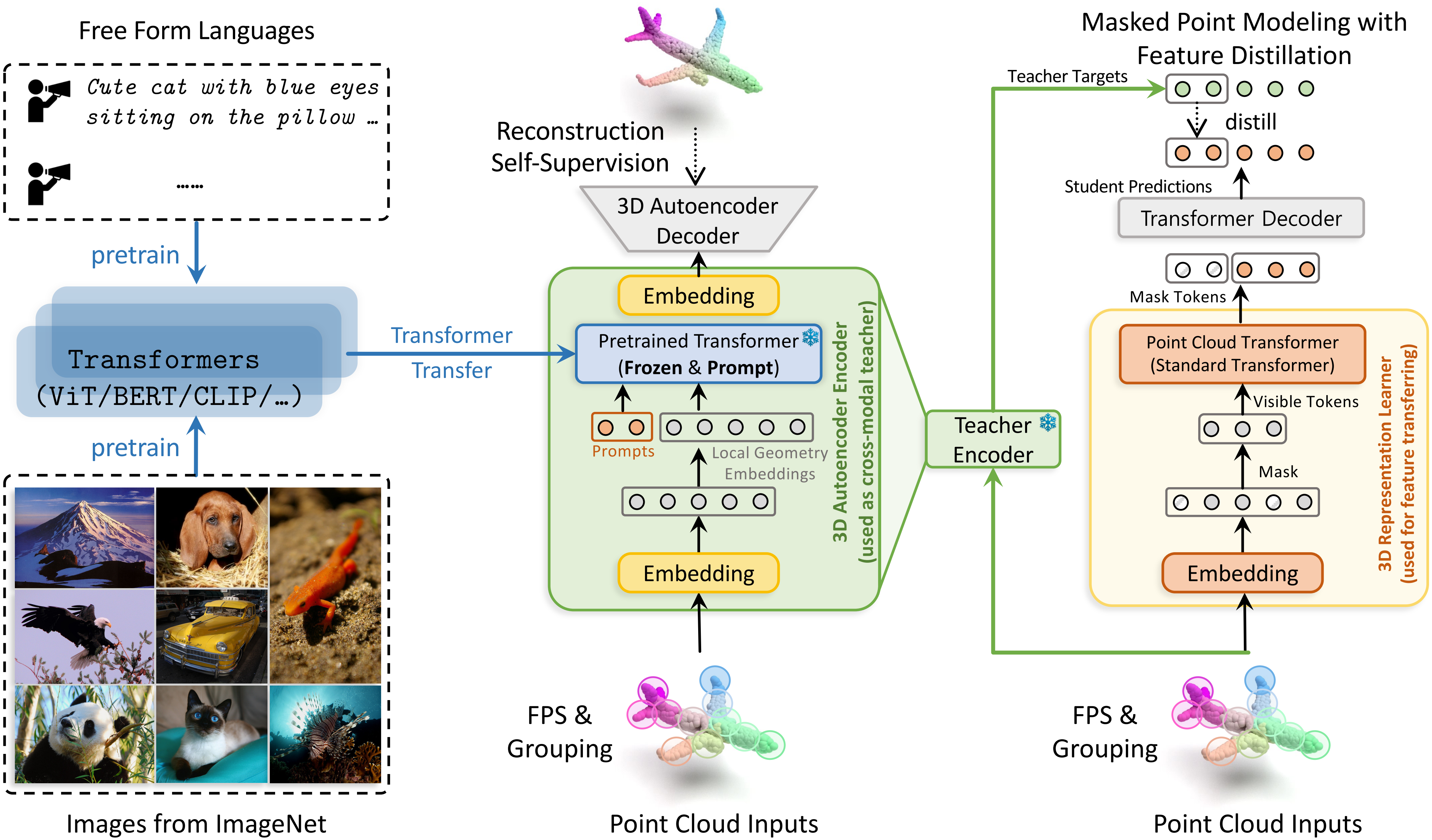
ACT: Autoencoders as Cross-Modal Teachers



ACT: Autoencoders as Cross-Modal Teachers



ACT: Autoencoders as Cross-Modal Teachers



Experimental Validation

Table 2: Classification results on ScanObjectNN. Ours[1]: results trained with no data augmentation. Ours[2]: results trained with simple point cloud rotation. DA: data augmentation is used during fine-tuning training. The overall accuracy, *i.e.*, OA (%) is reported.

Method	#Params(M)	DA	OBJ_BG	OBJ_ONLY	PB_T50_RS
Supervised Learning Only					
PointNet (Qi et al., 2017a)	3.5	✓	73.3	79.2	68.0
SpiderCNN (Xu et al., 2018)	-	✓	77.1	79.5	73.7
PointNet++ (Qi et al., 2017b)	1.5	✓	82.3	84.3	77.9
DGCNN (Wang et al., 2019)	1.8	✓	82.8	86.2	78.1
PointCNN (Li et al., 2018)	0.6	✓	86.1	85.5	78.5
BGA-DGCNN (Uy et al., 2019a)	1.8	✓	-	-	79.7
BGA-PN++ (Uy et al., 2019a)	1.5	✓	-	-	80.2
DRNet (Qiu et al., 2021)	-	✓	-	-	80.3
GBNet (Qiu et al., 2022)	8.8	✓	-	-	80.5
SimpleView (Goyal et al., 2021)	-	✓	-	-	80.5±0.3
PRANet (Cheng et al., 2021)	2.3	✓	-	-	81.0
MVTN (Hamdi et al., 2021)	-	✓	-	-	82.8
PointMLP (Ma et al., 2022)	13.2	✓	-	-	85.4±0.3
with Self-Supervised Representation Learning (FULL)					
Transformer (Vaswani et al., 2017)	22.1	✓	79.86	80.55	77.24
OcCo (Wang et al., 2021)	22.1	✓	84.85	85.54	78.79
Point-BERT (Yu et al., 2022)	22.1	✓	87.43	88.12	83.07
MaskPoint (Liu et al., 2022a)	22.1	✓	89.30	88.10	84.30
Point-MAE (Pang et al., 2022)	22.1	✓	90.02	88.29	85.18
ACT (Ours[1])	22.1	×	**91.22**	**89.16**	**85.81**
ACT (Ours[2])	22.1	✓	**93.29**	**91.91**	**88.21**
Point-MAE (Pang et al., 2022)	22.1	✓	89.31±0.41	87.88±0.36	84.35±0.31
ACT (Ours[1])	22.1	×	**90.06**±0.56	**89.02**±0.22	**85.33**±0.27
ACT (Ours[2])	22.1	✓	**92.48**±0.59	**91.57**±0.37	**87.88**±0.36
with Self-Supervised Representation Learning (MLP-LINEAR)					
Point-MAE (Pang et al., 2022)	22.1	✓	82.58±0.58	83.52±0.41	73.08±0.30
ACT (Ours[1])	22.1	×	**82.71**±0.45	**84.34**±0.29	**74.17**±0.05
ACT (Ours[2])	22.1	✓	**85.20**±0.83	**85.84**±0.15	**76.31**±0.26
with Self-Supervised Representation Learning (MLP-3)					
Point-MAE (Pang et al., 2022)	22.1	✓	84.29±0.55	85.24±0.67	77.34±0.12
ACT (Ours[1])	22.1	×	**85.67**±0.29	**86.79**±0.30	**78.89**±0.22
ACT (Ours[2])	22.1	✓	**87.14**±0.22	**88.90**±0.40	**81.52**±0.19

Table 5: Few-shot classification on ModelNet40, overall accuracy (%) 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
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
Point-MAE	96.3 ± 2.5	97.8 ± 1.8	92.6 ± 4.1	95.0 ± 3.0
ACT (Ours)	**96.8 ± 2.3**	**98.0 ± 1.4**	**93.3 ± 4.0**	**95.6 ± 2.8**
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 (Ours)	**91.8 ± 4.7**	**93.1 ± 4.2**	**84.5** ± 6.4	**90.7** ± 4.3
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 (Ours)	**95.9 ± 2.2**	**97.7 ± 1.8**	**92.4** ± 5.0	**94.7 ± 3.9**

Table 11: 3D object detection on the ScanNetV2 dataset. The detection performance using mean Average Precision (mAP) at two different IoU thresholds of 0.50 and 0.25, *i.e.*, AP_{50} and AP_{25} are reported. xyz: point cloud coordinates are used.

Method	SSL	Input	AP_{50}	AP_{25}
VoteNet (Qi et al., 2019)	×	xyz	33.5	58.6
PointContrast (Xie et al., 2020)	✓	xyz	38.0	59.2
STRL (Huang et al., 2021)	✓	xyz	38.4	59.5
RandomRooms (Rao et al., 2021)	✓	xyz	36.2	61.3
DepthContrast (Zhang et al., 2021)	✓	xyz	-	61.3
3DETR (Misra et al., 2021)	×	xyz	37.9	62.1
Point-BERT (Yu et al., 2022)	✓	xyz	38.3	61.0
MaskPoint (Liu et al., 2022a)	✓	xyz	40.6	63.4
ACT (Ours)	✓	xyz	**42.1**	**63.8**

Autoencoders as Cross-Modal Teachers: Can Pretrained 2D Image Transformers Help 3D Representation Learning?

Runpei Dong · Zekun Qi · Linfeng Zhang · Junbo Zhang · Jianjian Sun · Zheng Ge
Li Yi · Kaisheng Ma

GitHub

Thanks!

Please Stop by at MH1-2-3-4 #75 for more details
or contact me if you have any questions.
runpei.dong@gmail.com

ArChip-Lab
Algorithm ARchitecture Chip

MEGVII 旷视

上海 期智研究院
SHANGHAI QI ZHI INSTITUTE

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