JUNE 18-22, 2023
CVPR
VANCOUVER, CANADA

PiMAE: Point Cloud and Image Interactive Masked Autoencoders for 3D Object Detection

Anthony Chen[1,2,*] · Kevin Zhang[1,2,*] · Renrui Zhang[3]
Zihan Wang[1,2] · Yuheng Lu[1,2,4] · Yandong Guo[5] · Shanghang Zhang[1,2,†]

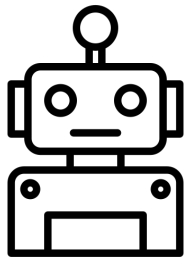
CVPR2023 Paper Tag: **TUE-PM-110**

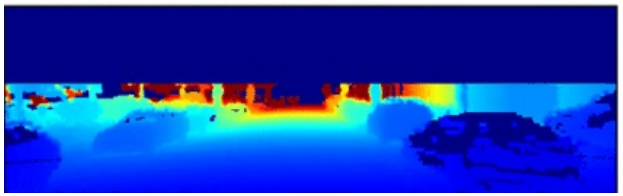
[1]National Key Laboratory for Multimedia Information Processing
[2]Peking University, [3]The Chinese University of Hong Kong,
[4]Wukong Lab, iKingtec, [5]Beijing University of Posts and Telecommunications

Background

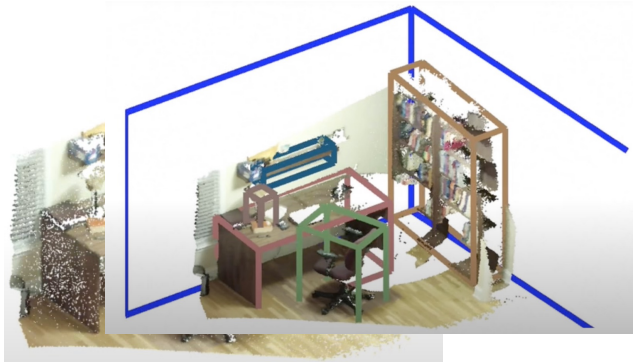
- Unlabeled point cloud and image data

Collects





Expensive Annotation!



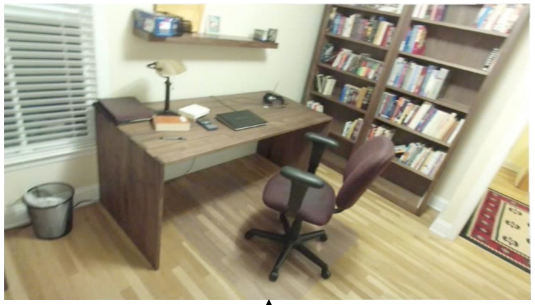
Contribution

JUNE 18-22, 2023
CVPR
VANCOUVER, CANADA

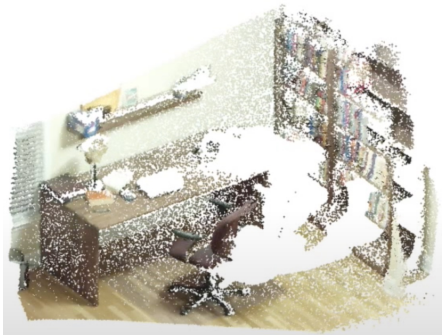
- We propose a novel multi-modal self-supervised pretraining scheme on unlabeled point cloud and image data.
- We **first** extend Masked Autoencoders (MAE) from image pretraining to point cloud & image multi-modal pretraining with three novel cross-modal interaction designs, including a complementary cross-modal masking strategy, a modal shared-decoder, and a cross-modal reconstruction task.
- Our pretrained models boosts performance of 2D & 3D detectors by a large margin.

Motivation

- Naturally paired point cloud and image data collected by RGBD sensor

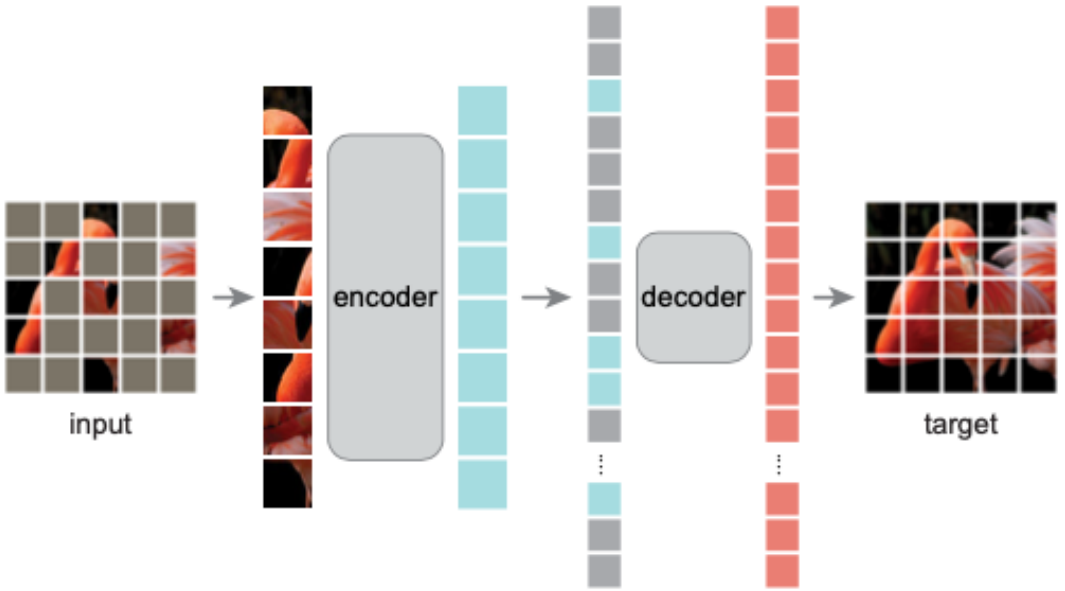


Paired



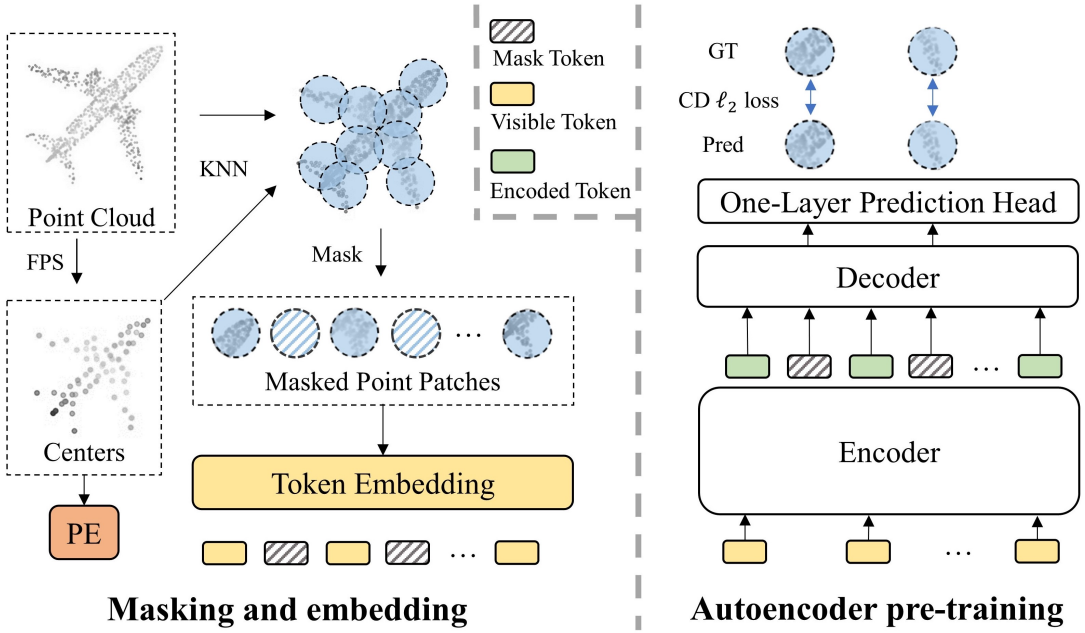
Motivation

- A suitable self-supervised method
- Contrastive Methods? No
- Masked Autoencoders? Yes



Motivation

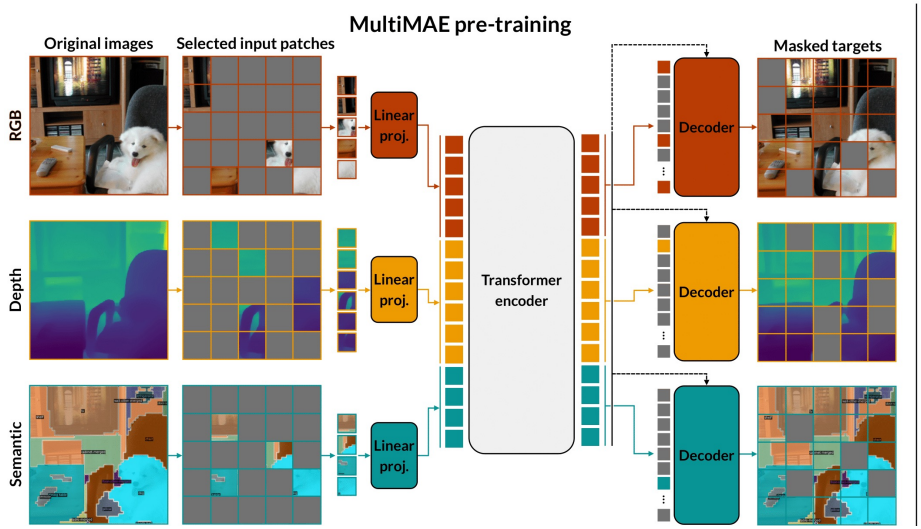
- Masked Autoencoders on Point cloud & Image pretraining not yet exist



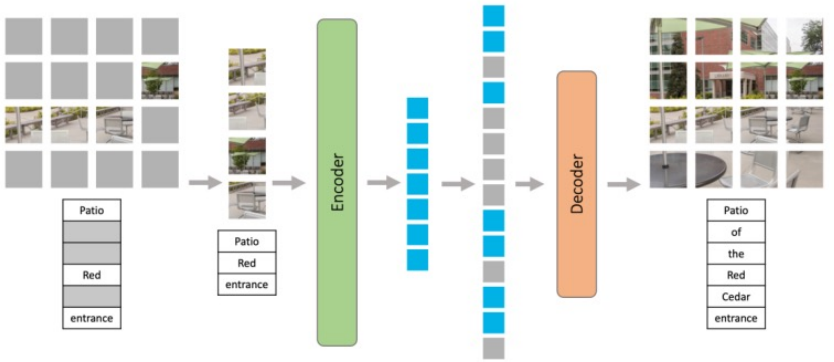
PointMAE (ECCV2022)

Motivation

- Current Multi-modal Masked Autoencoders methods are lack of modality interactions

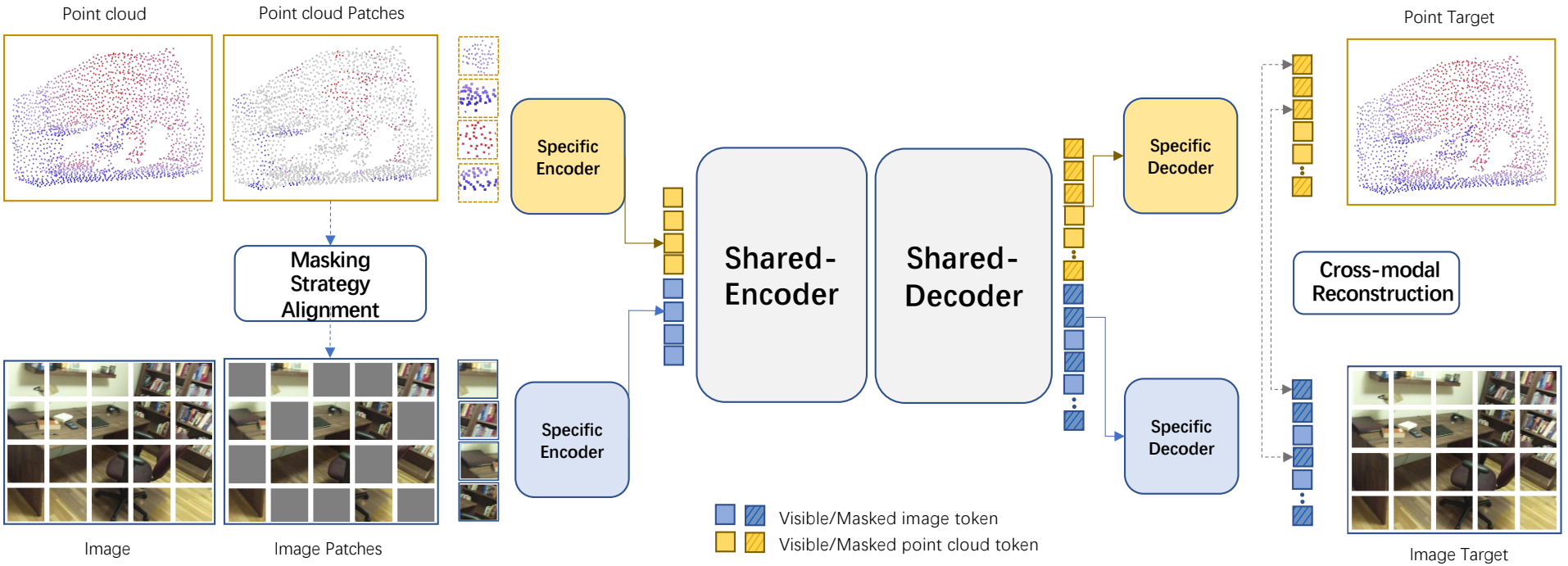


MultiMAE (ECCV2022)

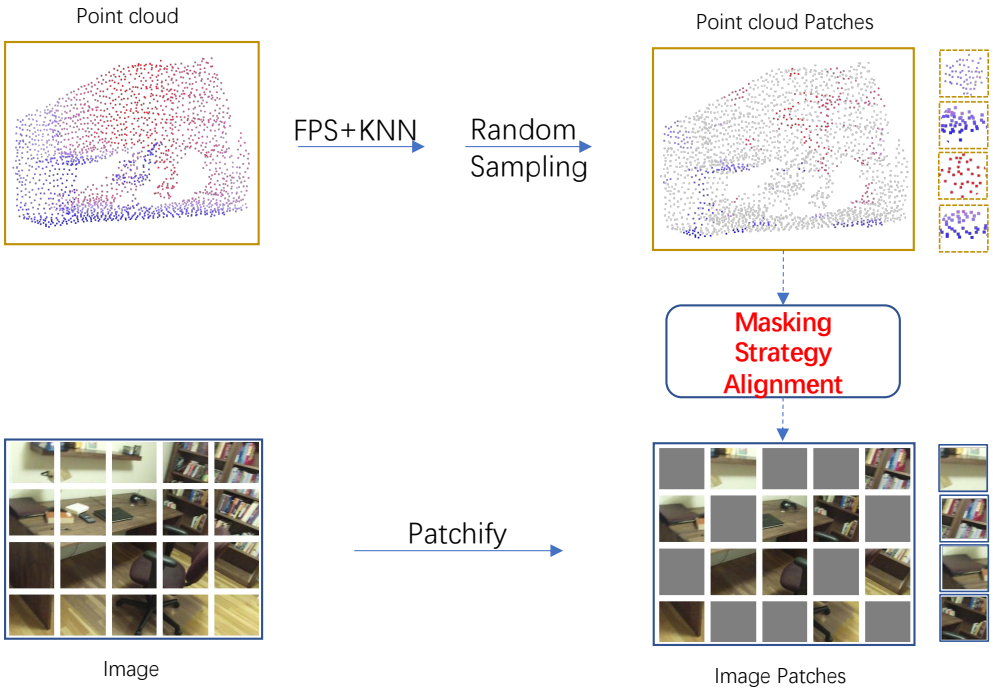


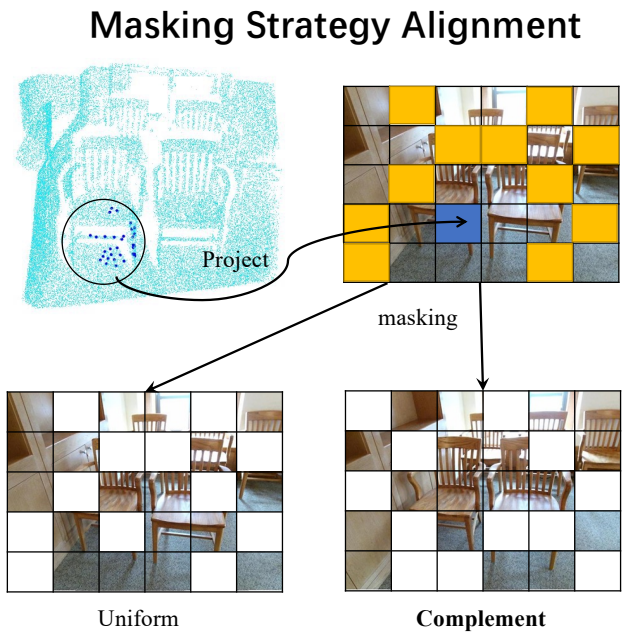
M3AE

Methods Overview

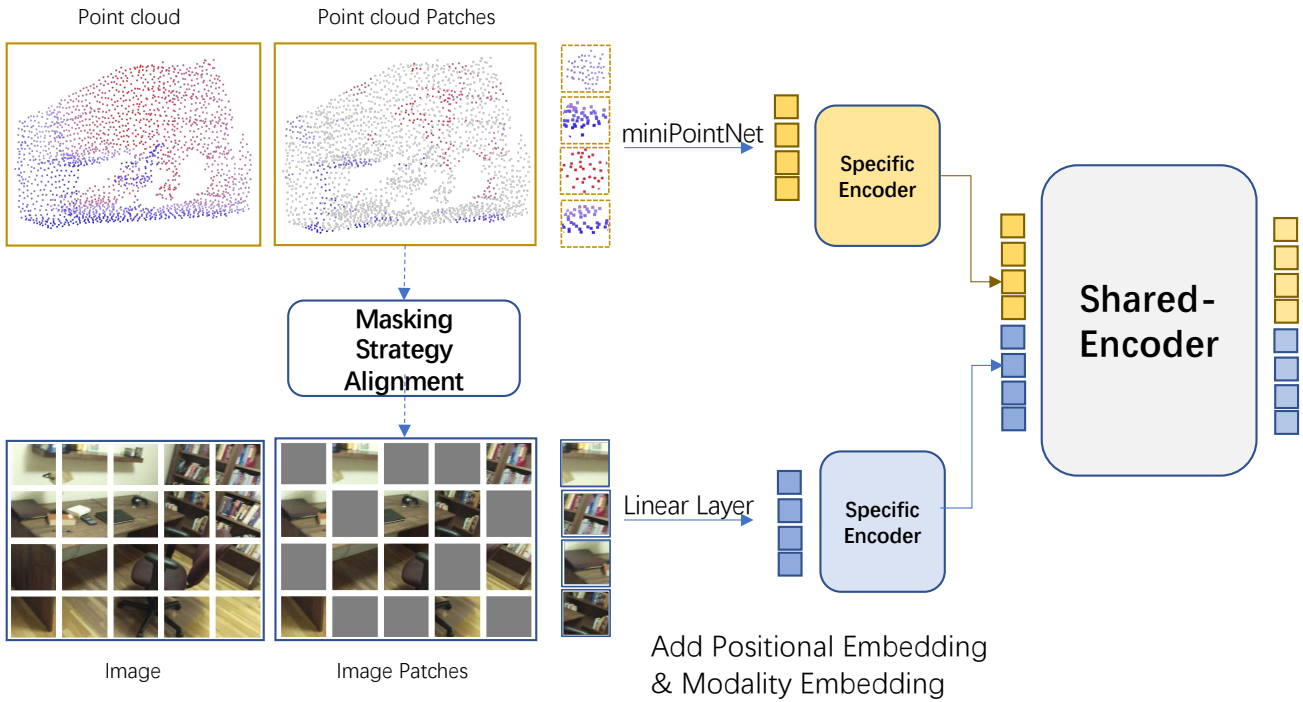


Step1: Embedding & Masking



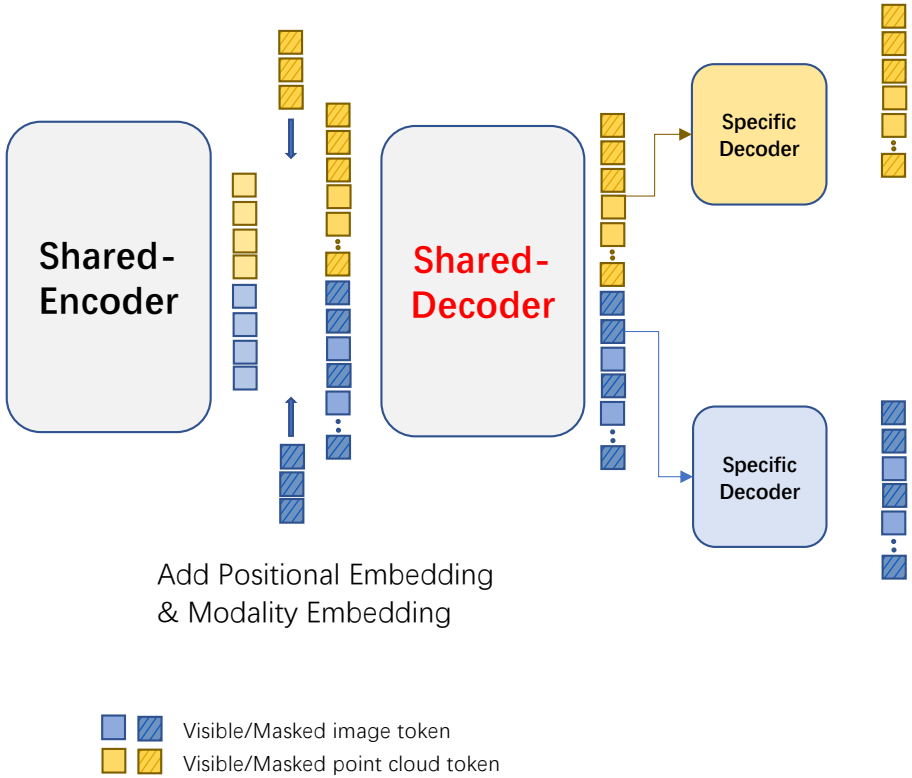


Step2: Encoding

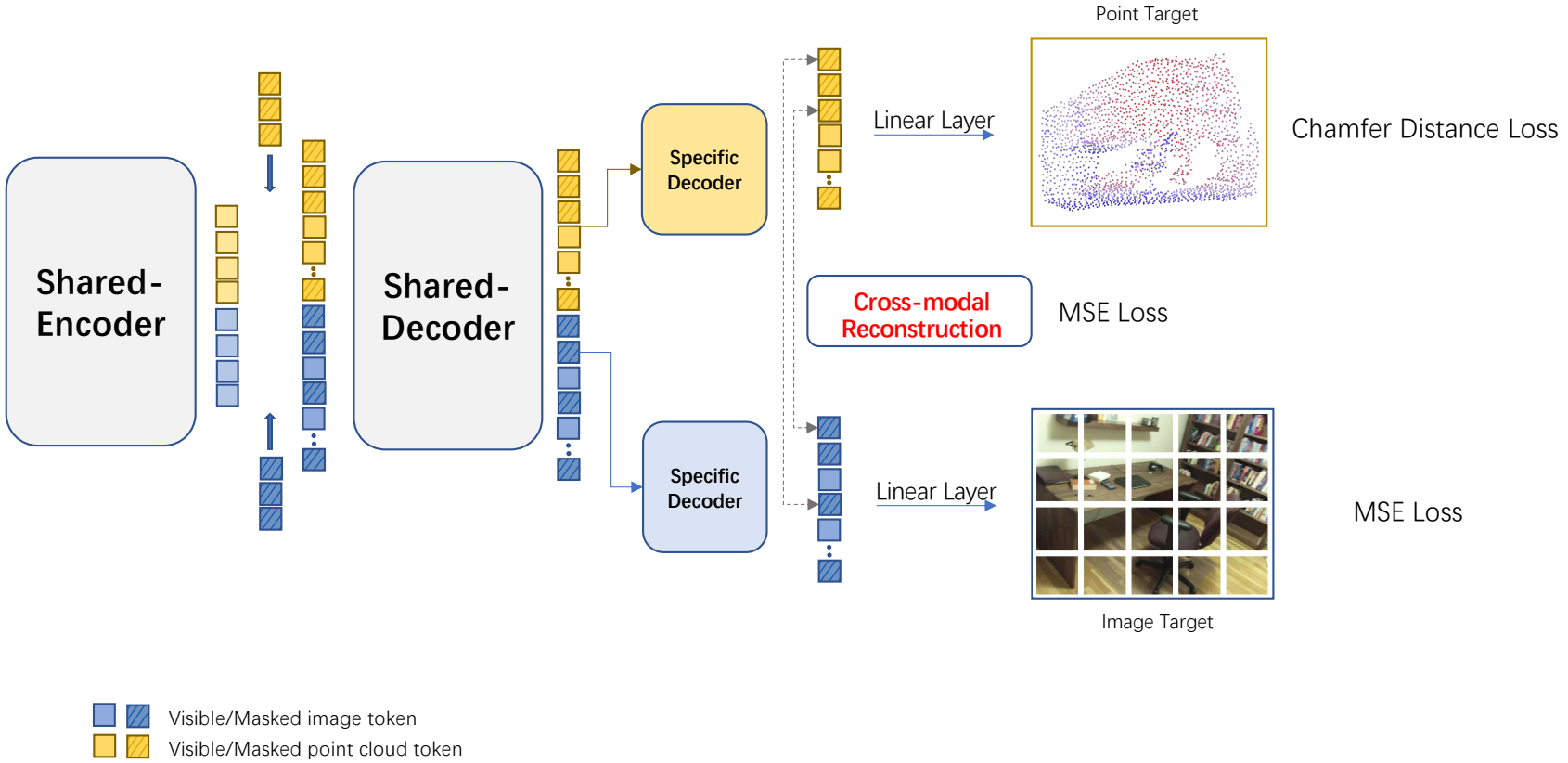


JUNE 18-22, 2023
CVPR
VANCOUVER, CANADA

Step3: Decoding



Step4: Reconstructing



Experiments

- Unsupervised Pretraining on Point Cloud & Image dataset
  - Dataset: SUNRGB-D
  - Epochs: 400

- Downstream Finetuning
  - Dataset: SUNRGB-D, ScanNetV2, KITTI, CIFAR-FS, miniImageNet, FC100
  - Baseline: DETR, 3DETR, MonoDETR, GroupFree3D
  - Tasks: 2D/3D Object Detection, Few-shot Image Classification

Experiments

- 3D Object Detection

Methods	Pre-trained	SUN RGB-D		ScanNetV2	
		AP_{25}	AP_{50}	AP_{25}	AP_{50}
DSS [52]	*None*	42.1	-	15.2	6.8
PointFusion [61]	*None*	45.4	-	-	-
3D-SIS [23]	*None*	-	-	40.2	22.5
VoteNet [43]	*None*	57.7	32.9	58.6	33.5
3DETR [39]	*None*	58.0	30.3	62.1	37.9
+Ours(from scratch)	*None*	58.7	31.7	59.7	40.0
+Ours	SUN RGB-D	59.4(+1.4)	33.2(+2.9)	62.6(+0.5)	39.4(+1.5)
GroupFree3D [35]	*None*	63.0	45.2	67.3	48.9
+Ours(from scratch)	*None*	61.2	44.7	65.5	47.4
+Ours	SUN RGB-D	**64.6(+1.6)**	**46.2(+1.0)**	**67.6(+0.3)**	**49.7(+0.8)**

Experiments

- 2D Object Detection on ScanNetV2
- Monocular 3D Object Detection on KITTI (Out of Distribution data)

Methods	AP_{50}	AP_{75}	AP
*DETR [5]	39.8	26.2	25.3
+ PiMAE	**46.5(+6.7)**	**30.3(+4.1)**	**29.5(+4.2)**

Methods	Easy	Mod.	Hard
*MonoDETR [66]	23.1	17.3	14.5
+ PiMAE	**26.6(+3.5)**	**18.8(+1.5)**	**15.5(+1.0)**

Experiments

- Few-shot Image Classification

Method	CIFAR-FS 5-way		FC100 5-way		miniImageNet 5-way	
	1-shot	5-shot	1-shot	5-shot	1-shot	5-shot
MAML [12]	58.9	71.5	-	-	48.7	63.1
Matching Networks [55]	-	-	-	-	43.6	55.3
Prototypical Network [50]	55.5	72.0	35.3	48.6	49.3	68.2
Relation Network [53]	55.0	69.3	-	-	50.4	65.3
CrossPoint [1]	64.5	80.1	-	-	-	-
PiMAE From Scratch	62.4	76.6	37.3	50.5	50.1	66.7
PiMAE Pre-trained	**66.9**	**80.7**	**39.0**	**53.3**	**55.3**	**70.2**

Experiments

JUNE 18-22, 2023
CVPR
VANCOUVER, CANADA

- Ablation Study (Masking Strategy[1], Reconstruction Target[2], Modality Influence[3], Data Efficiency[4]…)

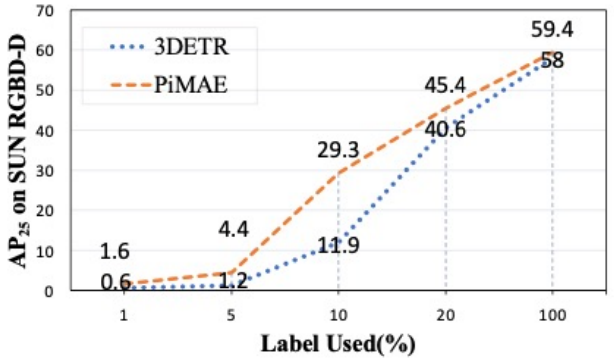
1

Masking Strategy	AP_{25}	AP_{50}
Random	58.0	32.9
Uniform	58.1	32.6
Complement	**59.0**	**33.0**

2

Point Cloud			RGB	AP_{25}	AP_{50}
3D Geo	2D feat	2D pix	2D pix		
✓			✓	59.0	33.0
✓		✓	✓	58.0	31.6
✓	✓		✓	**59.4**	**33.2**

4



3

Input	3D Object Detection		Few-shot image classification	
	AP_{25}	AP_{50}	5-way 1-shot	5-way 5-shot
RGB	-	-	66.3	79.5
Geo	58.4	32.3	-	-
RGB+Geo	**59.4**	**33.2**	**66.9**	**80.7**

Experiments

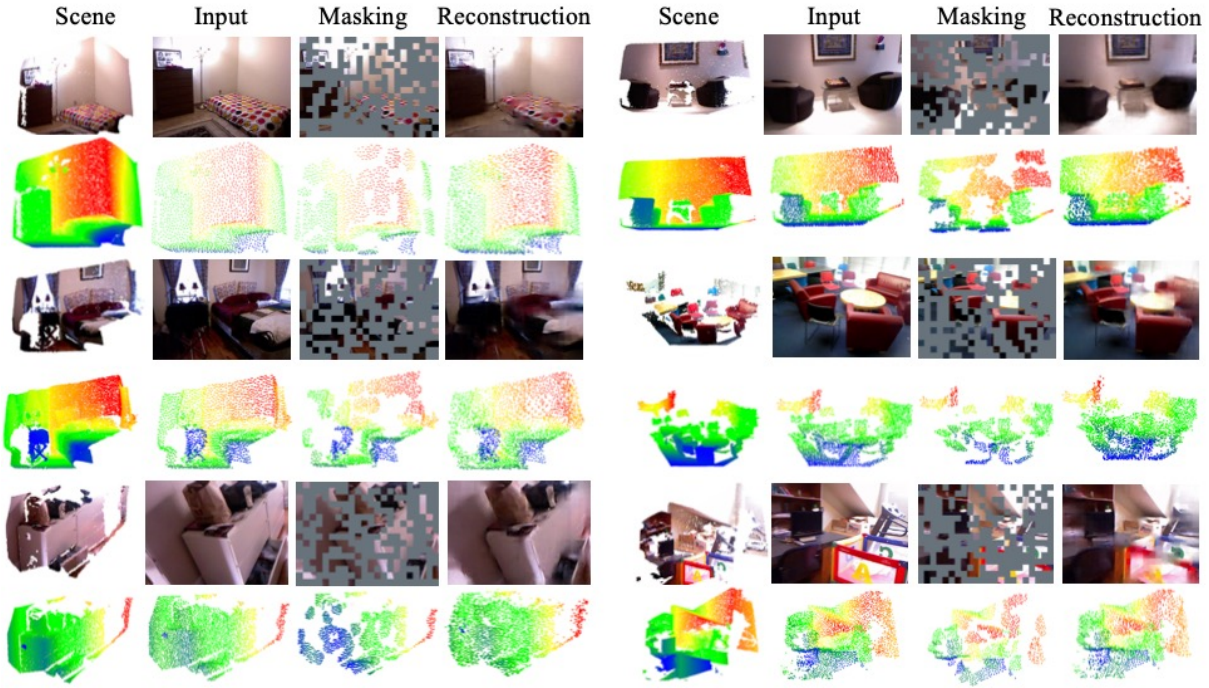
- Ablation Study (Masking Strategy[1], Reconstruction Target[2], Modality Influence[3], Data Efficiency[4]…)

Encoder	Decoder	AP_{25}	AP_{50}
3+3	0+3	58.0	30.2
3+3	1+2	**59.4**	**33.2**
3+3	1+3	58.1	32.8
2+2	1+2	57.5	30.8

Mask Ratio	AP_{25}	AP_{50}
50%	58.7	33.1
60%	**59.4**	**33.2**
70%	58.4	33.0
80%	57.5	32.4

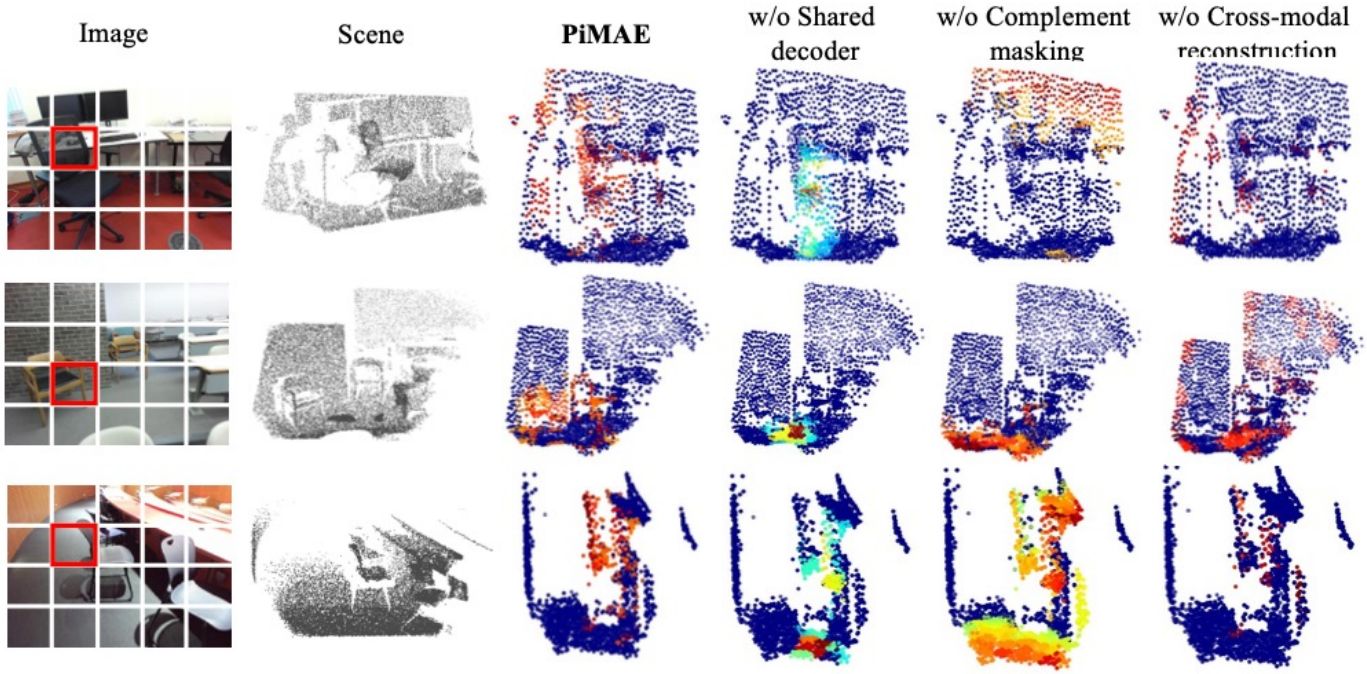
Visualization

- Mask Reconstruction, Cross-modal Attention Features, 3D Object Detection



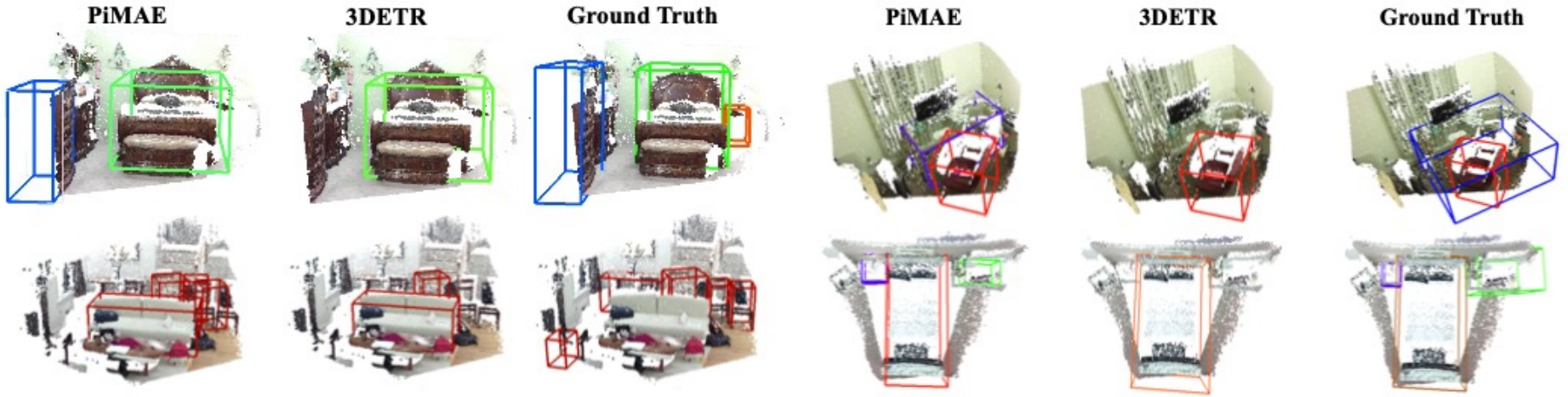
Visualization

- Mask Reconstruction, Cross-modal Attention Features, 3D Object Detection



Visualization

- Mask Reconstruction, Cross-modal Attention Features, 3D Object Detection



Conclusion

- We are the first to explore pre-training MAE with point cloud and RGB modalities interactively with three novel schemes, including a complementary cross-modal masking strategy, a modal shared-decoder, and a cross-modal reconstruction task.

- In our extensive experiments and ablation studies performed on datasets of both modalities, we discover that PiMAE has great potential, improving multiple baselines and tasks.

thanks for watching!

Anthony Chen
Contact: antonchen@pku.edu.cn