Harnessing Text-to-Image Diffusion Models for Point Cloud Self-Supervised Learning

Yiyang Chen[1], Shanshan Zhao[2†], Lunhao Duan[2], Changxing Ding[1,3†], Dacheng Tao[4]
[1]South China University of Technology, [2]Alibaba International Digital Commerce Group, [3]Pazhou Lab, [4]Nanyang Technological University

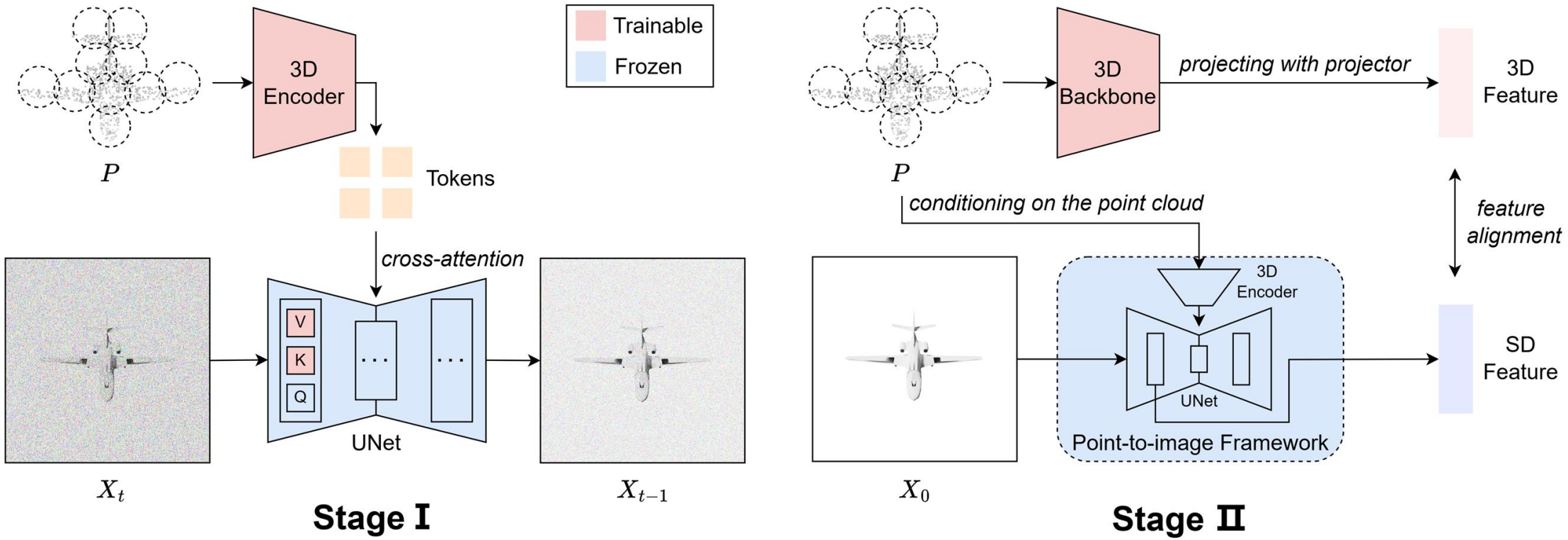
Introduction

Motivation:
1) Diffusion models excel in 2D text-to-image generation and 2D representation learning.
2) Extending them to 3D self-supervised learning is limited by small 3D datasets.
3) Large-scale text-to-image models, like Stable Diffusion (SD), may enhance 3D representation learning.

Our approach propose PointSD, leveraging SD for 3D self-supervised learning.

Method

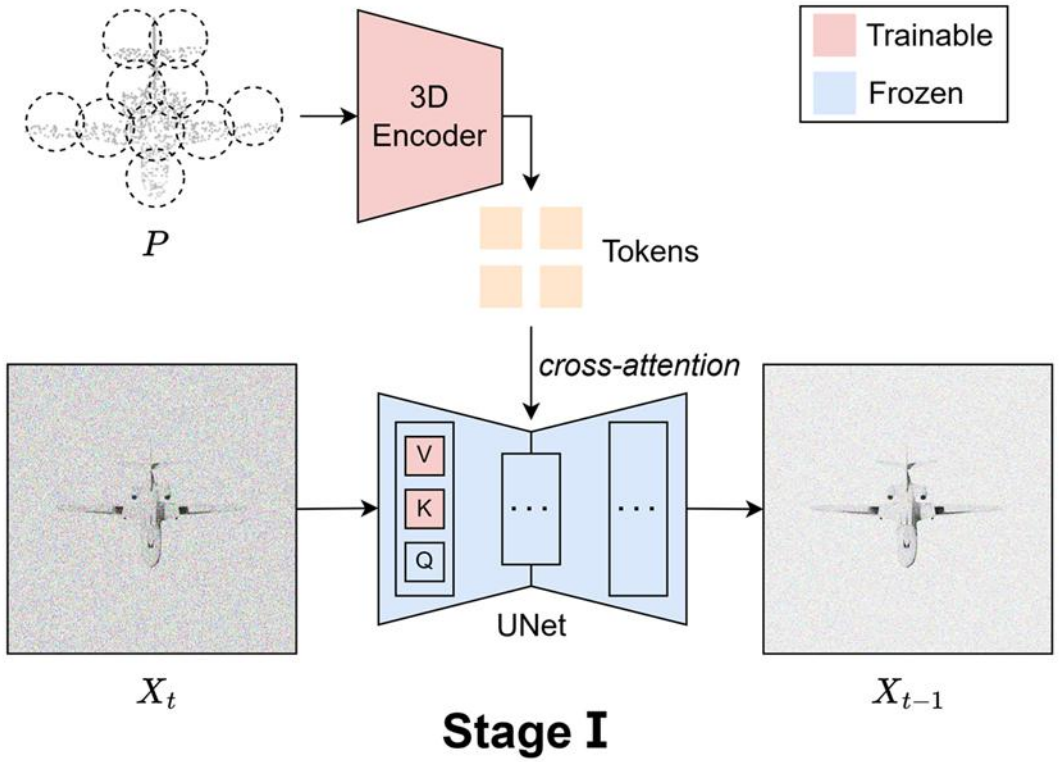
- **Framework**



Stage I: 3D encoder extracts features to guide denoising, forming a point-to-image diffusion framework.
Stage II: Feature alignment is performed between 3D features and SD features to boost 3D representation learning.

Method

- **Stage I: Text-to-image ⇒ Point-to-image**



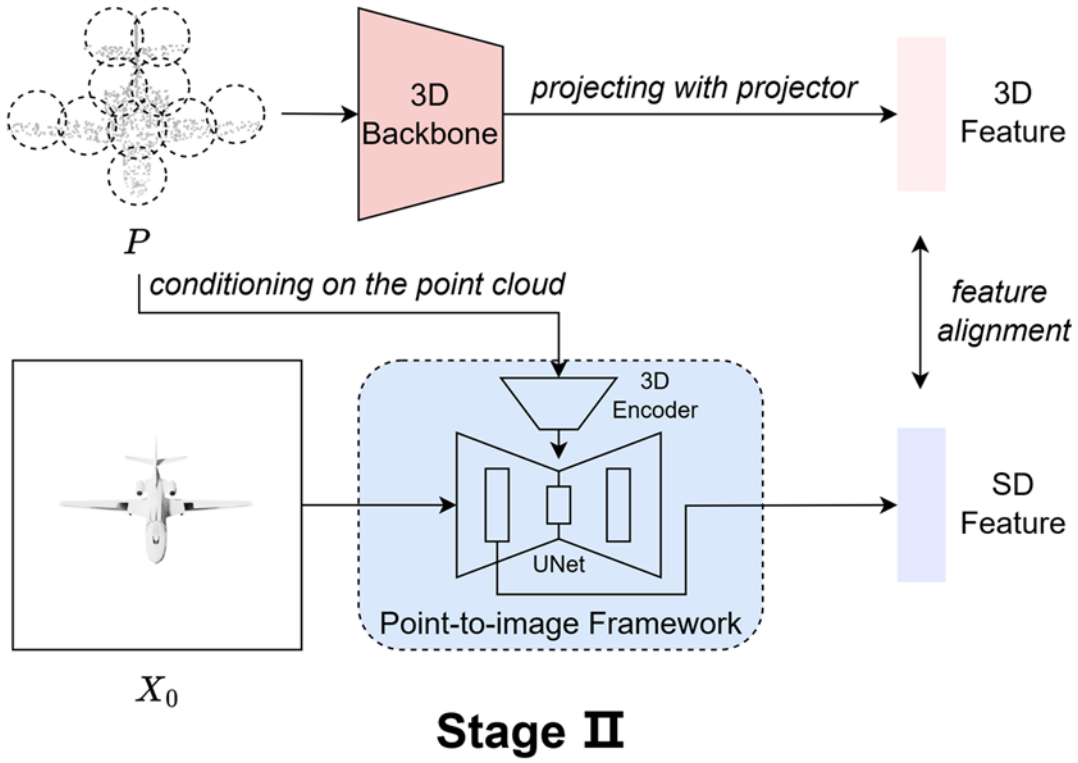
- Replace the text encoder with a 3D encoder and train a point-to-image diffusion model based on Stable Diffusion.

- Use cross-attention to let 3D features guide the denoising of images rendered from point clouds.

- **Limitation**: Image denoising inevitably captures low-level textures, which may distract the 3D backbone from high-level semantic features.

Method

- **Stage II: Point Cloud Learning via Feature Alignment**

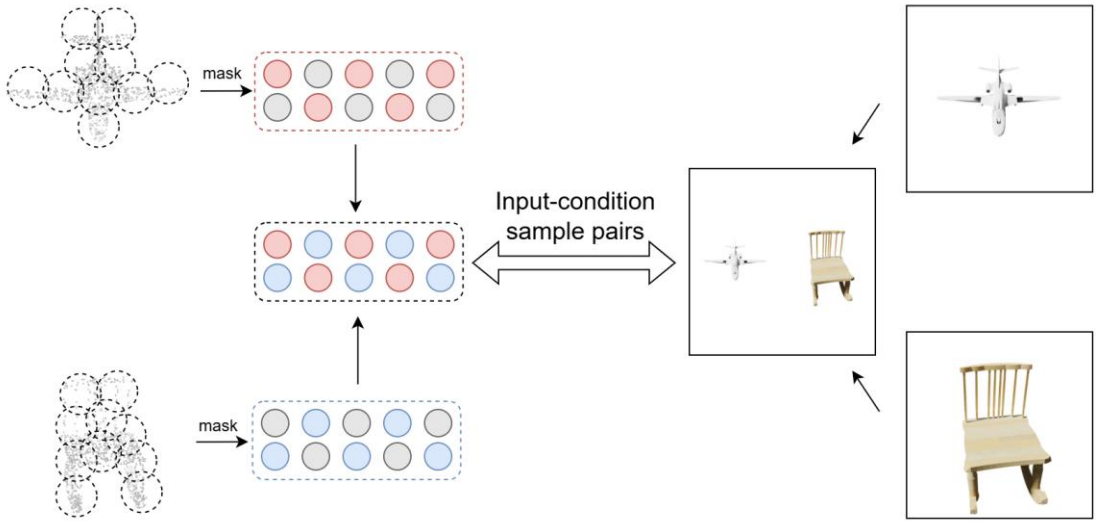


- Train a 3D backbone by aligning 3D object features with image features obtained from the pre-trained point-to-image framework in Stage I.

- **Advantage**: This alignment can leverage the rich semantics encapsulated in the SD model, enhancing the 3D backbone's ability to understand the object shapes and structures.

Method

- **Data Augmentation**



Point Cloud Mixing:

$$\tilde{P} = M \odot P + (1 - M) \odot P'$$

- M :binary mask for patch selection

Image Mixing:

$$\tilde{X} = X \cup X'$$

- Concatenate two images along width

Experiment

Object Classification

Method	Efficiency			ScanObjectNN			ModelNet40
	#Params	#FLOPs	Time	OBJ-BG	OBJ-ONLY	PB-T50-RS	
Supervised Learning Only							
PointNet [33]	3.5	-	-	73.3	79.2	68.0	89.2
PointNet++ [34]	1.5	-	-	82.3	84.3	77.9	90.7
Transformer [59]	22.1	-	-	79.9	80.6	77.2	91.4
DGCNN [47]	1.8	-	-	82.8	86.2	78.1	92.9
PointCNN [23]	0.6	-	-	86.1	85.5	78.5	92.2
SimpleView [14]	-	-	-	-	-	80.5	93.9
MVTN [18]	11.2	-	-	-	-	82.8	93.8
PointMLP [29]	12.6	-	-	-	-	85.4 ± 0.3	**94.1**
PointNeXt [36]	1.4	-	-	-	-	87.7 ± 0.4	93.7
P2P-HorNet [49]	195.8	-	-	-	-	89.3	94.0
Single-Modal Self-Supervised Learning							
Point-BERT [59]	22.1	4.8	2.0	87.43	88.12	83.07	92.7
MaskPoint [24]	22.1	4.8	2.0	89.30	88.10	84.30	-
Point-MAE [31]	22.1	4.8	2.0	90.02	88.29	85.18	93.2
Point-M2AE [62]	15.3	3.6	5.0	91.22	88.81	86.43	93.4
PointDif [67]	22.1	4.8	2.0	93.29	91.91	87.61	-
Point-FEMAE [61]	27.4	14.2	3.1	**95.18**	93.29	90.22	94.0
Cross-Modal Self-Supervised Learning							
ACT [13]	22.1	4.8	2.0	93.29	91.91	88.21	93.2
Joint-MAE [17]	-	-	-	90.94	88.86	86.07	-
I2P-MAE [63]	15.3	3.6	5.0	94.15	91.57	90.11	93.7
ReCon* [35]	43.6	5.3	7.0	**95.18**	**93.63**	**90.63**	**94.1**
TAP [50]	22.1	4.8	2.0	90.36	89.50	85.67	-
Ours	22.1	4.8	2.0	**95.18**	**93.63**	90.08	93.7

Few-shot learning

Method	5-way		10-way	
	10-shot	20-shot	10-shot	20-shot
Supervised Learning Only				
DGCNN [47]	31.6 ± 2.8	40.8 ± 4.6	19.9 ± 2.1	16.9 ± 1.5
Transformer [59]	87.8 ± 5.2	93.3 ± 4.3	84.6 ± 5.5	89.4 ± 6.3
Single-Modal Self-Supervised Learning				
DGCNN-OcCo [47]	90.6 ± 2.8	92.5 ± 1.9	82.9 ± 1.3	86.5 ± 2.2
Transformer-OcCo [59]	94.0 ± 3.6	95.9 ± 2.3	89.4 ± 5.1	92.4 ± 4.6
Point-BERT [59]	94.6 ± 3.1	96.3 ± 2.7	91.0 ± 5.4	92.7 ± 5.1
MaskPoint [24]	95.0 ± 3.7	97.2 ± 1.7	91.4 ± 4.0	93.4 ± 3.5
Point-MAE [31]	96.3 ± 2.5	97.8 ± 1.8	92.6 ± 4.1	95.0 ± 3.0
Point-M2AE [62]	96.8 ± 1.8	98.3 ± 1.4	92.3 ± 4.5	95.0 ± 3.0
Point-FEMAE [61]	97.2 ± 1.9	98.6 ± 1.3	**94.0 ± 3.3**	95.8 ± 2.8
Cross-Modal Self-Supervised Learning				
ACT [13]	96.8 ± 2.3	98.0 ± 1.4	93.3 ± 4.0	95.6 ± 2.8
Joint-MAE [17]	96.7 ± 2.2	97.9 ± 1.8	92.6 ± 3.7	95.1 ± 2.6
I2P-MAE [63]	97.0 ± 1.8	98.3 ± 1.3	92.6 ± 5.0	95.5 ± 3.0
ReCon* [35]	97.3 ± 1.9	98.9 ± 1.2	93.3 ± 3.9	95.8 ± 3.0
TAP [50]	97.3 ± 1.8	97.8 ± 1.7	93.1 ± 2.6	95.8 ± 1.0
Ours	**97.7 ± 1.8**	**99.0 ± 0.9**	93.8 ± 3.6	**95.9 ± 2.6**

Experiment

- **Part Segmentation**

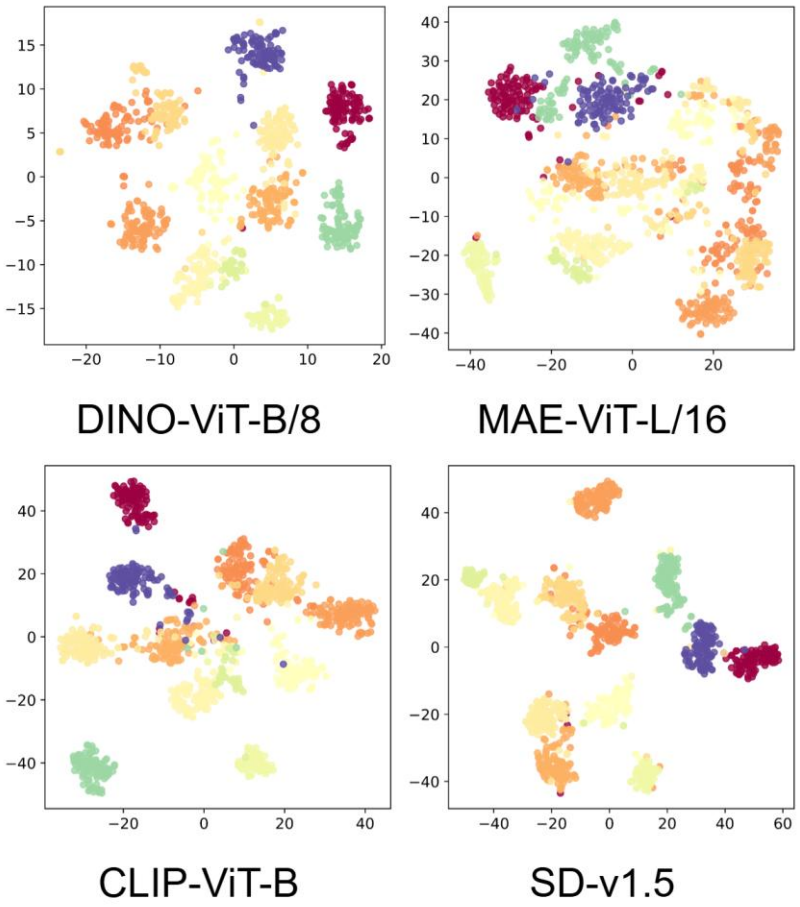
Method	Efficiency			ScanObjectNN			ModelNet40
	#Params	#FLOPs	Time	OBJ-BG	OBJ-ONLY	PB-T50-RS	
Supervised Learning Only							
PointNet [33]	3.5	-	-	73.3	79.2	68.0	89.2
PointNet++ [34]	1.5	-	-	82.3	84.3	77.9	90.7
Transformer [59]	22.1	-	-	79.9	80.6	77.2	91.4
DGCNN [47]	1.8	-	-	82.8	86.2	78.1	92.9
PointCNN [23]	0.6	-	-	86.1	85.5	78.5	92.2
SimpleView [14]	-	-	-	-	-	80.5	93.9
MVTN [18]	11.2	-	-	-	-	82.8	93.8
PointMLP [29]	12.6	-	-	-	-	85.4 ± 0.3	**94.1**
PointNeXt [36]	1.4	-	-	-	-	87.7 ± 0.4	93.7
P2P-HorNet [49]	195.8	-	-	-	-	89.3	94.0
Single-Modal Self-Supervised Learning							
Point-BERT [59]	22.1	4.8	2.0	87.43	88.12	83.07	92.7
MaskPoint [24]	22.1	4.8	2.0	89.30	88.10	84.30	-
Point-MAE [31]	22.1	4.8	2.0	90.02	88.29	85.18	93.2
Point-M2AE [62]	15.3	3.6	5.0	91.22	88.81	86.43	93.4
PointDif [67]	22.1	4.8	2.0	93.29	91.91	87.61	-
Point-FEMAE [61]	27.4	14.2	3.1	**95.18**	93.29	90.22	94.0
Cross-Modal Self-Supervised Learning							
ACT [13]	22.1	4.8	2.0	93.29	91.91	88.21	93.2
Joint-MAE [17]	-	-	-	90.94	88.86	86.07	-
I2P-MAE [63]	15.3	3.6	5.0	94.15	91.57	90.11	93.7
ReCon* [35]	43.6	5.3	7.0	**95.18**	**93.63**	**90.63**	**94.1**
TAP [50]	22.1	4.8	2.0	90.36	89.50	85.67	-
Ours	22.1	4.8	2.0	**95.18**	**93.63**	90.08	93.7

- **Object Detection**

Method	Pre-training Dataset	$AP_{0.5}$
3DETR [30]	-	37.9
Point-BERT [59]	ScanNet-Medium	38.3
MaskPoint [24]	ScanNet-Medium	40.6
TAP [50]	ShapeNet	41.4
Ours	ShapeNet	**42.4**

Experiment

- t-SNE visualization



- Ablation study on pretrained models

Pre-trained Models	ScanObjectNN
DINO-ViT-B/8 [4]	89.38
MAE-ViT-L/16 [19]	88.72
CLIP-ViT-B [37]	89.35
SD-v1.5 [39]	90.08

Conclusion

- Motivated by previous works, this paper studies how to harness the SD model to enhance point cloud self-supervised learning.
- We develop PointSD, a framework that leverages existing SD models to assist point cloud pre-training in two stages, where a point-to-image framework is built to extract features from SD, and then feature alignment is performed to encourage the 3D backbone to learn robust representations.