Meta

CVPR Nashville JUNE 11-15, 2025

Sonata : Self-Supervised Learning of Reliable Point Representations (2025)

Paper Review

2025.10.01

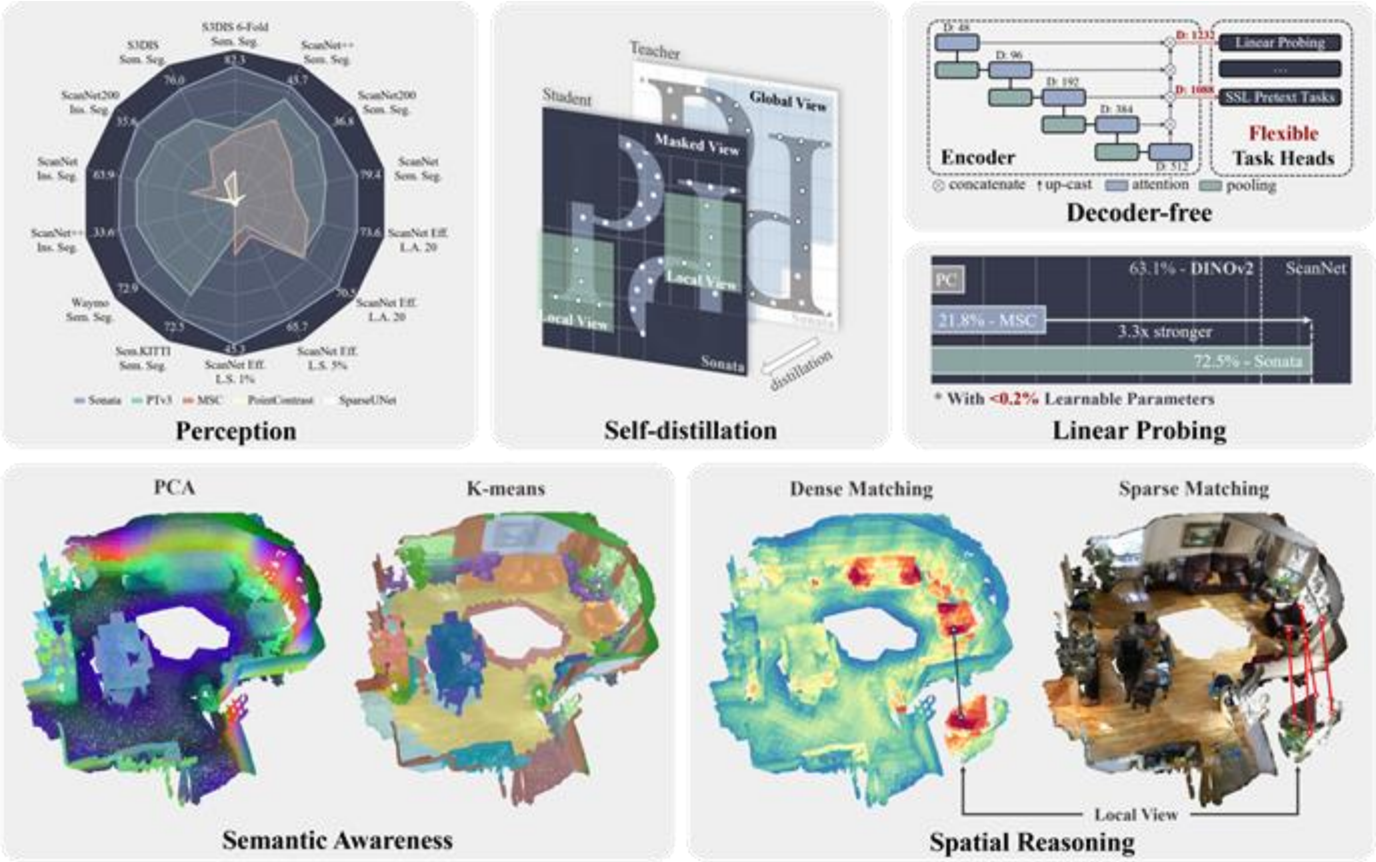
Changseon Yu

CAU 중앙대학교

GSPS Graduate Student Paper Seminar

Contents

1. Introduction
2. Purpose
3. Methodology
4. Result
5. Conclusion
6. Research

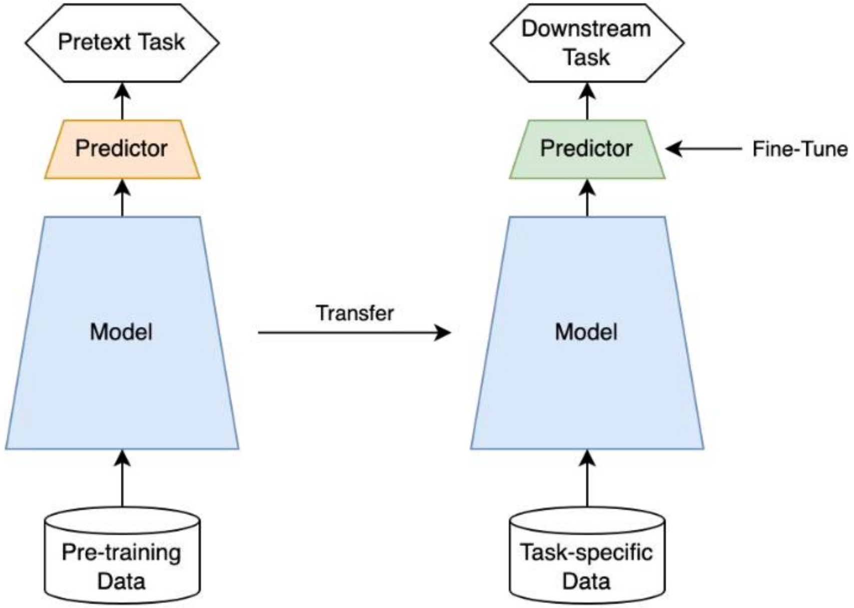


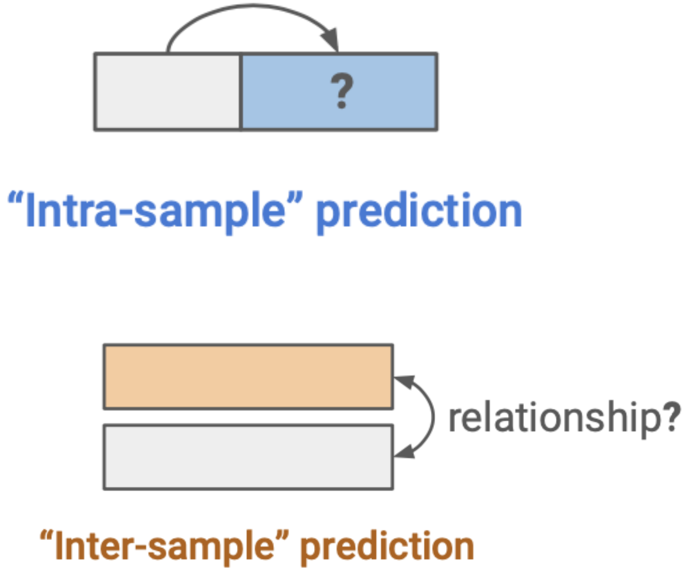
Introduction

Background(1)

❖ Self Supervised Learning (Image)

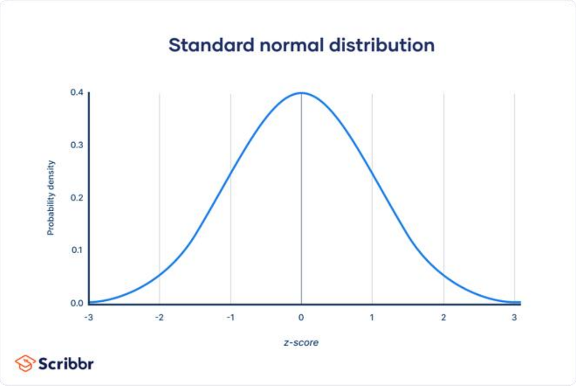
- Learning method to obtain good representation from unlabeled data
- Learn the model in the supervision method by setting the task as a target within the input (x) without label (y)
- In some cases, **linear probing surpasses full fine tuning**





Introduction

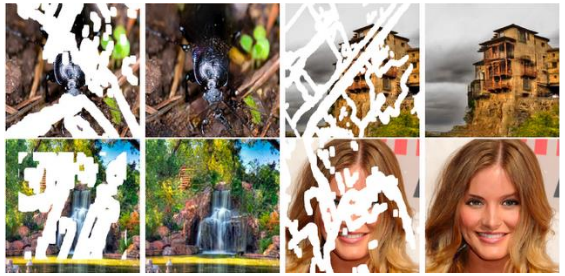
Background(1)



❖ Self Supervised Learning (Image)

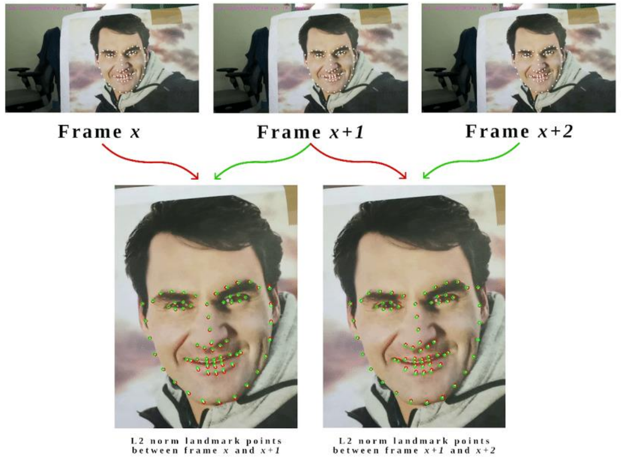
- Learning method to obtain good representation from unlabeled data
- Learn the model in the supervision method by setting the task as a target within the input (x) without label (y)
- In some cases, **linear probing surpasses full fine tuning**

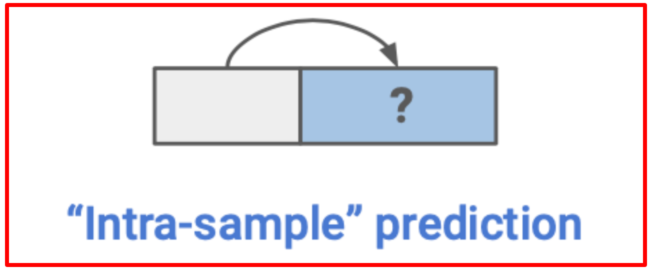
Inpaint



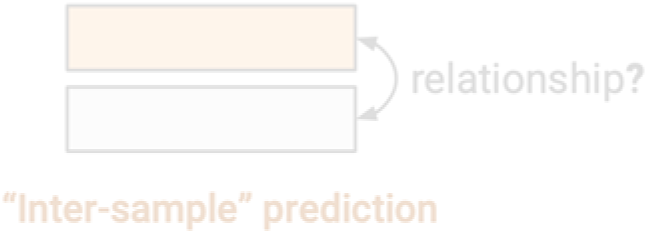
Jittering

$$\hat{P} = P + N(0, \sigma 2)$$


Frame x
Frame $x+1$
Frame $x+2$
L2 norm landmark points between frame x and $x+1$
L2 norm landmark points between frame $x+1$ and $x+2$



"Intra-sample" prediction



"Inter-sample" prediction

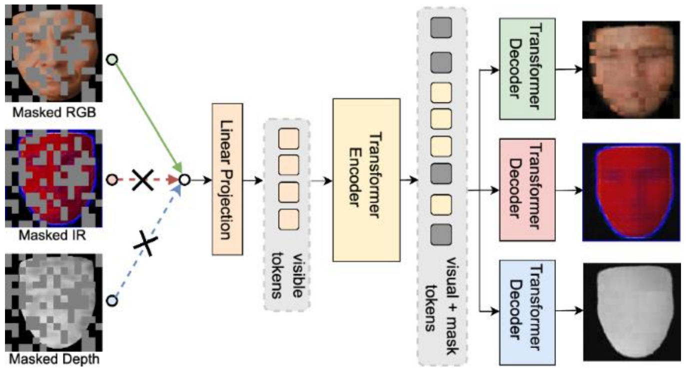
Introduction

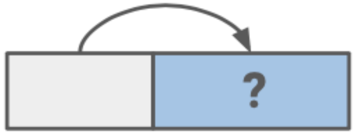
Background(1)

❖ Self Supervised Learning (Image)

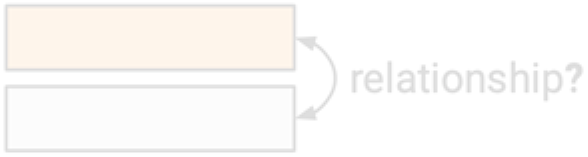
- Learning method to obtain good representation from unlabeled data
- Learn the model in the supervision method by setting the task as a target within the input (x) without label (y)
- In some cases, **linear probing surpasses full fine tuning**

Masked autoencoder





“Intra-sample” prediction



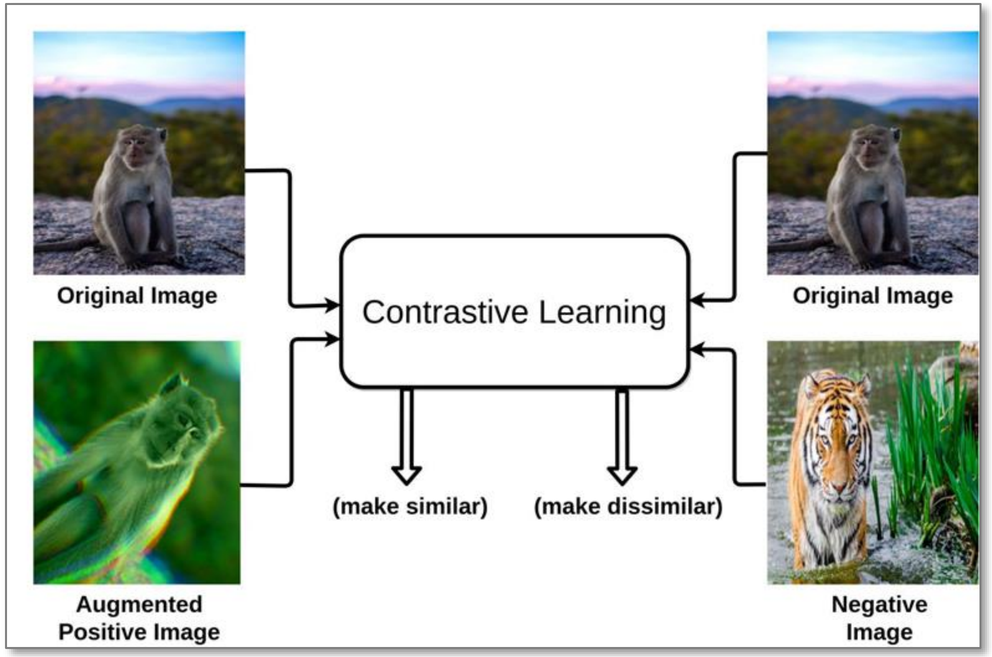
“Inter-sample” prediction

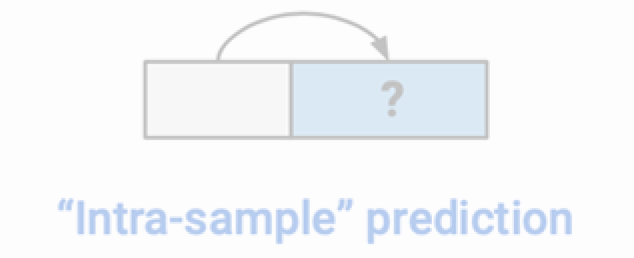
Introduction

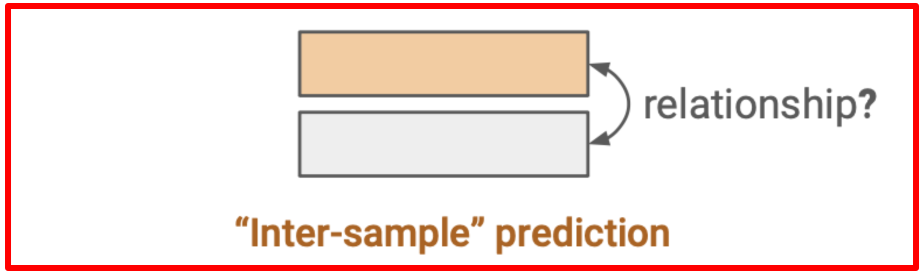
Background(1)

❖ Self Supervised Learning (Image)

- Learning method to obtain good representation from unlabeled data
- Learn the model in the supervision method by setting the task as a target within the input (x) without label (y)
- In some cases, **linear probing surpasses full fine tuning**







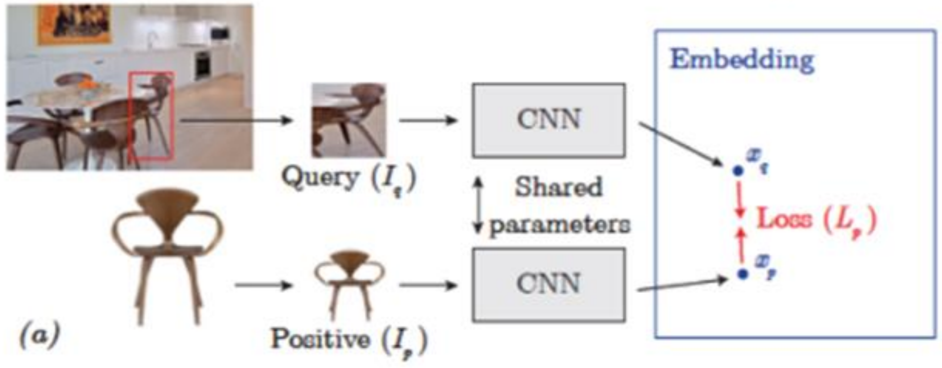
Introduction

Background(1)

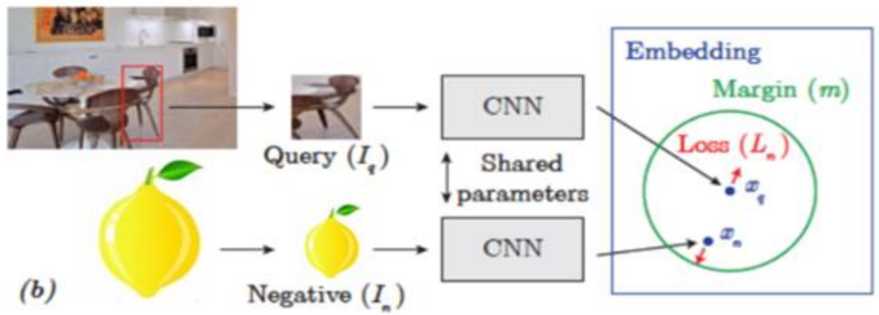
❖ Objective of SSL

- Reducing the difference between the original and the pseudo lables
- Higher performance with quality pseudo labels

Typical positive pair (x_p, x_q) loss: $L(x_p, x_q) = ||x_p - x_q||^2$ (Euclidian Loss)



Typical negative pair (x_n, x_q) loss :
$L(x_n, x_q) = \max(0, m^2 - ||x_n - x_q||^2)$ (Hinge Loss)

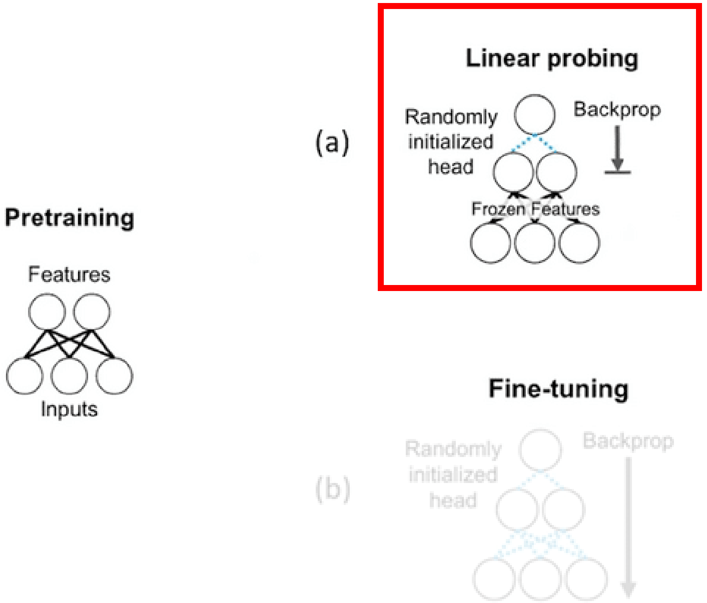


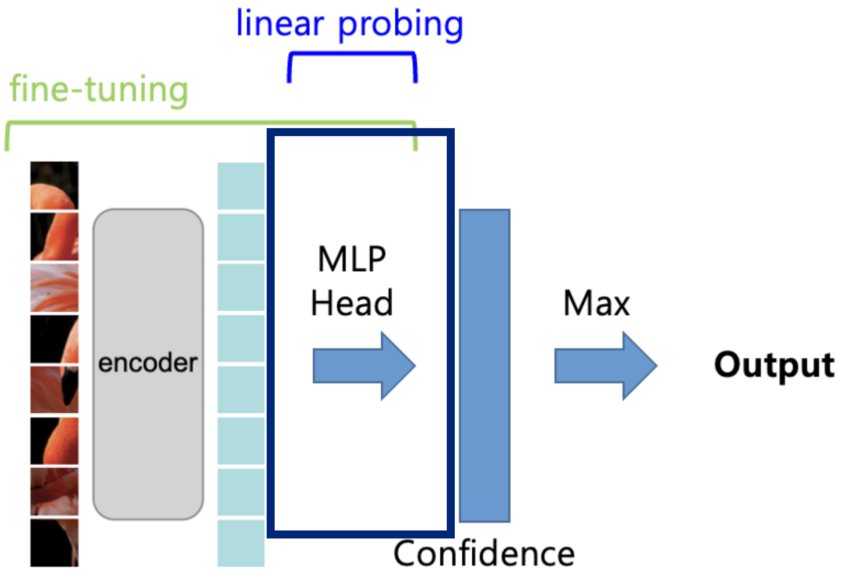
Introduction

Background(2)

❖ Linear probing

- Fixed the weight of the pre-trained feature extractor (encoder)
- Randomly initialized head, i.e., learning by **adding only one linear classifier**
- Evaluate how well pre-trained features contain the semantic information needed to solve downstream tasks



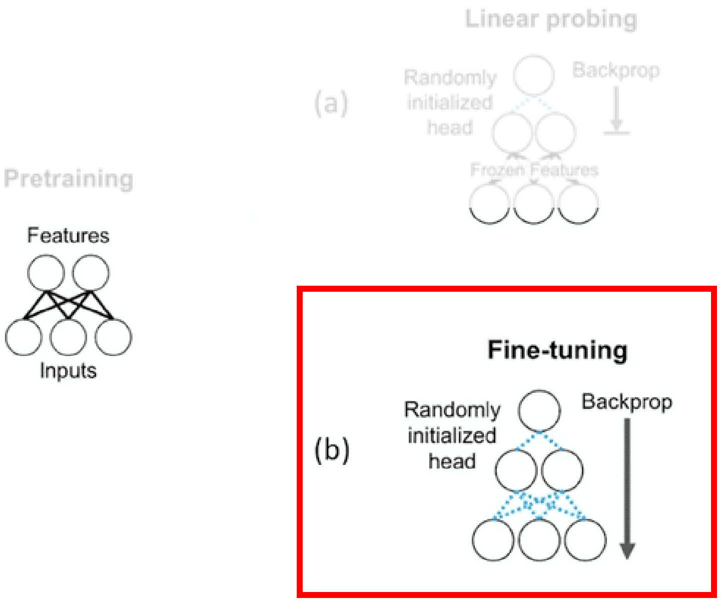


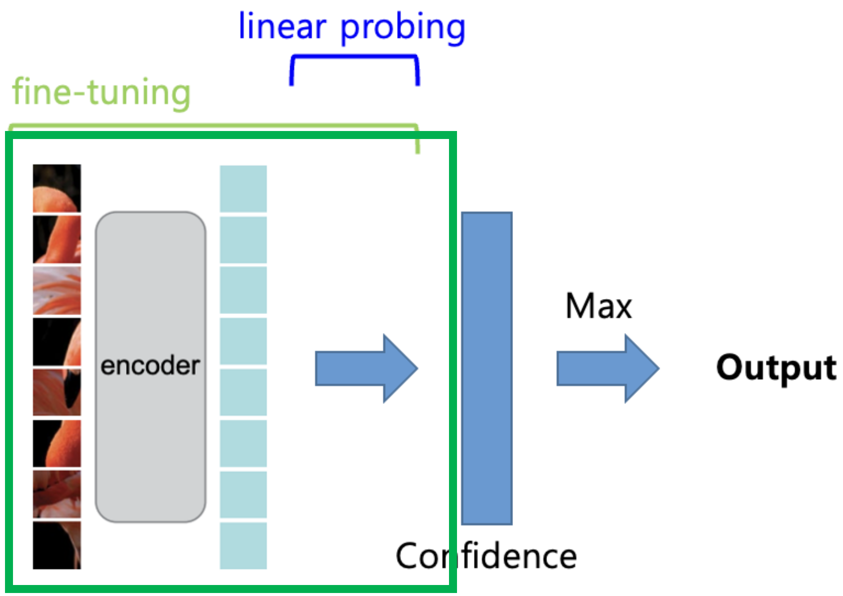
Introduction

Background(3)

❖ Fine tuning (Full fine tuning)

- Add randomly initialized heads over pre-trained encoders
- **Re-learn every layer** of the entire network to the target task (downstream task)
- Standard way to achieve the best performance for a particular task





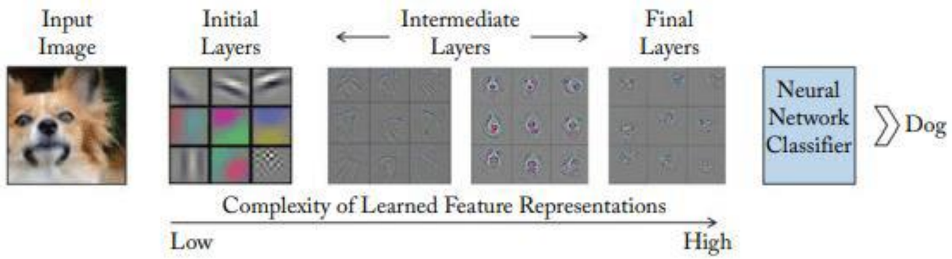
Introduction

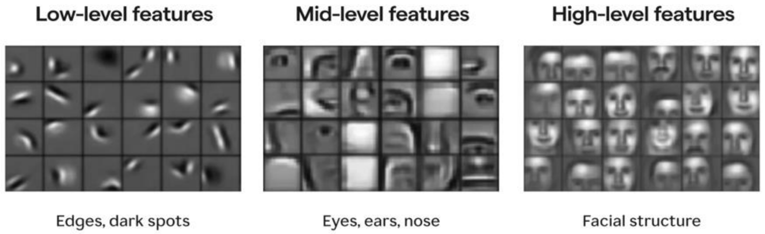
Background(3)

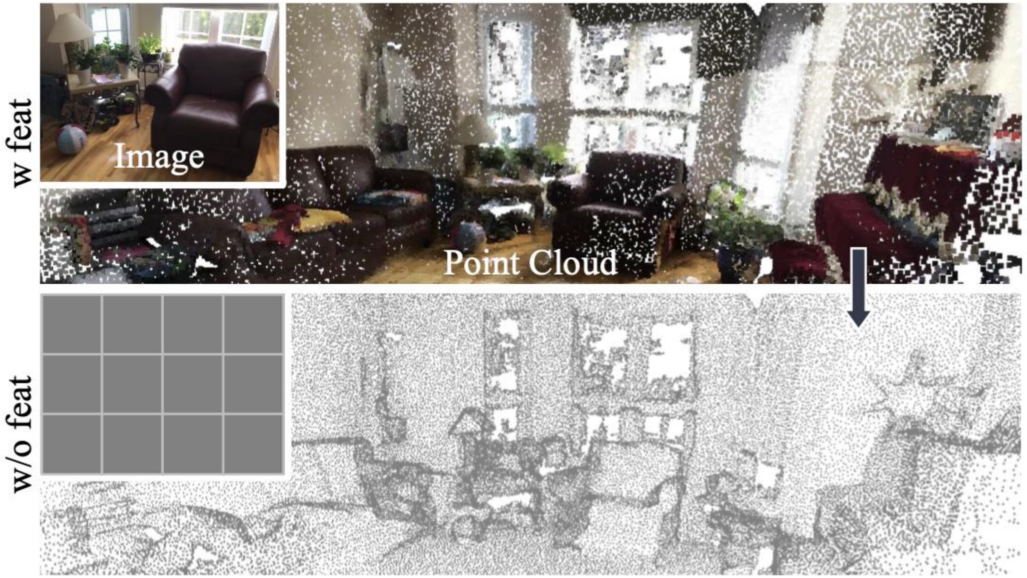
Low level features : color, edge, point, corner
High level features : object, scene, action, interaction

❖ Geometric shortcut

- Point cloud SSL is still limited (compared with 2D SSL)
- 2D Image has RGB, 3D PC has surface normal, height (cause **overfitting**)
- Results biased towards **low-level feature** extraction over Semantic information


Input Image
Initial Layers
Intermediate Layers
Final Layers
Neural Network Classifier
Dog
Complexity of Learned Feature Representations
Low
High



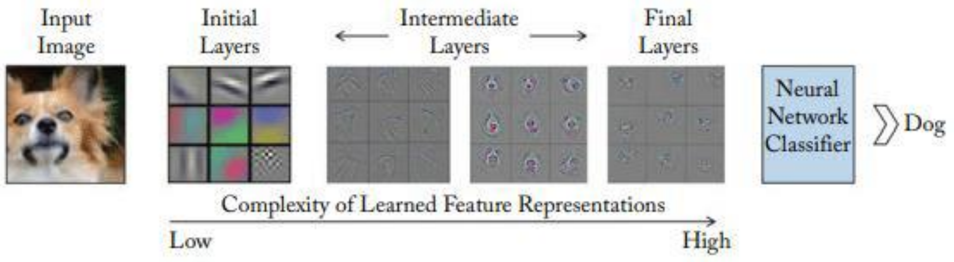


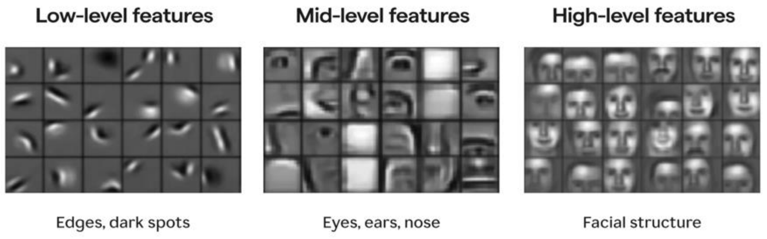
Introduction

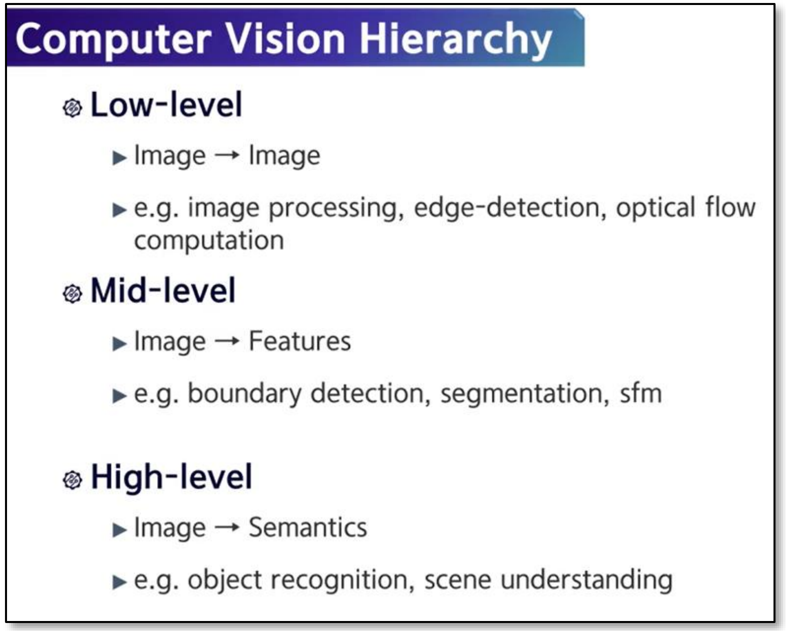
Background(3)

❖ Geometric shortcut

- Point cloud SSL is still limited (compared with 2D SSL)
- 2D Image has RGB, 3D PC has surface normal, height (cause **overfitting**)
- Results biased towards **low-level feature** extraction over Semantic information







Computer Vision Hierarchy

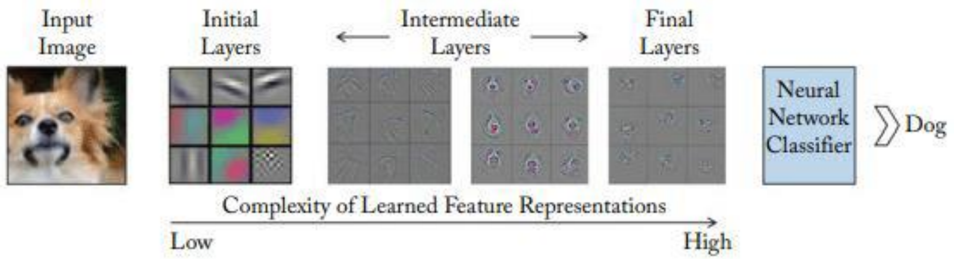
- **Low-level**
  - Image → Image
  - e.g. image processing, edge-detection, optical flow computation
- **Mid-level**
  - Image → Features
  - e.g. boundary detection, segmentation, sfm
- **High-level**
  - Image → Semantics
  - e.g. object recognition, scene understanding

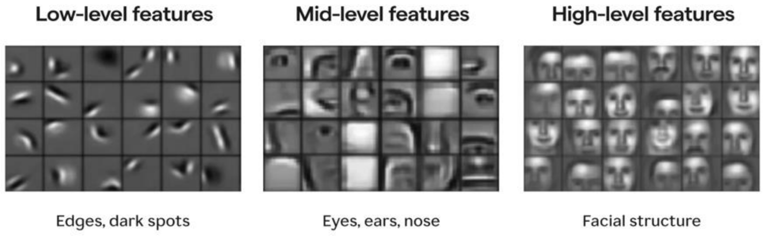
Introduction

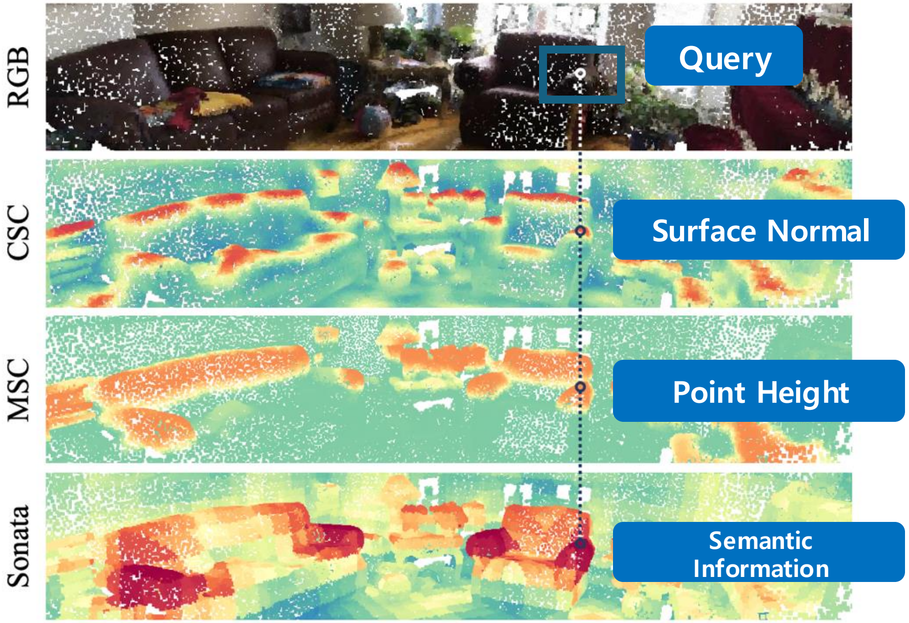
Background(3)

❖ Geometric shortcut

- Point cloud SSL is still limited (compared with 2D SSL)
- 2D Image has RGB, 3D PC has surface normal, height (cause **overfitting**)
- Results biased towards **low-level feature** extraction over Semantic information


Input Image
Initial Layers
Intermediate Layers
Final Layers
Neural Network Classifier
Dog
Complexity of Learned Feature Representations
Low
High



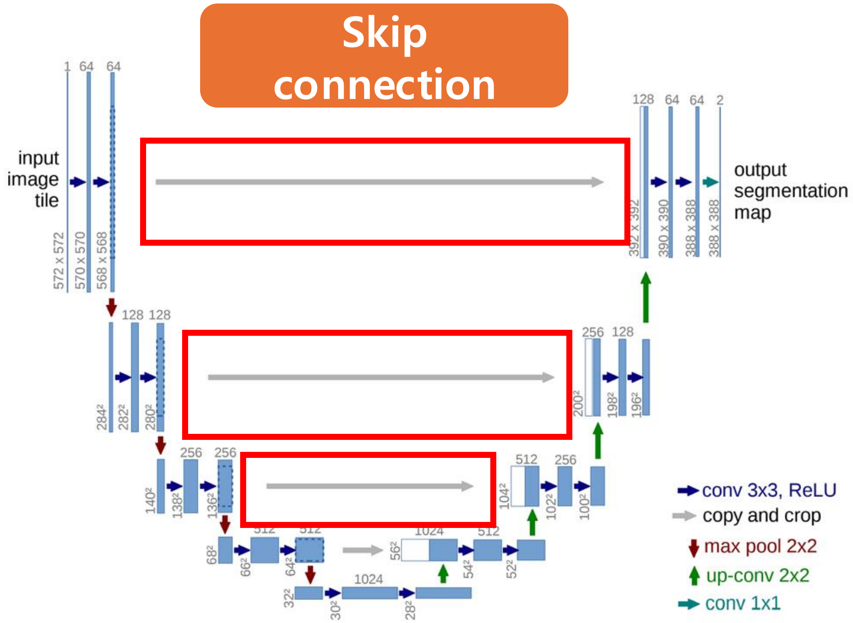


● similarity

Purpose

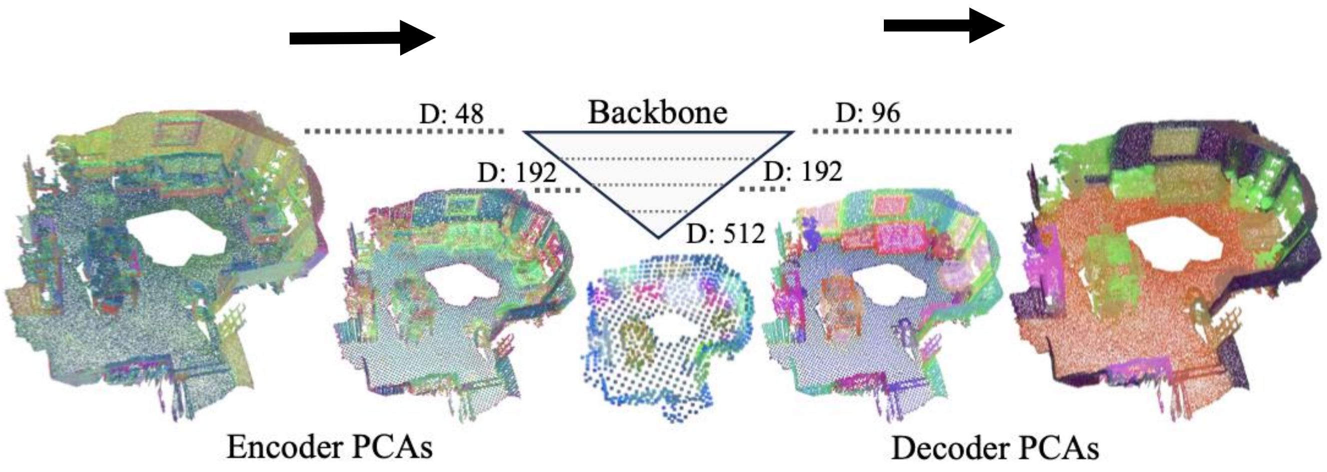
❖ Encoder Only

- The tight structure of the **encoder decoder reduces flexibility**, generalization ability
- Sonata focuses only on encoder for SSL
- Allow the downstream tasks to add only the needed Flexible Task Heads onto the encoder for use


Skip connection

Increase Channel Feature

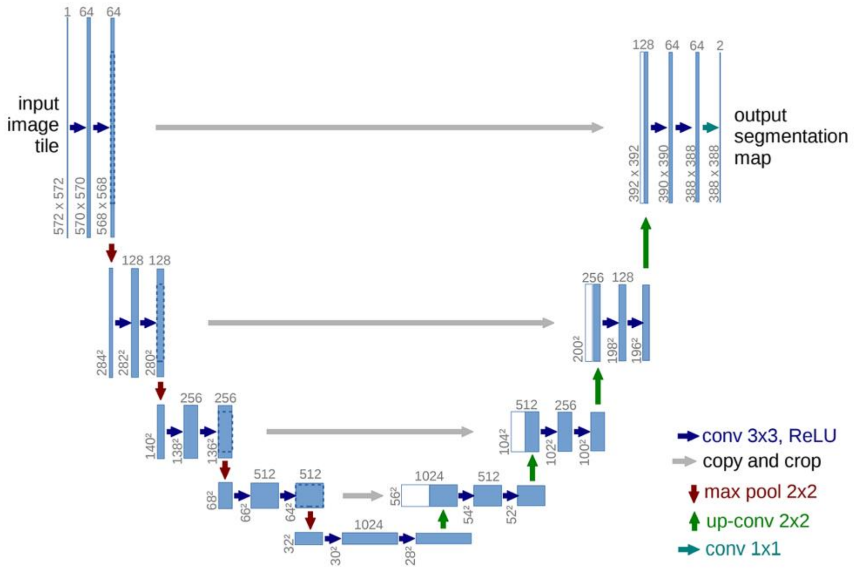
Decrease Channel Feature



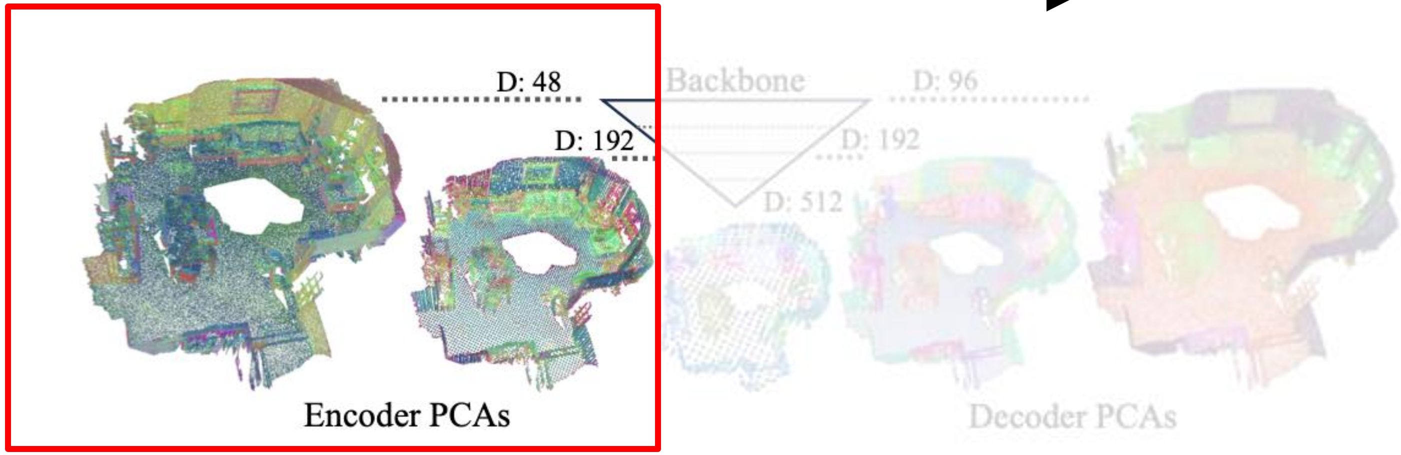
Purpose

❖ Encoder Only

- The tight structure of the encoder decoder reduces flexibility, generalization ability
- Sonata focuses only on encoder for SSL
- Allow the downstream tasks to add only the needed Flexible Task Heads onto the encoder for use



Shallow Channel Feature

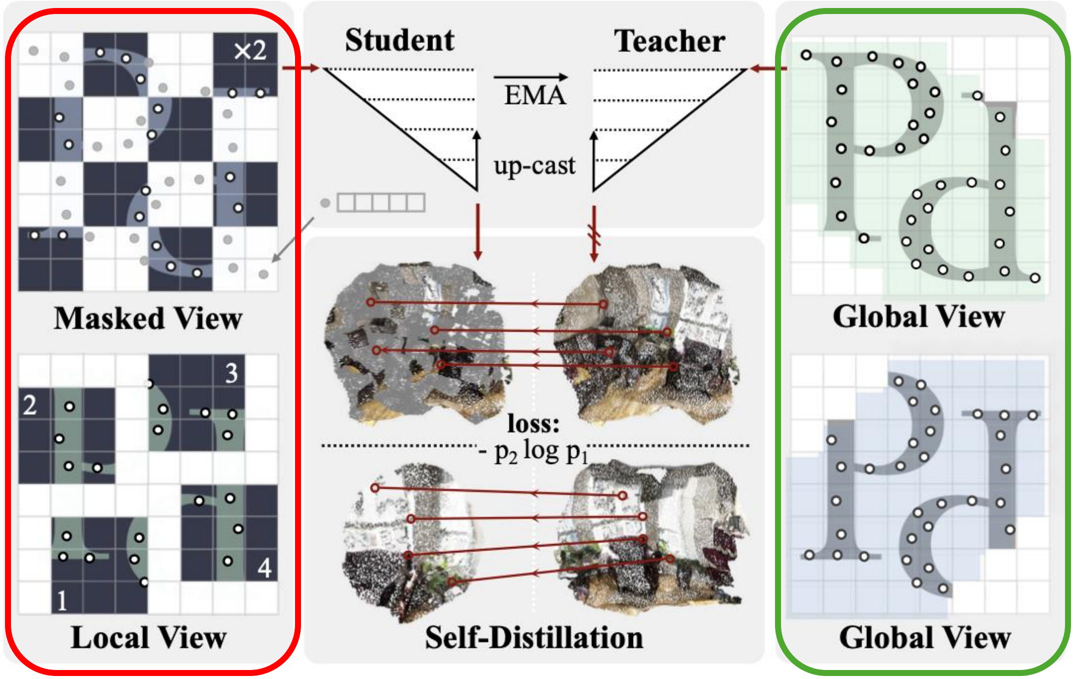


Purpose

❖ Point Self-distillation

- Learns rich and reliable representations to provide robust linear probing results
- Matched to the corresponding points in Global View based on the original spatial distance
- perform extremely challenging SSL tasks with **reliable guidance** from the teacher

- **Local View :** Extremely hard view
- **Masked View :** Apply high-rate patch mask



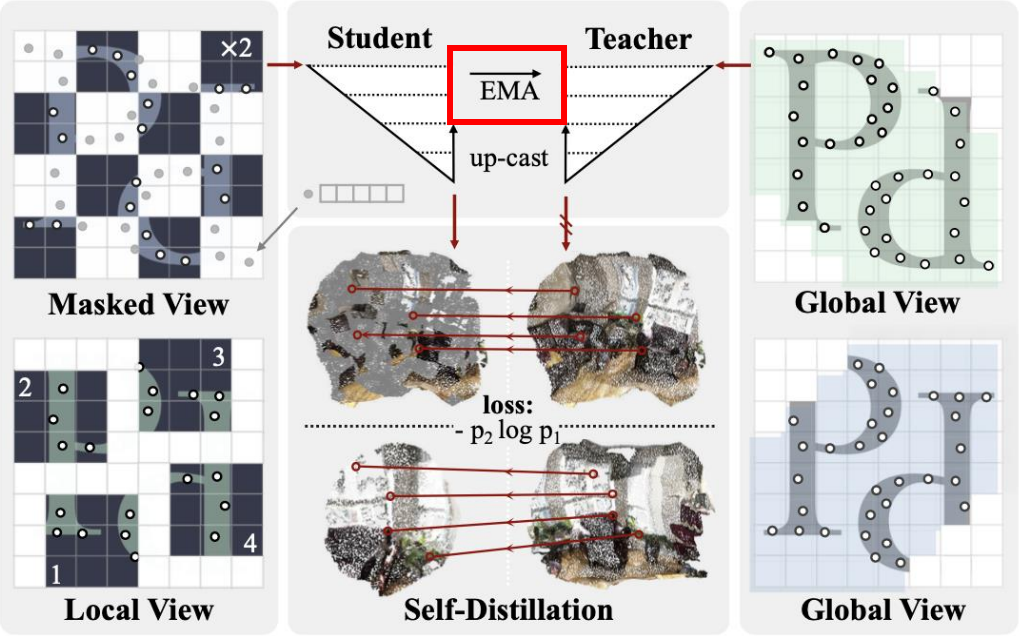
- **Stable view (goal)**
- **Less Augmentation**
- **Encoded by Teacher model**

Method

❖ EMA (Exponential Moving Average)

- The higher the difficulty of prior work, the greater the risk of model collapse
- Sonata uses an asymmetric encoding approach, EMA (exponential moving average)

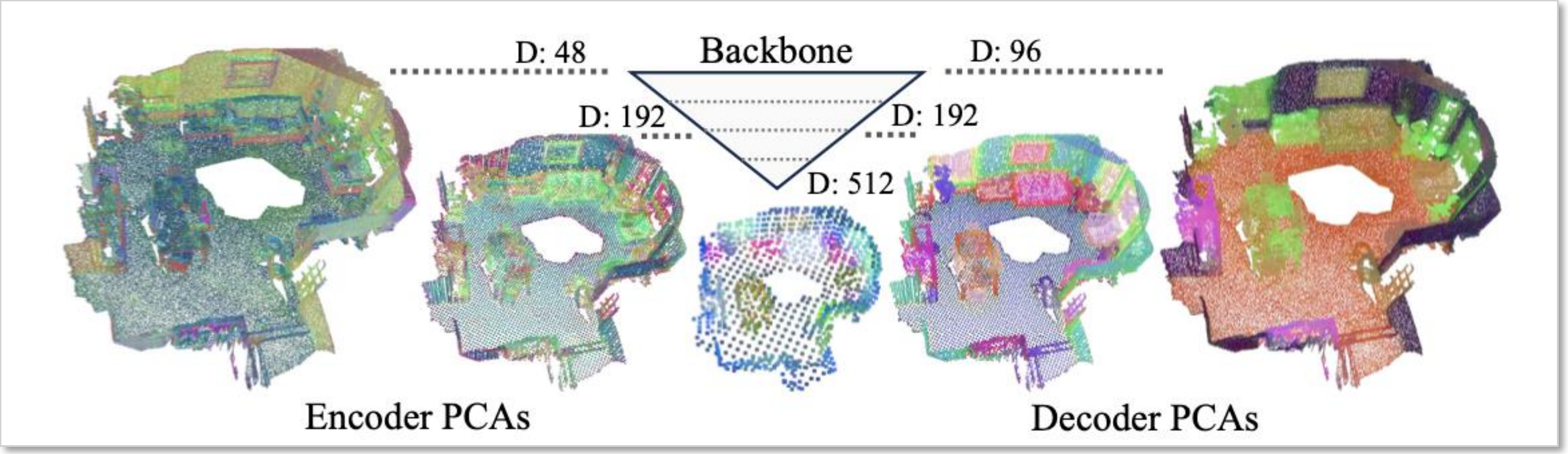
$$\theta t \leftarrow m\theta t + (1 - m)\theta s$$



Method

❖ **Decoder removal**

- Sonata's Micro Design to Solve Geometric Shortcuts
- Since spatial information is deeply tied to point coordinates, masking is difficult
- Obscures spatial information and highlights input features

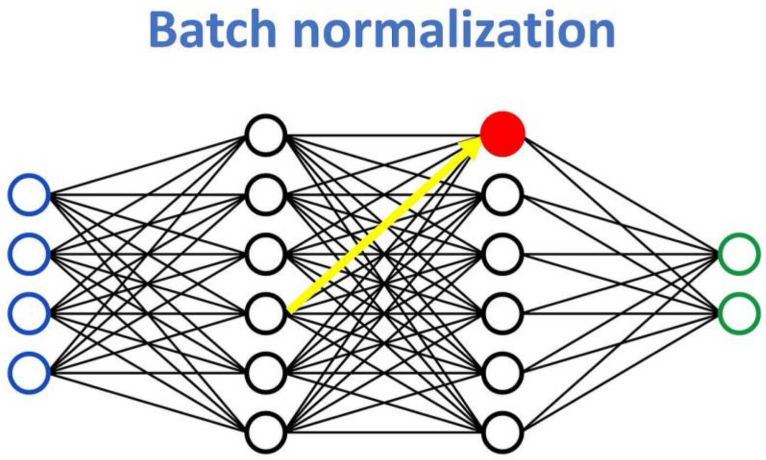


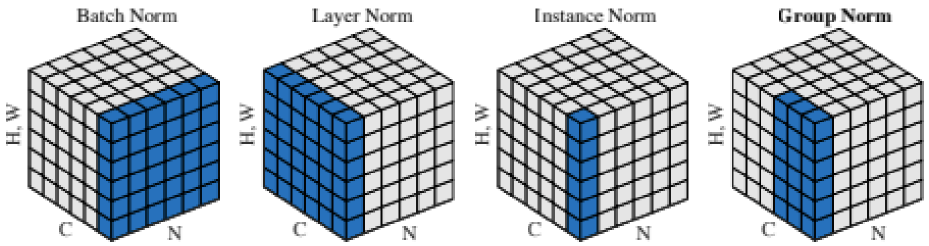
Result

❖ **Environment setting**

- BN: Batch Normalization → LN : Layer Normalization
- BN : Normalize features using the mean and variance of all samples in a batch
- LN : Normalize by calculating the mean and variance of features within a single individual sample (layer)

- Advantage for large-scale distributed learning or small batch sizes
- Less sensitive to statistical differences between datasets when learning different datasets





Result

- **Zero-shot comparison with DINOv2.**
- DINOv2 excels at capturing optical details, and sonatas better distinguish spatial information

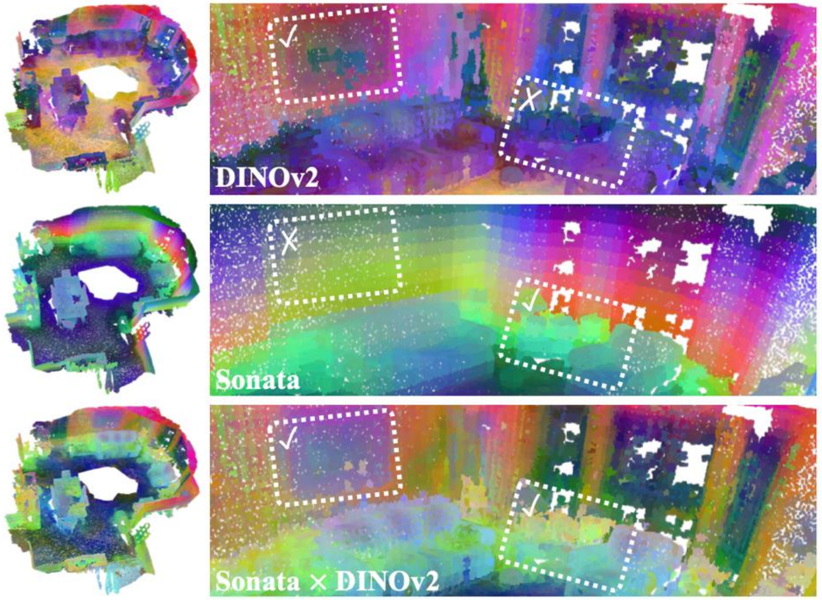
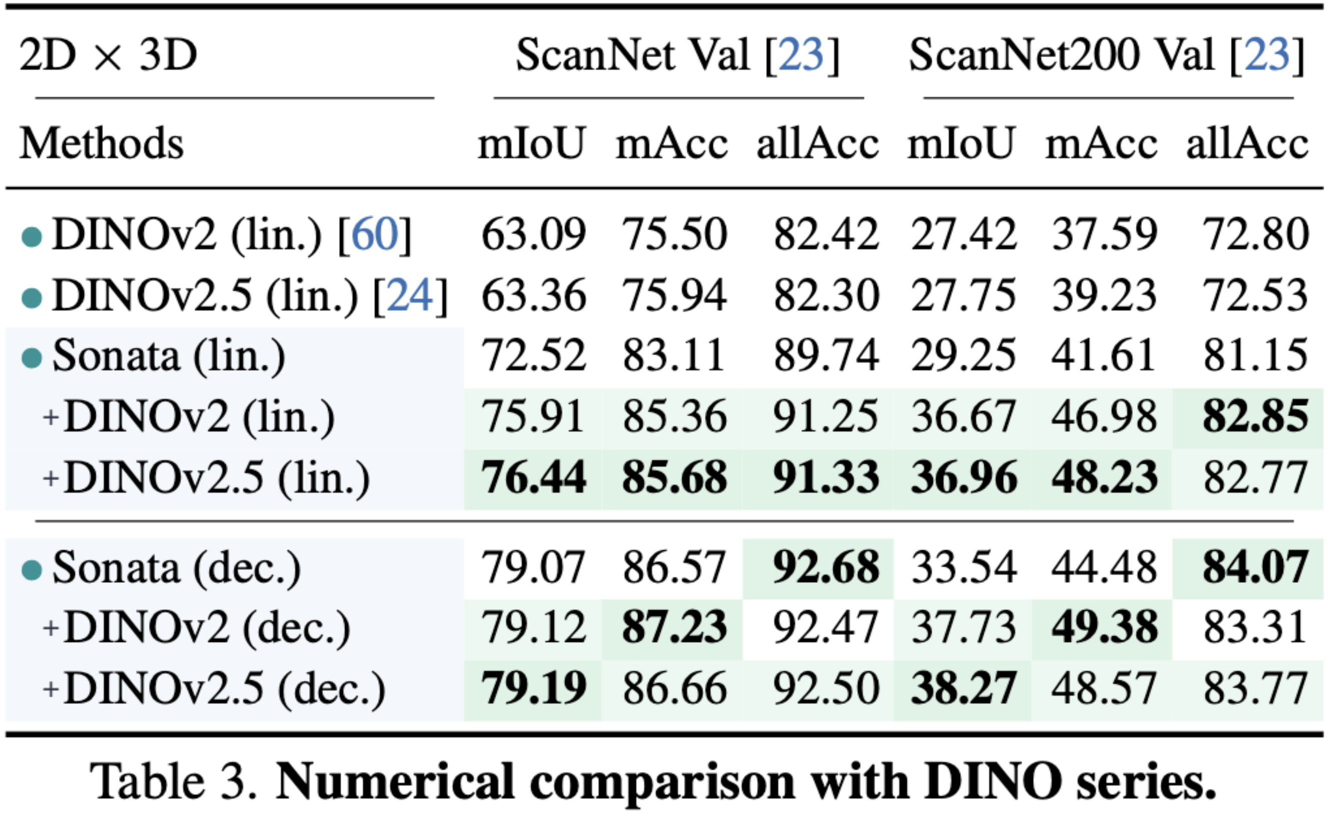


Figure 7. **Zero-shot comparison with DINOv2.** We compare the PCA visualizations of DINOv2, Sonata, and their combined feature representation. DINOv2 excels at capturing photometric details, while Sonata better distinguishes spatial information. The combined model demonstrates improved coherence and detail, showcasing the complementary strengths of both models.

Result

- **Numerical comparison with DINO series**
- When linear probing was performed by combining the features of Sonata with those of DINOv2.5, ScanNet mIoU achieved 76.44%
- Learn unique 3D geor

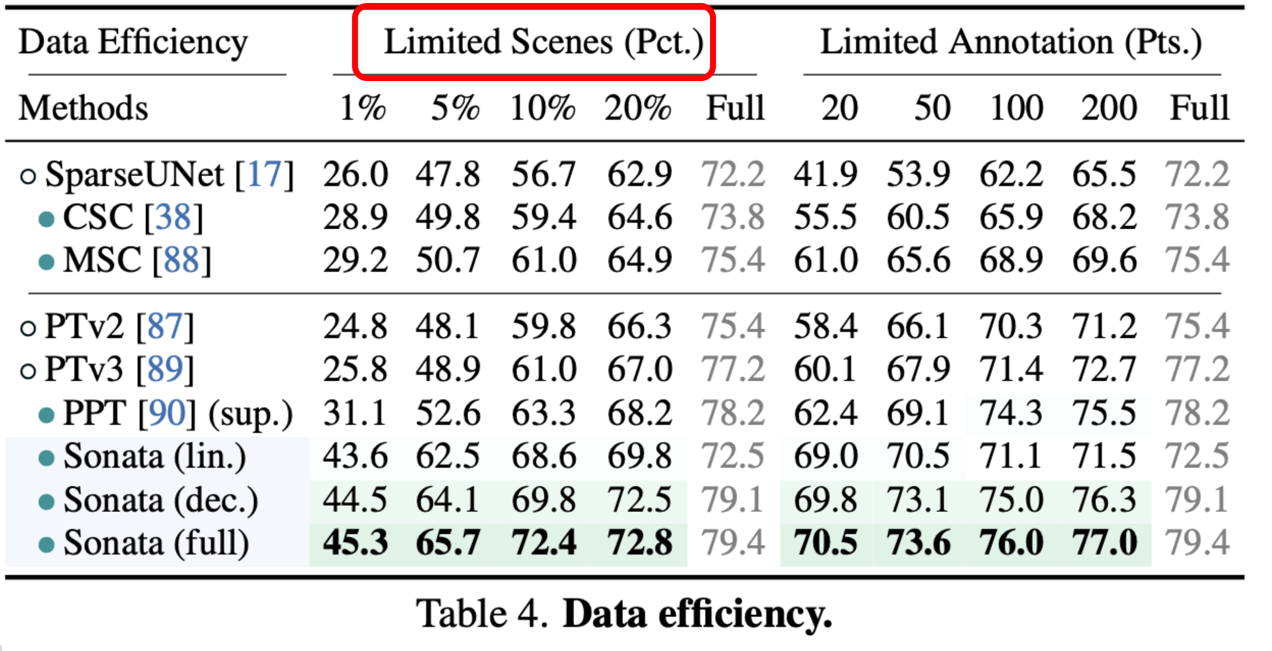


2D × 3D	ScanNet Val [23]			ScanNet200 Val [23]		
Methods	mIoU	mAcc	allAcc	mIoU	mAcc	allAcc
● DINOv2 (lin.) [60]	63.09	75.50	82.42	27.42	37.59	72.80
● DINOv2.5 (lin.) [24]	63.36	75.94	82.30	27.75	39.23	72.53
● Sonata (lin.)	72.52	83.11	89.74	29.25	41.61	81.15
+DINOv2 (lin.)	75.91	85.36	91.25	36.67	46.98	**82.85**
+DINOv2.5 (lin.)	**76.44**	**85.68**	**91.33**	**36.96**	**48.23**	82.77
● Sonata (dec.)	79.07	86.57	**92.68**	33.54	44.48	**84.07**
+DINOv2 (dec.)	79.12	**87.23**	92.47	37.73	**49.38**	83.31
+DINOv2.5 (dec.)	**79.19**	86.66	92.50	**38.27**	48.57	83.77

Table 3. **Numerical comparison with DINO series.**

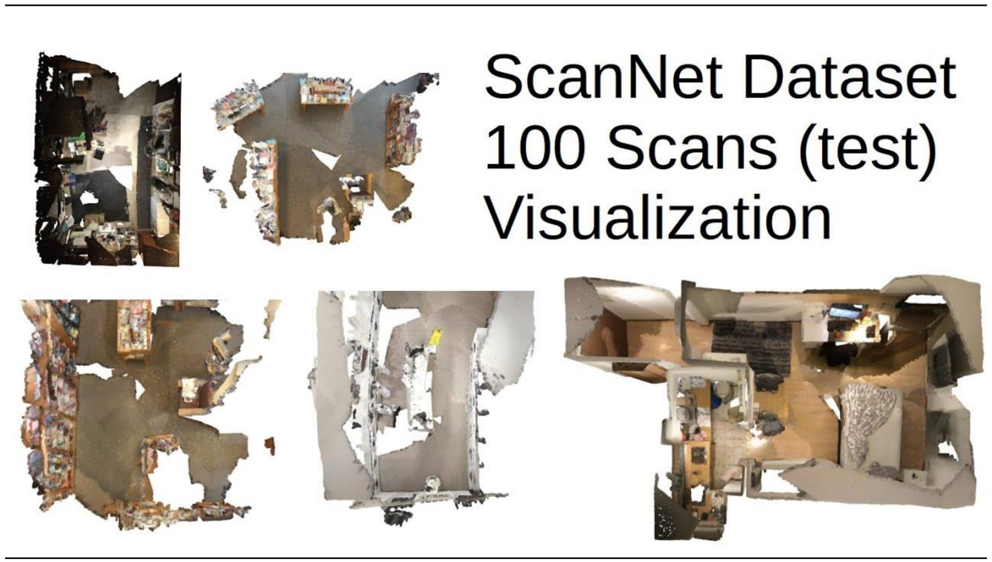
Result

- **Data efficiency**
- Compare Sonata's performance under limited data conditions to demonstrate its data efficiency
- **Limited Scenes**: Performance when trained using only a fraction of ScanNet's total data (1%, 5%, 10%, 20%)
- Limited annotation: performance when limiting the number of point annotations (Pts.) per scene
- Learned representations are highly efficient and have superior generalization capabilities



Data Efficiency	Limited Scenes (Pct.)					Limited Annotation (Pts.)				
Methods	1%	5%	10%	20%	Full	20	50	100	200	Full
○ SparseUNet [17]	26.0	47.8	56.7	62.9	72.2	41.9	53.9	62.2	65.5	72.2
● CSC [38]	28.9	49.8	59.4	64.6	73.8	55.5	60.5	65.9	68.2	73.8
● MSC [88]	29.2	50.7	61.0	64.9	75.4	61.0	65.6	68.9	69.6	75.4
○ PTv2 [87]	24.8	48.1	59.8	66.3	75.4	58.4	66.1	70.3	71.2	75.4
○ PTv3 [89]	25.8	48.9	61.0	67.0	77.2	60.1	67.9	71.4	72.7	77.2
● PPT [90] (sup.)	31.1	52.6	63.3	68.2	78.2	62.4	69.1	74.3	75.5	78.2
● Sonata (lin.)	43.6	62.5	68.6	69.8	72.5	69.0	70.5	71.1	71.5	72.5
● Sonata (dec.)	44.5	64.1	69.8	72.5	79.1	69.8	73.1	75.0	76.3	79.1
● Sonata (full)	**45.3**	**65.7**	**72.4**	**72.8**	79.4	**70.5**	**73.6**	**76.0**	**77.0**	79.4

Table 4. **Data efficiency.**



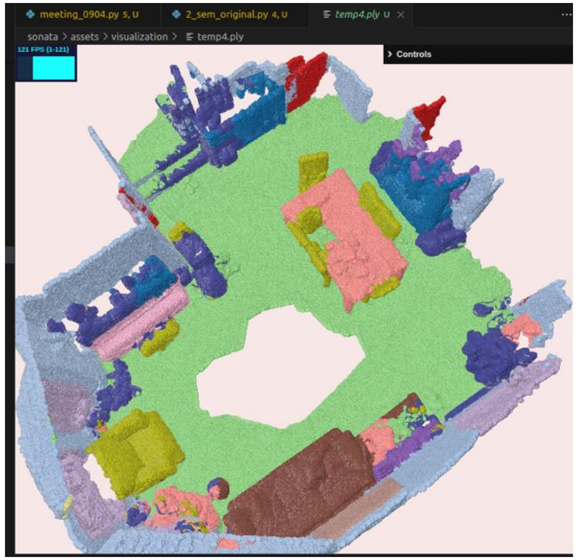
ScanNet Dataset
100 Scans (test)
Visualization

Result

- **Data efficiency**
- Compare Sonata's performance under limited data conditions to demonstrate its data efficiency
- Limited Scenes: Performance when trained using only a fraction of ScanNet's total data (1%, 5%, 10%, 20%)
- **Limited annotation**: performance when limiting the number of point annotations (Pts.) per scene
- Learned representations are highly efficient and have superior generalization capabilities

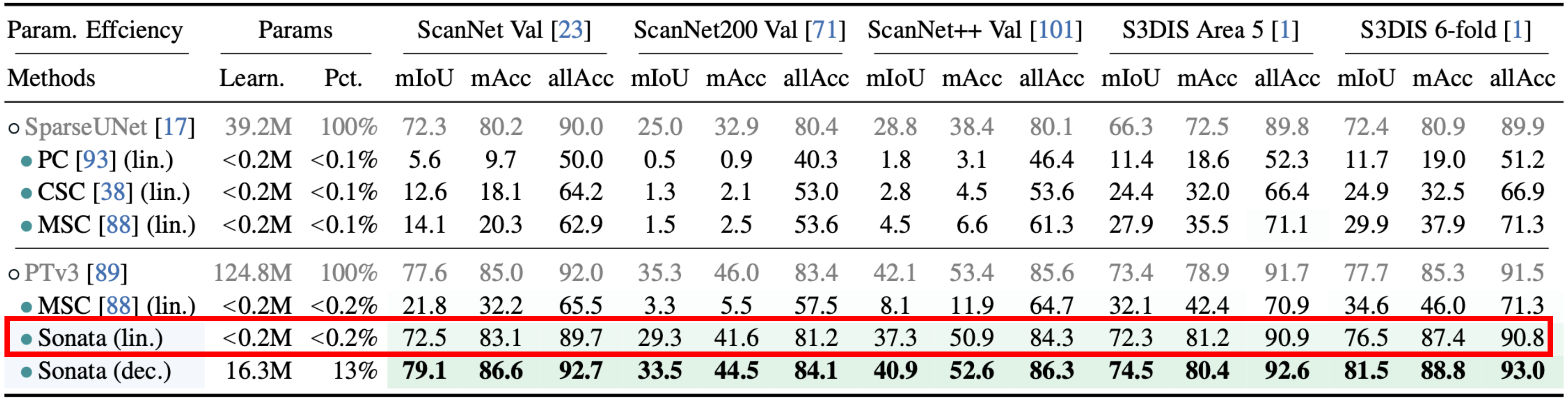
🟩 floor 🟦 wall 🟦 cabinet 🟧 bed 🟨 chair 🟫 sofa 🟥 table 🟥 door 🟪 window 🟪 bookshelf 🟫 picture
🟦 counter 🟪 desk 🟨 curtain 🟧 refrigerator 🟪 bathtub 🟦 shower curtain 🟩 toilet ⬛ sink 🟪 otherfurniture

Data Efficiency	Limited Scenes (Pct.)					Limited Annotation (Pts.)				
Methods	1%	5%	10%	20%	Full	20	50	100	200	Full
○ SparseUNet [17]	26.0	47.8	56.7	62.9	72.2	41.9	53.9	62.2	65.5	72.2
● CSC [38]	28.9	49.8	59.4	64.6	73.8	55.5	60.5	65.9	68.2	73.8
● MSC [88]	29.2	50.7	61.0	64.9	75.4	61.0	65.6	68.9	69.6	75.4
○ PTv2 [87]	24.8	48.1	59.8	66.3	75.4	58.4	66.1	70.3	71.2	75.4
○ PTv3 [89]	25.8	48.9	61.0	67.0	77.2	60.1	67.9	71.4	72.7	77.2
● PPT [90] (sup.)	31.1	52.6	63.3	68.2	78.2	62.4	69.1	74.3	75.5	78.2
● Sonata (lin.)	43.6	62.5	68.6	69.8	72.5	69.0	70.5	71.1	71.5	72.5
● Sonata (dec.)	44.5	64.1	69.8	72.5	79.1	69.8	73.1	75.0	76.3	79.1
● Sonata (full)	**45.3**	**65.7**	**72.4**	**72.8**	79.4	**70.5**	**73.6**	**76.0**	**77.0**	79.4



Result

- **Parameter efficiency**
- SparseUNet / PTv3 : Performance of supervised whole model without pre-training
- PC/CSC/MSC (lin.) : linear probing failure of the existing SSL model
- Sonata(lin.) : Improves previous SOTA (21.8%) by 3.3x with minimal parameters, and approaches PTv3 overall learning (77.6%)



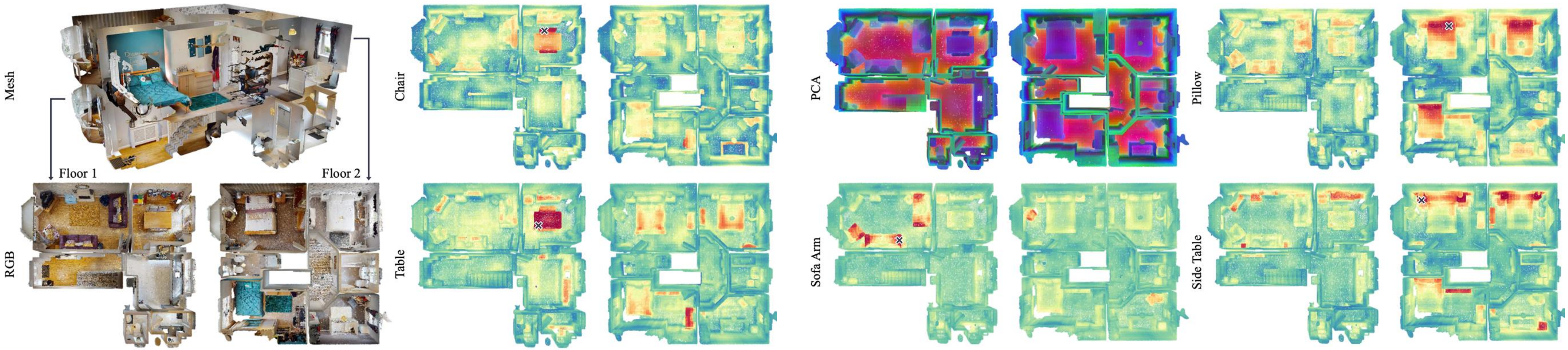
Param. Effciency	Params		ScanNet Val [23]			ScanNet200 Val [71]			ScanNet++ Val [101]			S3DIS Area 5 [1]			S3DIS 6-fold [1]		
Methods	Learn.	Pct.	mIoU	mAcc	allAcc	mIoU	mAcc	allAcc	mIoU	mAcc	allAcc	mIoU	mAcc	allAcc	mIoU	mAcc	allAcc
○ SparseUNet [17]	39.2M	100%	72.3	80.2	90.0	25.0	32.9	80.4	28.8	38.4	80.1	66.3	72.5	89.8	72.4	80.9	89.9
● PC [93] (lin.)	<0.2M	<0.1%	5.6	9.7	50.0	0.5	0.9	40.3	1.8	3.1	46.4	11.4	18.6	52.3	11.7	19.0	51.2
● CSC [38] (lin.)	<0.2M	<0.1%	12.6	18.1	64.2	1.3	2.1	53.0	2.8	4.5	53.6	24.4	32.0	66.4	24.9	32.5	66.9
● MSC [88] (lin.)	<0.2M	<0.1%	14.1	20.3	62.9	1.5	2.5	53.6	4.5	6.6	61.3	27.9	35.5	71.1	29.9	37.9	71.3
○ PTv3 [89]	124.8M	100%	77.6	85.0	92.0	35.3	46.0	83.4	42.1	53.4	85.6	73.4	78.9	91.7	77.7	85.3	91.5
● MSC [88] (lin.)	<0.2M	<0.2%	21.8	32.2	65.5	3.3	5.5	57.5	8.1	11.9	64.7	32.1	42.4	70.9	34.6	46.0	71.3
● Sonata (lin.)	<0.2M	<0.2%	72.5	83.1	89.7	29.3	41.6	81.2	37.3	50.9	84.3	72.3	81.2	90.9	76.5	87.4	90.8
● Sonata (dec.)	16.3M	13%	**79.1**	**86.6**	**92.7**	**33.5**	**44.5**	**84.1**	**40.9**	**52.6**	**86.3**	**74.5**	**80.4**	**92.6**	**81.5**	**88.8**	**93.0**

Table 5. **Parameter efficiency.**

Result

❖ Visualization

- Zero-shot representation across scenes
- Sonata consistently delivers semantically rich and informative representations across diverse indoor environments



Conclusion

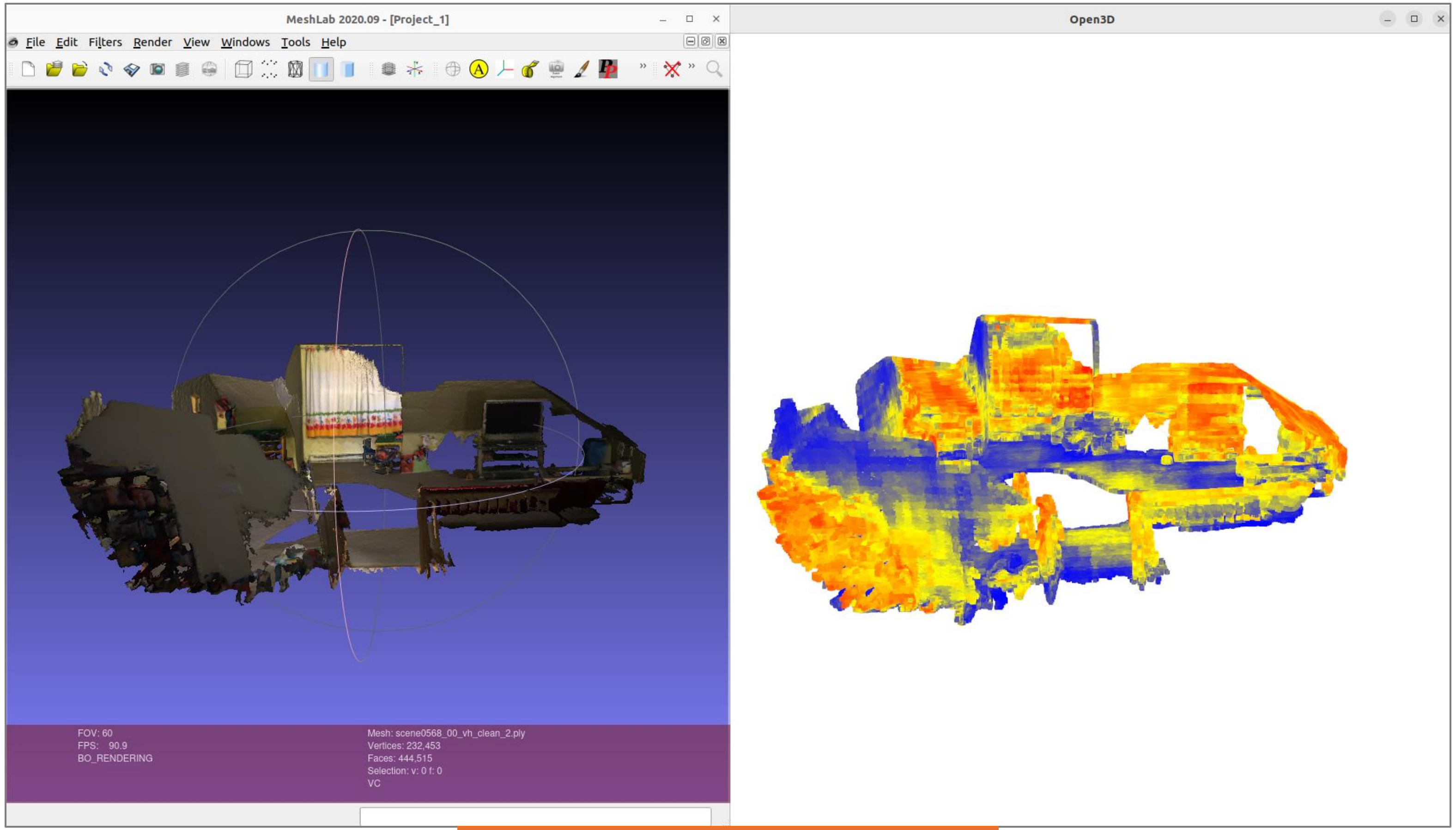
❖ Achievements

1. Sonata's Achievement: This work advances on robust and reliable 3D representation learning with instance-level semantic responses
2. geometric shortcut problem of traditional SSL by linear probing, and address it through a point distillation framework
3. Sonata scales to a 140k point cloud for semantically significant zero-shot visualization and exceptional parameters and data efficiency

❖ Limitations

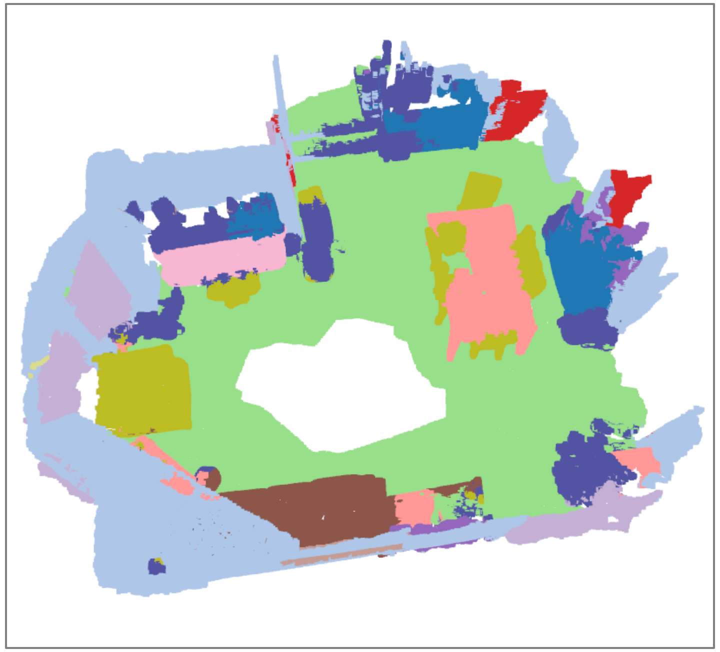
1. Lack of train datasets variance & OOD performance limitations (SONATA mIoU 32.0% vs **PTv3 34.9%)**
2. Potention loss of Decoder-free approach in complex task (ScanNet200 & ScanNet++)
3. Still exists gap : semantic segmentation (Scannet 7.1% & S3DIS 5.8%)

Research


MeshLab 2020.09 - [Project_1]
File Edit Filters Render View Windows Tools Help
Open3D

```
FOV: 60
FPS:    90.9
BO_RENDERING
```

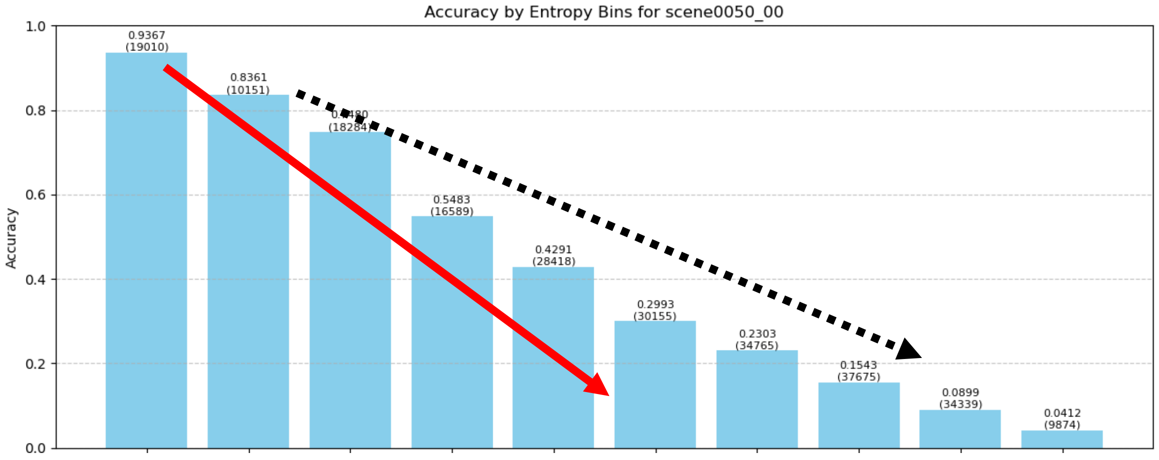
```
Mesh: scene0568_00_vh_clean_2.ply
Vertices: 232,453
Faces: 444,515
Selection: v: 0 f: 0
VC
```

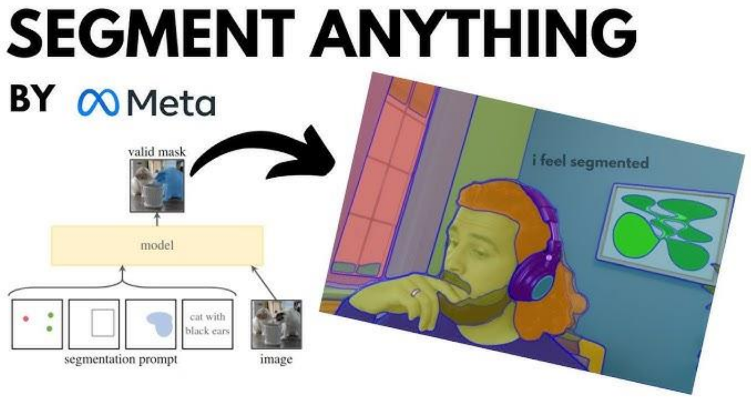


Scannet Dataset

Semantic segmentation

Research


Accuracy by Entropy Bins for scene0050_00
Accuracy
1.0
0.8
0.6
0.4
0.2
0.0
0.9367
(19010)
0.8361
(10151)
0.7480
(18284)
0.5483
(16589)
0.4291
(28418)
0.2993
(30155)
0.2303
(34765)
0.1543
(37675)
0.0899
(34339)
0.0412
(9874)


SEGMENT ANYTHING
BY Meta
valid mask
model
cat with black ears
segmentation prompt
image
i feel segmented

ENCORD
DINOv2:
Self-supervised
Model
Explained

감사합니다.

End of Document

CHUNG-ANG
UNIVERSITY