Research Report

Point Representations for Spatial Cognition

Xiaoyang Wu

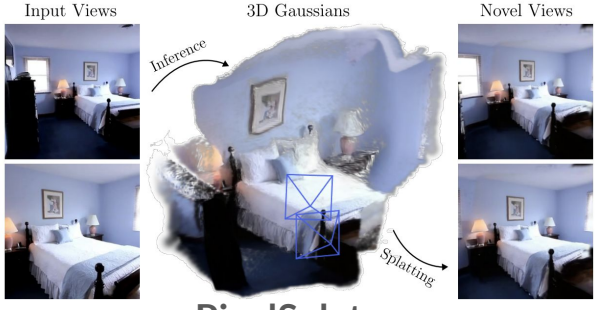
June 23, 2025

香 港 大 學
THE UNIVERSITY OF HONG KONG

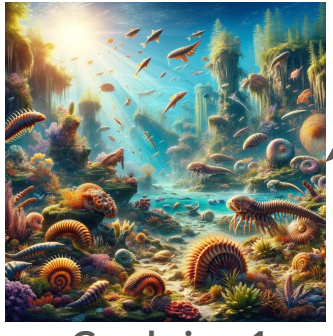
POINTCEPT
Point Cloud Perception Codebase

DINOv2 is changing how we conduct research with images

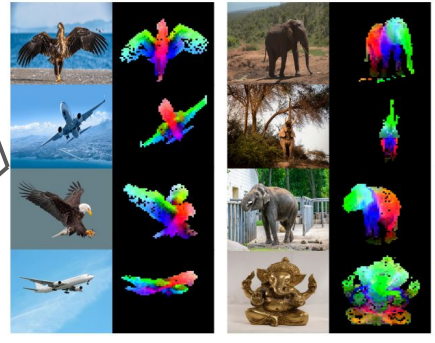
Now, if we processing images, we will **definitely consider** using it.
And it rarely **disappoints us.**



PixelSplat
(Reconstruction)



Cambrian-1
(Multimodal)



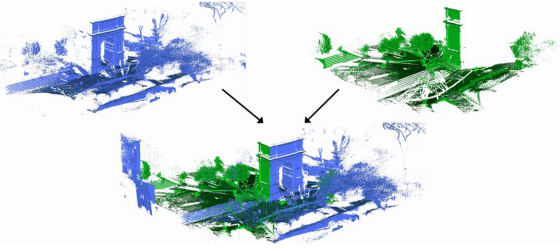
DINO Series
(Image SSL)

Depth Anything
(Perception)

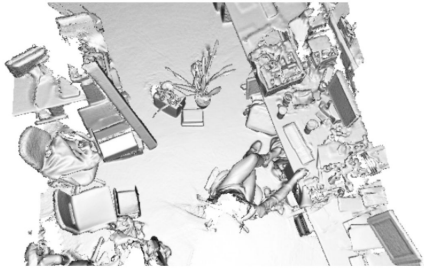
AnyDoor
(Generation)

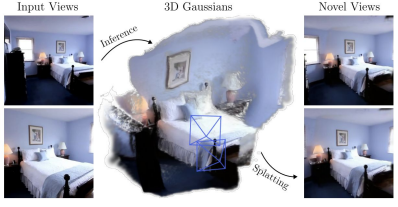
There are many tasks related to 3D space that we care

Do we have **one model** can benefit all of them, efficient and effectively?

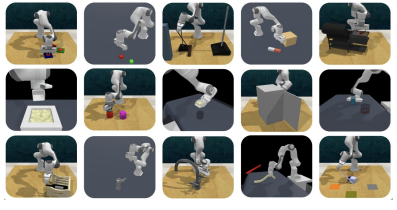


3D Localization

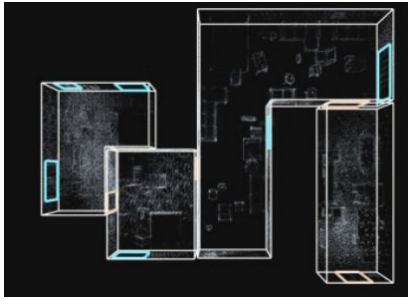


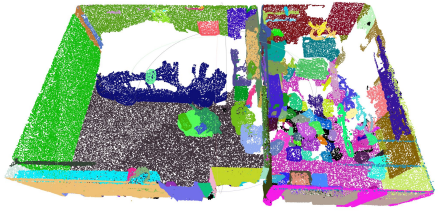


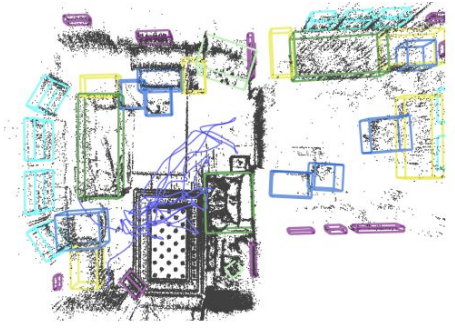
3D Reconstruction



3D Planning & Interaction

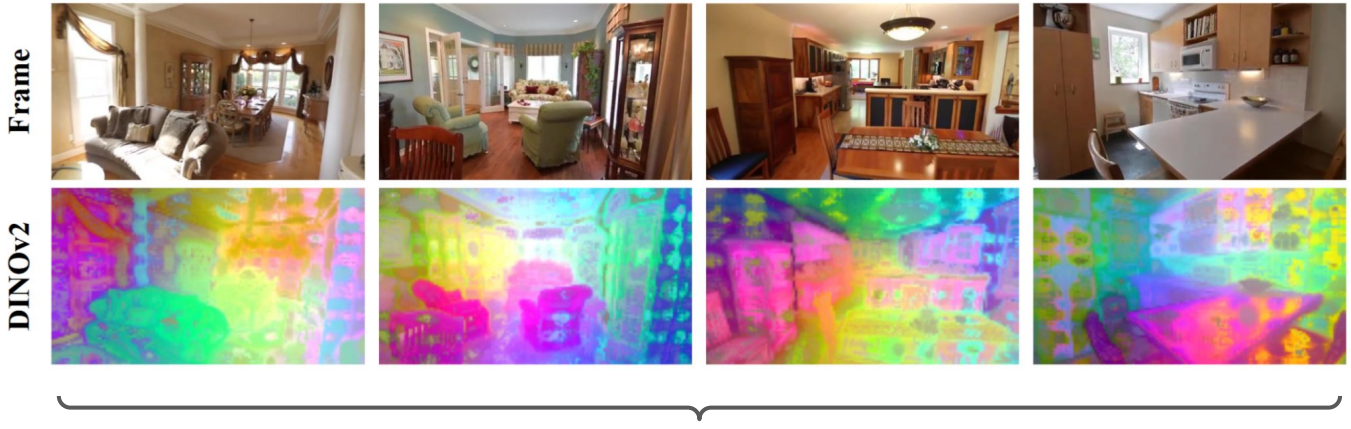


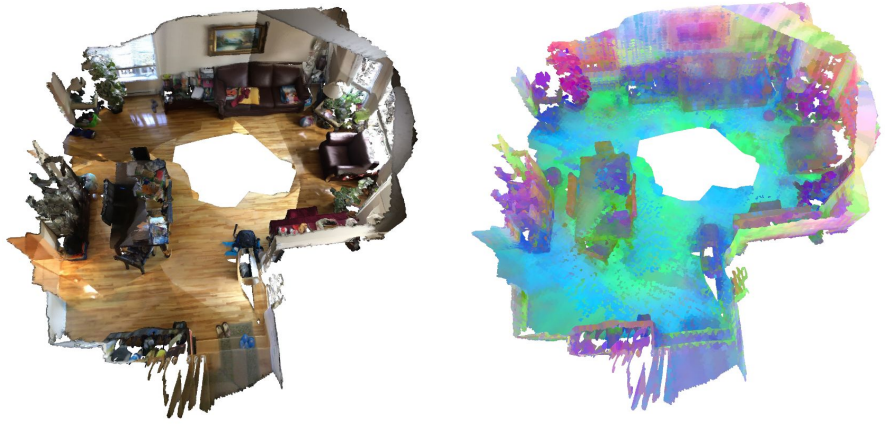




3D Perception

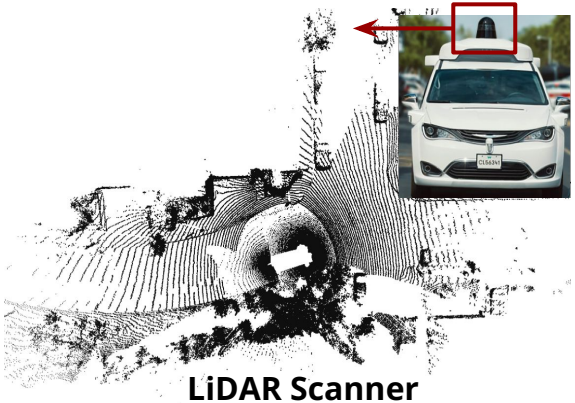
Image processing with DINOs might be good, but ...





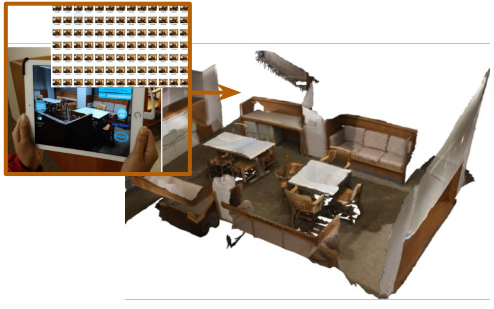
- It requires **significant inference time** to process a scene video with thousands of frames.
- It **lacks geometric awareness**, as it is not natively encoded in 3D space.

We believe the point cloud is an ideal data structure for spatial cognition.

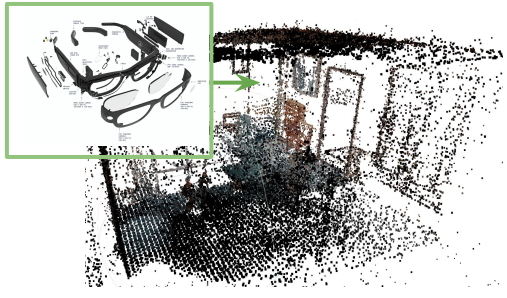


LiDAR Scanner

Laser Scanner



RGB-D Video

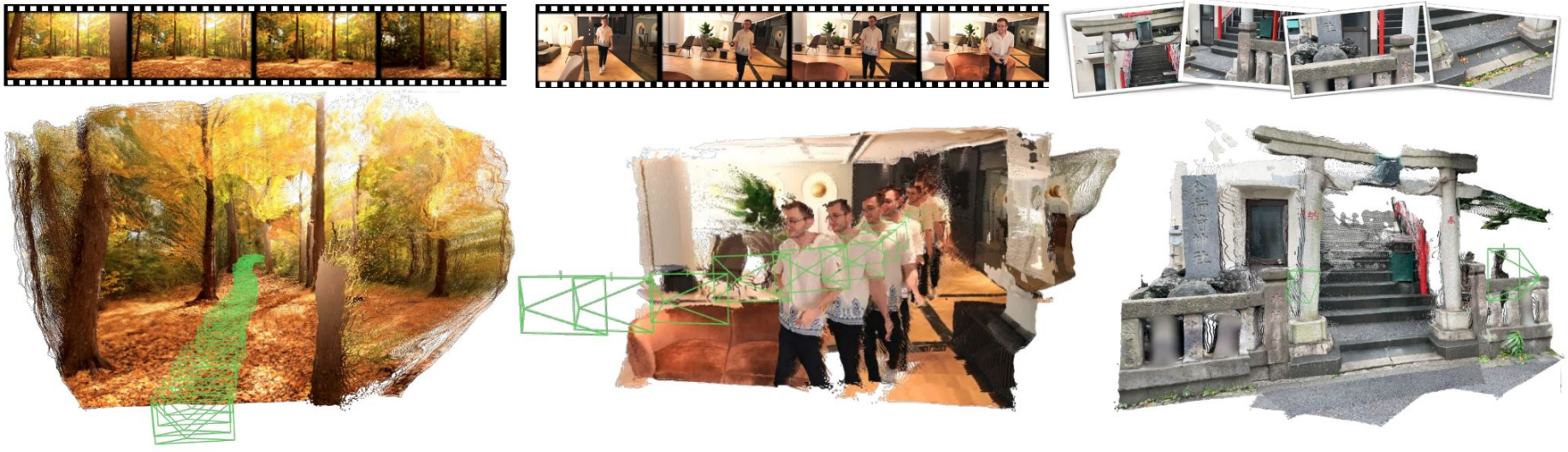


SLAM System

Existing Model Assets

Point cloud is everywhere and easy to acquire

We believe the point cloud is an ideal data structure for spatial cognition.



Static Scene | Dynamic Scene | Sparse Photo Collection

Dust3R, VGGT, Fast3R, Cut3R, Foundation Stereo …

We can lift point clouds from "unlimited" dynamic videos

We believe the point cloud is an ideal data structure for spatial cognition.

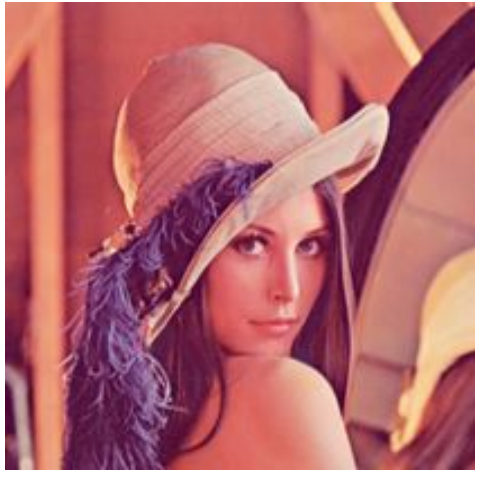
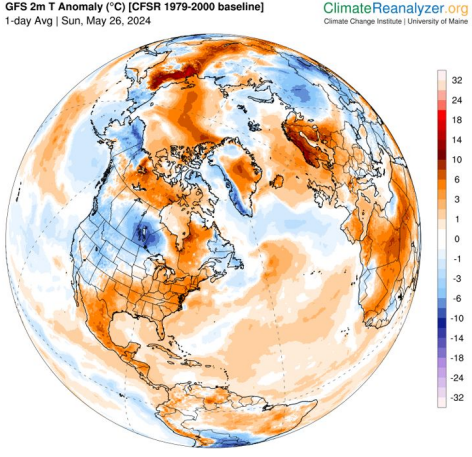
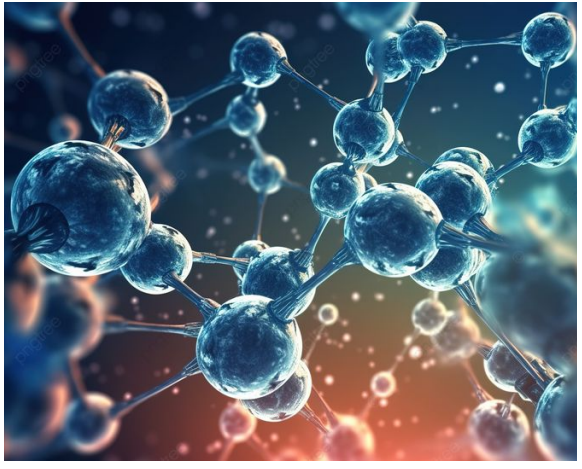


Image (2D)



Climate (2.5D)

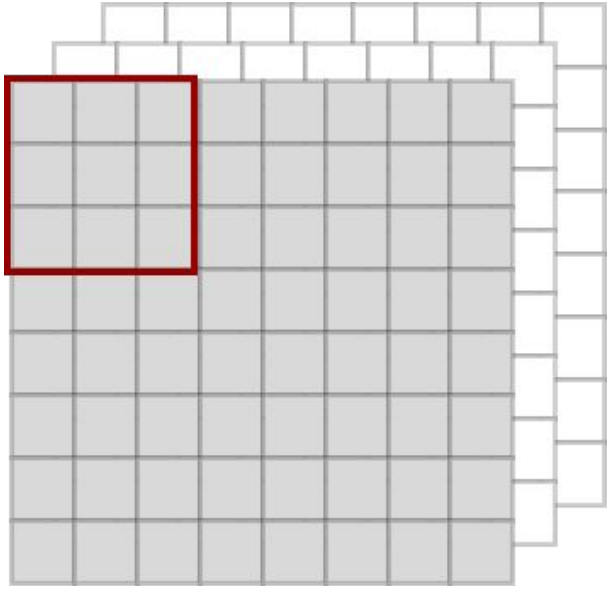


Molecular Structure (3D)

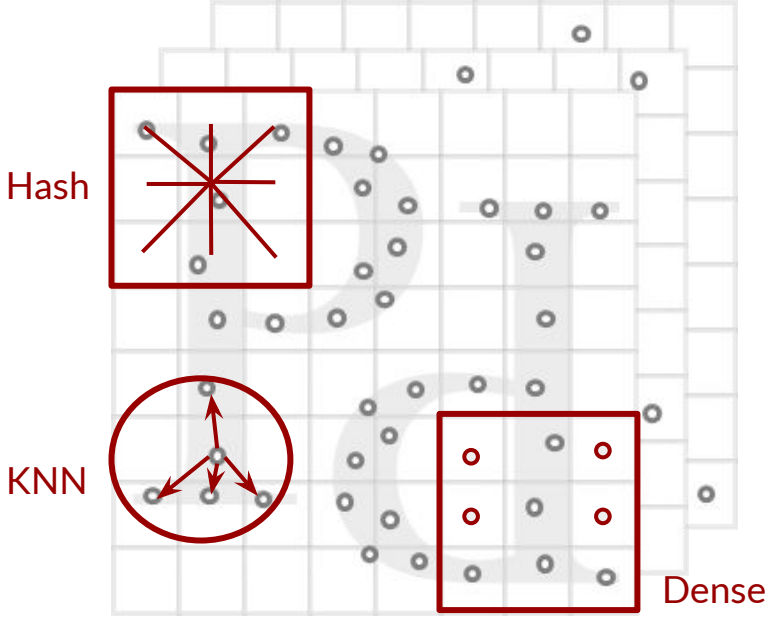
High dimensional sparse data

Or many modalities actually can be viewed as "point cloud"

The 1st key issue: efficient, scalable encoding with sparse point clouds



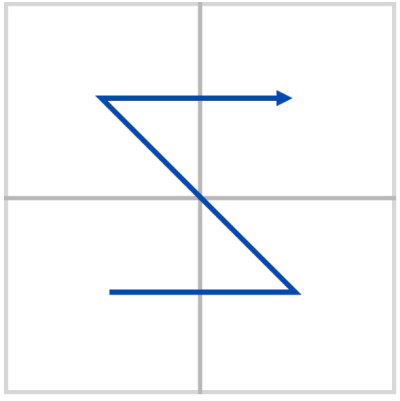
2D Image
(dense data)

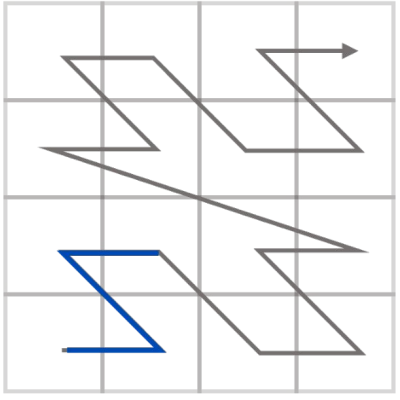


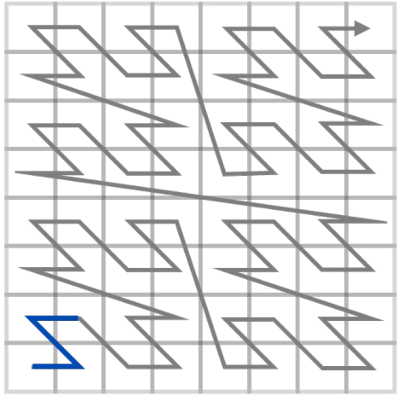
2D Point Cloud
(sparse data)

Space-Filling Curve

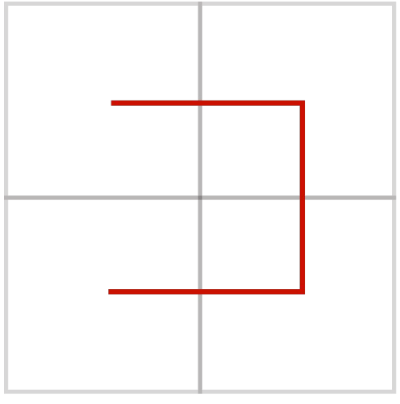
Z-Order

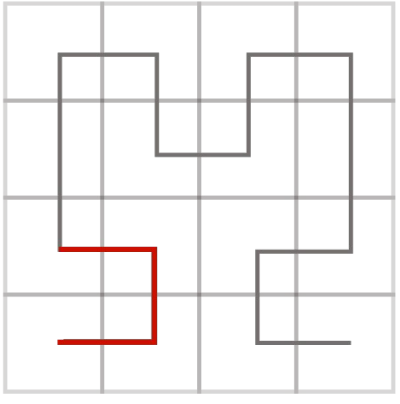


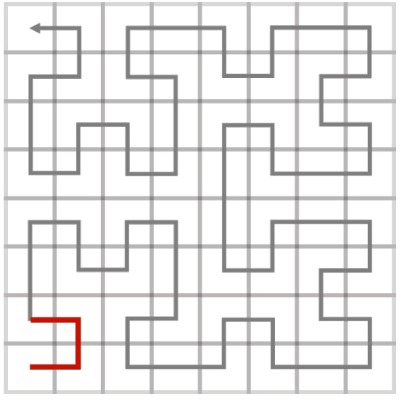




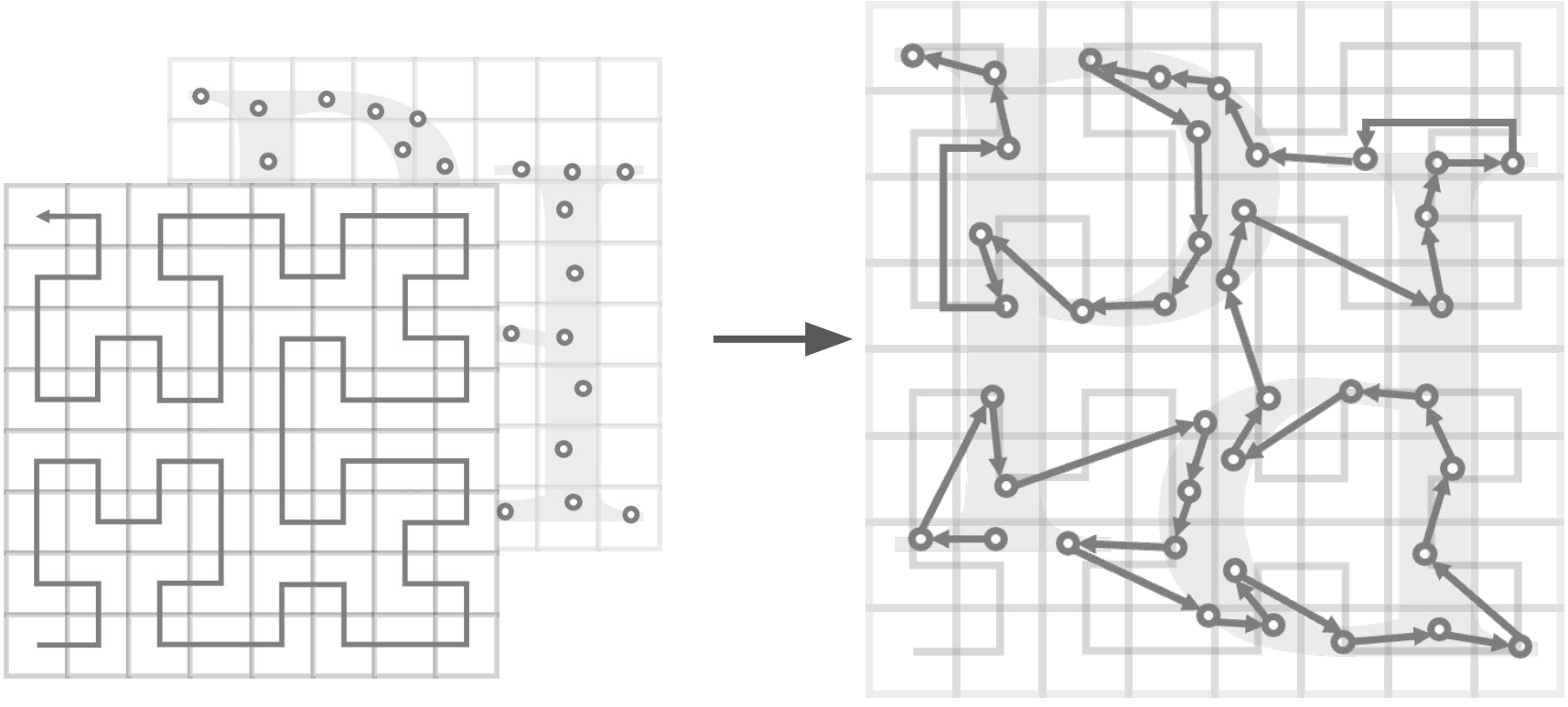
Hilbert Order



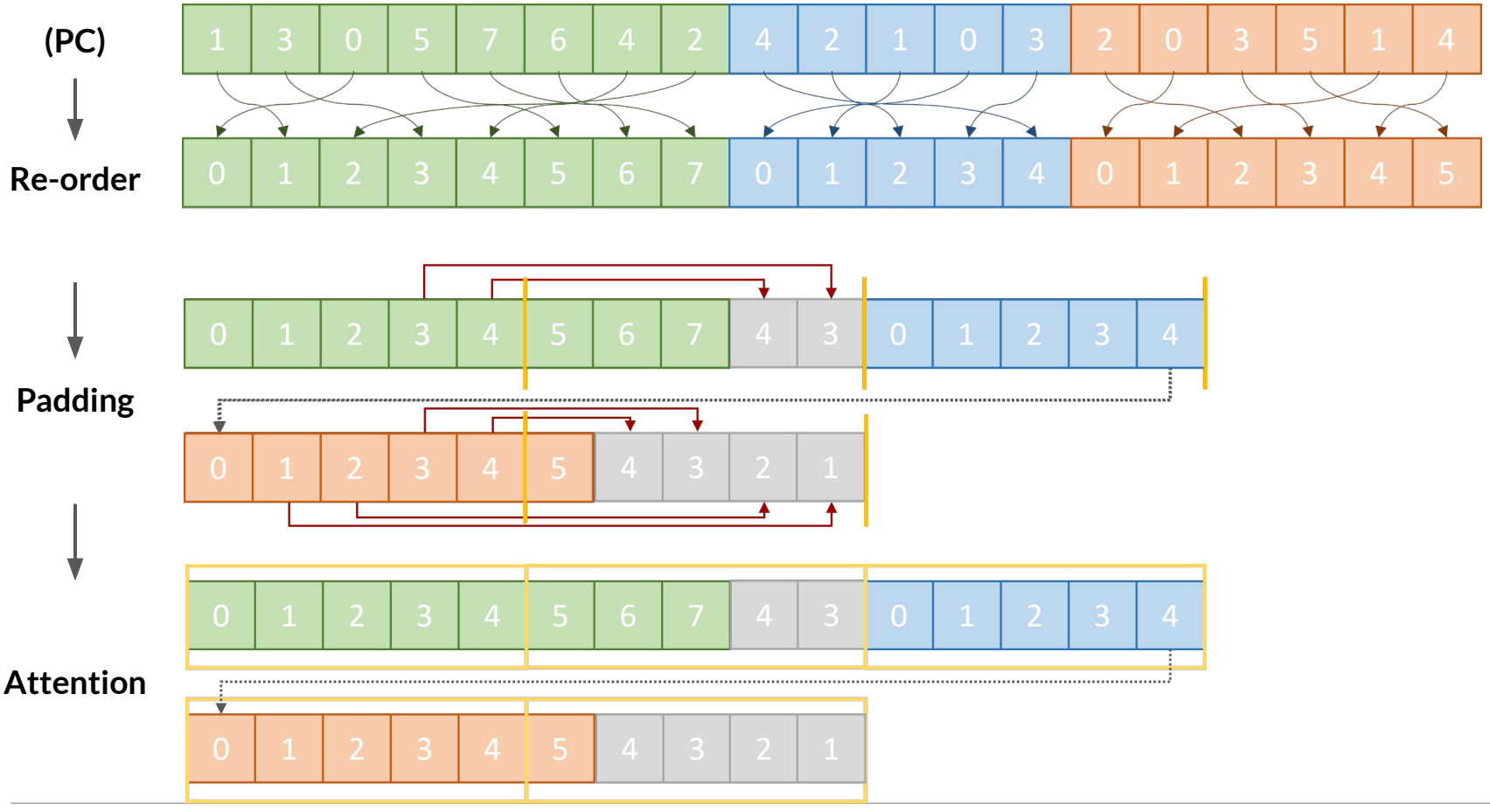




Serialization

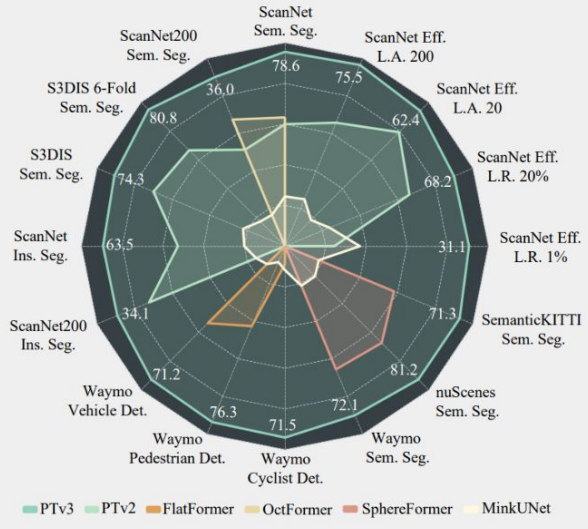


Serialized Attention

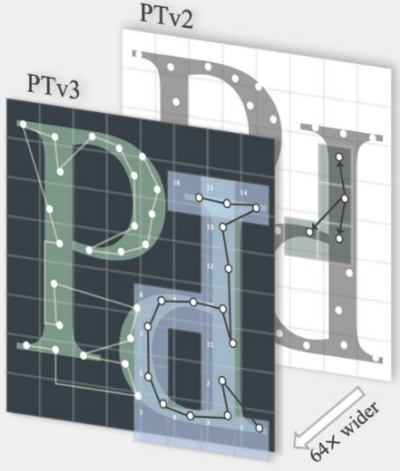


Point Transformer V3: Simpler, Faster, Stronger

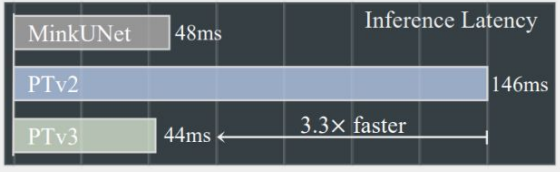
— CVPR 2024 Oral



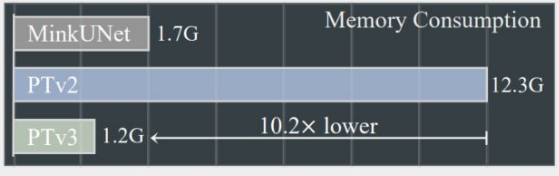
Stronger Performance



Wider Receptive Field



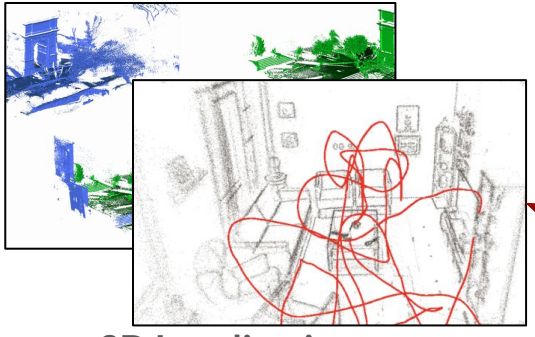
Faster Speed



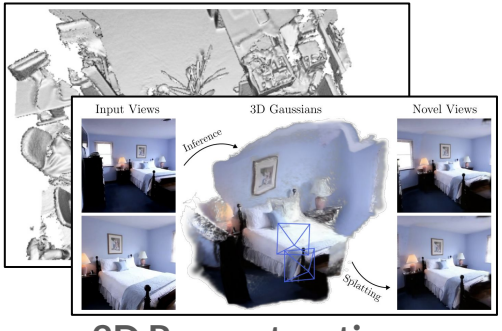
Lower Memory Consumption

We wish to build one point cloud model to shape people's trust as DINOs

It shows the potential to **benefit all 3D tasks**;

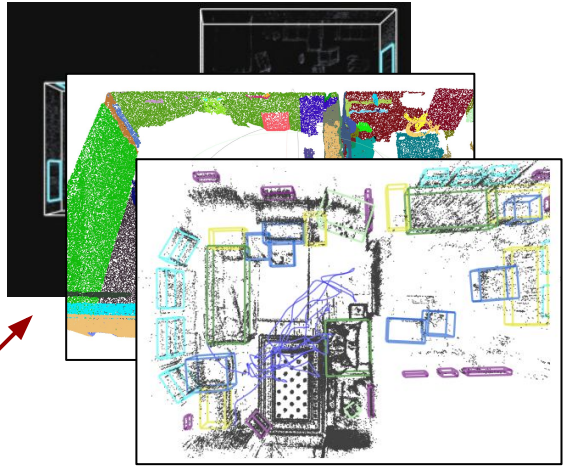


3D Localization

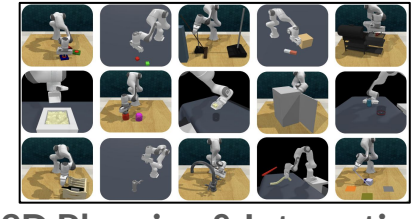


3D Reconstruction

Sonata
(Point Cloud SSL)

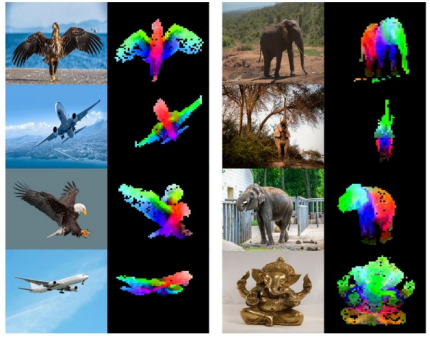


3D Perception



3D Planning & Interaction

Two key factors contribute to the powerful capabilities of representation



DINO Series
(Image SSL)

The magic of data:

The foundation of the intelligence

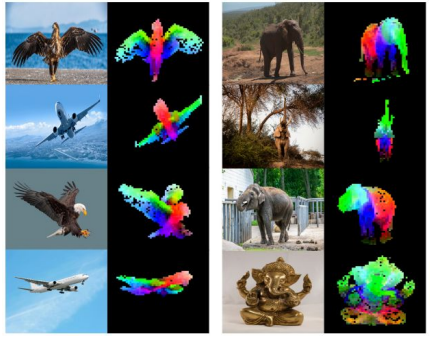
142M High Quality Images

Task	Dataset / Split	Images	Retrieval	Retrieved	Final
classification	ImageNet-22k / –	14,197,086	as is	–	14,197,086
classification	ImageNet-22k / –	14,197,086	sample	56,788,344	56,788,344
classification	ImageNet-1k / train	1,281,167	sample	40,997,344	40,997,344
fine-grained classif.	Caltech 101 / train	3,030	cluster	2,630,000	1,000,000
fine-grained classif.	CUB-200-2011 / train	5,994	cluster	1,300,000	1,000,000
fine-grained classif.	DTD / train1	1,880	cluster	1,580,000	1,000,000
fine-grained classif.	FGVC-Aircraft / train	3,334	cluster	1,170,000	1,000,000
fine-grained classif.	Flowers-102 / train	1,020	cluster	1,060,000	1,000,000
fine-grained classif.	Food-101 / train	75,750	cluster	21,670,000	1,000,000
fine-grained classif.	Oxford-IIIT Pet / trainval	3,680	cluster	2,750,000	1,000,000
fine-grained classif.	Stanford Cars / train	8,144	cluster	7,220,000	1,000,000
fine-grained classif.	SUN397 / train1	19,850	cluster	18,950,000	1,000,000
fine-grained classif.	Pascal VOC 2007 / train	2,501	cluster	1,010,000	1,000,000
segmentation	ADE20K / train	20,210	cluster	20,720,000	1,000,000
segmentation	Cityscapes / train	2,975	cluster	1,390,000	1,000,000
segmentation	Pascal VOC 2012 (seg.) / trainaug	1,464	cluster	10,140,000	1,000,000
depth estimation	Mapillary SLS / train	1,434,262	as is	–	1,434,262
depth estimation	KITTI / train (Eigen)	23,158	cluster	3,700,000	1,000,000
depth estimation	NYU Depth V2 / train	24,231	cluster	10,850,000	1,000,000
depth estimation	SUN RGB-D / train	4,829	cluster	4,870,000	1,000,000
retrieval	Google Landmarks v2 / train (clean)	1,580,470	as is	–	1,580,470
retrieval	Google Landmarks v2 / train (clean)	1,580,470	sample	6,321,880	6,321,880
retrieval	AmsterTime / new	1,231	cluster	960,000	960,000
retrieval	AmsterTime / old	1,231	cluster	830,000	830,000
retrieval	Met / train	397,121	cluster	62,860,000	1,000,000
retrieval	Revisiting Oxford / base	4,993	cluster	3,680,000	1,000,000
retrieval	Revisiting Paris / base	6,322	cluster	3,660,000	1,000,000
					142,109,386

The magic of representation learning:

An art of learning knowledge from “infinite” data

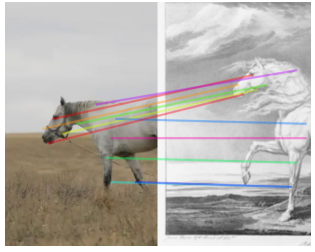
Two key factors lead to people's trust to DINOv2



DINO Series
(Image SSL)

Zero-shot Visualization:

See the representation with your own eye



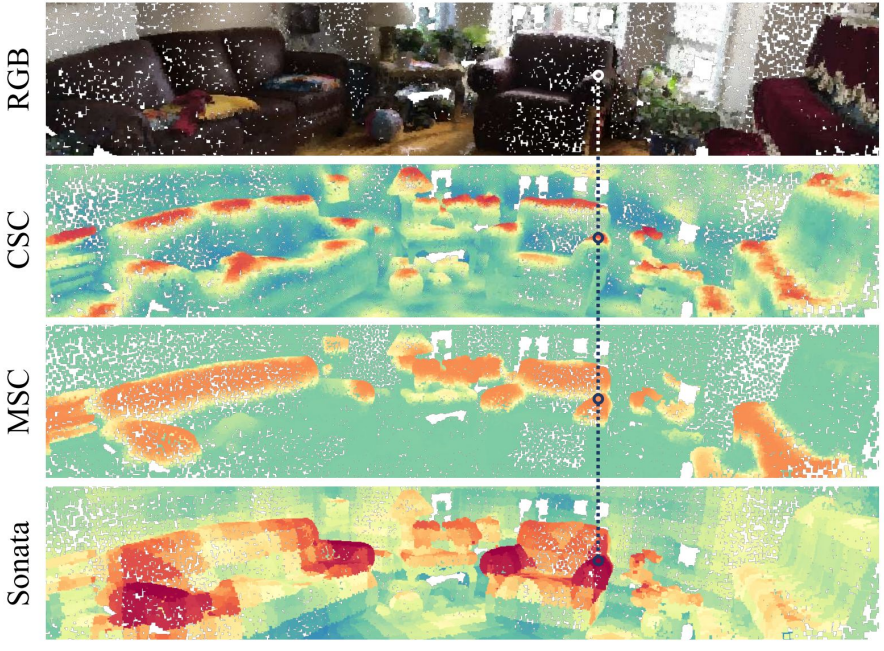
Linear Probing

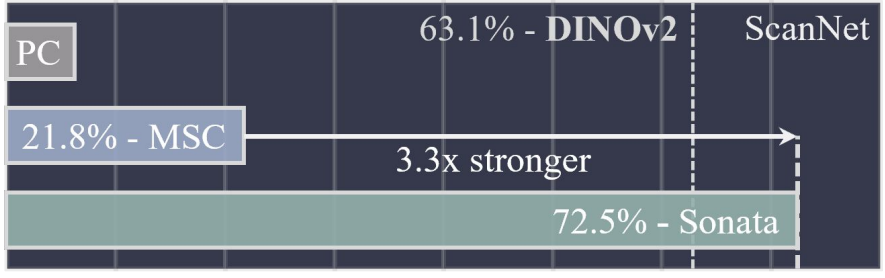
True robustness benchmarked by linear probing

	INet-1k k-NN	INet-1k linear
iBOT	72.9	82.3
+(our reproduction)	74.5 ↑ 1.6	83.2 ↑ 0.9
+LayerScale, Stochastic Depth	75.4 ↑ 0.9	82.0 ↓ 1.2
+128k prototypes	76.6 ↑ 1.2	81.9 ↓ 0.1
+KoLeo	78.9 ↑ 2.3	82.5 ↑ 0.6
+SwiGLU FFN	78.7 ↓ 0.2	83.1 ↑ 0.6
+Patch size 14	78.9 ↑ 0.2	83.5 ↑ 0.4
+Teacher momentum 0.994	79.4 ↑ 0.5	83.6 ↑ 0.1
+Tweak warmup schedules	80.5 ↑ 1.1	83.8 ↑ 0.2
+Batch size 3k	81.7 ↑ 1.2	84.7 ↑ 0.9
+Sinkhorn-Knopp	81.7 =	84.7 =
+Untying heads = DINOv2	82.0 ↑ 0.3	84.5 ↓ 0.2

We never adopt the two metric to point representation learning before

Let's remove the fig leaf to see how previous models perform ...





* With <0.2% Learnable Parameters

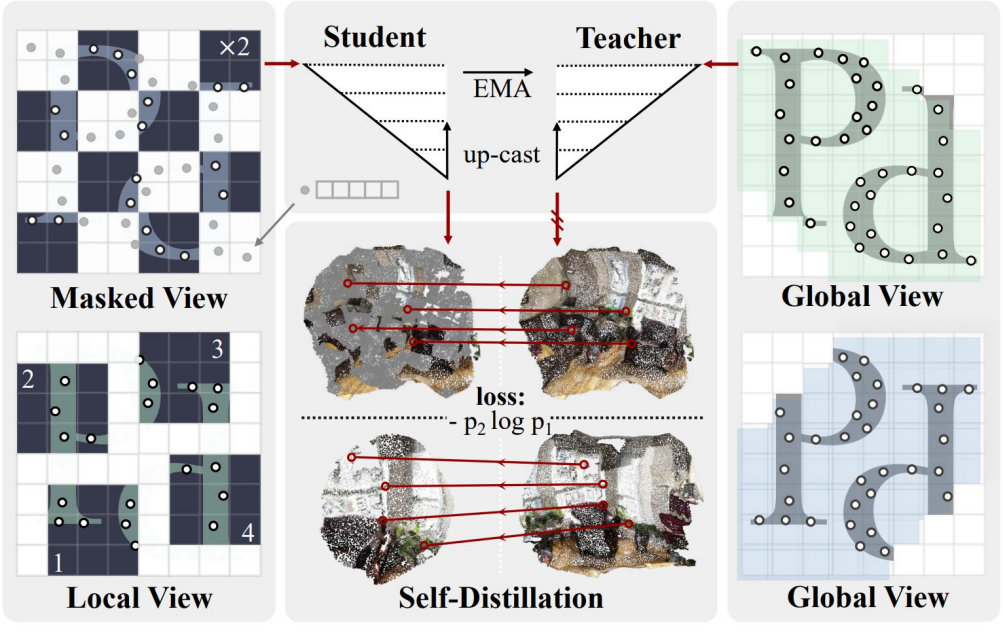
Linear Probing

True robustness benchmarked by linear probing

Zero-shot Visualization:

See the representation with your own eye

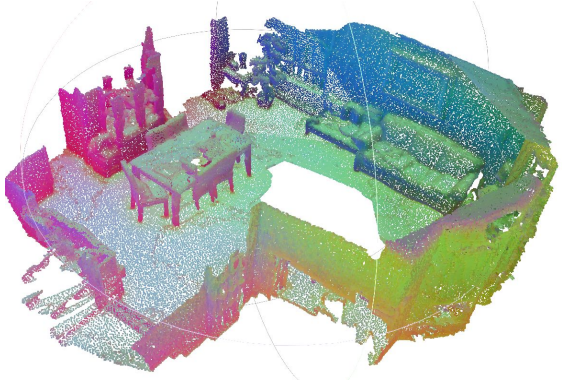
Every point cloud SSL research begin with imitate image SSL

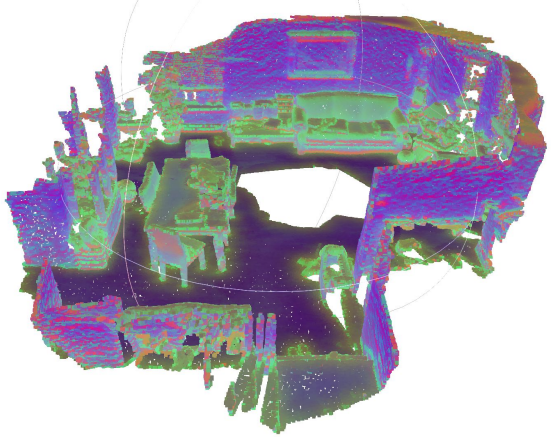


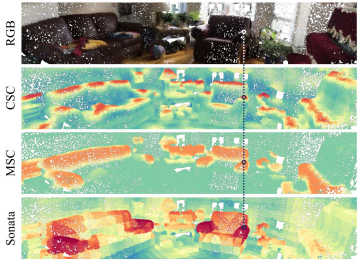
- Local-Global Alignment
- Mask-Unmask Alignment
- EMA
- Online Clustering
- Sinkhorn-Knopp centering
- KoLeo regularizer

But this won't lead to strong point representations; collapse still exists!

PCA Visualization from Prototype of Sonata



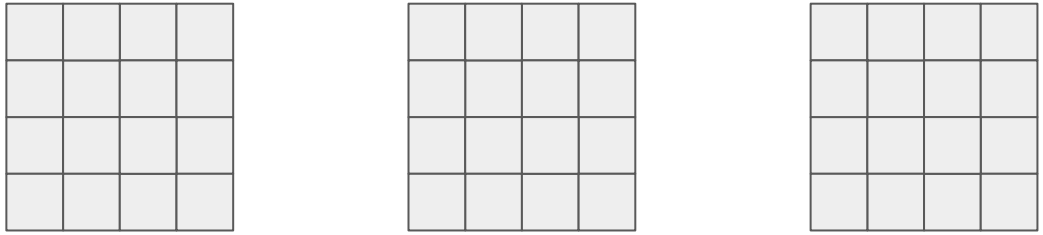




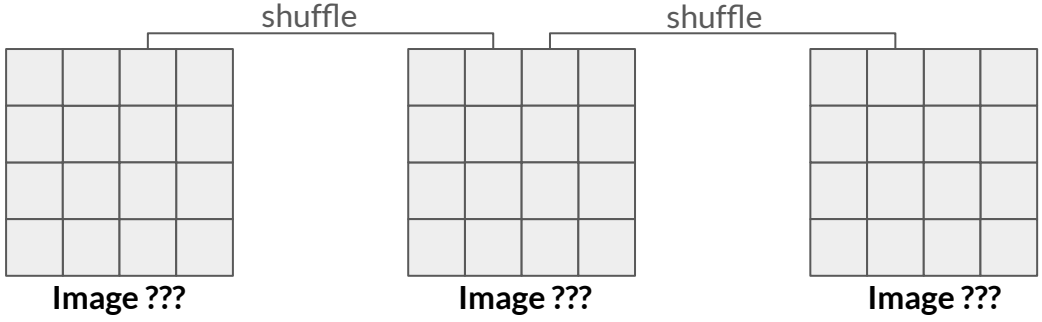
A unique shortcut must exist specific to the 3D Point Cloud
The Geometric Shortcuts

Why point cloud SSL easily collapse to geometric information

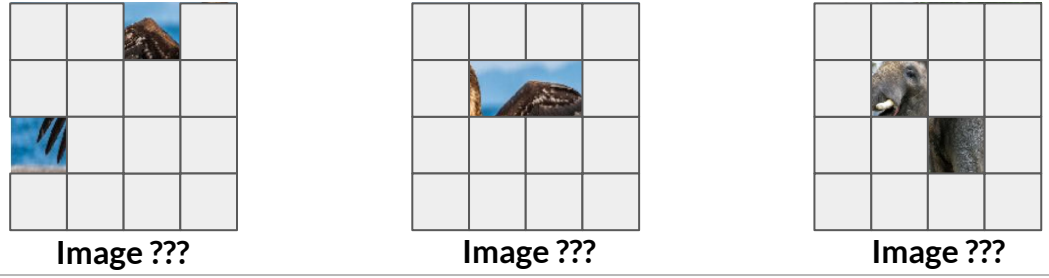
Let's understand it with a simple game.



> Image **A'** and Image **A** are same.
> Image **B** and Image **A** are different.
> Mask out **all the feature**



> **Shuffle** the order of three images.
> Can we distinguish **which is A and A'**?
> **No**, we can't without any feature



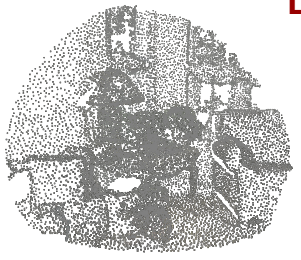
> **How about add some feature back?**
> **Yes**, now we can!!

>> Image SSL have to relies on feature and impossible to collapse into geometry information (coordinate)

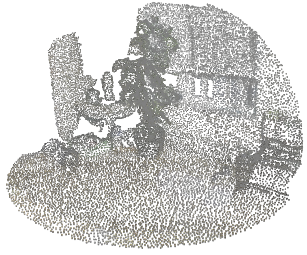
Why point cloud SSL easily collapse to geometric information

Let's understand it with a simple game.

Point Cloud A



Point Cloud A'

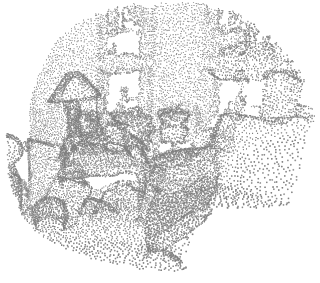


Point Cloud B

> PC **A'** and PC **A** are same.
> PC **B** and PC **A** are different.
> Mask out **all the feature**

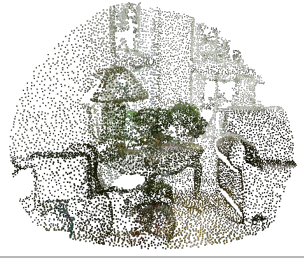
Point Cloud ~~A~~?

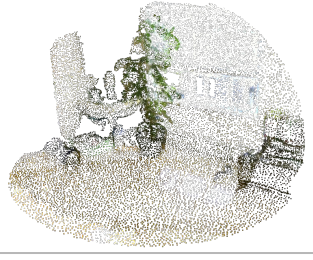
Point Cloud ~~B~~?

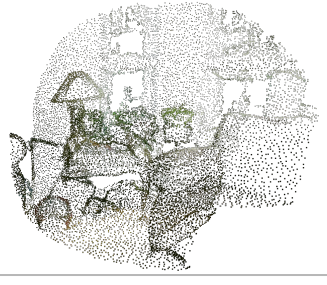


Point Cloud ~~A~~?

> **Shuffle** the order of three images.
> Can we distinguish **which is A and A'**?
> **Yes, we can. Easy!**



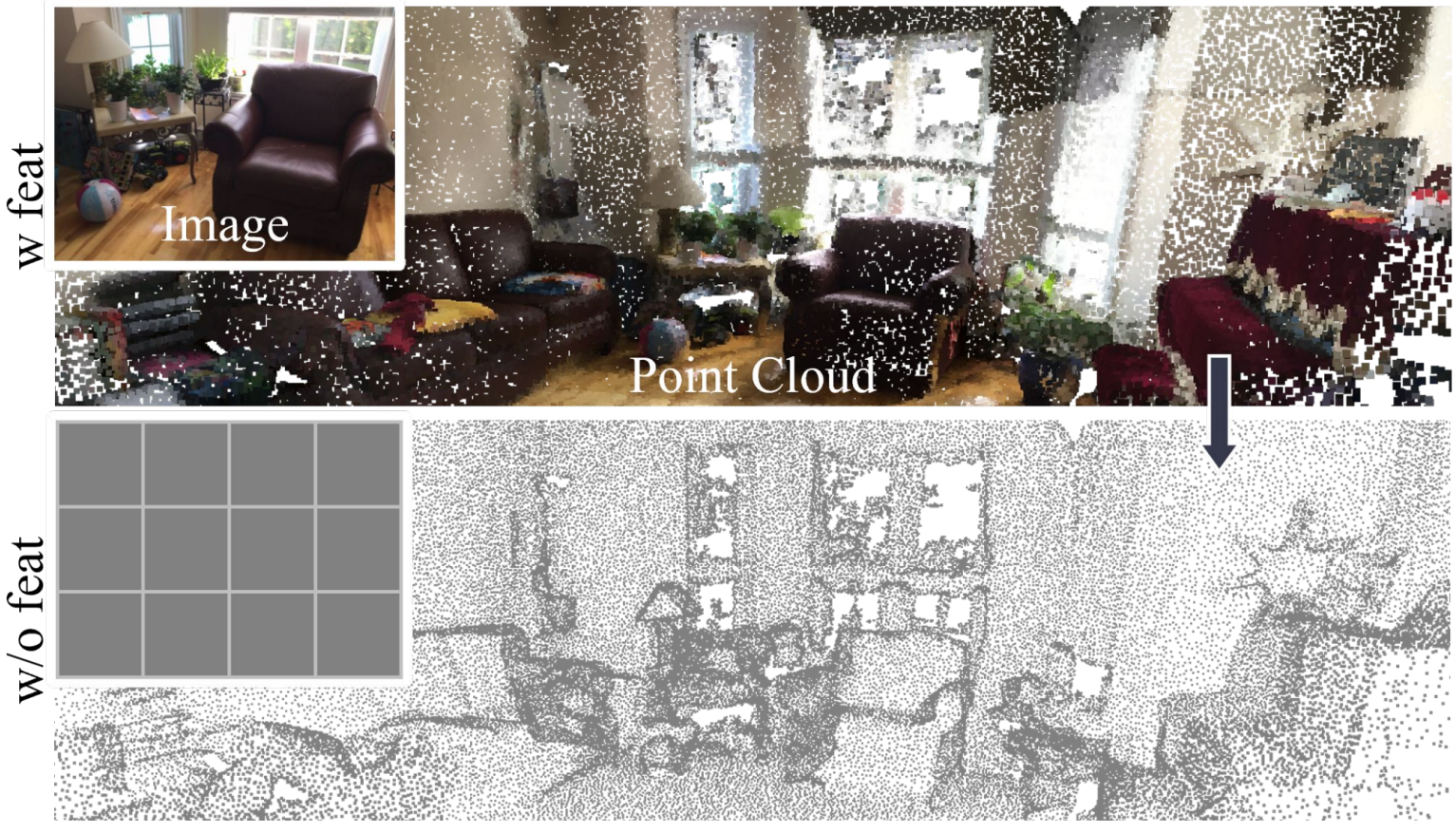




> Do we really need other feature?
> **No, no need.**

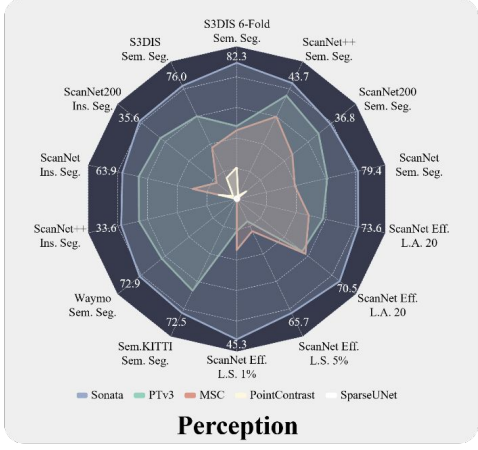
>> Point cloud SSL exist a shortcut that only rely on geometry information

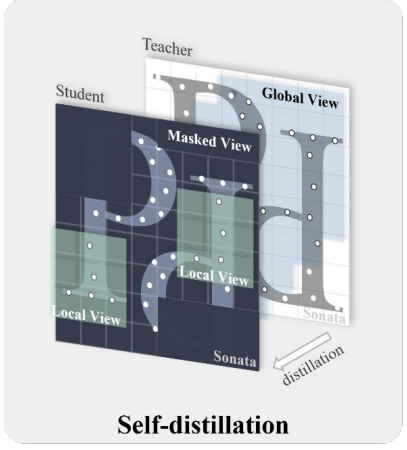
Why point cloud SSL easily collapse to geometric information

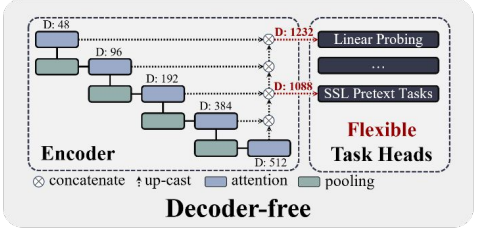


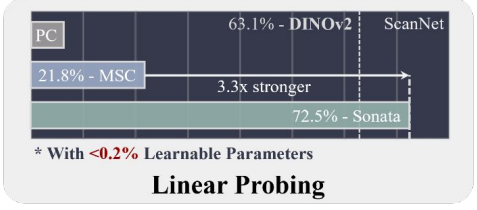
Sonata: Self-Supervised Learning of Reliable Point Representations

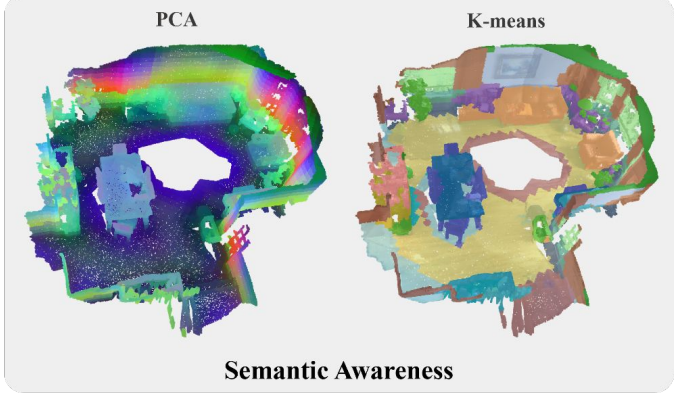
— CVPR 2025 Highlight

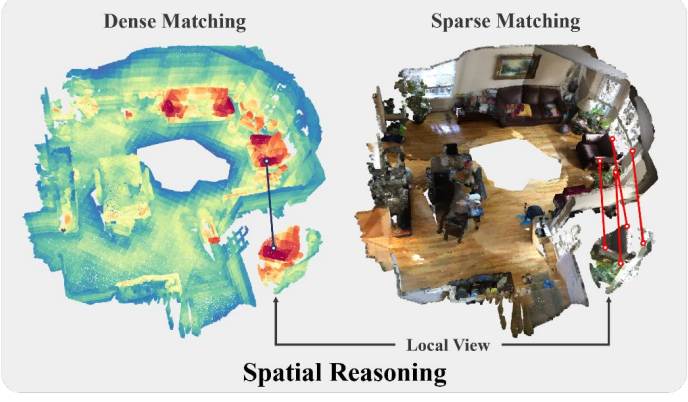




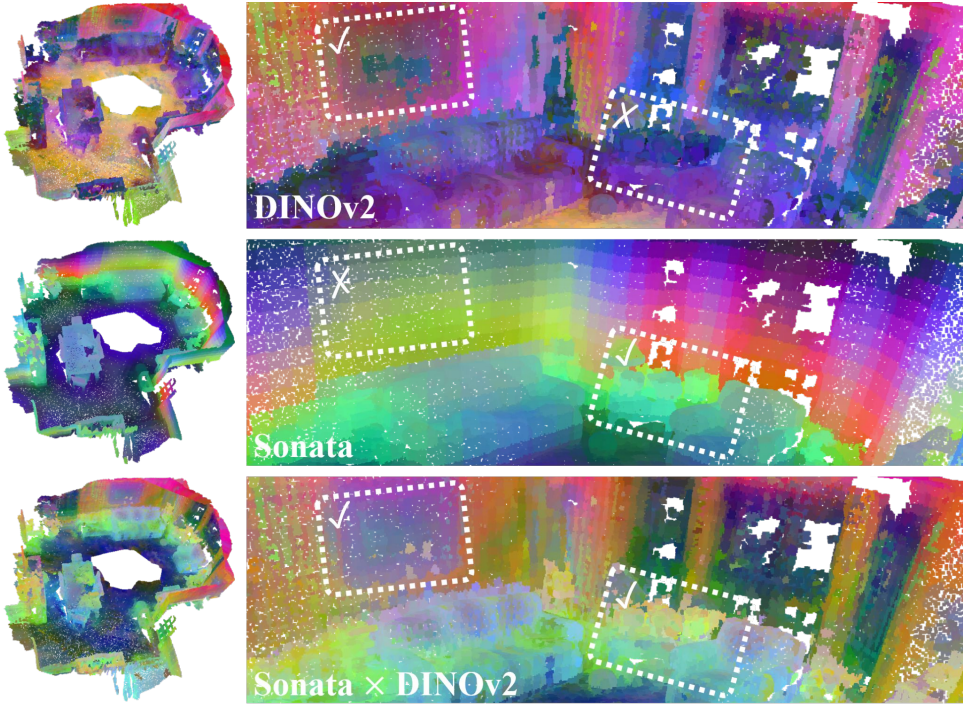








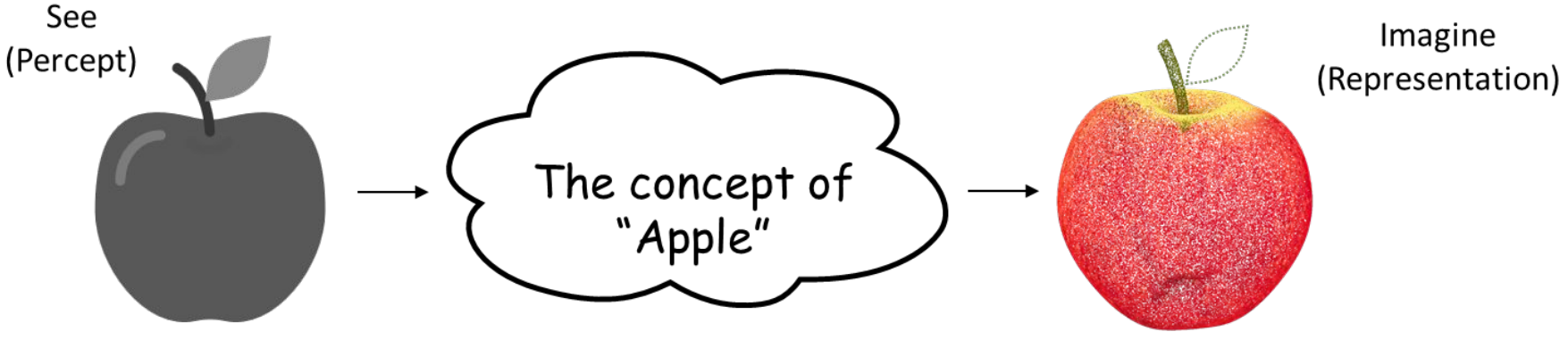
One more thing: A richer feature space exists!



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

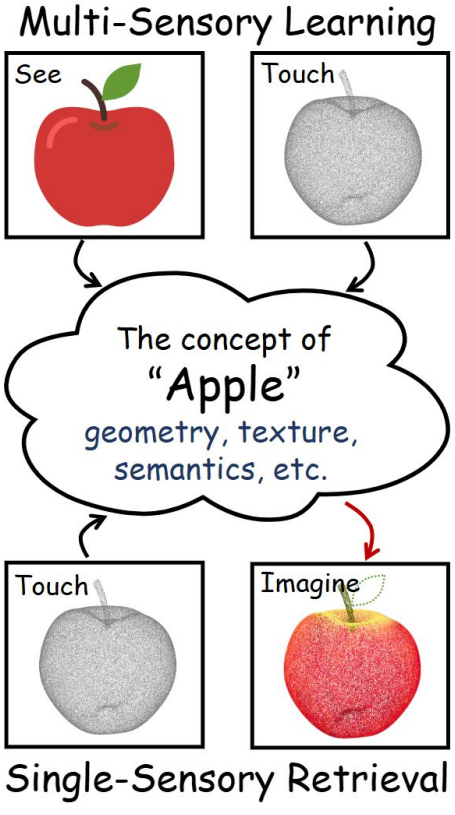
Concatenating features from DINOv2 and Sonata leads to improved linear probing performance
=> **each modality captures complementary, rather than redundant**, aspects of spatial information.

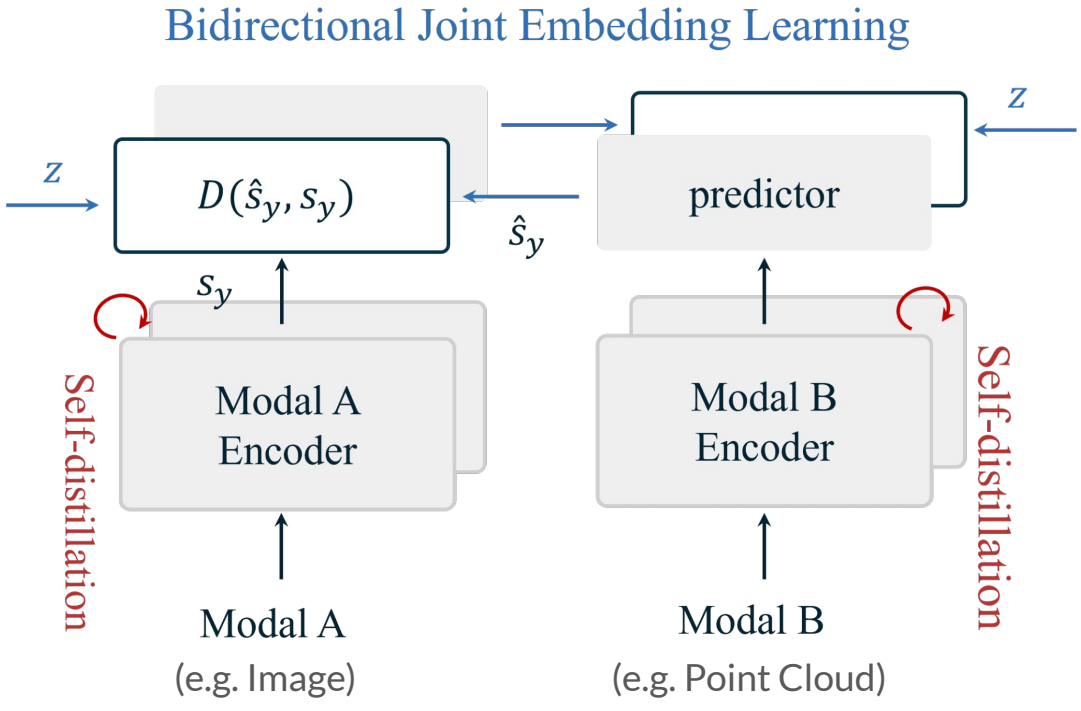
Humans are able to retrieve such a rich multisensory representation with just a single modal input.



How human learn such an informative representation?

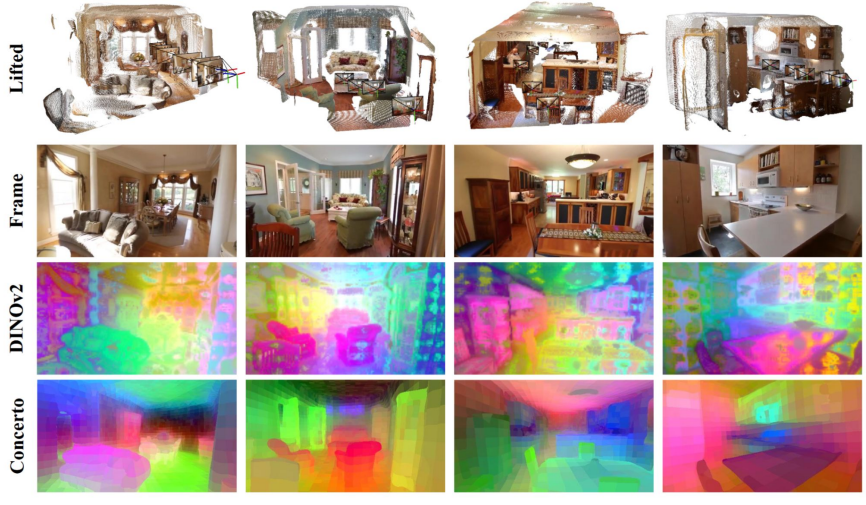
— with multisensory synergy!





Concerto: Joint 2D-3D Self-Supervised Learning Emerges Spatial Representations

Semantic Segmentation			ScanNet Val [15]			ScanNet200 Val [33]		
Method	Type	Encoder	mIoU	mAcc	allAcc	mIoU	mAcc	allAcc
DINOv2 [26]	2D Image SSL	ViT-G	63.09	75.50	82.42	27.42	37.59	72.80
Sonata [47]	3D Point SSL	PTv3-B	72.52	83.11	89.74	29.25	41.61	81.15
Sonata×DINOv2	3D SSL×2D SSL	Both	75.91	85.36	91.25	36.67	46.98	82.85
Concerto (ours)	2D-3D Joint SSL	PTv3-B	**77.32**	**86.58**	**91.74**	**37.41**	**49.49**	**83.29**



What are we fight for?

Building Strong

Point Representations for Spatial Cognition

Building a great community for better and better point cloud representations

POINTCEPT

Point Cloud Perception Codebase

Formatter passing

Pointcept is a powerful and flexible codebase for point cloud perception research. It is also an official implementation of the following paper:

- 🚀 **Sonata: Self-Supervised Learning of Reliable Point Representations**
  Xiaoyang Wu, Daniel DeTone, Duncan Frost, Tianwei Shen, Chris Xie, Nan Yang, Jakob Engel, Richard Newcombe, Hengshuang Zhao, Julian Straub
  IEEE Conference on Computer Vision and Pattern Recognition (**CVPR**) 2025 - Highlight
  [Pretrain] [Sonata] - [Project] [arXiv] [Bib] [Demo] [Weight] → here
- **Point Transformer V3: Simpler, Faster, Stronger**
  Xiaoyang Wu, Li Jiang, Peng-Shuai Wang, Zhijian Liu, Xihui Liu, Yu Qiao, Wanli Ouyang, Tong He, Hengshuang Zhao
  IEEE Conference on Computer Vision and Pattern Recognition (**CVPR**) 2024 - Oral
  [Backbone] [PTv3] - [arXiv] [Bib] [Project] → here
- **OA-CNNs: Omni-Adaptive Sparse CNNs for 3D Semantic Segmentation**
  Bohao Peng, Xiaoyang Wu, Li Jiang, Yukang Chen, Hengshuang Zhao, Zhuotao Tian, Jiaya Jia
  IEEE Conference on Computer Vision and Pattern Recognition (**CVPR**) 2024
  [Backbone] [OA-CNNs] - [arXiv] [Bib] → here
- **Towards Large-scale 3D Representation Learning with Multi-dataset Point Prompt Training**
  Xiaoyang Wu, Zhuotao Tian, Xin Wen, Bohao Peng, Xihui Liu, Kaicheng Yu, Hengshuang Zhao
  IEEE Conference on Computer Vision and Pattern Recognition (**CVPR**) 2024
  [Pretrain] [PPT] - [arXiv] [Bib] → here
- **Masked Scene Contrast: A Scalable Framework for Unsupervised 3D Representation Learning**
  Xiaoyang Wu, Xin Wen, Xihui Liu, Hengshuang Zhao
  IEEE Conference on Computer Vision and Pattern Recognition (**CVPR**) 2023
  [Pretrain] [MSC] - [arXiv] [Bib] → here
- **Learning Context-aware Classifier for Semantic Segmentation** (3D Part)
  Zhuotao Tian, Jiequan Cui, Li Jiang, Xiaojuan Qi, Xin Lai, Yixin Chen, Shu Liu, Jiaya Jia
  AAAI Conference on Artificial Intelligence (**AAAI**) 2023 - Oral
  [SemSeg] [CAC] - [arXiv] [Bib] [2D Part] → here
- **Point Transformer V2: Grouped Vector Attention and Partition-based Pooling**
  Xiaoyang Wu, Yixing Lao, Li Jiang, Xihui Liu, Hengshuang Zhao
  Conference on Neural Information Processing Systems (**NeurIPS**) 2022
  [Backbone] [PTv2] - [arXiv] [Bib] → here

We invited you to join our point cloud community!

Play with it, use it and trust it

Thank you!