Research Report

# Self-supervised Point Representations for Spatial Cognition

Xiaoyang Wu

June 9, 2025

香 港 大 學
THE UNIVERSITY OF HONG KONG

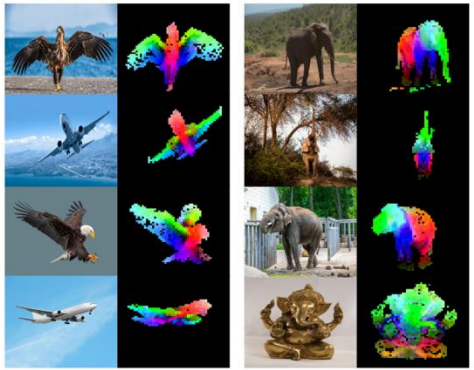
POINTCEPT
Point Cloud Perception Codebase

# DINOv2 is changing how we research

We were initially impressed by its **semantically meaningful visualization** of representations

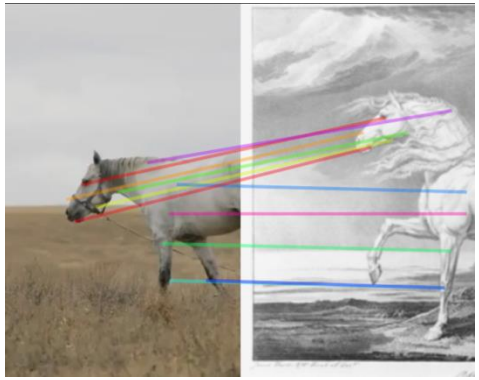
We found it to be super robust in many tasks, even with a **simple linear probing**.

PCA



Retrieval

Dense matching

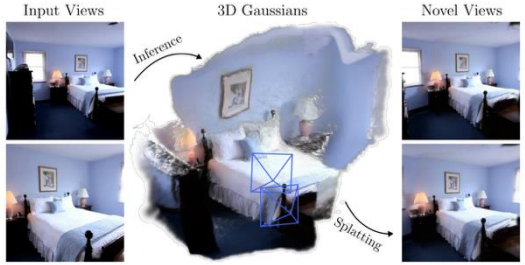


Sparse matching

# DINOv2 is changing how we research

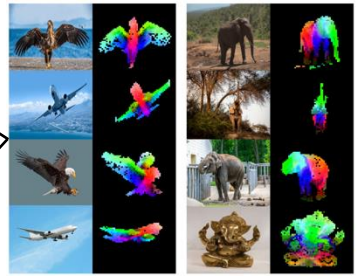
If we process images, we **absolutely consider** using it,
and it **rarely disappoints** us!

**This shapes trust and reliability in our minds!**




PixelSplat
(Reconstruction)

Cambrian-1
(Multimodal)

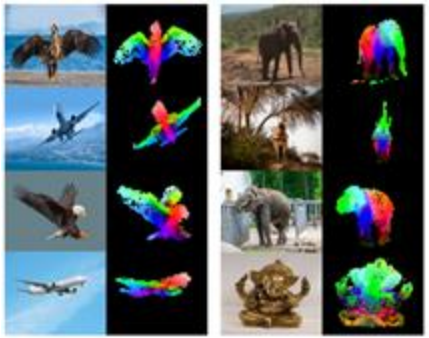


DINO Series
(Image SSL)

Depth Anything
(Perception)

AnyDoor
(Generation)

# Two key factors contribute to the powerful capabilities of representation



DINO Series
(Image SSL)

## The magic of data:

The foundation of the intelligence

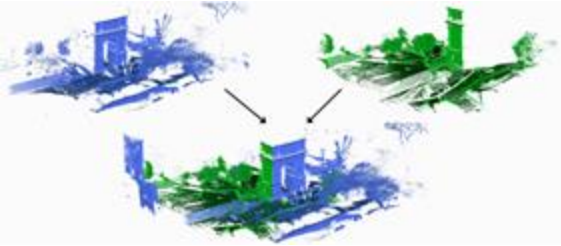
**142M** High Quality Images

| 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 | 18,910,000 | 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 | 1,280,000 | 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 | 4,970,000 | 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 | 3,060,000 | 1,000,000 |
| retrieval | Revisiting Paris / base | 6,322 | cluster | 3,060,000 | 1,000,000 |
| | | | | | 142,509,266 |

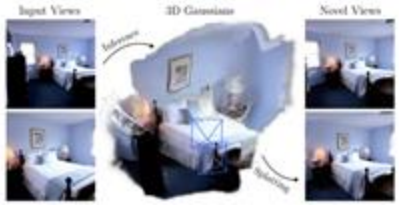
## The magic of self-supervised learning:

The potential of leverage “infinite” data

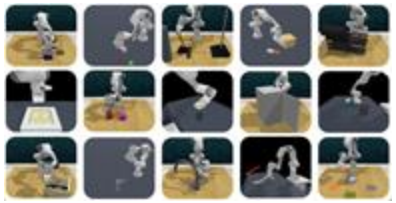
# There are so many spatial tasks we care about!



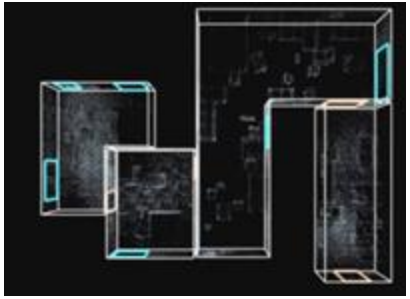
3D Localization

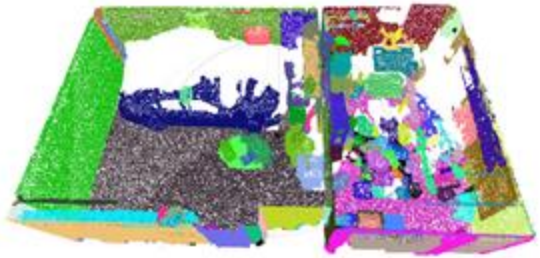


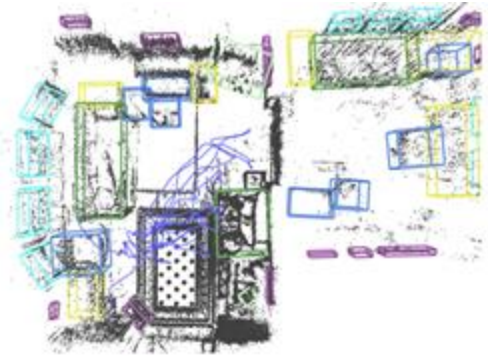

3D Reconstruction



3D Planning & Interaction







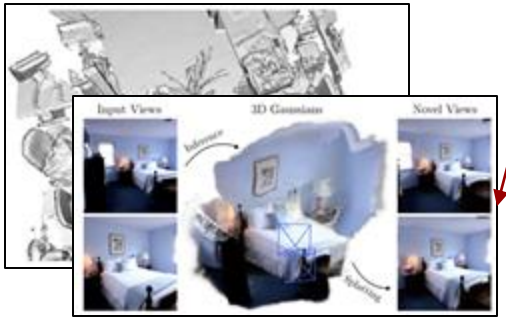
3D Perception

# What's the DINO time of 3D?

It leverages the **magic of data and learning**;

It (target at) **unify and benefit all 3D tasks**;

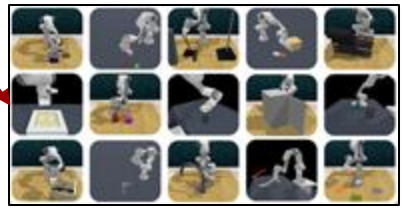
3D Localization




3D Reconstruction

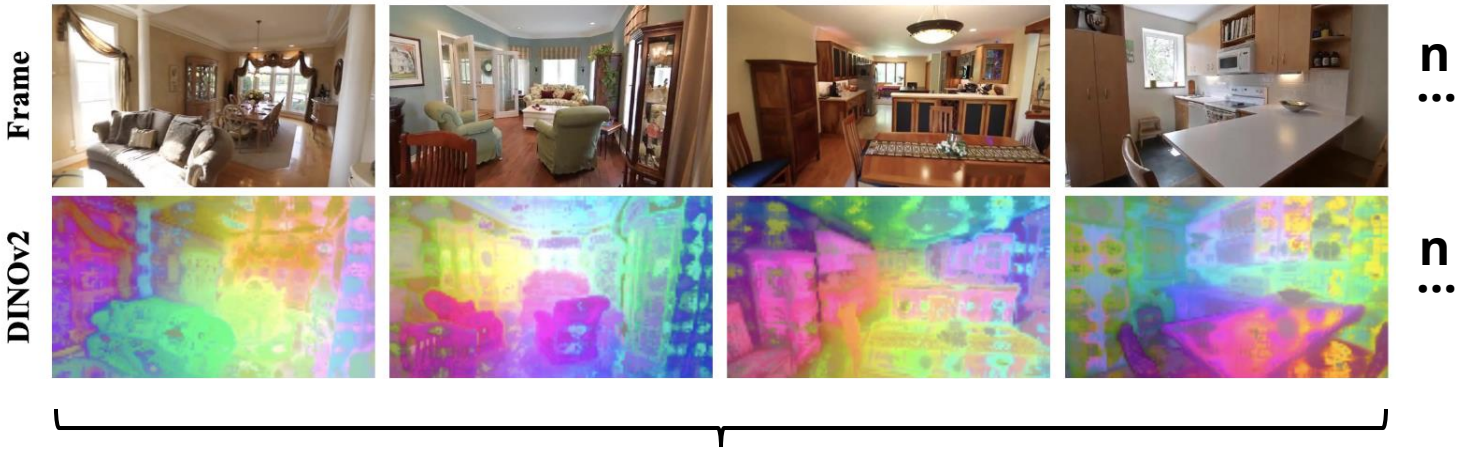
???

3D Perception



3D Planning & Interaction

## We can absolutely solve them using images, but ...




**While sometimes it is good, several restrictions remain:**

- 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.

Why point cloud?

- Point cloud is **everywhere** and **easy to acquire**

LiDAR Scanner


Laser Scanner

RGB-D Video

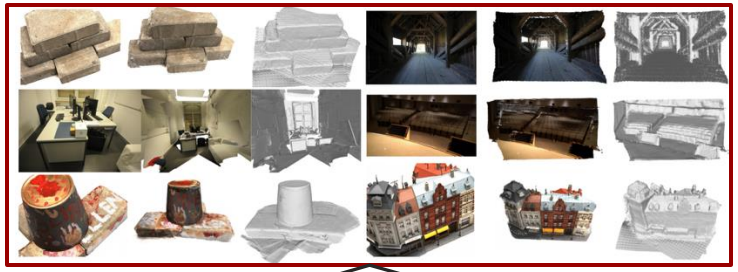
SLAM System

Existing Model Assets

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

Why point cloud?

- Now we can also get 3D/4D point cloud from dynamic videos

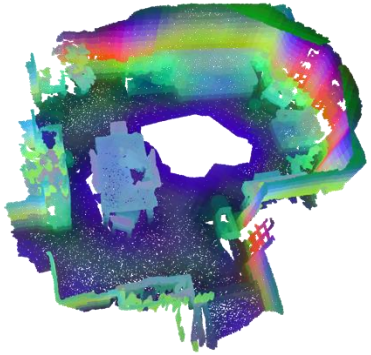


DUSt3R

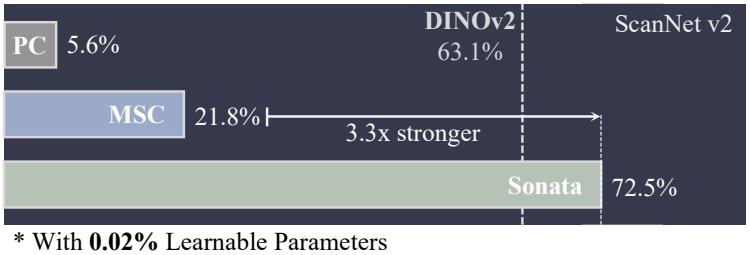
VGGT

CUT3R

# Let's see what kind of representations we can acquire from these sparse point clouds.

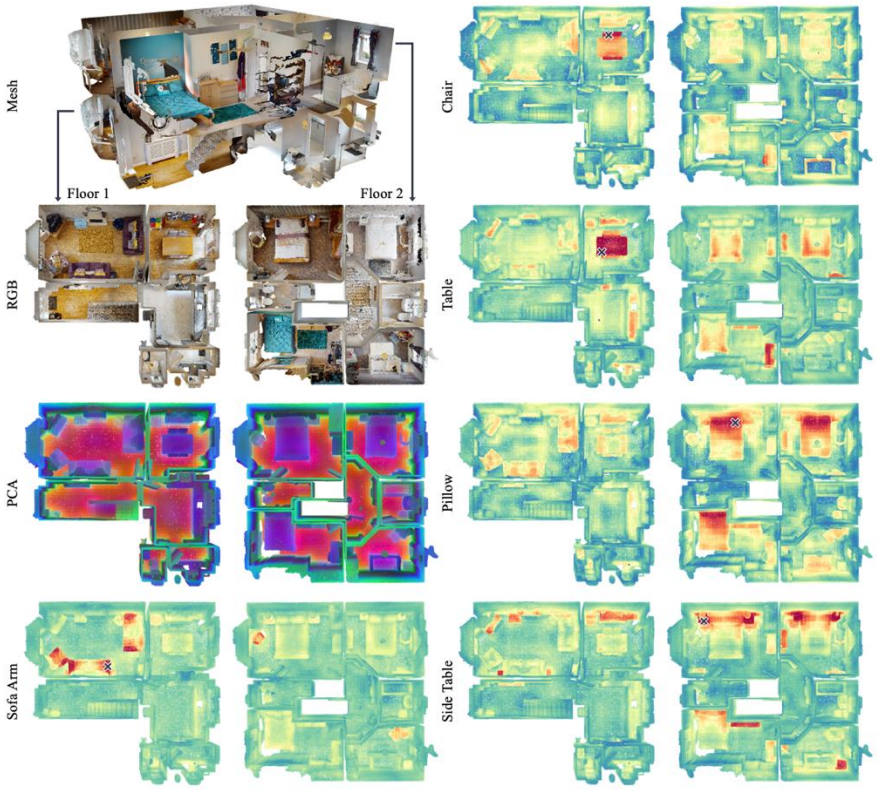


PCA Visualization




* With **0.02%** Learnable Parameters

Linear Probing




Similarity Heatmap

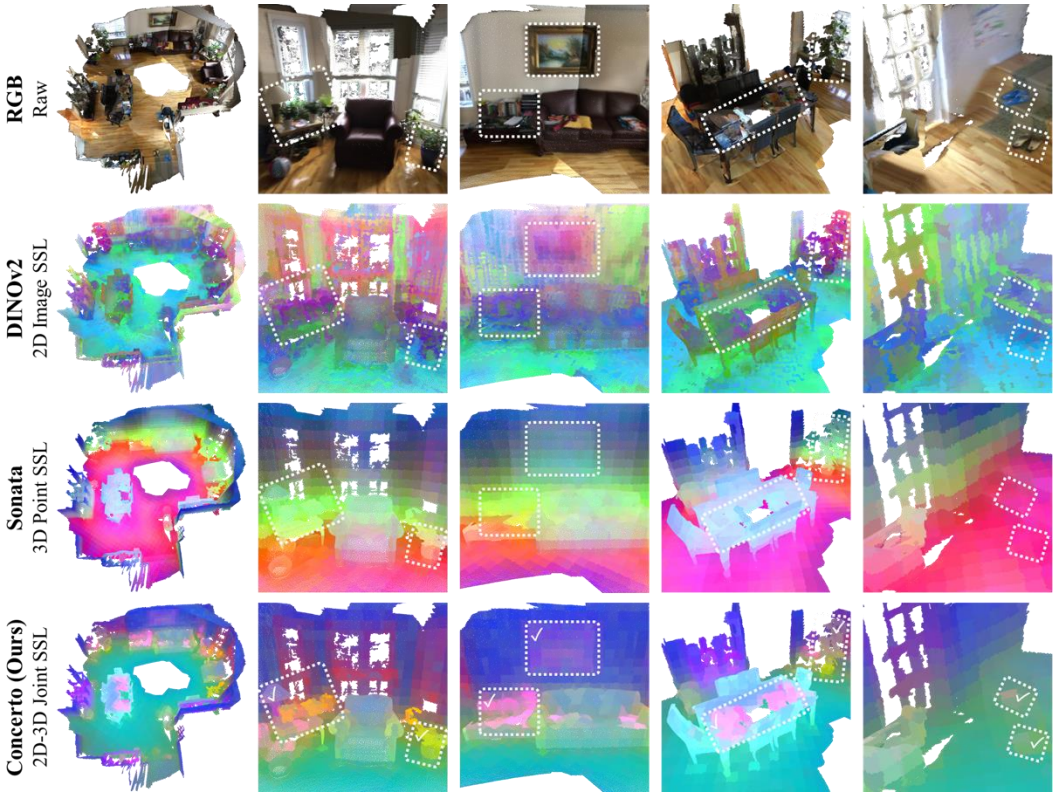


香 港 大 學
THE UNIVERSITY OF HONG KONG

POINTCEPT
Point Cloud Perception Codebase

# Let's see what kind of representations we can acquire from these sparse point clouds.

## In our latest research (will be released in July)



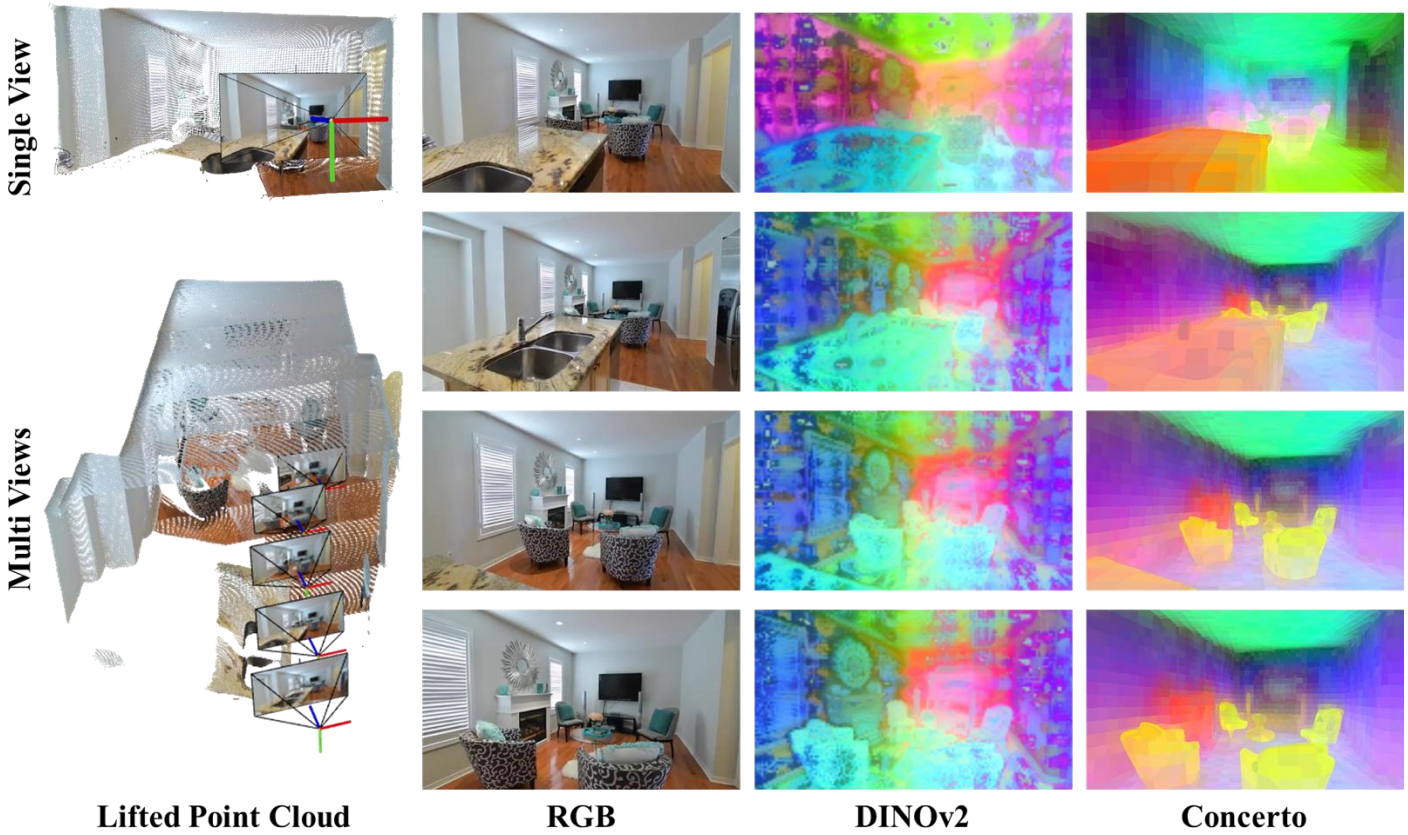

PCA Visualization

| Semantic Segmentation | | | ScanNet Val [15] | | |
|---|---|---|---|---|---|
| Method | Type | Encoder | mIoU | mAcc | allAcc |
| DINOv2 [26] | 2D Image SSL | ViT-G | 63.09 | 75.50 | 82.42 |
| Sonata [47] | 3D Point SSL | PTv3-B | 72.52 | 83.11 | 89.74 |
| Sonata×DINOv2 | 3D SSL×2D SSL | Both | 75.91 | 85.36 | 91.25 |
| **Concerto** (ours) | 2D-3D Joint SSL | PTv3-B | **77.32** | **86.58** | **91.74** |

Linear Probing

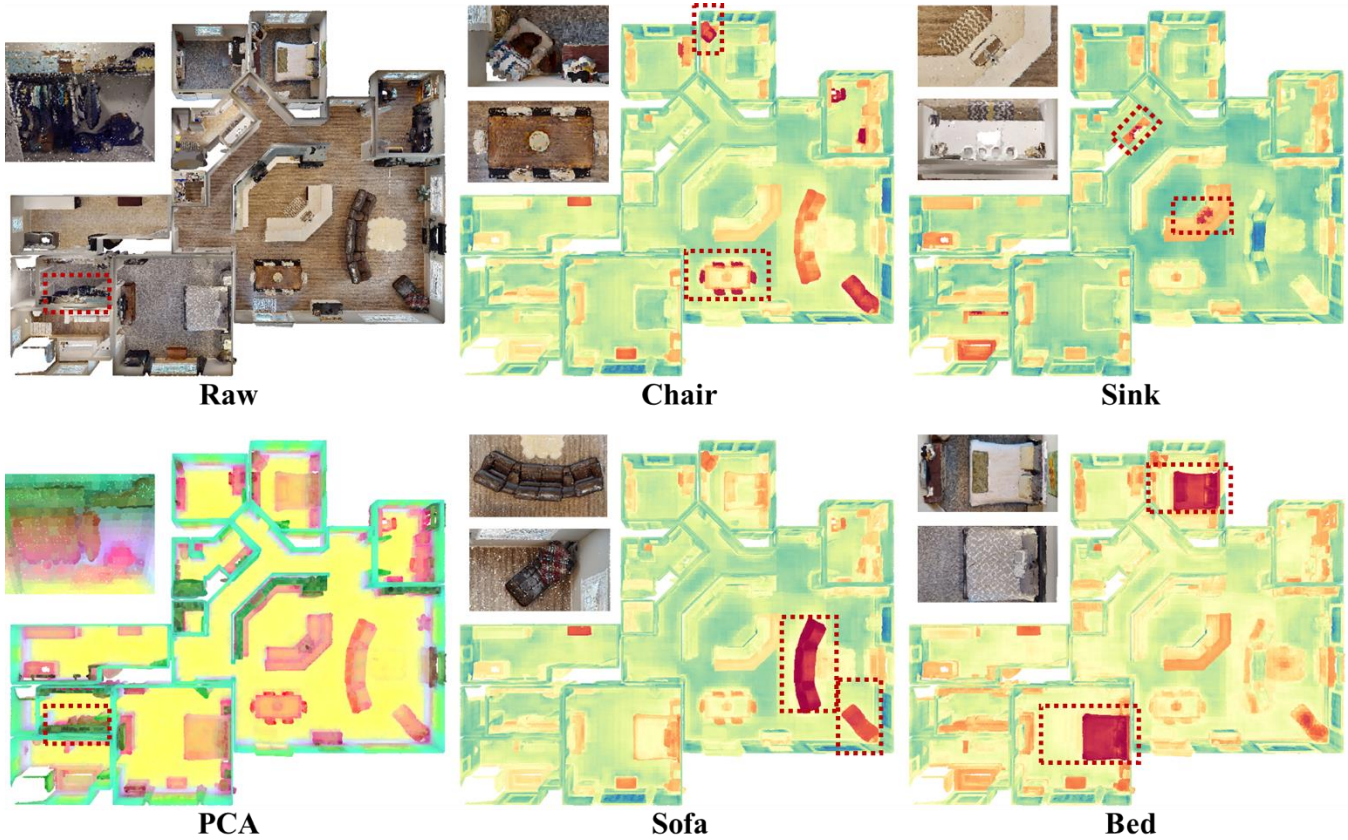
# Lifting and encoding images from scene videos

In our latest research (will be released in July)



# Probing representations into language space

## In our latest research (will be released in July)



# Excited about these results? Let's dive into the technical details!

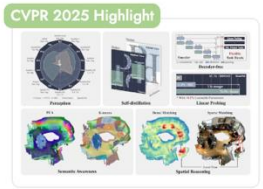
**Before 2023**

Point Transformer V1&V2, Masked Scene Contrast, Point Prompt Training, ...

**2023**


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
*Computer Vision and Pattern Recognition (**CVPR**) 2024 - **Oral** (0.78% acceptance rate)*
*Ranked **1st** Place on 2024 Waymo Open Dataset Challenge in Semantic Segmentation*
[Paper] [Code] [Bib] [Cert.] - Stars 1.2k

**2024**




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
*Computer Vision and Pattern Recognition (**CVPR**) 2025 - **Highlight** (3.0% acceptance rate)*
[Page] [Paper] [Code] [Bib] - Stars 340

**2025**

???

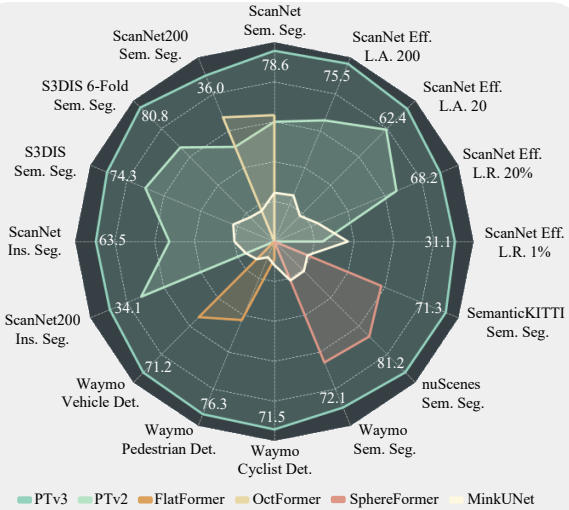
**Concerto: Joint 2D-3D Self-Supervised Learning Emerges Spatial Representations**

# Point Transformer V3: Simpler, Faster, Stronger

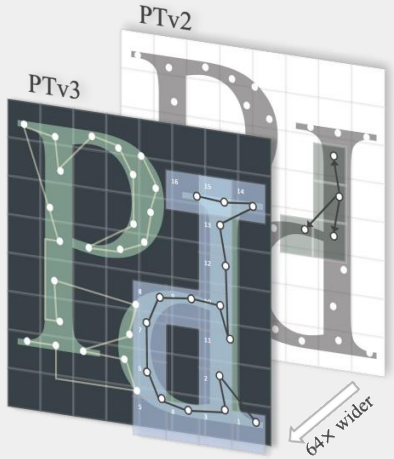
⇒ State-of-the-art performance on over **20** downstream tasks that span both **indoor** and **outdoor** scenarios.

⇒ Expanding the receptive field from **16** to **1024** points while remaining efficient.

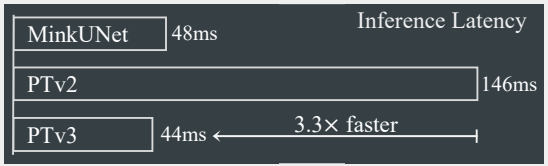
⇒ **3x** increase in processing speed and **10x** improvement in memory efficiency compared with PTv2



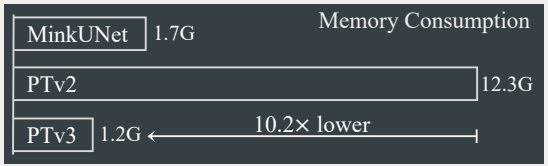

**Stronger Performance**




**Wider Receptive Field**




**Faster Speed**




**Lower Memory Consumption**

## We were suffering from an extremely small scale of data!

Approx. 100,000 points

≈

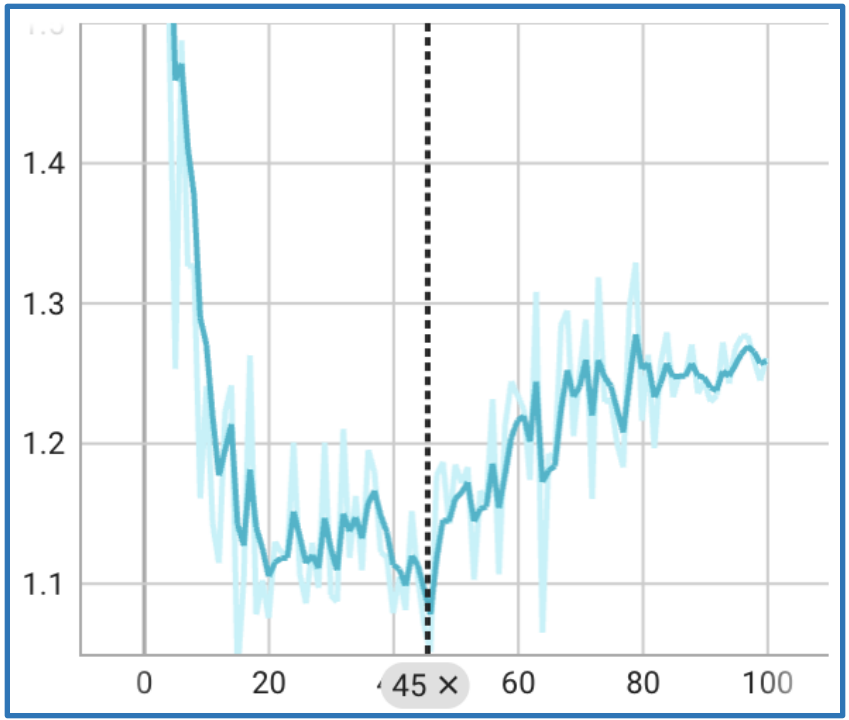
300 pixel

300 pixel

⇒ The **largest (2022)** indoor perception dataset, ScanNet v2, **only contains 1,613** scene-level point clouds.

⇒ **The largest dataset only worths 1,613 images with 300×300 resolution?**

# Backbone designs are also overfitting to complex designs!




Validation loss

⇒ Validation loss increase during training

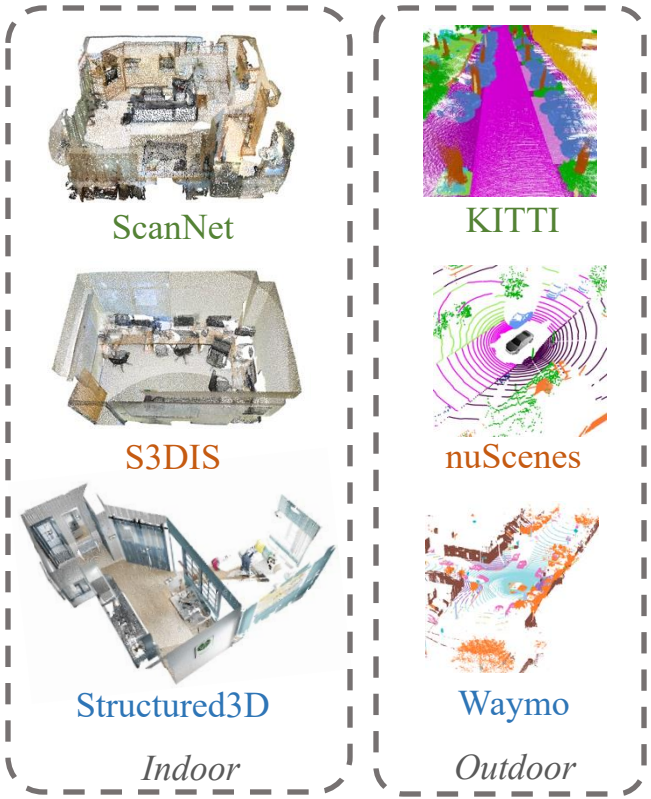
*(observed in PTv2, the issue is almost addressed after PPT and PTv3)*

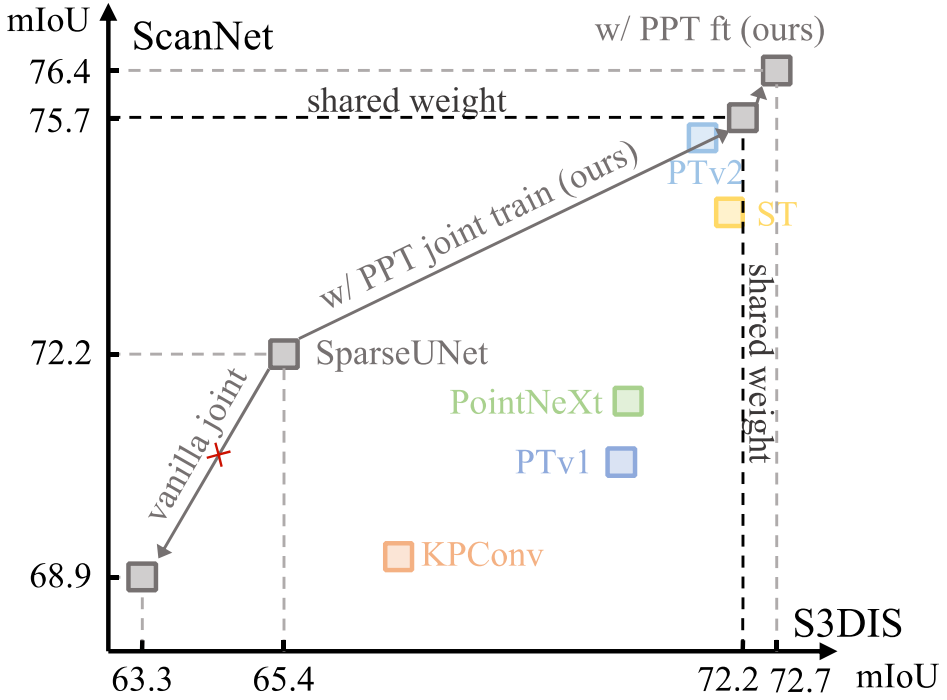
⇒ Why can not unify parameters for different tasks

⇒ Why backbones overfitting to complex designs

*Point Prompt Training, by **Xiaoyang Wu, et al.**; accepted by **CVPR 2024***

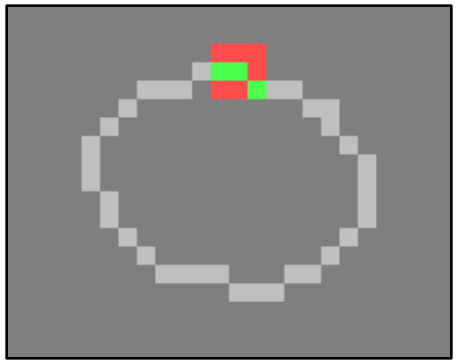
# Enlarging data scale and training scale with multi-dataset Point Prompt Training



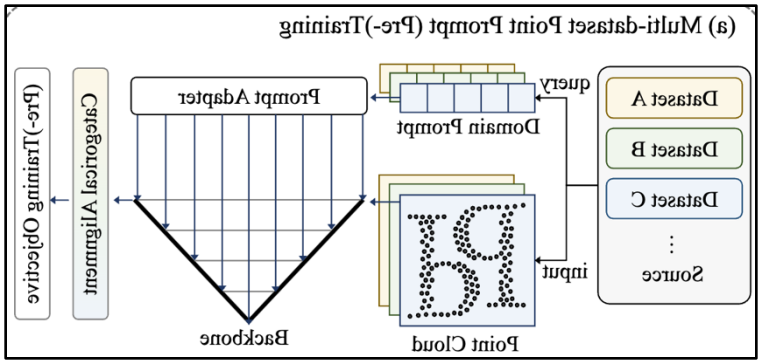

*Point Prompt Training, by **Xiaoyang Wu, et al.**; accepted by **CVPR 2024***

# PTv3: Scaling Principle for Backbone Design



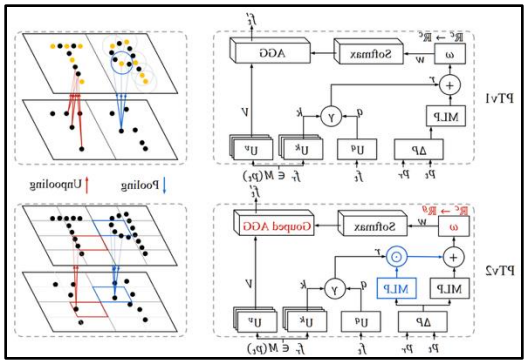
SparseUNet (2019)

+




PPT (Large-scale Training) (2023)

>




PTv2 (2022)

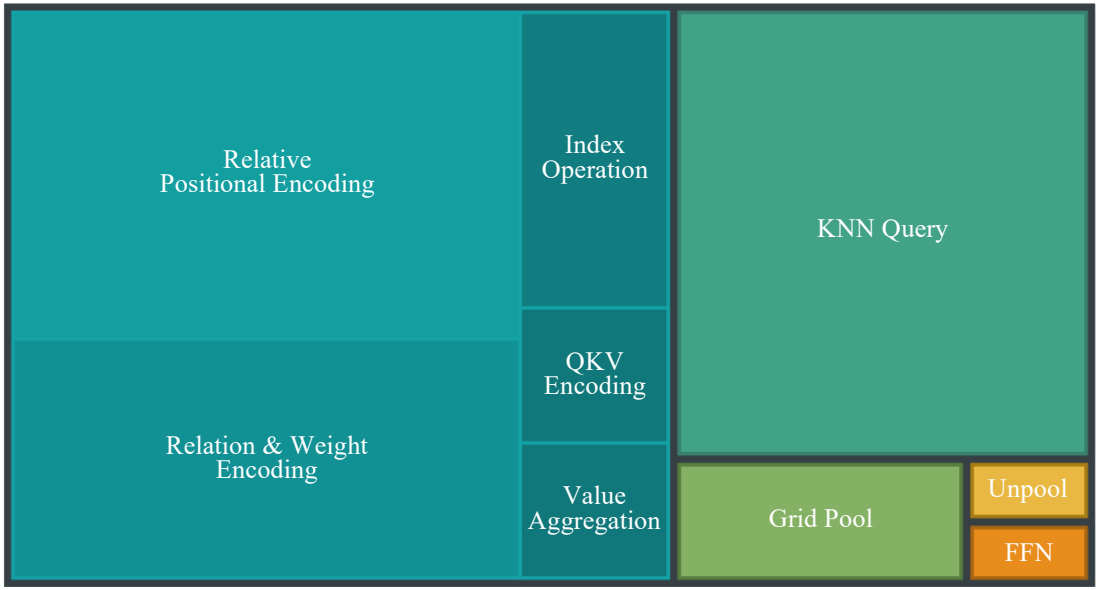
## Scaling Principle

⇒ Model performance is **more significantly influenced by scale** than by complex design;

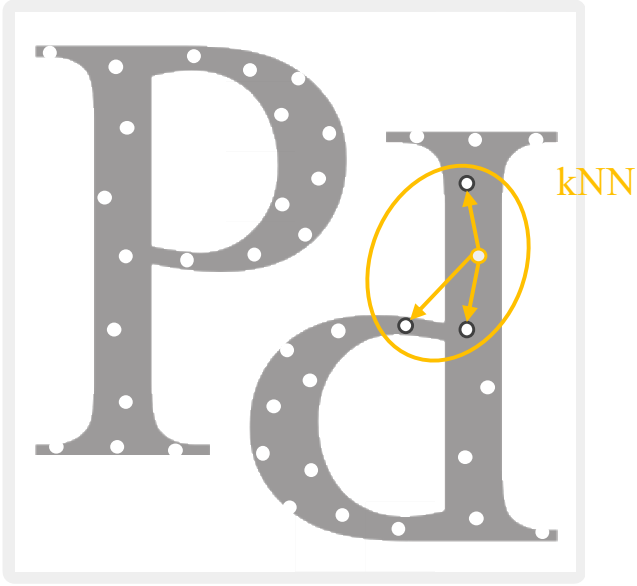
⇒ We should **prioritize simplicity and efficiency over the accuracy** of certain mechanisms;

⇒ **Efficiency enable scalability** and further enable stronger accuracy.

# Breaking the curse of permutation invariance




PTv2 Forward Latency

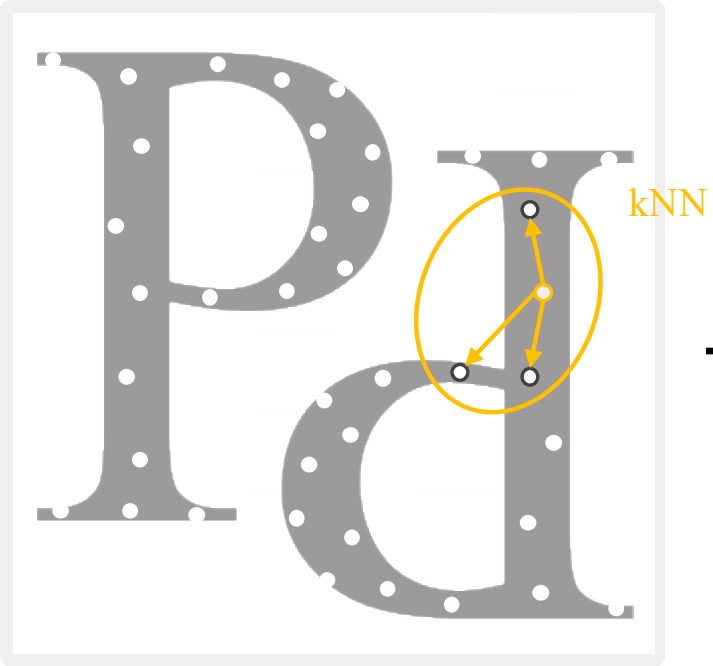



PTv2

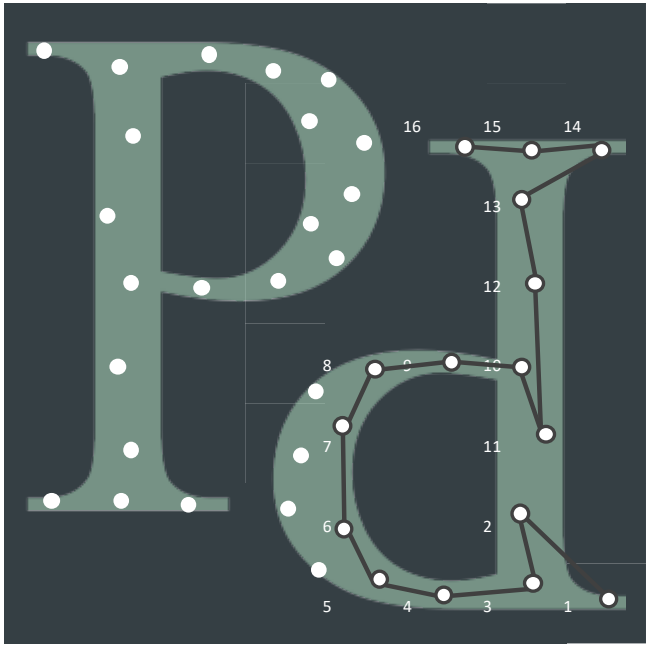
⇒ The unstructured nature of point cloud makes point-based methods **rely on K-nearest neighbor**.

⇒ However, K-nearest neighbor take up to **28%** forward latency!

# Breaking the curse of permutation invariance




Point Transformer V2




Point Transformer V3

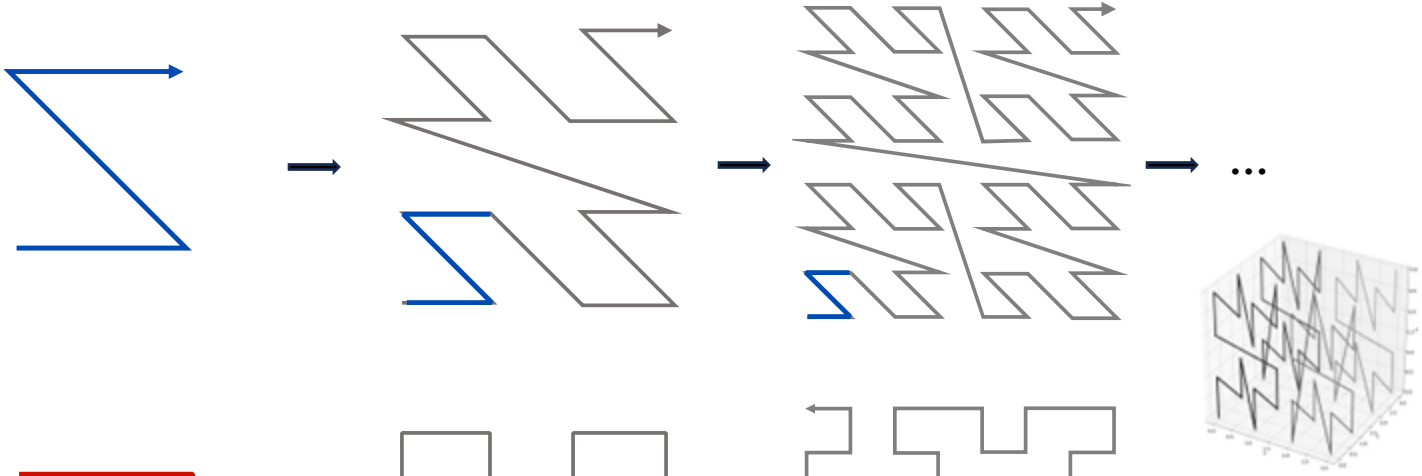
⇒ **Do we really need the accurate neighbors**?

⇒ **No**, attention is adaptive to kernel shape, large kernel >> precise neighbors;

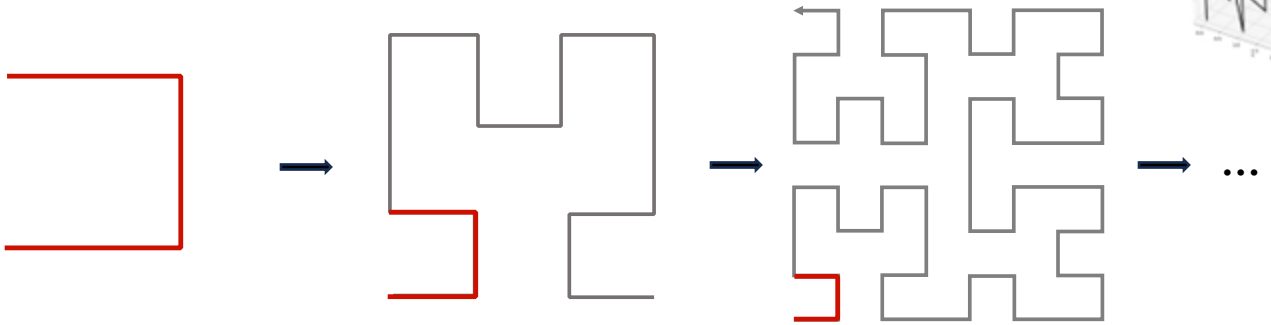
⇒ **Serialized point cloud into a structured format.**

# Space-filling Curve

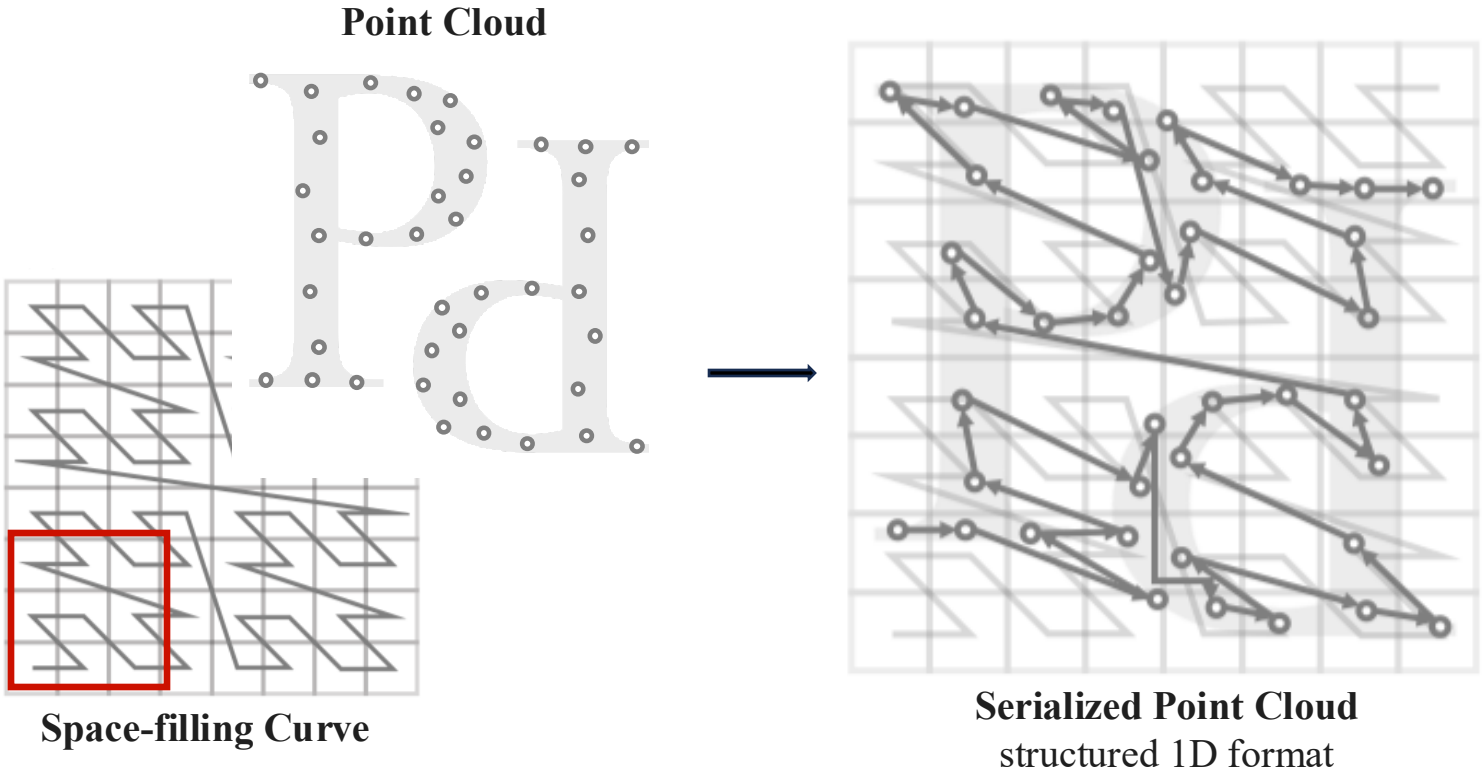
Z-order Curve



Hilbert Curve

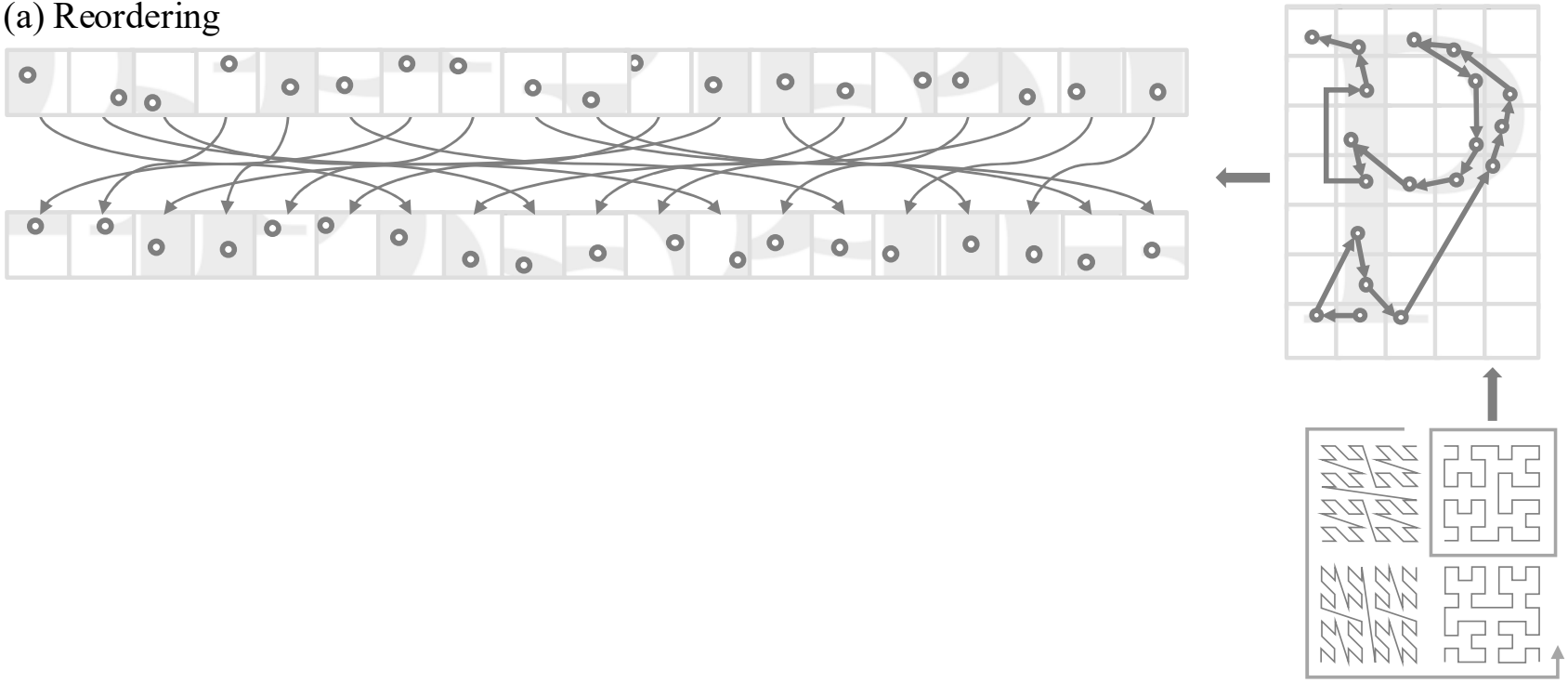


# Point Cloud Serialization

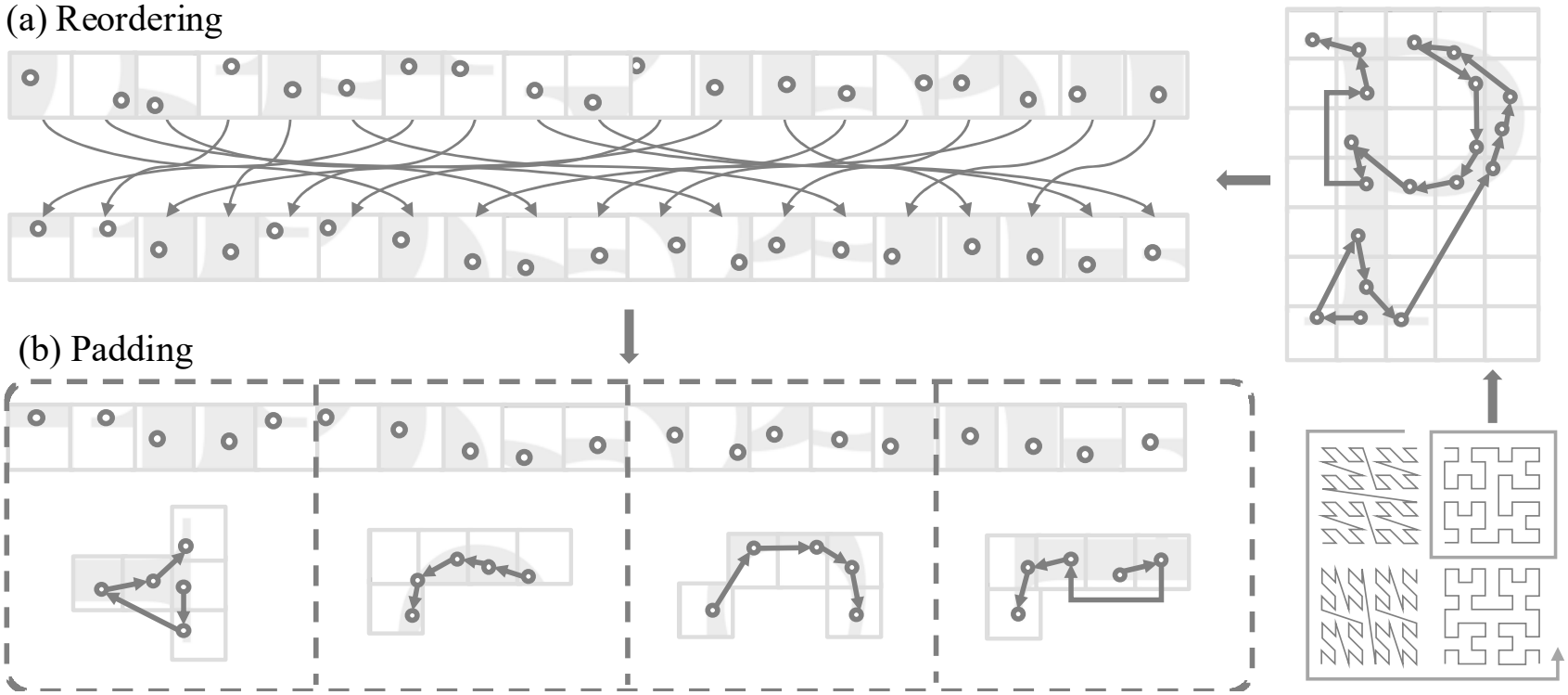


# Serialized Attention

(a) Reordering



# Serialized Attention



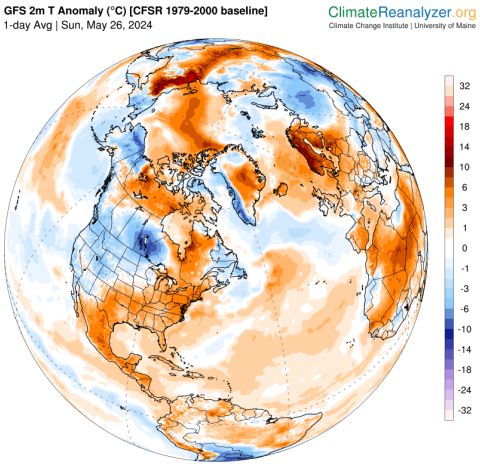
# The potential for multi-modal feature extraction

⇒ Point cloud high-dimensional sparse data.

⇒ **Many** high-dimensional data with positional information **can be view as** a point cloud.

⇒ Like image, climate and of course point cloud

Image (2D)




Climate (2.5D)

Point cloud (3D)

# What's the DINO time of 3D?

It leverages the **magic of data and learning**;

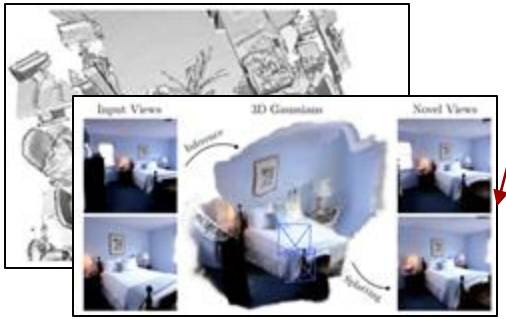
It (target at) **unify and benefit all 3D tasks**;

**It is Project Sonata**

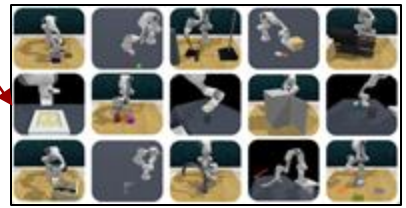
3D Localization

3D Perception

Sonata

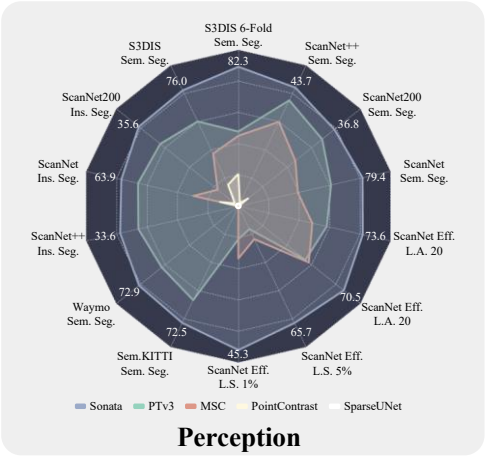



3D Reconstruction

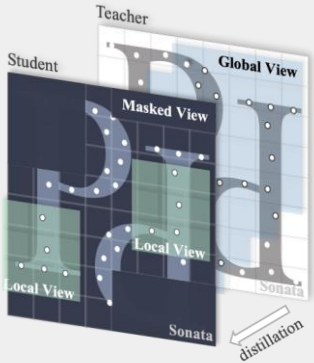


3D Planning & Interaction

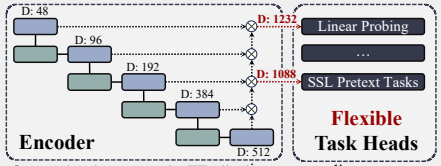
# Sonata: Self-Supervised Learning of Reliable Point Representations




**Perception**



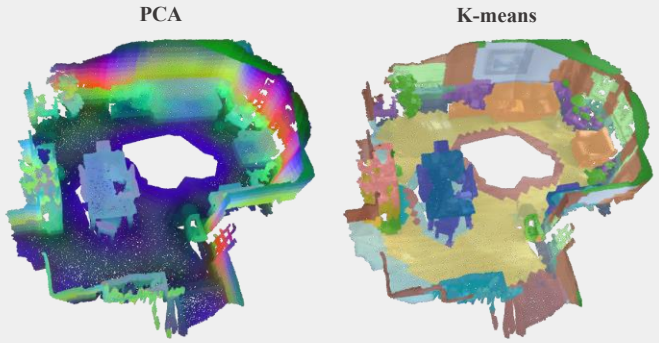

**Self-distillation**



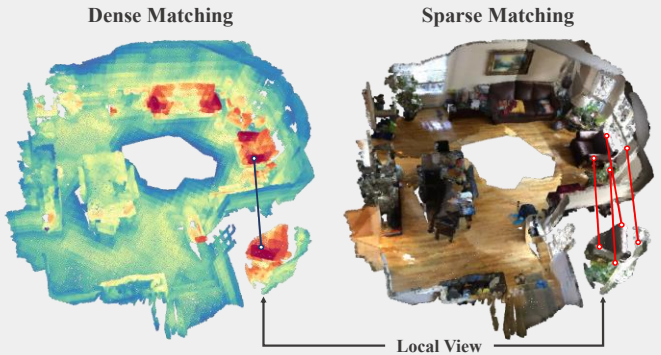

**Decoder-free**


* With <0.2% Learnable Parameters

**Linear Probing**




**Semantic Awareness**




**Spatial Reasoning**



香 港 大 學
THE UNIVERSITY OF HONG KONG

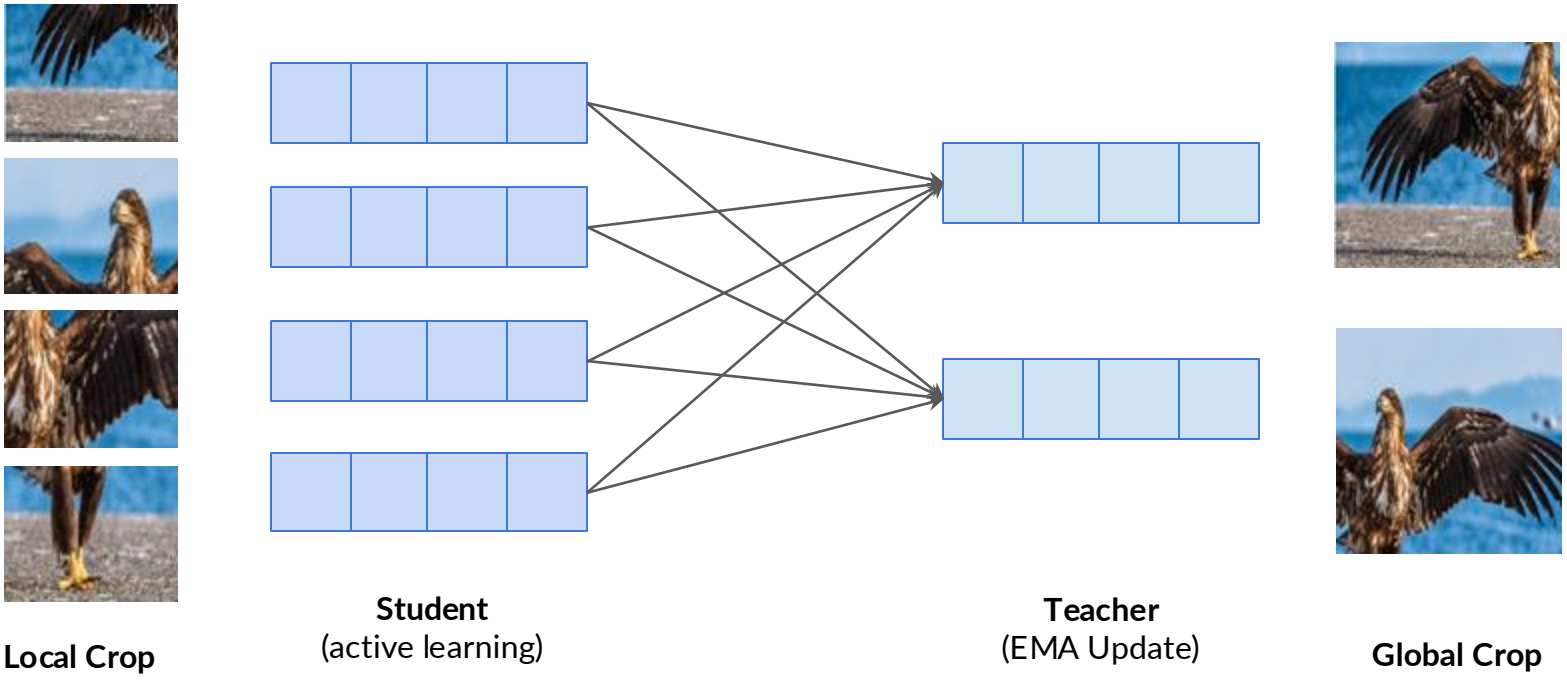
Meta

# What is Self-supervised Learning ?

=> Make things should be same, the same.

=> “No plain, no gain”, the more the model struggles, the better the representations it learns.

=> Just like a child, the model is good at finding shortcuts.

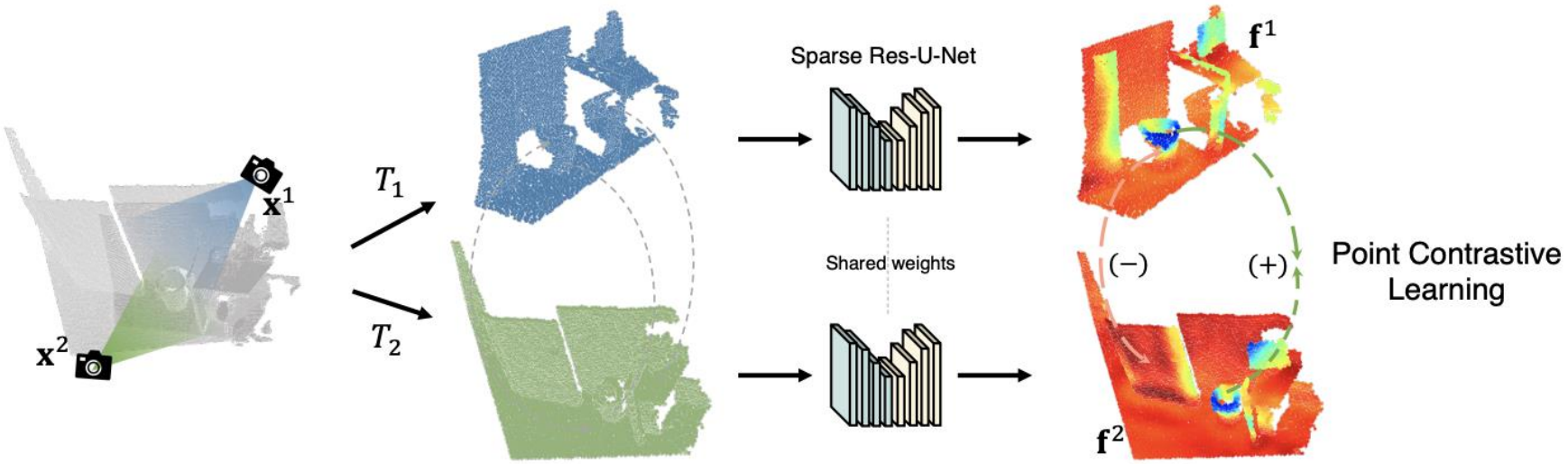



**Contrastive Learning as an example**

## What is Self-supervised Learning ?

=> Make things should be same, the same.

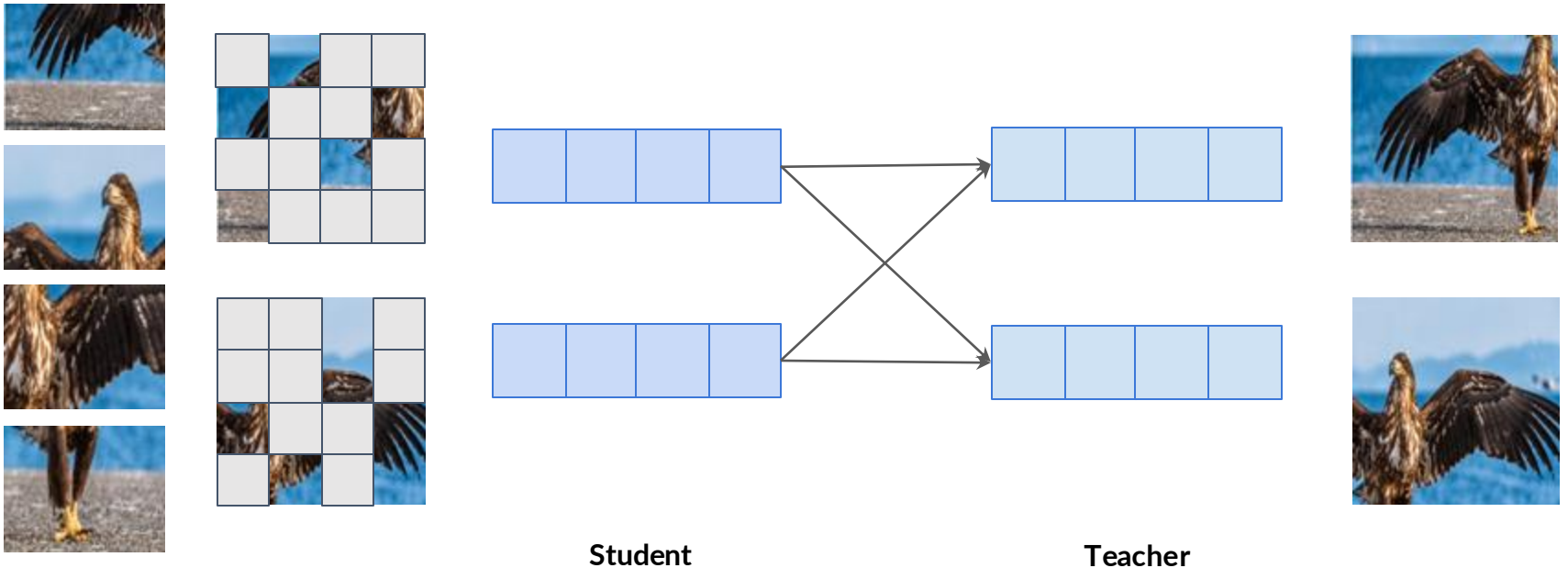
=> Just like a child, the model is good at finding shortcuts.




**Contrastive Learning as an example**

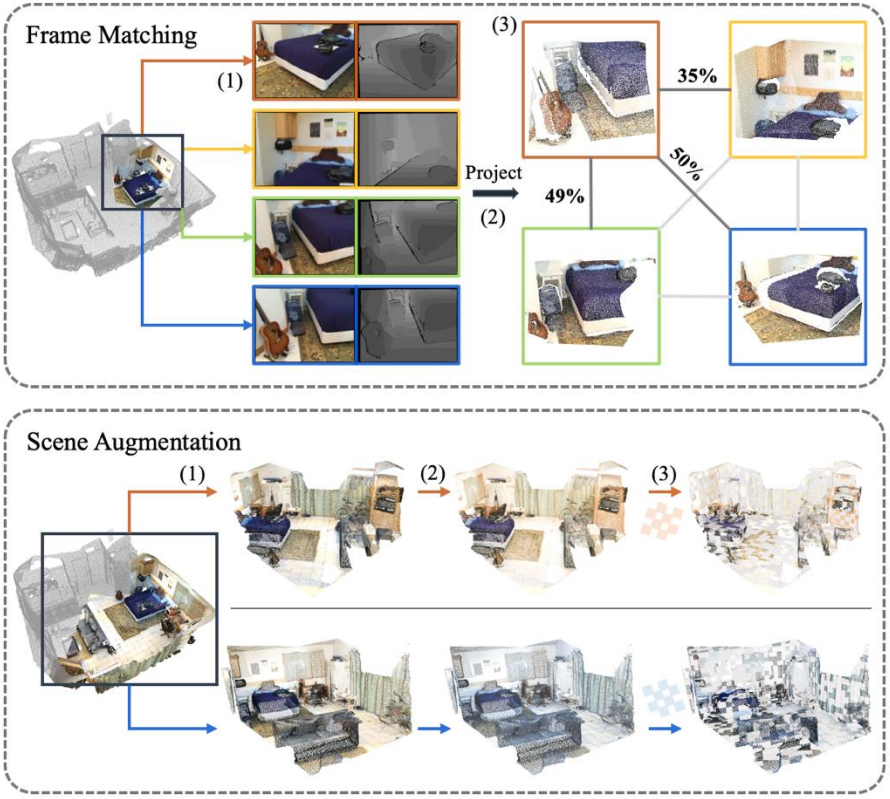
*PointContrast, by **Saining Xie, et al.***

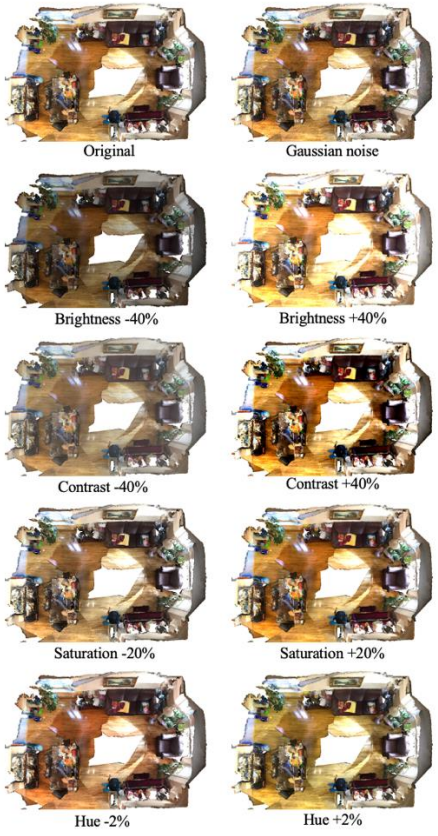
# Why we introduce strong augmentations and mask image/point modeling?




**Core Principle:** **continuously increasing the difficulty of pretext tasks as long as the model continues to converge**
**Local-Global Alignment:** The **"base"** in the **"cocktail"** of SSL, **setting** the foundation and **direction** for convergence.
**Masked-Unmasked Alignment:** The **"liqueur"** that **intensifies the challenge** of SSL, adding complexity.

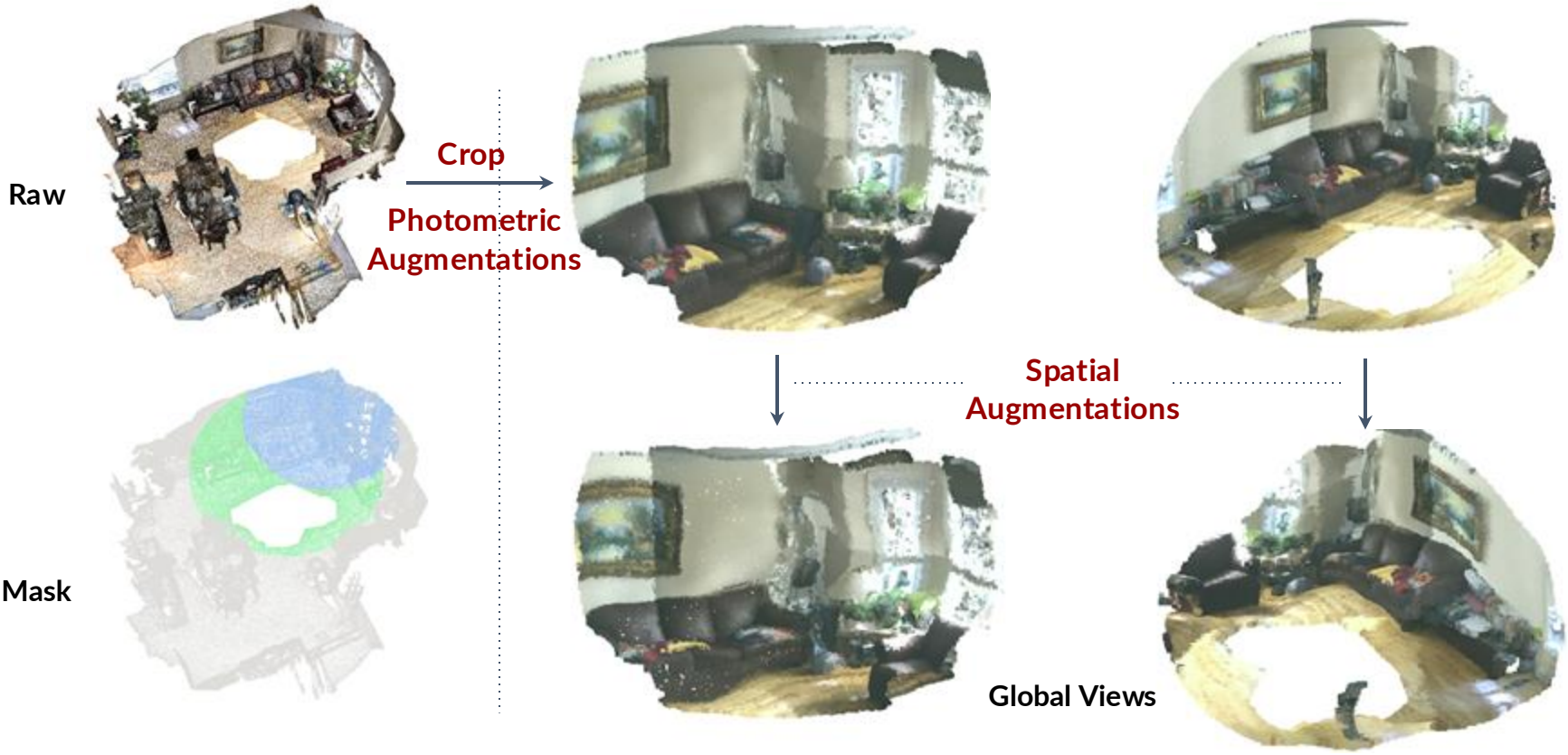
# Why we introduce strong augmentations and mask image/point modeling?



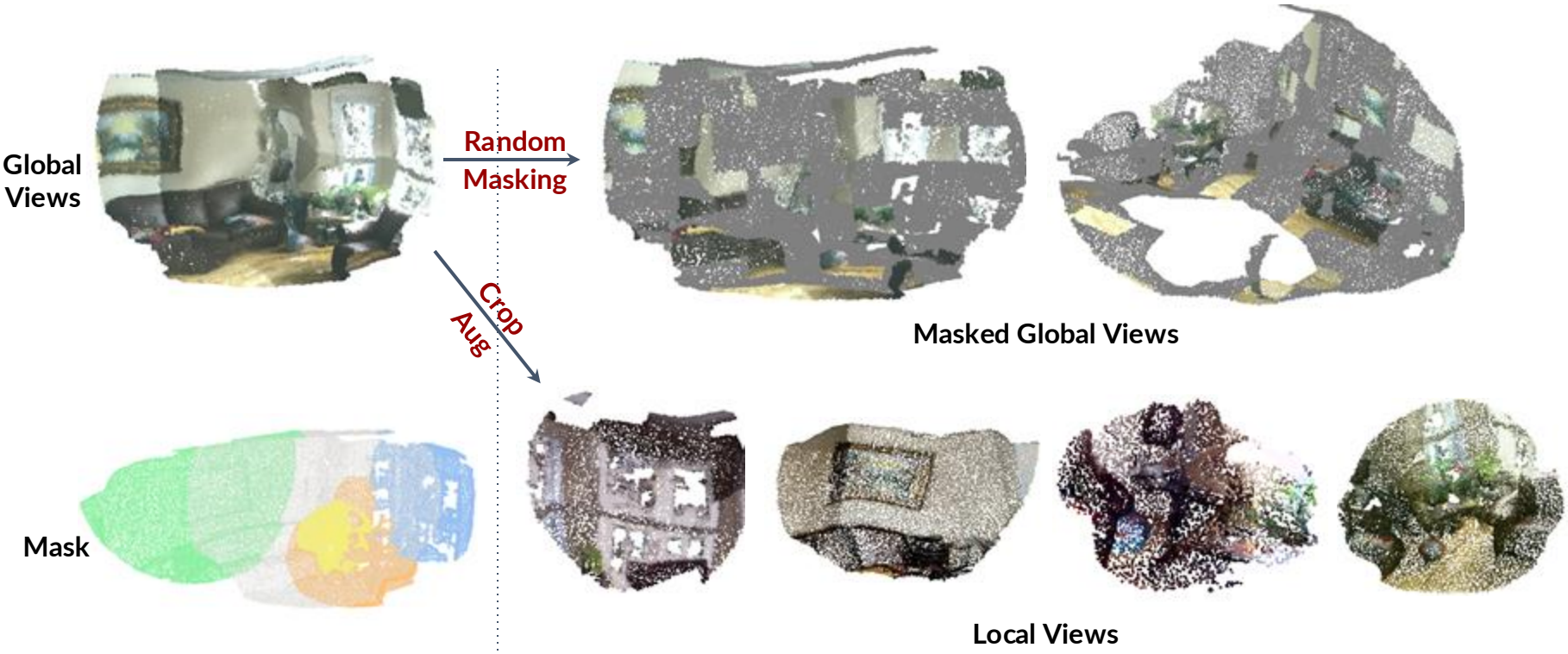




*Masked Scene Contrast, by **Xiaoyang Wu, et al.***

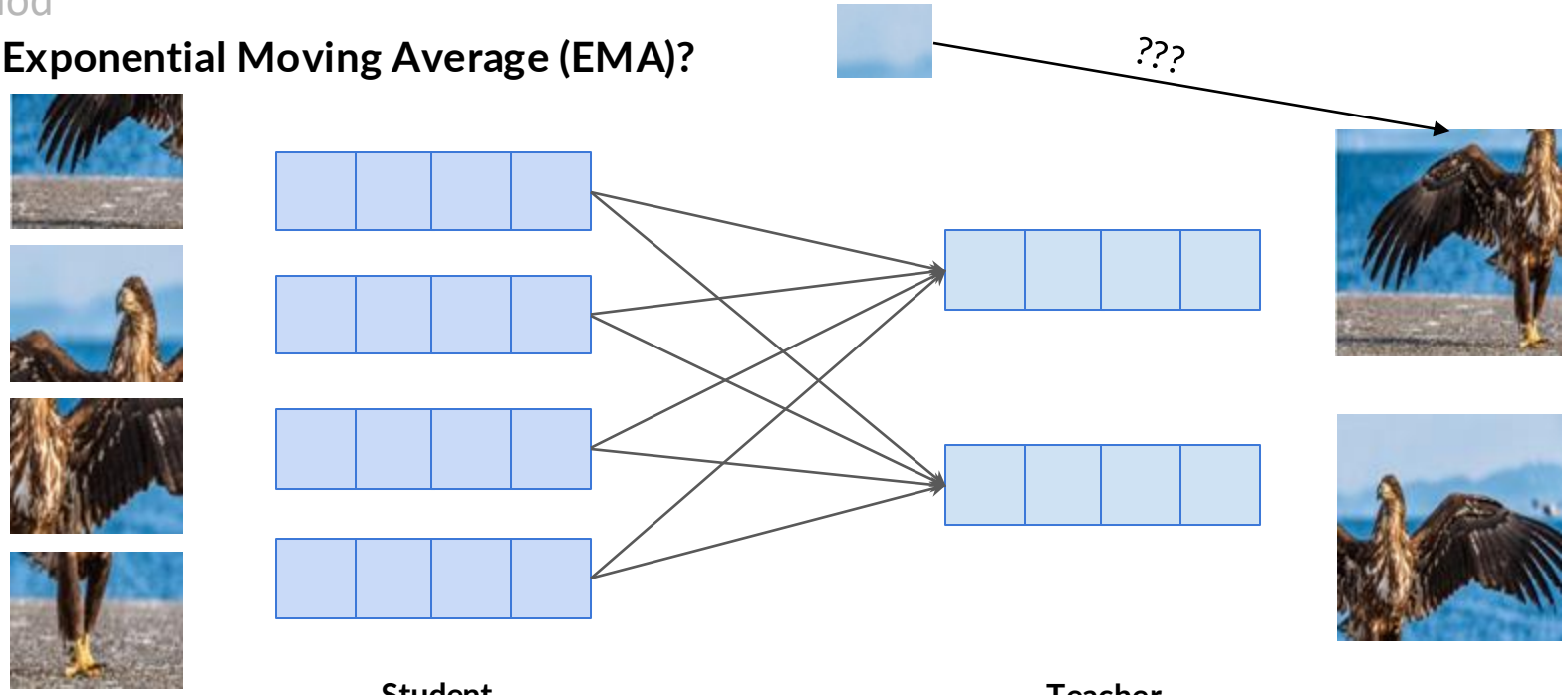
# Why we introduce strong augmentations and mask image/point modeling?



# Why we introduce strong augmentations and mask image/point modeling?



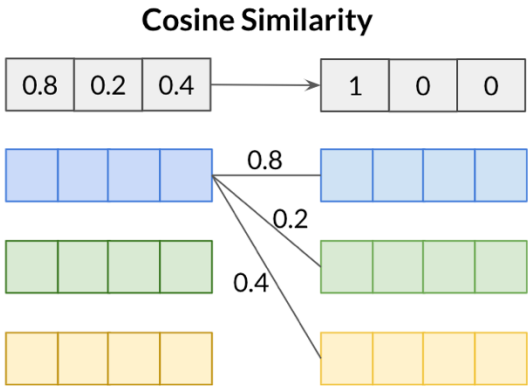
## Why Exponential Moving Average (EMA)?

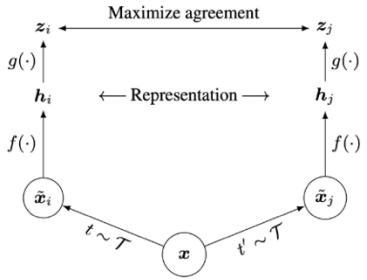


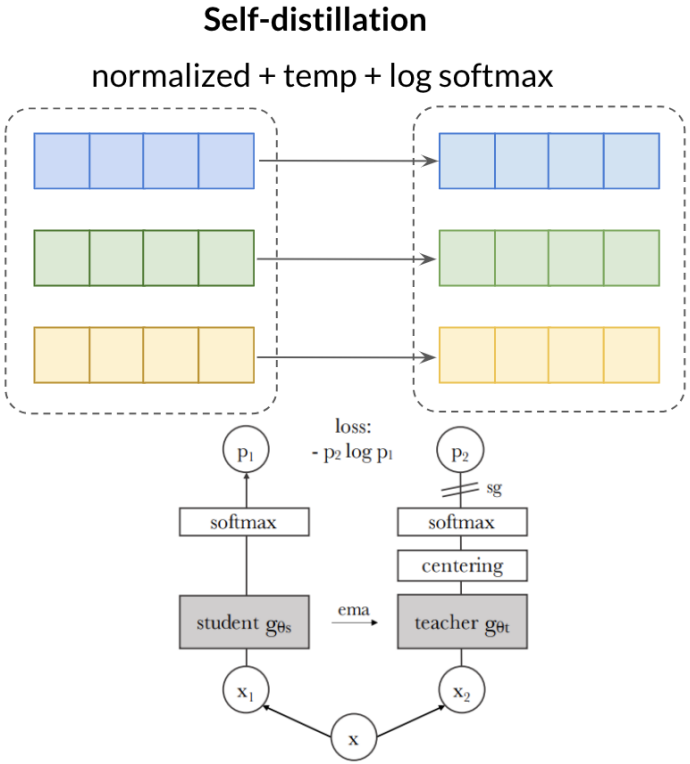

**Teacher (backup):** a teacher is actually a kind of "backup" and "smooth average" of a student;
**Student (spearhead):** the student is more like a "spearhead", tries something challenging and risky;

With the teacher preventing the student from being misled, the student is less likely to get lost in **"mission impossible"** and has a greater chance to discover treasure within "impossible"

# Why Self-distillation?



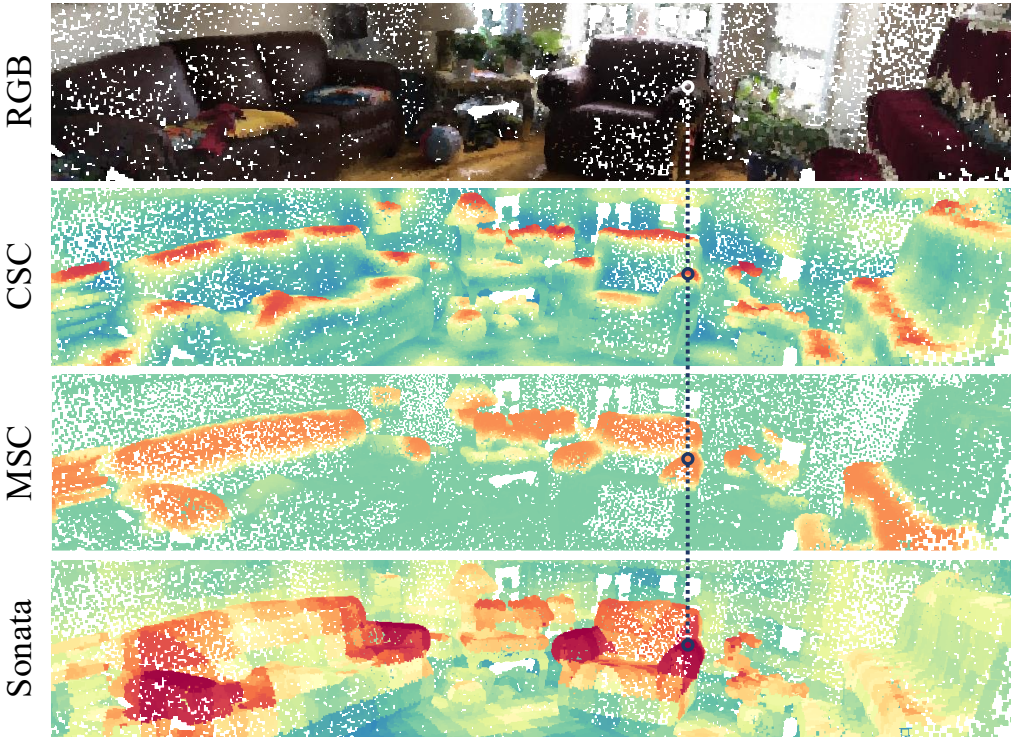



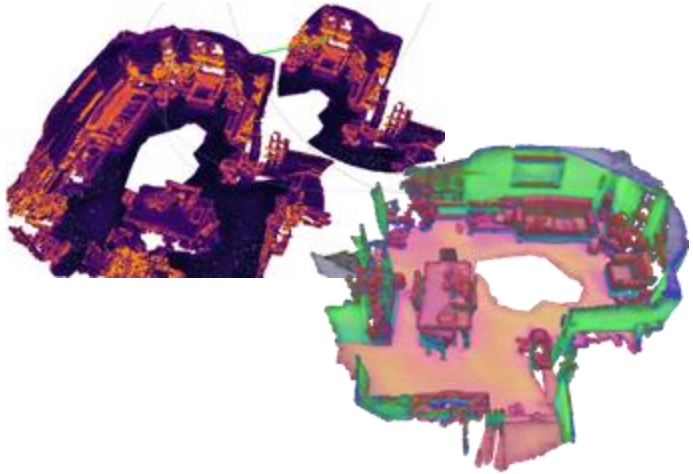

Now we lead to the DINOv2 Architecture…

# What is Point Cloud Self-supervised Learning?

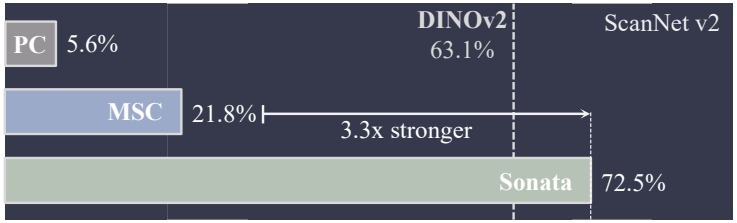
Is it just adapting Image SSL to point cloud data?

# What is the Geometric Shortcut



Sonata Prototype

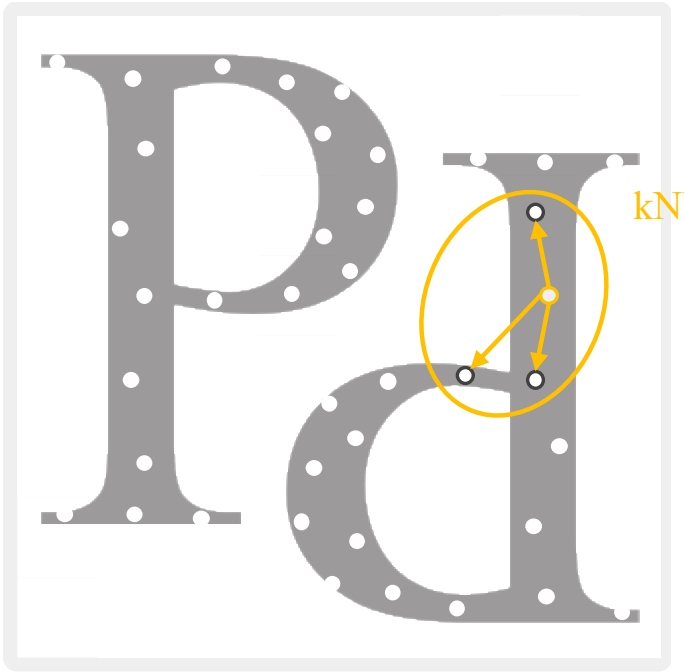



* With **0.02%** Learnable Parameters

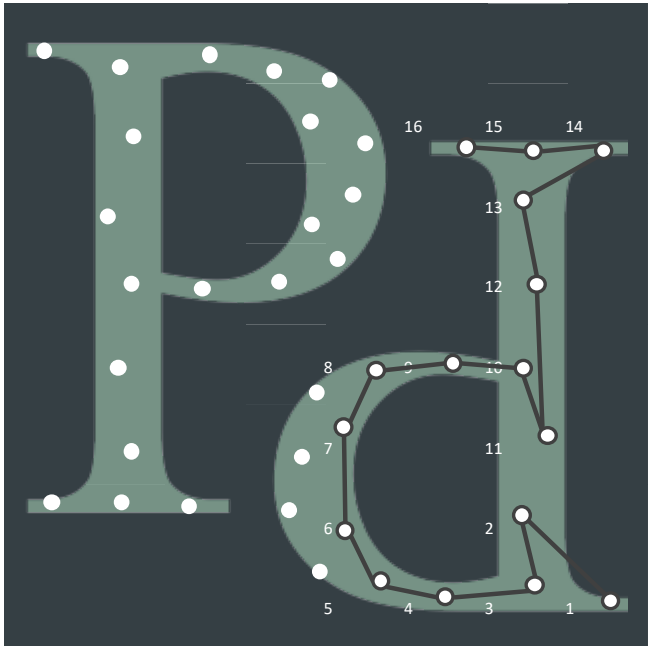
Linear Probing

Geometric shortcut **refers to** the tendency of the model to **collapse to easily accessible, low-level geometric cues**, such as normal direction or point height. Therefore, resulting low linear probing performance in downstream tasks

## How do geometric shortcuts occur




Point Transformer V2




Point Transformer V3

**Evidence**: The point cloud encoder can capture geometric information not only from input features but also from the operator kernel defined by coordinates.

# The Geometric Shortcut is a unique challenge to Point Cloud (Sparse Data)

**Why not happen with image SSL**

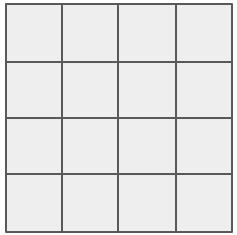


**Image A**

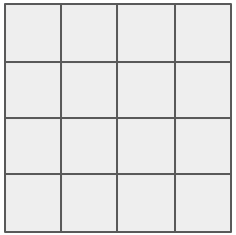


**Image A'**

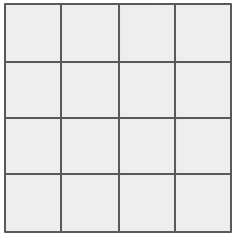


**Image B**

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

shuffle shuffle

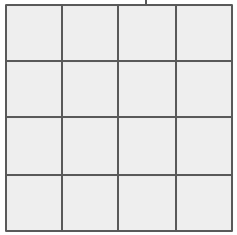


**Image ???**

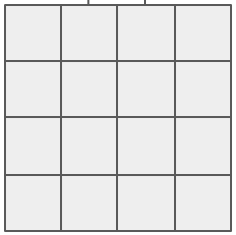


**Image ???**

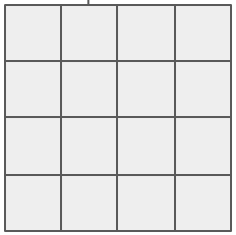


**Image ???**

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

**Image ???**

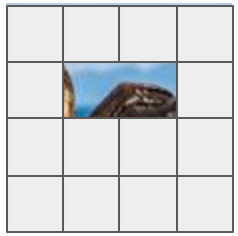


**Image ???**

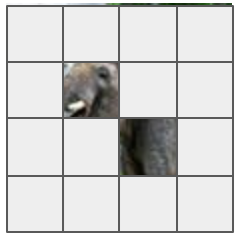


**Image ???**

> **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)**

# The Geometric Shortcut is a unique challenge to Point Cloud (Sparse Data)

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?

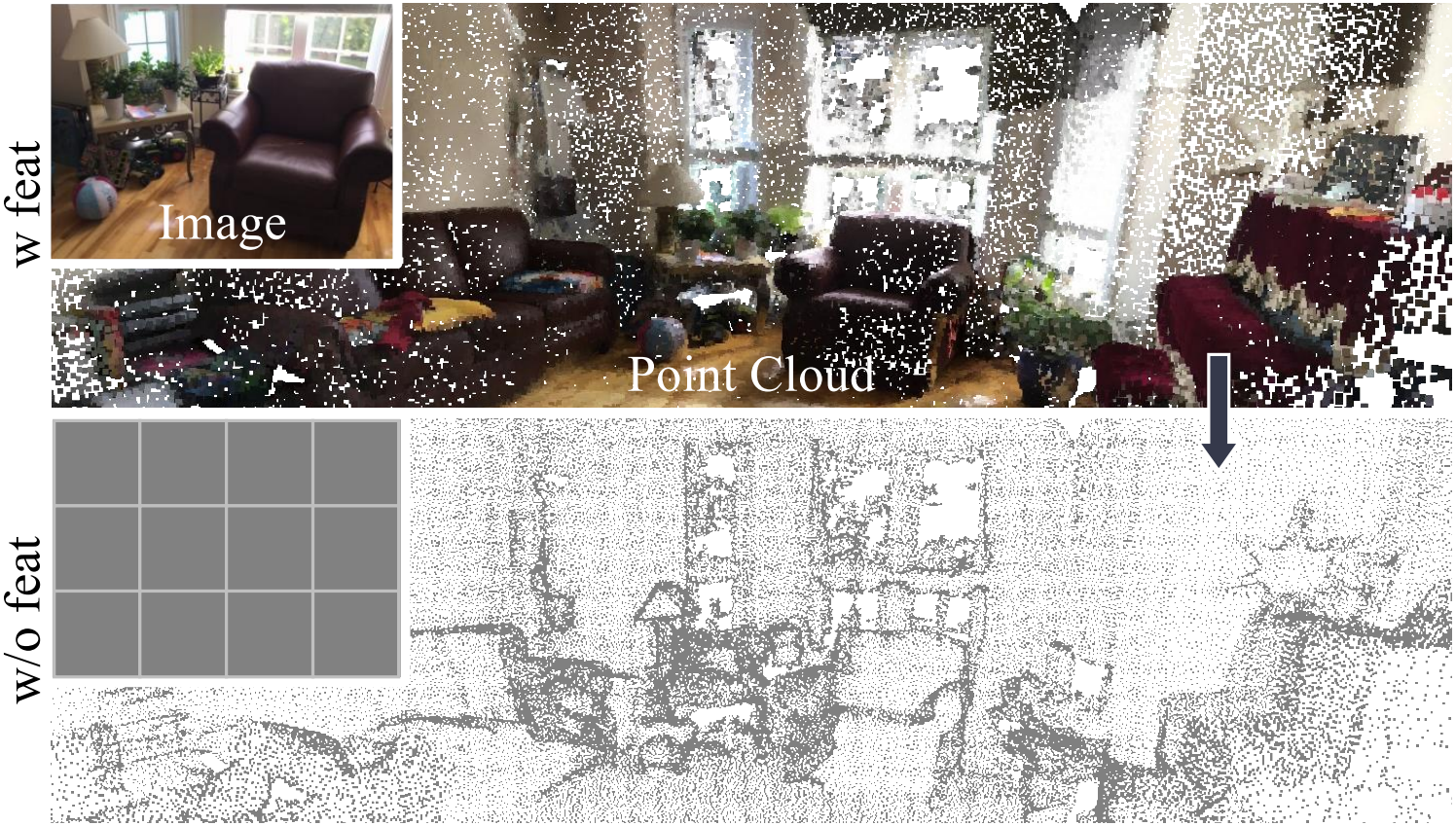
Point Cloud A?

> **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.**

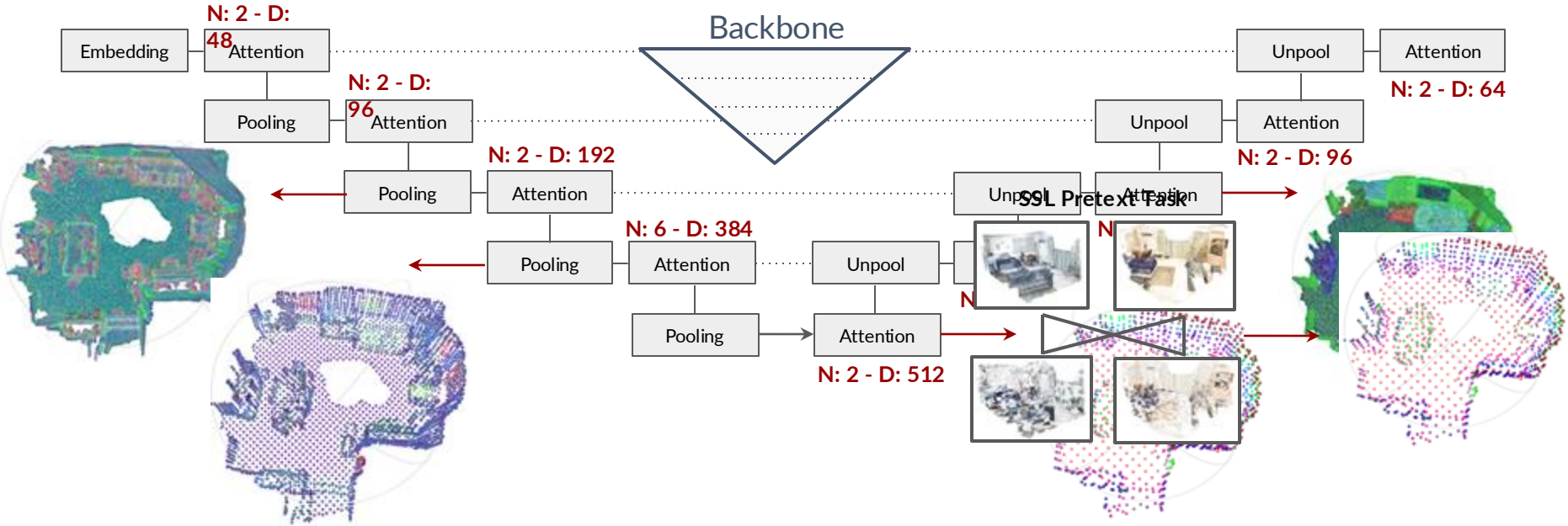
**>> Point cloud SSL exist a shortcut that only rely on geometry information**

# The Geometric Shortcut is a unique challenge to Point Cloud (Sparse Data)



# Key solutions to geometric shortcut

**Core concept:** Disturb on positional information to restraint collapse



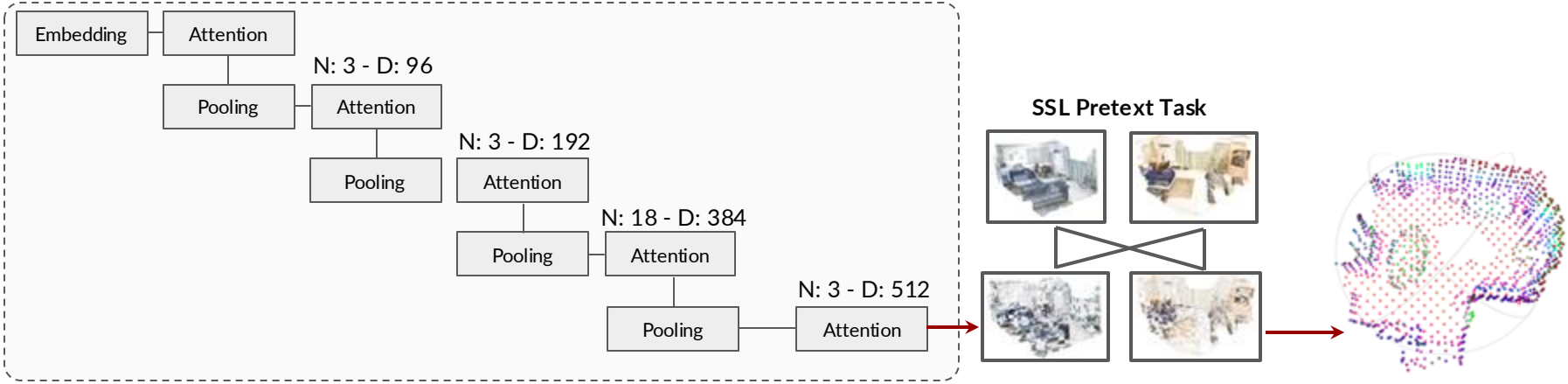



香 港 大 學
THE UNIVERSITY OF HONG KONG

Meta

# Key solutions to geometric shortcut

**Core concept:** Disturb on positional information to restraint collapse

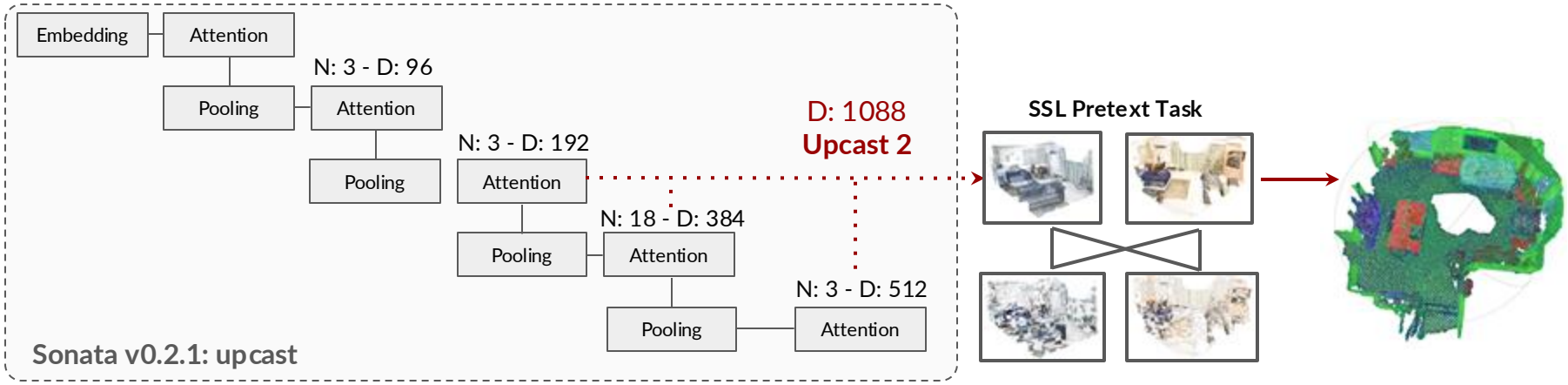



**Good:** Improve Linear probing from **20% to 60%** +

**Bad:** Here **only contains** global-wise embedding, also need some local information

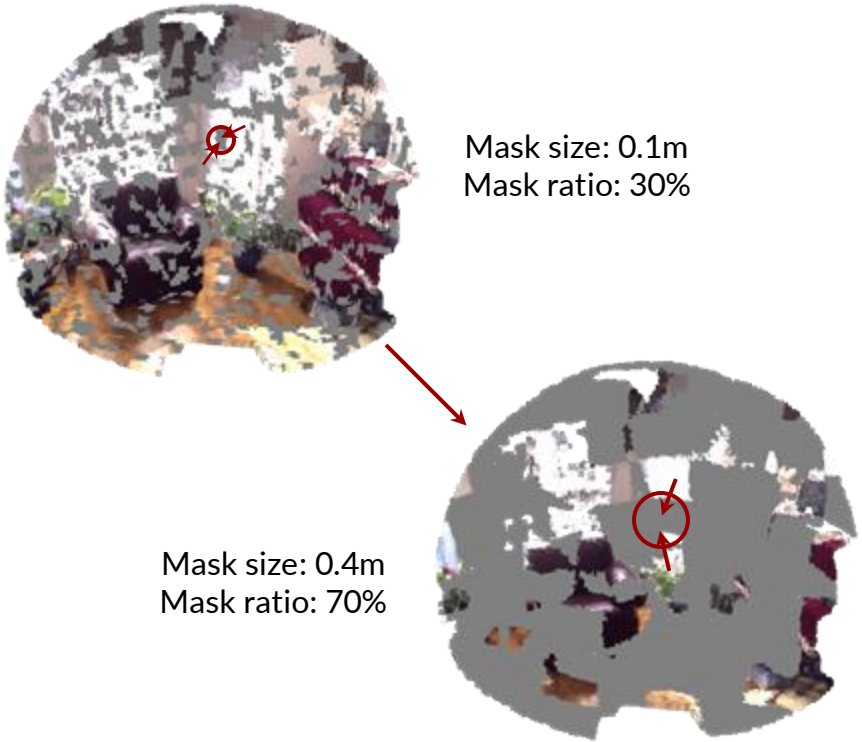
# Key solutions to geometric shortcut

**Core concept:** Disturb on positional information to restraint collapse

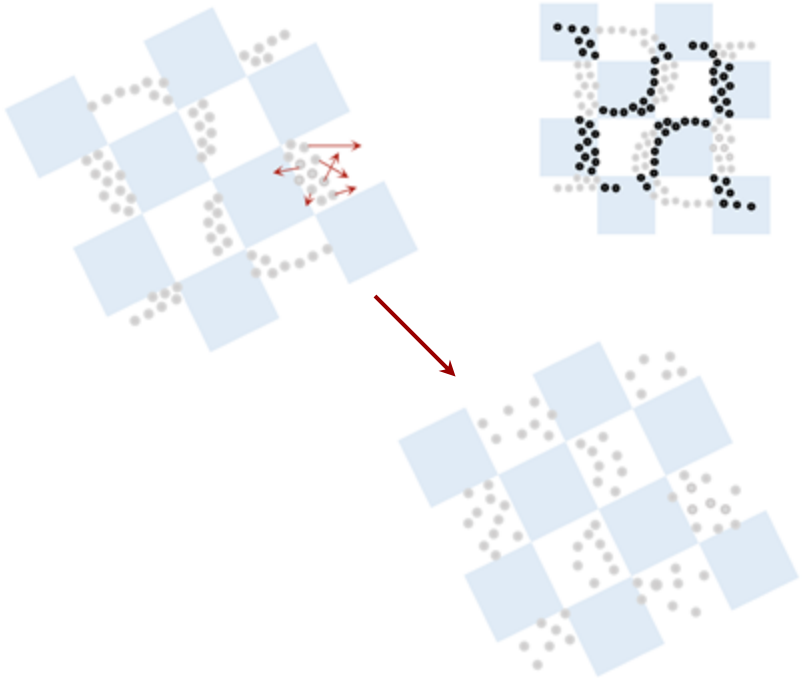


# Other solutions to geometric shortcut (Micro Designs)

**Mask ratio scheduler & Mask size scheduler:**
Set a trap to against geometric shortcut




**Random gaussian jitter with masked point:**
Break local geometric information

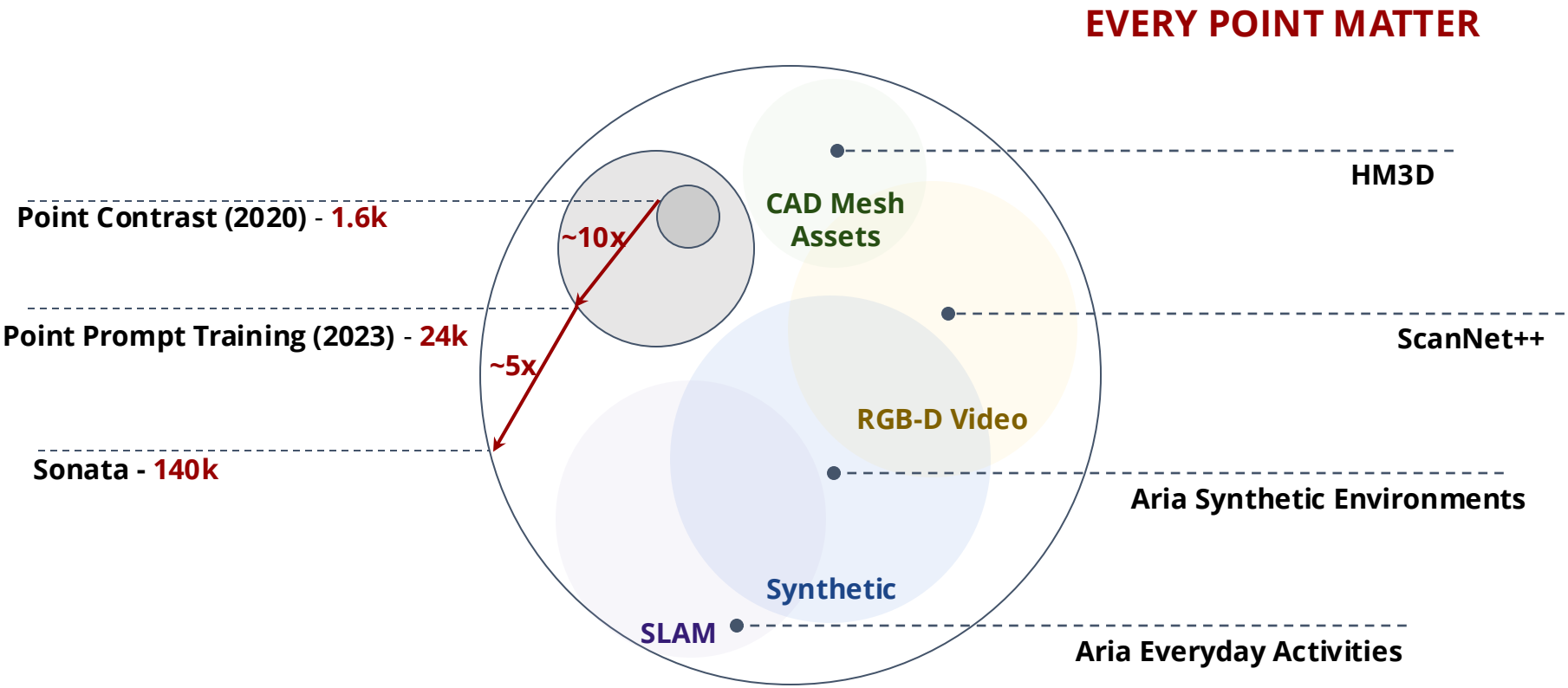




香 港 大 學
THE UNIVERSITY OF HONG KONG

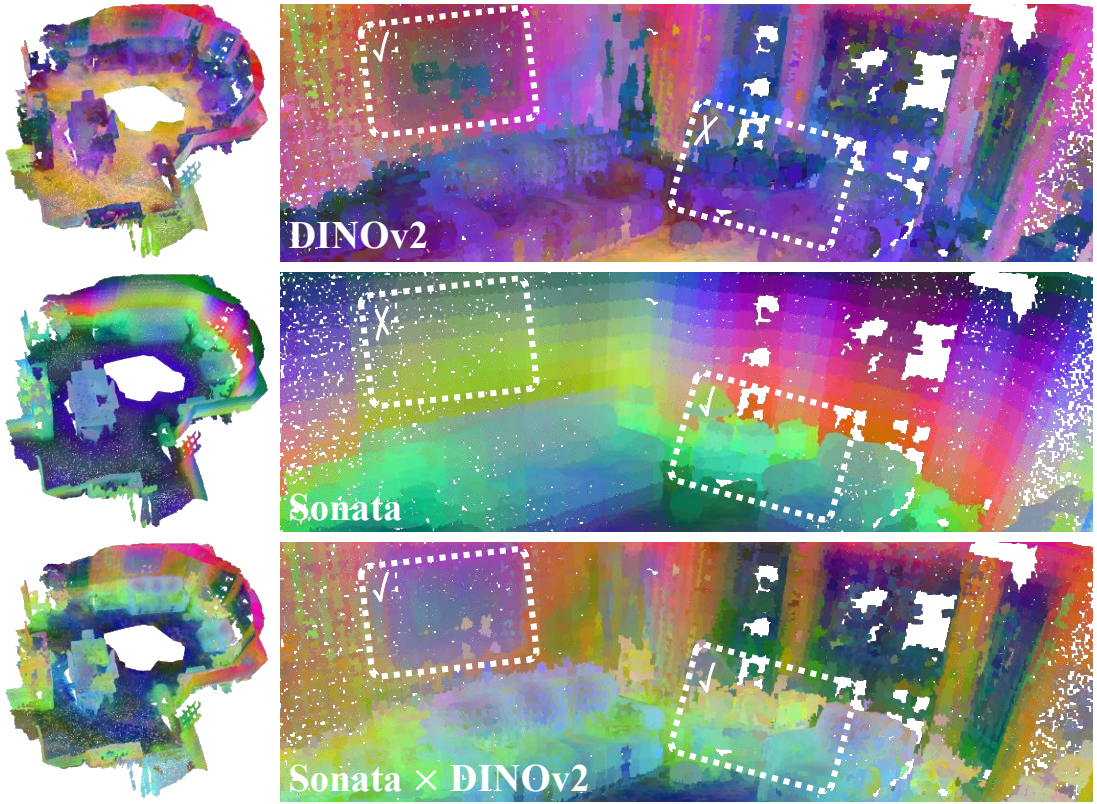
Meta

# Data Scaling

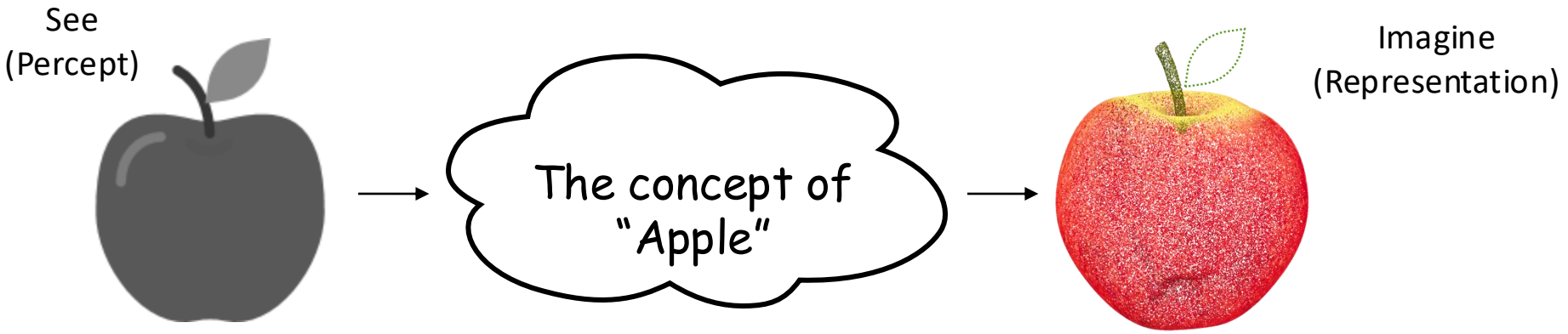


# Comparison with DINO in 3D Space

| Semantic Segmentation | | | ScanNet Val [15] | | |
|---|---|---|---|---|---|
| Method | Type | Encoder | mIoU | mAcc | allAcc |
| DINOv2 [26] | 2D Image SSL | ViT-G | 63.09 | 75.50 | 82.42 |
| Sonata [47] | 3D Point SSL | PTv3-B | 72.52 | 83.11 | 89.74 |
| Sonata×DINOv2 | 3D SSL×2D SSL | Both | 75.91 | 85.36 | 91.25 |



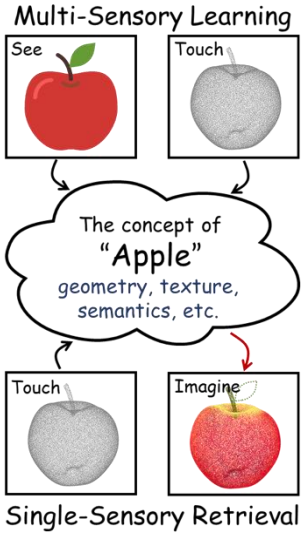
# The power of multi-sensory learning




- When we see an image of an apple, even in grayscale, it is easy for us to **imagine** its geometric **shape,** its possible **color,** and even its unique favors

*To see a World in a Grain of Sand And a Heaven in a Wildflower - William Blake*

# The power of multi-sensory learning

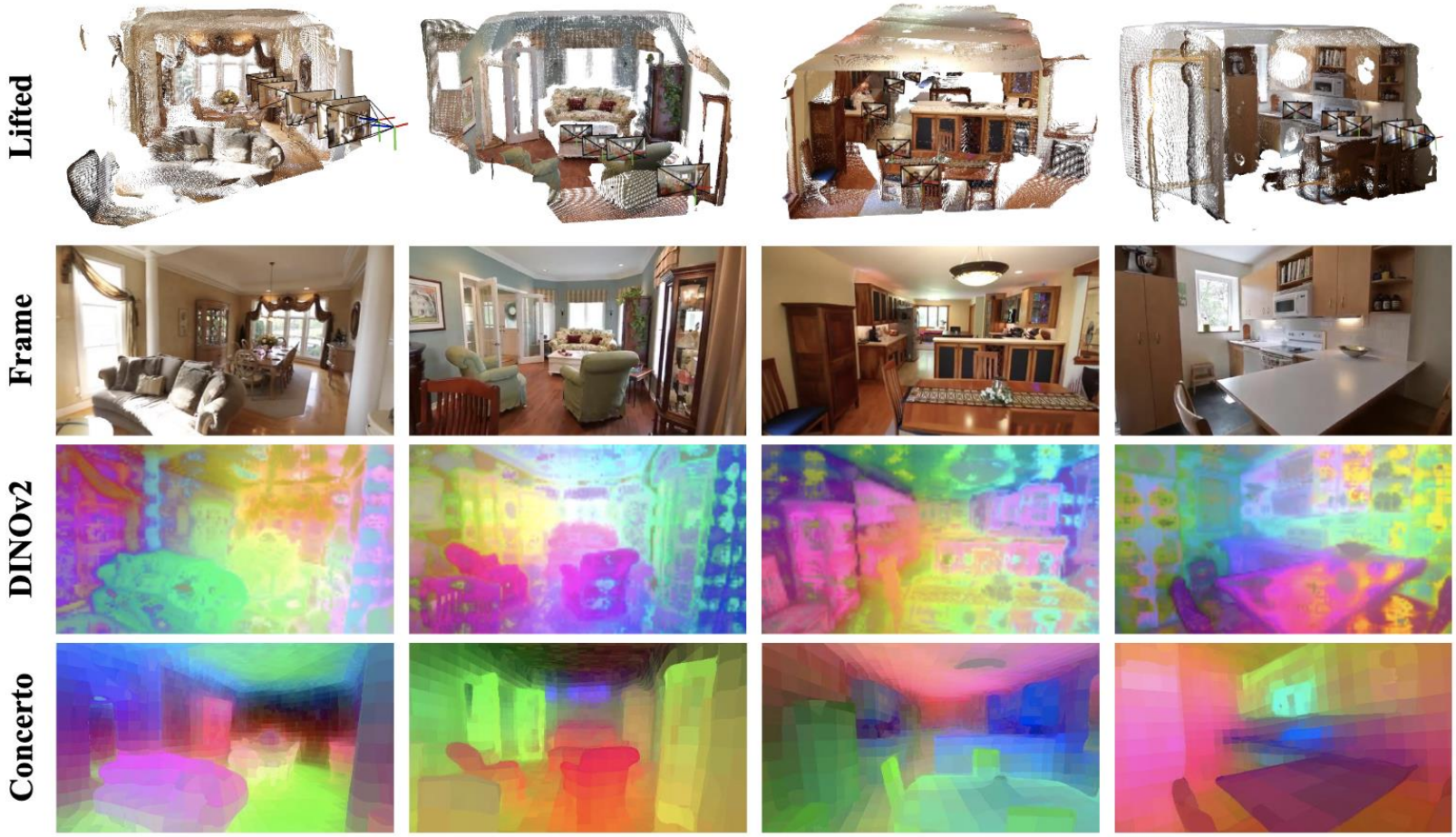



- **Why** human able to produce such powerful embedding?

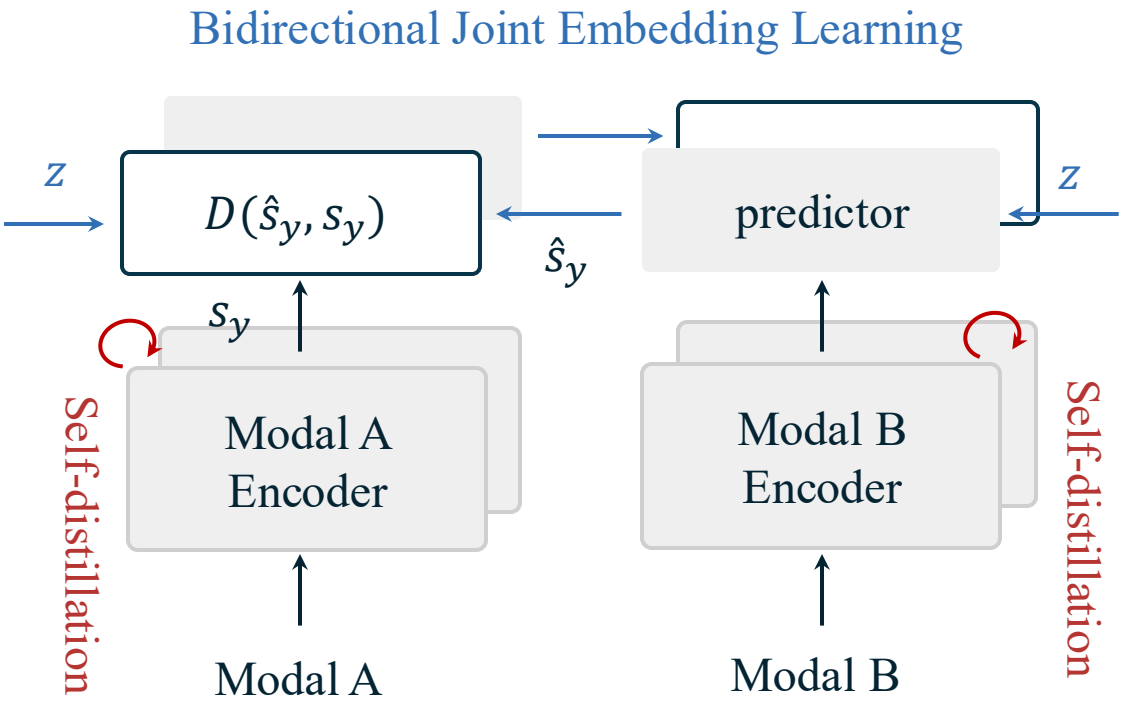
We **see** apple before, we **touch** apple before, we **eat** apple before. We know these multi sensory experience. We formulate a unified representation in our mind. (SSL)

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

# Further scaling up the training with video dataset

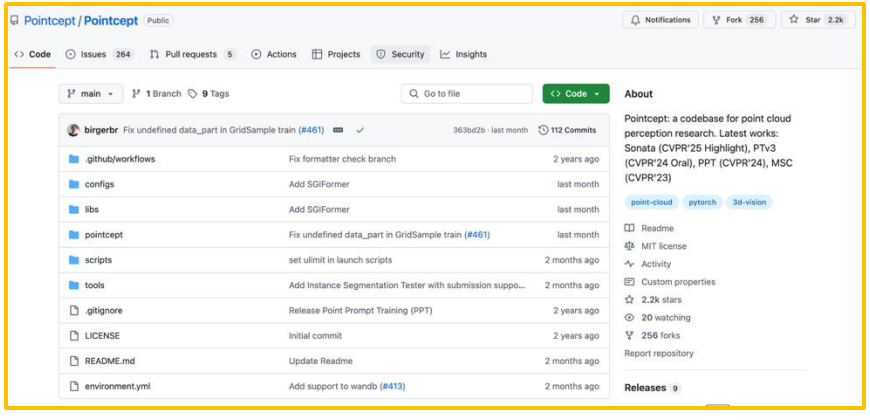


# Native 2D-3D Joint Self-supervised Learning




Two module can be asynchronous

# Thank you!




**pointcept/pointcept**
**(Pre-training)**




**facebookresearch/sonata**
**(Inference)**

香 港 大 學
THE UNIVERSITY OF HONG KONG

POINTCEPT
Point Cloud Perception Codebase