Toward a General-Purpose Foundation Model for Neutrino Physics

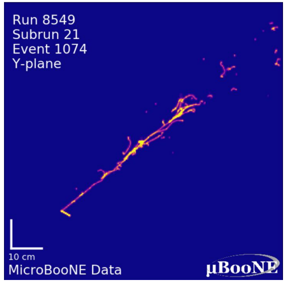
Samuel Young // NPML 2025

youngsam@stanford.edu | youngsm.com

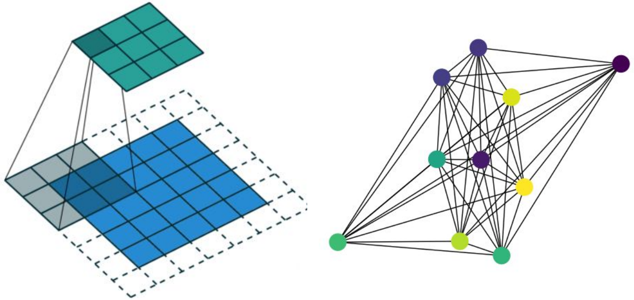
NATIONAL ACCELERATOR LABORATORY

Deep learning in neutrino physics

Current approach: deep neural network(s) trained used simulated datasets with labels via fully supervised learning.



→



→ ν_e/ν_μ

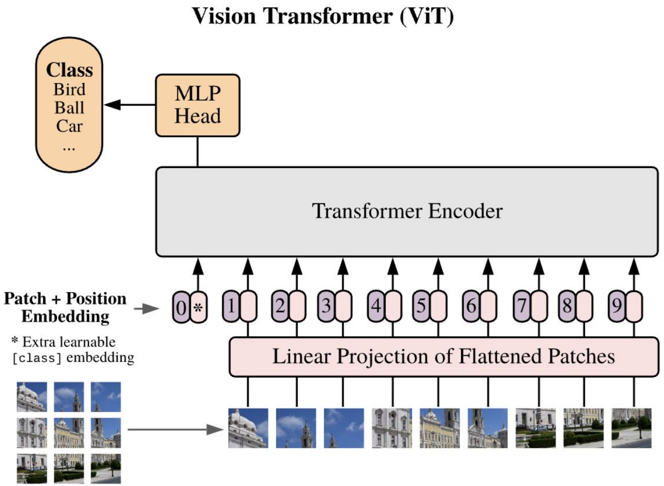
Extremely successful, but:

- Vulnerable to data-simulation discrepancies ⇒ heavy calibration efforts
- "Smart vs. big" models: can removing hand-crafted physical priors result in better performance with the introduction of more compute? [1]
- Task-specific, trained separately from scratch.

[1] Richard Sutton, *The Bitter Lesson*. http://www.incompleteideas.net/IncIdeas/BitterLesson.html

Task specificity

Cat/dog classifier will not learn anything about the difference between trees and flowers.



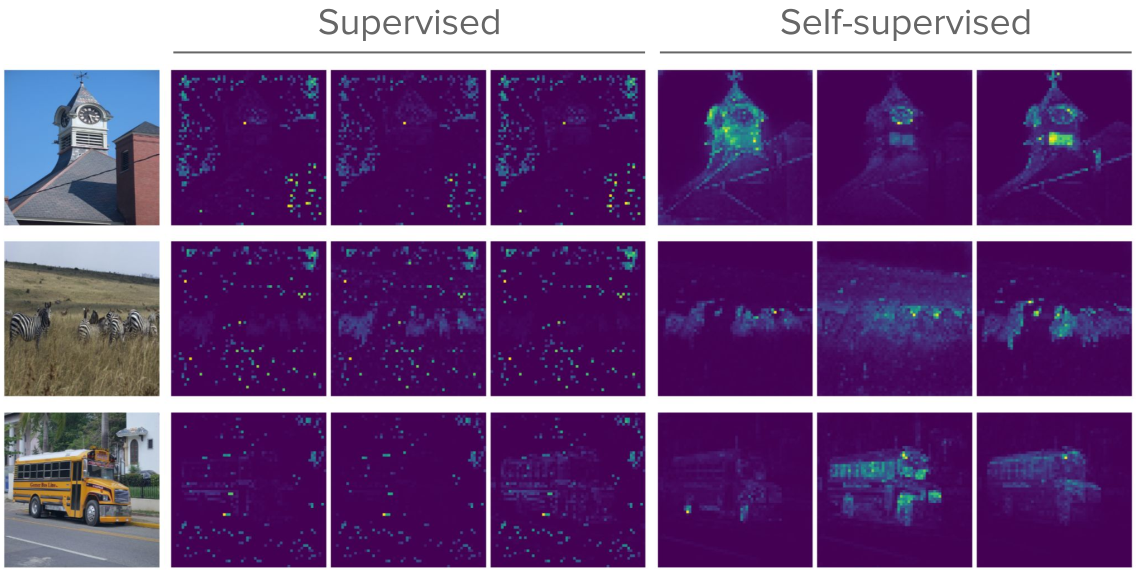
Attention maps

Attention maps from image classification in a vision transformer
[DINO (2104.14294)]

$$Attention(Q, K, V) = \underbrace{\mathrm{softmax}(\frac{QK^T}{\sqrt{d_k}})}_{\text{scores}} V$$

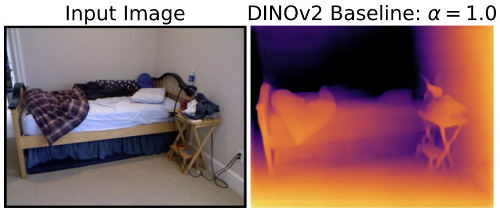
Foundation models are generalists

FM = learn more than the task requires so you can reuse it later

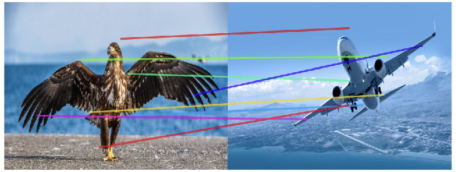


Attention maps from image classification and self-supervised tasks in a vision transformer
[DINO (2104.14294)](https://arxiv.org/abs/2104.14294)

Foundation models are generalists

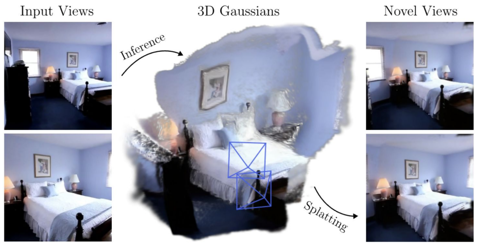


Monocular depth estimation [1]



Point Correspondence [2]

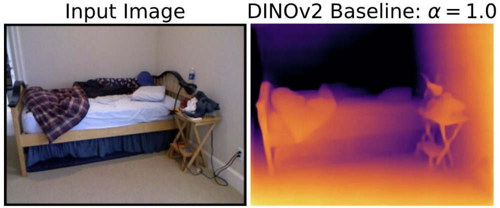
DINO (SSL) [2]



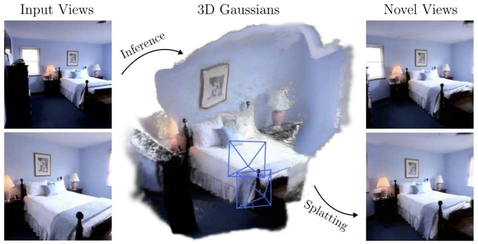
Multi-view Reconstruction [3]

Video Tracking [4]

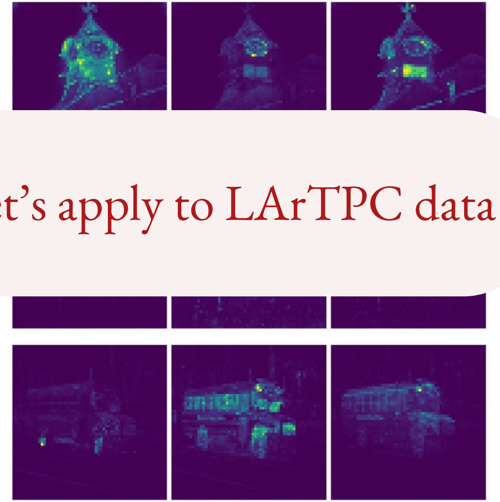
Foundation models are generalists



Monocular depth estimation [1]

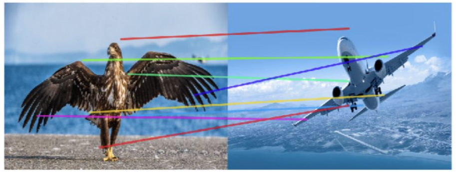


Multi-view Reconstruction [3]



Let's apply to LArTPC data

DINO (SSL) [2]



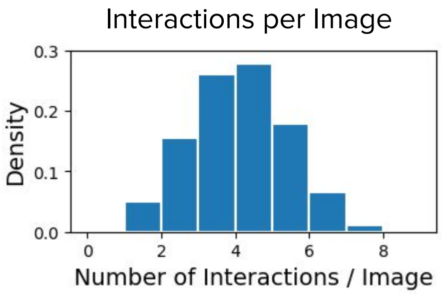
Point Correspondence [2]

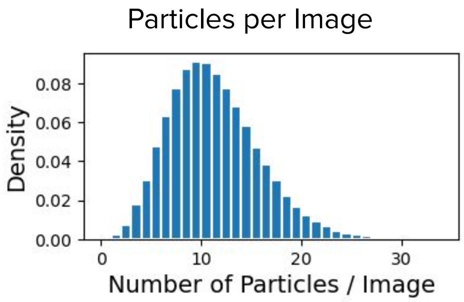
Video Tracking [4]

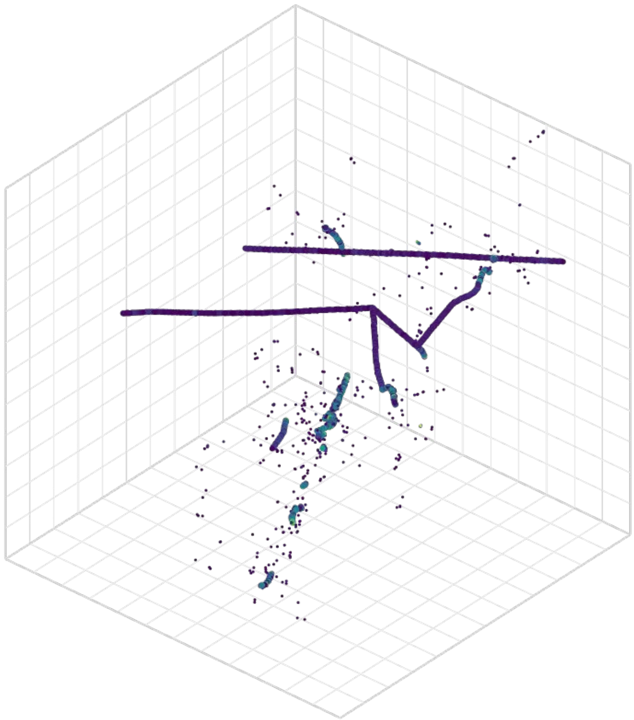
Dataset: PILArNet-Medium

Open data!

- Simulated dataset of 1.2M 3D events
- $(2.3\ \text{m})^3$ cube $(768\ \text{px})^3$. ~5B non-zero voxels.
- +1M events on top of previous open dataset, PILArNet (2020).
- Simply 3D energy depositions, equivalent to "digital hits" from a LArTPC (e.g., DUNE Near Detector)
- 1024 - 30,000 voxels/event





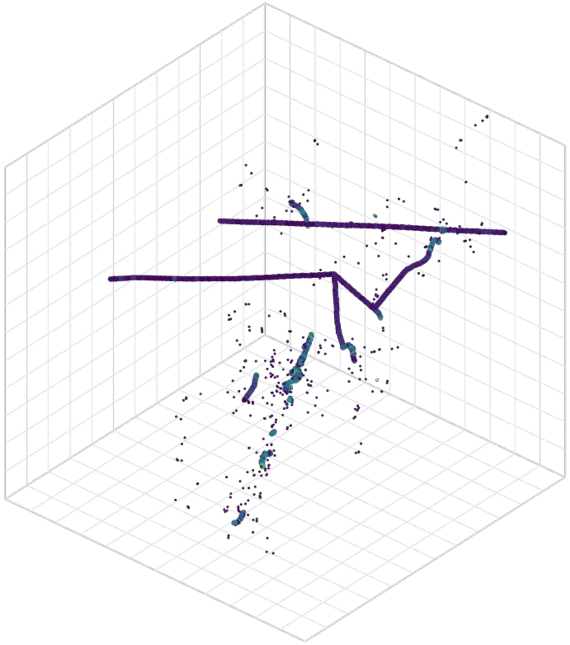


Raw Depositions

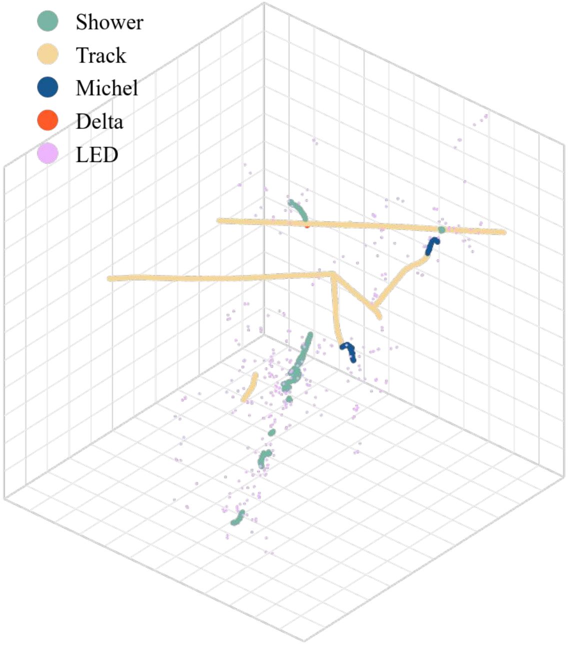
Young, S. et. al, *Particle Trajectory Representation Learning with Masked Point Modeling* (2025)

Semantic Segmentation Labels

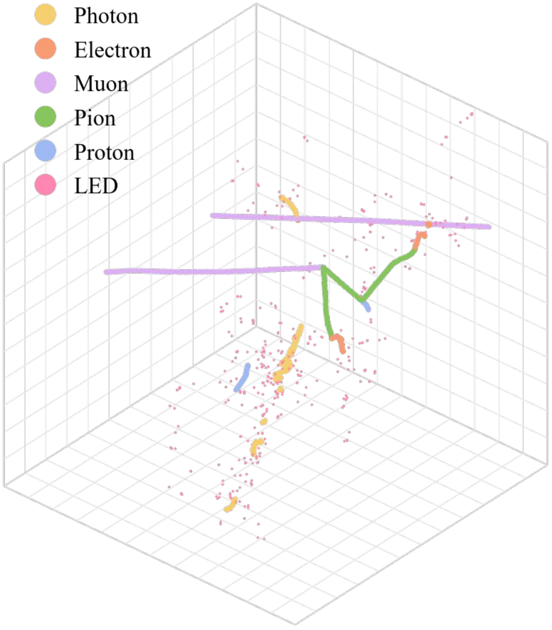
Open data!



Raw Depositions



Semantics Reconstruction

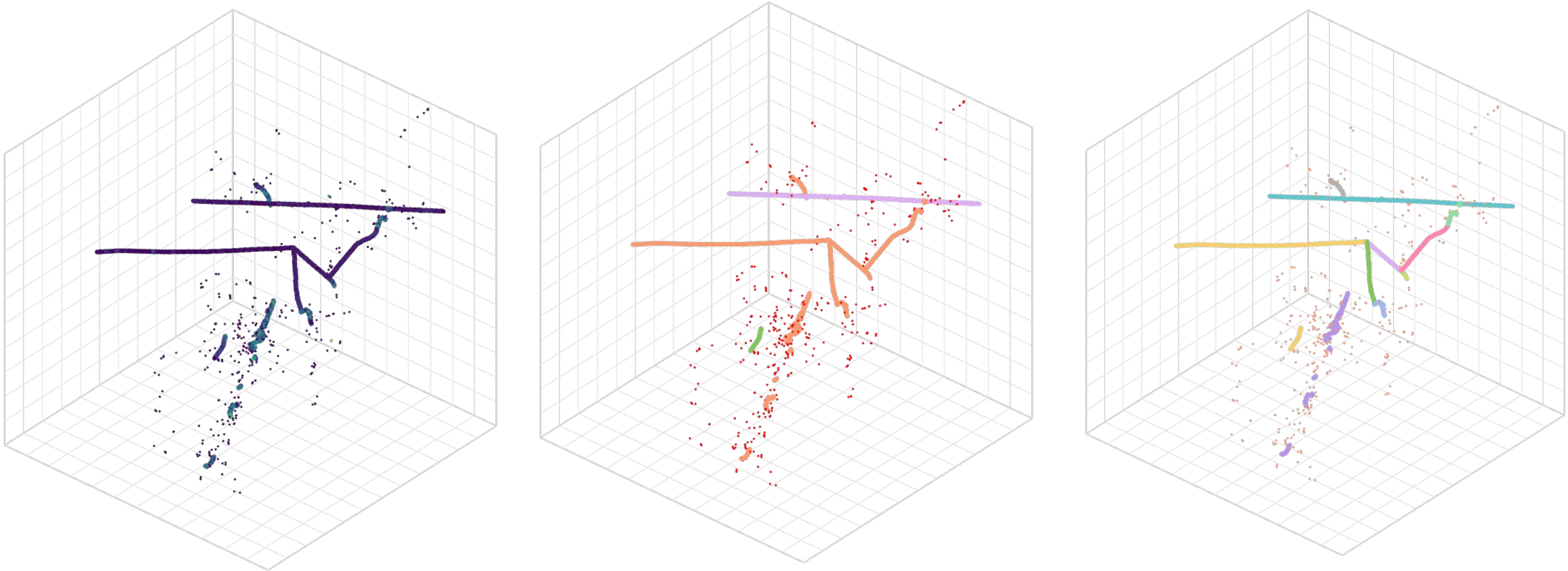


Particle ID Reconstruction

Young, S. et. al, *Particle Trajectory Representation Learning with Masked Point Modeling* (2025)

Instance Labels

Open data!



Raw Depositions | Interaction-level | Particle-level

Young, S. et. al, *Particle Trajectory Representation Learning with Masked Point Modeling* (2025)

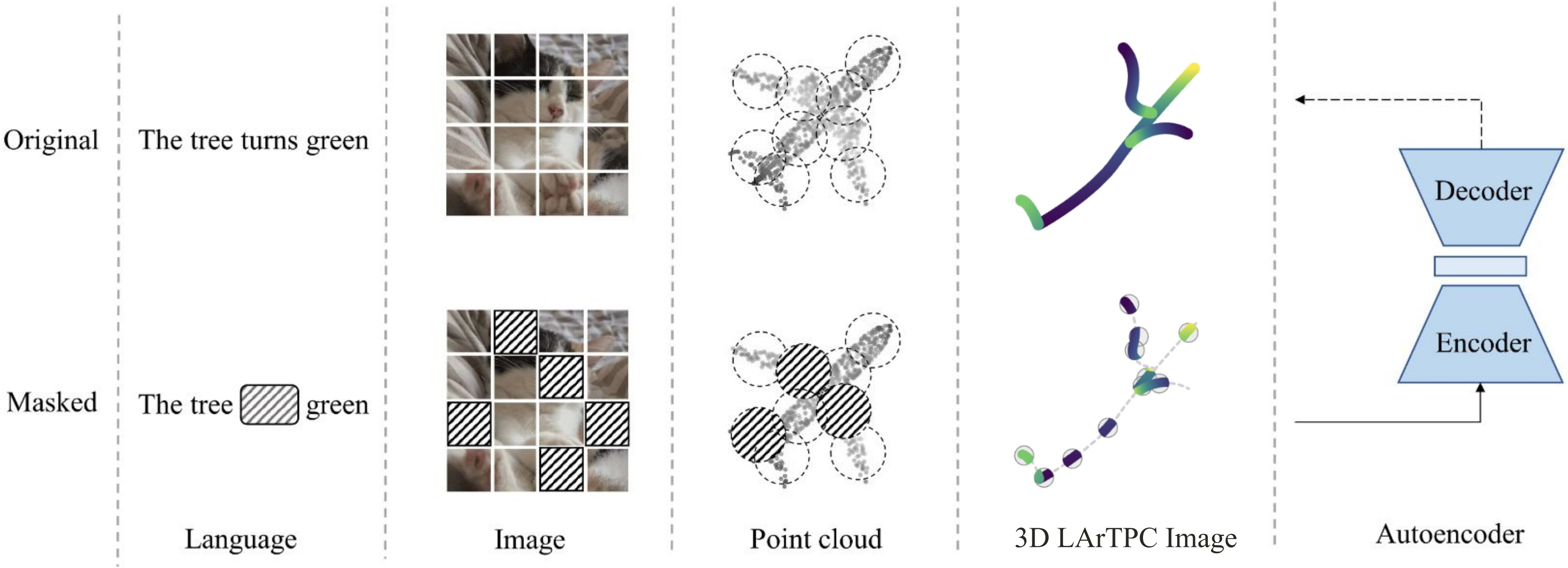
Masked Autoencoders

BERT
(1810.04805)

MAE
(2111.06377)

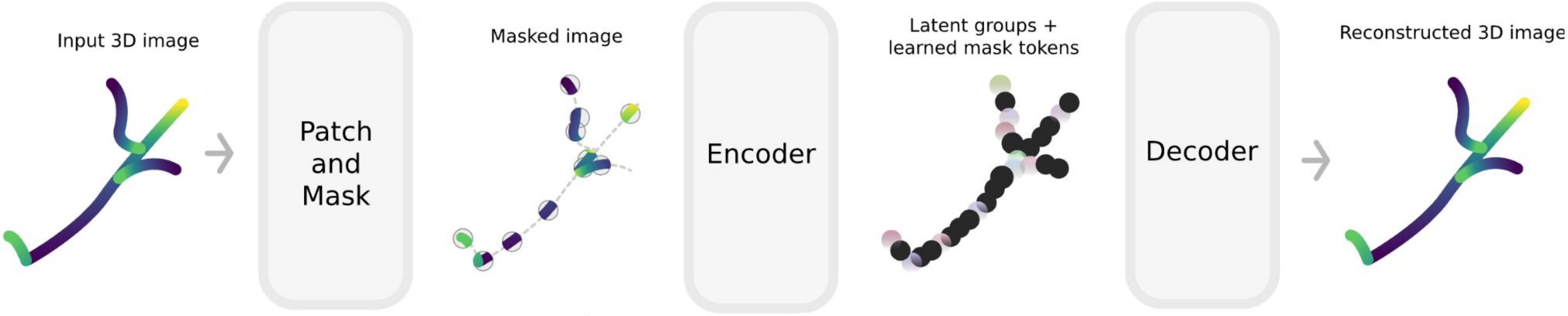
Point-MAE
(2203.06604)

PoLAr-MAE
(2502.02558)



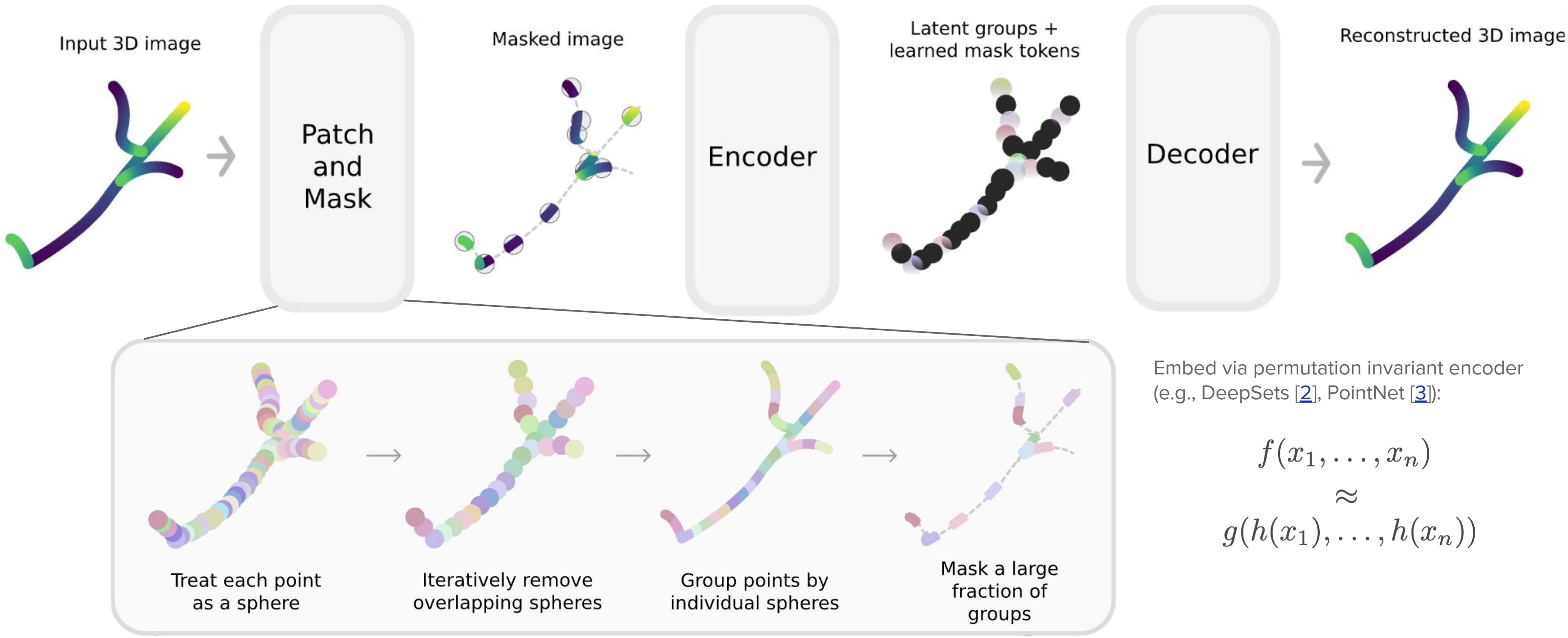
[adapted from PointMAE: arXiv:2203.06604]

PoLAr-MAE: Point-based LAr Masked Autoencoder [1]



- Encoder-decoder is **asymmetric**, i.e. encoder params $\gg$ decoder params.

PoLAr-MAE: Point-based LAr Masked Autoencoder [1]



Embed via permutation invariant encoder (e.g., DeepSets [2], PointNet [3]):

$$f(x_1, \dots, x_n) \approx g(h(x_1), \dots, h(x_n))$$

Young, S. et. al, *Particle Trajectory Representation Learning with Masked Point Modeling* (2025)

PoLAr-MAE: Point-based LAr Masked Autoencoder [1]

(Chamfer distance)

I $\qquad$ R

$$\mathcal{L} = d_{CD}(I, R)$$

$$= \sum_{x\in I} \min_{y\in R} ||x - y||_2^2 + \sum_{x\in R} \min_{y\in I} ||x - y||_2^2$$

Young, S. et. al, *Particle Trajectory Representation Learning with Masked Point Modeling* (2025)

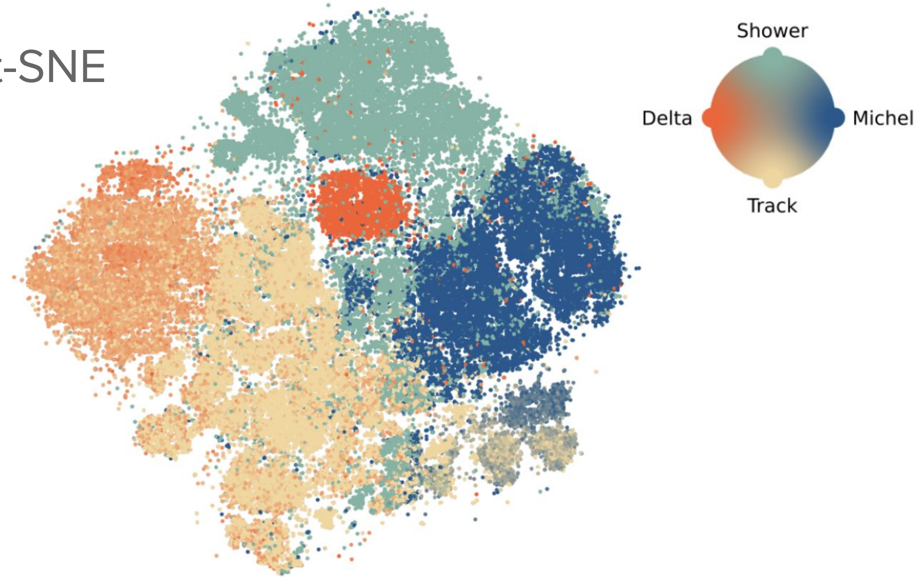
Patch Representations

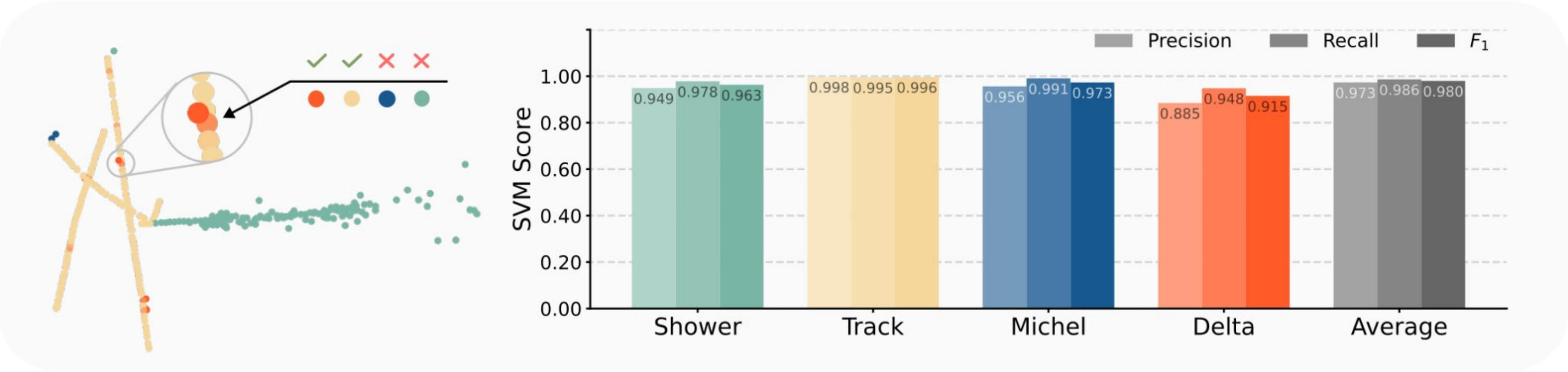
A look at patch representations.

Remember: one patch contains a group of pixels, so can contain >1 particle type.

Patch makeup semantic segmentation

t-SNE

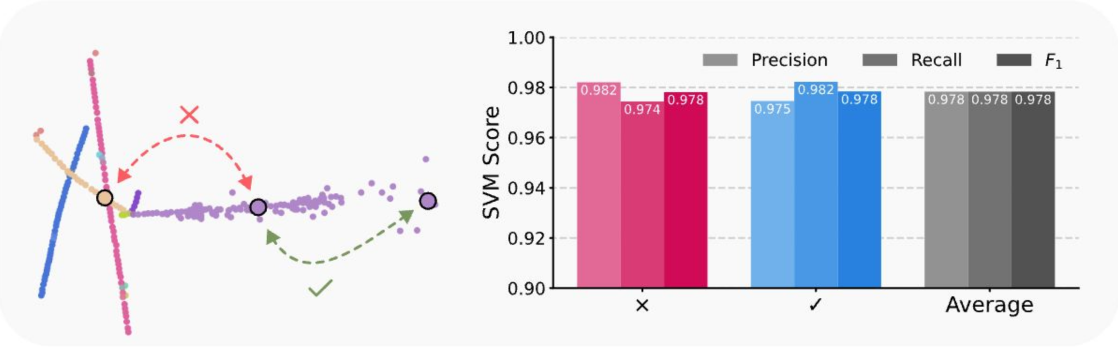




Young, S. et. al, *Particle Trajectory Representation Learning with Masked Point Modeling* (2025)

Instance and Vertex Patch Classification

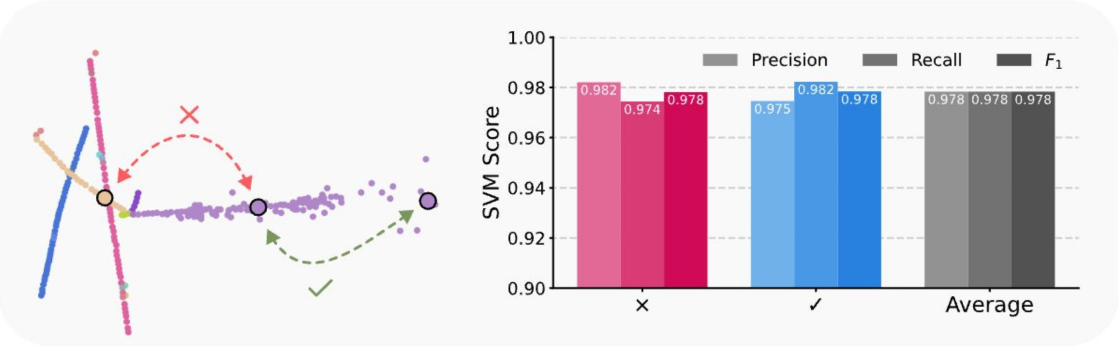
Instance Sharing



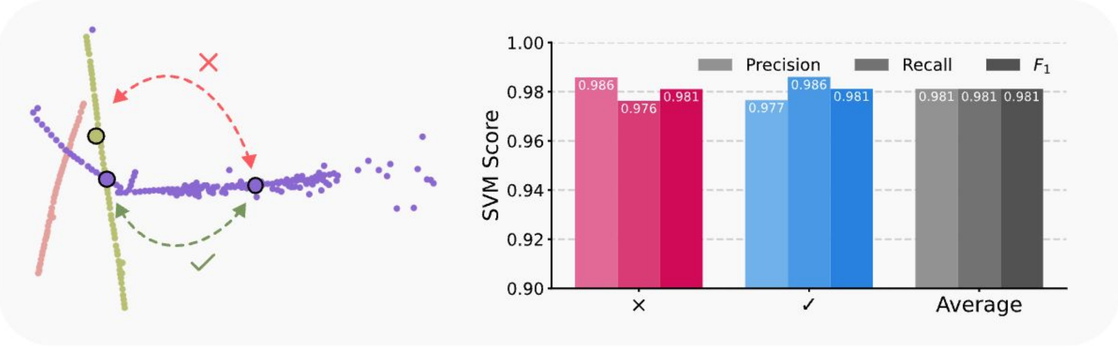
Young, S. et. al, *Particle Trajectory Representation Learning with Masked Point Modeling* (2025)

Instance and Vertex Patch Classification

Instance Sharing



Vertex Sharing

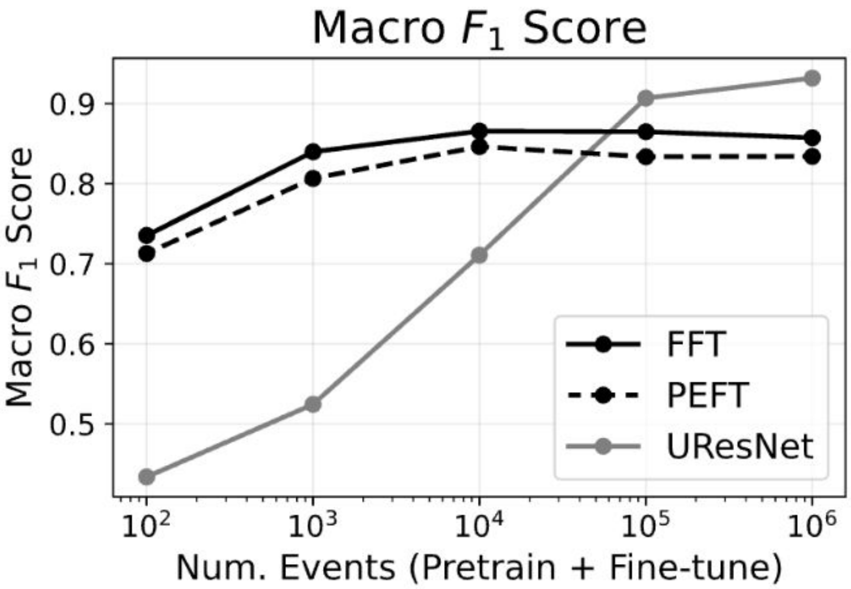


Young, S. et. al, *Particle Trajectory Representation Learning with Masked Point Modeling* (2025)

Comparison to State of the Art (UResNet): Semantic Segmentation

What we care about: per-pixel classification

- Beats state-of-the-art in data-constrained environment, but not in the limit of many events.

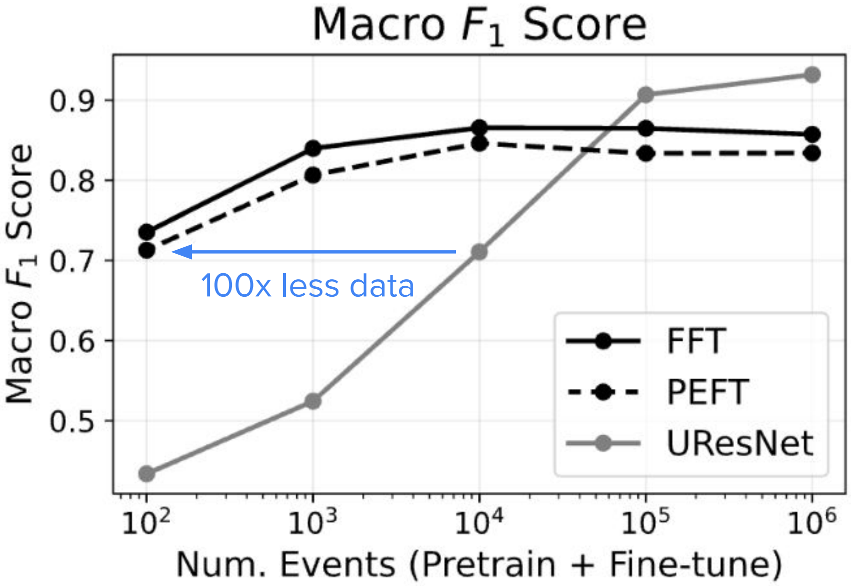


Young, S. et. al, *Particle Trajectory Representation Learning with Masked Point Modeling* (2025)

Comparison to State of the Art (UResNet): Semantic Segmentation

What we care about: per-pixel classification

- Beats state-of-the-art in data-constrained environment, but not in the limit of many events.

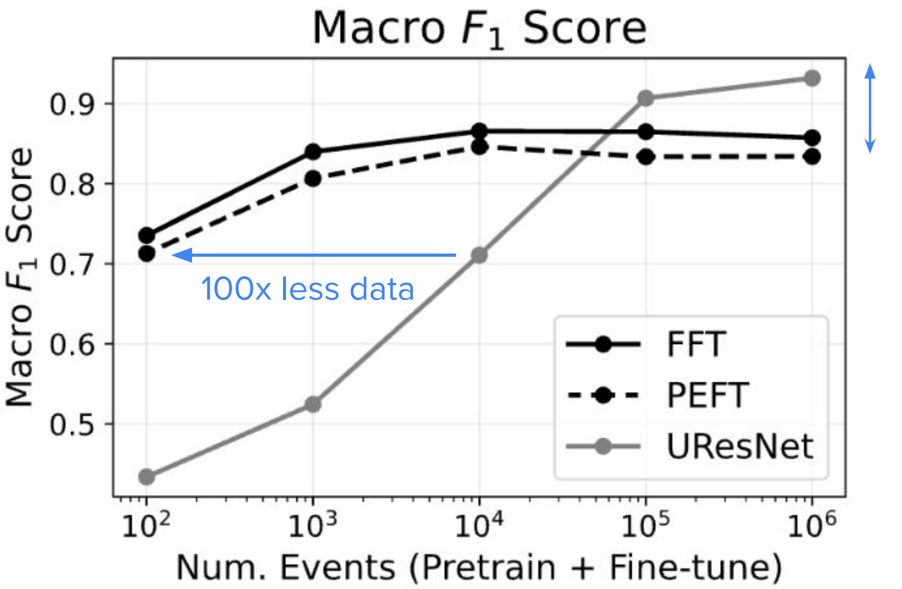


Young, S. et. al, *Particle Trajectory Representation Learning with Masked Point Modeling* (2025)

Comparison to State of the Art (UResNet): Semantic Segmentation

What we care about: per-pixel classification

- Beats state-of-the-art in data-constrained environment, but not in the limit of many events.



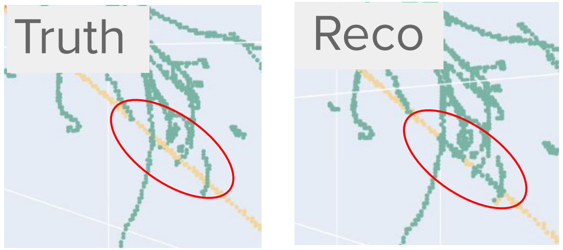
does not beat UResNet at high event counts.

➡ fundamental limit in PoLAr-MAE architecture.

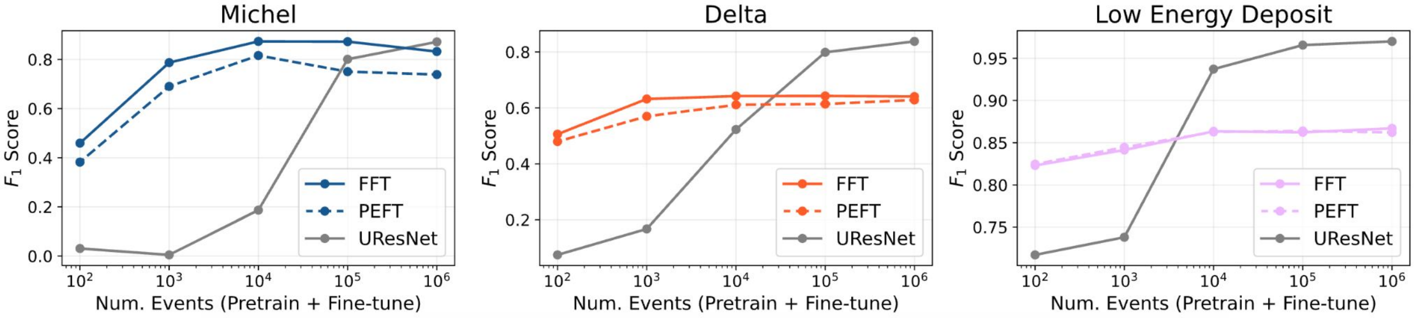
Young, S. et. al, *Particle Trajectory Representation Learning with Masked Point Modeling* (2025)

Comparison to State of the Art (UResNet): Semantic Segmentation

- Small features poorly modeled, i.e. “paint brush” classification.
- This is due to **single-scale** patches being used, which smears tiny structures.



“Small” classes



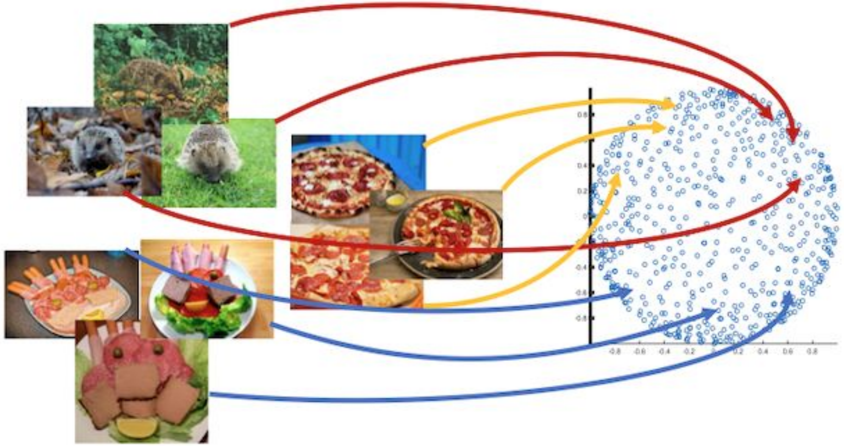
Young, S. et. al, *Particle Trajectory Representation Learning with Masked Point Modeling* (2025)

Panda: Self-distillation and hierarchy

instead of **reconstructing masked portions of image directly**,

let's **predict where they would would end up on a unit sphere** (i.e., classify them),

by enforcing **consistency between global and local views** of the same image.

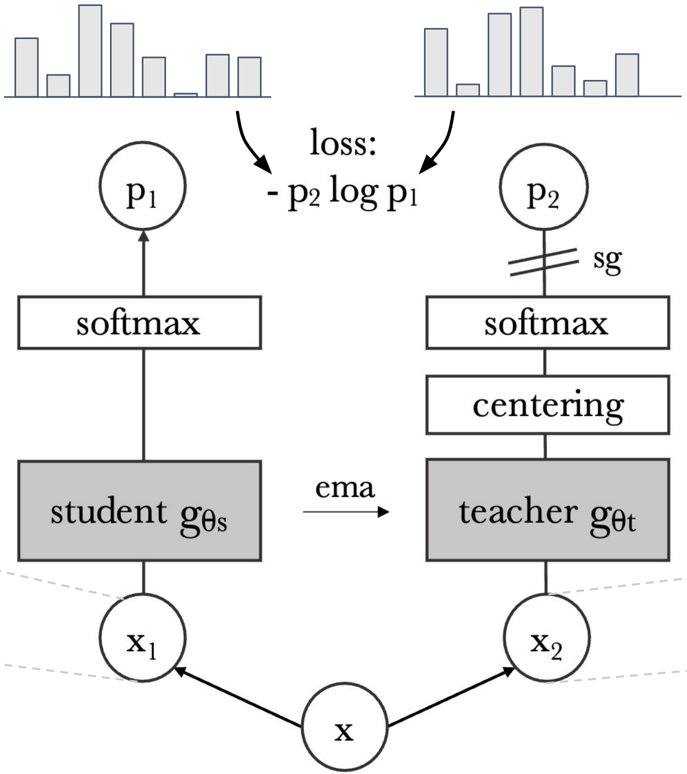


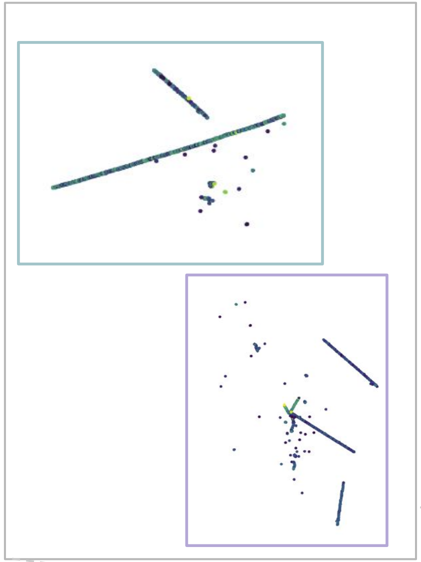
RESEARCH IN PROGRESS; RESULTS SUBJECT TO CHANGE

Young, S. et al., *Panda: Self-distillation of Reusable Sensor-level Representations for High Energy Physics* (2025)

Panda: Self-distillation and hierarchy

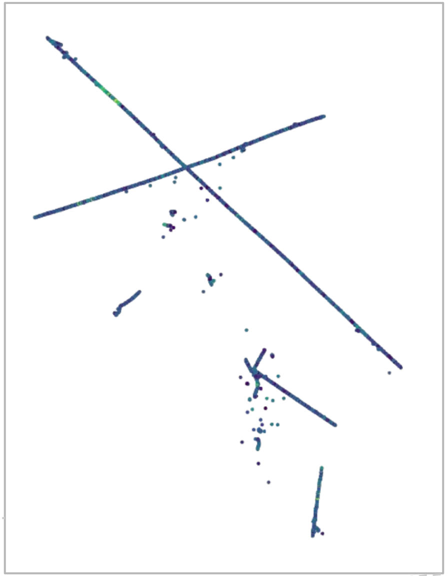
Based on Sonata [1]
Figure from DINO [1,2]





Local Crops

+ Translations & Rotations in 3D
+ O(0.5 voxel) Point Jitter

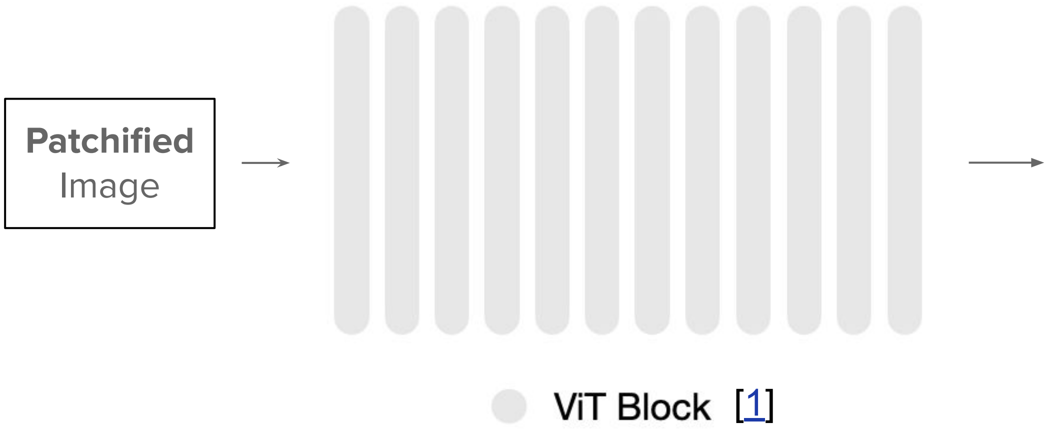


Global View

Young, S. et al., *Panda: Self-distillation of Reusable Sensor-level Representations for High Energy Physics* (2025)

Panda: Self-distillation and hierarchy

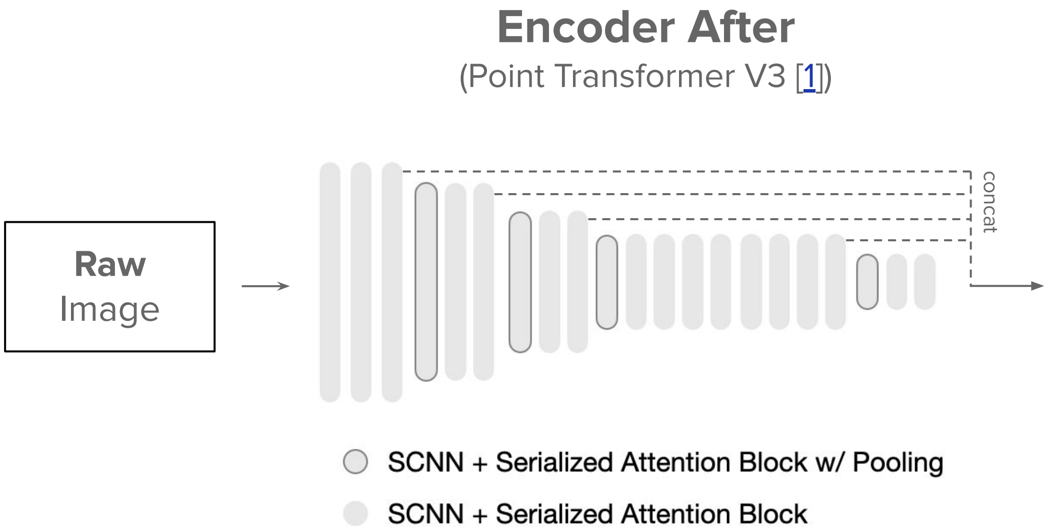
Encoder Before

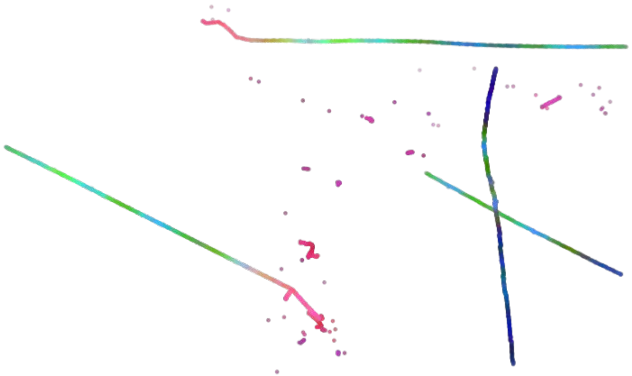


Example Encoded Event

Young, S. et al., *Panda: Self-distillation of Reusable Sensor-level Representations for High Energy Physics* (2025)

Panda: Self-distillation and hierarchy





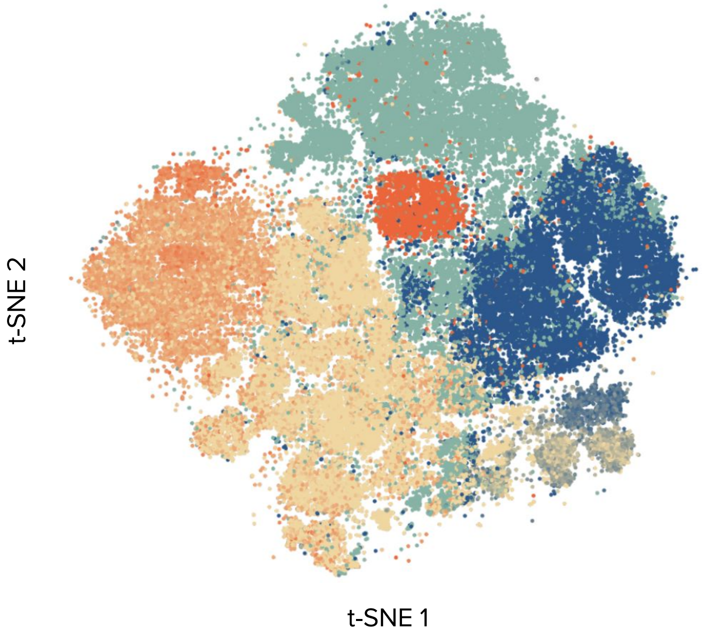
Example Encoded Event

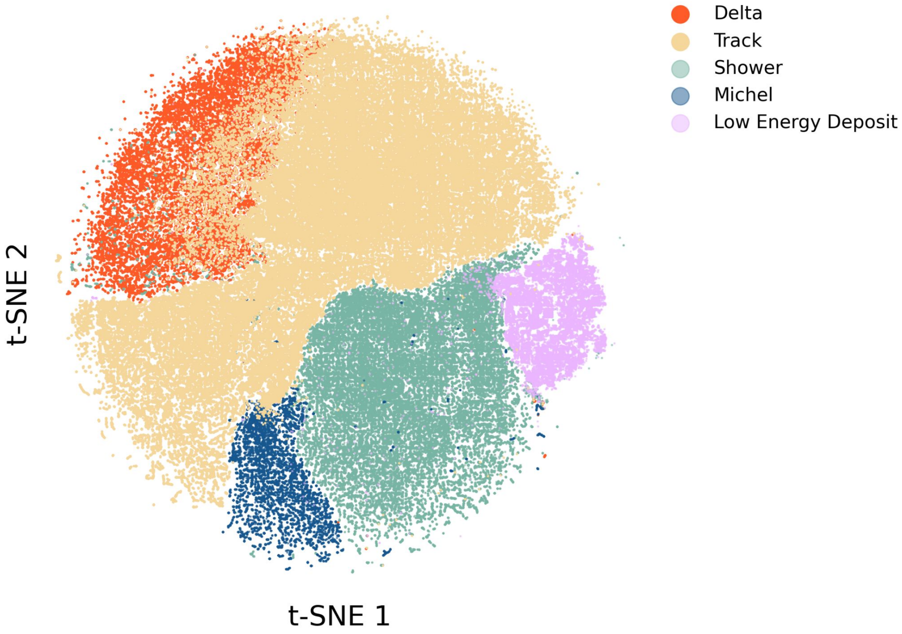
RESEARCH IN PROGRESS; RESULTS SUBJECT TO CHANGE

Young, S. et al., *Panda: Self-distillation of Reusable Sensor-level Representations for High Energy Physics* (2025)

Structure within the Feature Manifold

Pre-trained on 1M events



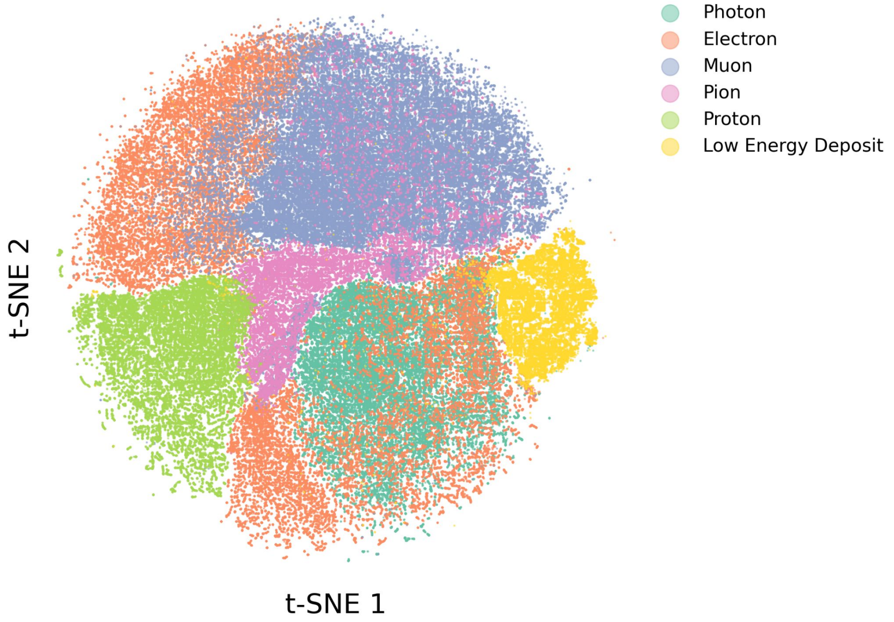


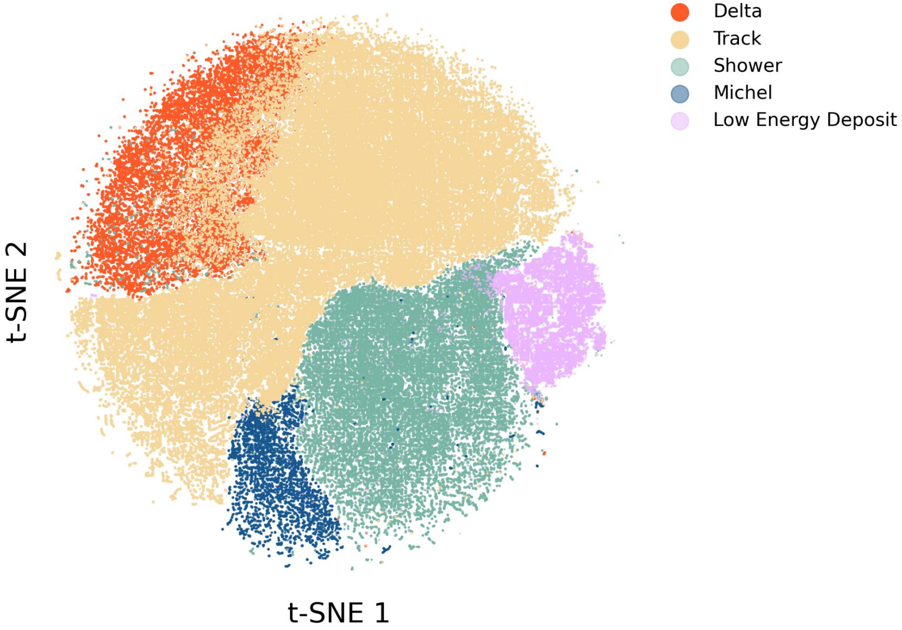
RESEARCH IN PROGRESS; RESULTS SUBJECT TO CHANGE

Young, S. et al., *Panda: Self-distillation of Reusable Sensor-level Representations for High Energy Physics* (2025)

Structure within the Feature Manifold

Pre-trained on 1M events



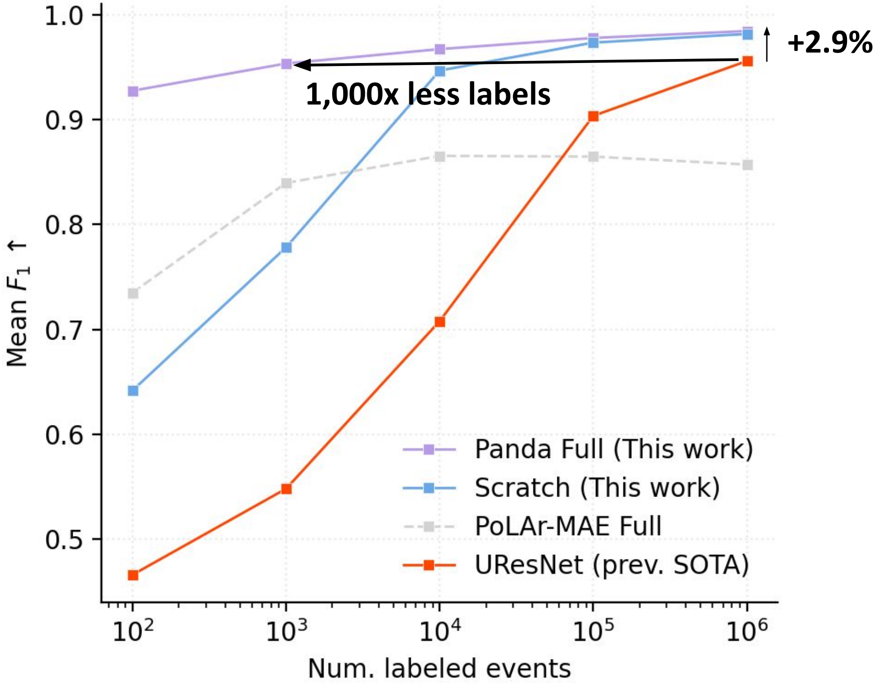


RESEARCH IN PROGRESS; RESULTS SUBJECT TO CHANGE

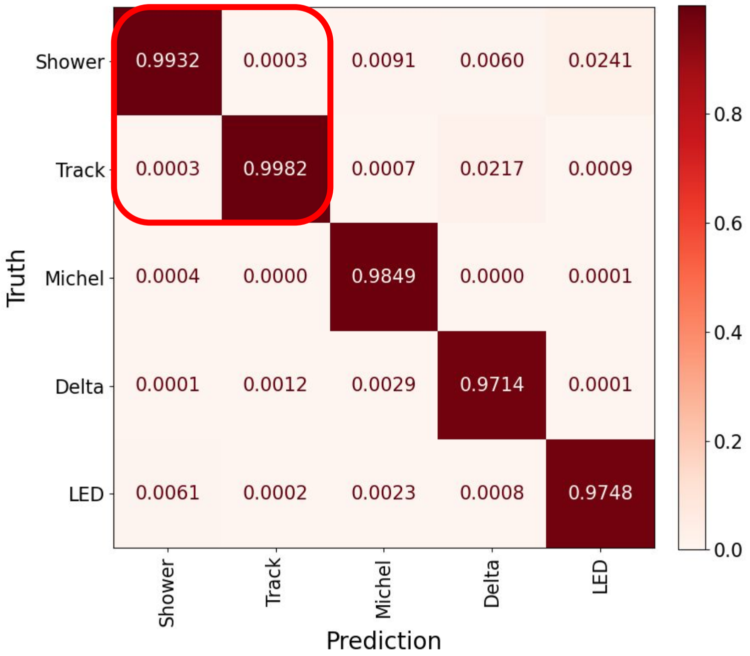
Young, S. et al., *Panda: Self-distillation of Reusable Sensor-level Representations for High Energy Physics* (2025)

Semantic Segmentation: Motif

Pre-trained on 1M events



Normalized by Prediction Column
(diag.=precision)

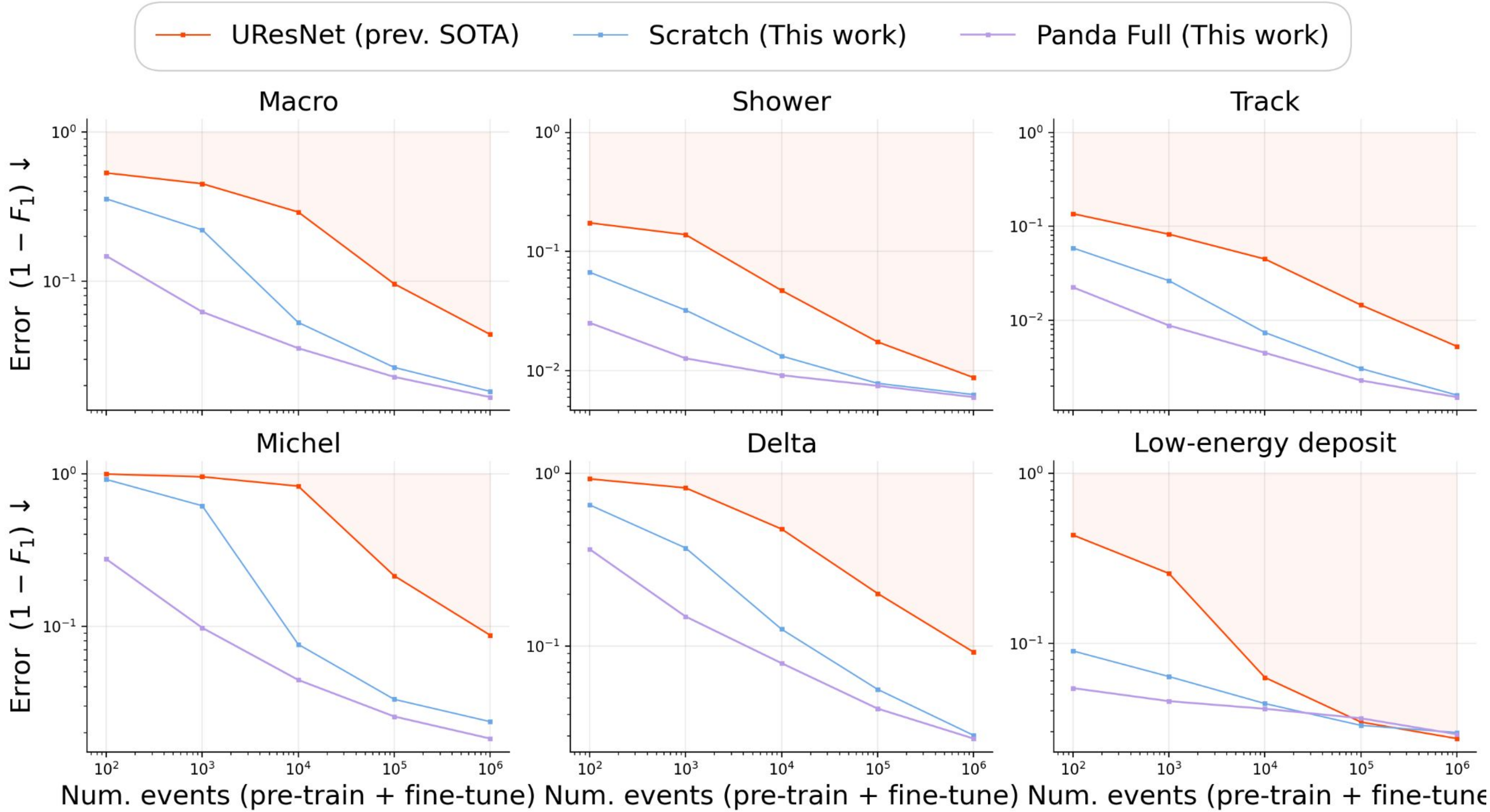


Truth \ Prediction	Shower	Track	Michel	Delta	LED
Shower	0.9932	0.0003	0.0091	0.0060	0.0241
Track	0.0003	0.9982	0.0007	0.0217	0.0009
Michel	0.0004	0.0000	0.9849	0.0000	0.0001
Delta	0.0001	0.0012	0.0029	0.9714	0.0001
LED	0.0061	0.0002	0.0023	0.0008	0.9748

RESEARCH IN PROGRESS; RESULTS SUBJECT TO CHANGE

Young, S. et al., *Panda: Self-distillation of Reusable Sensor-level Representations for High Energy Physics* (2025)

Error $(1 - F_1)$ vs. fine-tune set size — Pre-train + Fine-tune

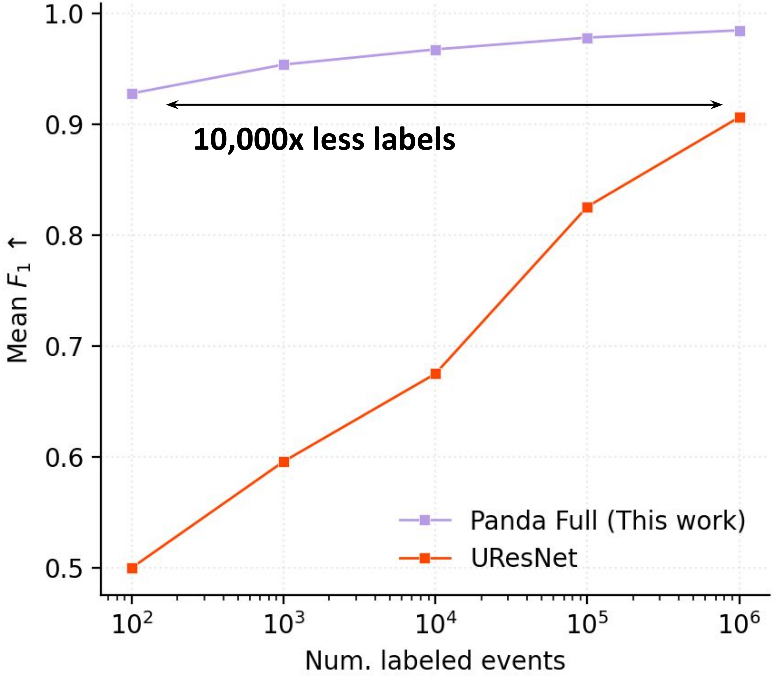


RESEARCH IN PROGRESS; RESULTS SUBJECT TO CHANGE

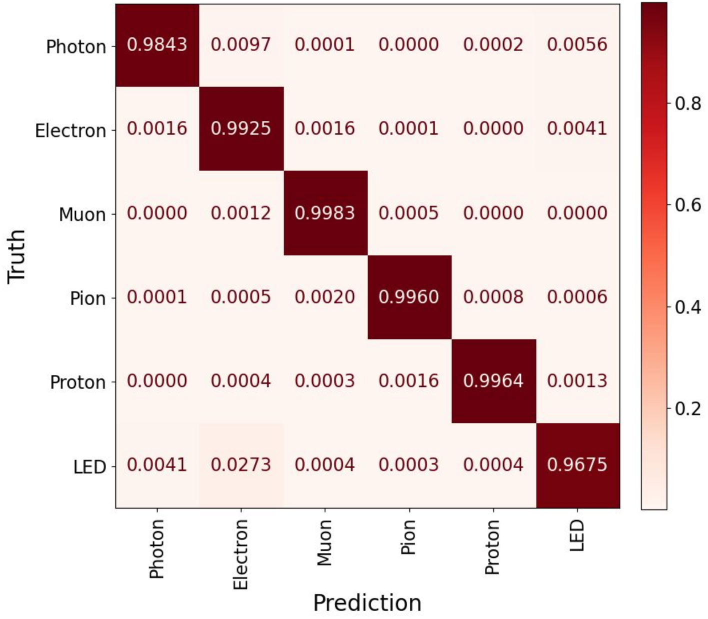
Young, S. et al., *Panda: Self-distillation of Reusable Sensor-level Representations for High Energy Physics* (2025)

Semantic Segmentation: Particle

Pre-trained on 1M events



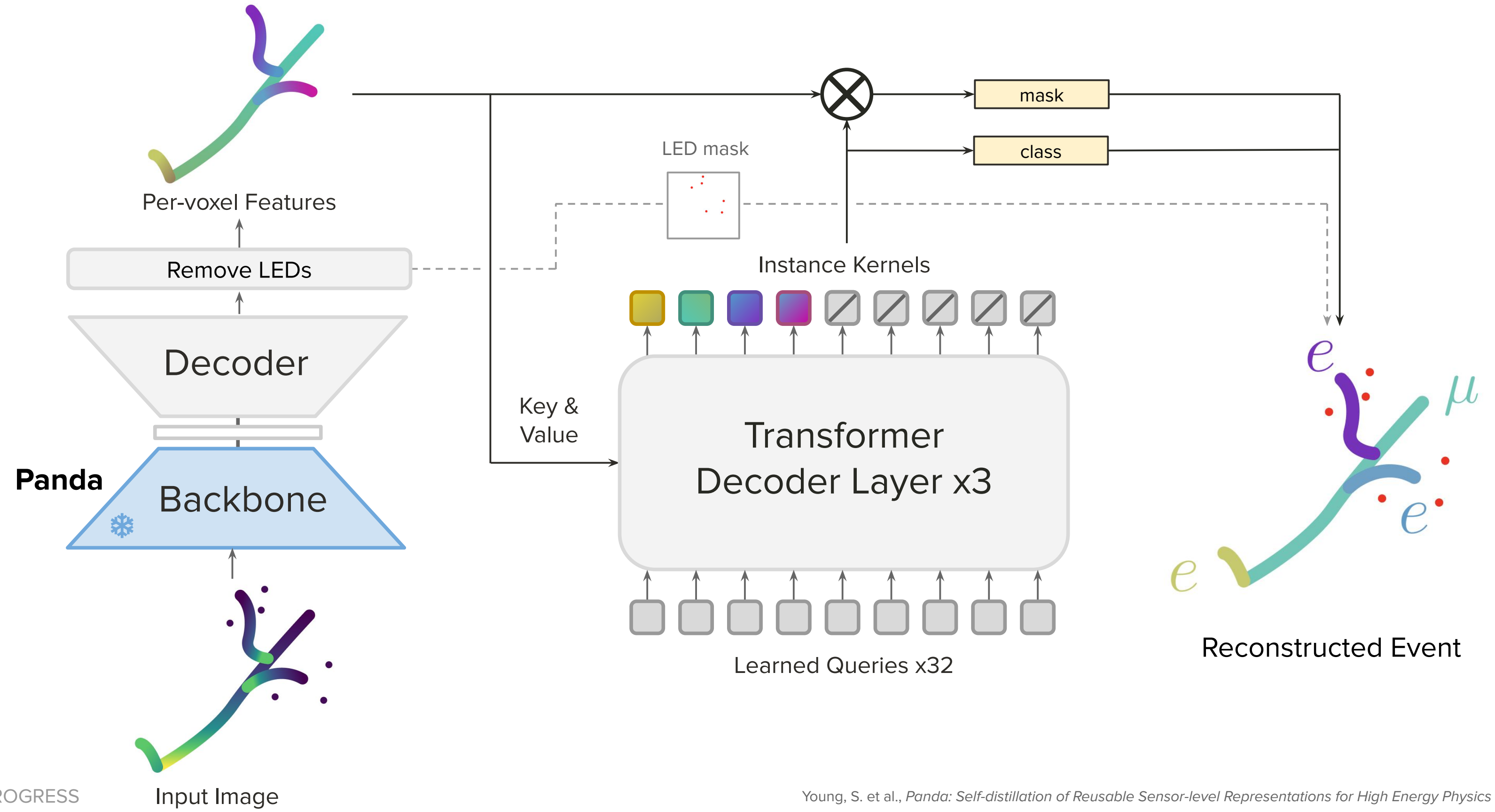
Normalized by Truth Column
(diag.=recall/efficiency), 1M fine-tune



RESEARCH IN PROGRESS; RESULTS SUBJECT TO CHANGE

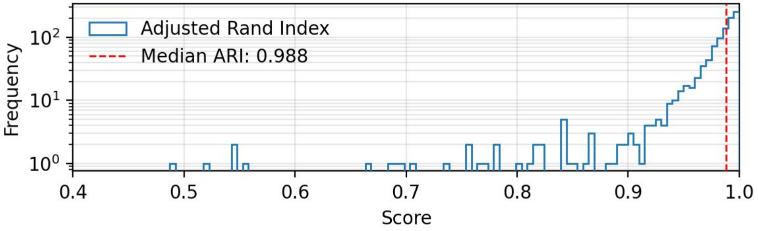
Young, S. et al., *Panda: Self-distillation of Reusable Sensor-level Representations for High Energy Physics* (2025)

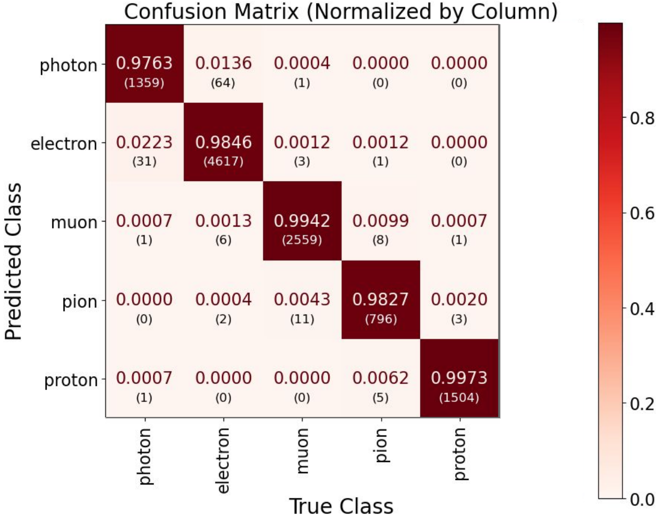
Instance Segmentation: separating particles from one another

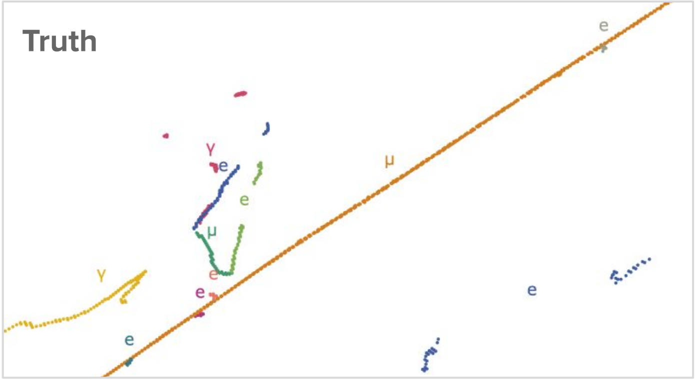


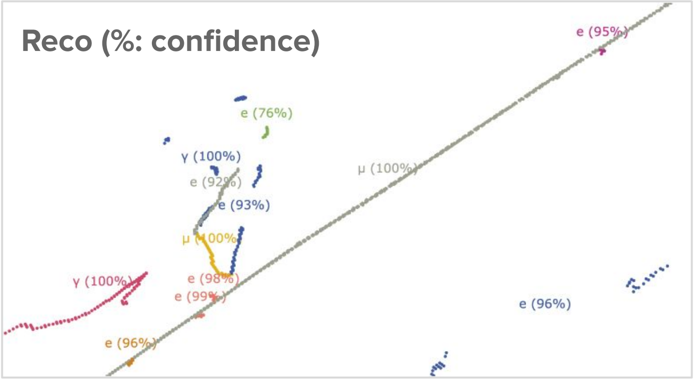
Young, S. et al., *Panda: Self-distillation of Reusable Sensor-level Representations for High Energy Physics* (2025)

Instance Segmentation: Particle





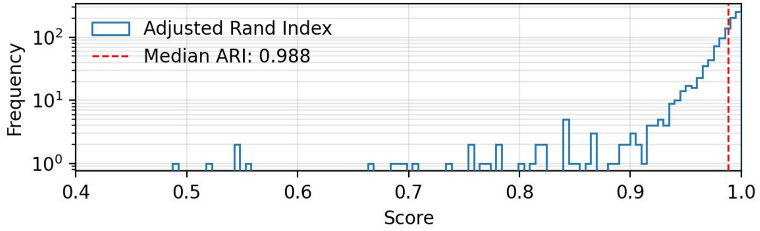


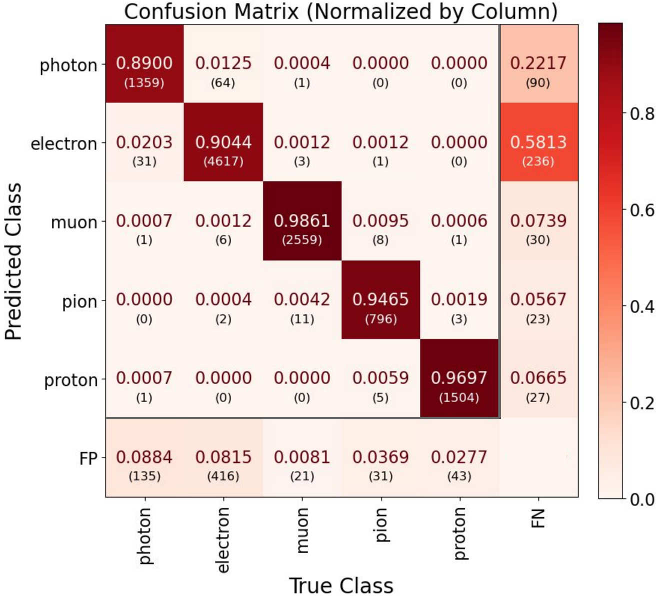


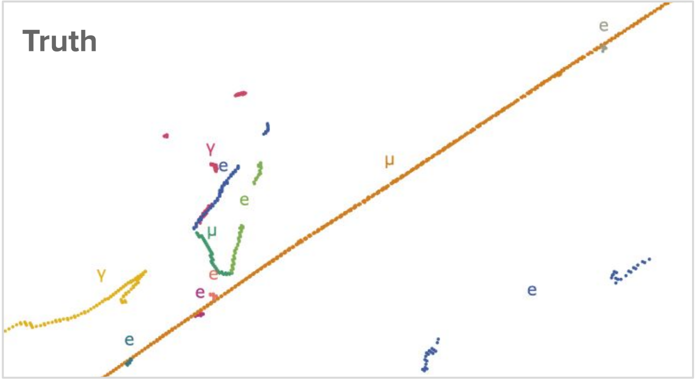
RESEARCH IN PROGRESS; RESULTS SUBJECT TO CHANGE

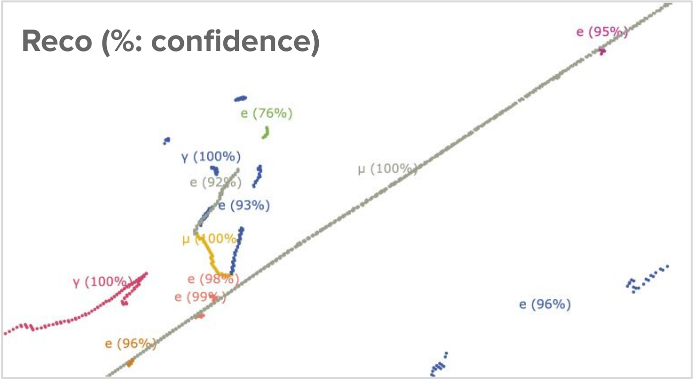
Young, S. et al., *Panda: Self-distillation of Reusable Sensor-level Representations for High Energy Physics* (2025)

Instance Segmentation: Particle





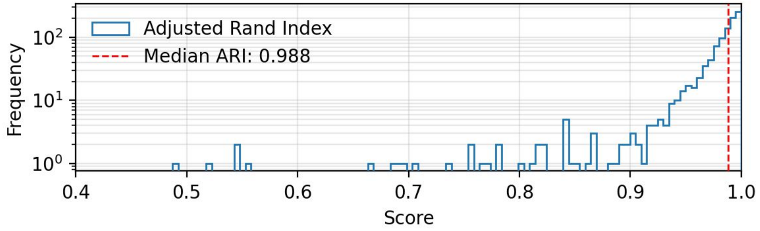


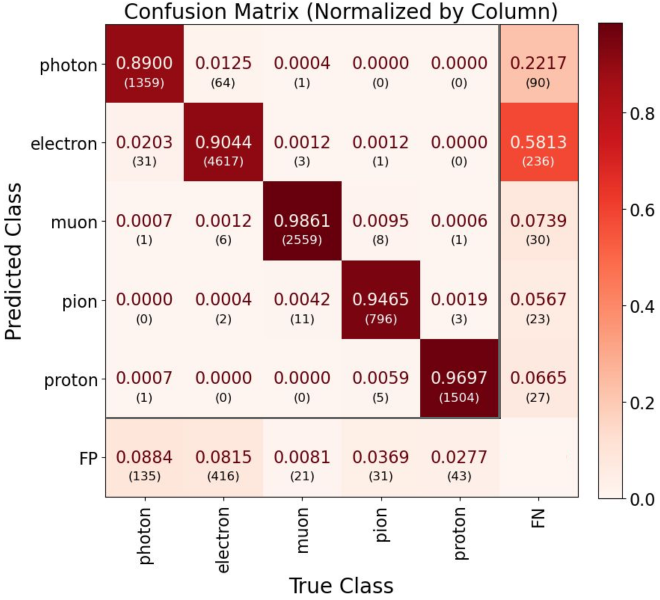


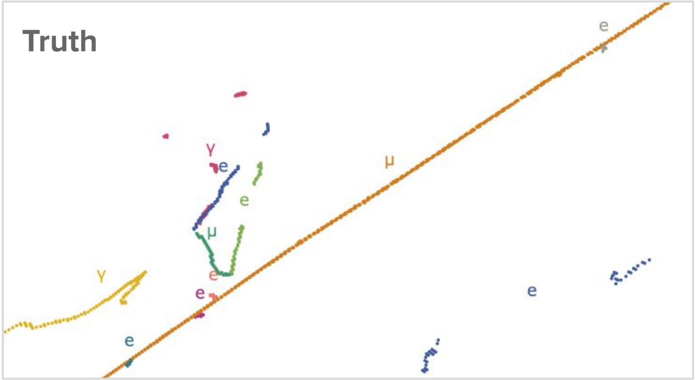
RESEARCH IN PROGRESS; RESULTS SUBJECT TO CHANGE

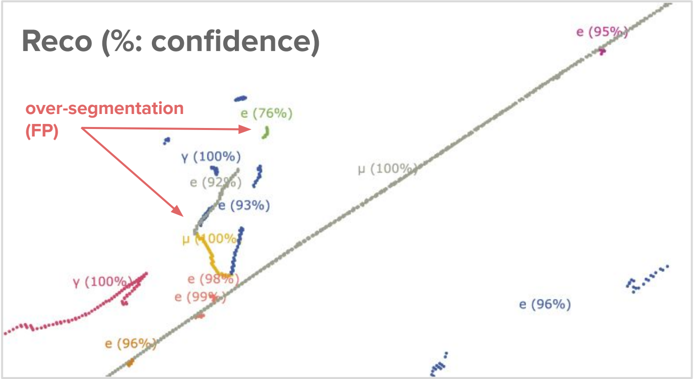
Young, S. et al., *Panda: Self-distillation of Reusable Sensor-level Representations for High Energy Physics* (2025)

Instance Segmentation: Particle





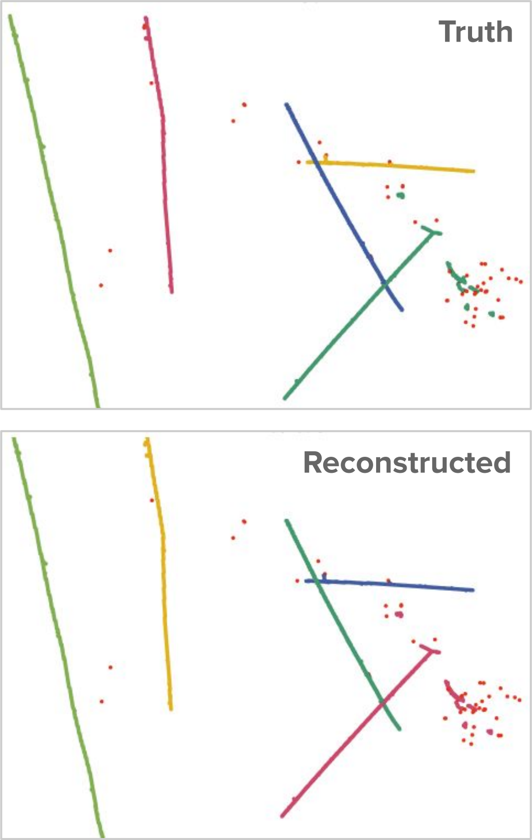


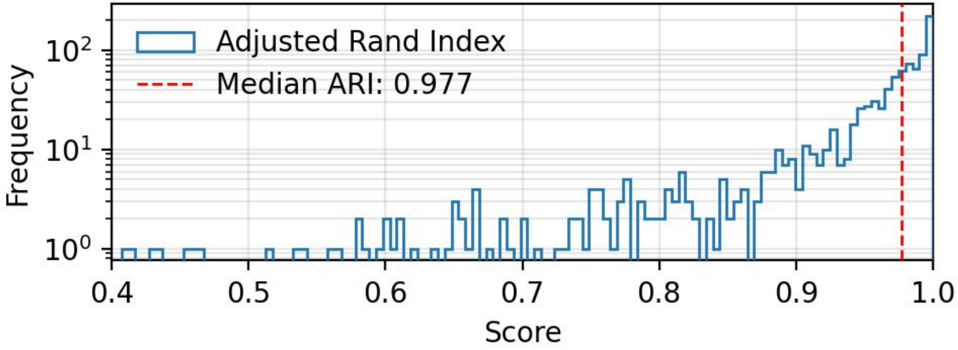


RESEARCH IN PROGRESS; RESULTS SUBJECT TO CHANGE

Young, S. et al., *Panda: Self-distillation of Reusable Sensor-level Representations for High Energy Physics* (2025)

Instance Segmentation: Interaction





RESEARCH IN PROGRESS; RESULTS SUBJECT TO CHANGE

Young, S. et al., *Panda: Self-distillation of Reusable Sensor-level Representations for High Energy Physics* (2025)

Takeaways

- Self-supervised learning works, and we're just getting started.
- A generic feature extractor unlocks new possibilities that were simply not possible before:
  - **Few-shot learning w/o well-calibrated sim**: track/shower, Michel tagging, particle ID, …
  - **Reasoning over images/captioning** with language (human-in-the-loop)
  - **Content-retrieval at scale**: "find events like this" in this dataset.
  - **Cross-experiment datasets** → invariant embeddings across detector conditions, easy adaptation.
- **Future paths**: multi-modal sensors (optical, 2D projection, …), incorporating (neuro-)symbolic reasoning, cross-experiment datasets, detector-related effects.
- Reproducing the results of other models is really hard and time-consuming (specific data formats, dependencies, closed-source). **Common benchmark(s) would go a long way in accelerating the work we do.**

The future is exciting!

Questions? Want to collaborate? Email: youngsam@stanford.edu

Extras

Sharpening + Centering

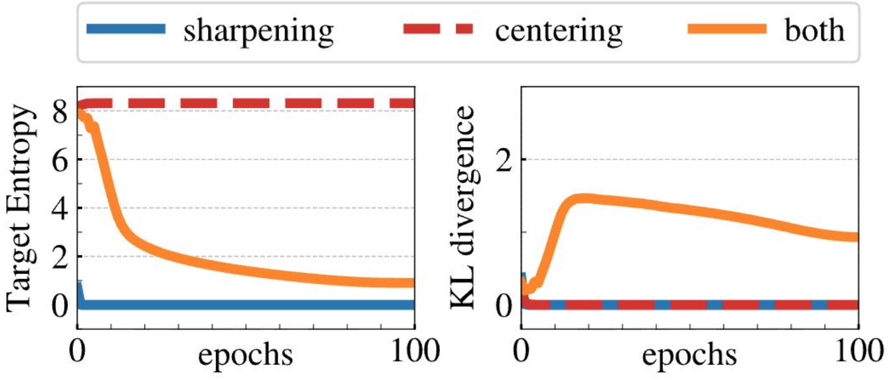


Figure 7: **Collapse study. (left)**: evolution of the teacher's target entropy along training epochs; **(right)**: evolution of KL divergence between teacher and student outputs.

There are two forms of collapse: regardless of the input, the model output is uniform along all the dimensions or dominated by one dimension. The centering avoids the collapse induced by a dominant dimension, but encourages an uniform output. Sharpening induces the opposite effect. We show this complementarity by decomposing the cross-entropy H into an entropy h and the Kullback-Leibler divergence ("KL") D_{KL}:

$$H(P_t, P_s) = h(P_t) + D_{KL}(P_t|P_s). \quad (5)$$

A KL equal to zero indicates a constant output, and hence a collapse.

Sharpening:

$$P_s(x)^{(i)} = \frac{\exp(g_{\theta_s}(x)^{(i)}/\tau_s)}{\sum_{k=1}^{K} \exp(g_{\theta_s}(x)^{(k)}/\tau_s)}, \quad (1)$$

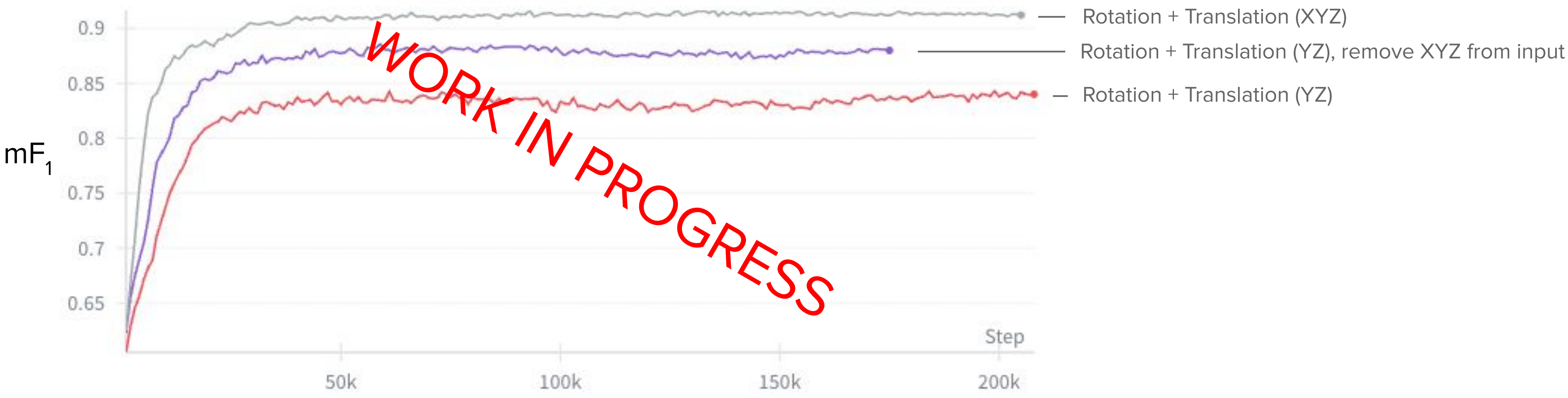
with $\tau_s > 0$ a temperature parameter

From *Emerging Properties in Self-Supervised Vision Transformers* (arXiv:2104.14294)

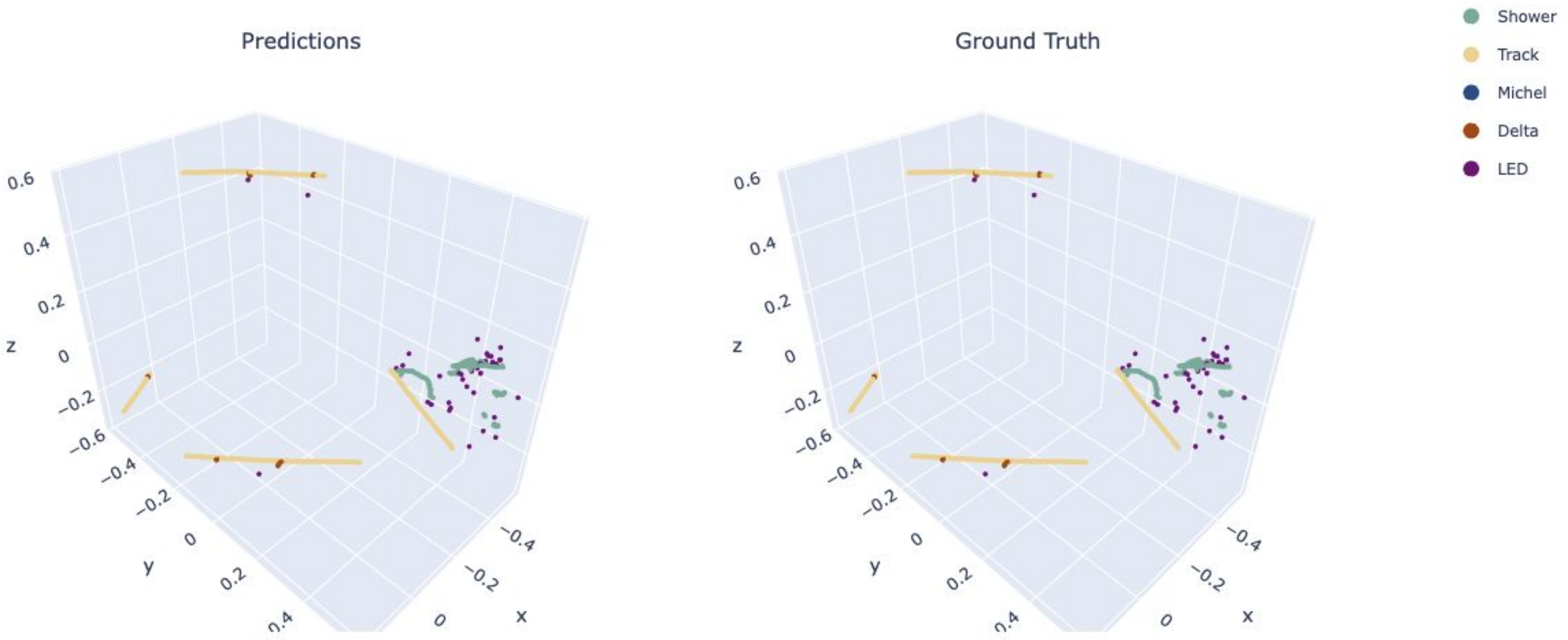
Augmentations

What about diffusion/attenuation?

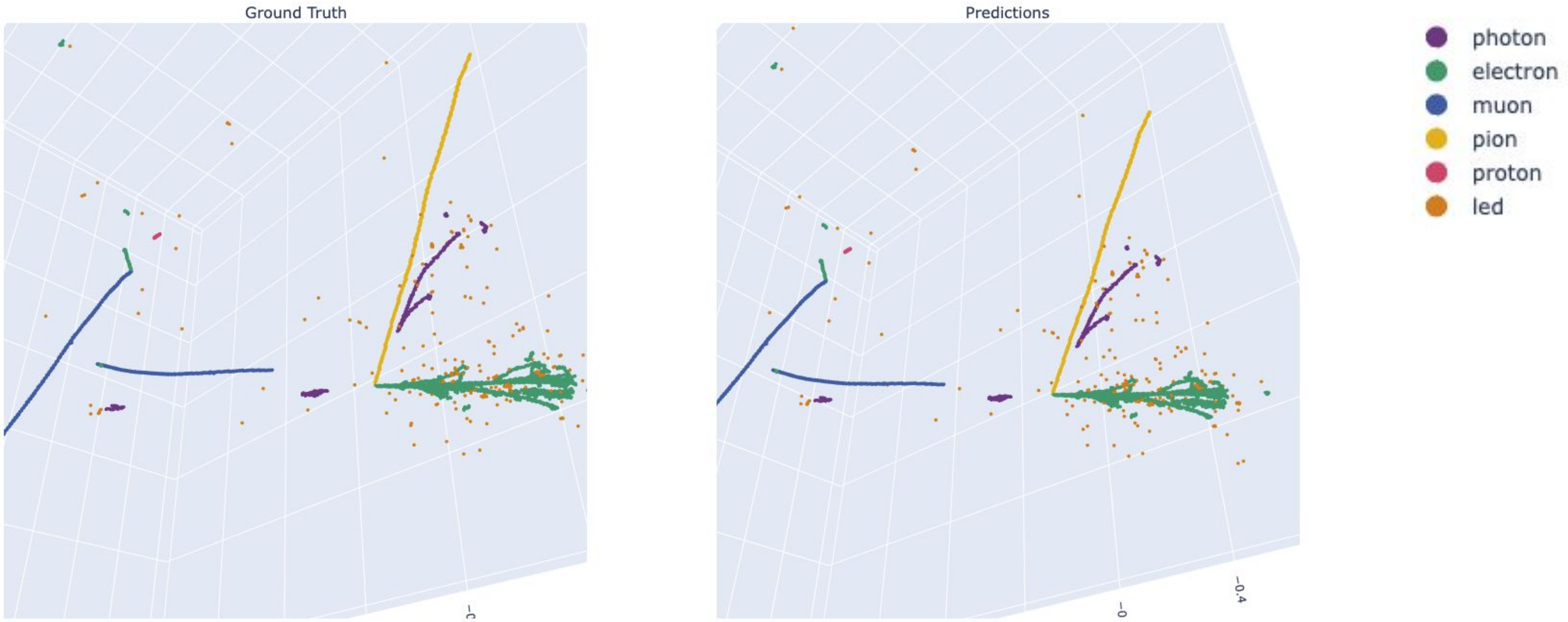
Linear Evaluation on pre-training – 1M events



Example Sem Seg - Motif

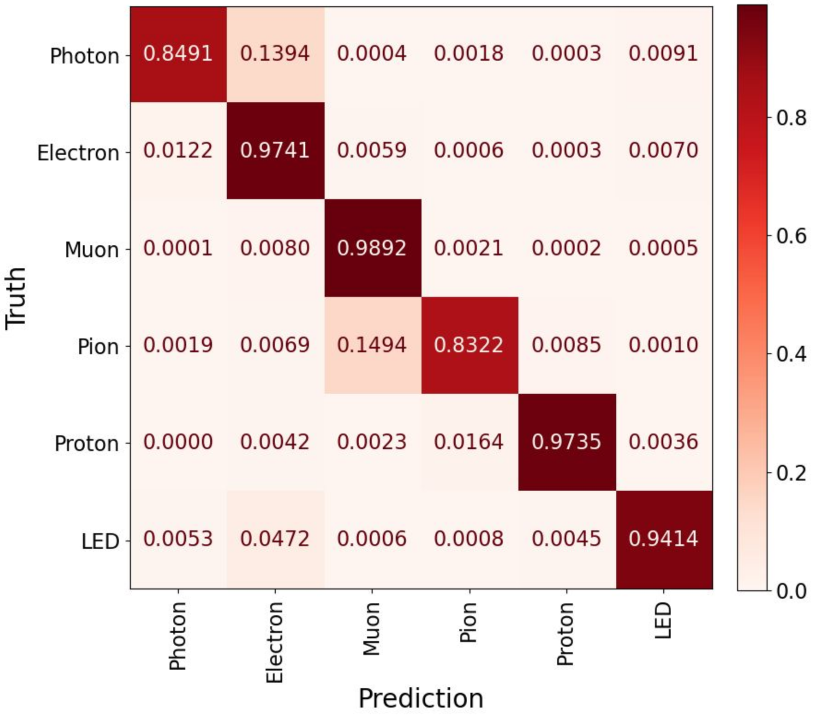


Example Sem Seg - PID

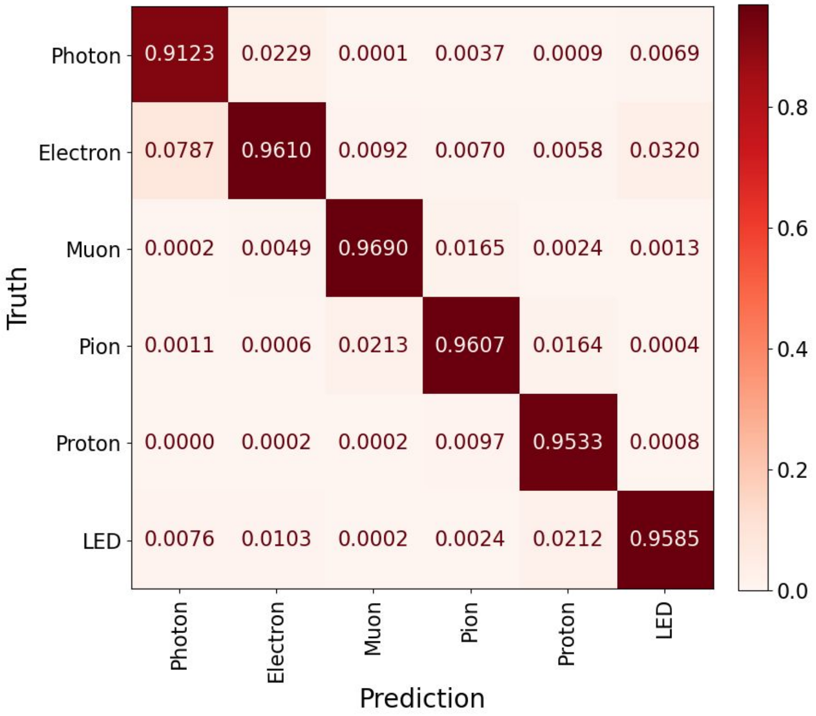


Sem Seg PID – 100 events

Normalized by Truth Row
(diag.=recall/efficiency)

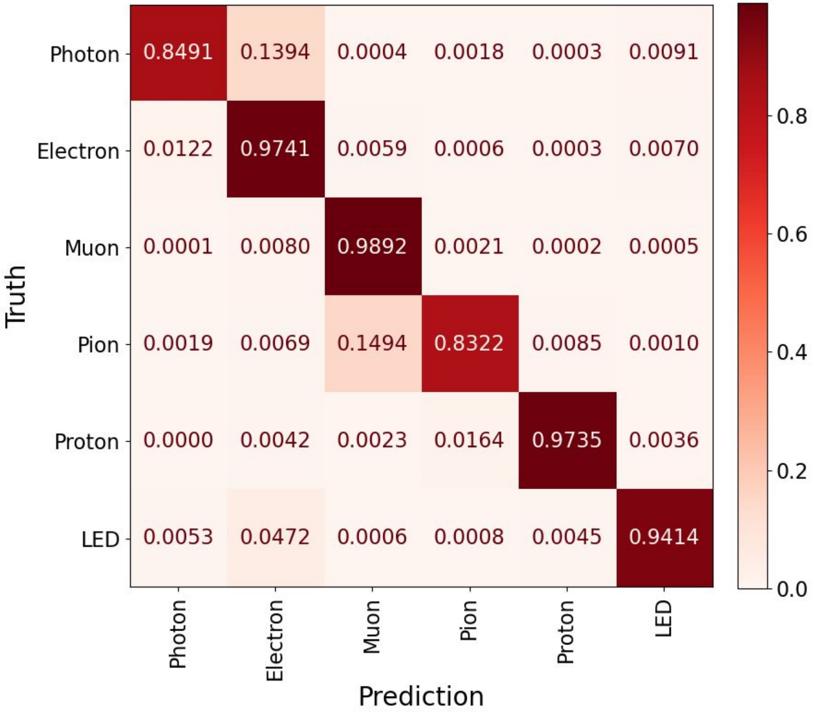


Normalized by Prediction Column
(diag.=precision)

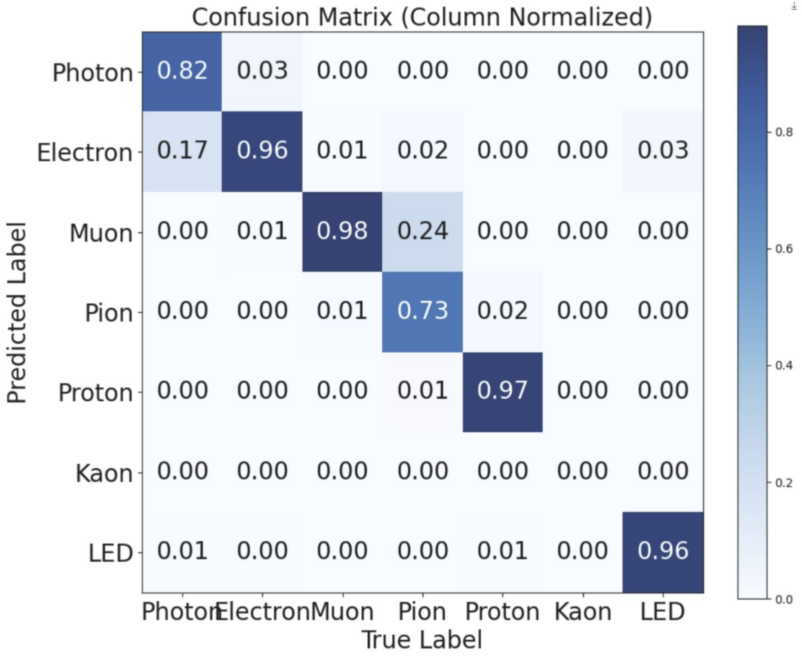


Sem Seg PID – UResNet Confusion

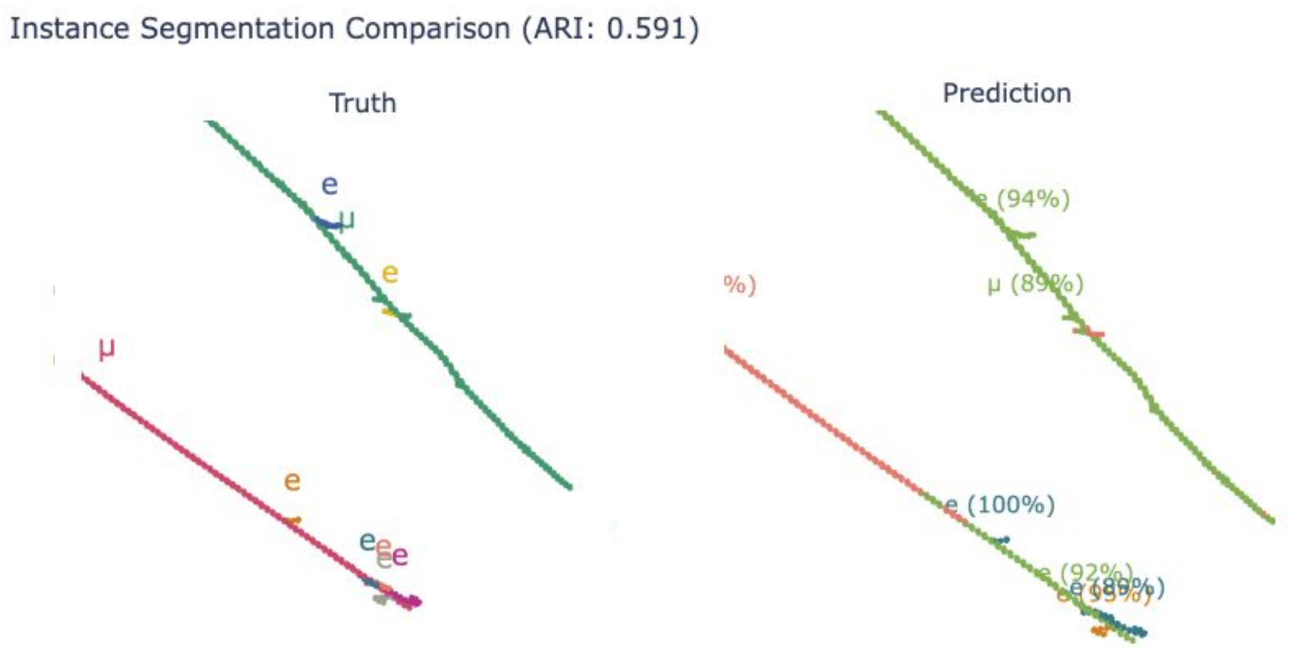
Normalized by Truth Row
(diag.=recall/efficiency)



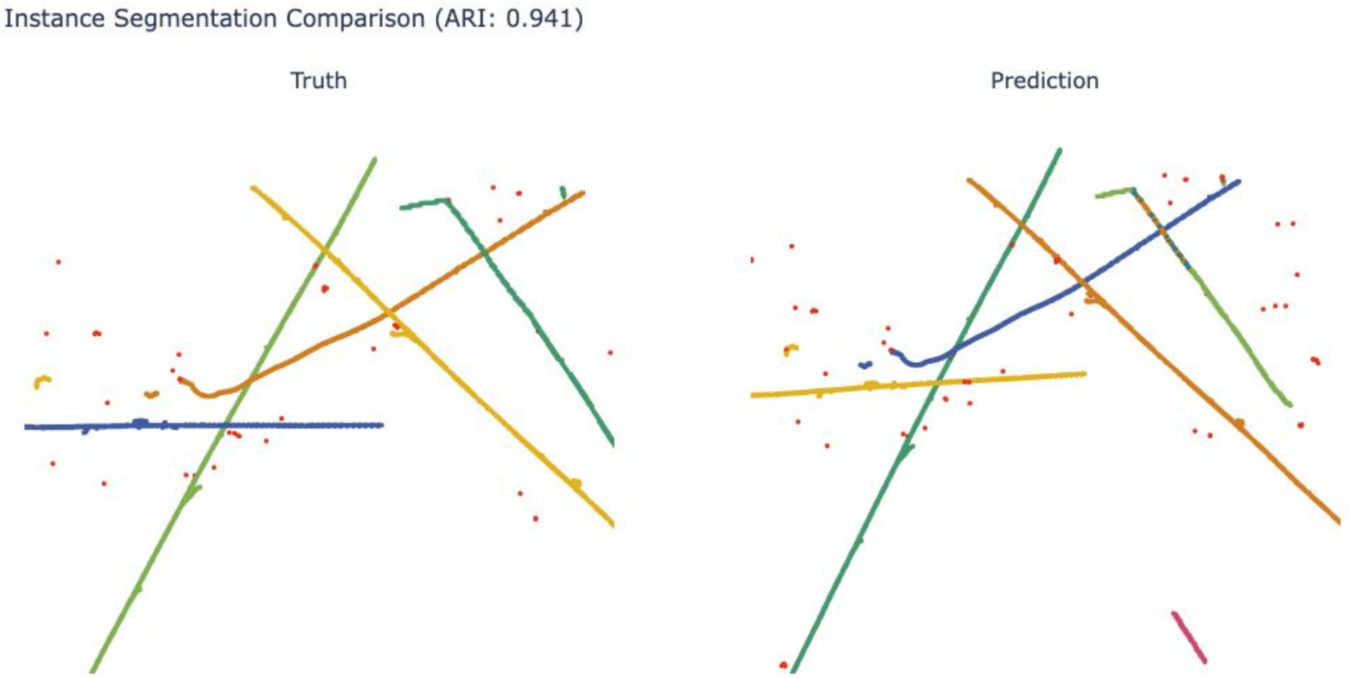
Normalized by Truth Column (sorry)
(diag.=recall)



Poor instance reconstruction – particle

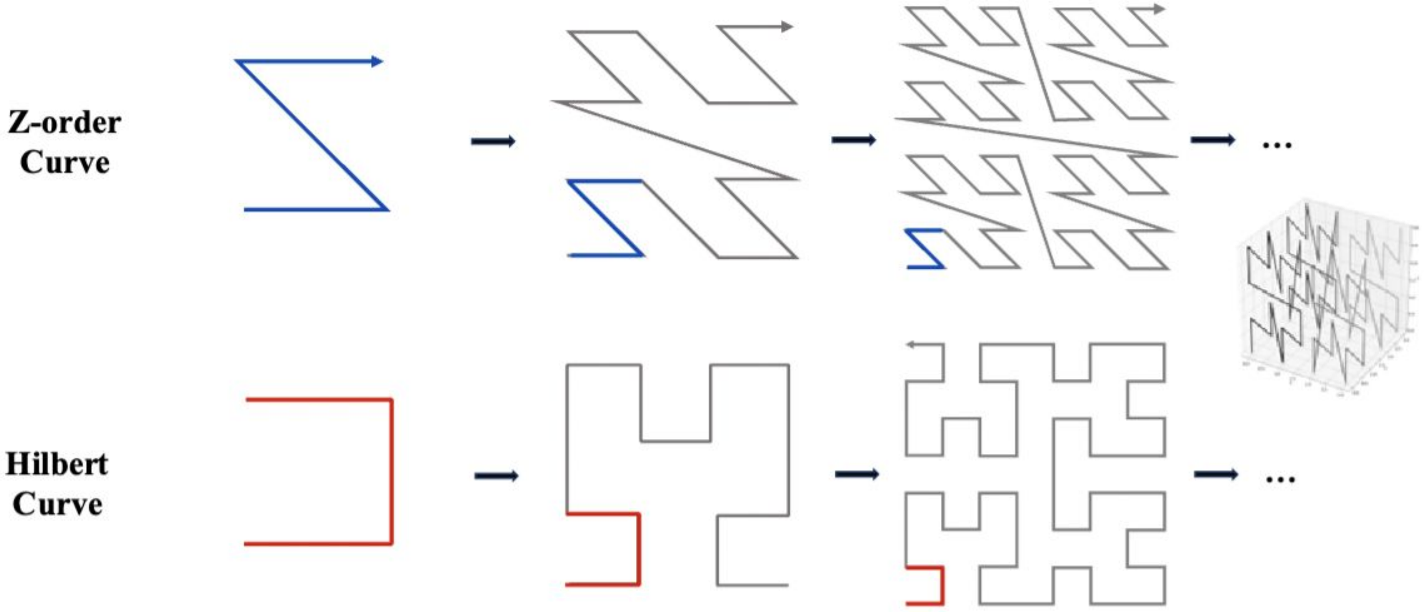


Poor instance reconstruction - interaction



Serialized Attention

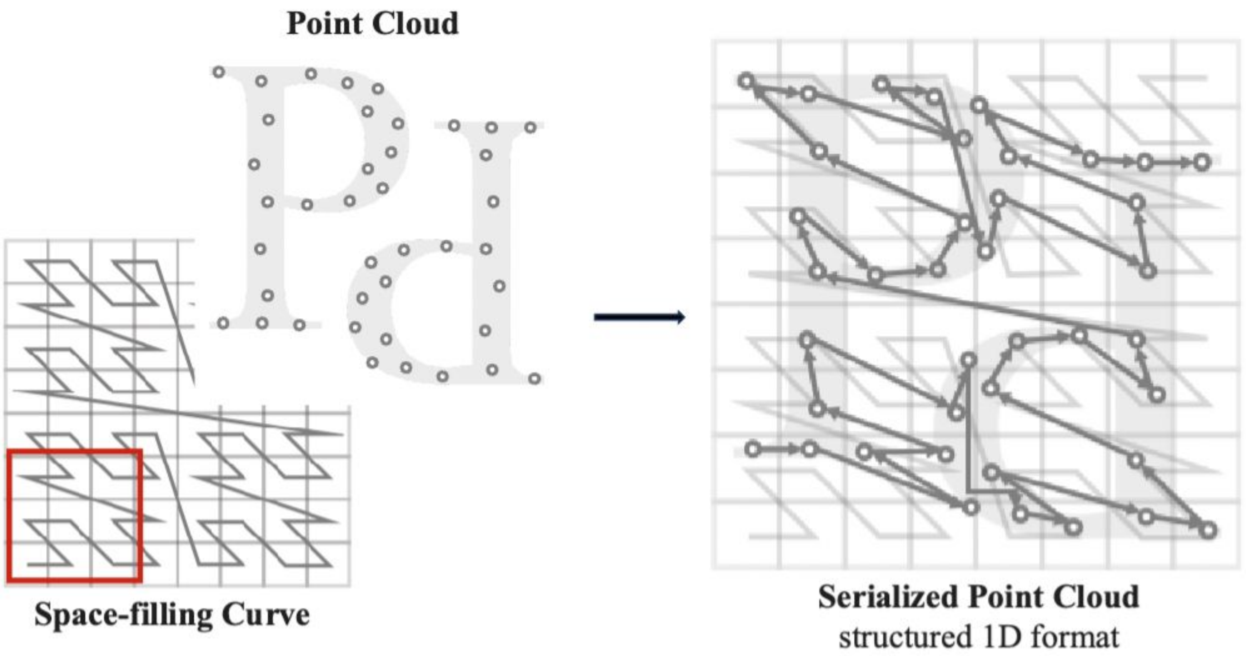
See Point Transformer V3 paper (arXiv:2312.10035) for more detail



From https://xywu.me/talk/20250609_Stanford_Spatial_Representations.pdf

Serialized Attention

See Point Transformer V3 paper ([arXiv:2312.10035](https://arxiv.org/abs/2312.10035)) for more detail

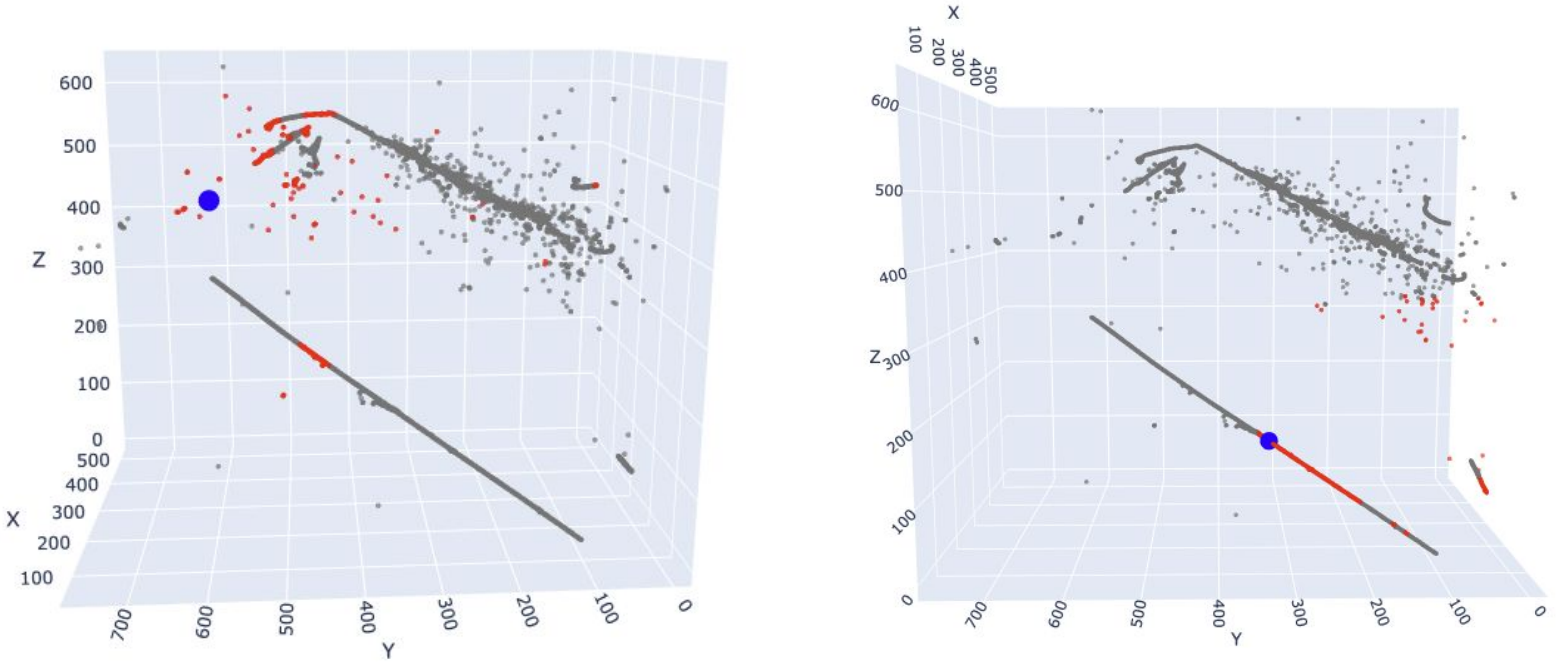


From https://xywu.me/talk/20250609_Stanford_Spatial_Representations.pdf

Serialized Attention – 256 voxel patches

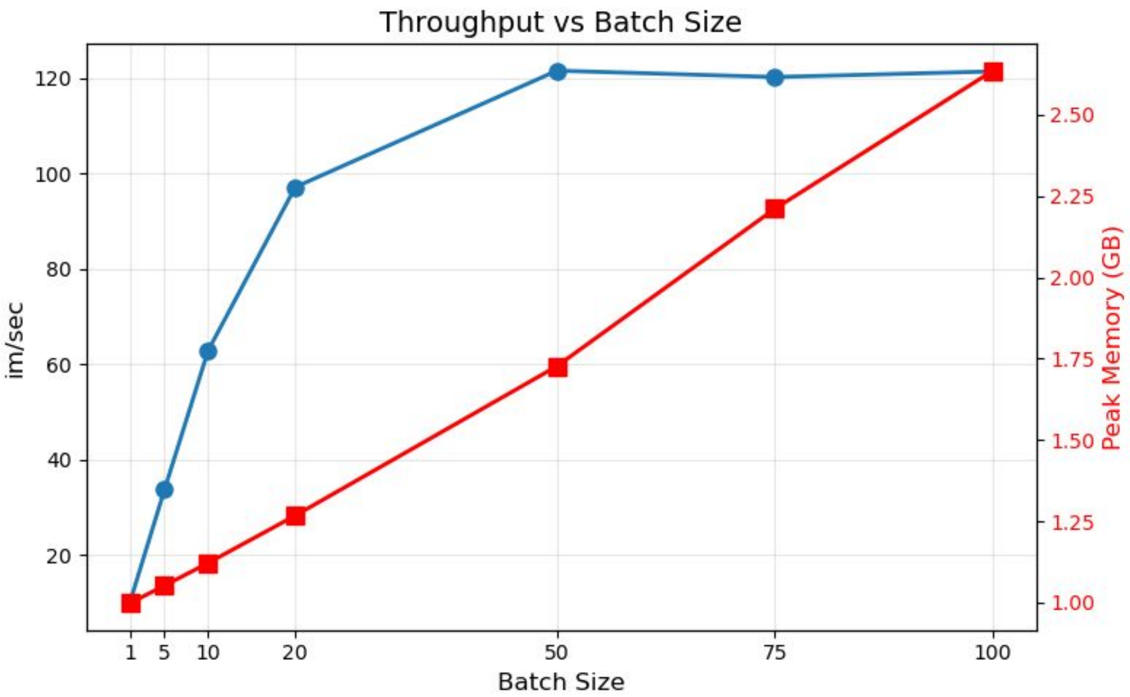
Receptive field at a single point at stage 0

- not perfect, but good enough with enough depth.



Scalability

Semantic Segmentation (measured on single A100)



Peak throughput:

8.3 ms/image

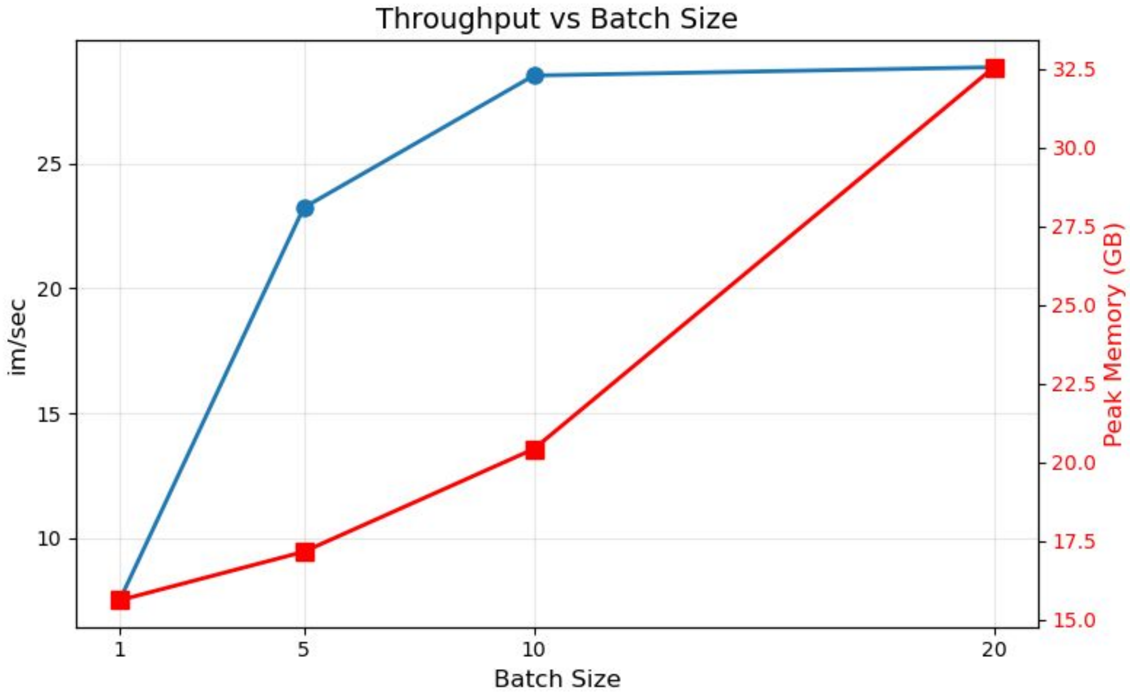
120 img/sec

Mem@peak:

~1.75 GB

Scalability

Instance/Panoptic Segmentation (model forward + NMS post-processing)



Peak throughput:

34.5 ms / image

29 img/sec

Mem@peak:

~20 GB

NMS post-processing is serial, but parallelized via multiprocessing

Scaling Model Params