Learning Image Representations Without Manual Annotations and Related Applications

Piotr Bojanowski, Senior Research Scientist Manager, FAIR, Meta

Ascona, Switzerland

October 31st 2023

Meta AI

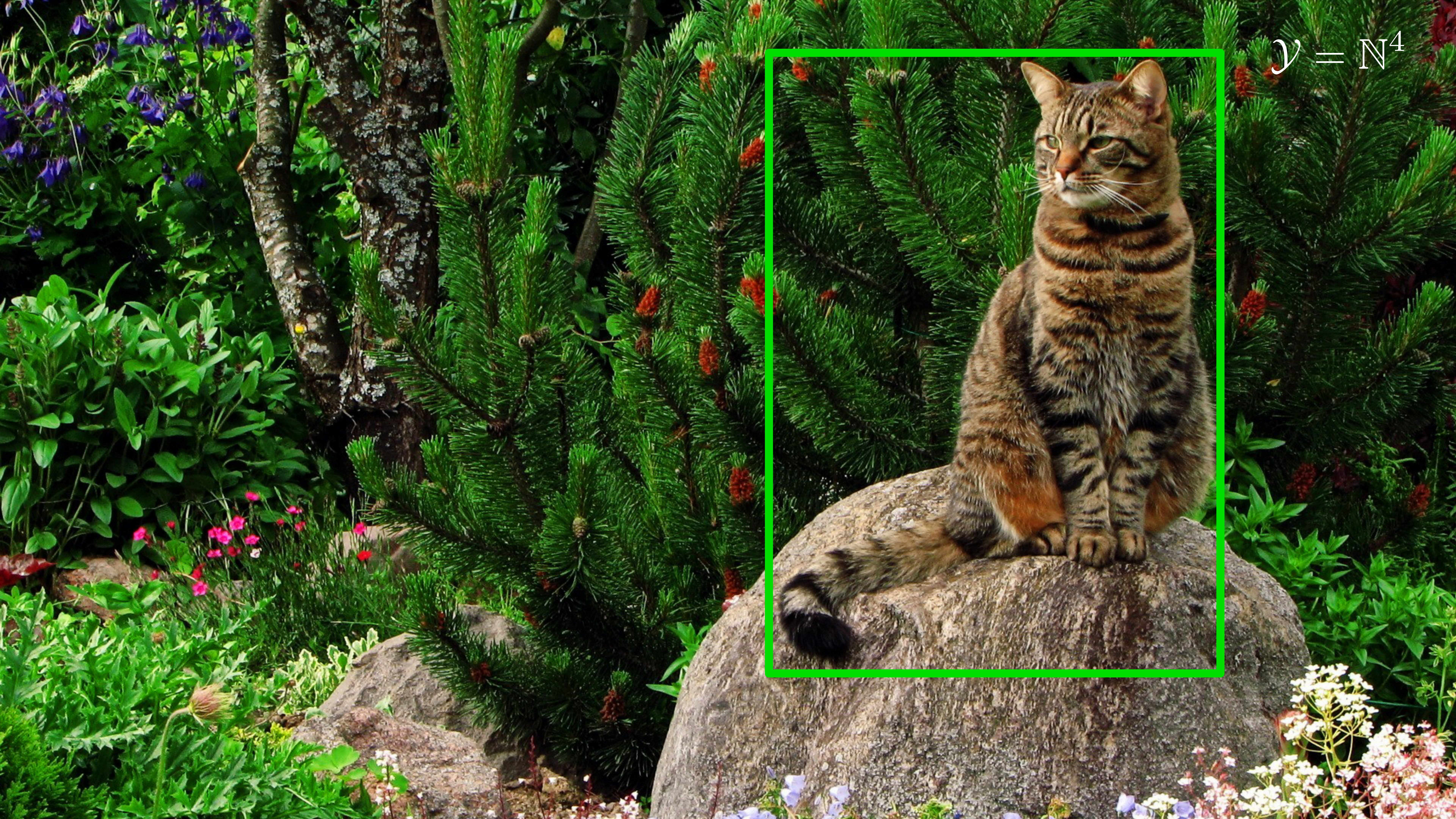

IMAGINED WITH AI IMAGINED WITH AI

Introduction

The Deep Learning Revolution

$$\mathcal{Y} = \{0, 1\}$$

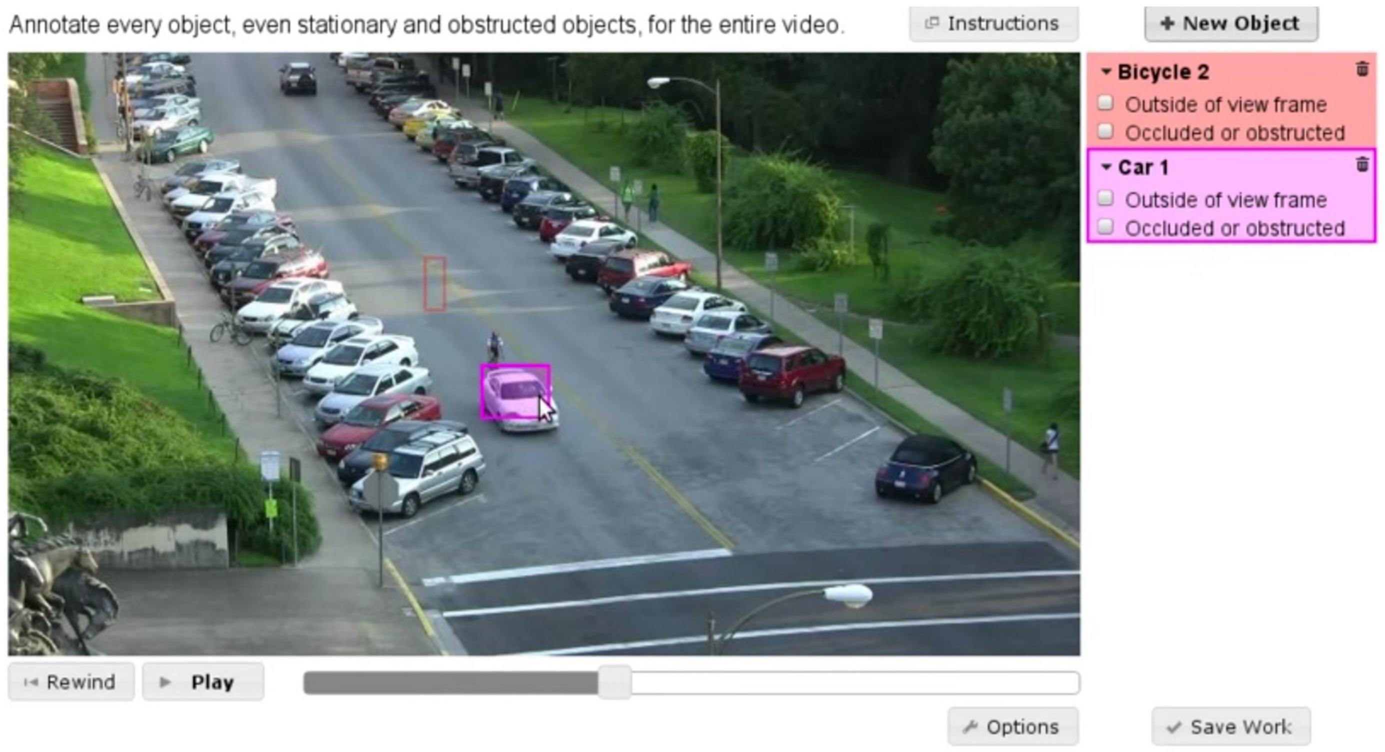
+1

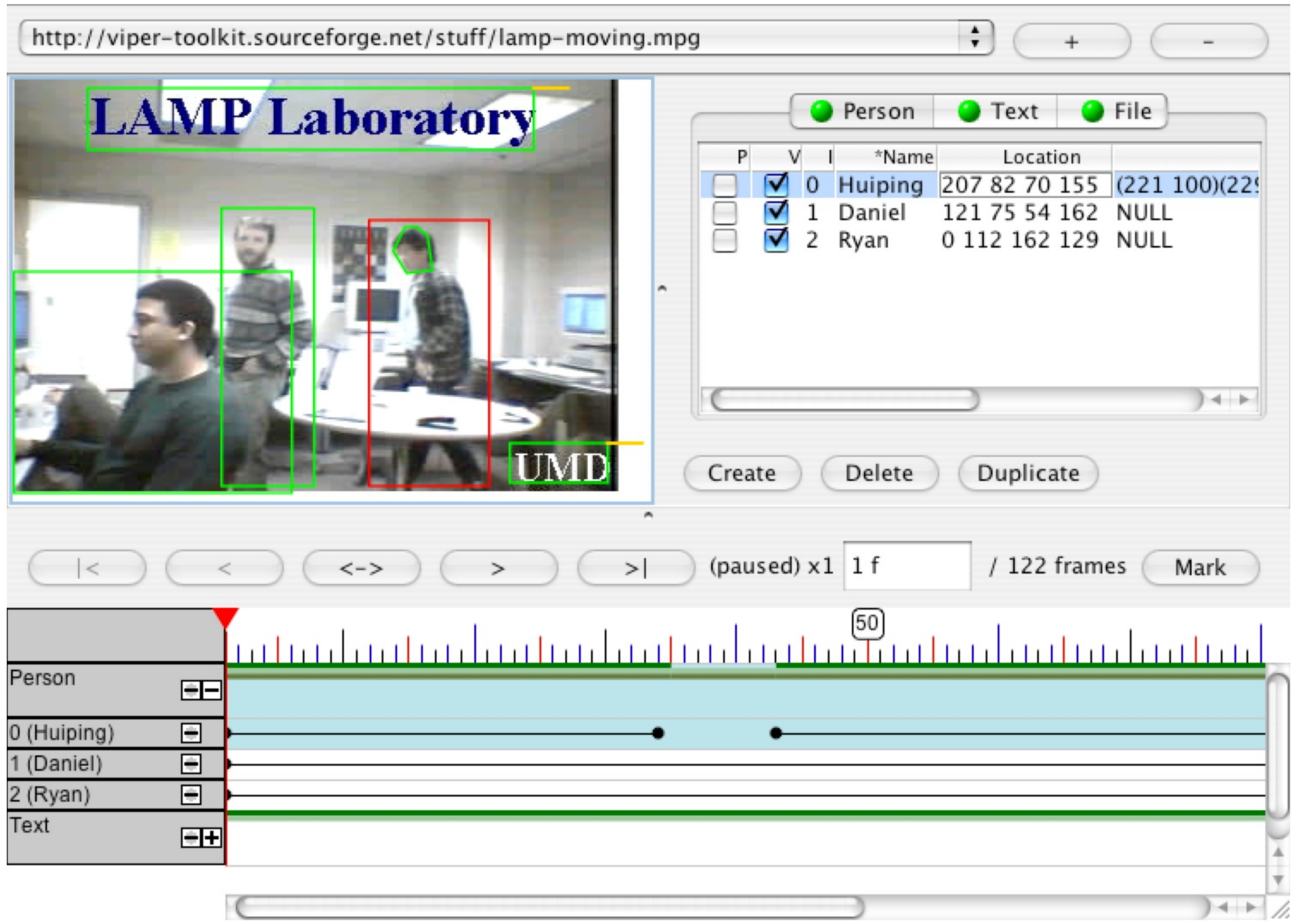

$\mathcal{Y} = \mathrm{N}^4$


$\mathcal{Y} = \{0,1\}^{w \times h}$

Manual Annotation of Video

- Manual annotation was* very tedious
  - Vatic (http://web.mit.edu/vondrick/vatic/)
  - Viper-GT (http://viper-toolkit.sourceforge.net/)
- Definition of an action is imprecise
  - Not as simple as physical extent of solids!
  - When does an action begin / end?

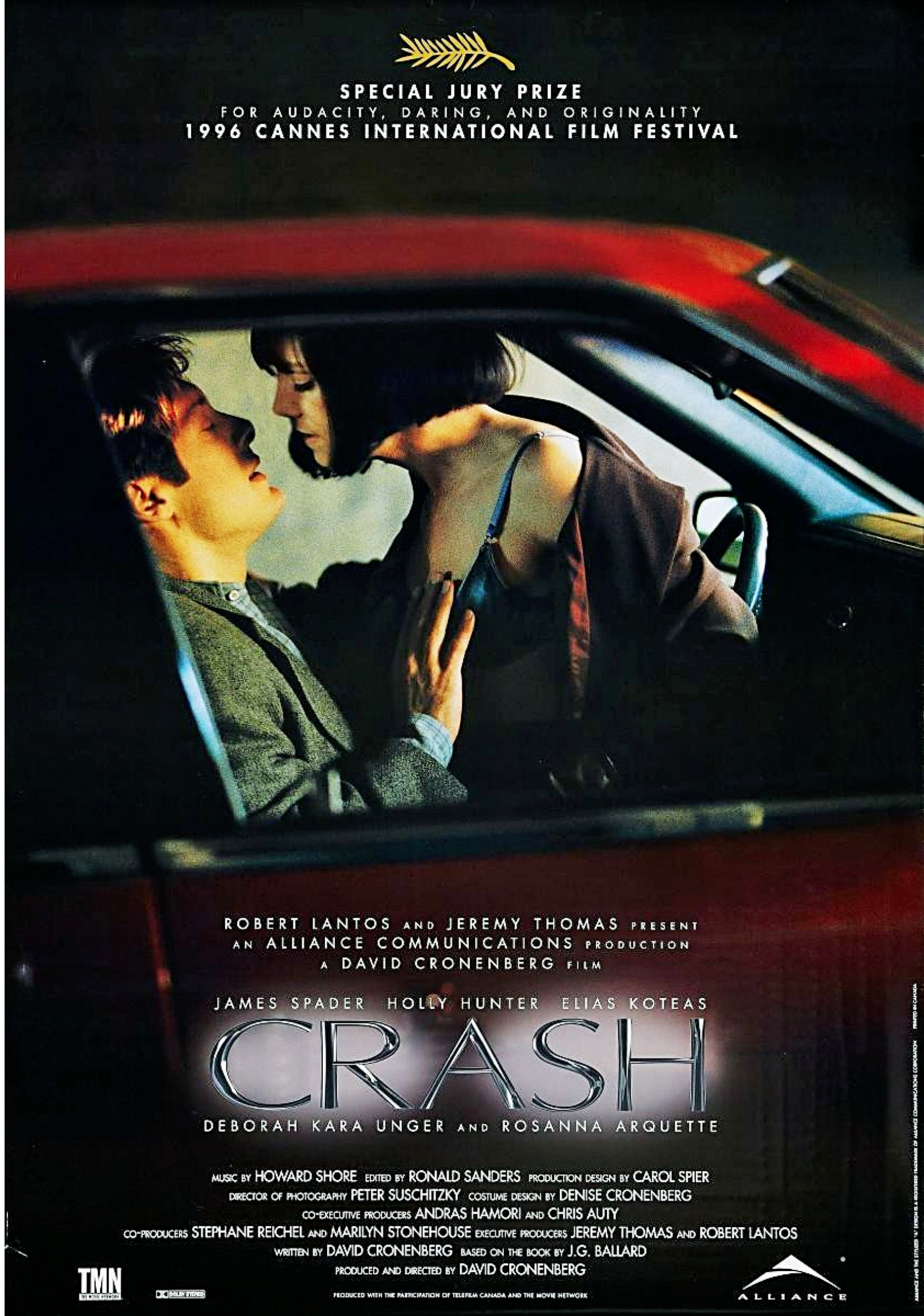




Manual annotations…

- Hollywood 2
  - Marcin Marszalek, Ivan Laptev, and Cordelia Schmid. "Actions in context." In *CVPR 2009.*
  - 810 + 884 videos
  - 12 actions
  - 69 Hollywood movies

- Hollywood 3 ?
  - Annotate all movies exhaustively
  - In charge of one of the movies

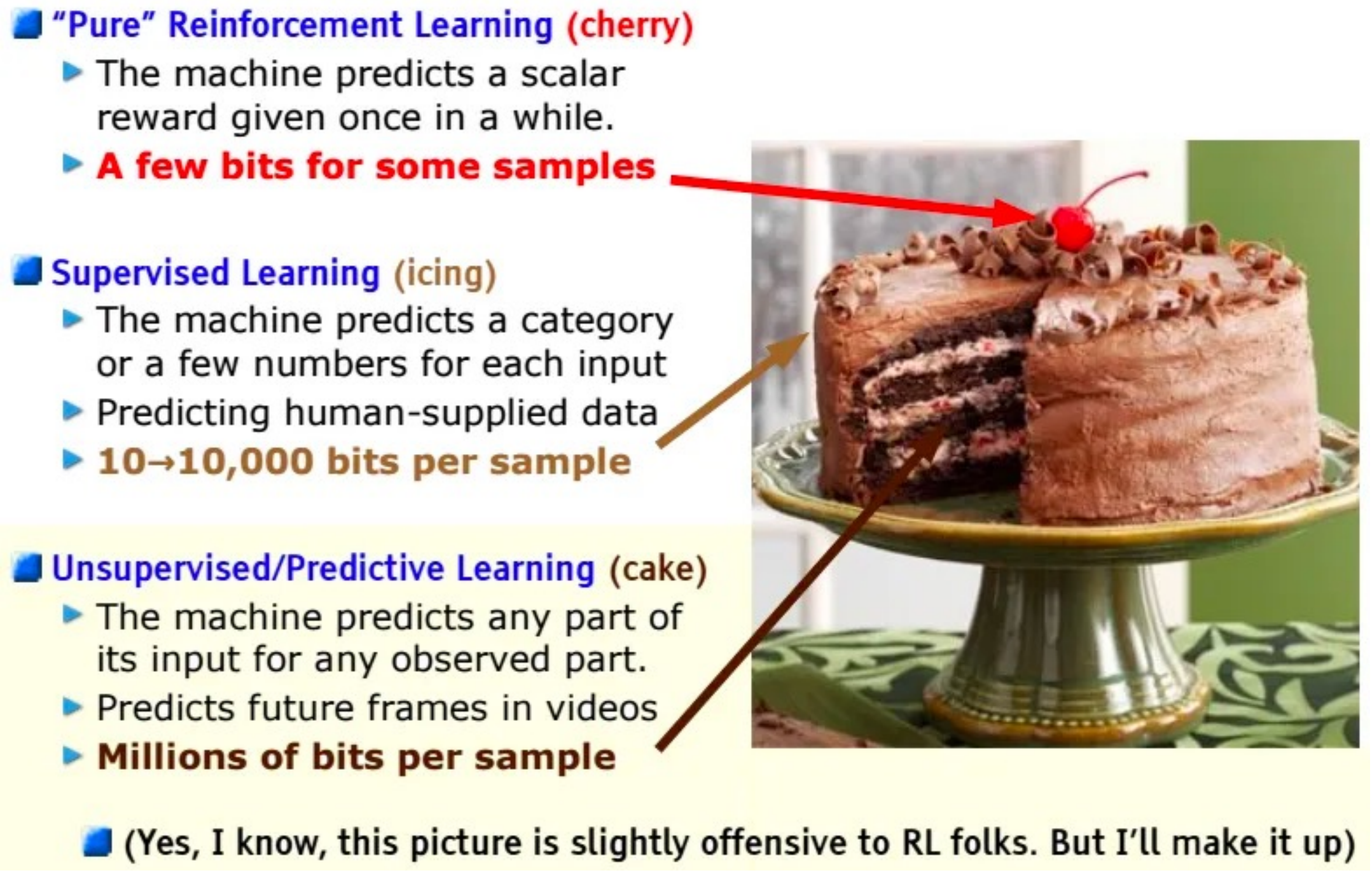


Manual Annotations

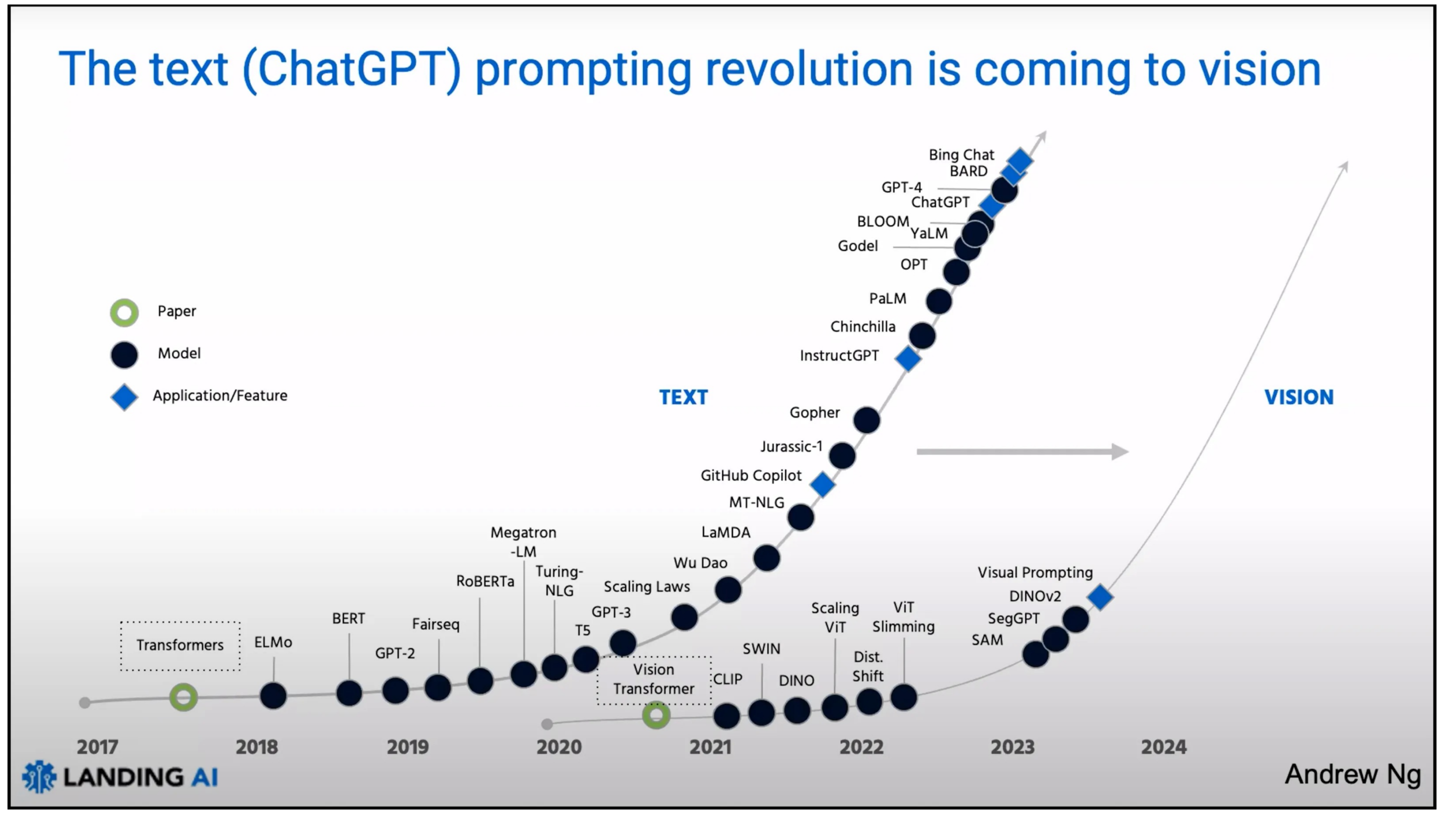
- Are expensive (if high quality)
- Are ambiguous
- Class definitions are not static
- Intractable with increasing complexity of the task

Baking the Cake

- Supervised Learning is needed!
- Unsupervised learning should do the heavy lifting
- Modern success of LLMs follows this exact recipe…
- Is the vision cake ready?



Yann LeCun, ~2016



Andrew Ng, 2023

Outline

1. Introduction
2. Large-Scale Self-Supervised Learning
3. Applications
4. Conclusion and Future Work

History of Self-Supervised Learning

~2015 a boom of creativity

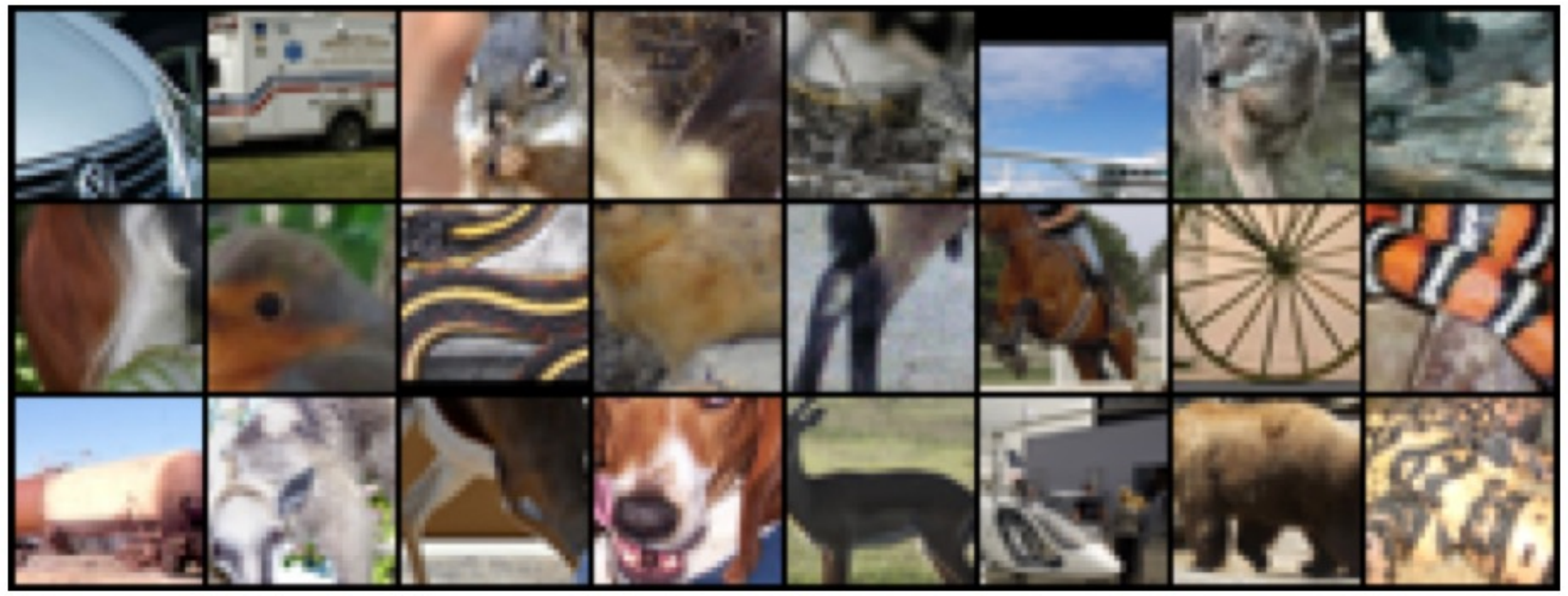
- Idea: reuse previous machinery
- Generate labels from raw data!

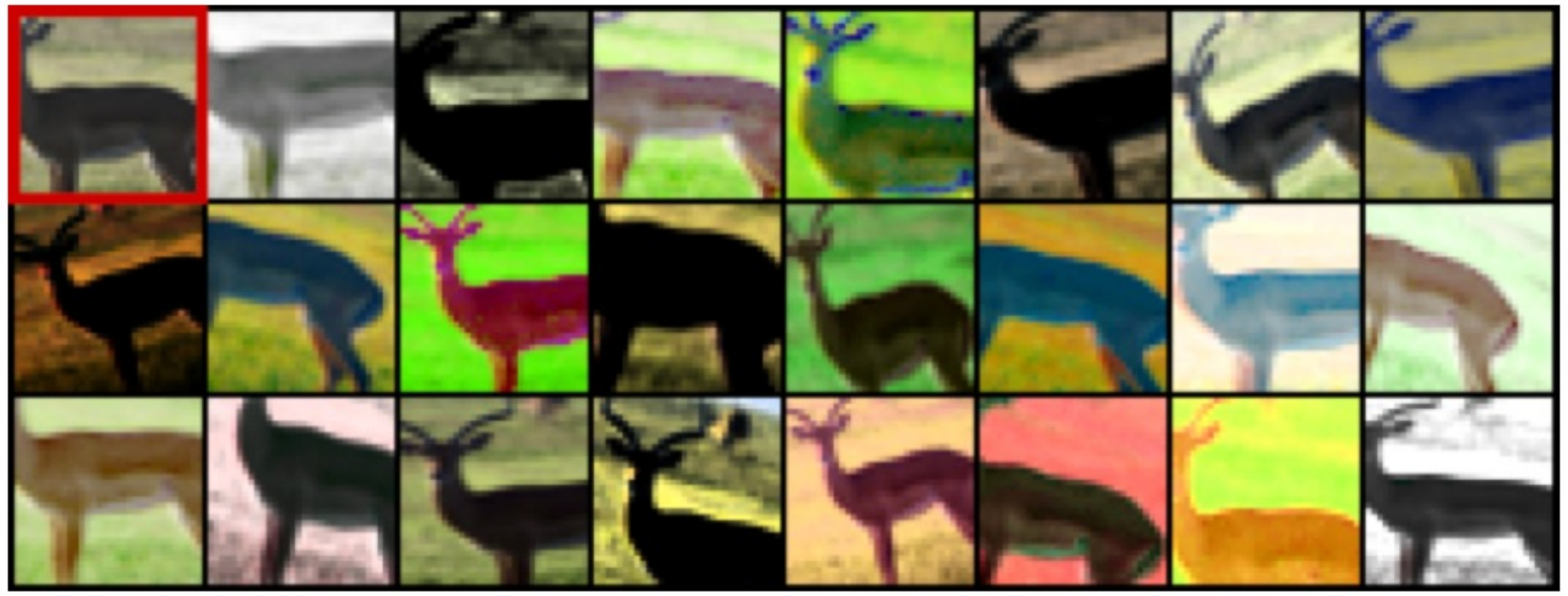
$$y : \mathcal{X} \to \mathcal{Y}$$
$$x \mapsto y(x)$$

- Then resort to good ol' fashioned Supervised Learning

$$\min_{\theta} \sum_{n=1}^{N} \ell \left(f_\theta(x_n), y(x_n) \right)$$

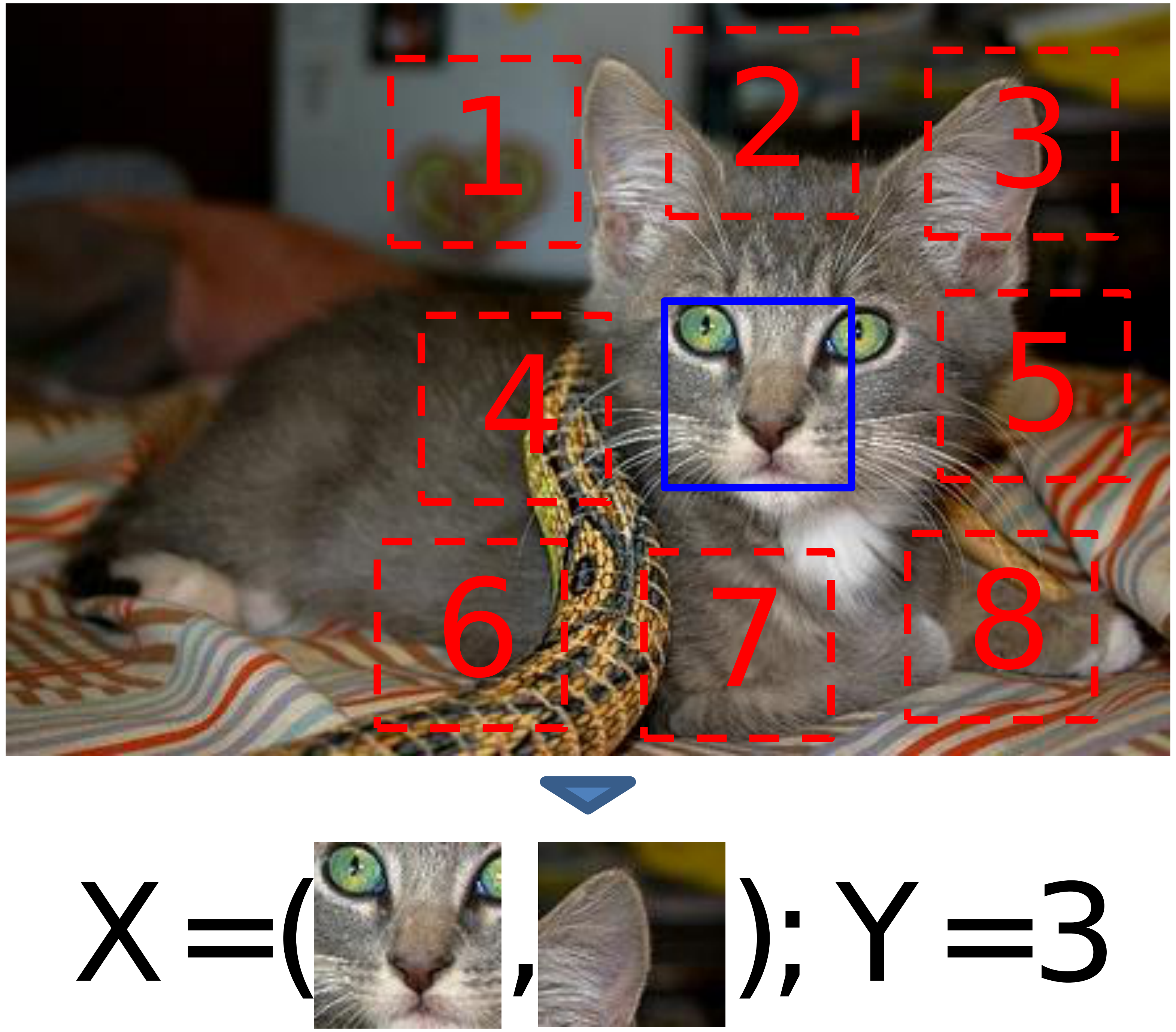
Instance Discrimination





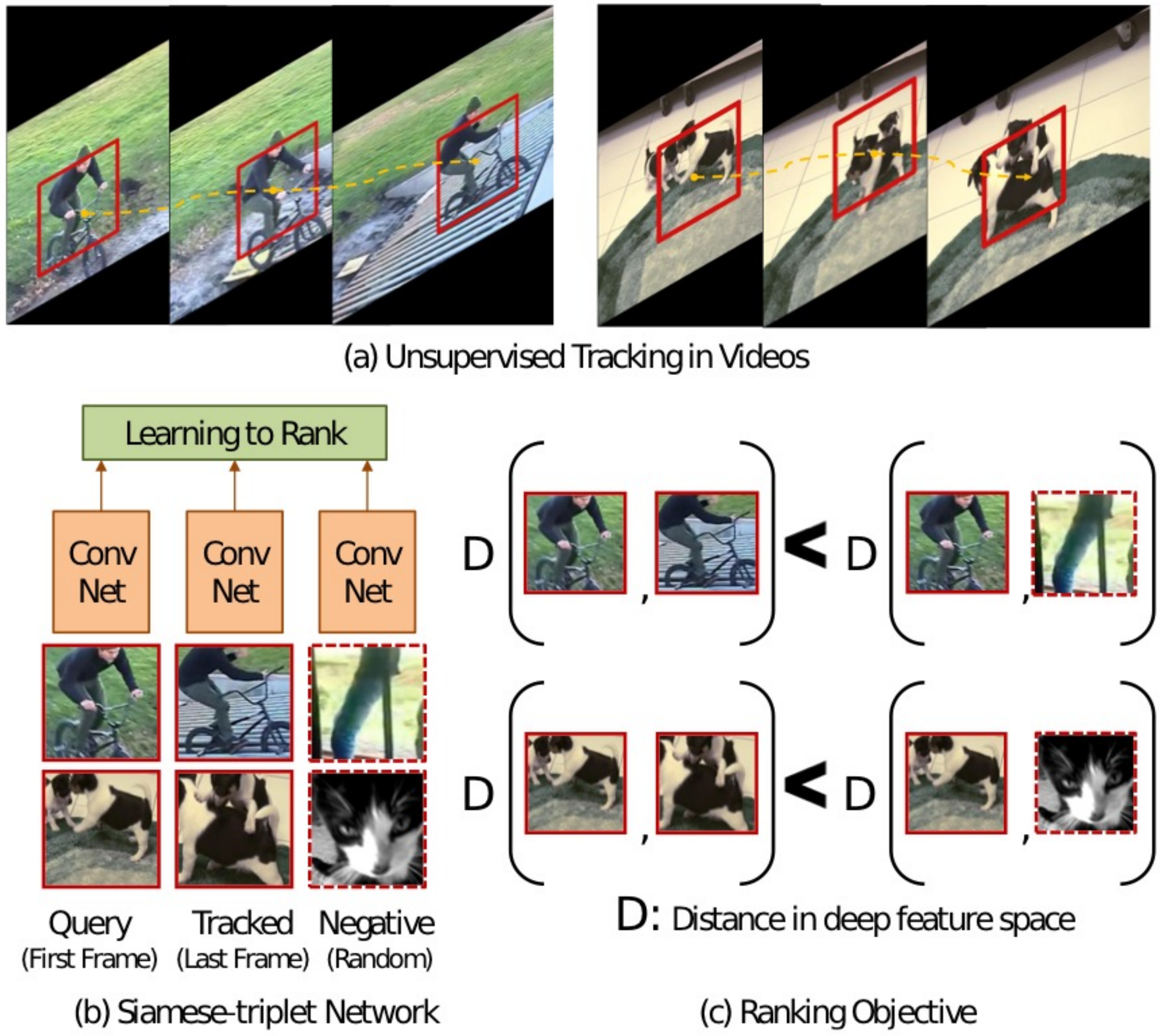
Dosovitskiy, Alexey, et al. "Discriminative unsupervised feature learning with convolutional neural networks." *Advances in neural information processing systems* 27 (2014).

Jigsaw Puzzles



Doersch, Carl, Abhinav Gupta, and Alexei A. Efros. "Unsupervised visual representation learning by context prediction." *Proceedings of the IEEE international conference on computer vision*. 2015.

Tracking



Wang, Xiaolong, and Abhinav Gupta. "Unsupervised learning of visual representations using videos." *Proceedings of the IEEE international conference on computer vision*. 2015.

Colorization

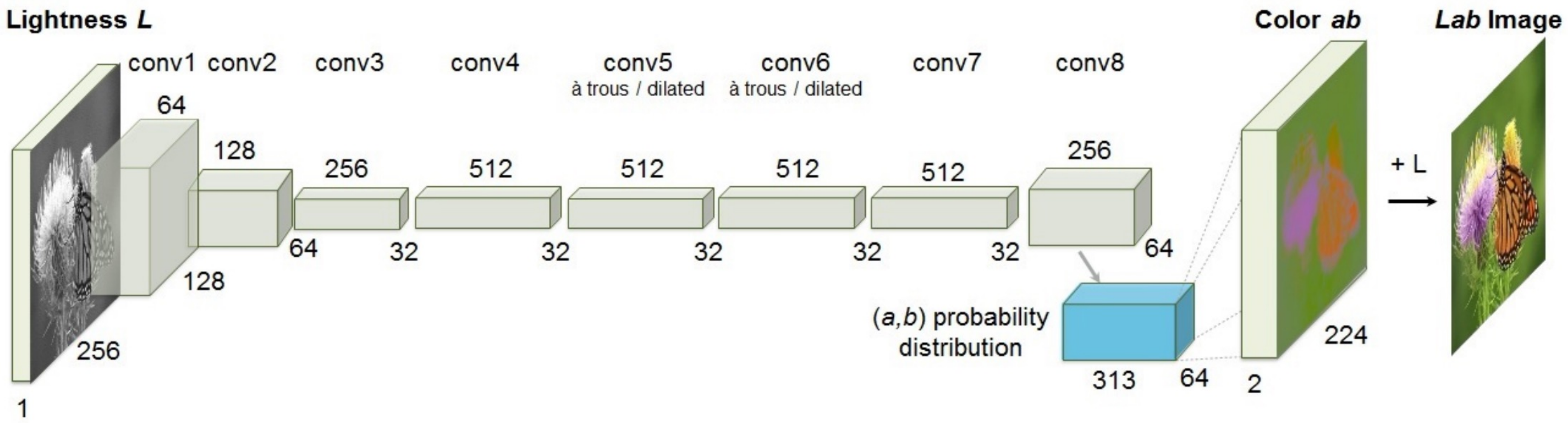
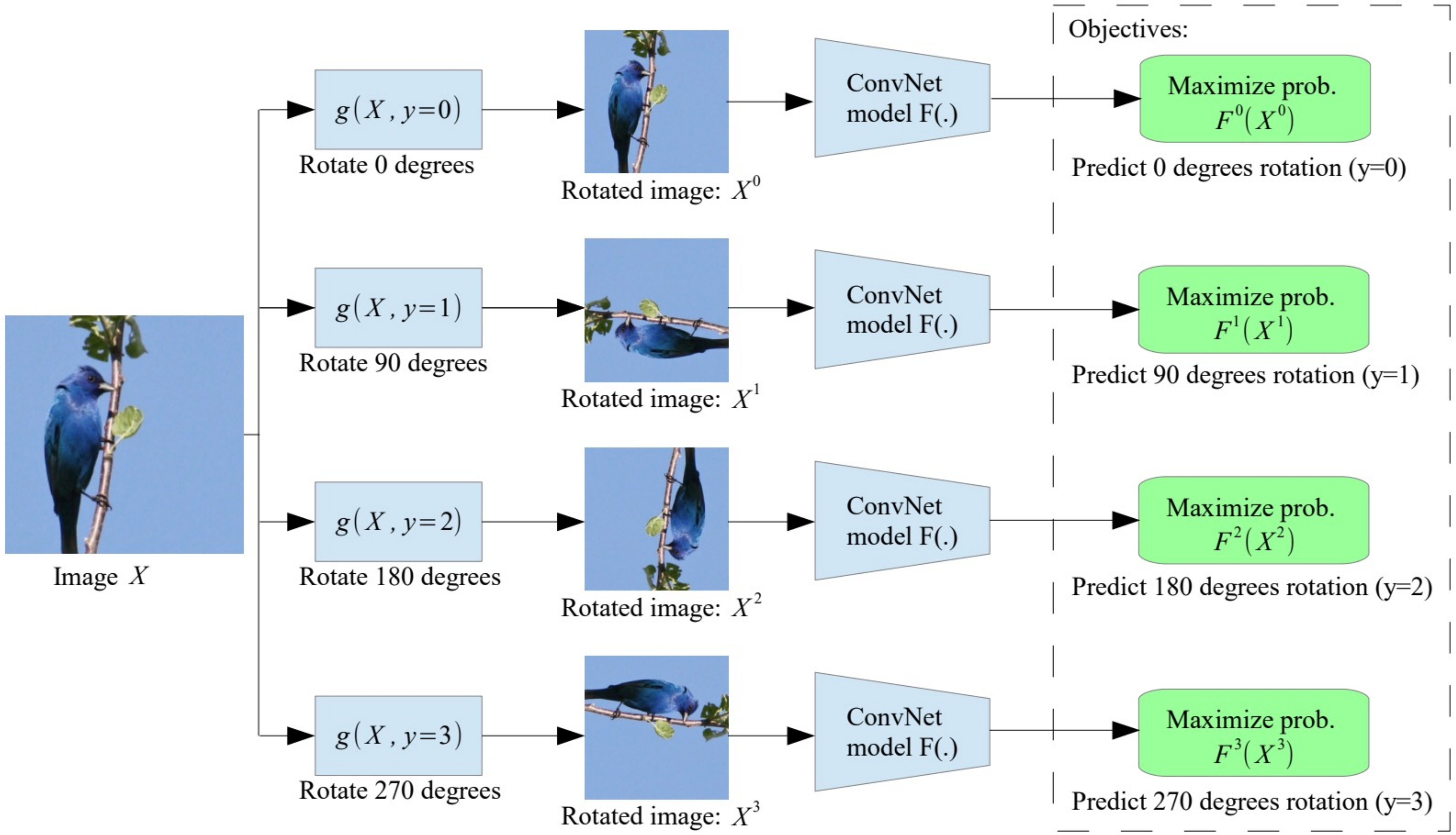


Fig. 2. Our network architecture. Each **conv** layer refers to a block of 2 or 3 repeated **conv** and **ReLU** layers, followed by a **BatchNorm** [30] layer. The net has no **pool** layers. All changes in resolution are achieved through spatial downsampling or upsampling between **conv** blocks.

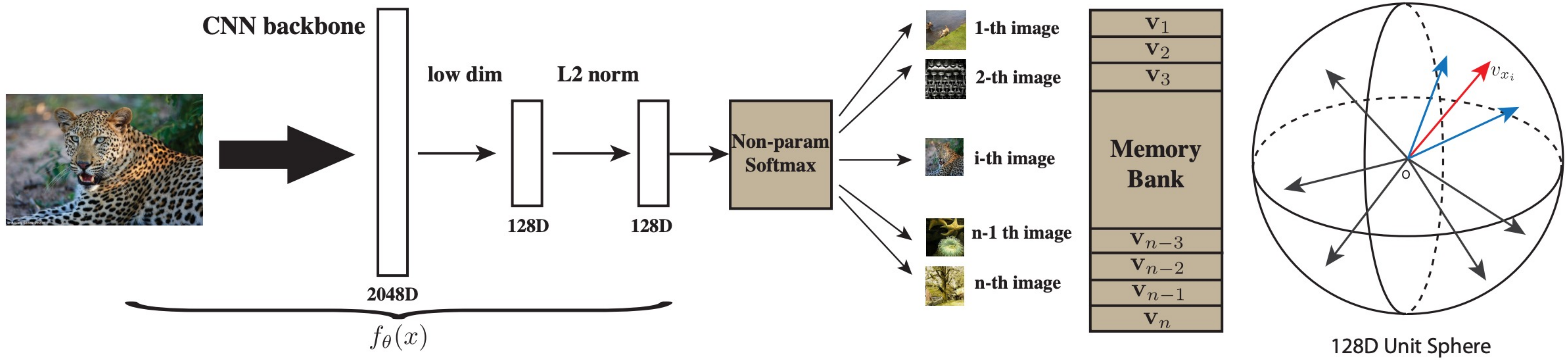
Zhang, Richard, Phillip Isola, and Alexei A. Efros. "Colorful image colorization." *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part III 14*. Springer International Publishing, 2016.

RotNet



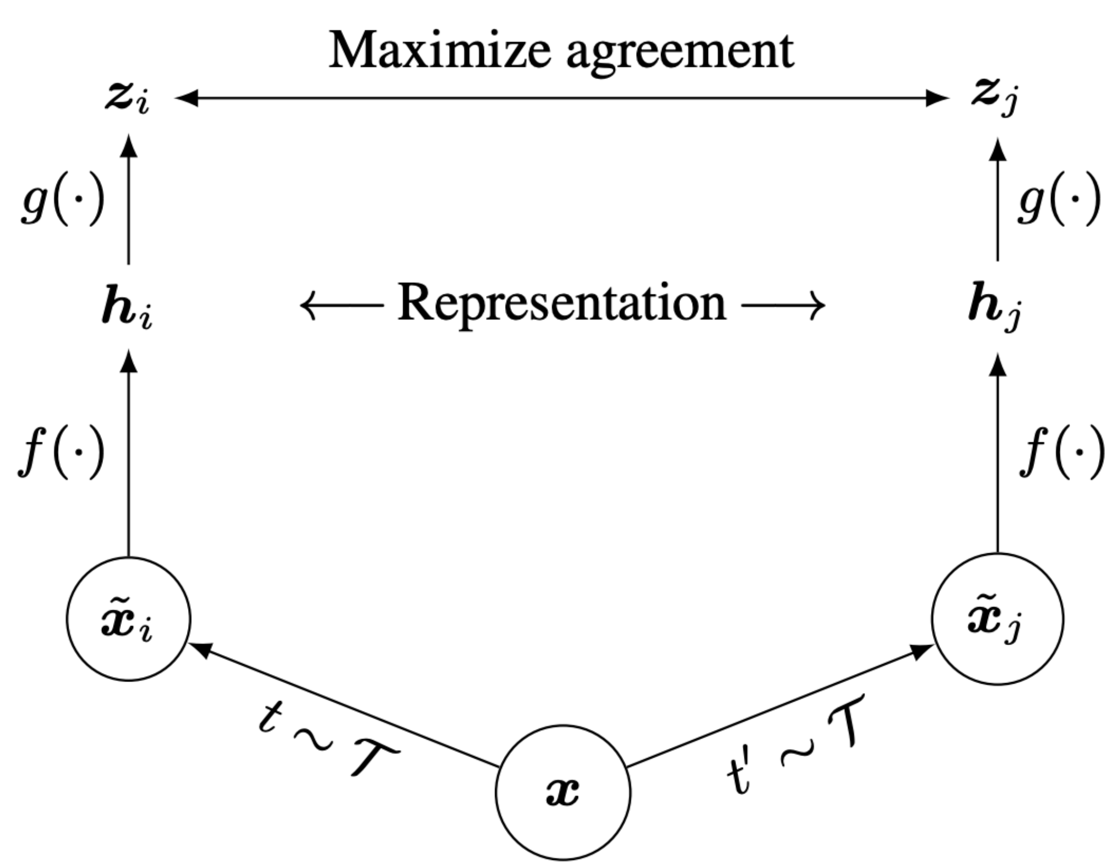
Gidaris, Spyros, Praveer Singh, and Nikos Komodakis. "Unsupervised Representation Learning by Predicting Image Rotations." *International Conference on Learning Representations*. 2018.

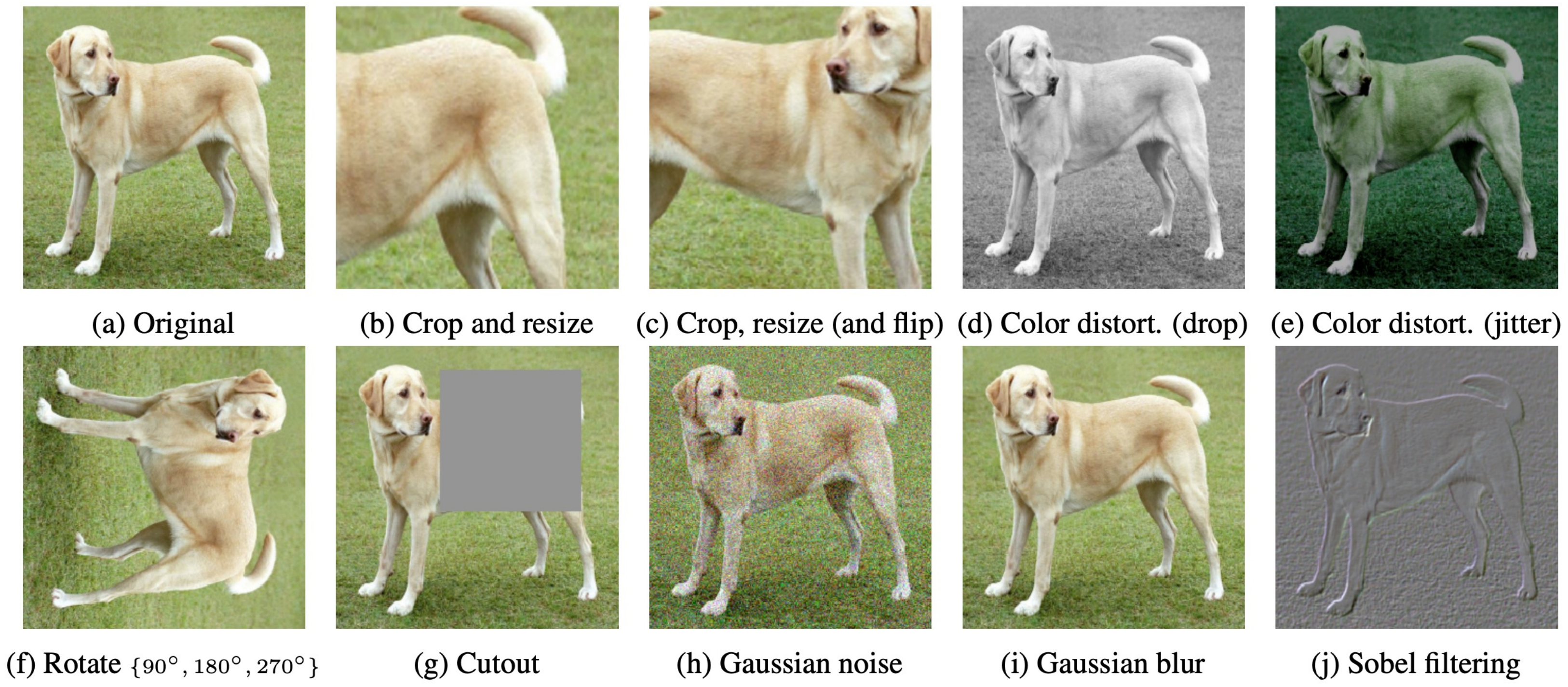
Non-parametric instance discrimination



Wu, Zhirong, et al. "Unsupervised feature learning via non-parametric instance discrimination." *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2018.

Joint-Embedding Architectures - SimCLR





$$\min -\frac{1}{N}\sum_{n=1}^{N}\log\frac{e^{z_{ni}^\top z_{nj}}}{e^{z_{ni}^\top z_{nj}}+\sum_{n'\neq n}e^{z_{ni}^\top z_{n'}}}$$

Chen, Ting, et al. "A simple framework for contrastive learning of visual representations." *International conference on machine learning*. PMLR, 2020.

CPCv2, SELA, MoCo, PIRL, SimCLR, MoCov2, PCL, BYOL, Barlow Twins, SimCLRv2, NN-CLR, VicReg...

Large-Scale Self-Supervised Learning








Meta










Clustering-Inspired SSL

Discriminative clustering

- Group samples and train a discriminative model of groups
- Generative / discriminative clustering

$$\min_{Y,C} \frac{1}{N} \|X - YC\|_F^2$$

$$\min_{Y,W} \frac{1}{N} \|XW - Y\|_F^2 + \lambda \|W\|_F^2$$

- Can we train a CNN with this objective?

$$\min_{Y,\theta} \frac{1}{N} \|f_\theta(X) - Y\|_F^2$$

Bach, F., & Harchaoui, Z. (2007). Diffrac: a discriminative and flexible framework for clustering. *Advances in Neural Information Processing Systems*, *20*.

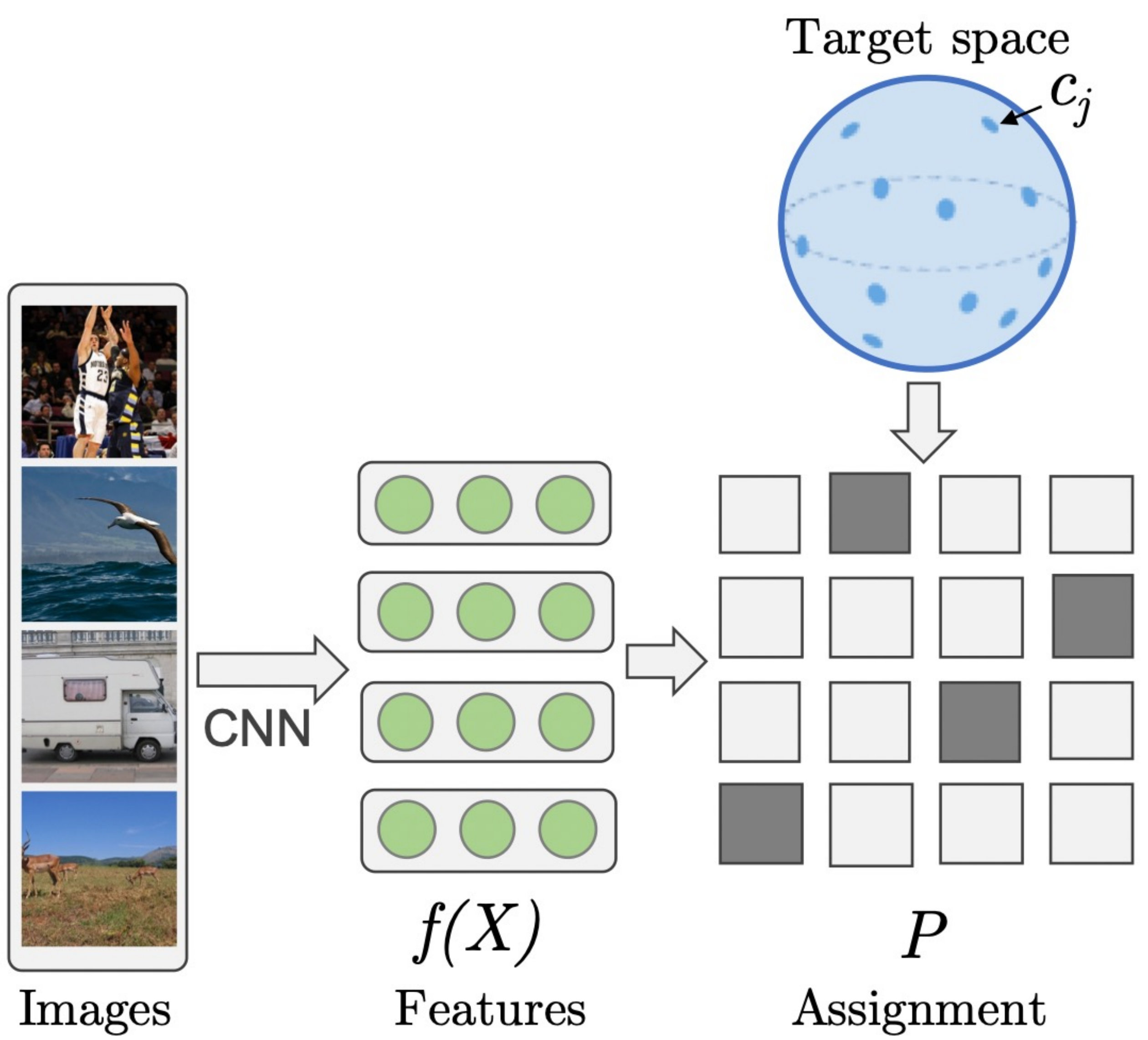
Joulin, A., Bach, F., & Ponce, J. (2010, June). Discriminative clustering for image co-segmentation. In *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*(pp. 1943-1950). IEEE.

NAT

- Main issue : avoid trivial solutions
- Solution : constrain Y as P x C

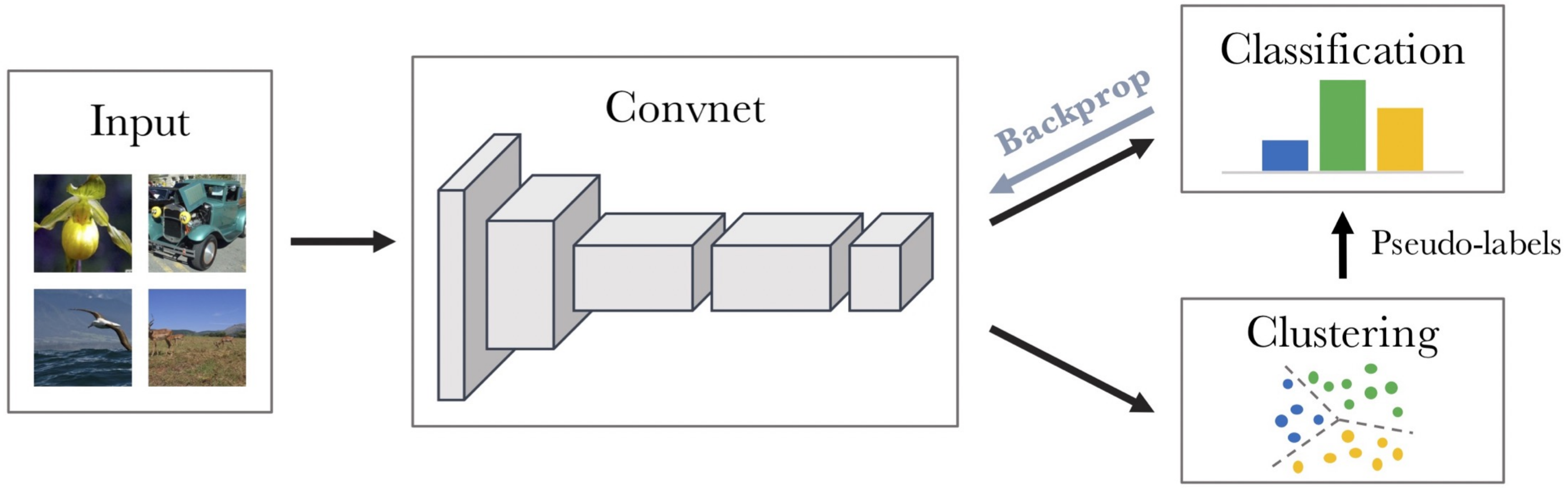
$$\min_{\theta,P} \|f_\theta(X) - PC\|_F^2$$

- C defines the neighborhood a priori (in Nxd)
- P is a NxN permutation matrix
- Used uniform distribution on a sphere for C



DeepCluster

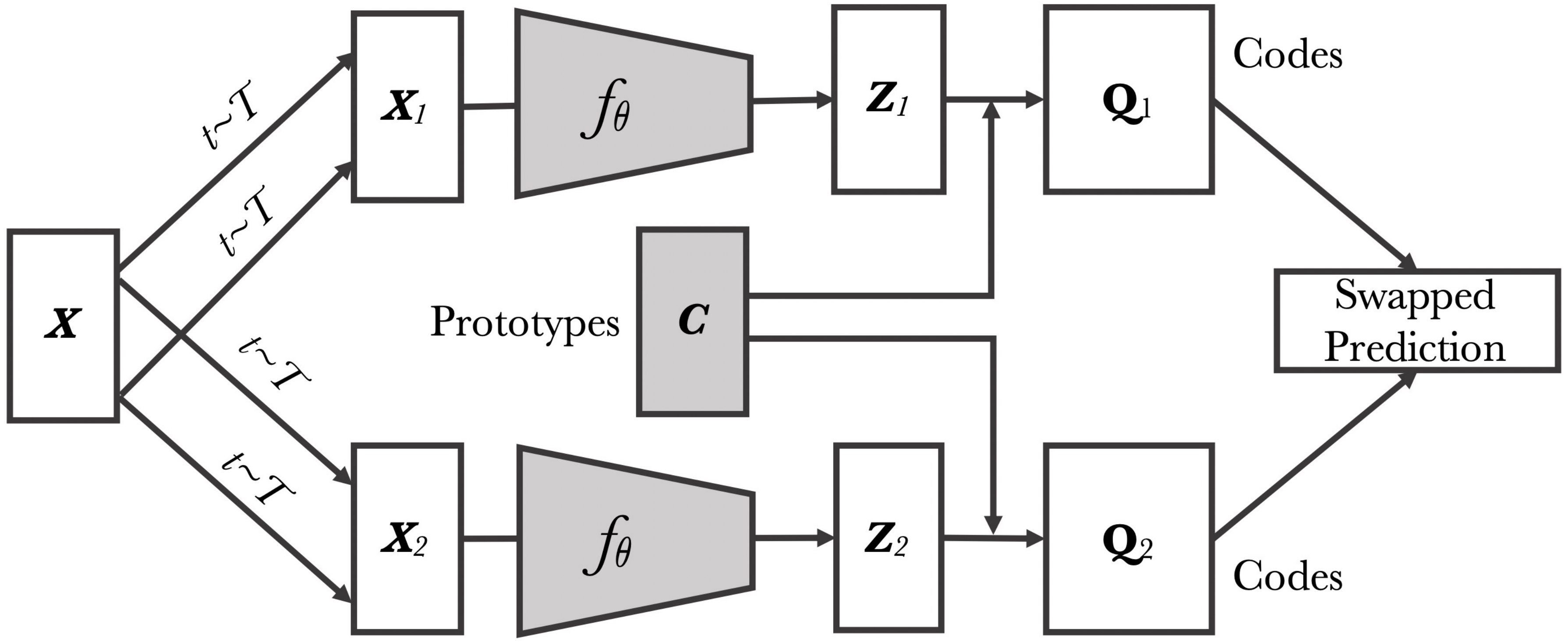
- Stochastic optimization of permutation matrices is hard
- Define a simpler algorithm!
- Key observation: a random AlexNet provides decent features
- Cluster initial features, using k-Means
- Treat the cluster assignments as labels and train with logistic loss
- Iterate…



SwAV

- On-line version of DeepCluster
- Prototypes are the equivalent of centroids
- Better use of codebook using assignment

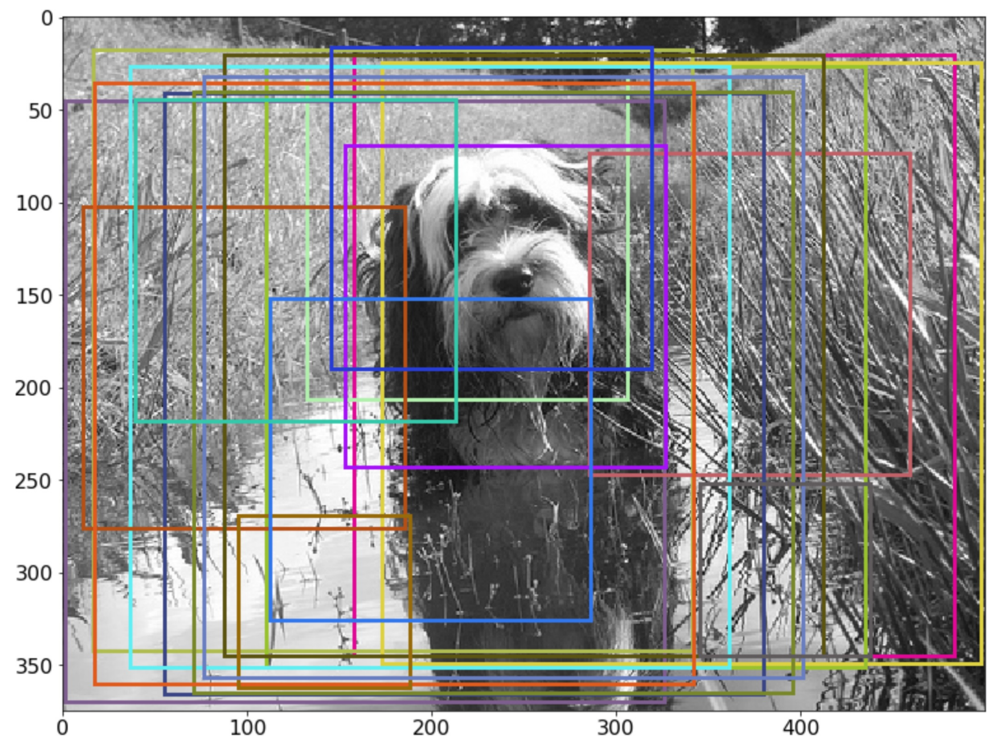
- Soft assignments instead of k-means
- Not actually solving assignment, just a few steps of SK…

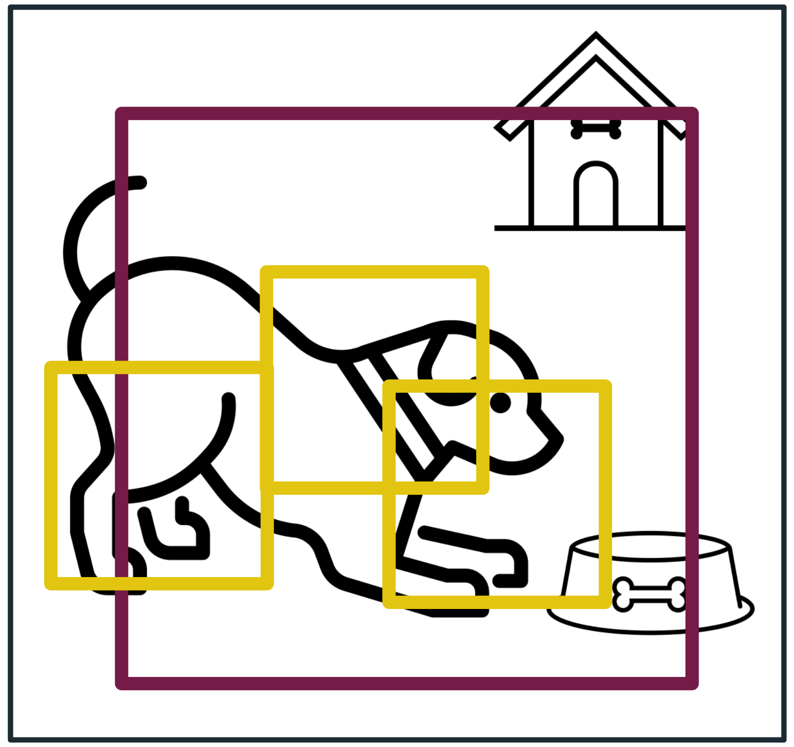


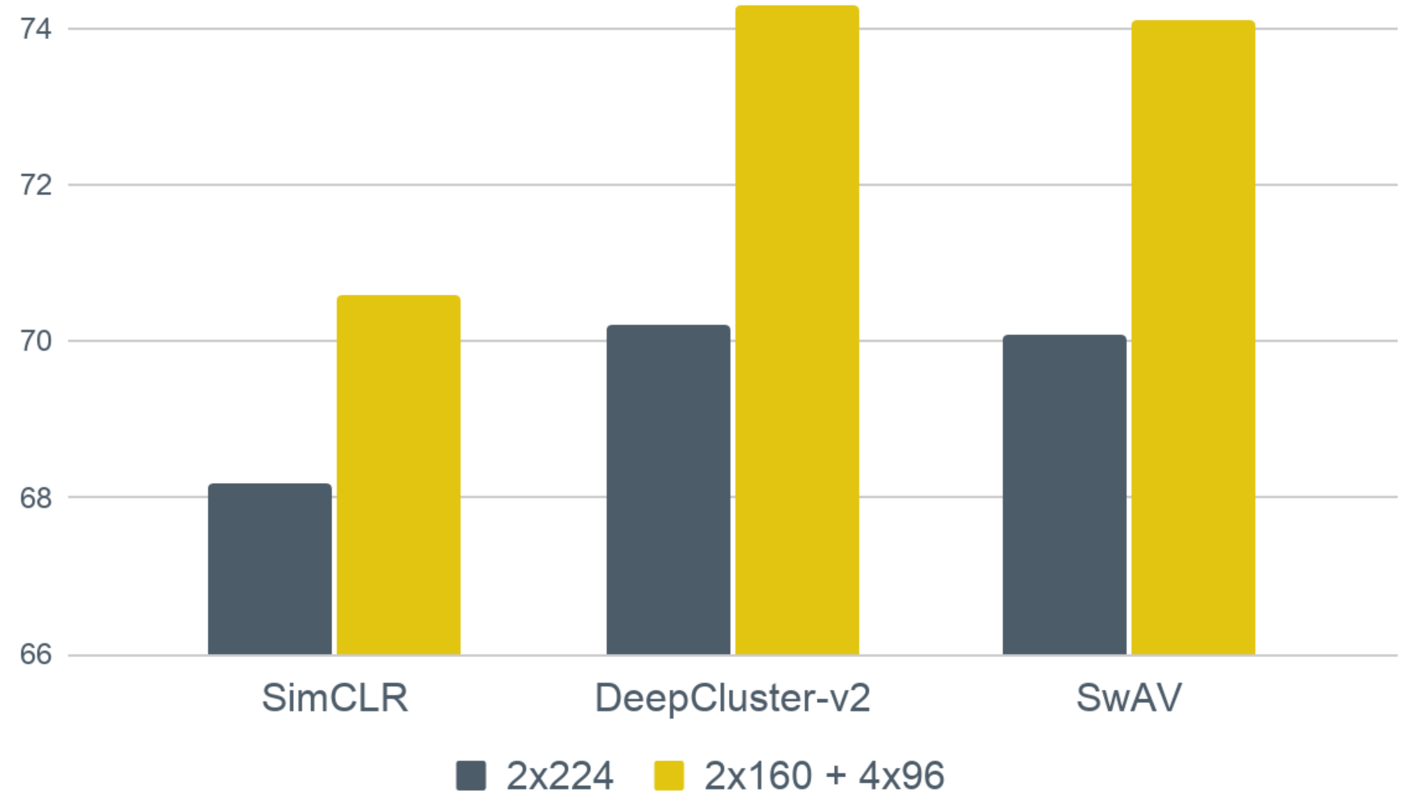
Mutlicrop

- Working on custom Cropping function
- Training with smaller images : speed up and small perf loss
- Mixing scales and resolutions was bringing non trivial boost

Courtesy of Mathilde Caron



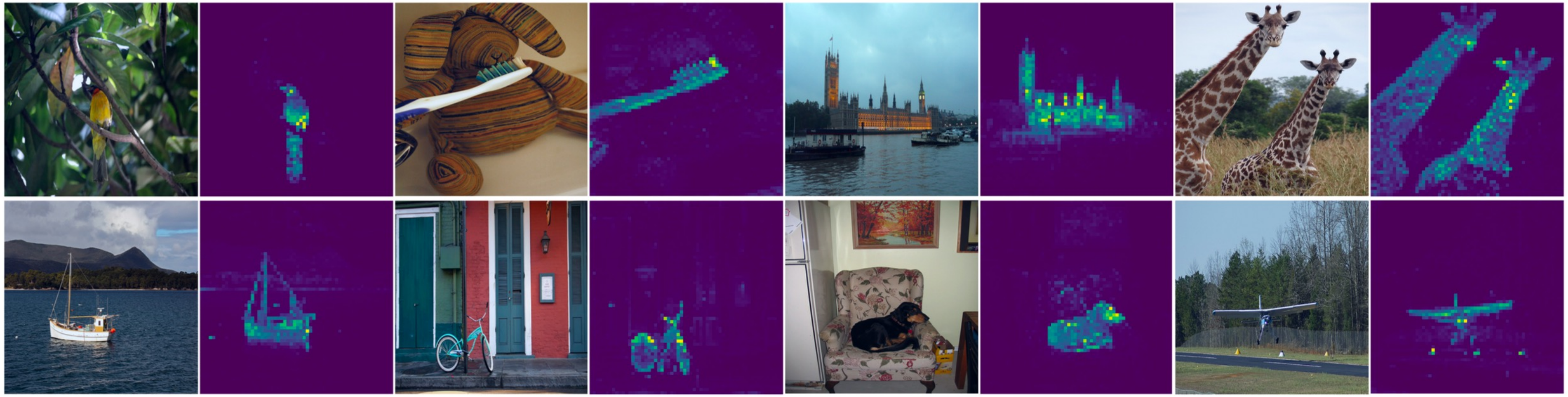




All runs are with ResNet-50 and trained for 400 epochs

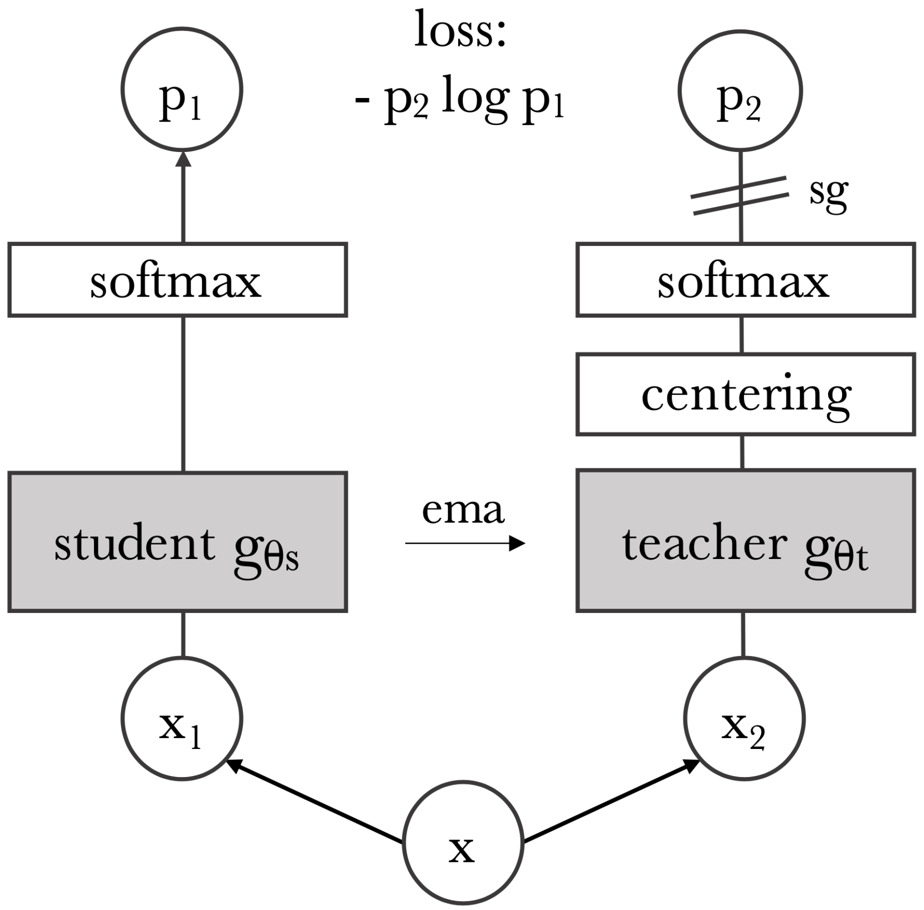
DINO

- Adapting SwAV to Vision Transformers
- Stripping the method down until it breaks



- Sample two data augmentation of same image x_1 and x_2
- Compute the representations z_1 and z_2
- Compute the output

$$f_\theta(x)^{(k)} = \frac{e^{w_k^\top z}}{\sum_{k'=1}^{K} e^{w_{k'}^\top z}}$$



- Compute the loss

$$L(\theta) = -\sum_{k=1}^{K} f_\eta(x_1)^{(k)} \log f_\theta(x_2)^{(k)}$$

- Update parameters

$$\theta_{i+1} = \theta_i - \alpha \, \nabla_\theta L(\theta_i) \qquad \text{(SGD)}$$

$$\eta_{i+1} = \mu \, \eta_i + (1-\mu) \, \theta_i \qquad \text{(EMA)}$$

Some ugly details

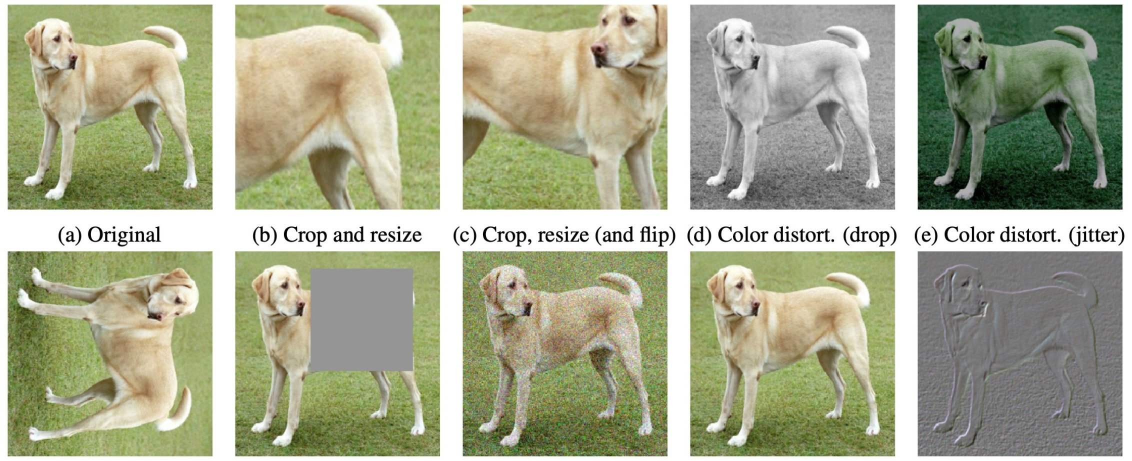
- For the teacher, we center the representation
- For both outputs, we use a softmax with low temperature

$$f_\eta(x_1)^{(k)} = \frac{e^{\frac{w_k^\top (z_1 - \bar{z})}{\tau}}}{\sum_{k'=1}^{K} e^{\frac{w_{k'}^\top (z_1 - \bar{z})}{\tau}}}$$

- Those two tricks avoid collapse

DINOv2

Was the modeling effort worth it?



(a) Original (b) Crop and resize (c) Crop, resize (and flip) (d) Color distort. (drop) (e) Color distort. (jitter)
(f) Rotate {90°, 180°, 270°} (g) Cutout (h) Gaussian noise (i) Gaussian blur (j) Sobel filtering

ℓ_2 norm?	τ	Entropy	Contrastive acc.	Top 1
Yes	0.05	1.0	90.5	59.7
	0.1	4.5	87.8	64.4
	0.5	8.2	68.2	60.7
	1	8.3	59.1	58.0
No	10	0.5	91.7	57.2
	100	0.5	92.1	57.0

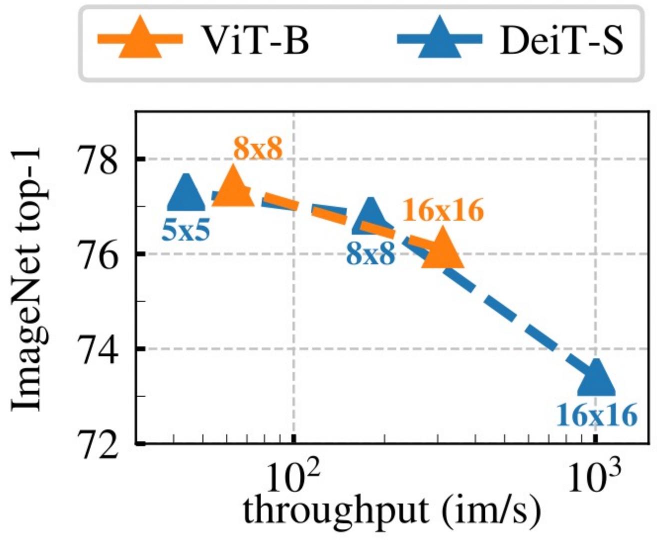
Target	τ_{base}	Top-1
Constant random network	1	18.8$\pm$0.7
Moving average of online	0.999	69.8
Moving average of online	0.99	**72.5**
Moving average of online	0.9	68.4
Stop gradient of online†	0	0.3

Chen, Ting, et al. "A simple framework for contrastive learning of visual representations." *International conference on machine learning*. PMLR, 2020.

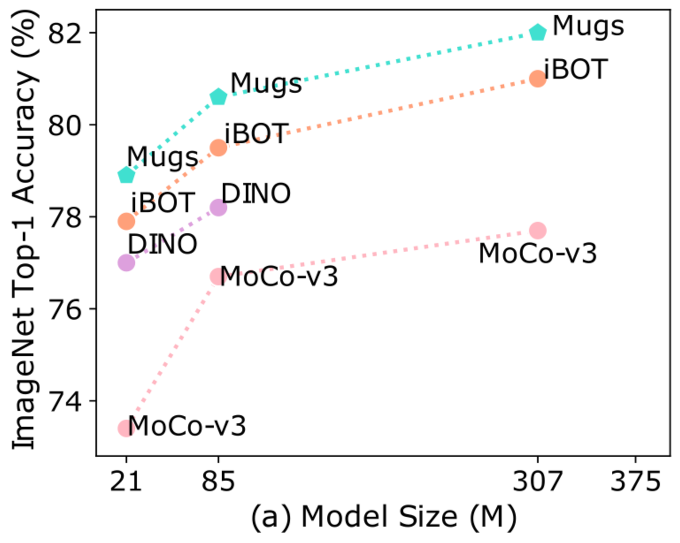
Grill, Jean-Bastien, et al. "Bootstrap your own latent-a new approach to self-supervised learning." *Advances in neural information processing systems* 33 (2020): 21271-21284.

Meta AI

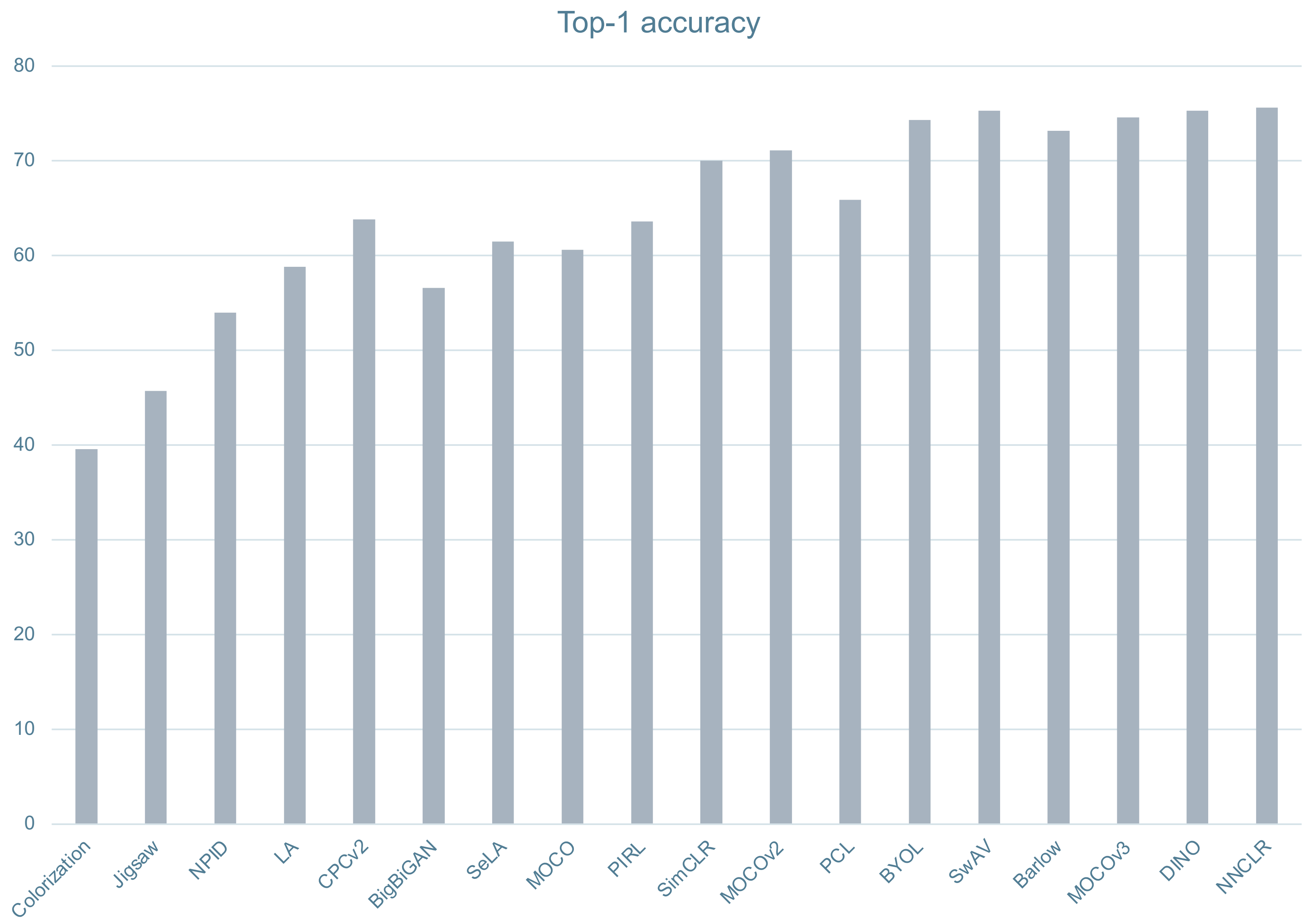
Plateau in Performance



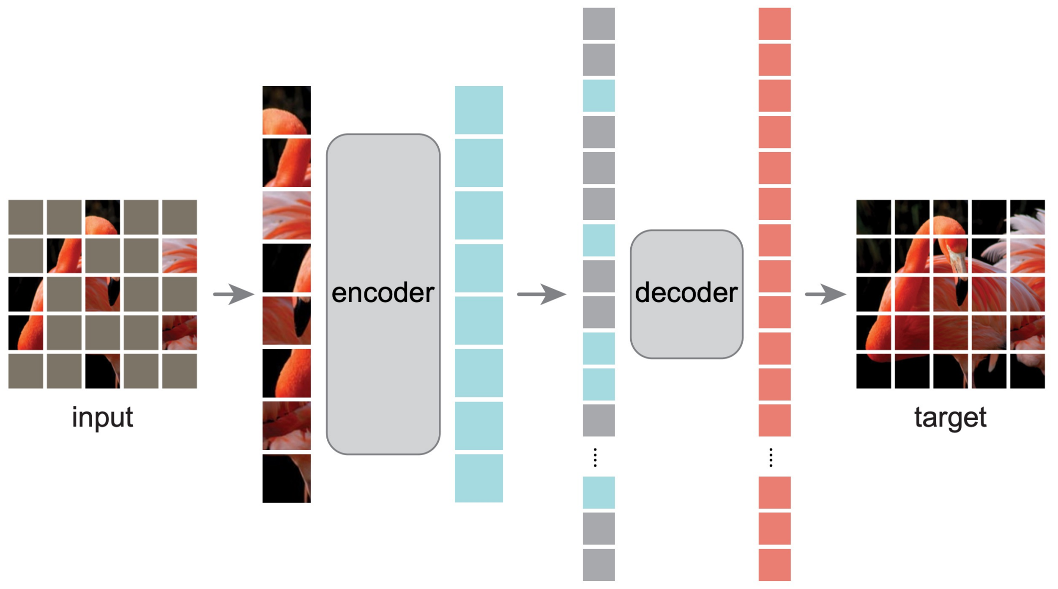
Caron, Mathilde, et al. "Emerging properties in self-supervised vision transformers." *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 2021.



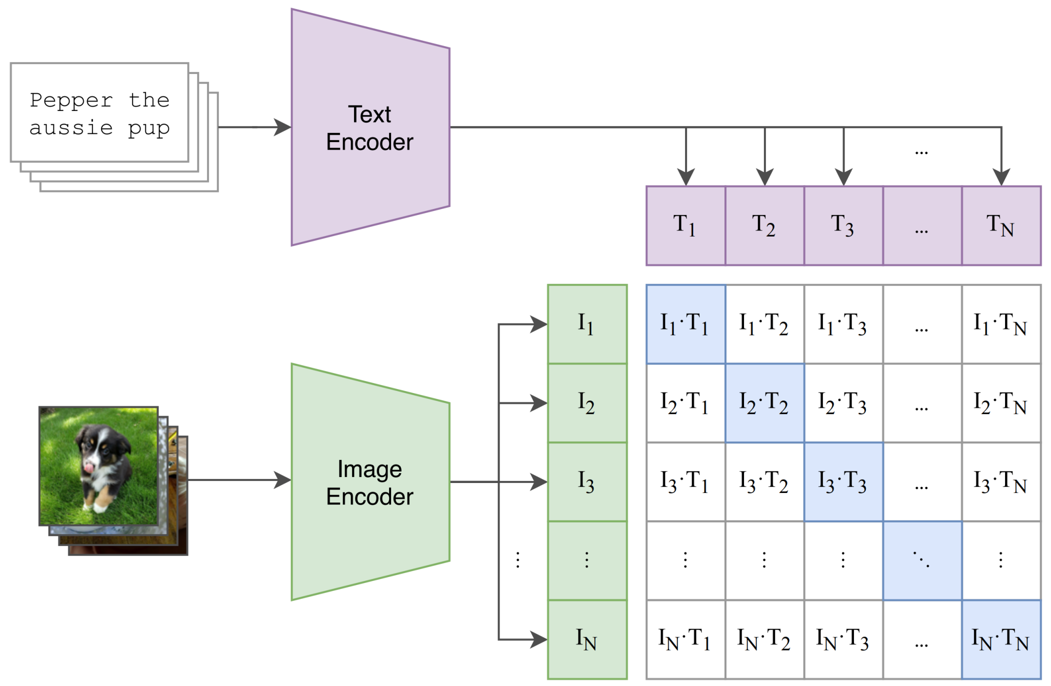
Zhou, Pan, et al. "Mugs: A Multi-Granular Self-Supervised Learning Framework." *arXiv preprint arXiv:2203.14415* (2022).



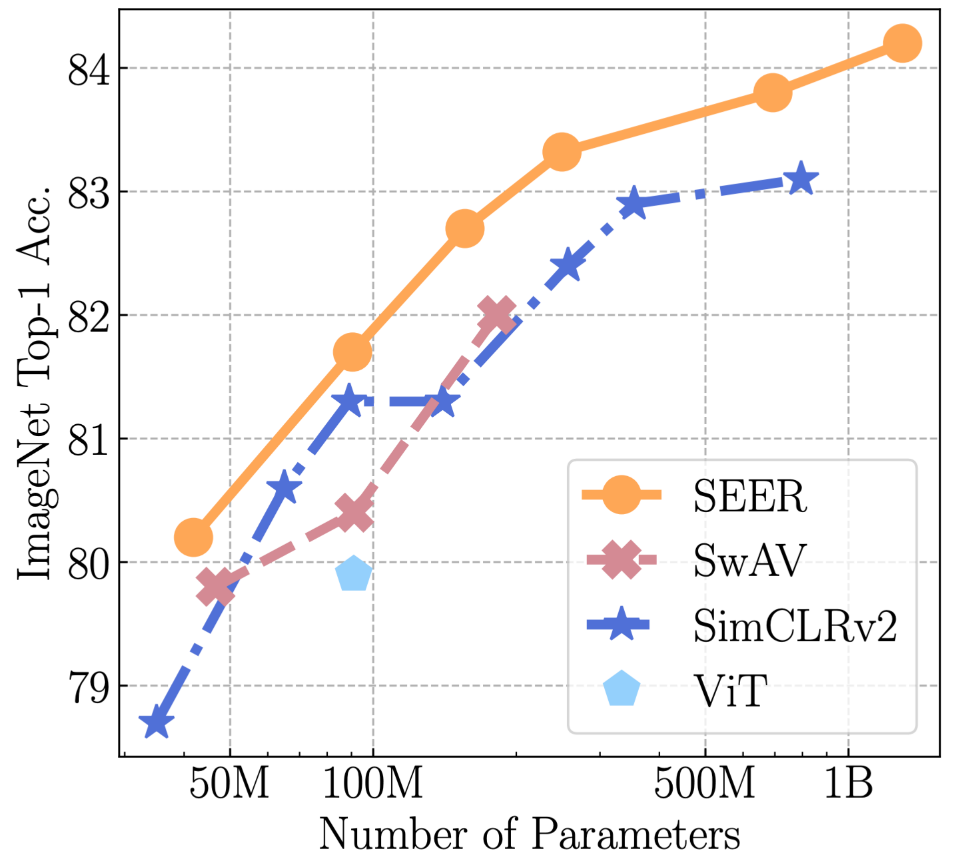
Motivations for DINOv2



He, Kaiming, et al. "Masked autoencoders are scalable vision learners." *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 2022.

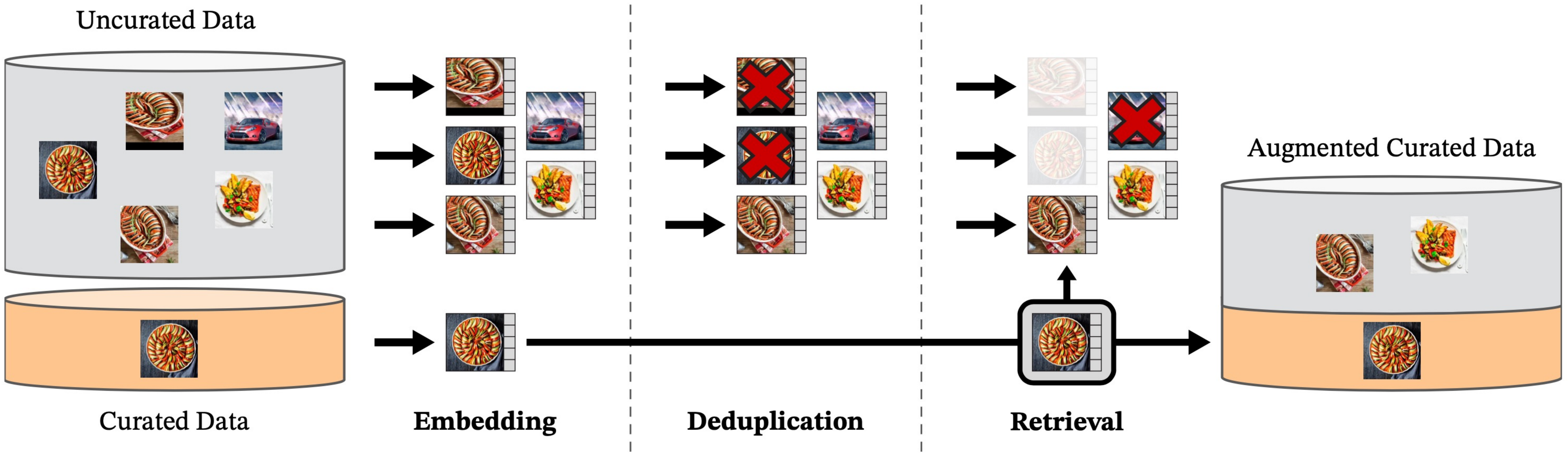


Radford, Alec, et al. "Learning transferable visual models from natural language supervision." *International conference on machine learning*. PMLR, 2021.



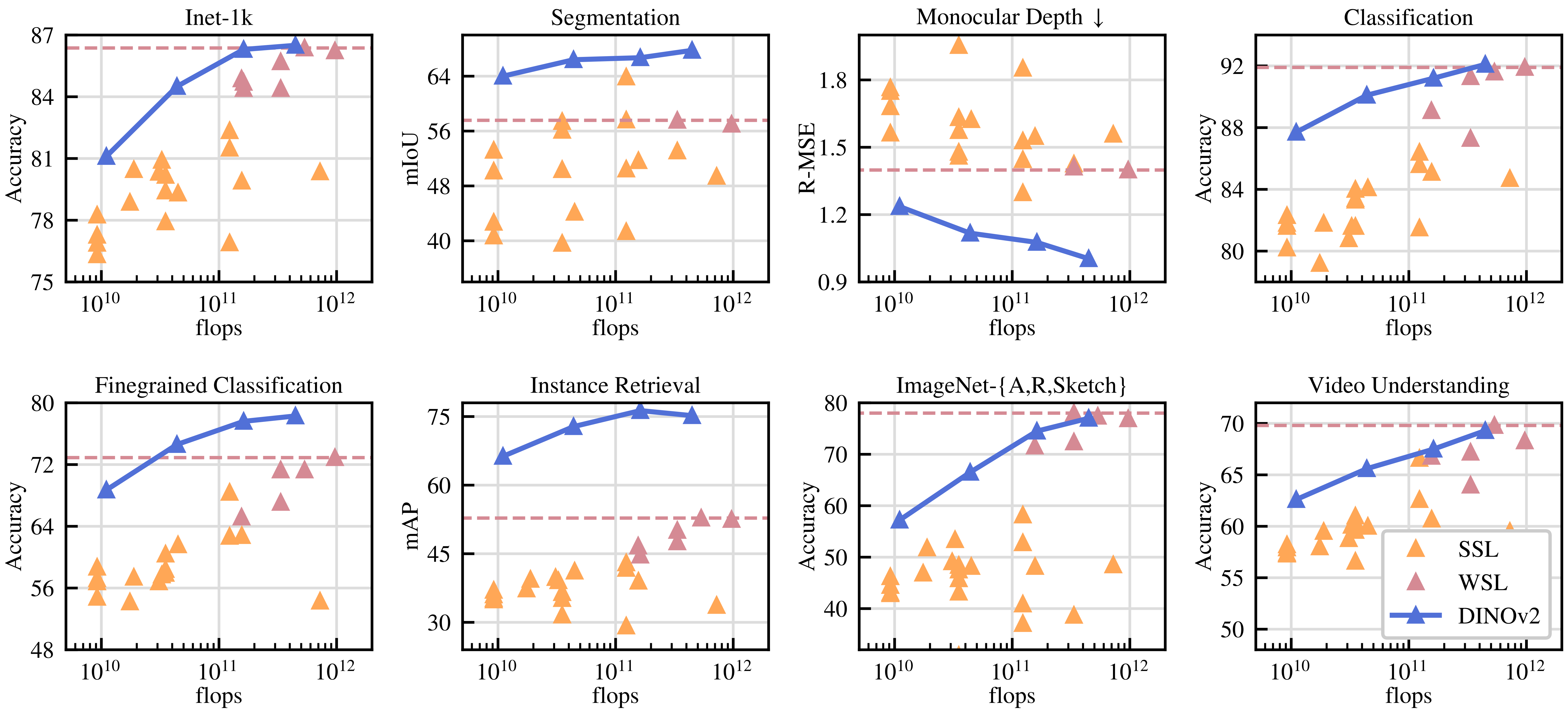
Goyal, Priya, et al. "Self-supervised pretraining of visual features in the wild." *arXiv preprint arXiv:2103.01988* (2021).

Data Curation



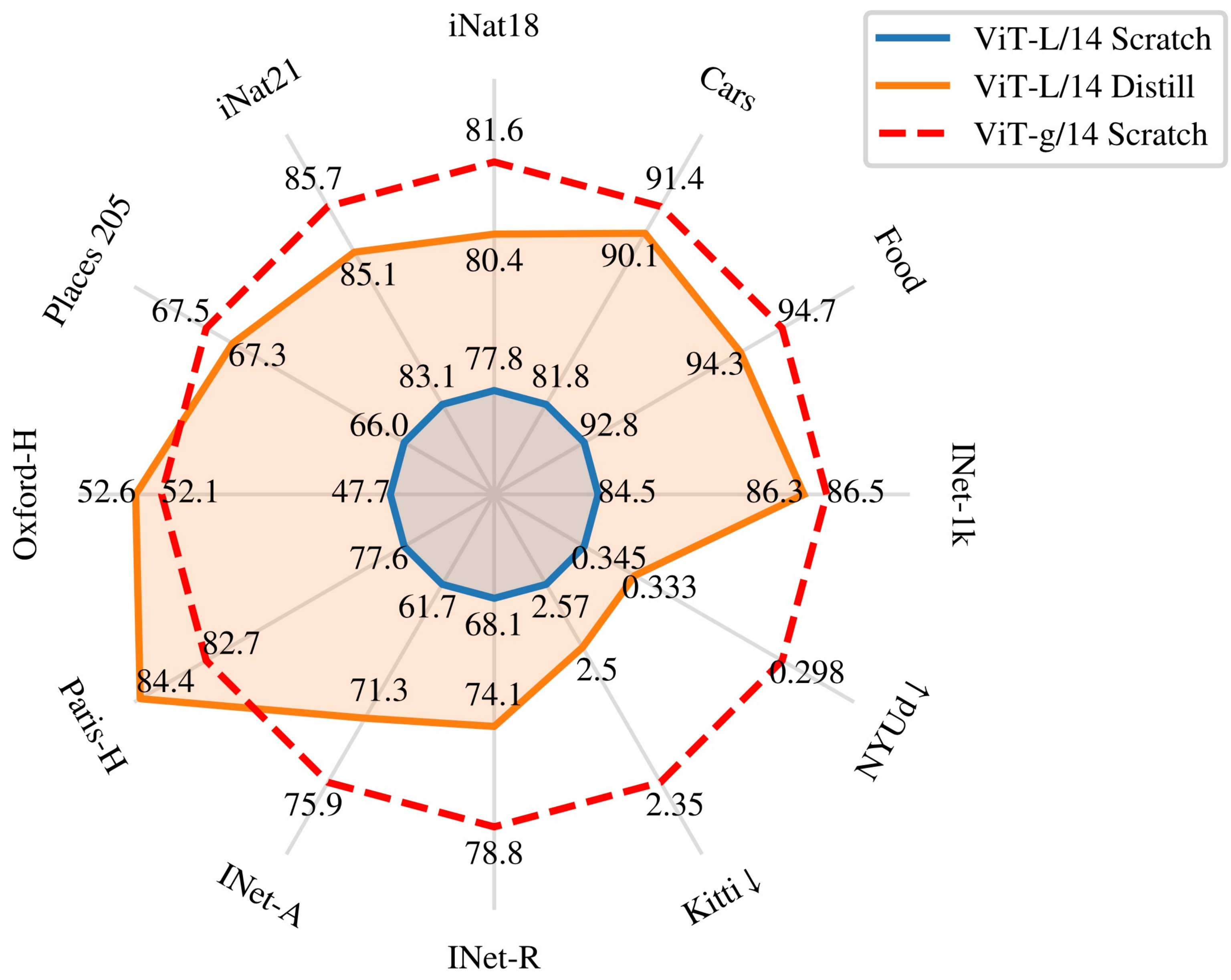
Model Scaling and Stability

	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


Inet-1k
Segmentation
Monocular Depth ↓
Classification
Finegrained Classification
Instance Retrieval
ImageNet-{A,R,Sketch}
Video Understanding
Accuracy
mIoU
R-MSE
mAP
flops
SSL
WSL
DINOv2

Distillation

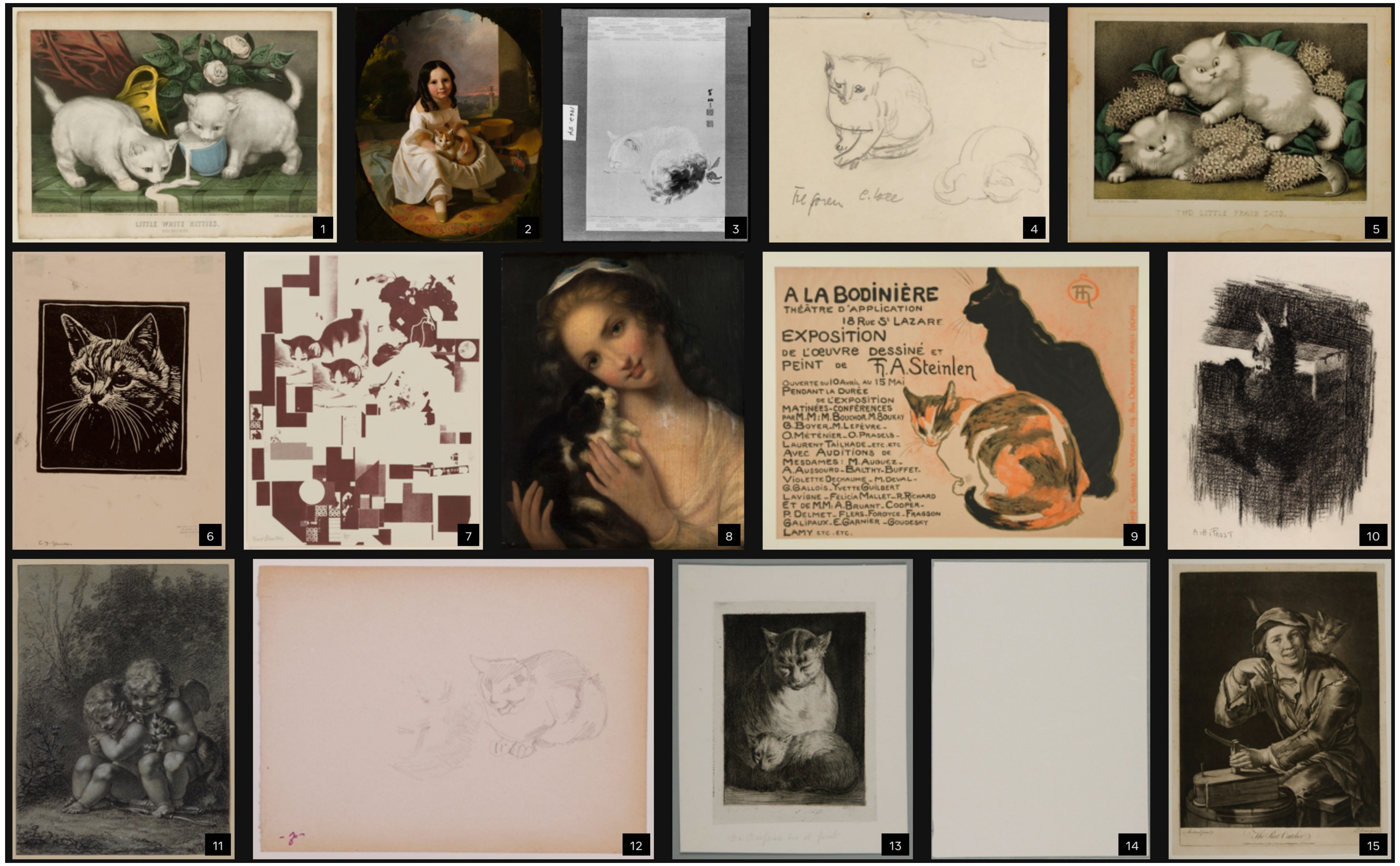
- Instead of training a family of model, train one → the largest one!
- Obtain smaller models using distillation
- Our training loss is perfectly suited for this, stop the teacher!
- Trained ViT-{S, B, L} from the ViT-g (1.1B params)
- Interestingly, the ViT-L distilled works better than from scratch!



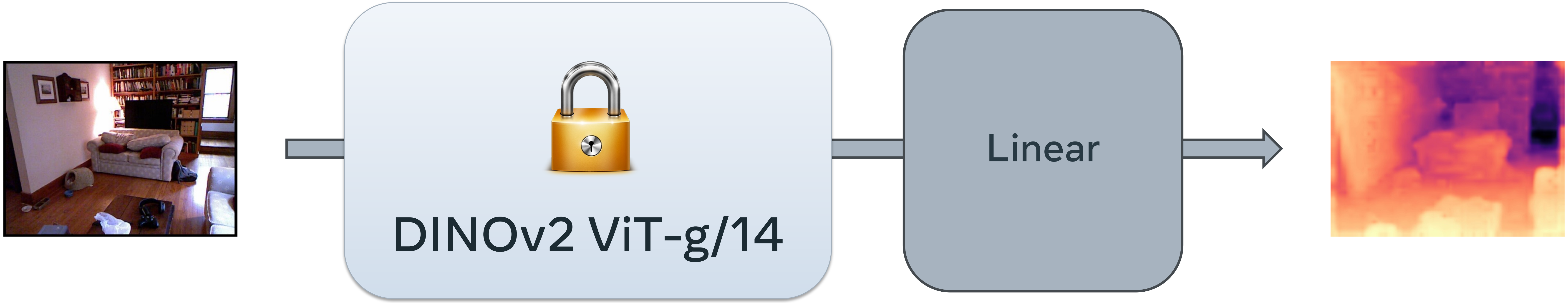
Retrieval

Feature	Arch	Oxford		Paris		Met			AmsterTime
		M	H	M	H	GAP	GAP-	ACC	mAP
OpenCLIP	ViT-G/14	50.7	19.7	79.2	60.2	6.5	23.9	34.4	24.6
MAE	ViT-H/14	11.7	2.2	19.9	4.7	7.5	23.5	30.5	4.2
DINO	ViT-B/8	40.1	13.7	65.3	35.3	17.1	37.7	43.9	24.6
iBOT	ViT-L/16	39.0	12.7	70.7	47.0	25.1	54.8	58.2	26.7
DINOv2	ViT-S/14	68.8	43.2	84.6	68.5	29.4	54.3	57.7	43.5
	ViT-B/14	72.9	49.5	90.3	78.6	36.7	63.5	66.1	45.6
	ViT-L/14	**75.1**	**54.0**	**92.7**	**83.5**	**40.0**	68.9	71.6	**50.0**
	ViT-g/14	73.6	52.3	92.1	82.6	36.8	**73.6**	**76.5**	46.7

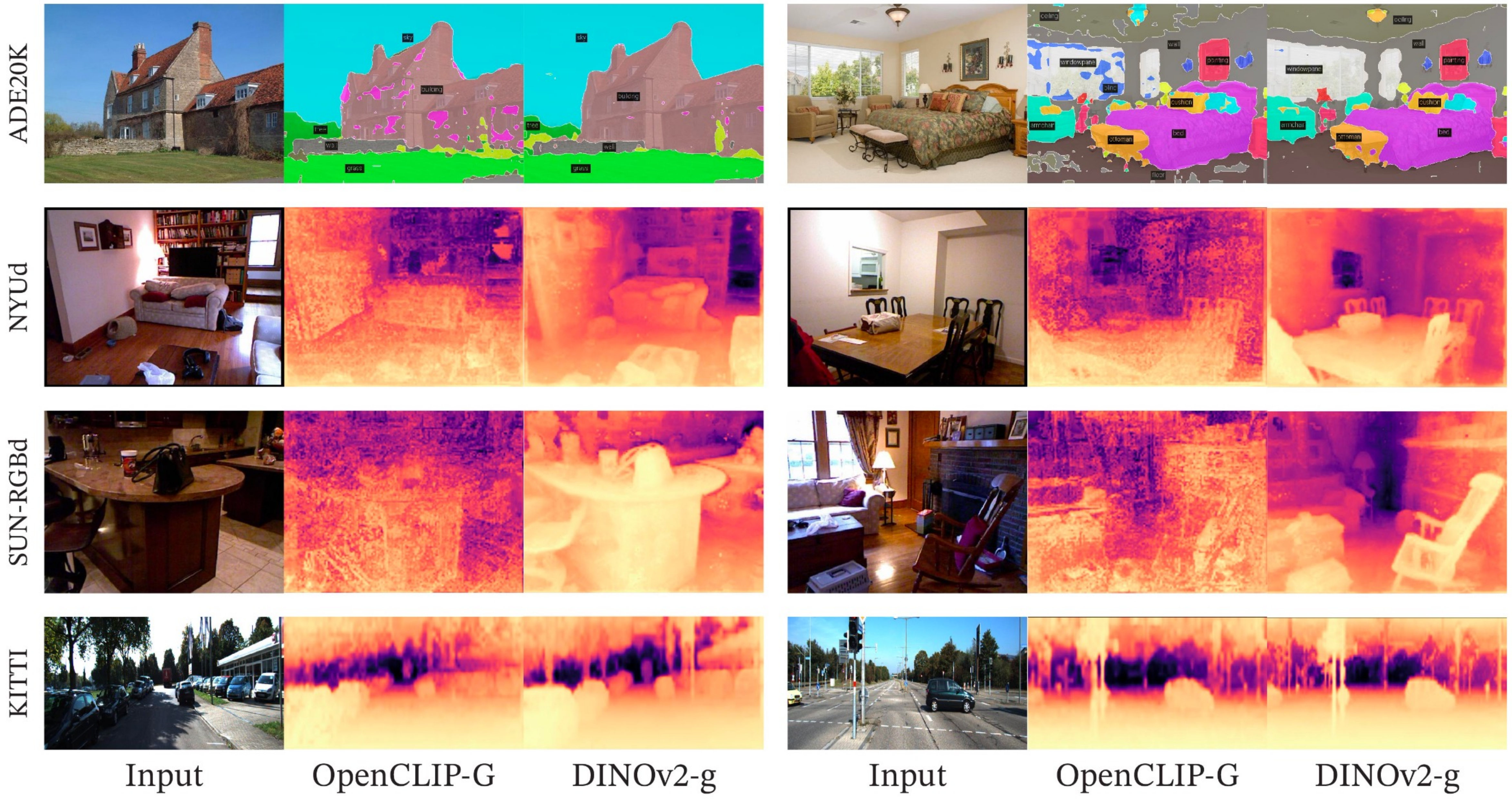
Query



Dense Prediction Tasks

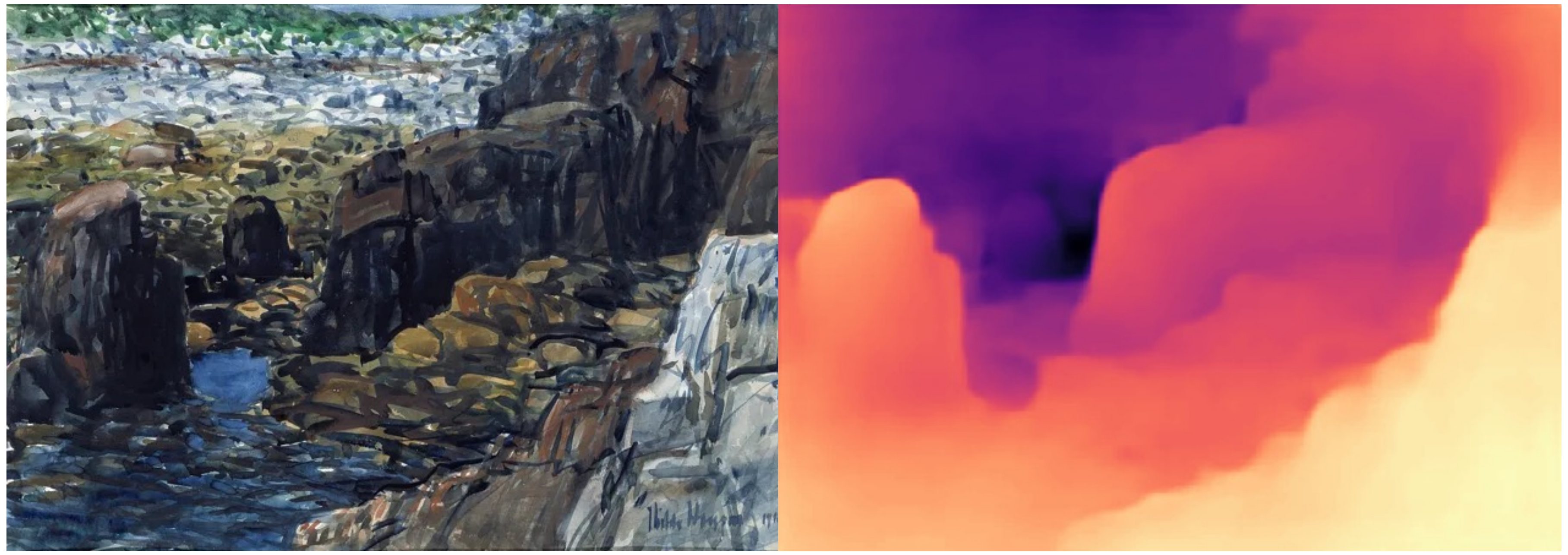


Dense Prediction Tasks



OOD Generalization

Method	Arch	Data	Im-A	Im-R	Im-C↓	Sketch
OpenCLIP	ViT-G/14	LAION	63.8	**87.8**	45.3	**66.4**
MAE	ViT-H/14	INet-1k	10.2	34.4	61.4	21.9
DINO	ViT-B/8	INet-1k	23.9	37.0	56.6	25.5
iBOT	ViT-L/16	INet-22k	41.5	51.0	43.9	38.5
DINOv2	ViT-S/14	LVD-142M	33.5	53.7	54.4	41.2
	ViT-B/14	LVD-142M	55.1	63.3	42.7	50.6
	ViT-L/14	LVD-142M	71.3	74.4	31.5	59.3
	ViT-g/14	LVD-142M	**75.9**	78.8	**28.2**	62.5

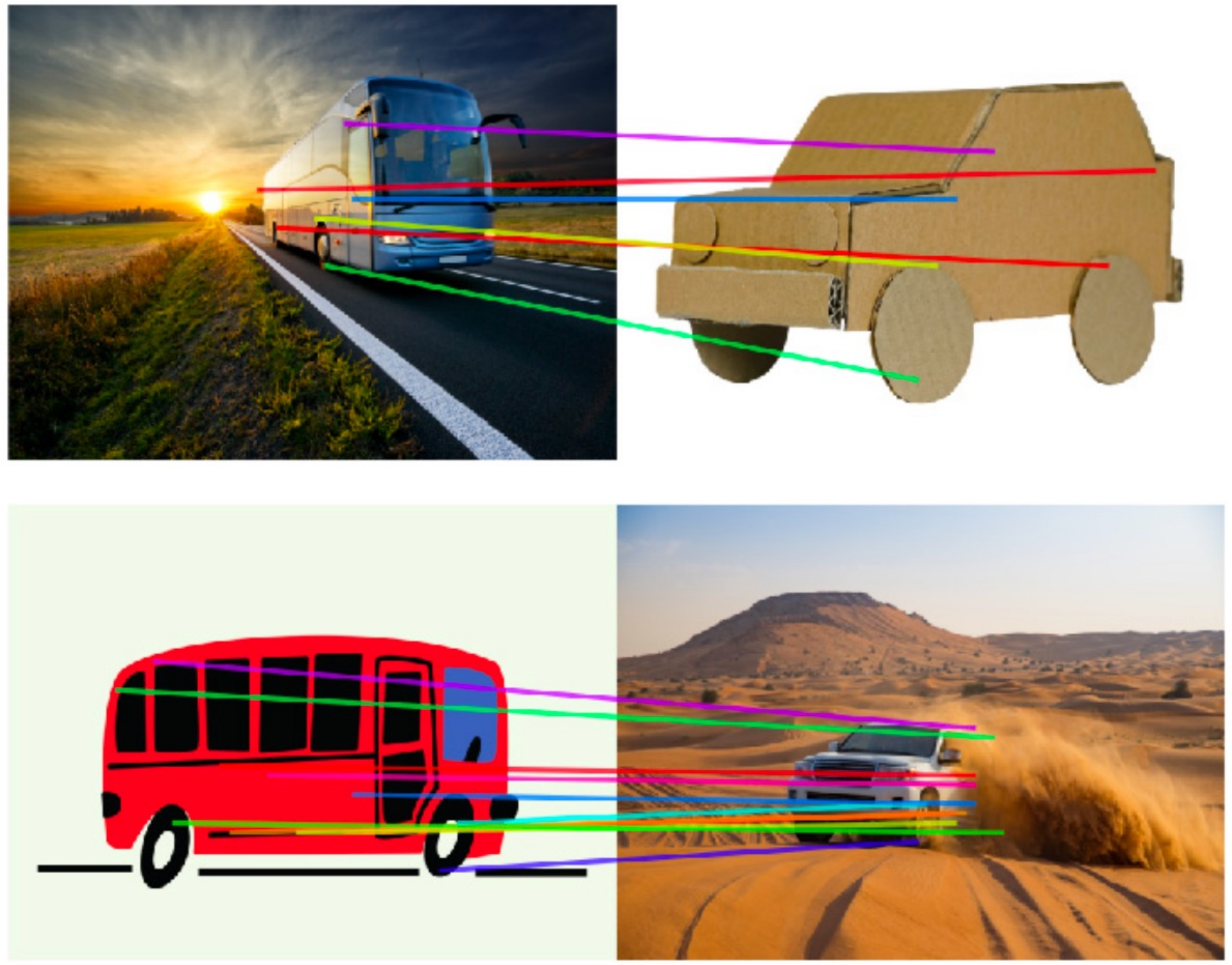


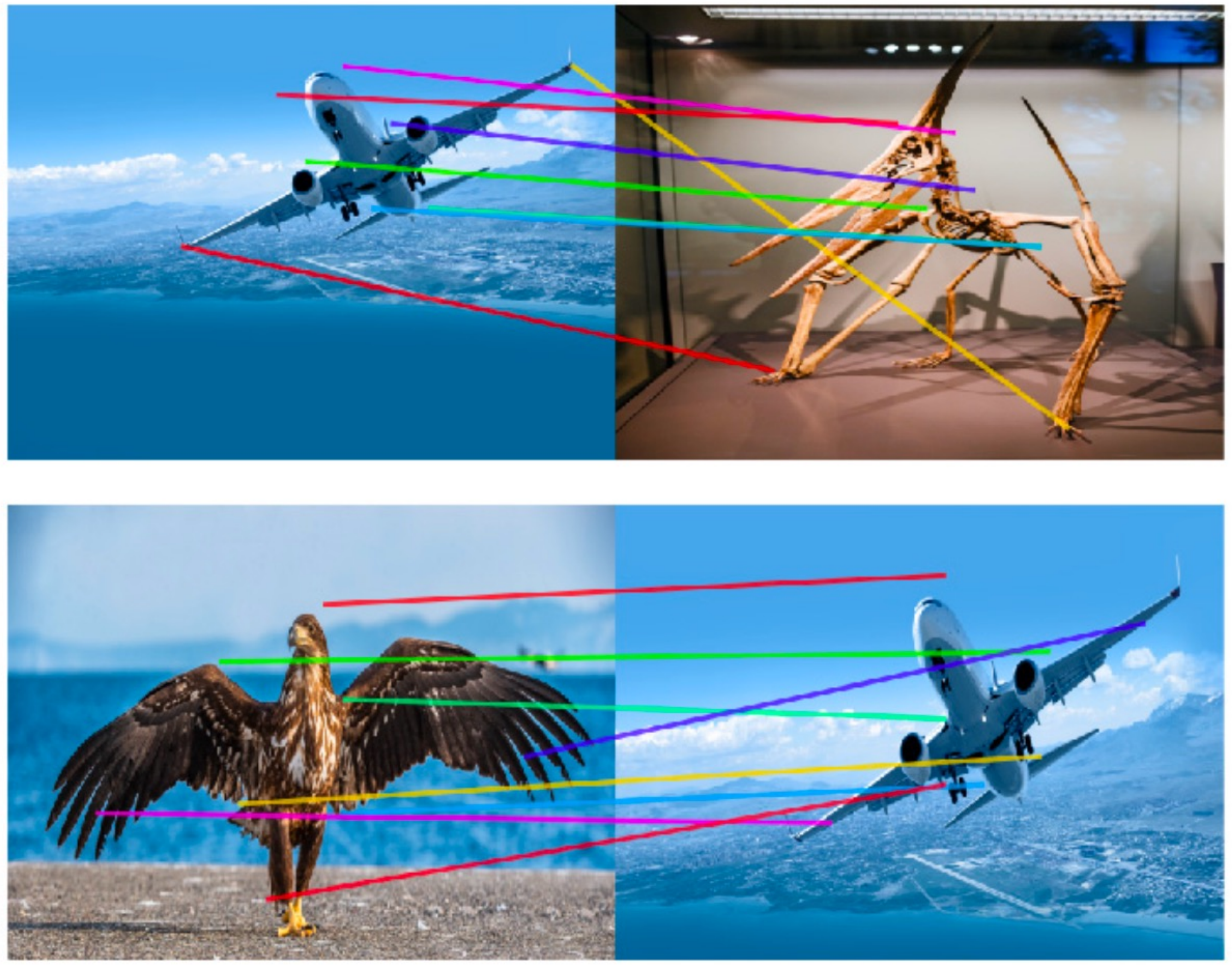
Emerging Properties

(a) (b) (c) (d)

Meta AI

Surprising Matching

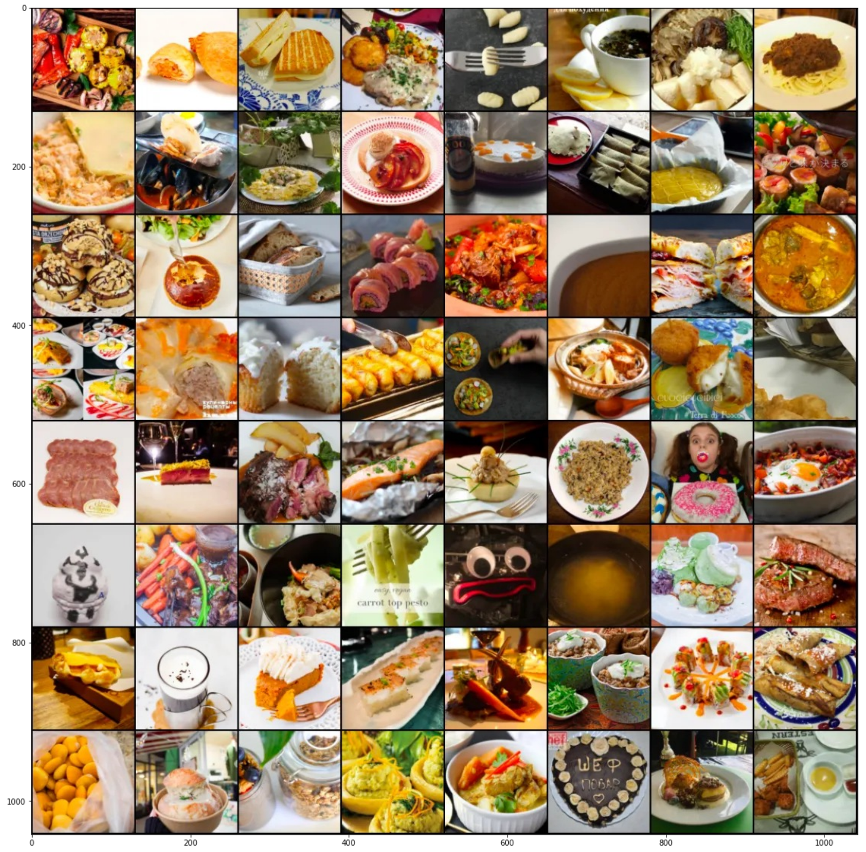
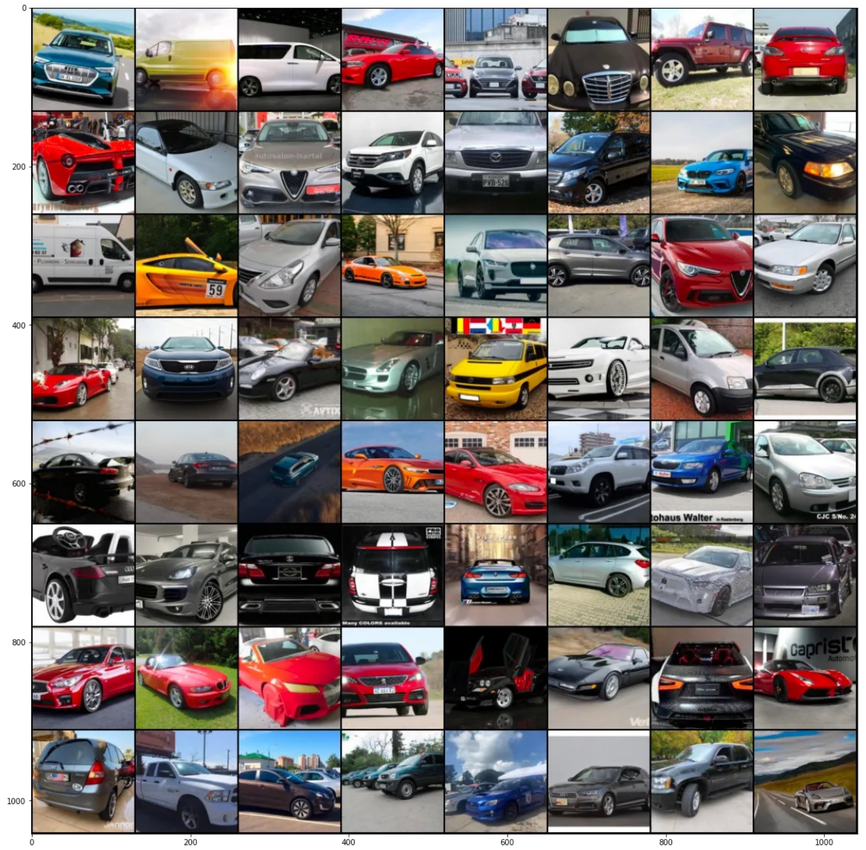
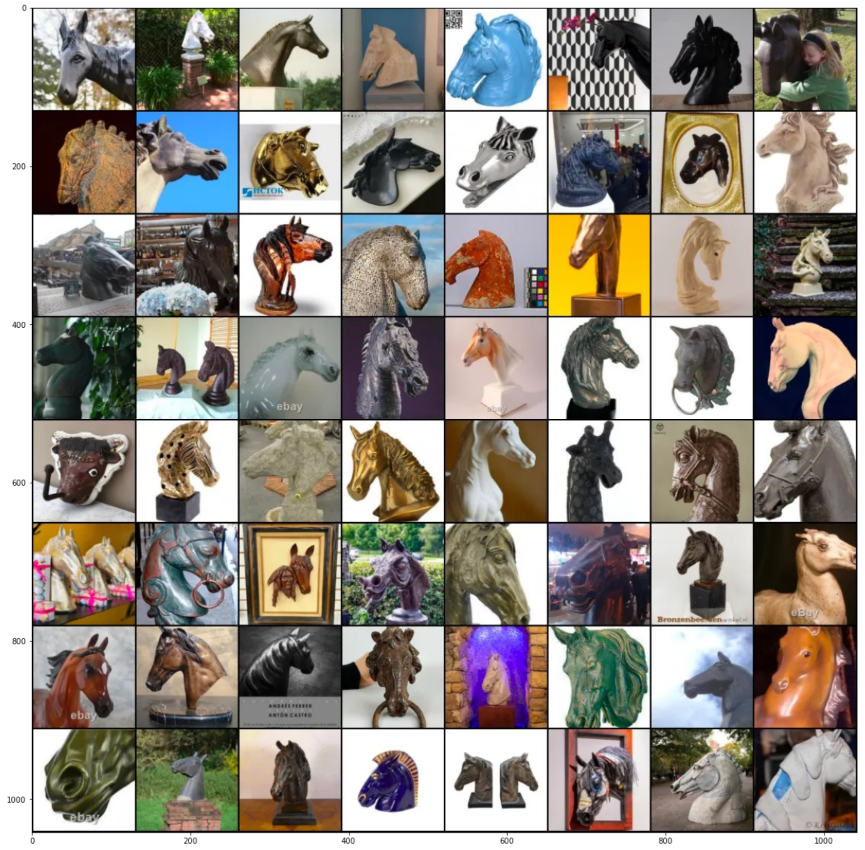

(Vehicles)


(Birds / Airplanes)

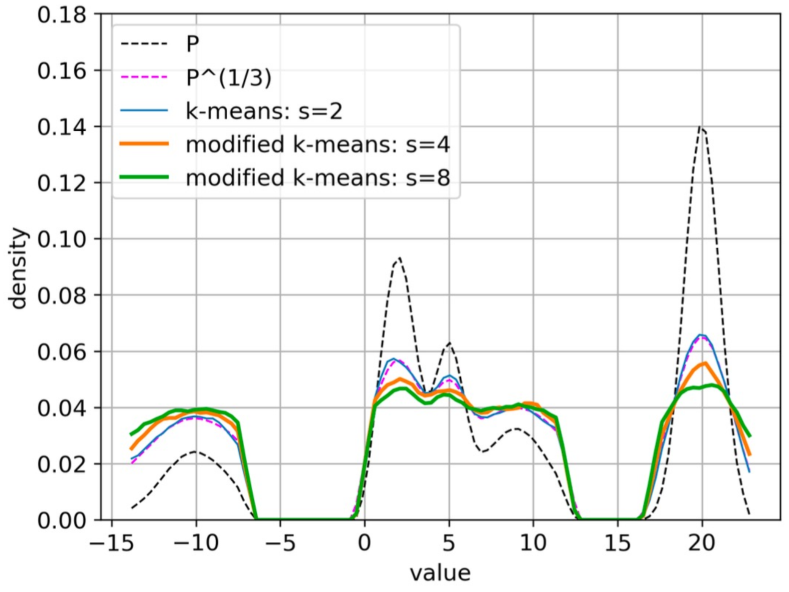
https://dinov2.metademolab.com/

Recent Improvements

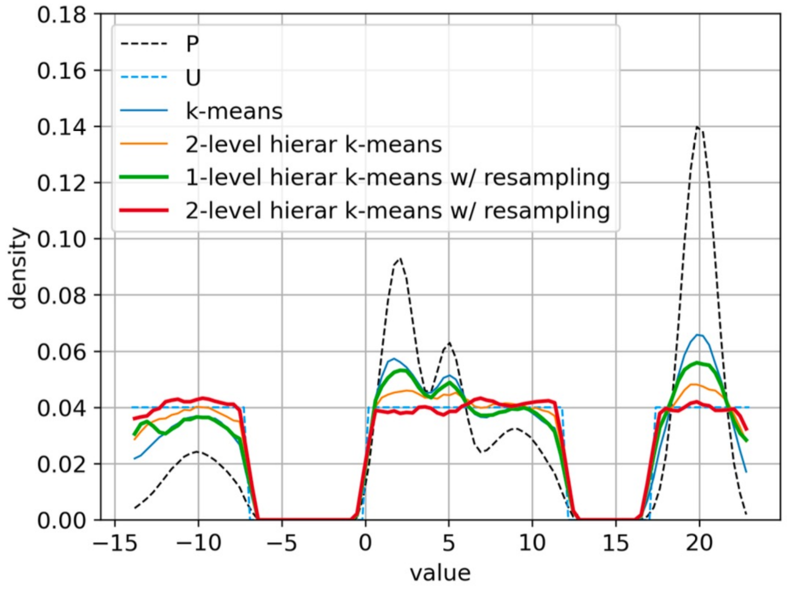
Automatic Data Curation

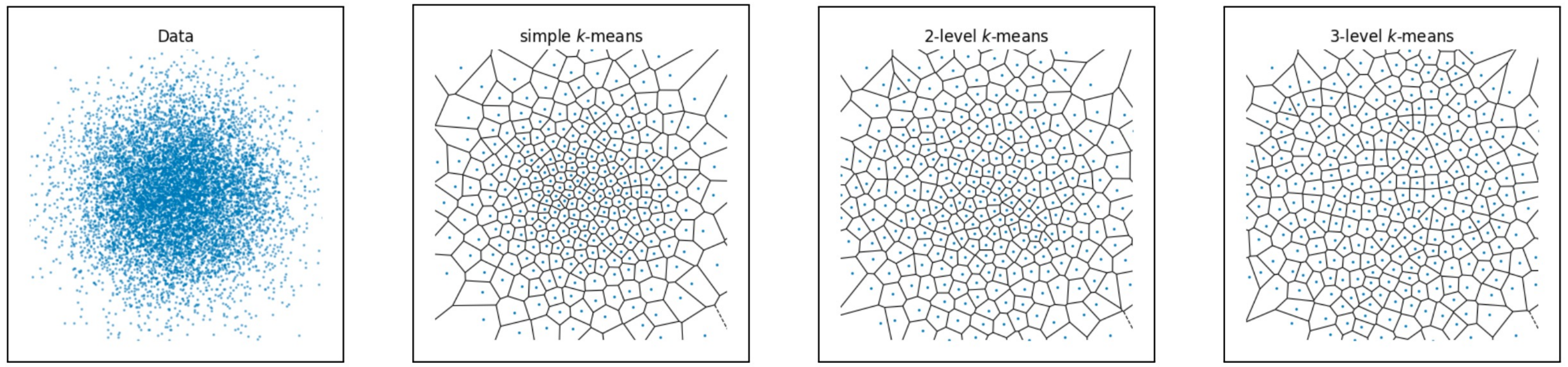

Hierarchical Clustering



(a) Modified k-means with $d = \|x - y\|^s$ in 1-D

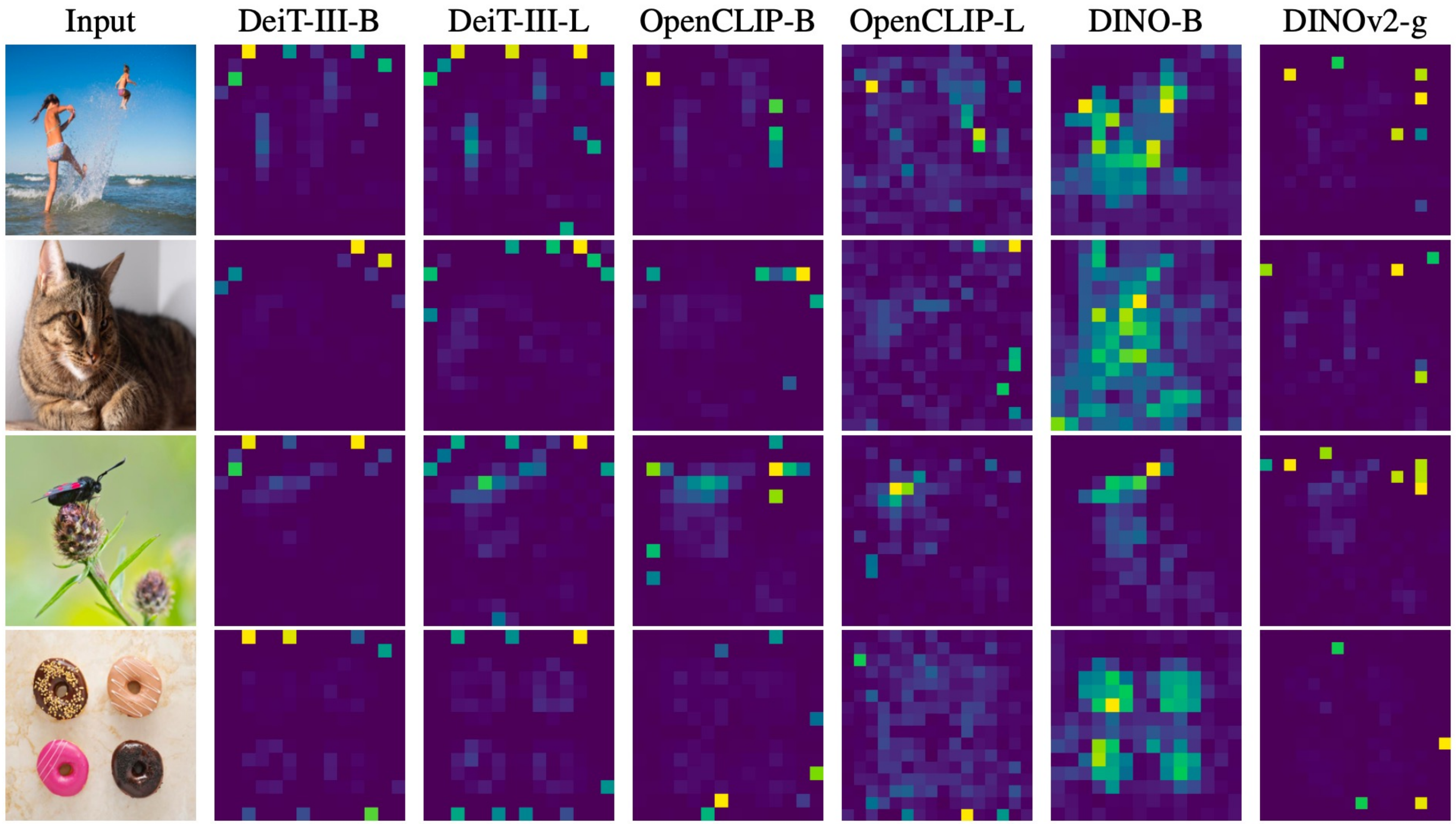


(b) Hierarchical k-means in 1-D

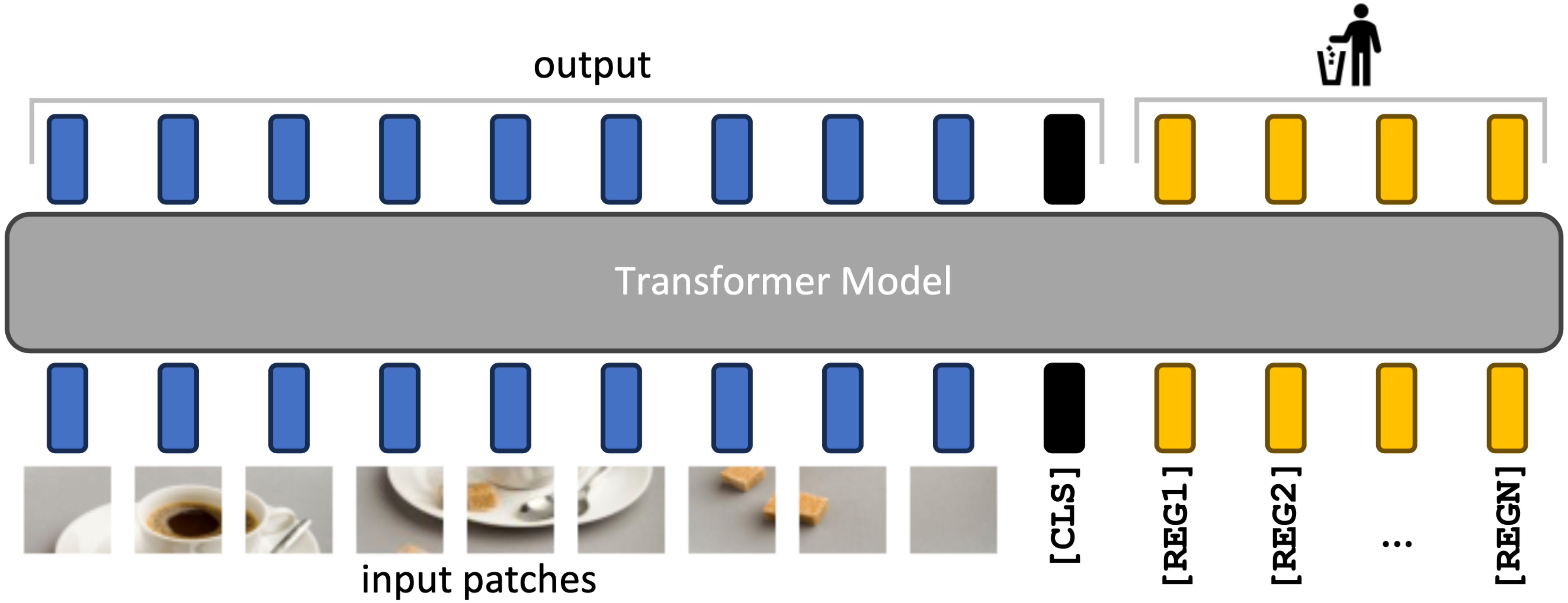


(c) Voronoi cells obtained from clusterings of a 2-D Gaussian

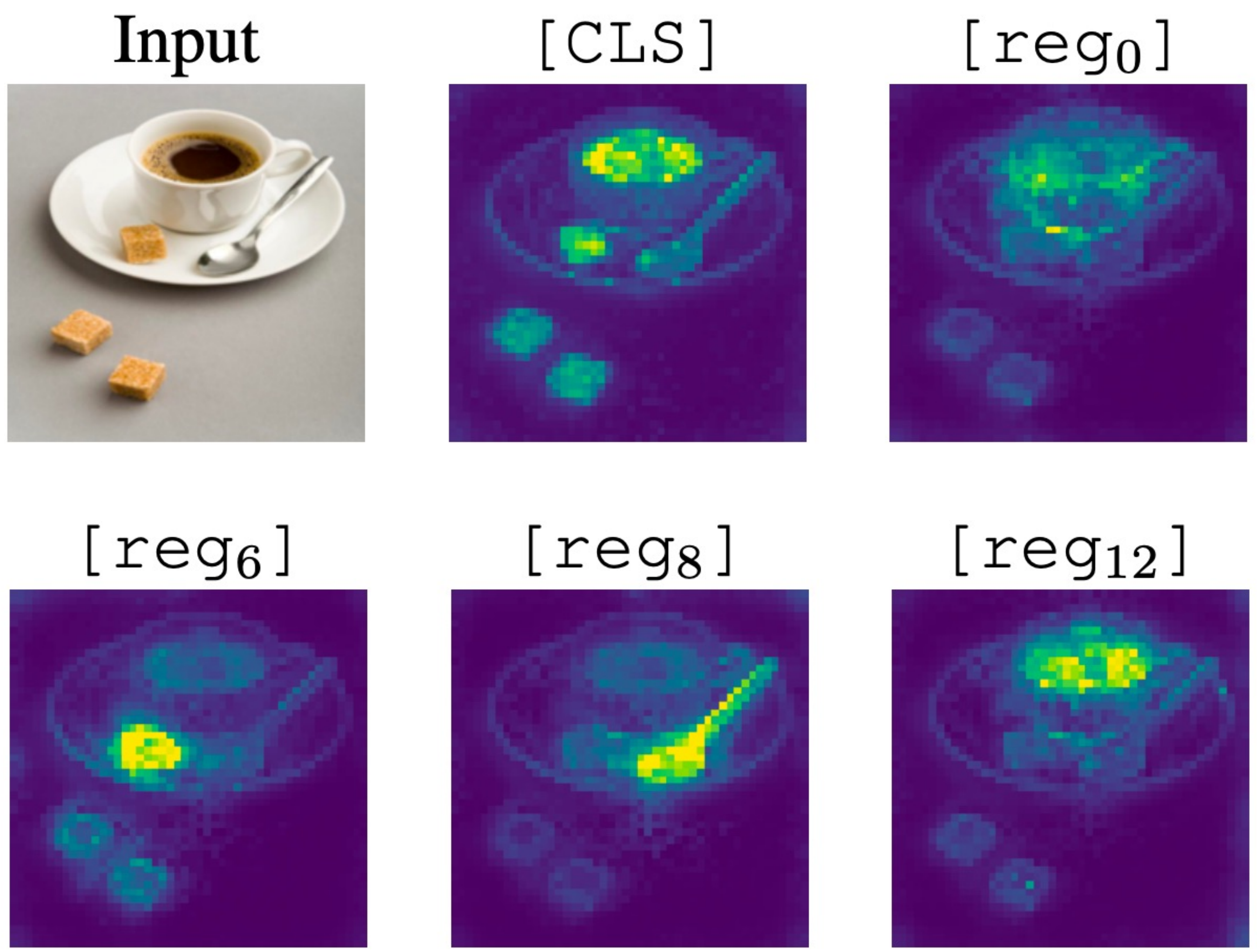
Fixing the attention maps of DINOv2

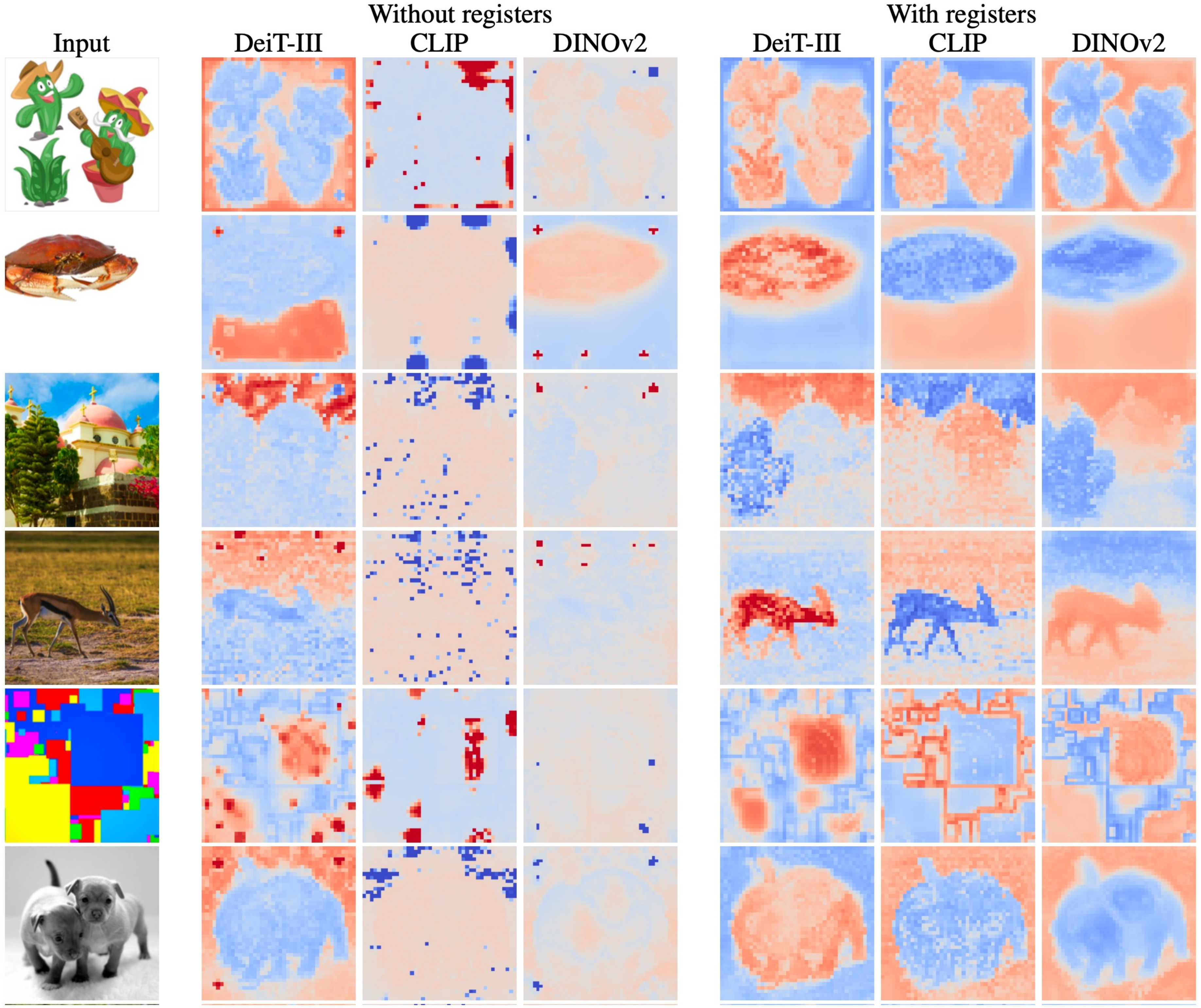


Registers



Results


Input
[CLS]
[reg_0]
[reg_6]
[reg_8]
[reg_{12}]


Input
Without registers
DeiT-III
CLIP
DINOv2
With registers
DeiT-III
CLIP
DINOv2

Applications

High-Resolution Canopy Height Estimation

Physical Modelling @ Meta

Jamie Tolan

Ben Nosarzewski

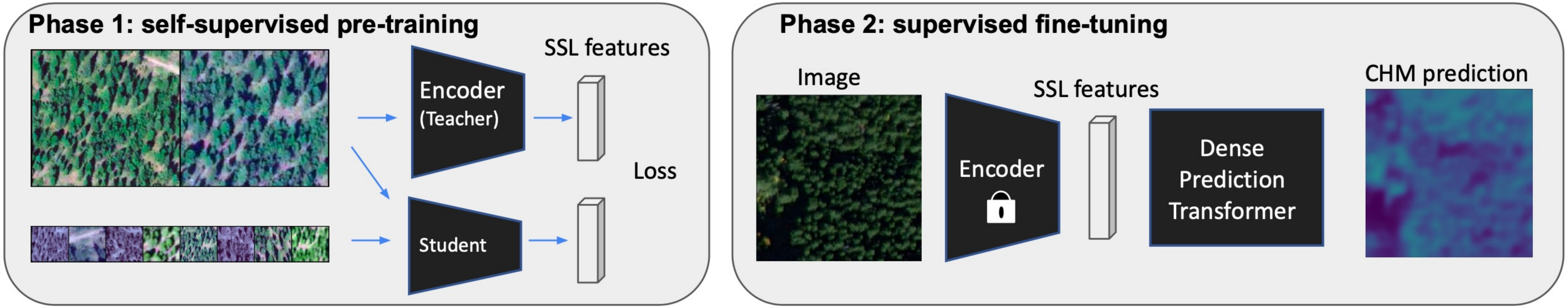
Tobias Tiecke

World Ressource Institute

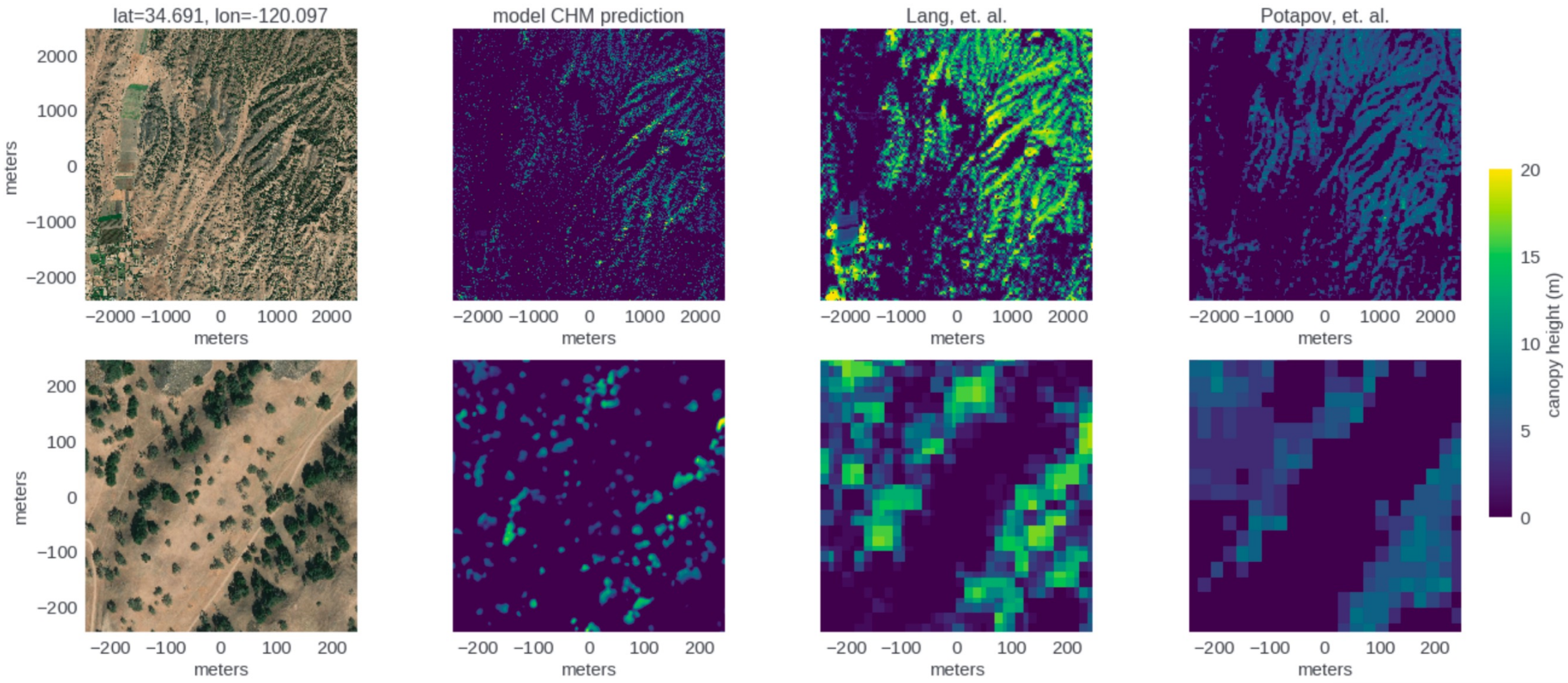
John Brandt

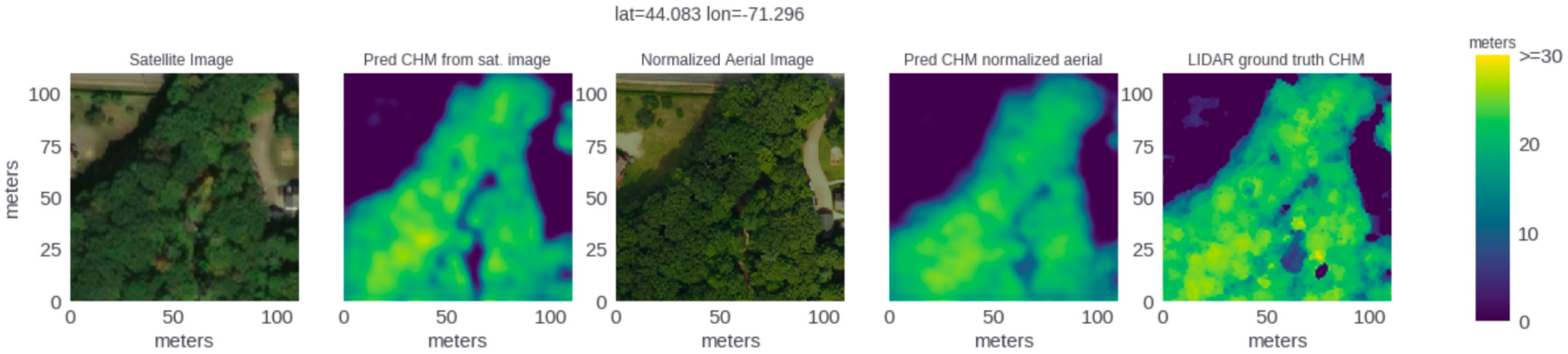
Justine Spore

	Coverage	Type	Channels	Beam
MAXAR	Global	Satellite	RGB	
GEDI	Near-Global	Satellite	RGB + LIDAR	25m
NEON	Small	Airborne	RGB + LIDAR	1m



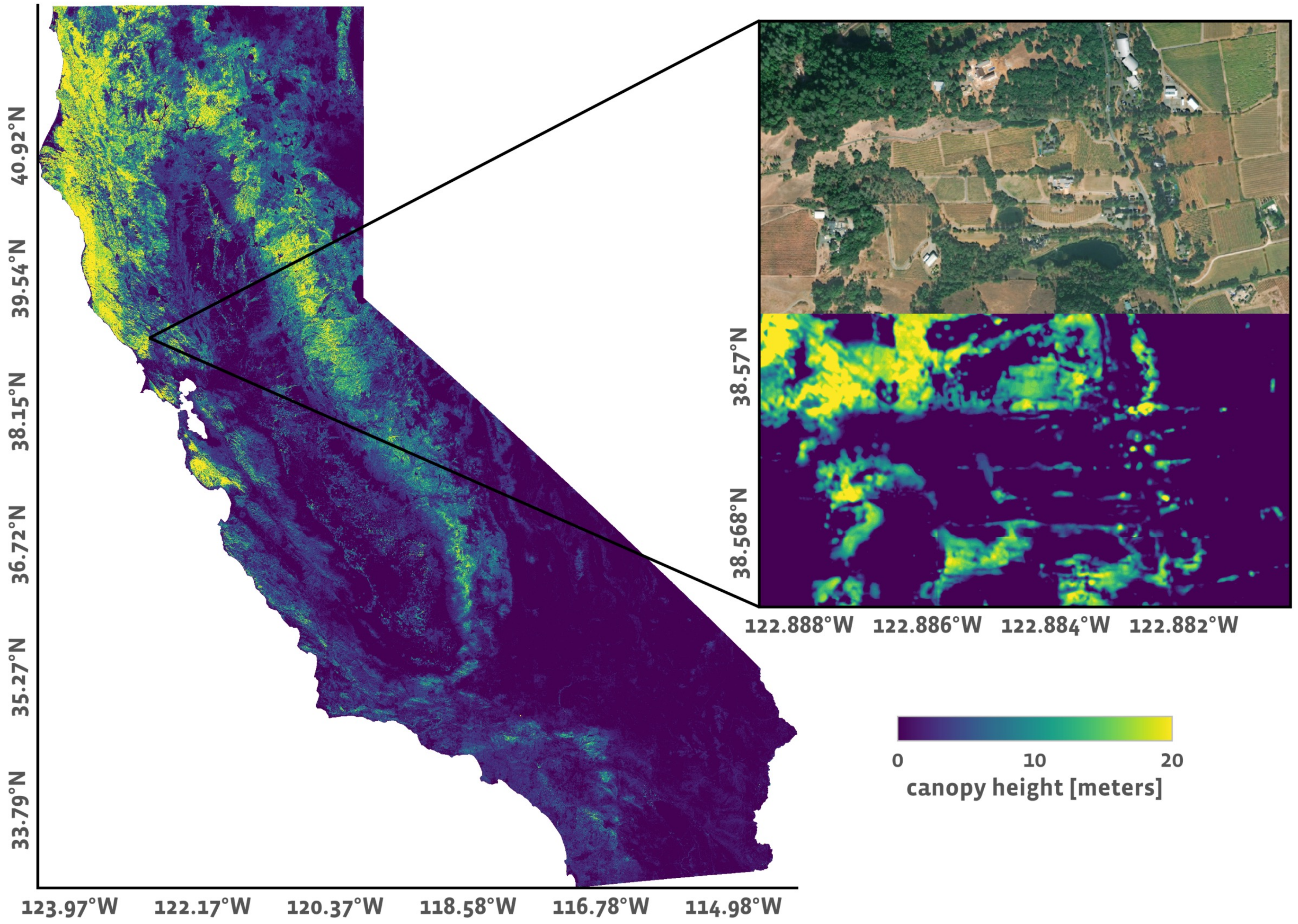
Canopy Height Estimation

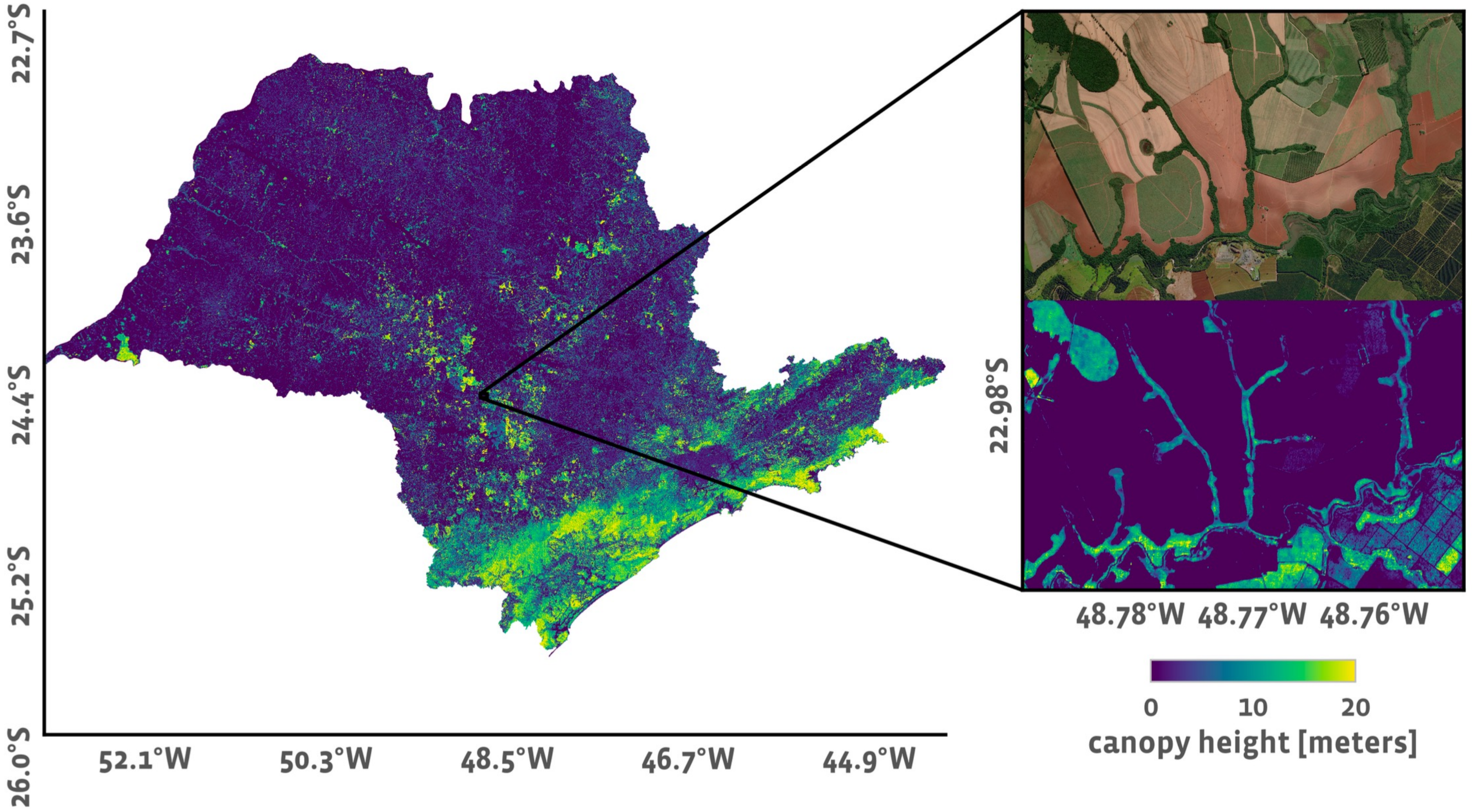




Meta AI

Canopy Height Estimation





Single-Cell Microscopy

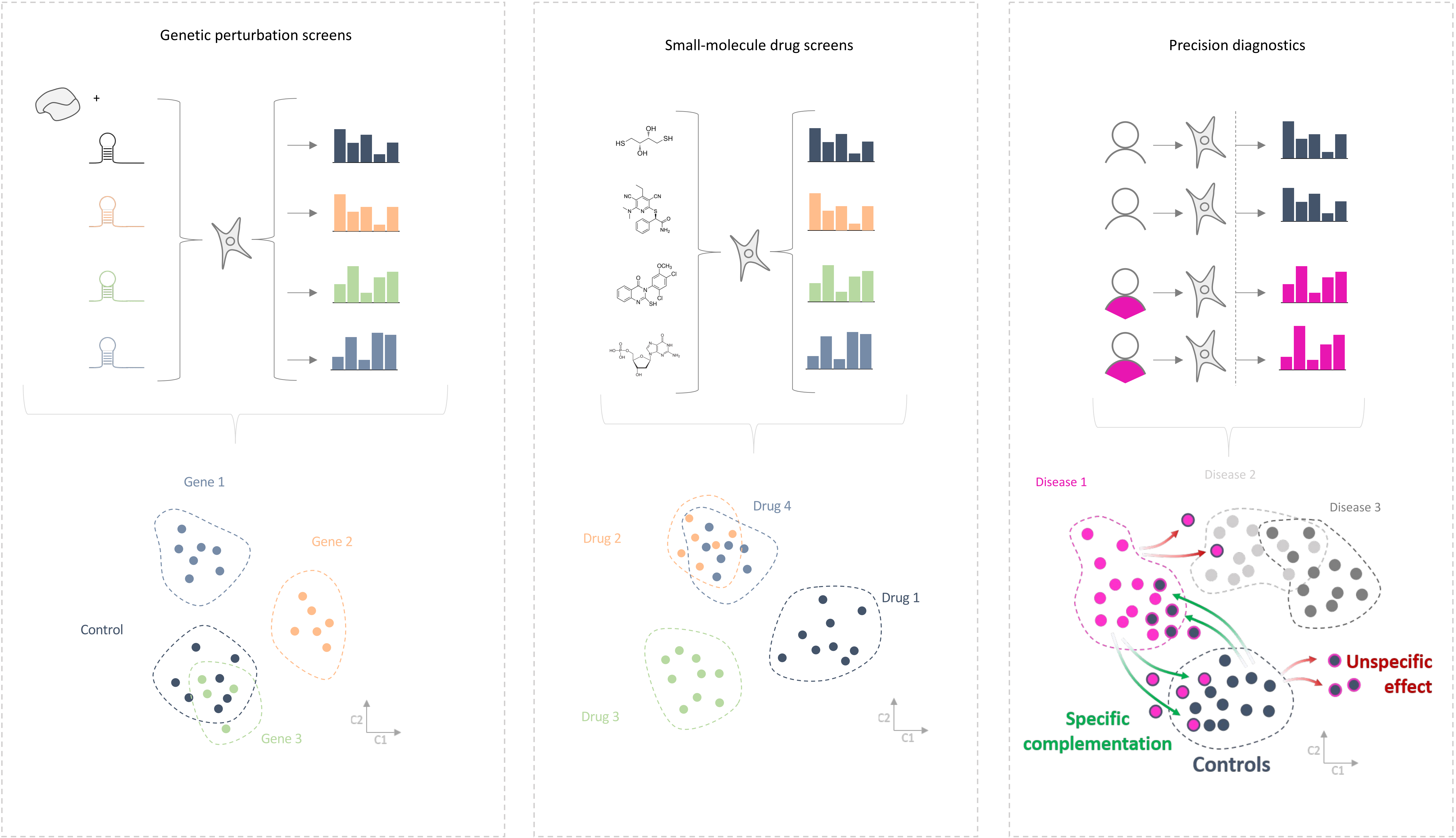
Juan C. Caicedo
University of Wisconsin-Madison /
Broad Institute of MIT

Wolfgang Pernice
Columbia University Irving Medical Center

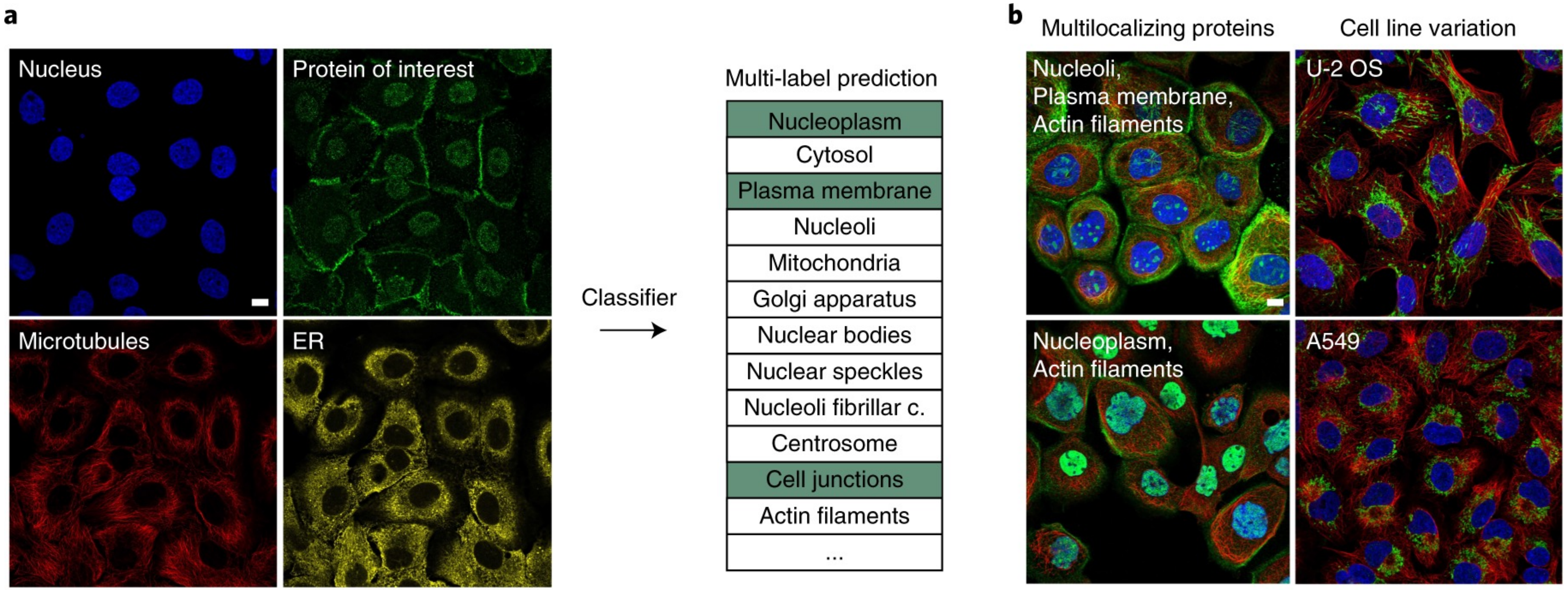
Michael Doron
Q.Ai / Broad Institute of MIT

Meta AI

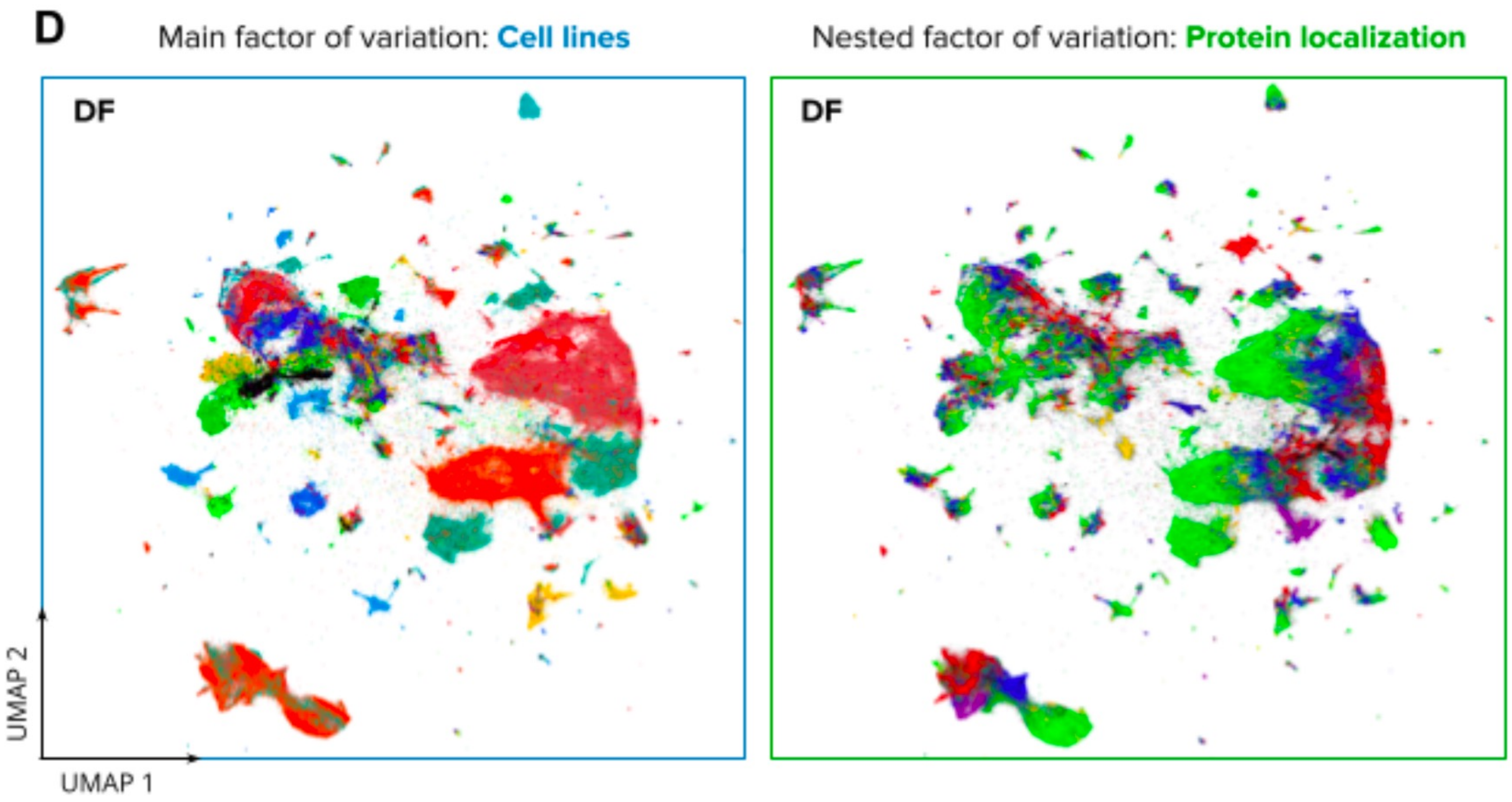
Single-Cell Microscopy

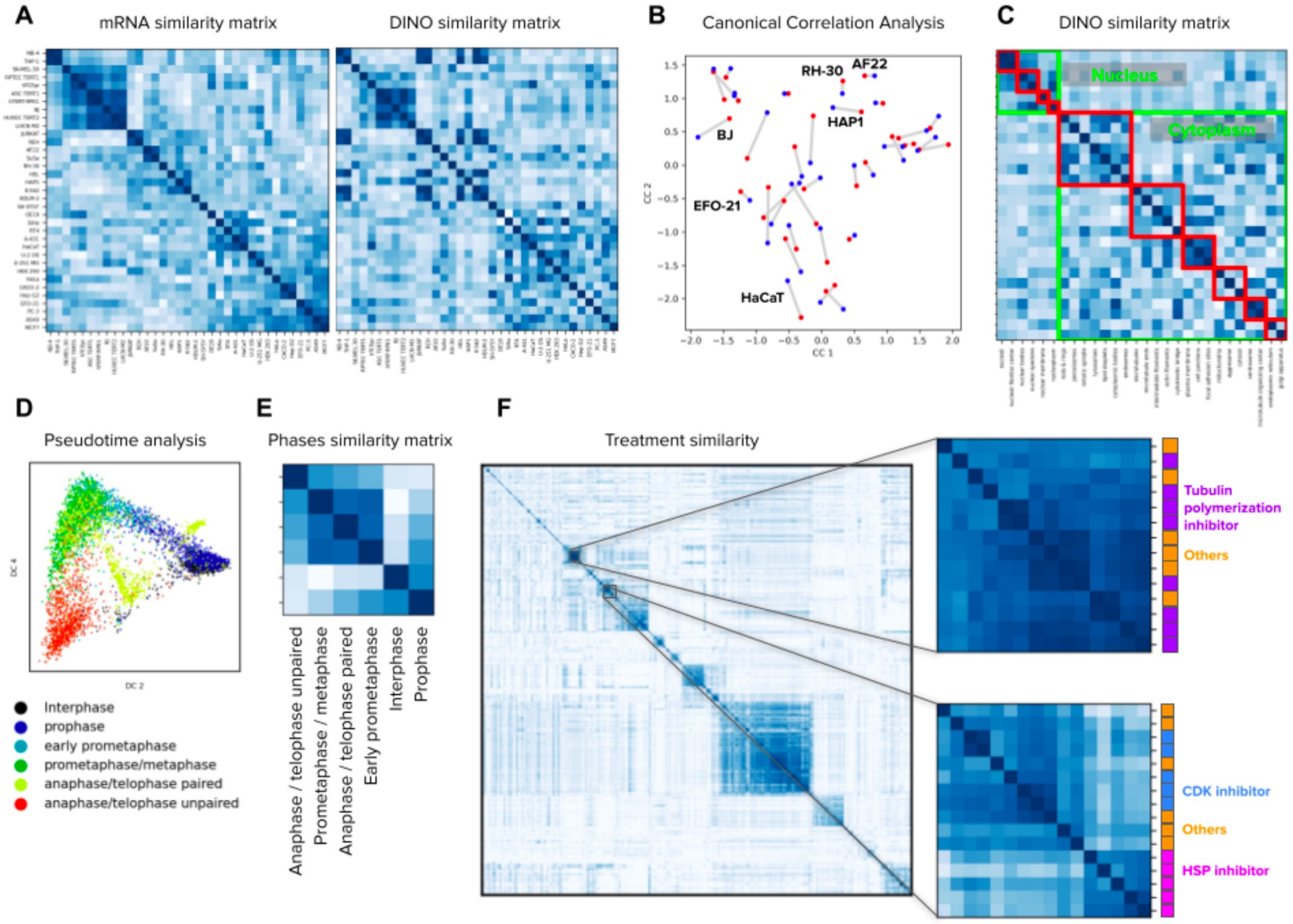


Meta AI



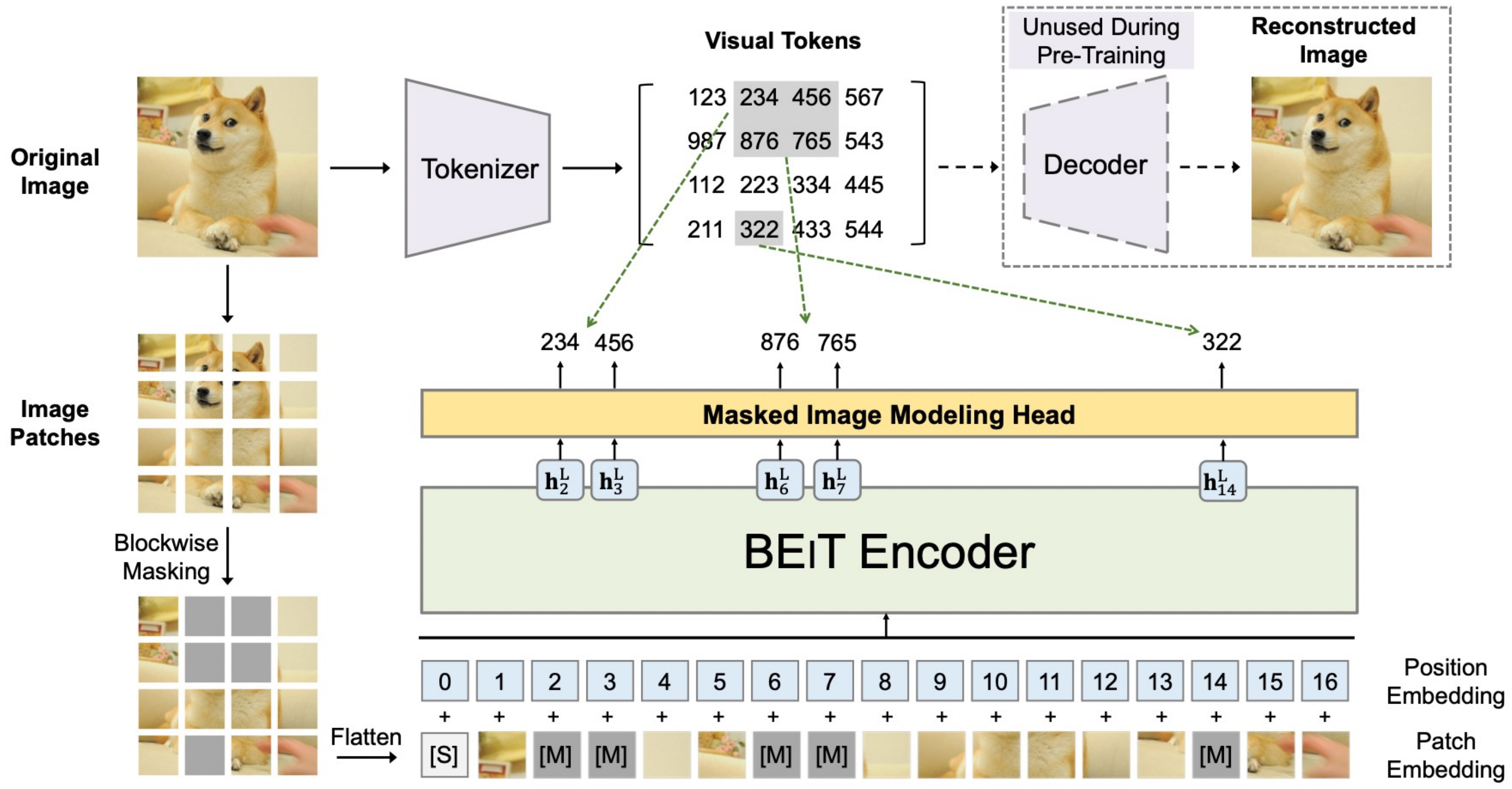
Uhlén, Mathias, et al. "Tissue-based map of the human proteome." *Science* 347.6220 (2015): 1260419.

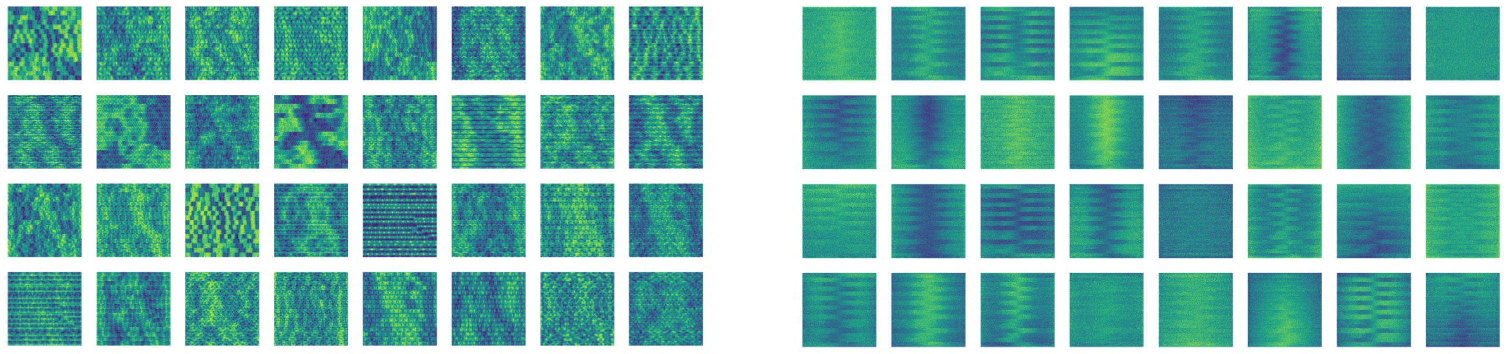

D
Main factor of variation: Cell lines
Nested factor of variation: Protein localization
DF
DF
UMAP 2
UMAP 1


A
mRNA similarity matrix
DINO similarity matrix
B
Canonical Correlation Analysis
RH-30
AF22
HAP1
BJ
EFO-21
HaCaT
CC 1
CC 2
C
DINO similarity matrix
Nucleus
Cytoplasm
D
Pseudotime analysis
Interphase
prophase
early prometaphase
prometaphase/metaphase
anaphase/telophase paired
anaphase/telophase unpaired
E
Phases similarity matrix
Anaphase / telophase unpaired
Prometaphase / metaphase
Anaphase / telophase paired
Early prometaphase
Interphase
Prophase
F
Treatment similarity
Tubulin polymerization inhibitor
Others
CDK inhibitor
Others
HSP inhibitor

Conclusion and Future Work

Masked image modeling



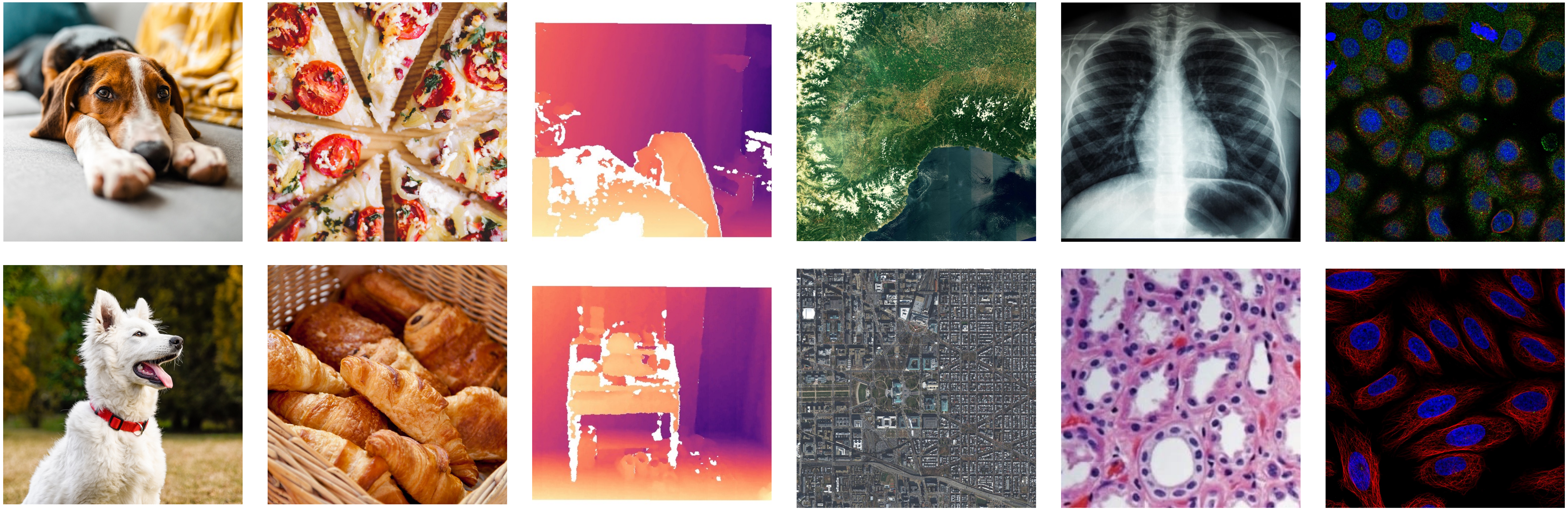


Carreira, Joao, et al. "Hierarchical perceiver." *arXiv preprint arXiv:2202.10890* (2022).

Bao, Hangbo, Li Dong, and Furu Wei. "Beit: Bert pre-training of image transformers." *arXiv preprint arXiv:2106.08254*(2021).

He, Kaiming, et al. "Masked autoencoders are scalable vision learners." *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2022.

Learning Universal Visual Representations



Physics data?

Meta AI