Self-Supervised Learning from Images with a Joint-Embedding Predictive Architecture

June 1st 2023

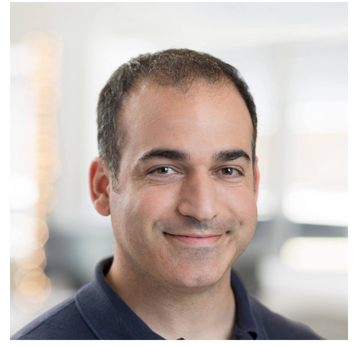
Mido Assran


Quentin Duval


Ishan Misra


Piotr Bojanowski


Pascal Vincent


Mike Rabbat


Yann LeCun


Nicolas Ballas

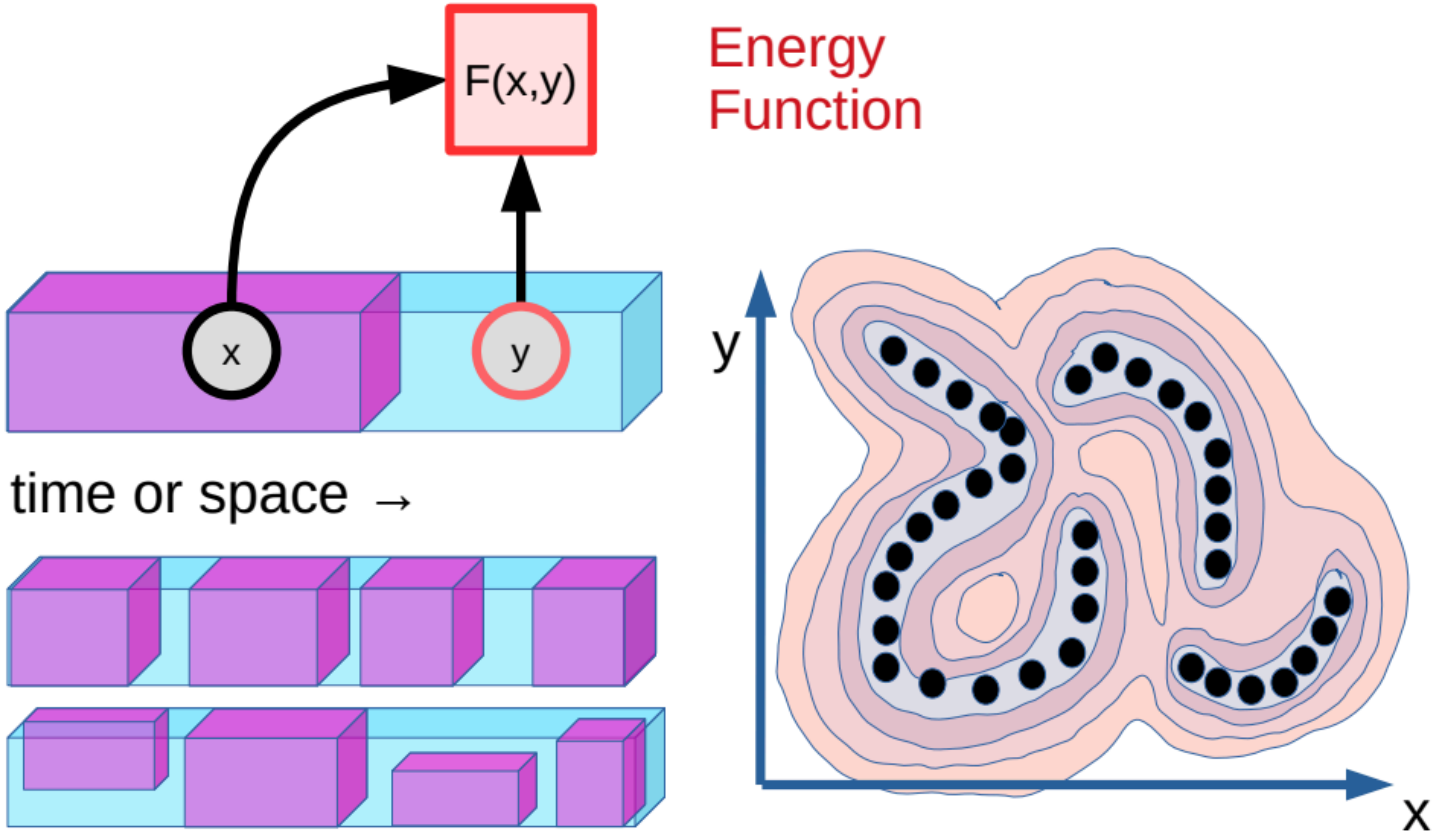
McGill Mila Meta AI

Self-Supervised Learning

Towards General Representations

Approach

Learn to represent data by capturing mutual dependencies between inputs

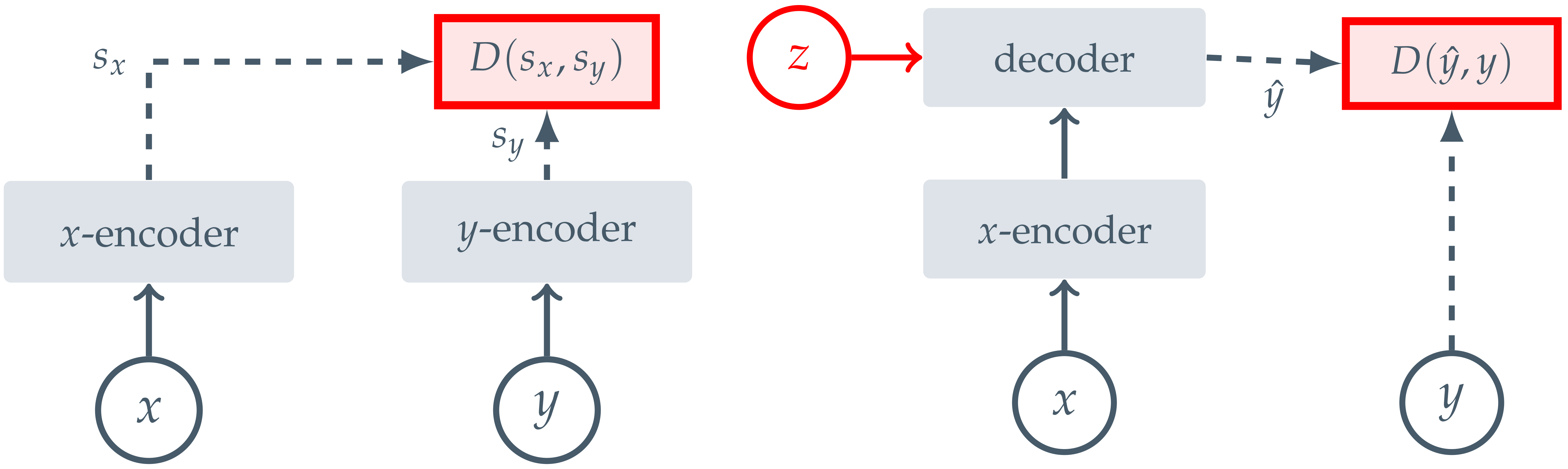


Yann LeCun
A Path Towards Autonomous Machine Intelligence
OpenReview, 2022.

What do we have so far?

Common Approaches for Visual Representation Learning

Learn representations by capturing mutual dependencies between inputs…



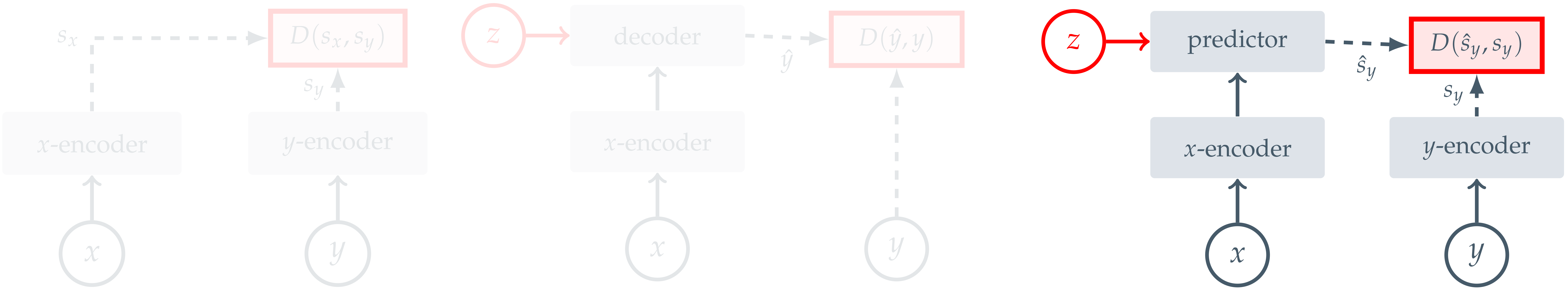
Joint-Embedding Architecture

Generative Architecture

What do we have so far?

Common Approaches for Visual Representation Learning

Learn representations by capturing mutual dependencies between inputs…



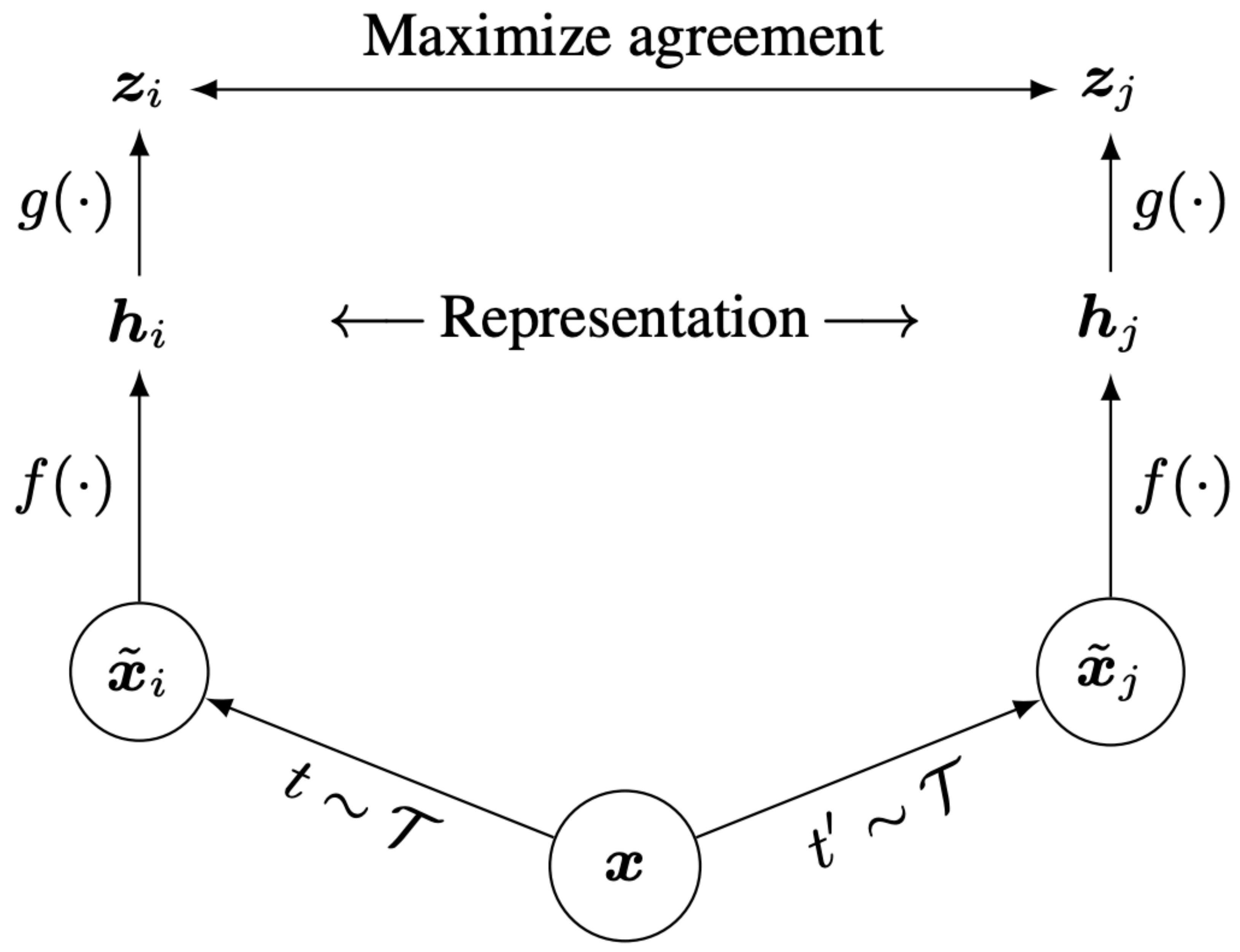
Joint-Embedding Architecture

Generative Architecture

Joint-Embedding Predictive Architecture

What do we have so far?

Common Approaches for Visual Representation Learning



Joint-Embedding Architecture

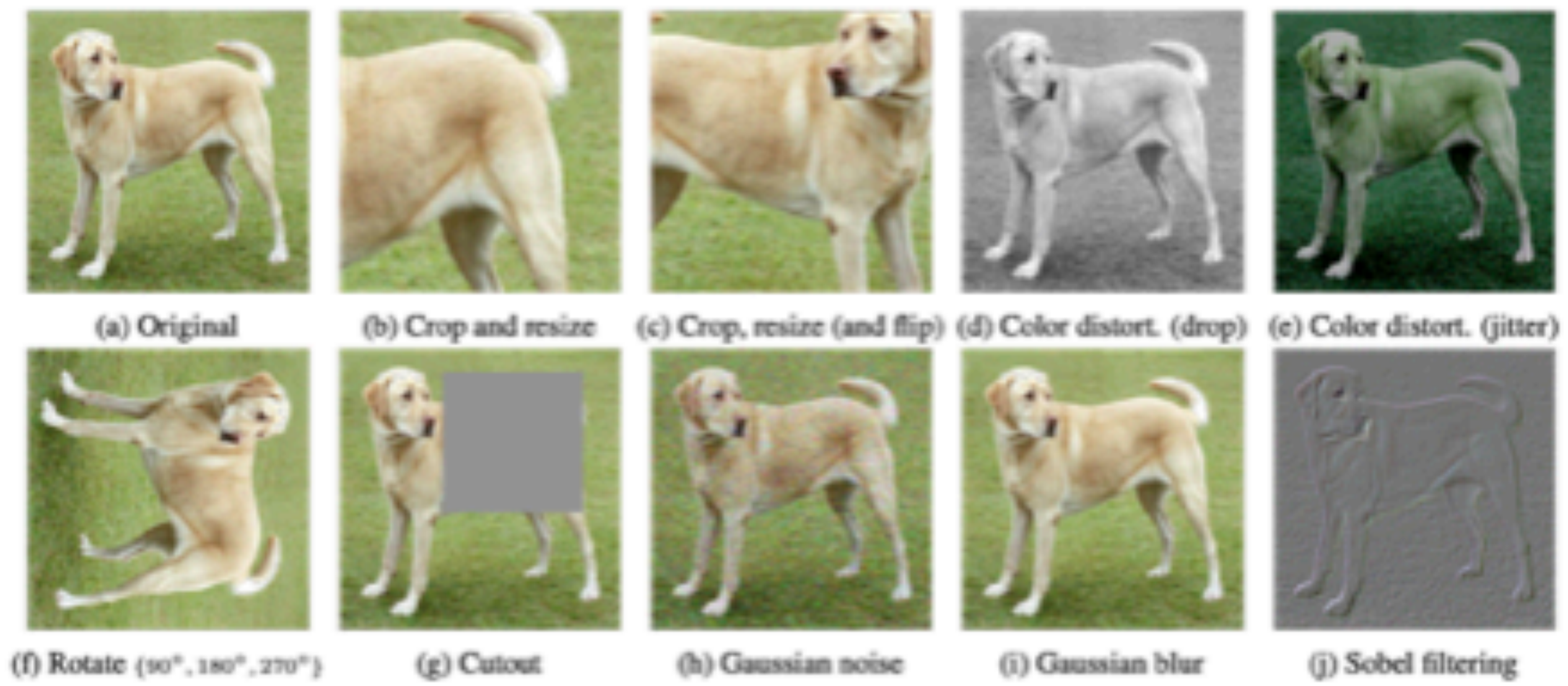


Figure 4. Illustrations of the studied data augmentation operators. Each augmentation can transform data stochastically with some internal parameters (e.g. rotation degree, noise level). Note that we *only* test these operators in ablation, the *augmentation policy used to train our models* only includes *random crop (with flip and resize), color distortion*, and *Gaussian blur*. (Original image cc-by: Von.grzanka)

Chen et al.,
A Simple Framework for Contrastive Learning of Visual Representations
ICML, 2020.

Self-Supervised Methods

Canonical Joint-Embedding Architecture

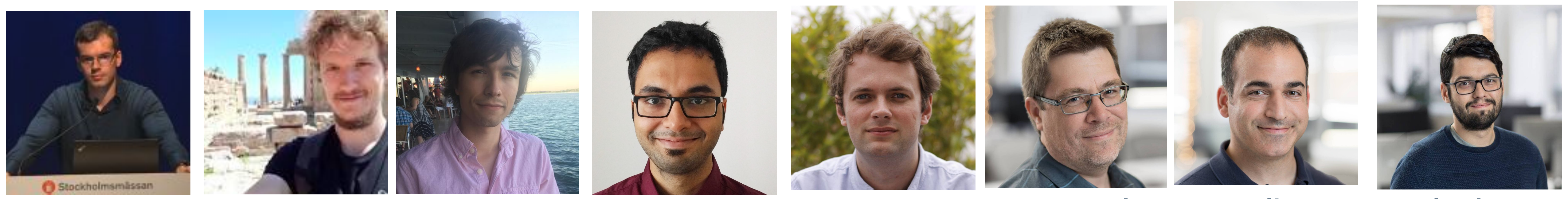
Limitations:

Semantic level of representations also depends on certain assumptions…

Published as a conference paper at ICLR 2023

THE HIDDEN UNIFORM CLUSTER PRIOR IN SELF-SUPERVISED LEARNING

Mahmoud Assran*[1,2,3], Randall Balestriero[1], Quentin Duval[1], Florian Bordes[1,3,4]
Ishan Misra[1], Piotr Bojanowski[1], Pascal Vincent[1,3,4], Michael Rabbat[1,3], Nicolas Ballas[1]
[1]Meta AI (FAIR) [2]McGill University, ECE [3]Mila, Quebec AI Institute [4]Universite de Montreal, DIRO



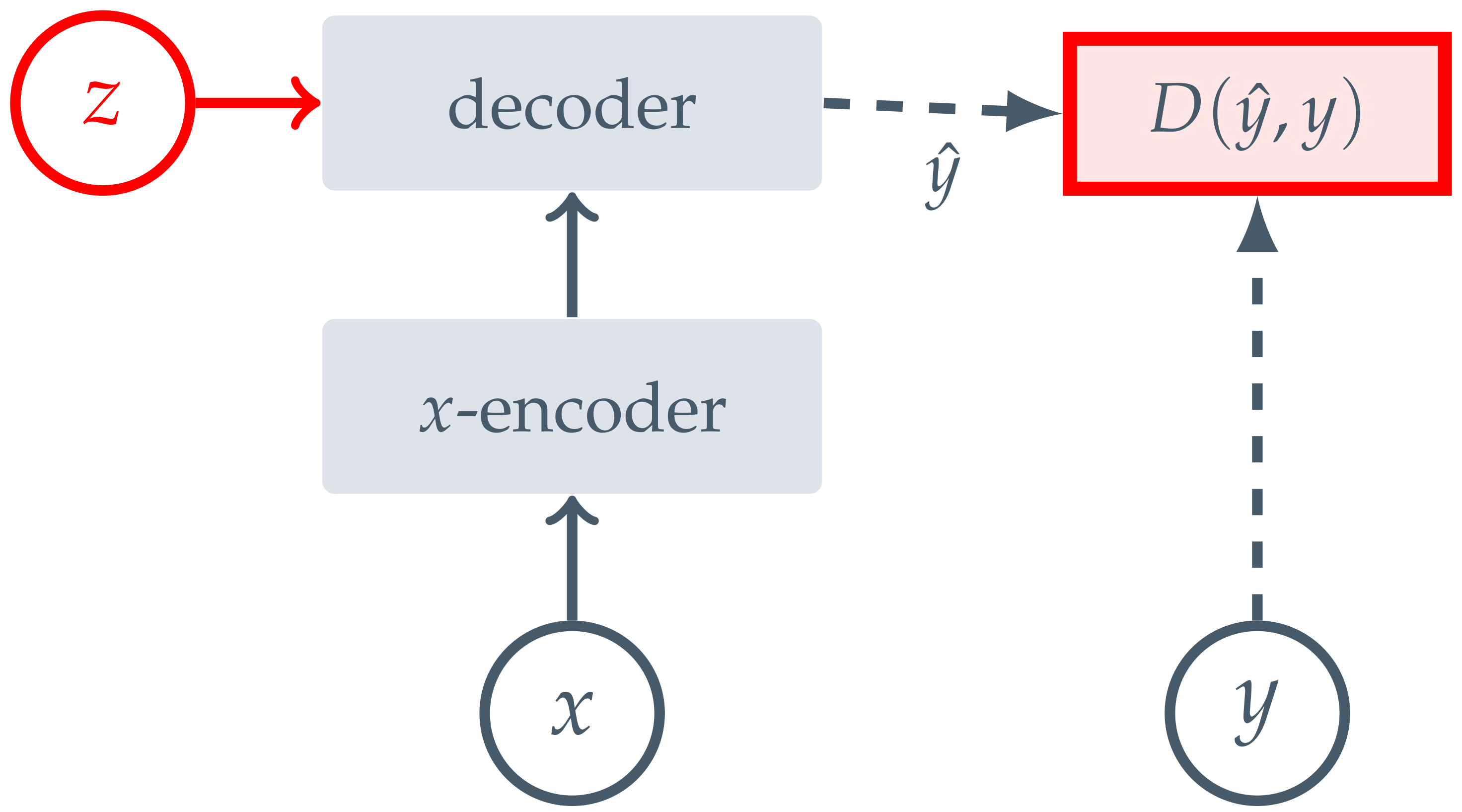
Randall Balestriero · Quentin Duval · Florian Bordes · Ishan Misra · Piotr Bojanowski · Pascal Vincent · Mike Rabbat · Nicolas Ballas

What do we have so far?

Common Approaches for Visual Representation Learning

Learn representations by capturing mutual dependencies between inputs...

Joint-Embedding Architecture



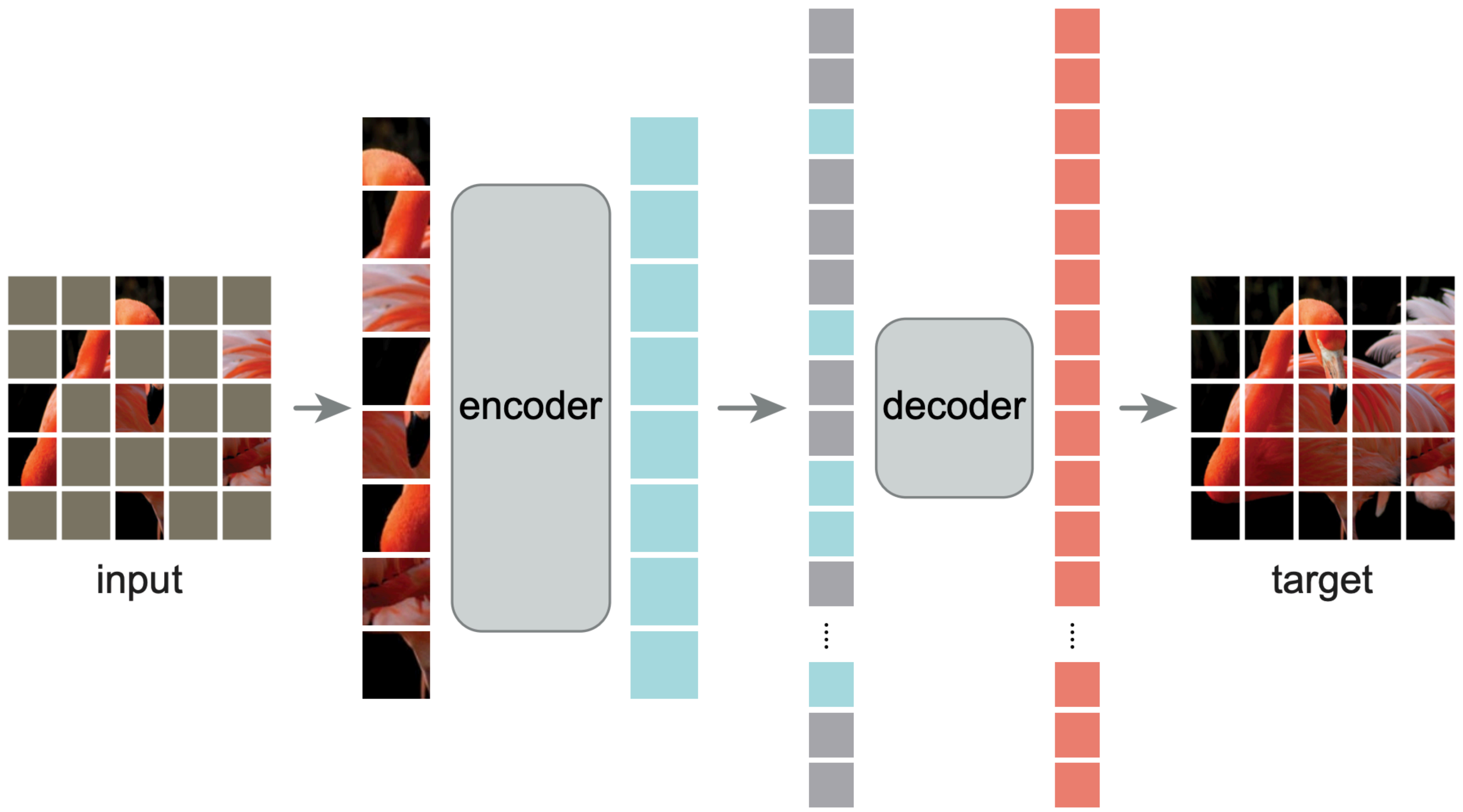
Generative Architecture

What do we have so far?

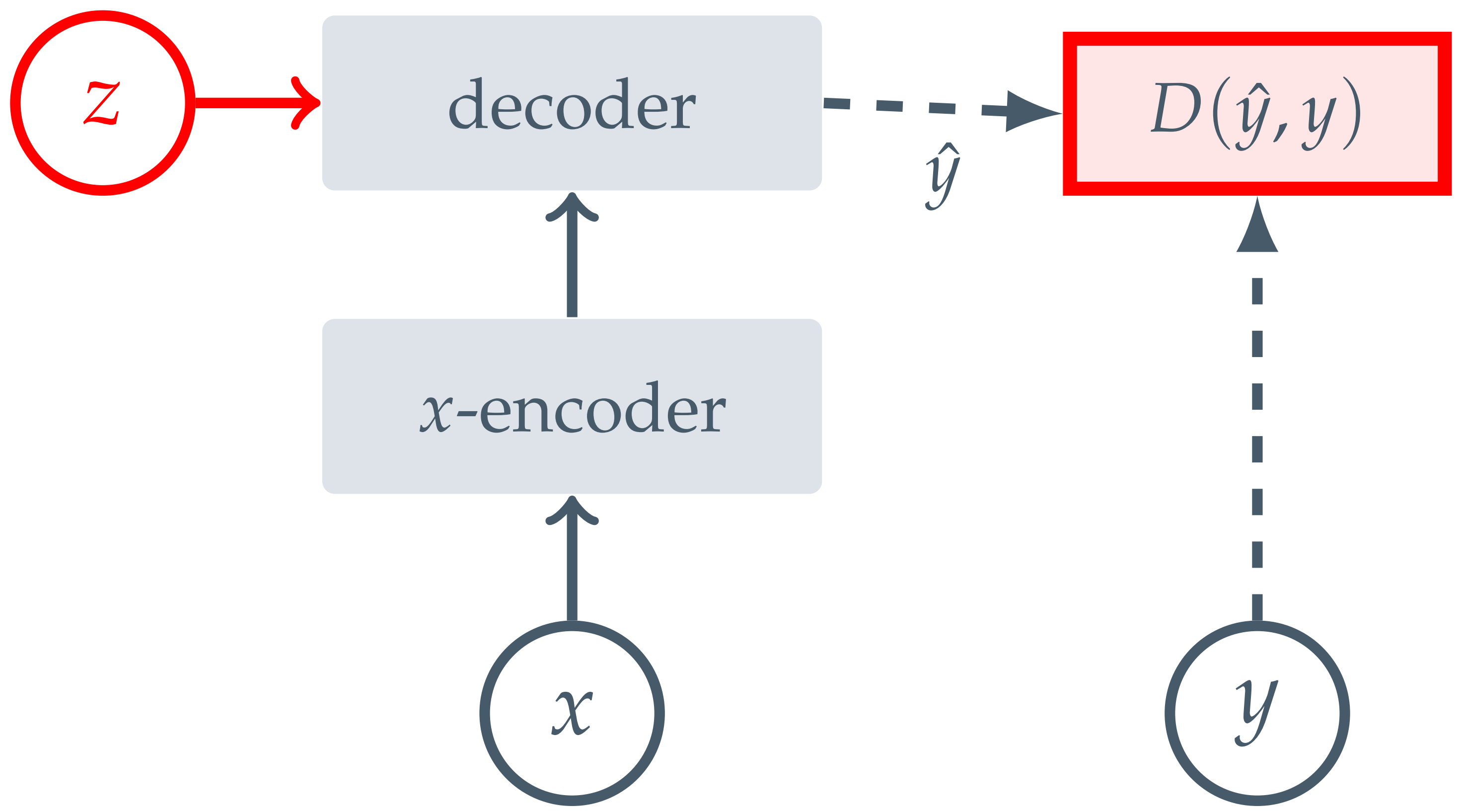
Common Approaches for Visual Representation Learning

Learn representations by capturing mutual dependencies between inputs…

Mask Denoising Architectures



He et al.,
Masked Autoencoders Are Scalable Vision Learners
arXiv, 2021.



Generative Architecture

Self-Supervised Learning

Generative Architectures

Generative architectures tend to learn representations of a lower semantic level…

Published as a conference paper at ICLR 2023

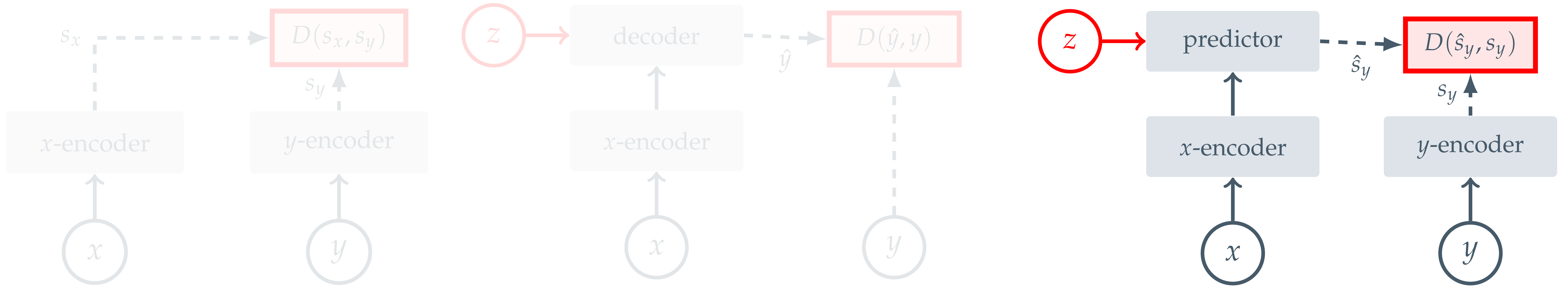
WHAT DO SELF-SUPERVISED VISION TRANSFORMERS LEARN?

Namuk Park[1*] **Wonjae Kim**[2] **Byeongho Heo**[2] **Taekyung Kim**[2] **Sangdoo Yun**[2]
[1]Prescient Design, Genentech [2]NAVER AI Lab
park.namuk@gene.com {wonjae.kim,bh.heo,taekyung.k,sangdoo.yun}@navercorp.com

What do we have so far?

Common Approaches for Visual Representation Learning

Learn representations by capturing mutual dependencies between inputs…



Joint-Embedding Architecture

Generative Architecture

Joint-Embedding Predictive Architecture

Image JEPA

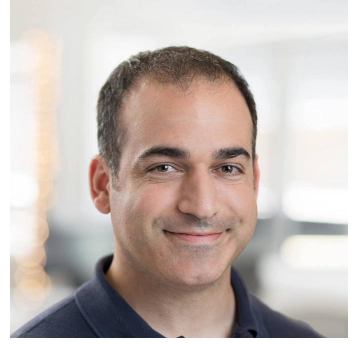
Towards More General Representations


Quentin Duval


Ishan Misra


Piotr Bojanowski


Pascal Vincent


Mike Rabbat


Yann LeCun


Nicolas Ballas

From a single context block, predict representations of various target blocks…

… no hand-crafted data augmentations!

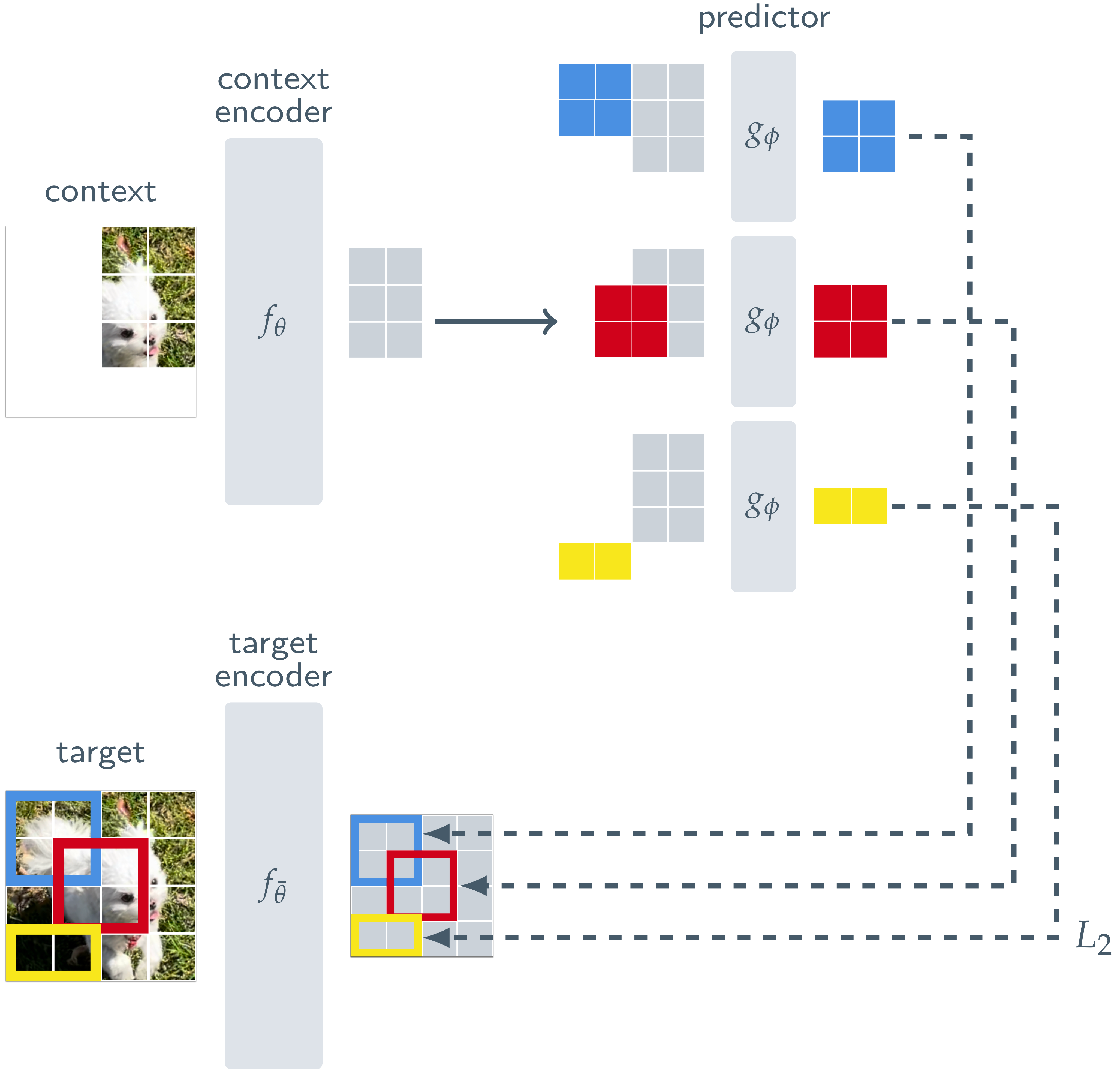
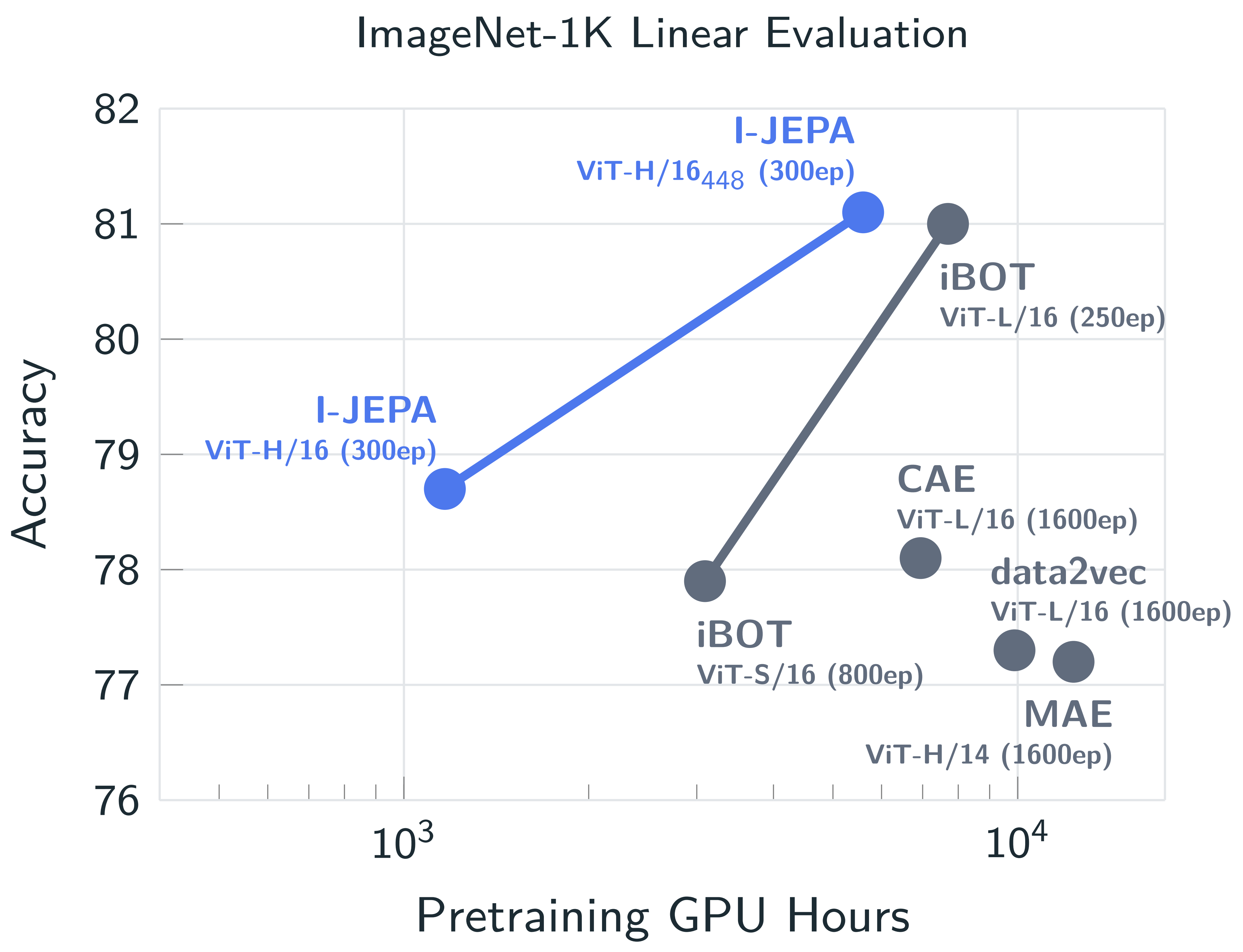


Image JEPA

Towards More General Representations



Very efficient…

Training ViT-Huge/16 with I-JEPA is faster than training ViT-Small/16 with iBOT!

Converges faster than generative methods, which require many epochs of pre-training (MAE)

Image JEPA

Towards More General Representations

Targets	Arch.	Epochs	Top-1
Target-Encoder Output	ViT-L/16	500	**66.9**
Pixels	ViT-L/16	800	40.7

Table 7. **Ablating targets**. Linear evaluation on ImageNet-1K using only 1% of the available labels; ablating the effect of the prediction targets during I-JEPA pretraining. To ensure convergence when predicting in pixel space, we trian the model for more epochs. The semantic level of the I-JEPA representations degrades significantly when the loss is applied in pixel space, rather than representation space, highlighting the importance of the target-encoder during pretraining.

I-JEPA is non-generative…

Same method in pixel space performs much worse on semantic classification tasks…

Image JEPA

Towards More General Representations

Image JEPA

Towards More General Representations

Method	Arch.	CIFAR100	Places205	iNat18
Methods without view data augmentations				
data2vec [7]	ViT-L/16	81.6	54.6	28.1
MAE [34]	ViT-H/14	77.3	55.0	32.9
I-JEPA	ViT-H/14	**87.5**	**58.4**	**47.6**
Methods using extra view data augmentations				
DINO [17]	ViT-B/8	84.9	57.9	55.9
iBOT [74]	ViT-L/16	**88.3**	**60.4**	**57.3**

Linear transfer to semantic image-level visual tasks

I-JEPA captures global semantics…

- Outperforms generative methods
- Closes gap with view-invariance methods

McGill Mila Meta AI

Image JEPA

Towards More General Representations

Method	Arch.	Clevr/Count	Clevr/Dist
Methods without view data augmentations			
data2vec [7]	ViT-L/16	85.3	71.3
MAE [34]	ViT-H/14	**90.5**	**72.4**
I-JEPA	ViT-H/14	86.7	**72.4**
Methods using extra data augmentations			
DINO [17]	ViT-B/8	86.6	53.4
iBOT [74]	ViT-L/16	85.7	62.8

Linear Transfer to Low-Level Visual Tasks

I-JEPA also captures local information…

- Outperforms view-invariance methods in low-level tasks (e.g., depth prediction)
- Comparable with generative methods

Image JEPA

Towards More General Representations

Method	Arch.	Epochs	Top-1
Methods without view data augmentations			
data2vec [7]	ViT-L/16	1600	73.3
MAE [34]	ViT-L/16	1600	67.1
	ViT-H/14	1600	71.5
I-JEPA	ViT-L/16	600	69.4
	ViT-H/14	300	73.3
	ViT-H/16_{448}	300	**77.3**
Methods using extra view data augmentations			
iBOT [74]	ViT-B/16	250	69.7
DINO [17]	ViT-B/8	300	70.0
SimCLR v2 [33]	RN151 (2×)	800	70.2
BYOL [33]	RN200 (2×)	800	71.2
MSN [3]	ViT-B/4	300	**75.7**

ImageNet 1% Semi-Supervised Evaluation

Scaling I-JEPA...

New SoTA for ImageNet semi-supervised eval...