Emergent structure in Representation Learning Application & Generalization

Samuel Lavoie

All the impressive achievements of deep learning amount to just curve fitting.

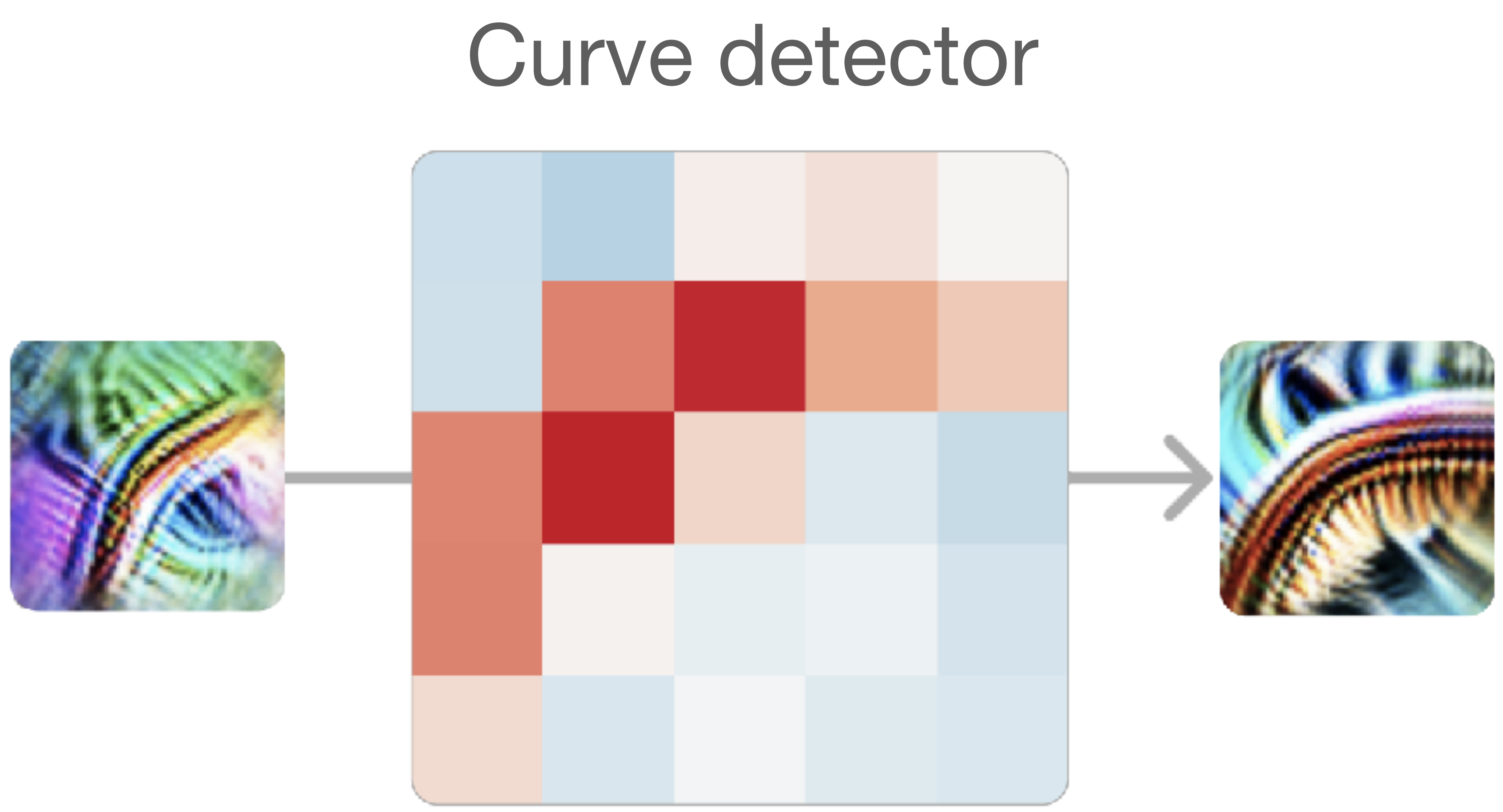
Judea pearl — https://www.quantamagazine.org/to-build-truly-intelligent-machines-teach-them-cause-and-effect-20180515/

Emergent property

Property P of a system S with micro-dynamics D is emergent iff P can be derived from D and the external conditions of S.

Mark bedau. Week Emergence* 1997

Example of emergent property



Chris Ollah et al. An Introduction to Circuits. 2020

Example of emergent property

Unsupervised Domain Translation

Jun-Yan Zhu et al. Unpaired Image-to-Image Translation using Cycle-Consistent Adversarial Networks. 2017

Properties of Unsupervised Domain Translation

- Preserve pose.
- Transfer textural properties.
- Requires very few samples.
  about 1000 for horse-zebra.

Shortcoming of Unsupervised Domain Translation

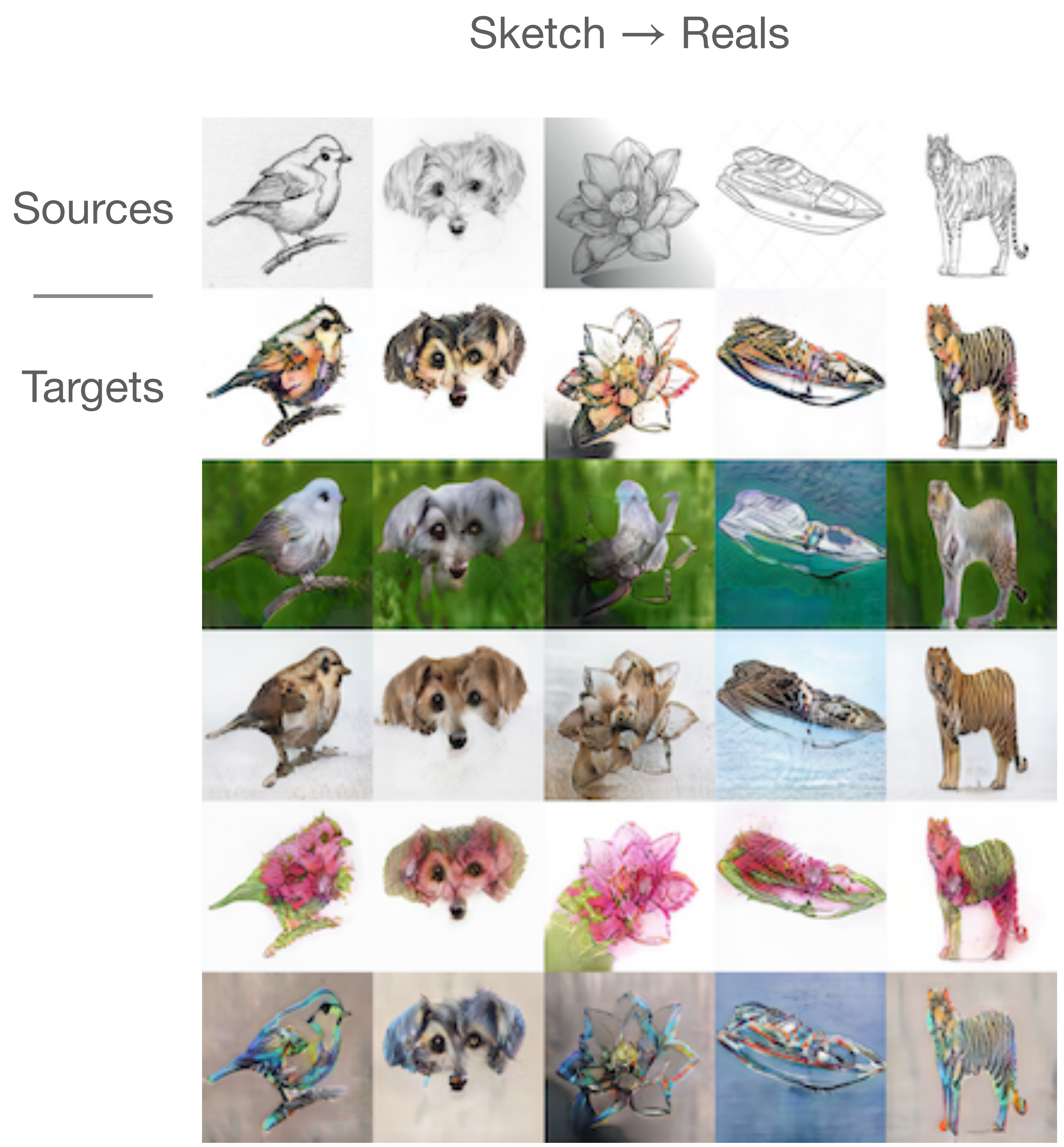
Does not preserve high-order attributes.

MNIST → SVHN

SVHN → MNIST

Shortcoming of Unsupervised Domain Translation

Inconsistent style generation.



Integrating Categorical Semantics into Unsupervised Domain Translation

In collaboration with Faruk Ahmed and Aaron Courville

Potential approaches

Supervised	Unsupervised
Objectives leveraging labels	Inductive bias via the architecture
Objectives leveraging pairing	Unsupervised objectives
Objectives leveraging pre-trained representation	Objective leveraging pre-trained representation without supervision

Potential approaches

Supervised

Objectives leveraging labels

Objectives leveraging pairing

Objectives leveraging pre-trained representation

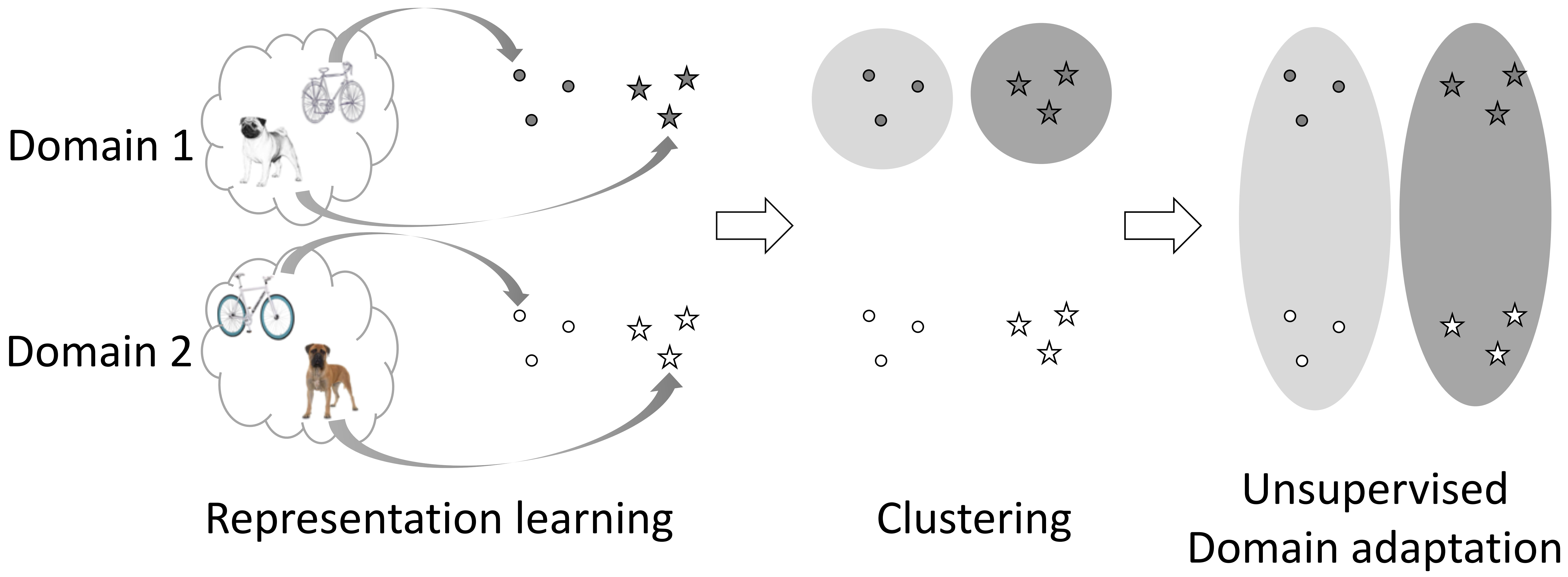
Unsupervised

Inductive bias via the architecture

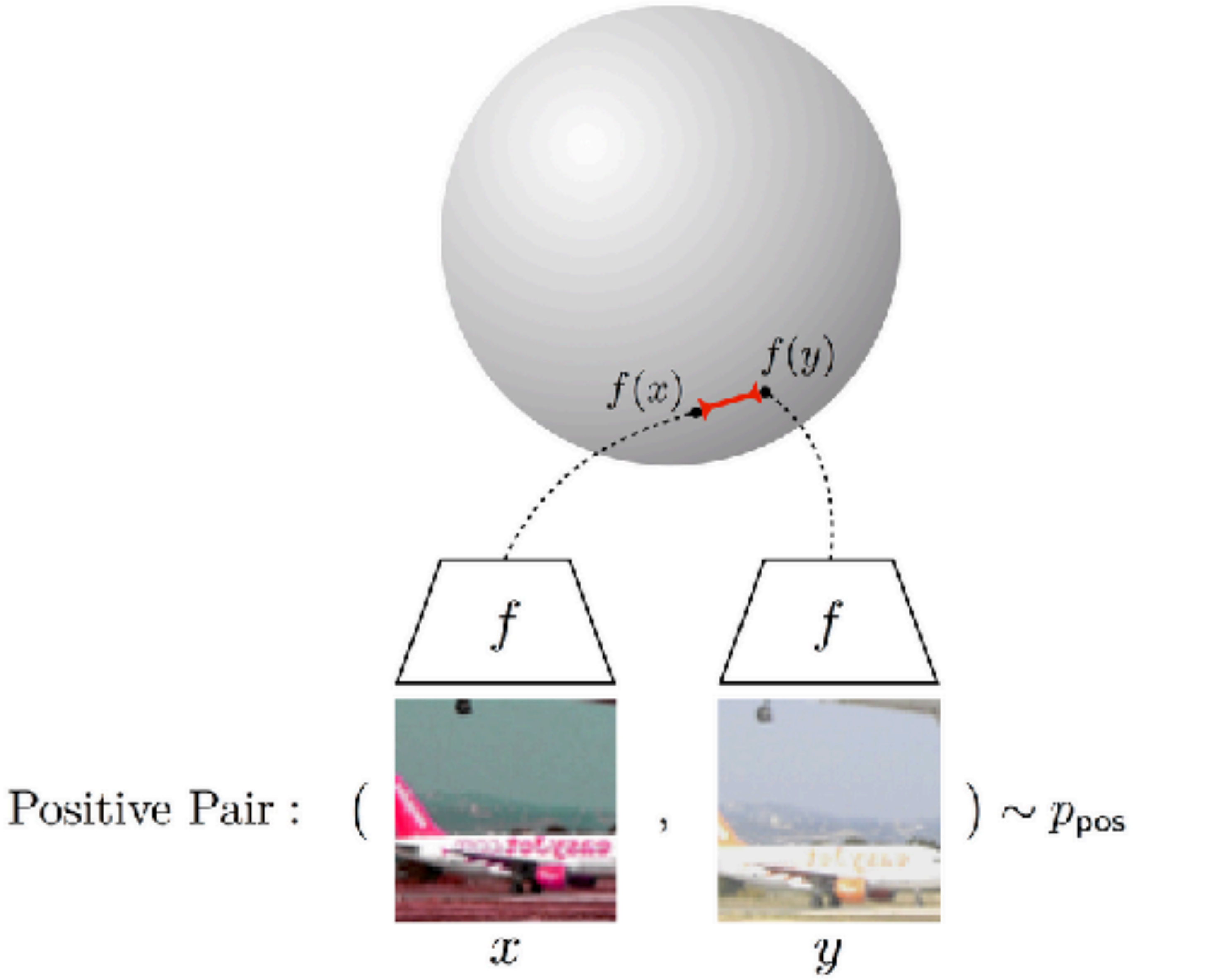
Unsupervised objectives

Objective leveraging pre-trained representation without supervision

Learning domain invariant representation without supervision



Representation learning — Self-supervised learning

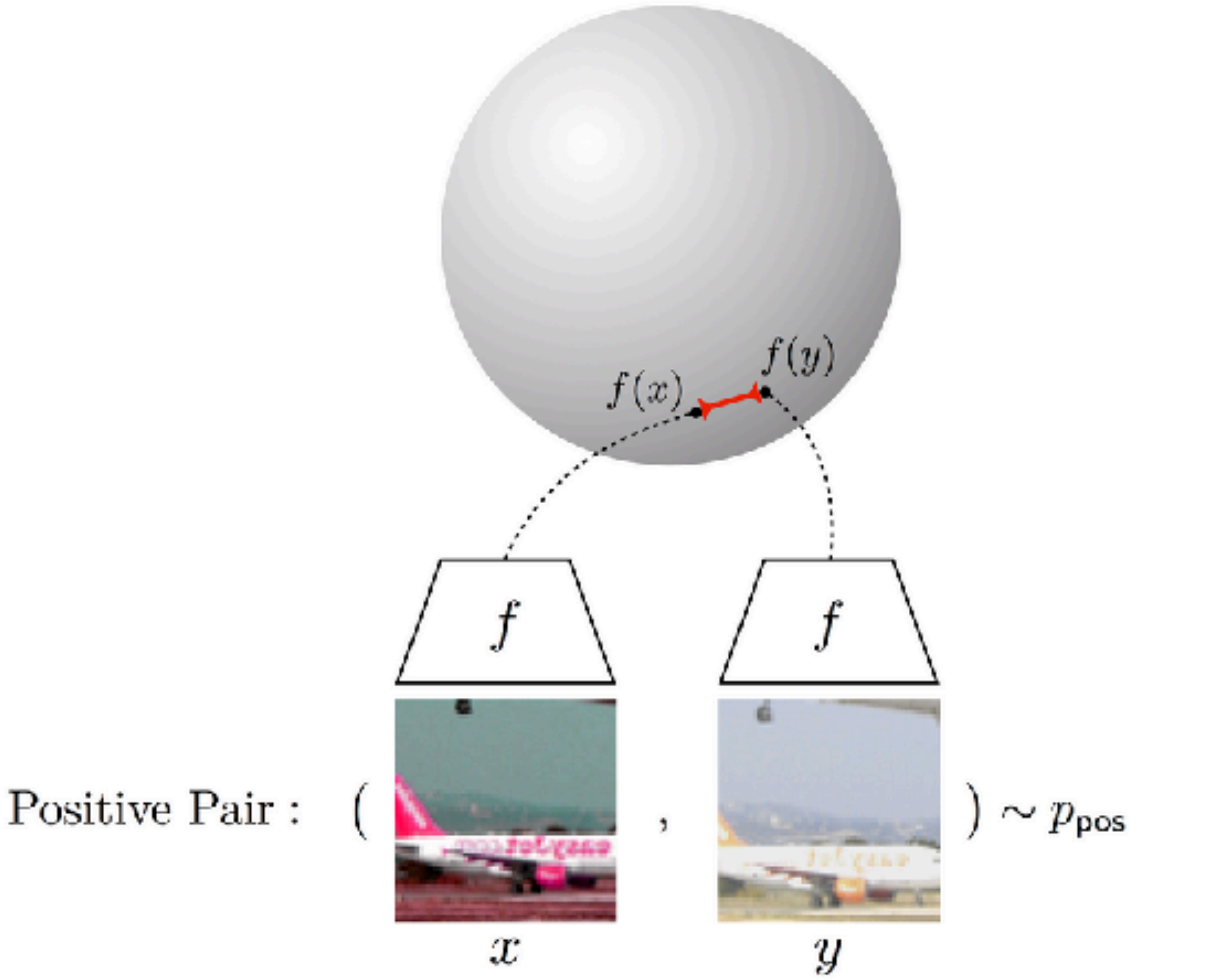


Alignment: Similar samples have similar features

Uniformity: Preserve maximal information

Tongzhou Wang and Phillip Isola. Understanding Contrastive Representation through Alignment and Uniformity on the Hypersphere. 2020

Representation learning — Self-supervised learning



Alignment: Similar samples have similar features

Uniformity: Preserve maximal information

Noise contrastive estimation

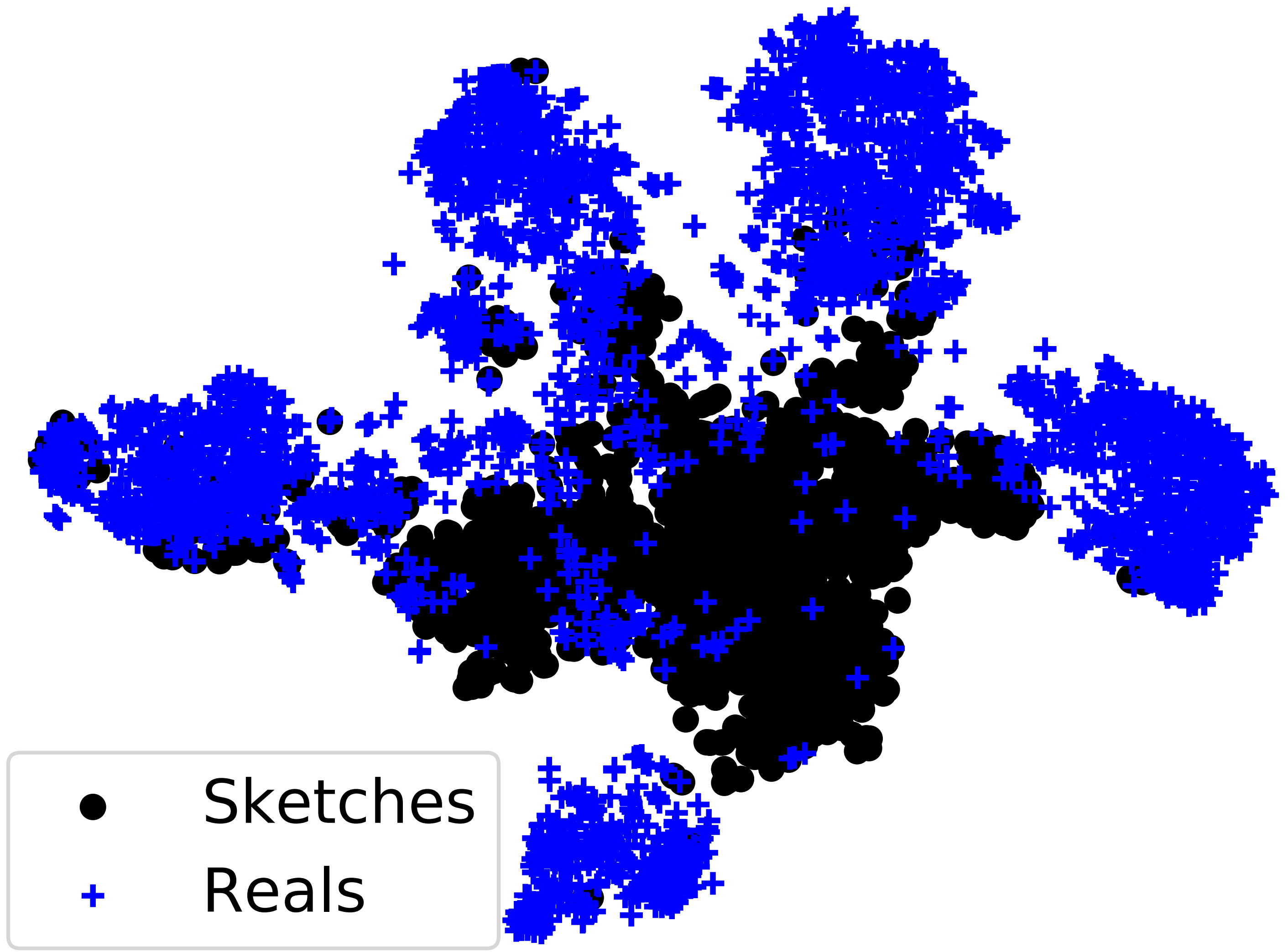
$$\mathcal{L}_{\text{nce}} := -\log \frac{\exp(d(f(x), f(y))/\tau)}{\sum_{\bar{x} \in \mathcal{X} \backslash x} \exp(d(f(x), f(\bar{x}))/\tau)}$$

Tongzhou Wang and Phillip Isola. Understanding Contrastive Representation through Alignment and Uniformity on the Hypersphere. 2020

Property of a model learned with contrastive learning

Samples cluster in dense region

Samples from different domain do not intersect

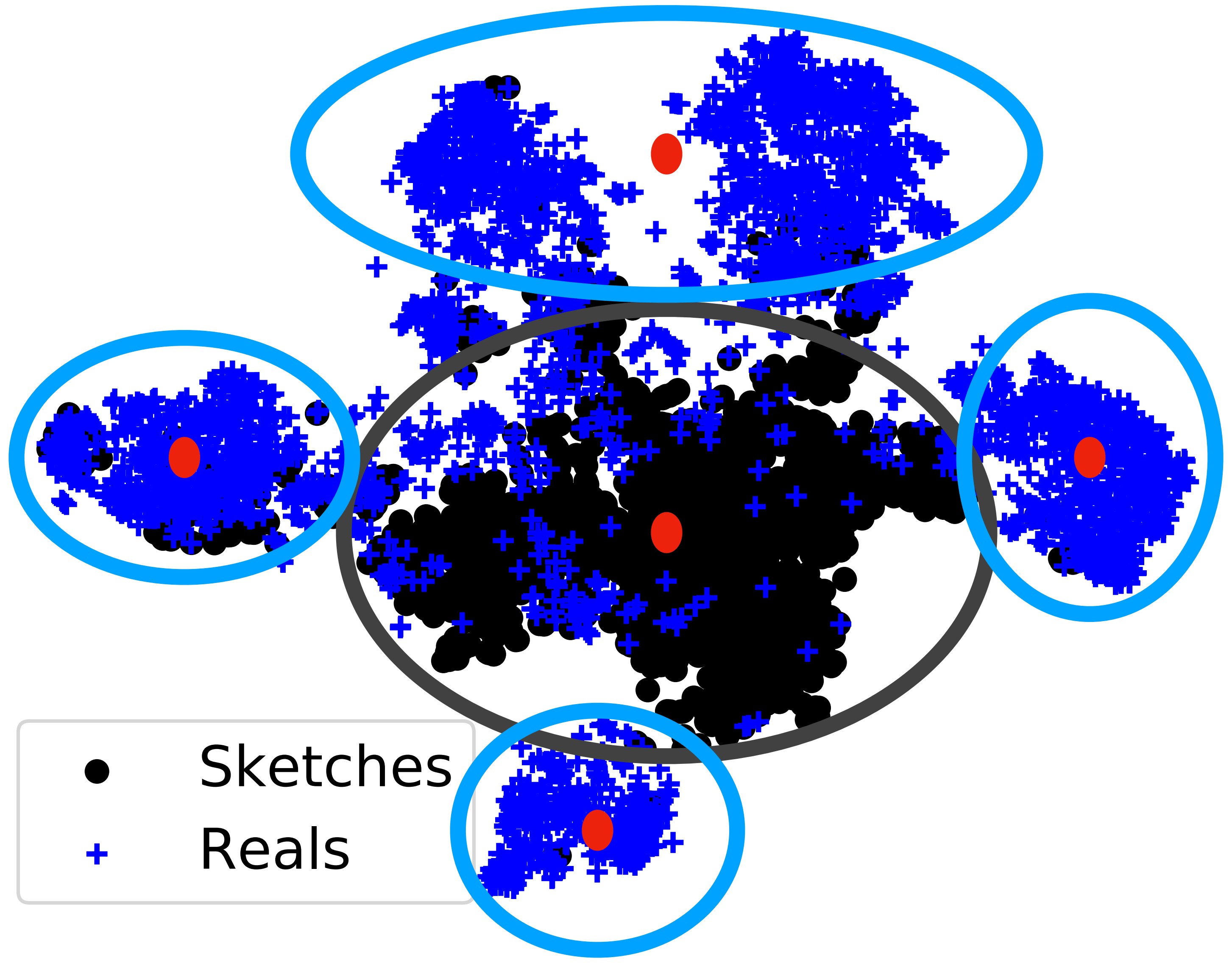


Embedding of 5 categories: bird, dog, flower, boat, tiger

Property of a model learned with contrastive learning

Samples cluster in dense region

Samples from different domain do not intersect

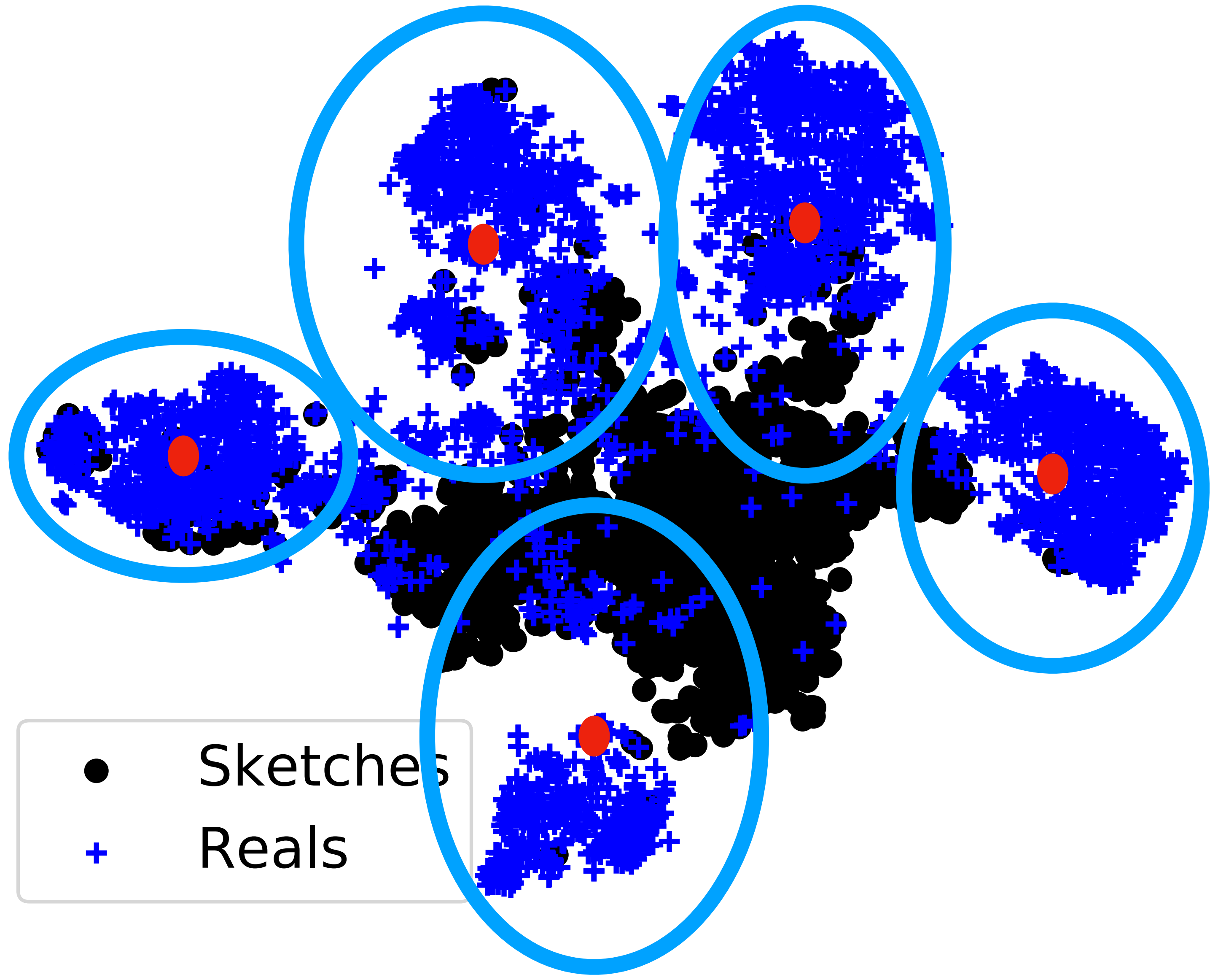


Embedding of 5 categories: bird, dog, flower, boat, tiger

Property of a model learned with contrastive learning

Samples cluster in dense region

Samples from different domain do not intersect

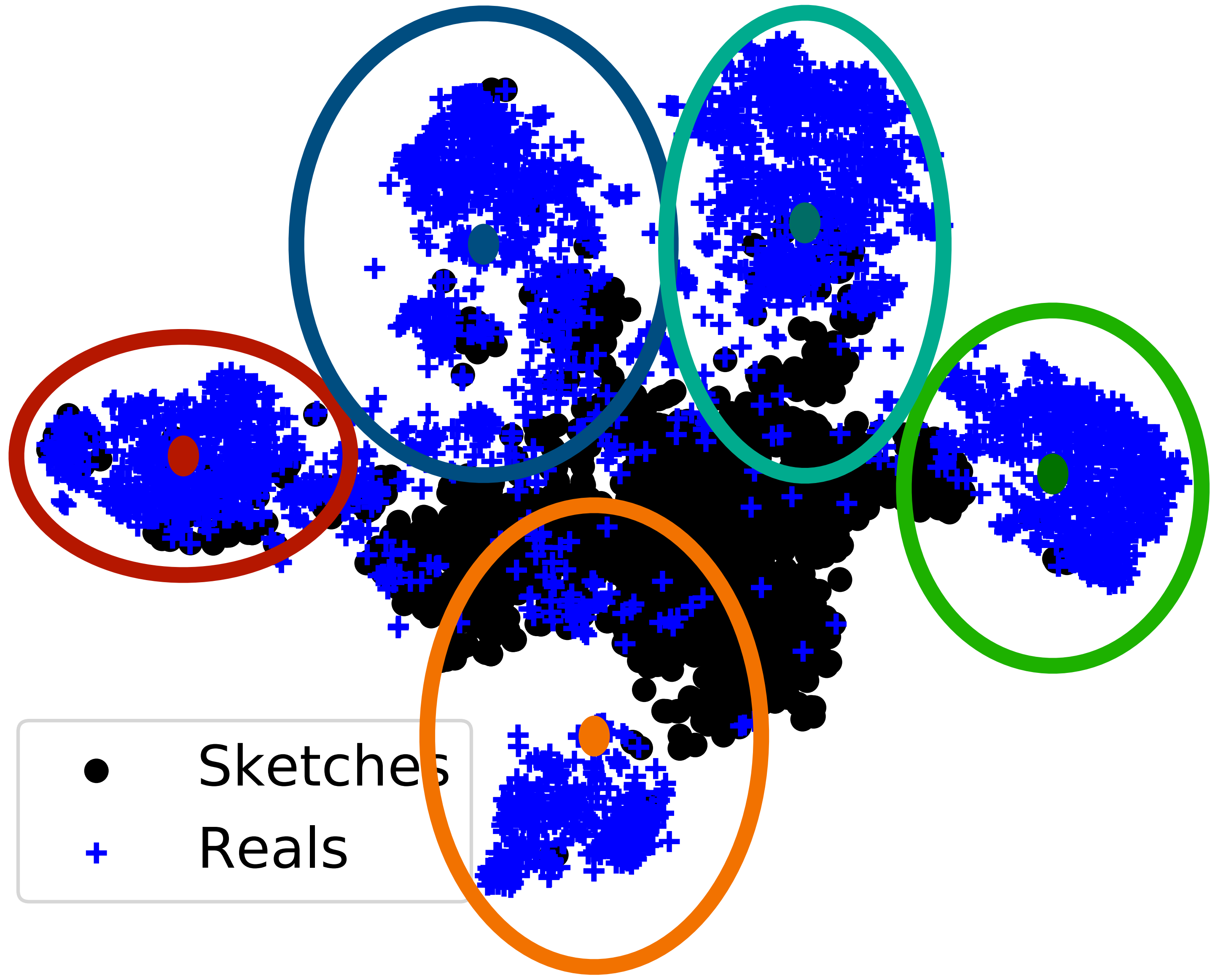


Embedding of 5 categories: bird, dog, flower, boat, tiger

Property of a model learned with contrastive learning

Samples cluster in dense region

Samples from different domain do not intersect

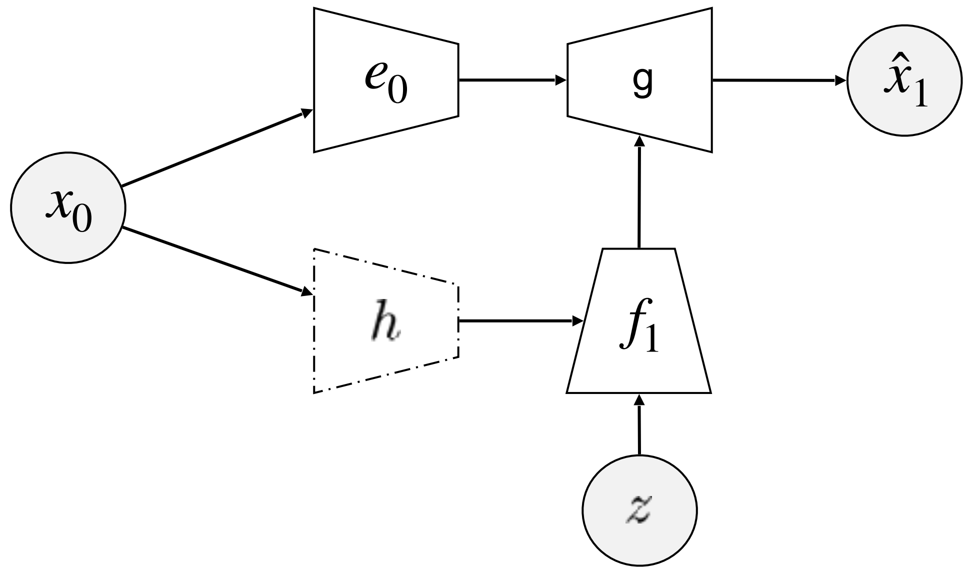


Embedding of 5 categories: bird, dog, flower, boat, tiger

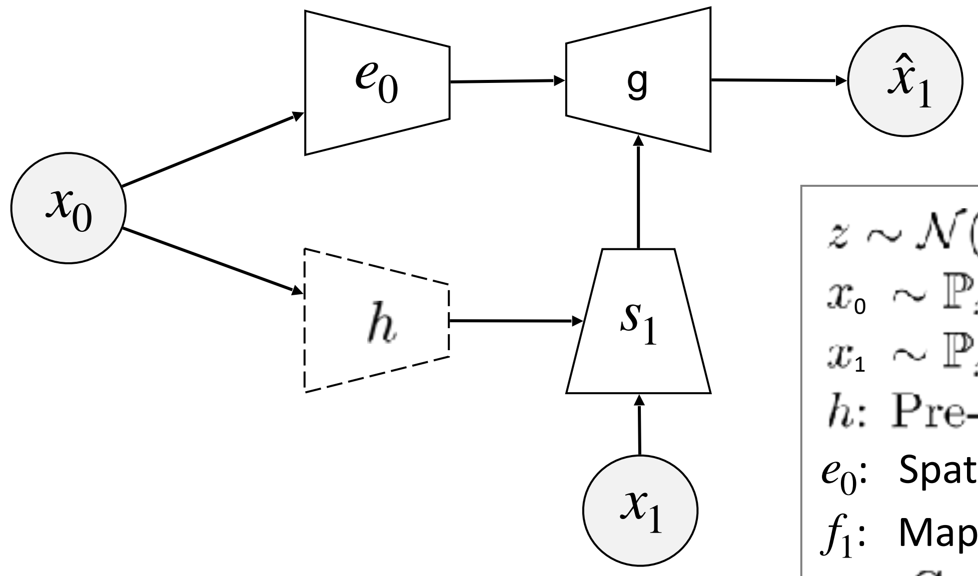
Define the clusters as pseudo-labels and adapt them to the sketches using Unsupervised Domain Adaptation

Integrate the learned semantics into Unsupervised Domain Translation

Condition style generation



Conditioning style on z

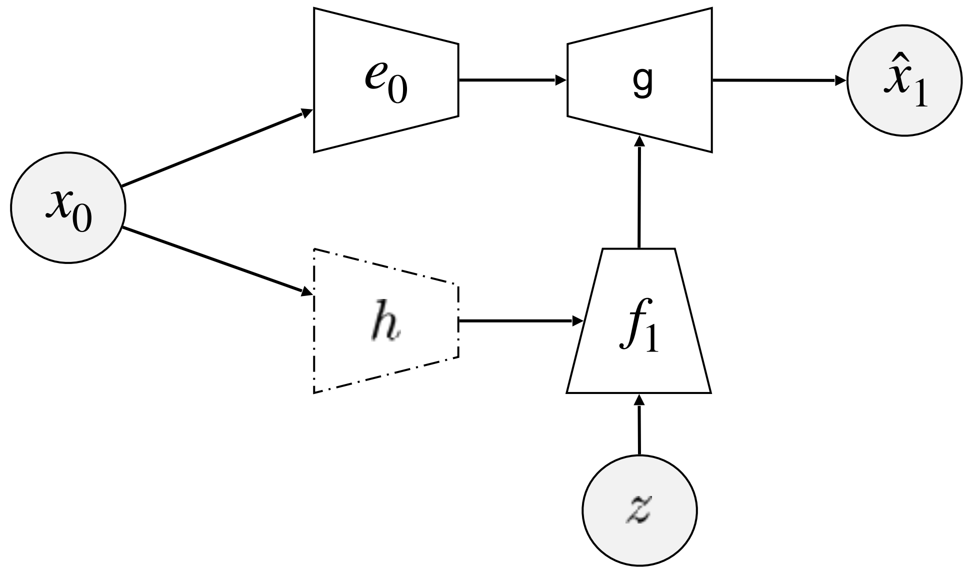


Conditioning style on x_1

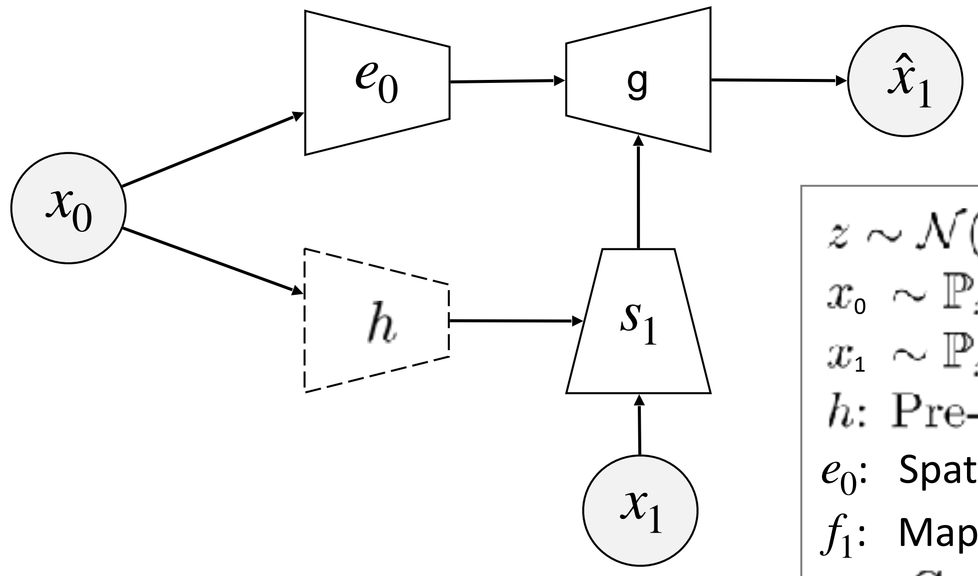
$z \sim \mathcal{N}(\mathbf{0}, \mathbf{1})$
$x_0 \sim \mathbb{P}_{x_0}$
$x_1 \sim \mathbb{P}_{x_1|x_0}$
h: Pre-trained semantics-encoder
e_0: Spatial encoder (domain 0)
f_1: Mapping network (domain 1)
g: Generator
s_1: Style encode (domain 1)
$\hat{x}_1$: Generated sample (domain 1)

Integrate the learned semantics into Unsupervised Domain Translation

Condition style generation



Conditioning style on z



Conditioning style on x_1

$z \sim \mathcal{N}(\mathbf{0}, \mathbf{1})$
$x_0 \sim \mathbb{P}_{x_0}$
$x_1 \sim \mathbb{P}_{x_1|x_0}$
h: Pre-trained semantics-encoder
e_0: Spatial encoder (domain 0)
f_1: Mapping network (domain 1)
g: Generator
s_1: Style encode (domain 1)
$\hat{x}_1$: Generated sample (domain 1)

Constraint mapping to preserve semantics

$$\mathscr{L} := -\sum_i h(x_0)_i \log(h(\hat{x}_1))_i$$

Results MNIST↔SVHN

MNIST→SVHN using our method

Results MNIST↔SVHN

MNIST→SVHN using our method

Sources
Style 1
Style 2
Style 3
Style 4
Style 5

Table 1: **Comparison with the baselines.** Domain translation accuracy and FID obtained on MNIST (M) ↔SVHN (S) for the different methods considered. The last column is the test classification accuracy of the classifier used to compute the metric. *: Using weak supervision.

	Data	CycleGAN	MUNIT	DRIT	Stargan-V2	EGSC-IT*	CatS-UDT	Target
Acc	M→S	10.89	10.44	13.11	28.26	47.72	**95.63**	98.0
	S→M	11.27	10.12	9.54	11.58	16.92	**76.49**	99.6
FID	M→S	46.3	55.15	127.87	66.54	72.43	**39.72**	-
	S→M	24.8	30.34	20.98	26.27	19.45	**6.60**	-

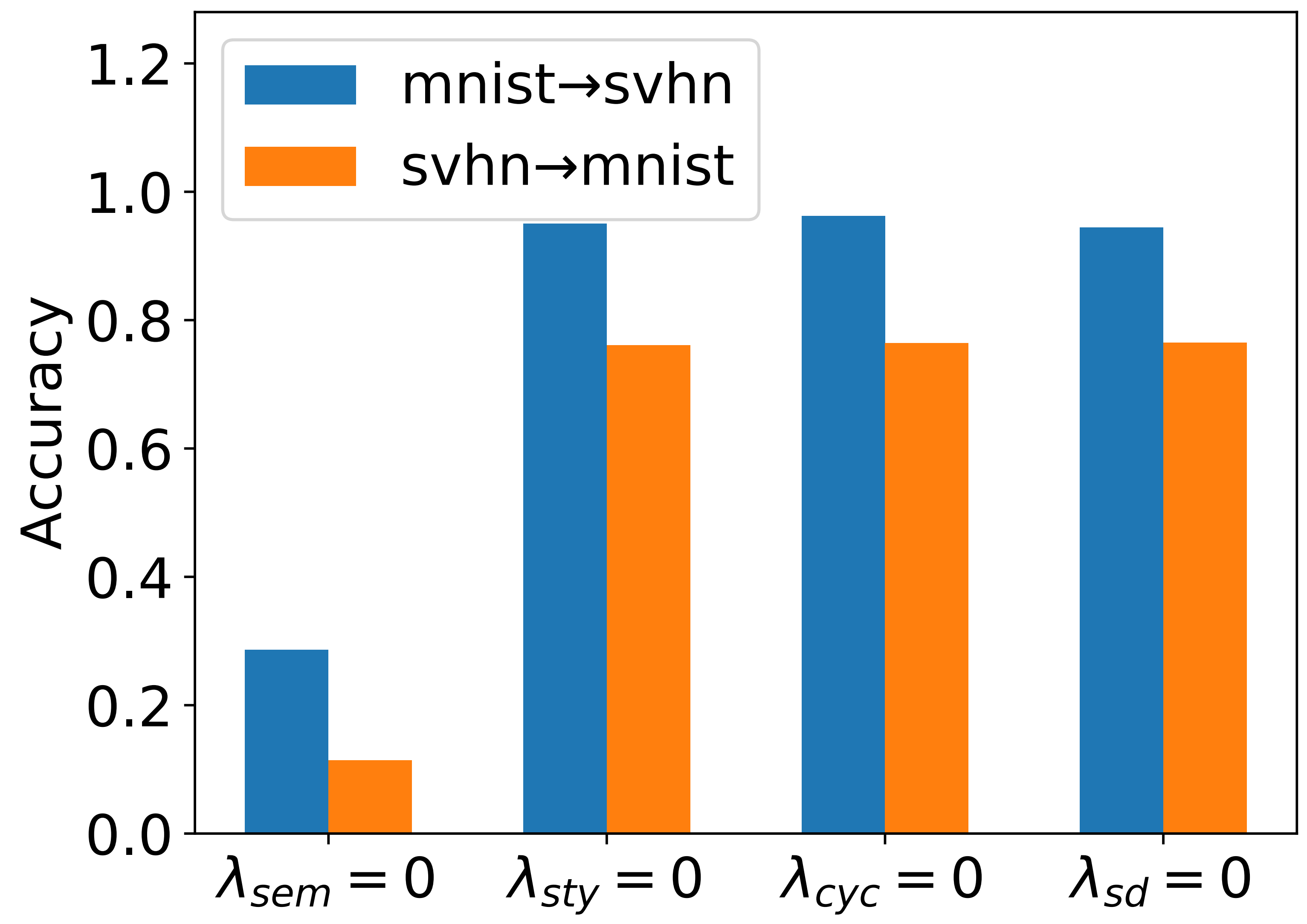
Results MNIST↔SVHN

Table 1: **Comparison with the baselines.** Domain translation accuracy and FID obtained on MNIST (M) ↔SVHN (S) for the different methods considered. The last column is the test classification accuracy of the classifier used to compute the metric. *: Using weak supervision.

	Data	CycleGAN	MUNIT	DRIT	Stargan-V2	EGSC-IT*	CatS-UDT	Target
Acc	M→S	10.89	10.44	13.11	28.26	47.72	**95.63**	98.0
	S→M	11.27	10.12	9.54	11.58	16.92	**76.49**	99.6
FID	M→S	46.3	55.15	127.87	66.54	72.43	**39.72**	-
	S→M	24.8	30.34	20.98	26.27	19.45	**6.60**	-

MNIST→SVHN using our method

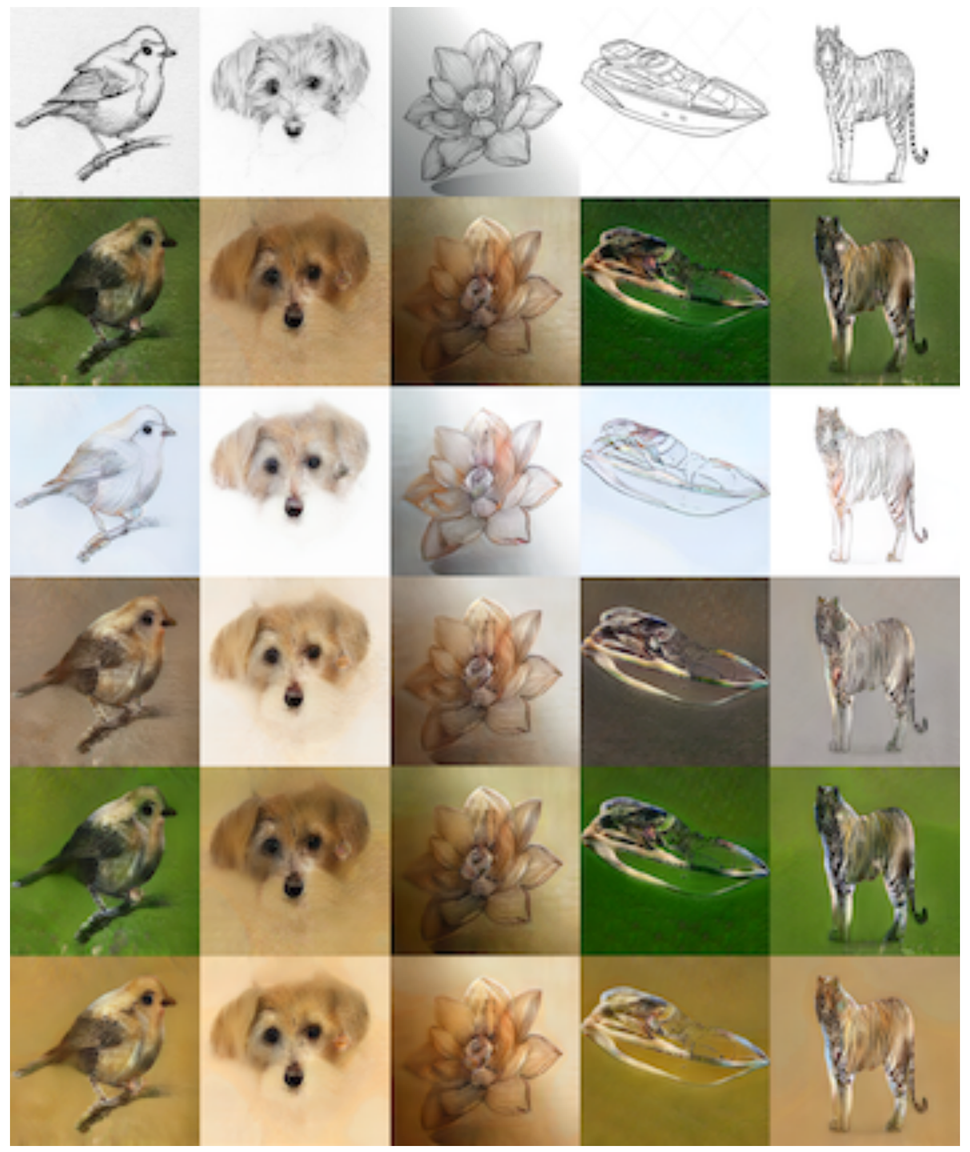
Ablating losses



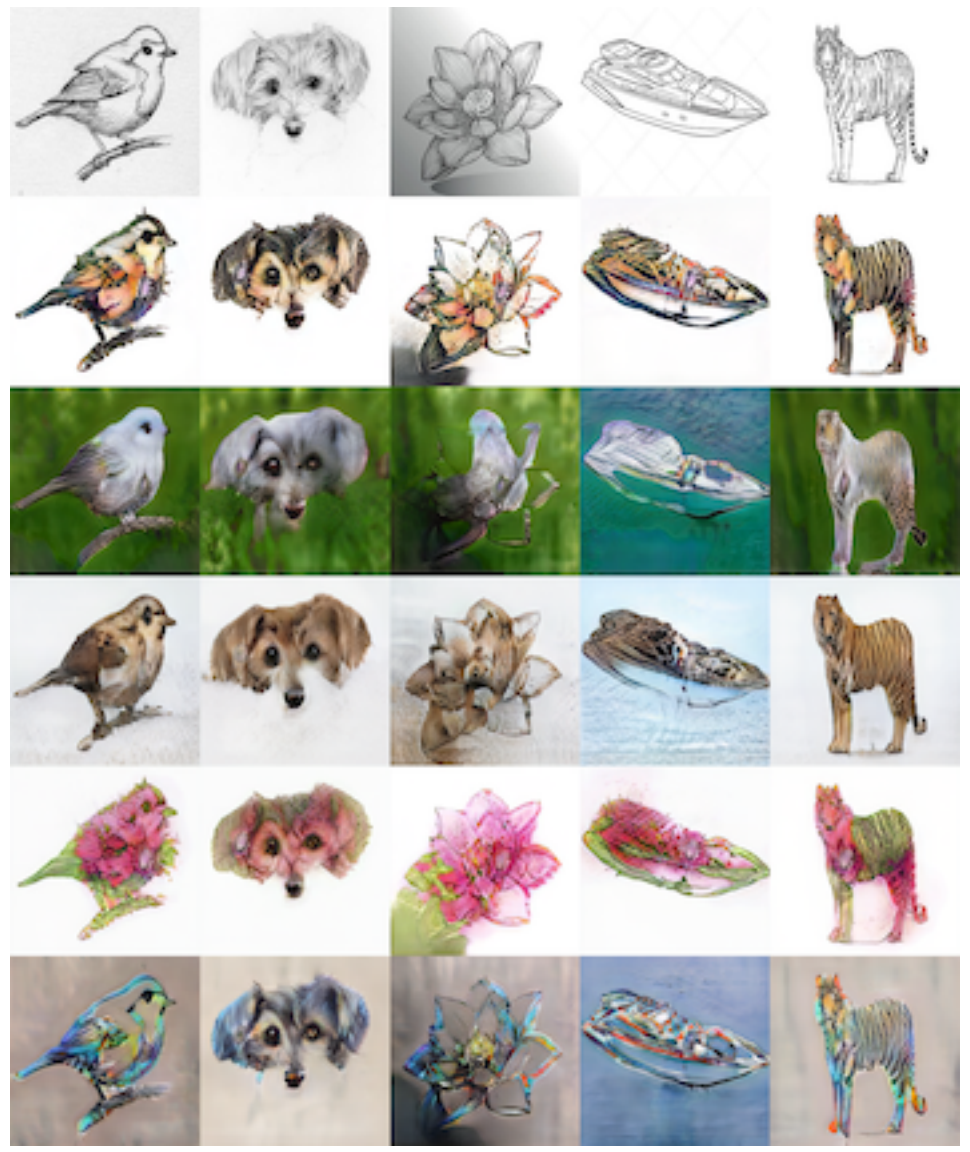
Results Sketches→Reals

DRIT

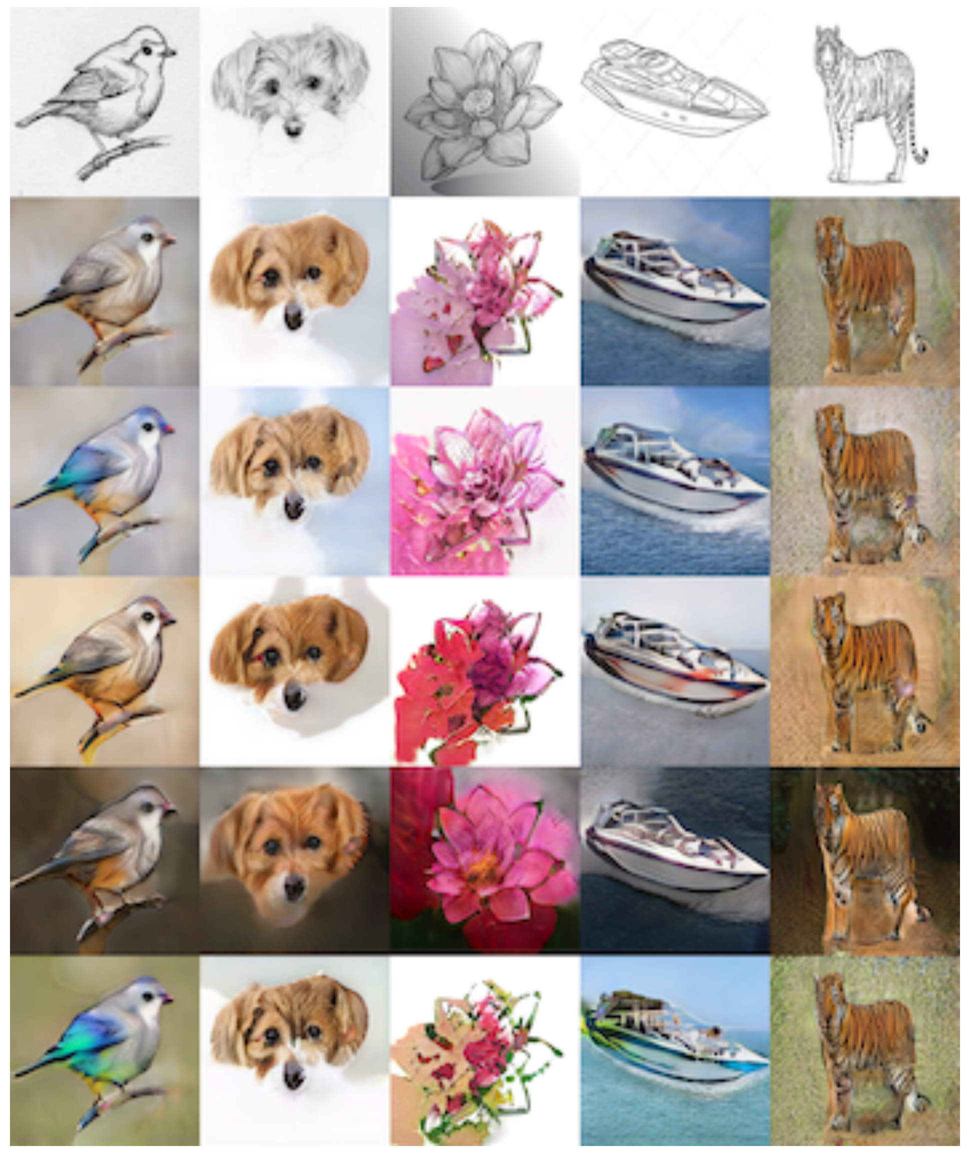
EGST-IT



StarGAN-V2



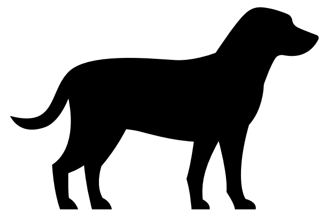
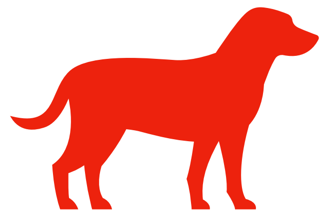
CatS-UDT (ours)



Emergence of structure in artificial language

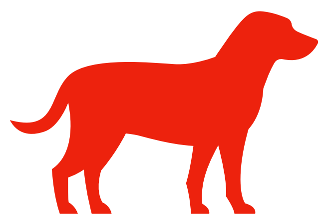
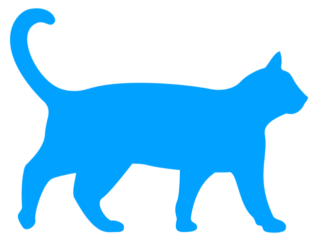
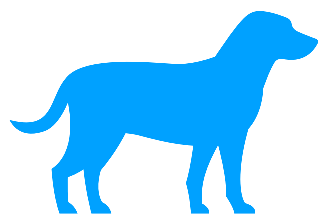
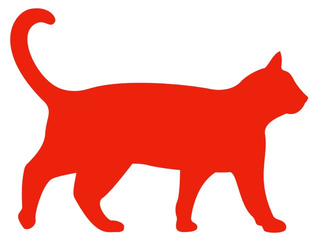
Compositionality

The meaning of a complex expression is determined by the meanings of its constituent expressions and the rules used to combine them.

Red +  = 

Systematicity

The capacity to understand a complex expression implies the capacity to understand structurally related expressions.

Understanding  and  → Understanding  and 

Language as the solution to a coordination problem

David Lewis. Convention. 1968.

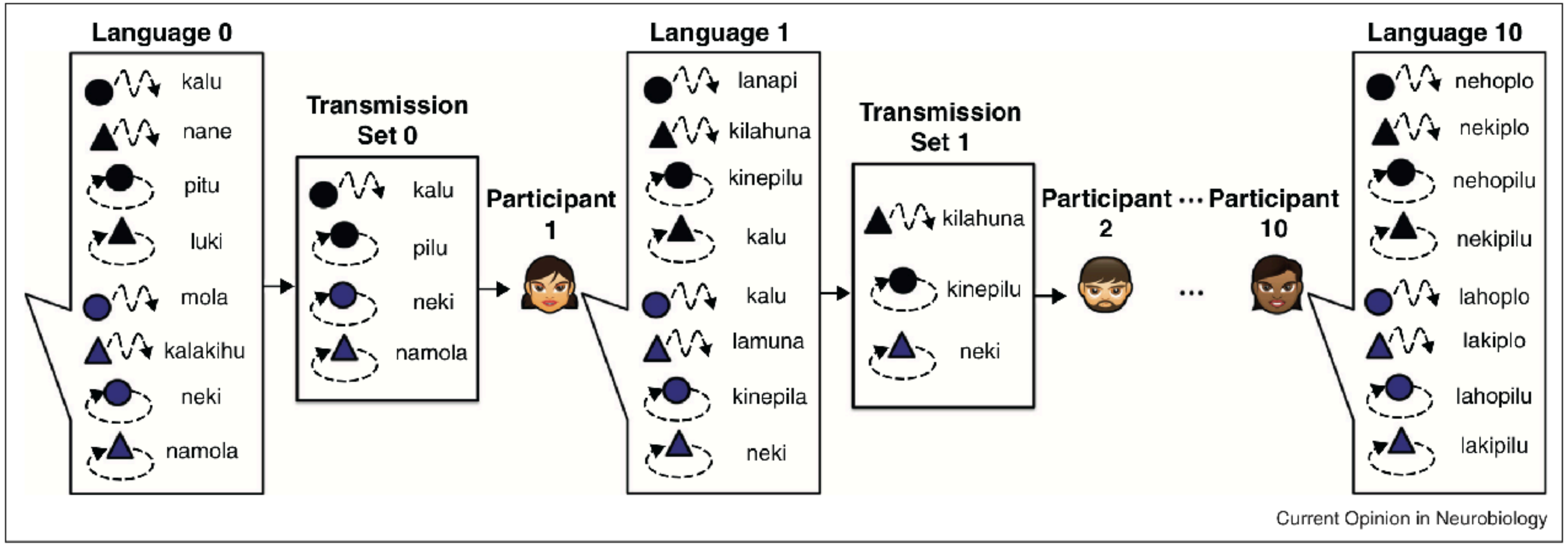
Language as the solution to a coordination problem

Vervet monkeys have a idiosyncratic call depending on the predator

Predator	Leopard	Eagle	Snake
Alarm	Loud barking	Short double syllable cough	Shutter

David Lewis. Convention. 1968.

Cultural transmission

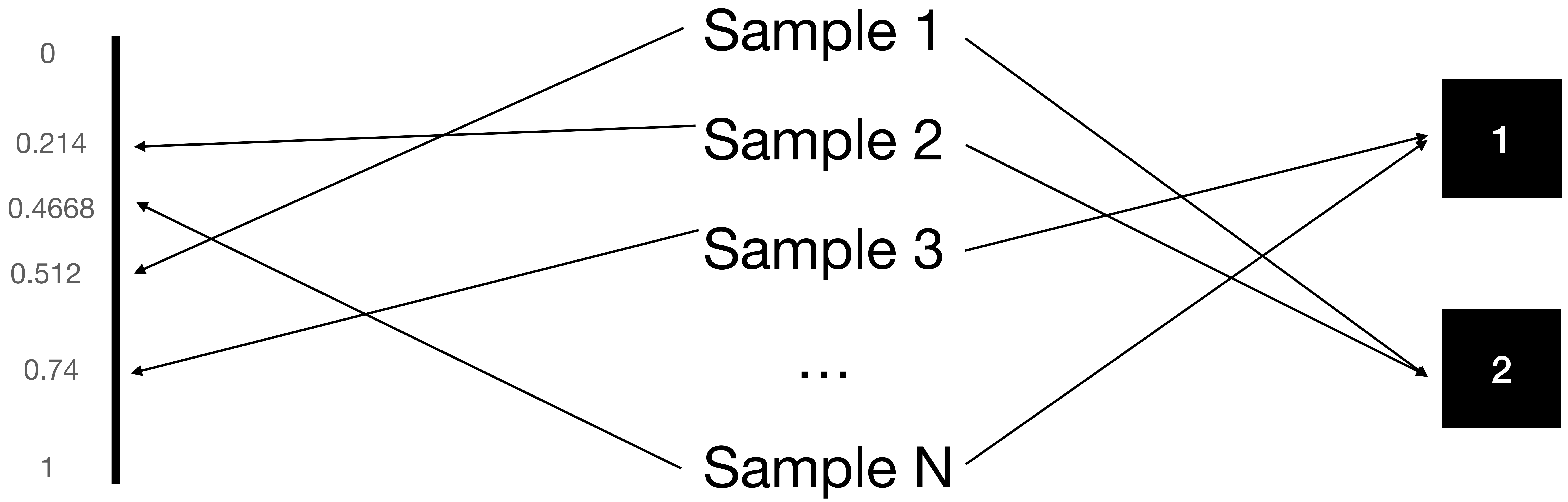


Simon Kirby, Tom Griffiths and Kenny Smith. Iterated Learning and the evolution of language. 2014

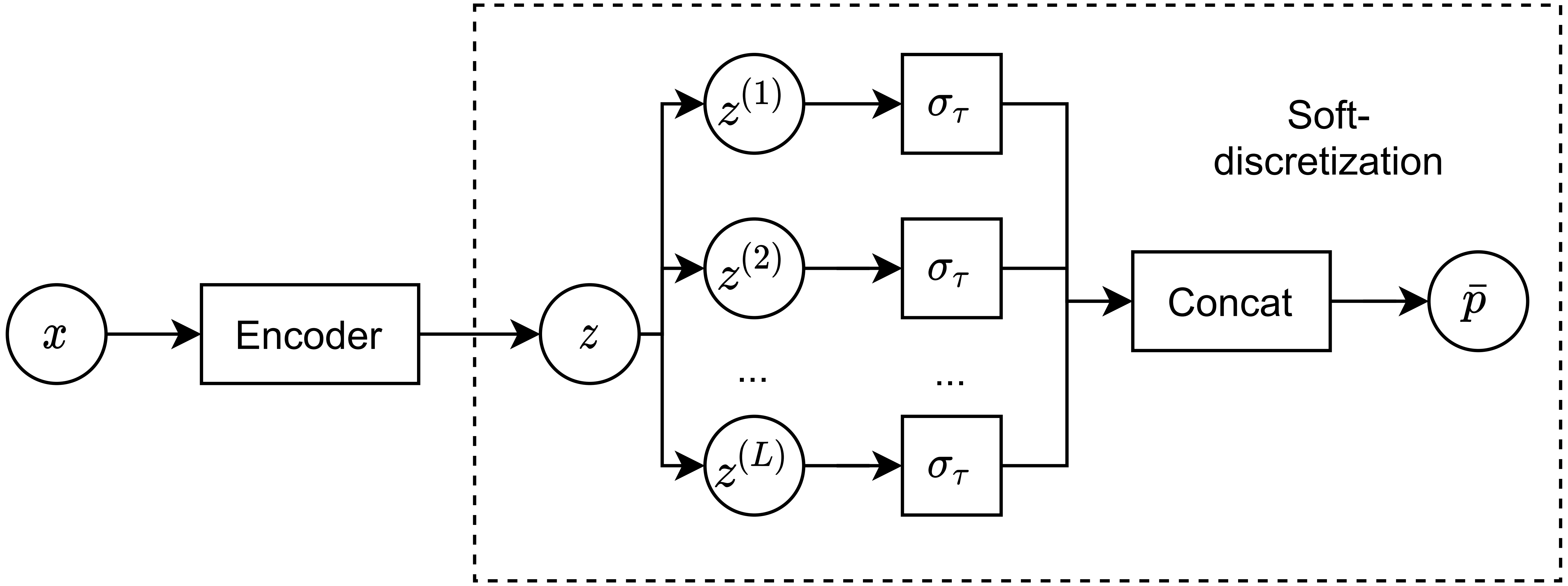
Soft-discretization bottleneck for self-supervised learning

In collaboration with Christos Tsirigotis, Max Schwarzer, Ankit Vani and Aaron Courville.

Continuous vs Discrete representation



Soft-discretization bottleneck

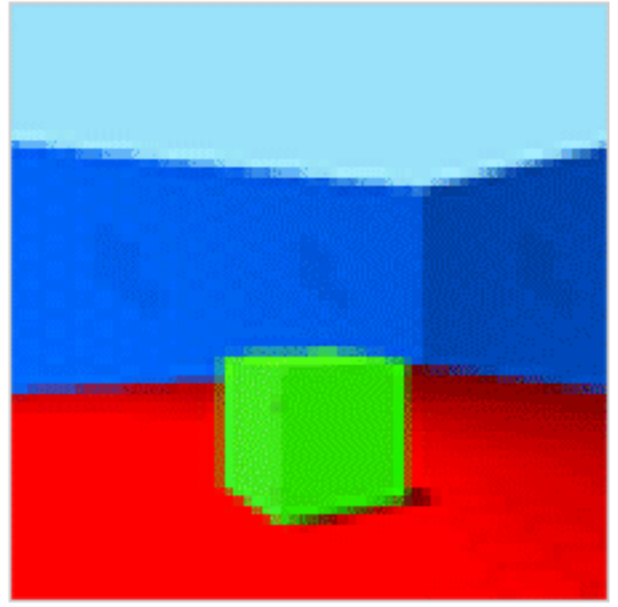


$$z^{(i)} \in \mathbb{R}^V.$$

$$\sigma_\tau(z^{(j)})_i := \frac{e^{z_i^{(j)}/\tau}}{\sum_{k=0}^{V} e^{z_k^{(j)}/\tau}}.$$

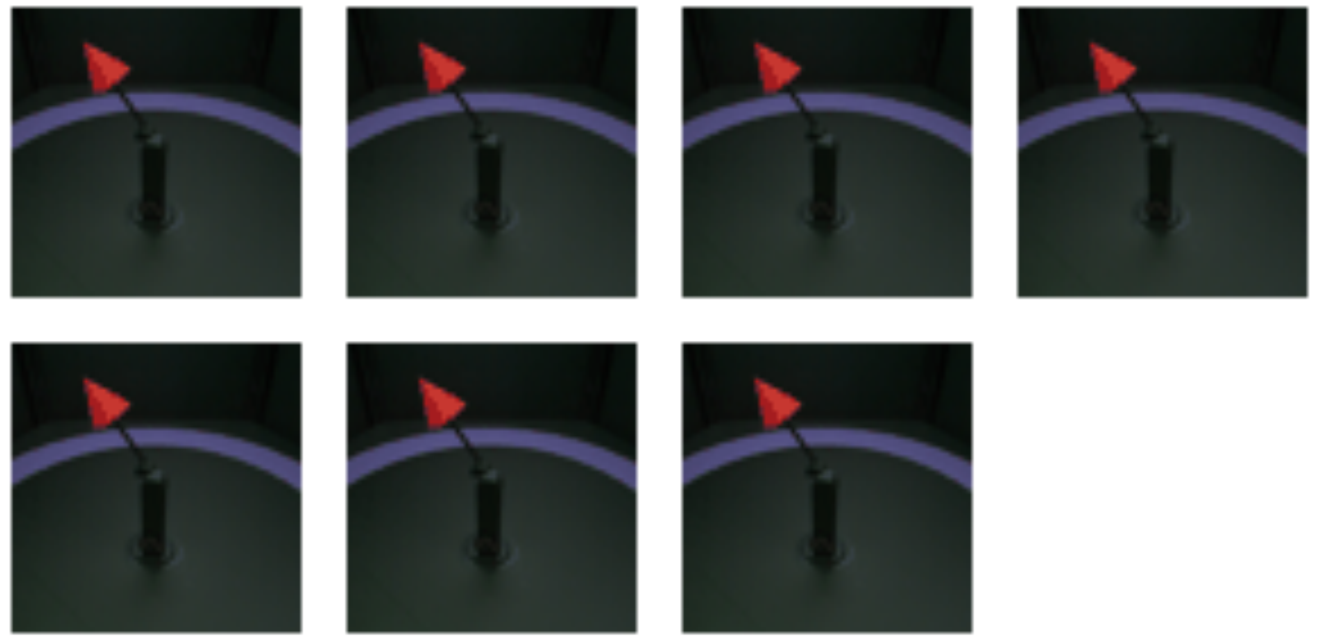
Systematic generalization

Shapes3d



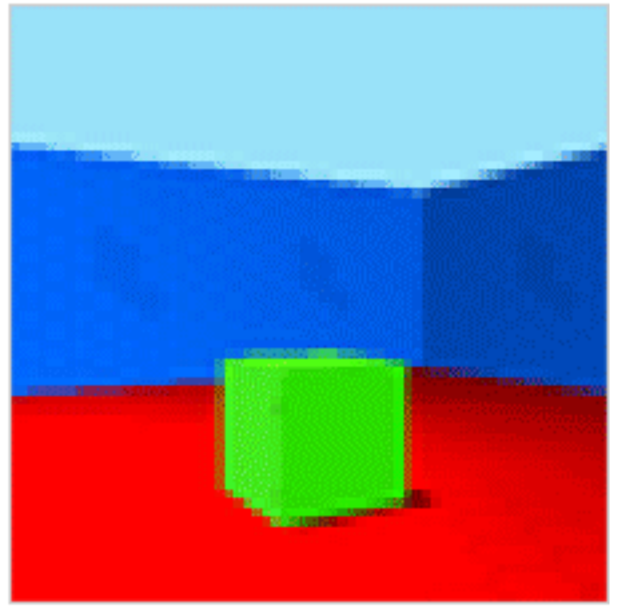
dSprites

MPI3D



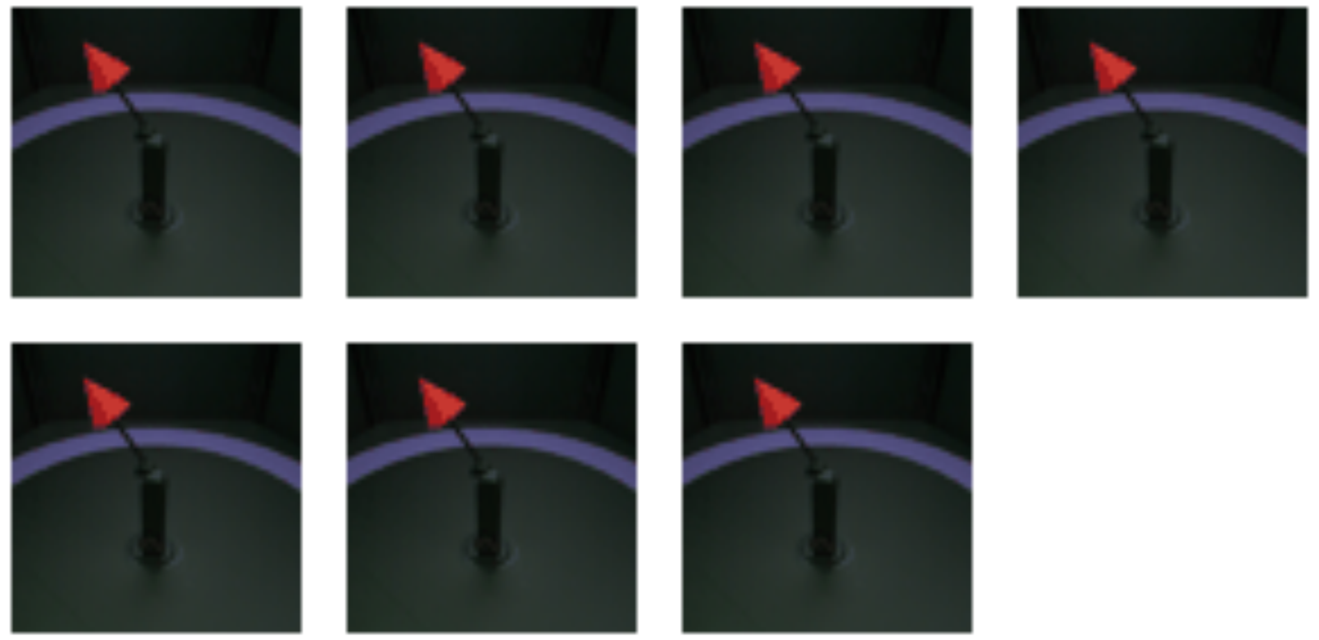
Systematic generalization

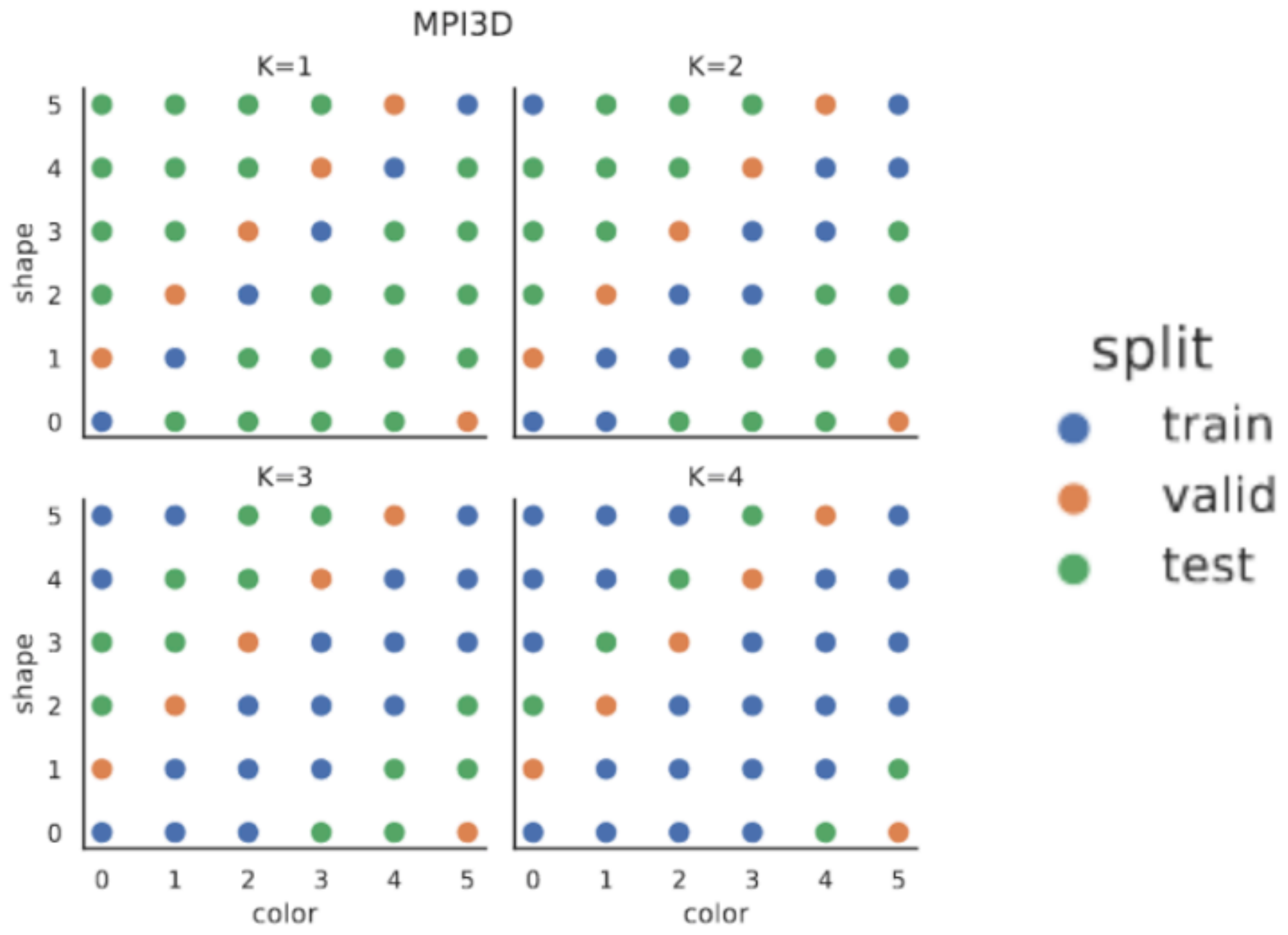
Shapes3d



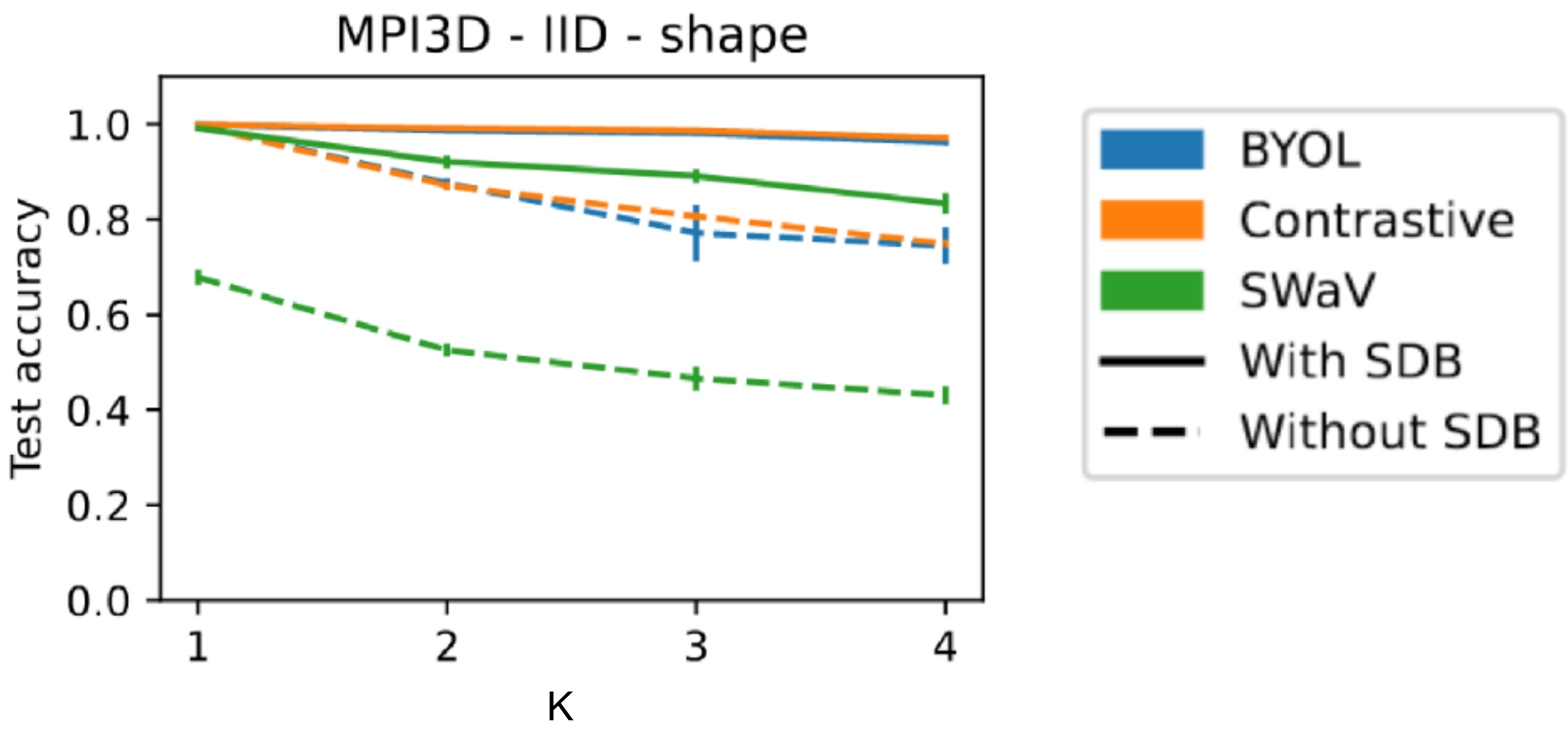
dSprites

MPI3D

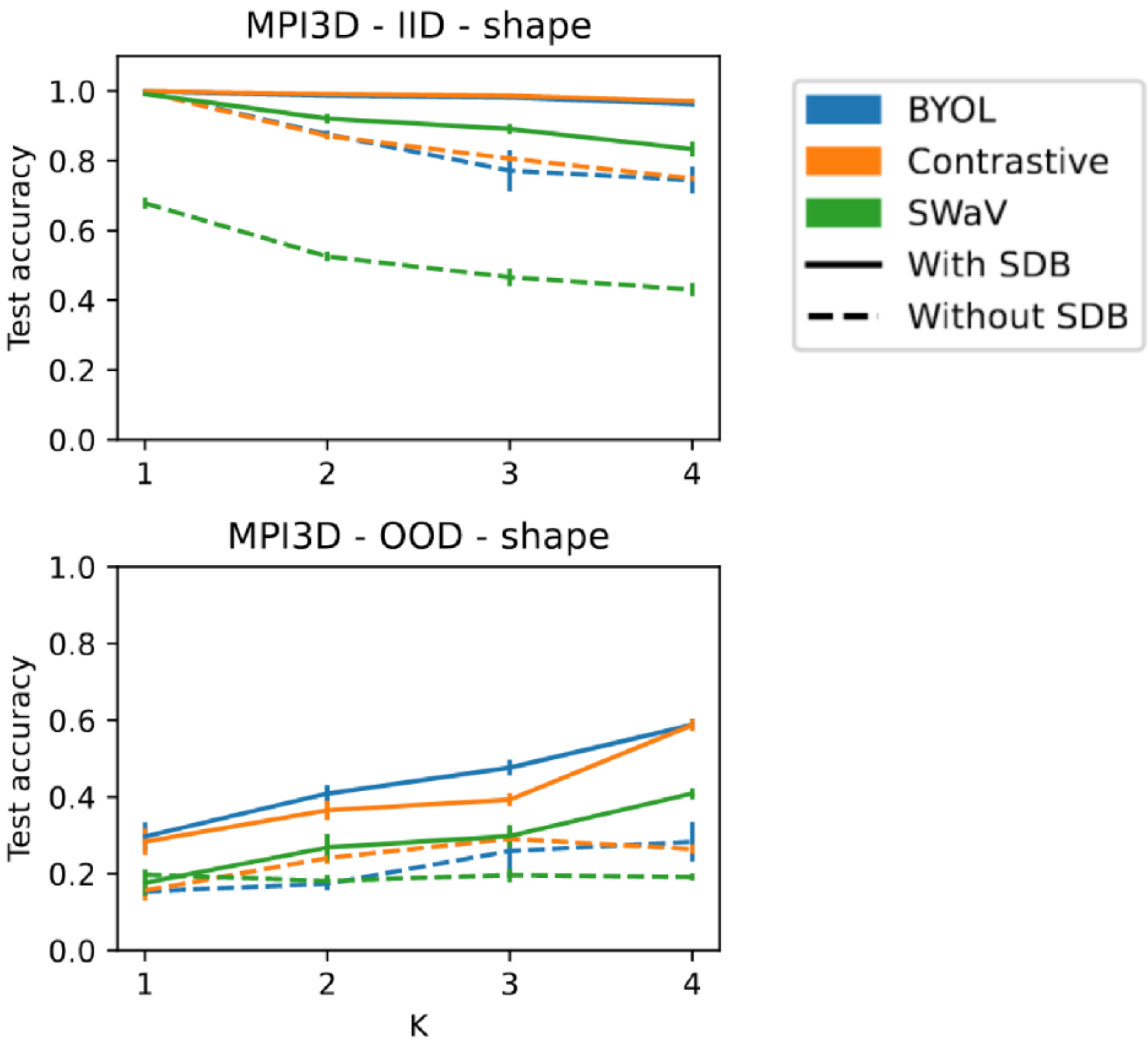




Results — Systematic generalization



Results — Systematic generalization



Results — Systematic generalization

	+ noise	Argmax	Softmax	VQ	Test accuracy
Baseline					0.29 ± 0.01
	✓				0.28 ± 0.01
Soft-Discrete			✓		0.52 ± 0.03
	✓		✓		0.49 ± 0.04
Hard-Discrete	✓	✓			0.32 ± 0.03
				✓	0.39

Table 1. Ablation of SSL-SB on MPI3D-K:3. We test the effect of adding noise, a hard discretization bottleneck via Gumbel-Softmax straight-through estimation and Vector Quantization and the soft discretization bottleneck.

Results — Systematic generalization

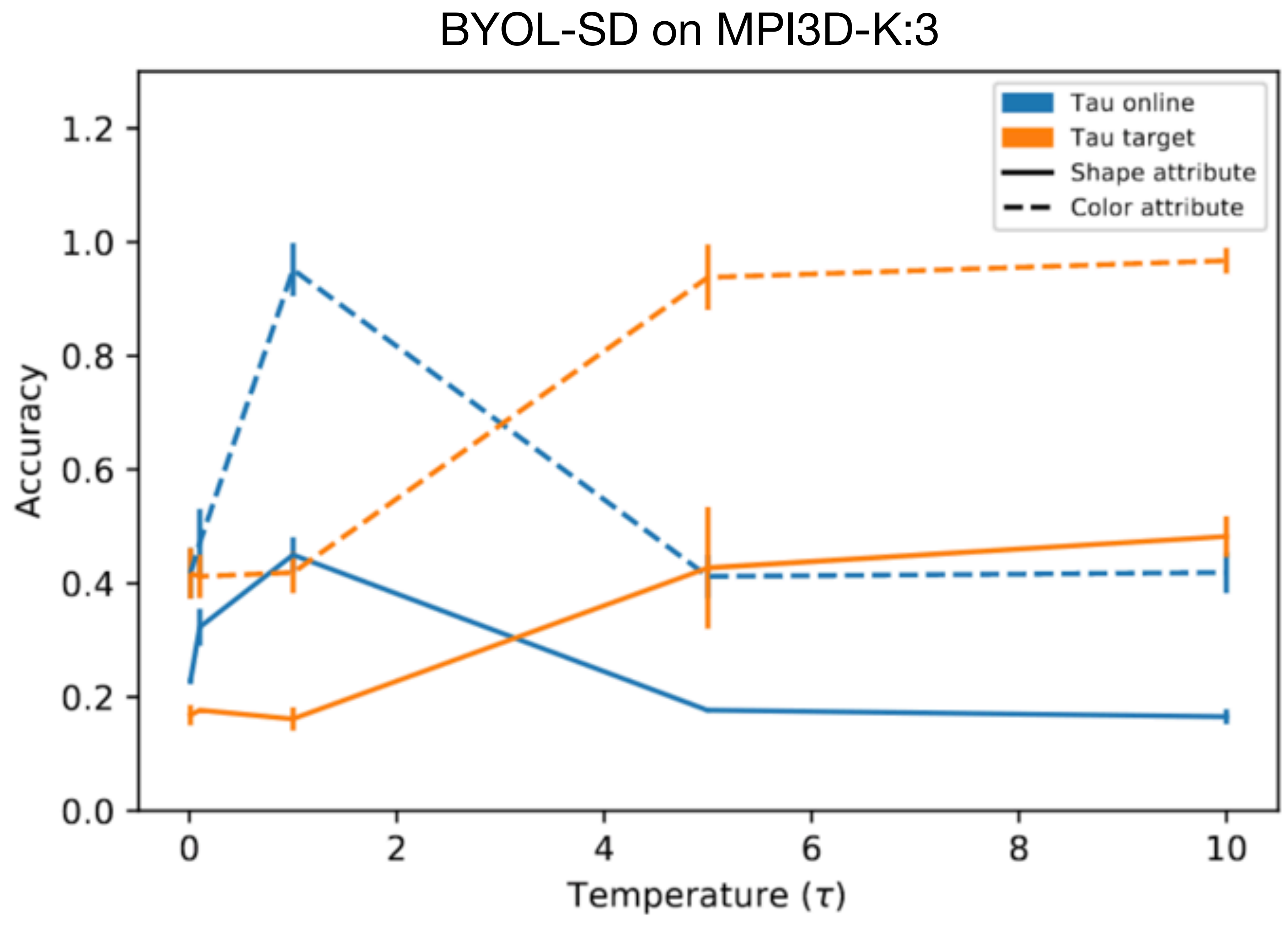


Figure 6. Study of the effect of the temperature parameter on the online (τ_O) and the target (τ_T) networks. We fix the temperature $\tau_O = 1.5$ when interpolating τ_T and $\tau_T = 4.0$ when interpolating τ_O.

Robustness to distribution shift

Train dataset

Test datasets

ImageNet

ImageNet-C

ImageNet-A

ImageNet-R

ImageNet-V2

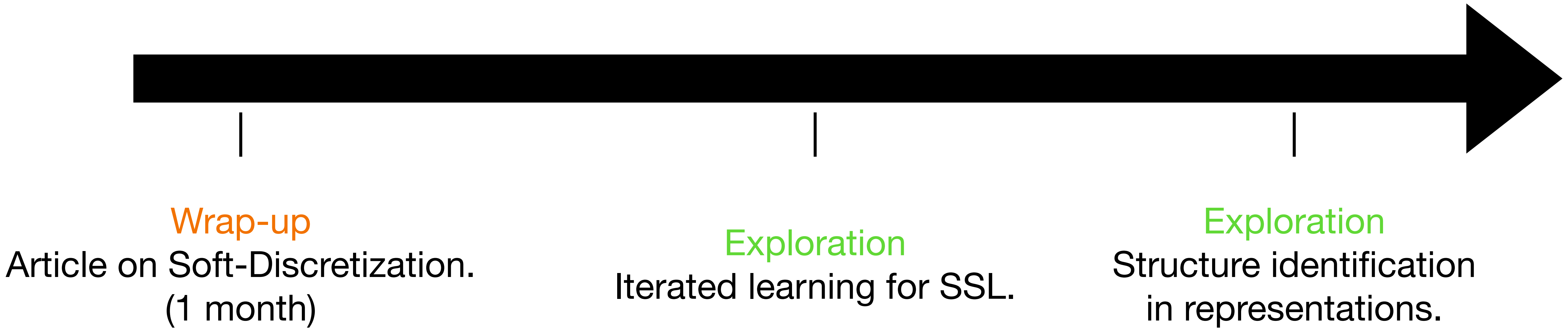
Dan Hendricks, Thomas Dietterich. Benchmarking Neural Networks Robustness to Common Corruptions and Perturbations. 2019.
Dan Hendricks et al. Natural Adversarial Examples. 2021.
Dan Hendricks et al. The Many Faces of Robustness. 2019.
Benjamin Recth et al. Do ImageNet Classifiers Generalize to ImageNet? 2019.

Results — Robustness

	Imagenet	Imagenet-v2	Imagenet-r	Imagenet-a	Imagenet-c
BYOL	67.16	53.96	15.35	0.87	33.32
BYOL + SDB	70.22	57.73	17.95	1.01	37.98

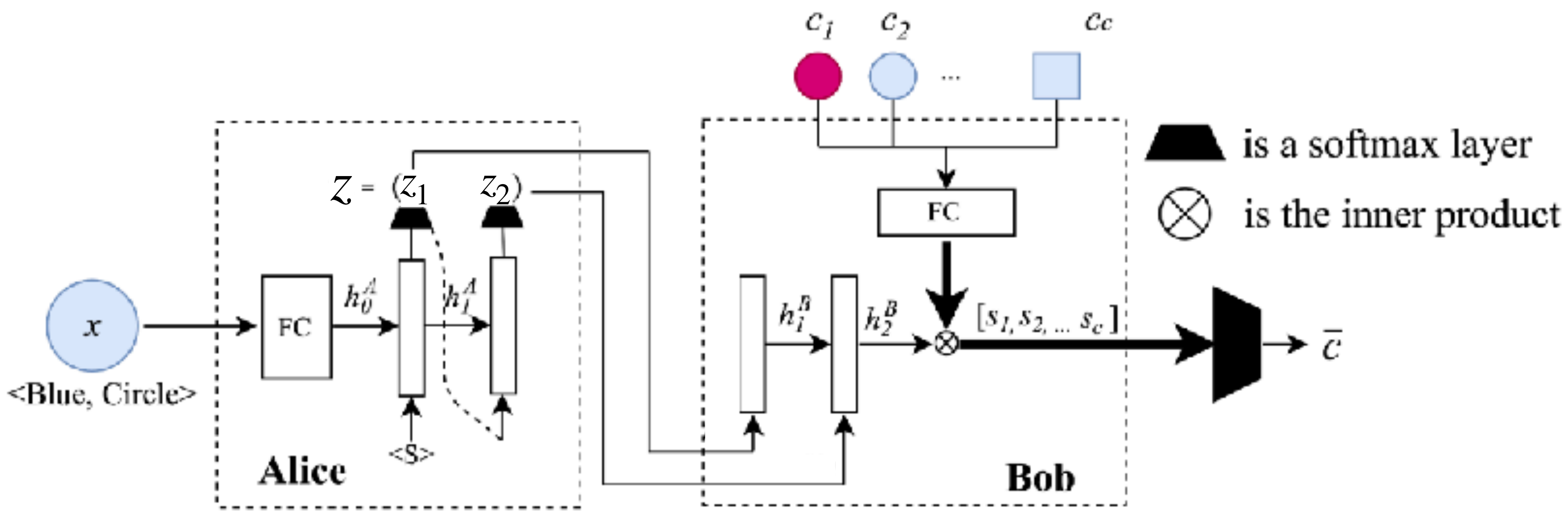
Future works

Future works



Iterated learning for communication games

Interaction: Object selection game



$$\nabla_\theta J := E\left[R(\bar{c}, \boldsymbol{x})\nabla_\theta \log p_f(\boldsymbol{z}|\boldsymbol{x}) + \lambda_f \nabla_\theta H[p_f(\boldsymbol{z}|\boldsymbol{x})]\right]$$

$$\nabla_\psi J := E\left[R(\bar{c}, \boldsymbol{x})\nabla_\psi \log p_g(\bar{c}|\boldsymbol{z}, c_1, \ldots, c_k) + \lambda_g \nabla_\psi H(p_g(\bar{c}|\boldsymbol{x}, c_1, \ldots, c_n))\right]$$

Generation

$$\mathcal{Z} := \{(\boldsymbol{x}_i, \boldsymbol{z}_i)\}_{i=1}^N$$

Distillation

$$\min_{\theta^{t+1}} E_{(\boldsymbol{x},\boldsymbol{z})\sim\mathcal{Z}}\, l(f_{\theta^{t+1}}(\boldsymbol{x}), \boldsymbol{z})$$

* l is defined as the cross-entropy

Algorithm 1 Neural Iterated Learning

Require: $\mathcal{X}$, f_{θ^0}, g_{ϕ^0}, N_{iter}, $M_{\text{interaction}}$.
- θ^0 randomly initialized
- ϕ^0 randomly initialized
- $t \leftarrow 0$
- **while** $N_{\text{iter}} \neq 0$ **do**
  - $S \leftarrow 0$
  - **while** $S \neq M$ **do**
    - $\theta^t \leftarrow \theta^t + \alpha \nabla_{\theta^t} J$
    - $\psi^t \leftarrow \psi^t + \alpha \nabla_{\psi^t} J$

    } Interaction
    - $S \leftarrow S + 1$
  - **end while**
  - $\mathcal{Z} \leftarrow$ Generation($\mathcal{X}$, f_{θ^t})
  - $\theta^{t+1} \leftarrow$ Distillation($\mathcal{Z}$, $f_{\theta^{t+1}}$)
  - ψ^{t+1} randomly initialized
  - $t \leftarrow t + 1$
  - $N_{\text{iter}} \leftarrow N_{\text{iter}} - 1$
- **end while**

Yi Ren et al. Compositional language emerges in a Neural Iterated Learning Model. 2020.

Iterated learning for self-supervised learning

Interaction: Self-supervised learning objective

Example: Noise contrastive estimation, BYOL

Generation

$$\mathcal{Z} := \{(\boldsymbol{x}_i, \boldsymbol{z}_i)\}_{i=1}^{N}$$

Distillation

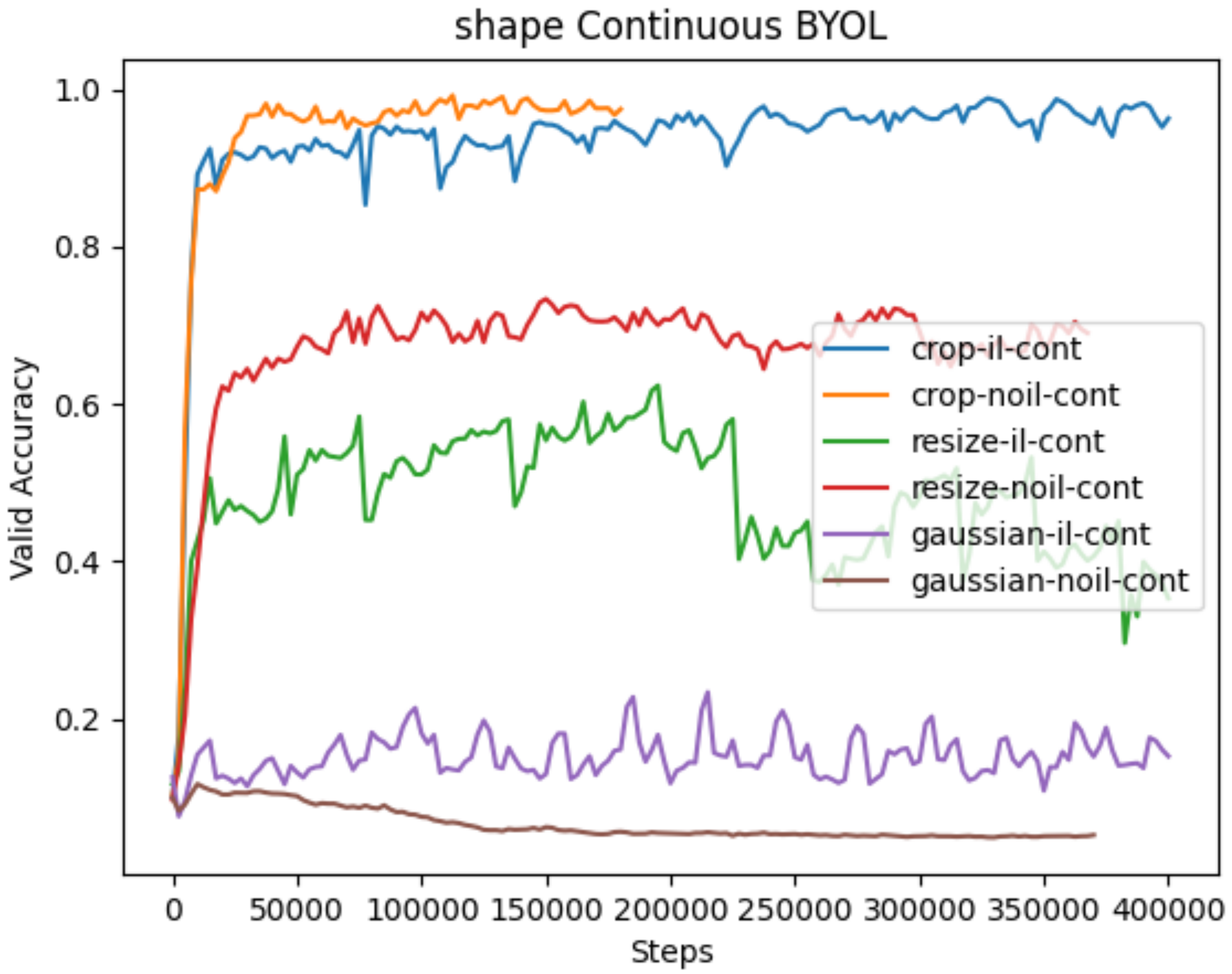
$$\min_{\theta^{t+1}} E_{(\boldsymbol{x},\boldsymbol{z})\sim\mathcal{Z}} l(f_{\theta^{t+1}}(\boldsymbol{x}), \boldsymbol{z})$$

* l is defined as ?

Algorithm 1 Neural Iterated Learning

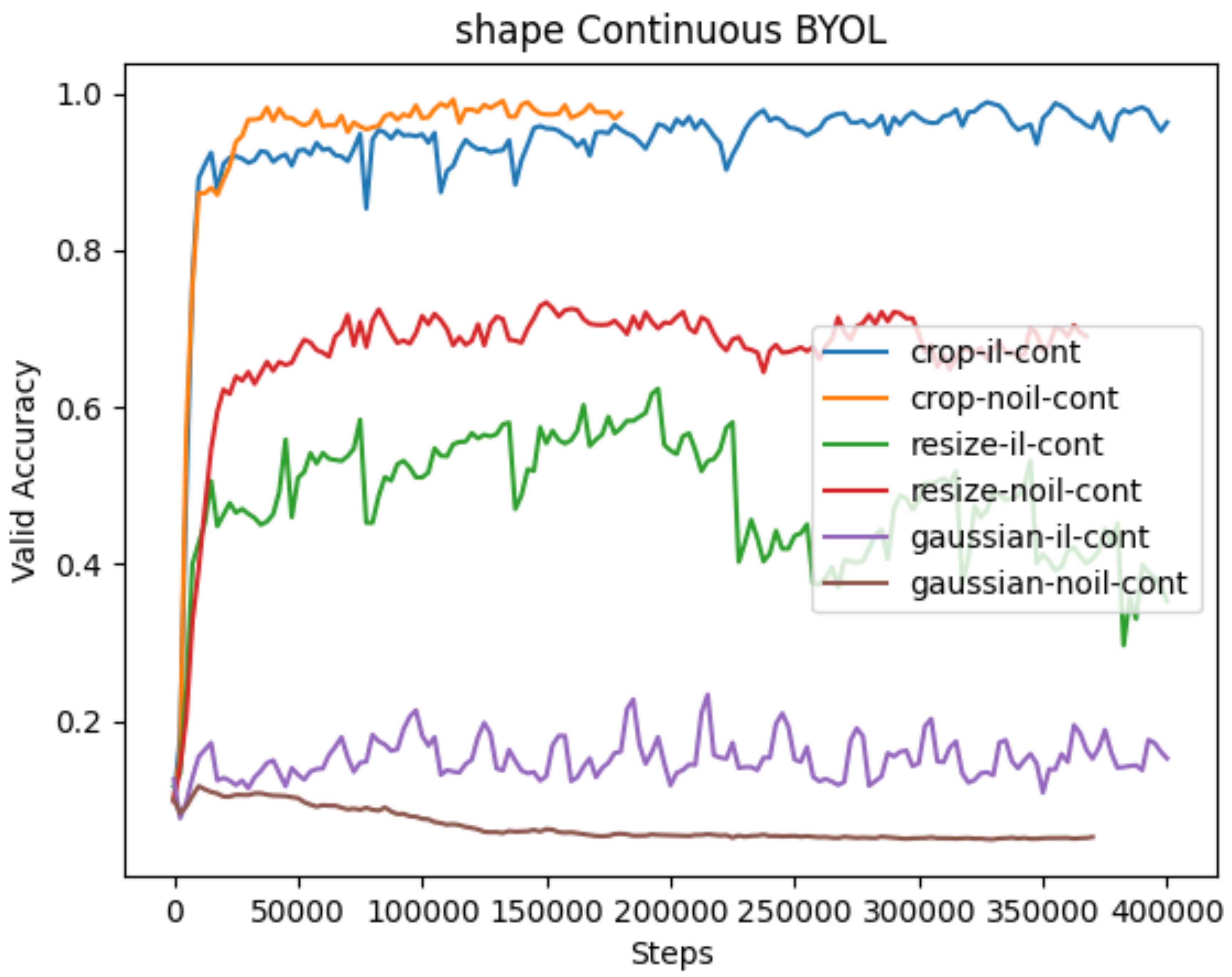
Require: $\mathcal{X}$, f_{θ^0}, g_{ϕ^0}, N, $M_{\text{interaction}}$.
 θ^0 randomly initialized
 ϕ^0 randomly initialized
 $t \leftarrow 0$
 while $N \neq 0$ **do**
  $S \leftarrow 0$
  while $S \neq M$ **do**
   $\theta^t \leftarrow \theta^t \cancel{+\alpha\nabla_{\theta^t}J} - \alpha\nabla_{\phi^t}J$
   $\psi^t \leftarrow \psi^t \cancel{+\alpha\nabla_{\psi^t}J} - \alpha\nabla_{\psi^t}J$
   $S \leftarrow S + 1$
  end while
  $\mathcal{Z} \leftarrow \text{Generation}(\mathcal{X}, f_{\theta^t})$
  $\theta^{t+1} \leftarrow \text{Distillation}(\mathcal{Z}, f_{\theta^{t+1}})$
  ψ^{t+1} randomly initialized
  $t \leftarrow t + 1$
  $N \leftarrow N - 1$
 end while

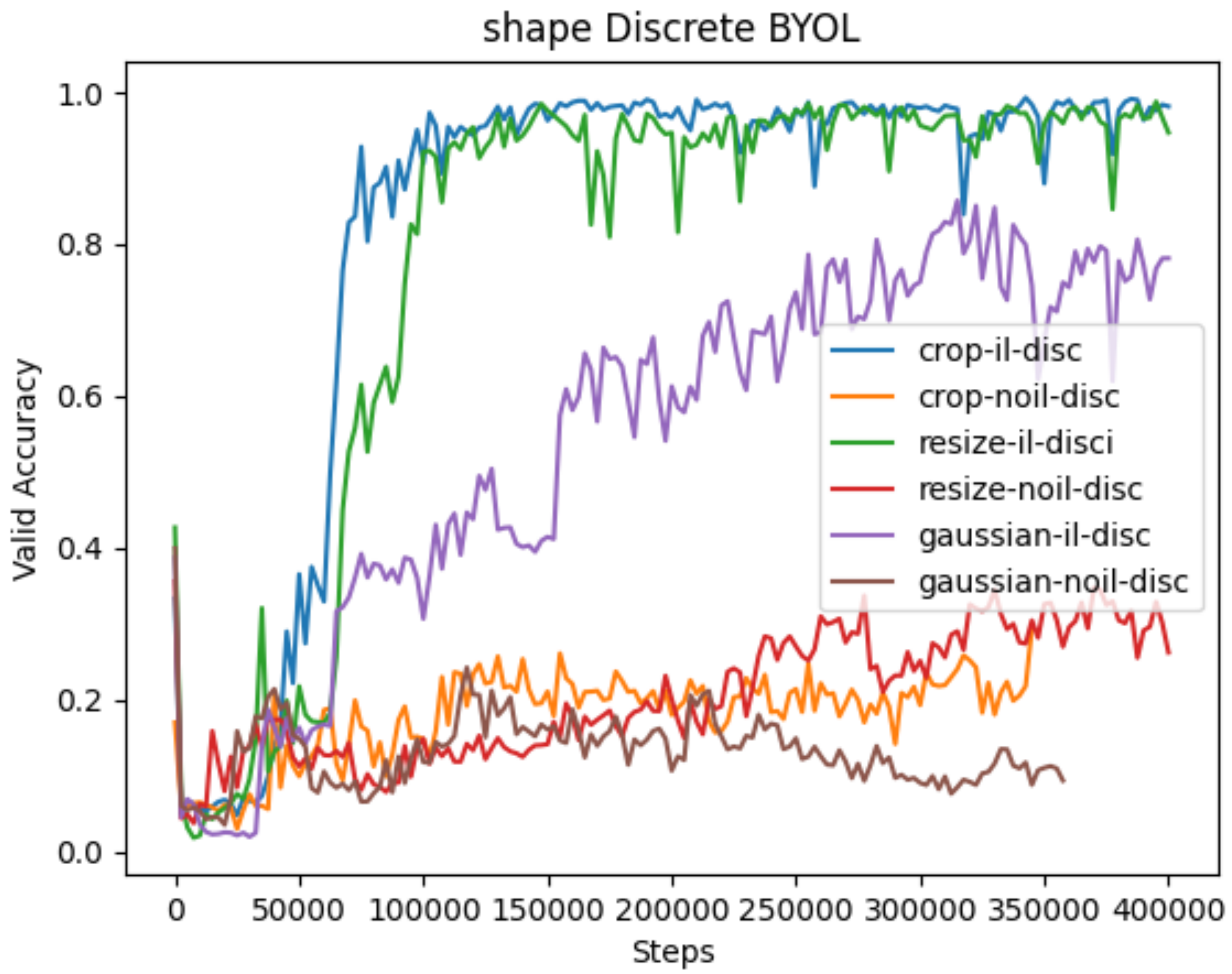
Iterated learning for self-supervised learning



Comparing continuous and discrete bottleneck on the systematic generalization task of predicting the shape of dSprites for K=2.

Iterated learning for self-supervised learning





Comparing continuous and discrete bottleneck on the systematic generalization task of predicting the shape of dSprites for K=2.

Conclusion





+ colleagues