

The self-supervised revolution: how to trade compute for labels?

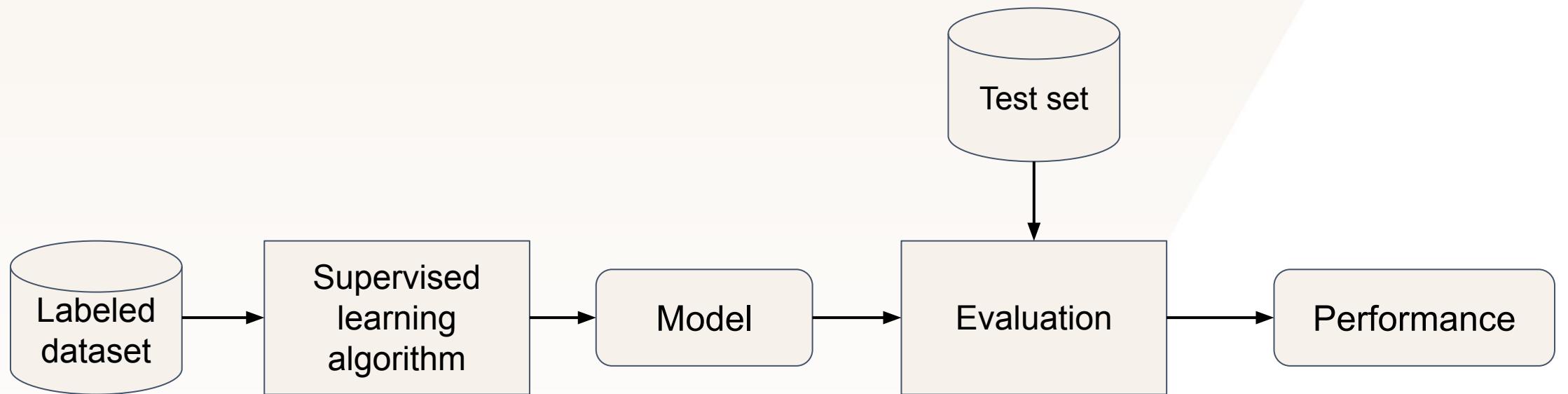
Gaétan Marceau Caron



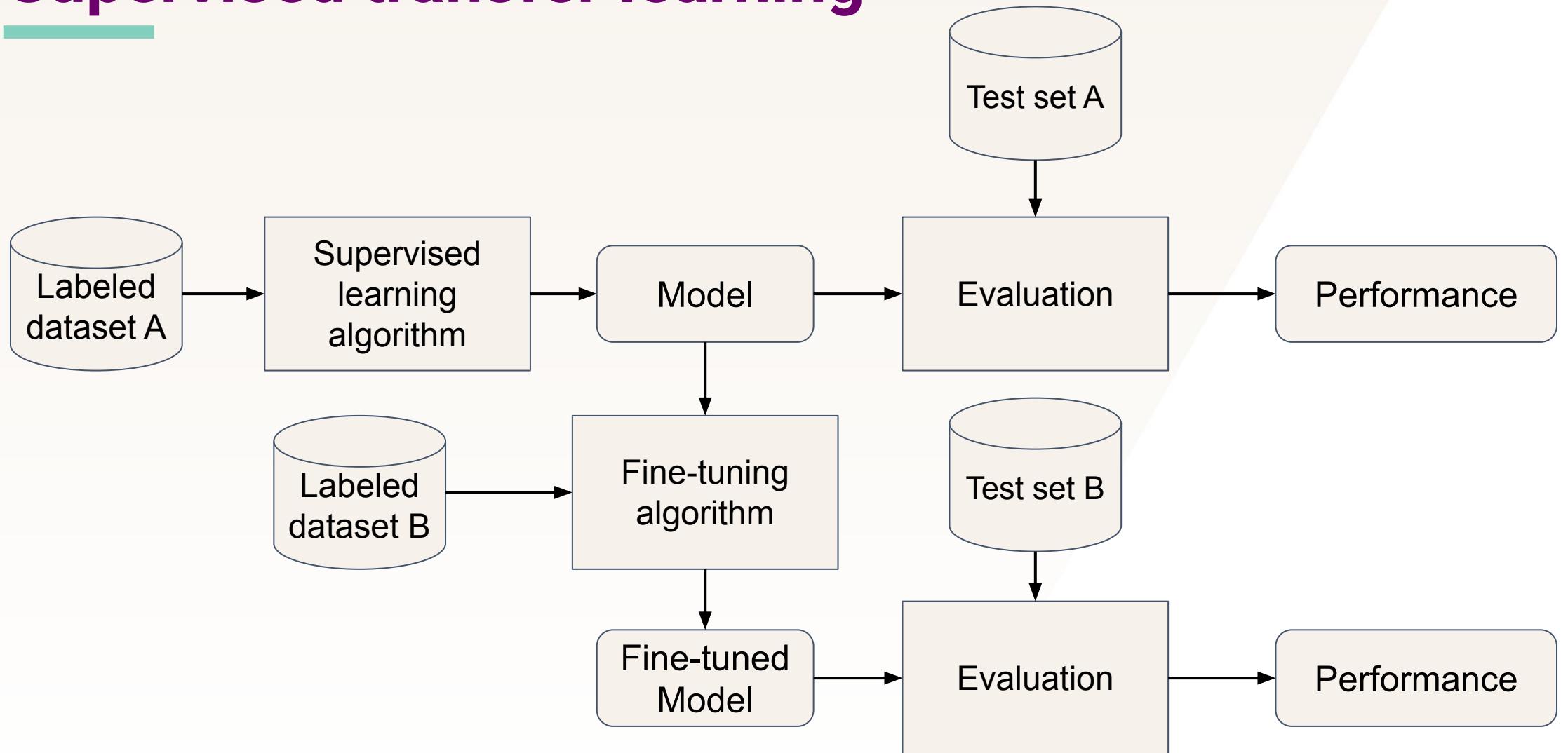
01

Representation learning

Supervised learning

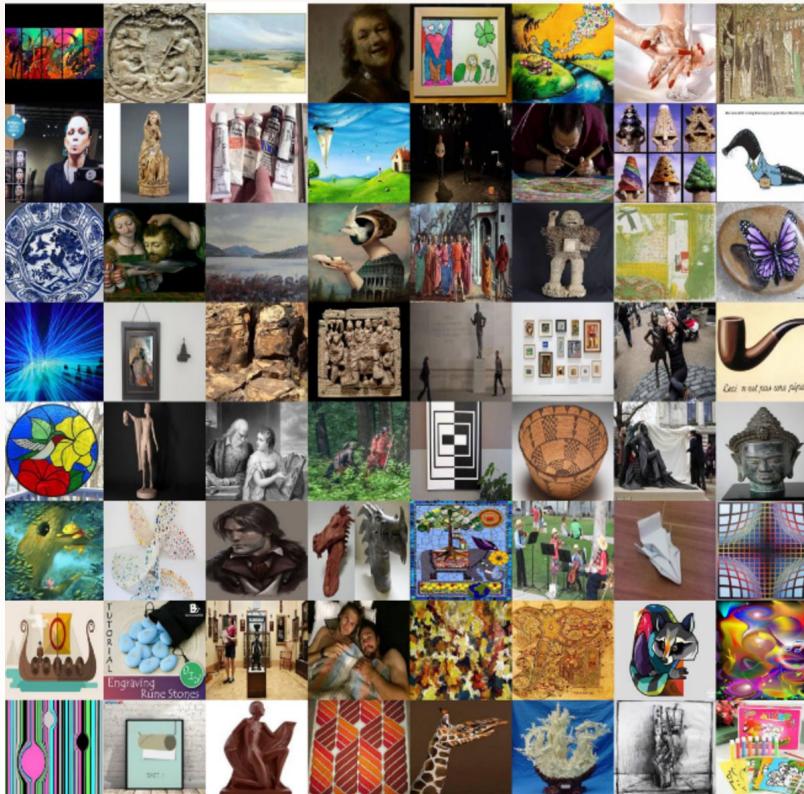


Supervised transfer learning

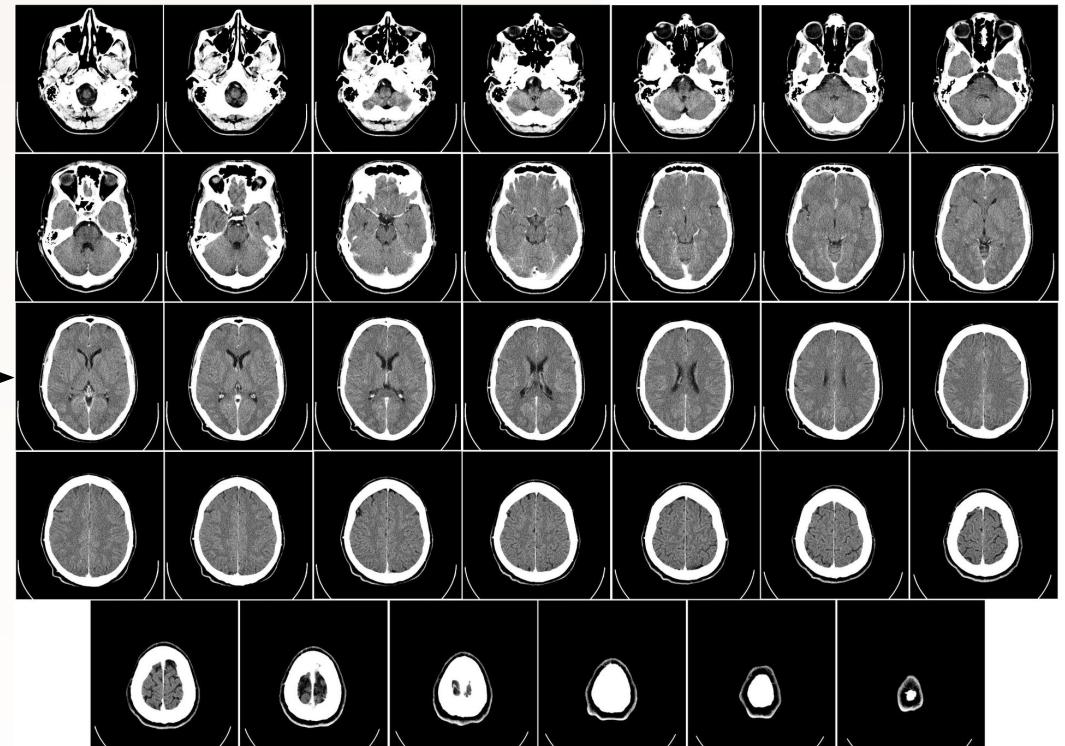


Limitations of supervised pre-training

Pre-training on ImageNet

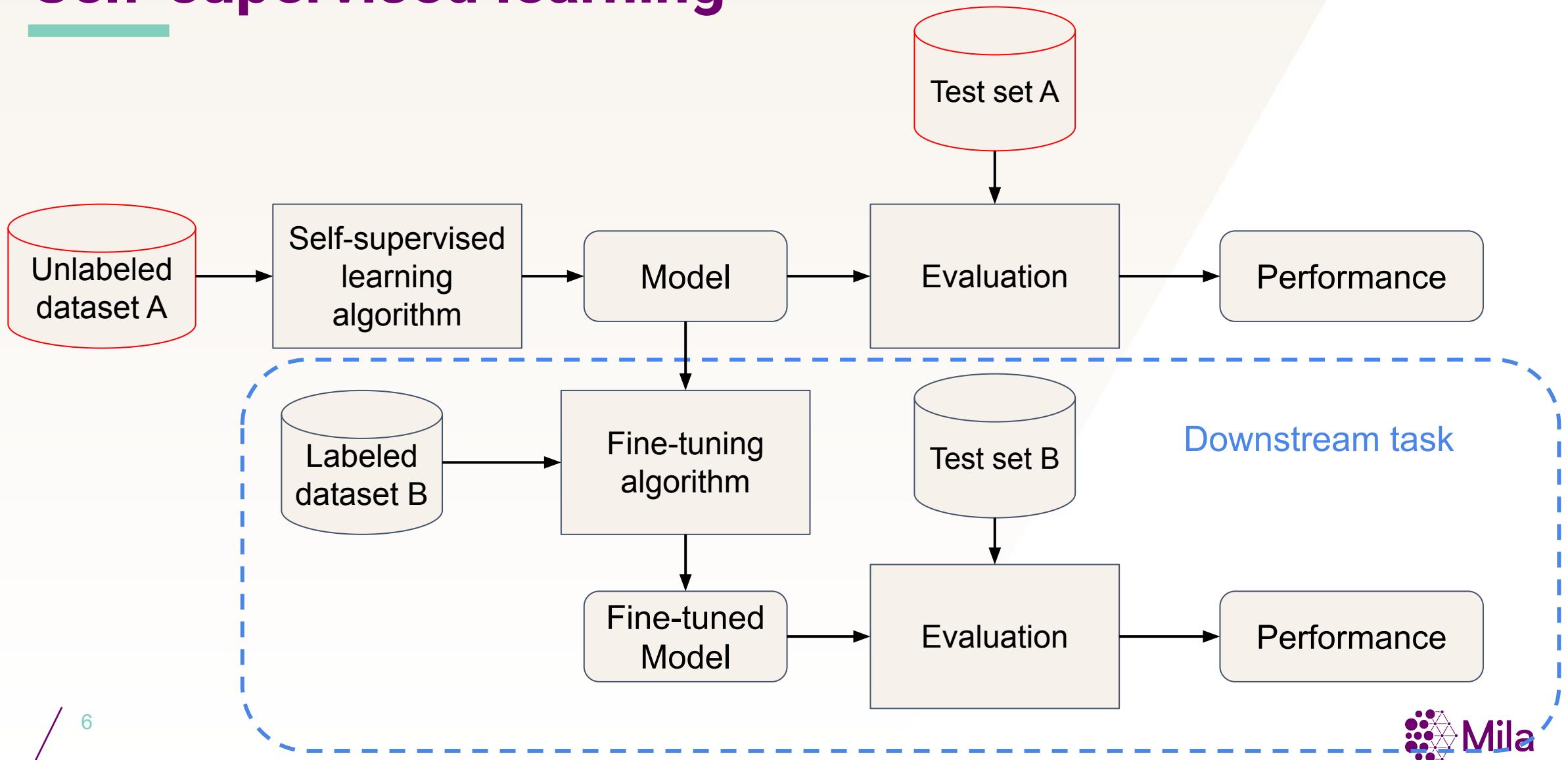


Downstream task on CT scan



Transfer?

Self-supervised learning



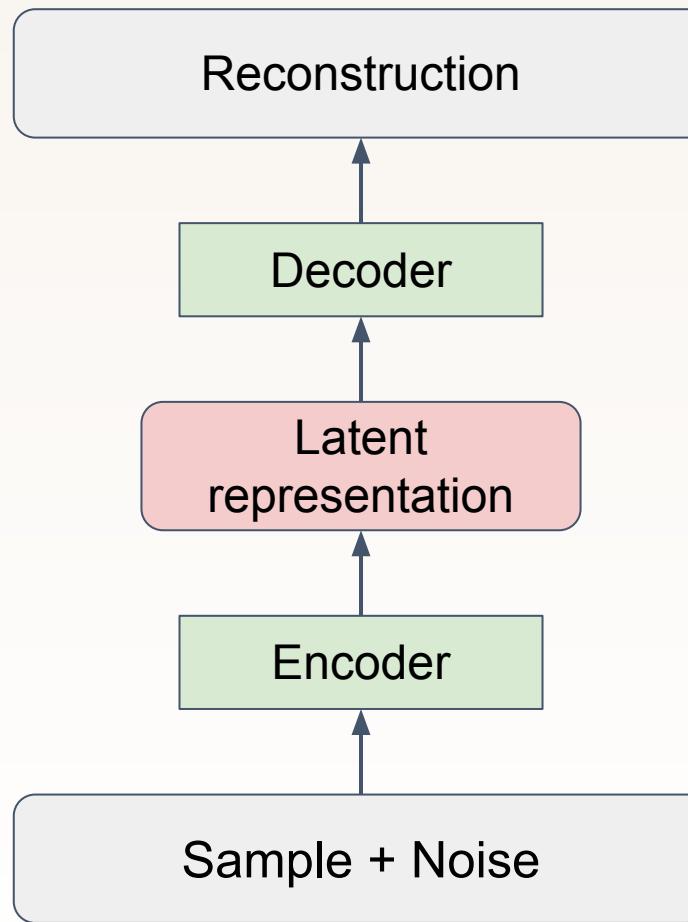
Objective

Find different techniques that leverage **the massive amount of unlabeled data** to learn representations that can **transfer well to downstream tasks**.

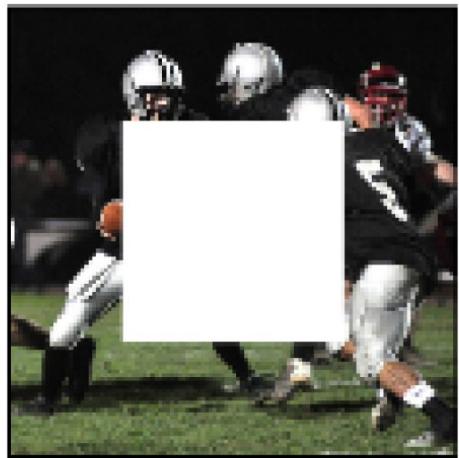
02

Pretext tasks with reconstruction

Denoising auto-encoder



Context encoders



Encoder

Encoder Features

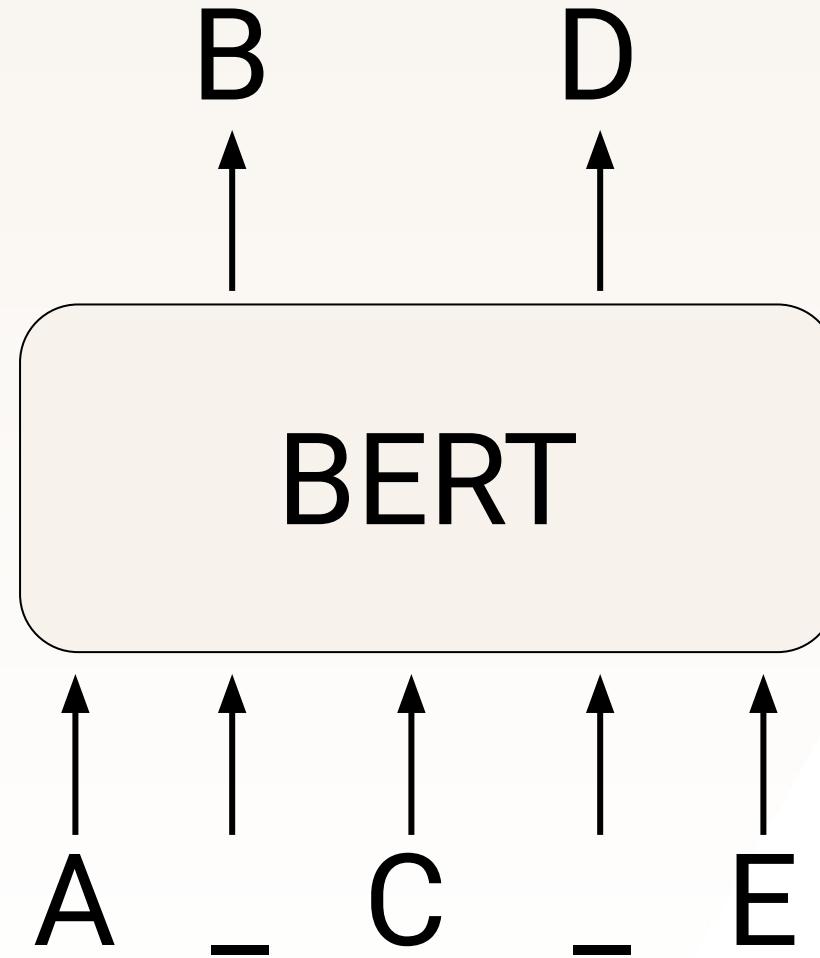
Channel-wise
Fully
Connected

Decoder Features

Decoder



Reconstruction in NLP



Jacob Devlin, Ming-Wei Chang, Kenton Lee, Kristina Toutanova:
BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. NAACL-HLT (1) 2019: 4171-4186

Colorization

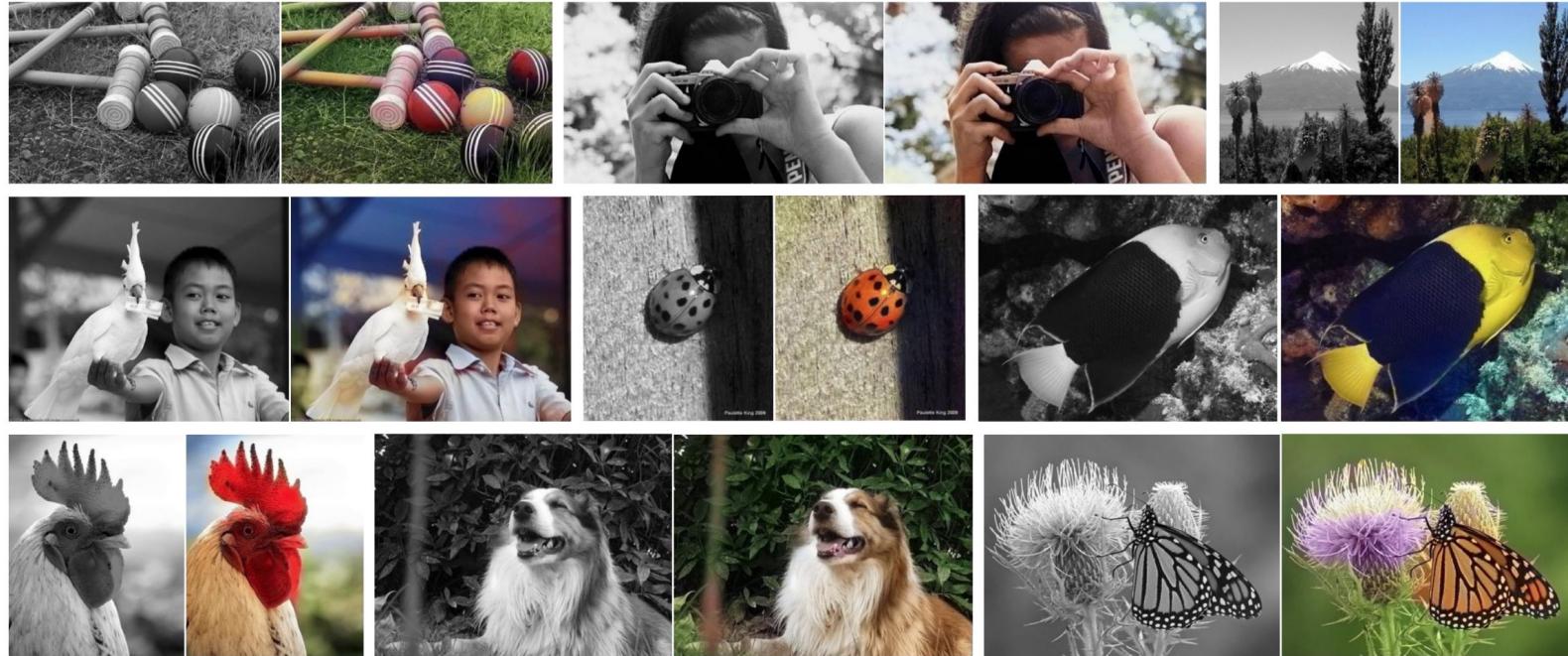
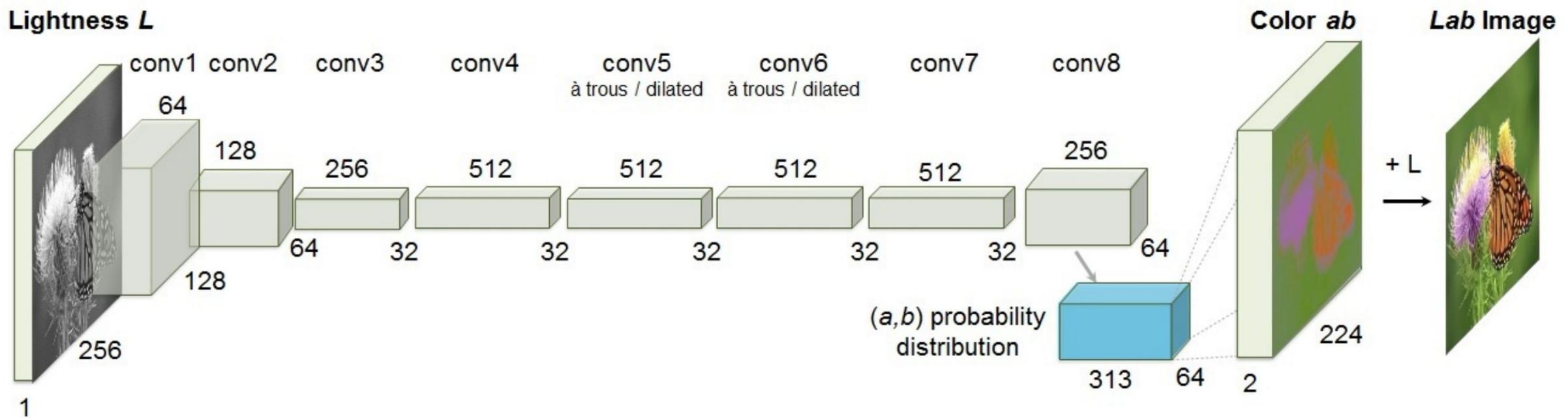


Fig. 1. Example input grayscale photos and output colorizations from our algorithm. These examples are cases where our model works especially well. Please visit <http://richzhang.github.io/colorization/> to see the full range of results and to try our model and code. Best viewed in color (obviously).

Colorization



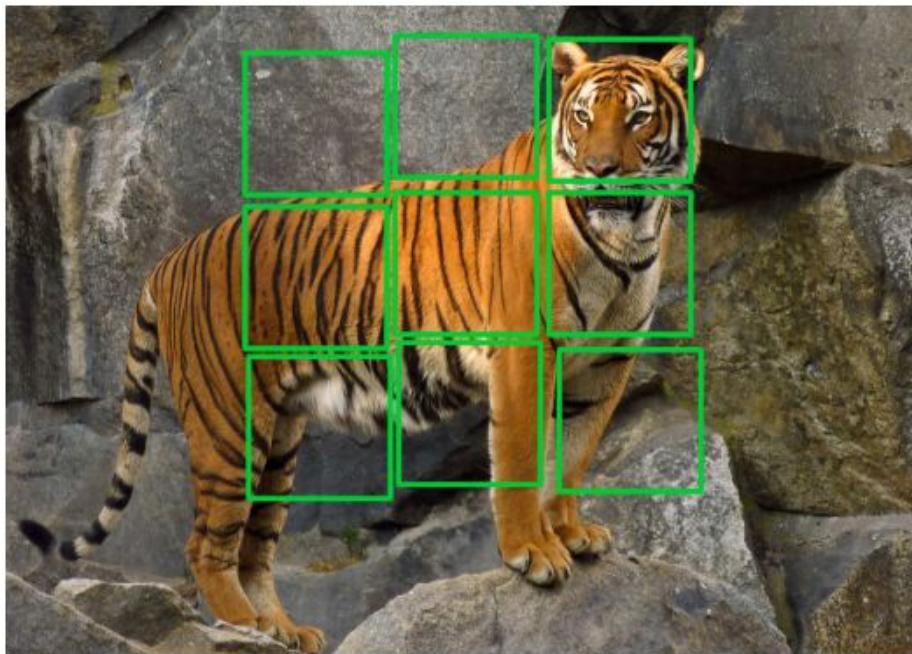
Discussion

- Make predictions in the high-dimensional input space.
- Work better for sequences of symbols than audio signals or images.

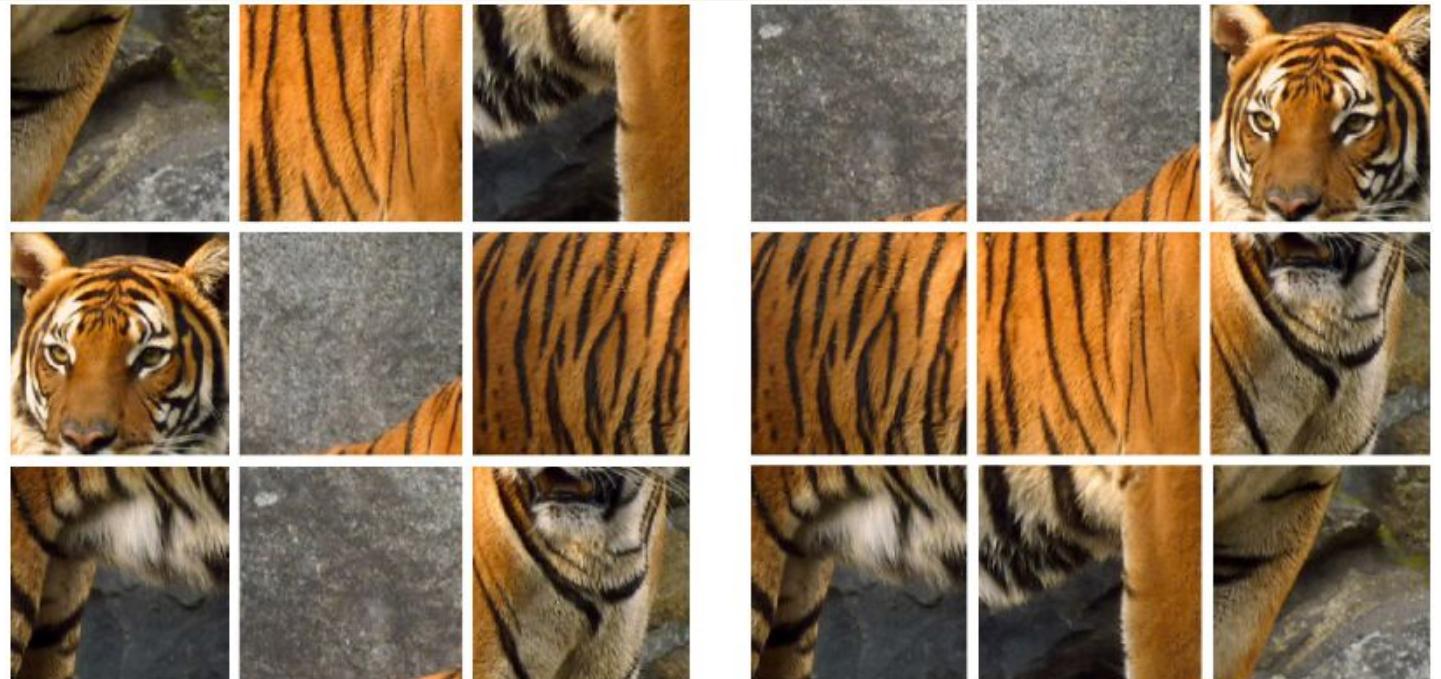
03

Pretext tasks with pseudo-labels

Jigsaw puzzle



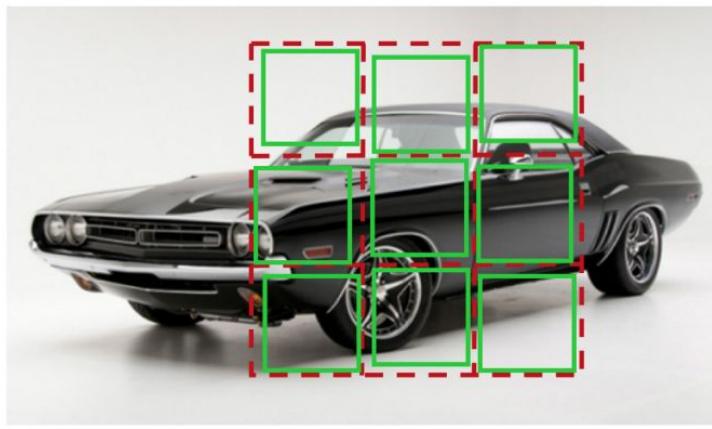
(a)



(b)

(c)

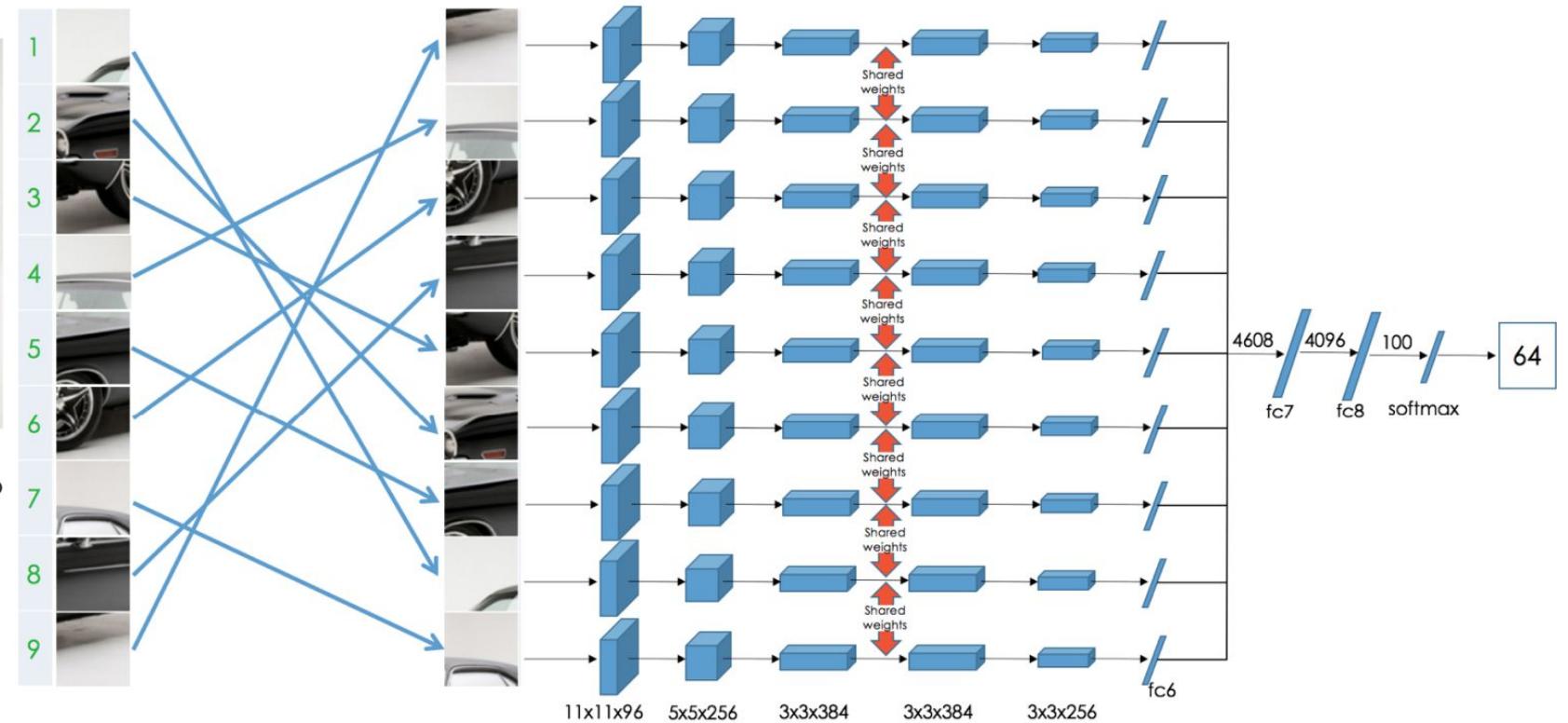
Jigsaw puzzle



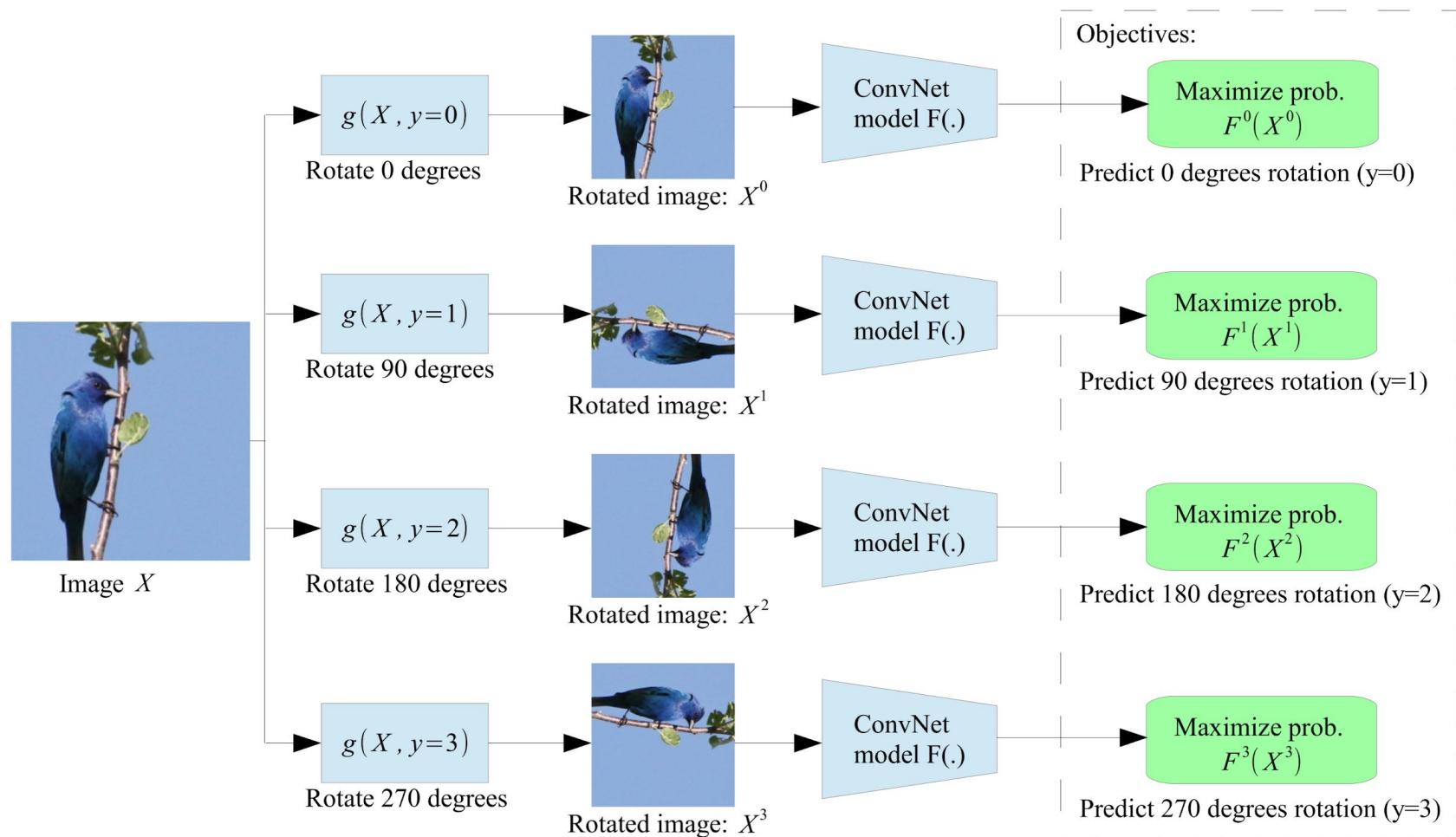
Permutation Set

index	permutation
64	9,4,6,8,3,2,5,1,7

Reorder patches according to the selected permutation



RotNet



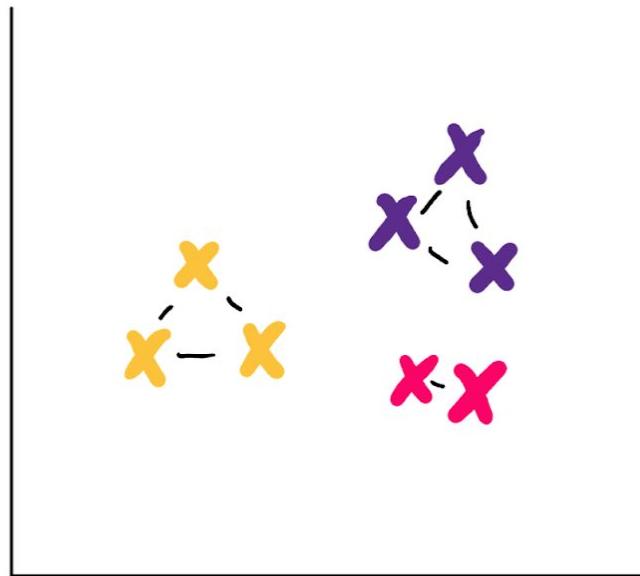
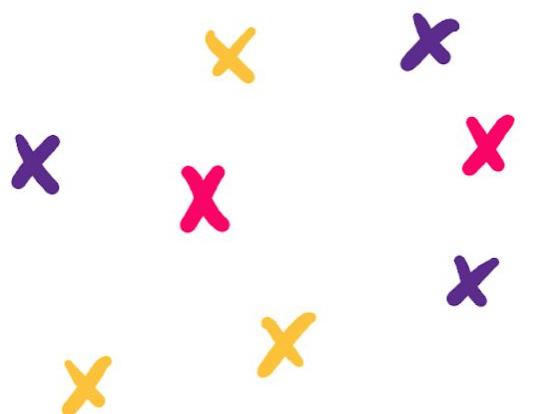
Discussion

- The pseudo-labels are generated with task-specific prior knowledge.
- Not competitive anymore compared to the following methods.

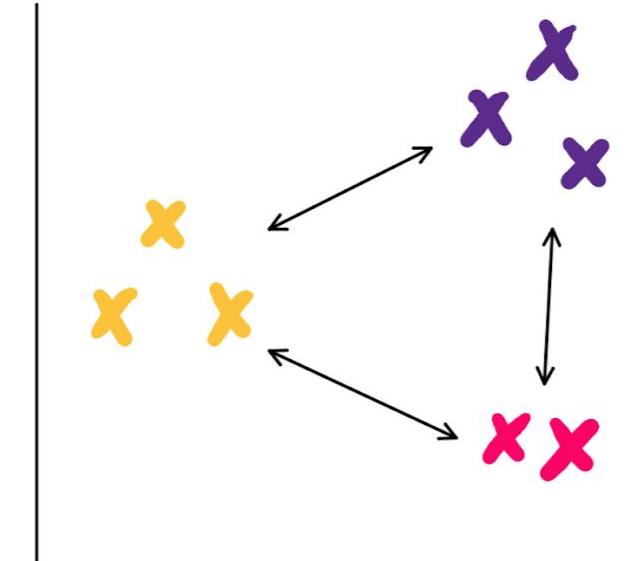
04

Contrastive learning

Contrastive learning



attract similar data



repulse
dissimilar data

Normalized Temperature-scaled Cross Entropy

$$\mathcal{L}(z^a, z^+, Z^-) = -\log \frac{\exp(\text{sim}(z^a, z^+)/\tau)}{\sum_{z^- \in Z^-} \exp(\text{sim}(z^a, z^-)/\tau)}$$

Diagram illustrating the components of the Normalized Temperature-scaled Cross Entropy loss function:

- Anchor**: Points to the variable z^a .
- Positive example**: Points to the variable z^+ .
- Negative examples**: Points to the set Z^- .
- Similarity function**: Points to the term $\text{sim}(z^a, z^-)/\tau$.
- Temperature**: Points to the parameter τ .

Similarity function

$$\text{sim}(z, z') = \frac{z \cdot z'}{\|z\| \|z'\|}$$

↓
Dot product

↑
l2-Norm

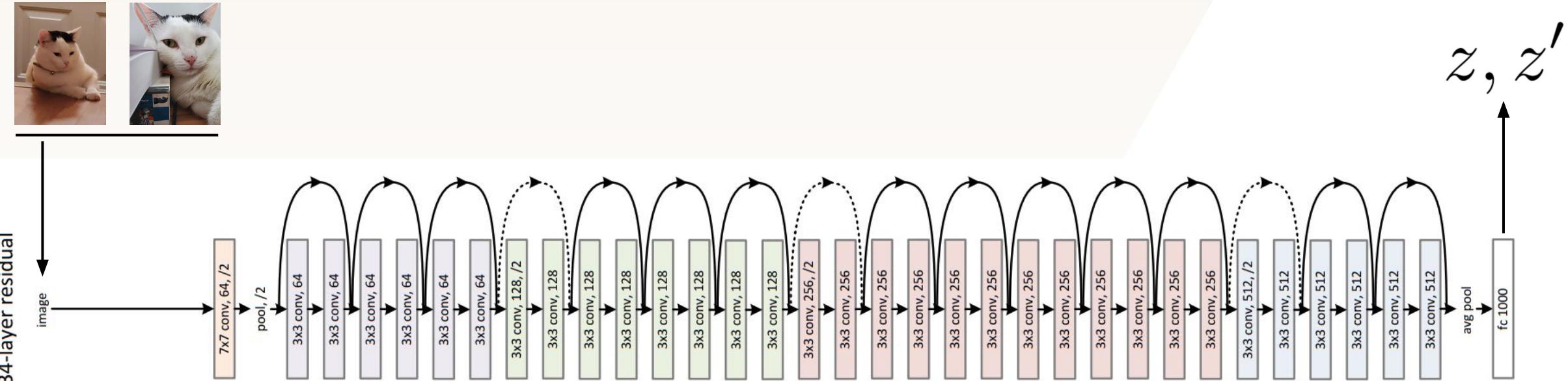
What are the inputs of the similarity function?

$$\text{sim}(z, z') = \frac{z \cdot z'}{\|z\| \|z'\|}$$


The equation shows the formula for calculating the similarity between two vectors z and z' . The inputs z and z' are represented by arrows pointing from two photographs of cats. A large question mark is positioned between the two images, indicating that the inputs are unknown or to be determined.

Backbone encoder

ResNet-50 or ResNet-161

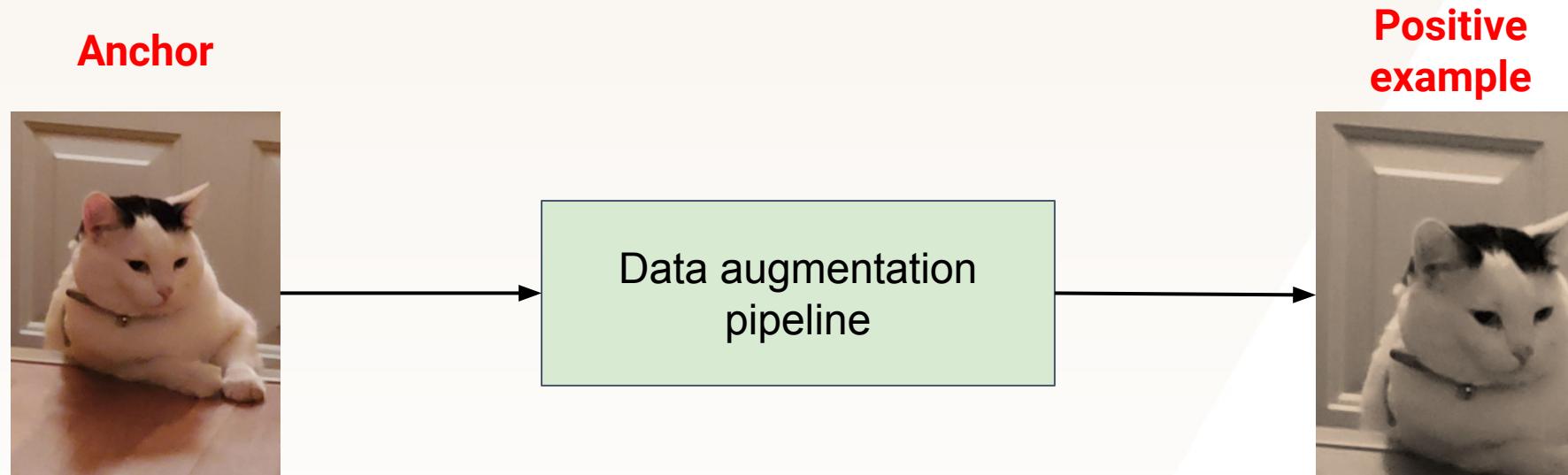


How to select positive and negative examples?

$$\mathcal{L}(z^a, z^+, Z^-) = -\log \frac{\exp(\text{sim}(z^a, z^+)/\tau)}{\sum_{z^- \in Z^-} \exp(\text{sim}(z^a, z^-)/\tau)}$$

Anchor Positive example Negative examples

Data augmentation



Data augmentation

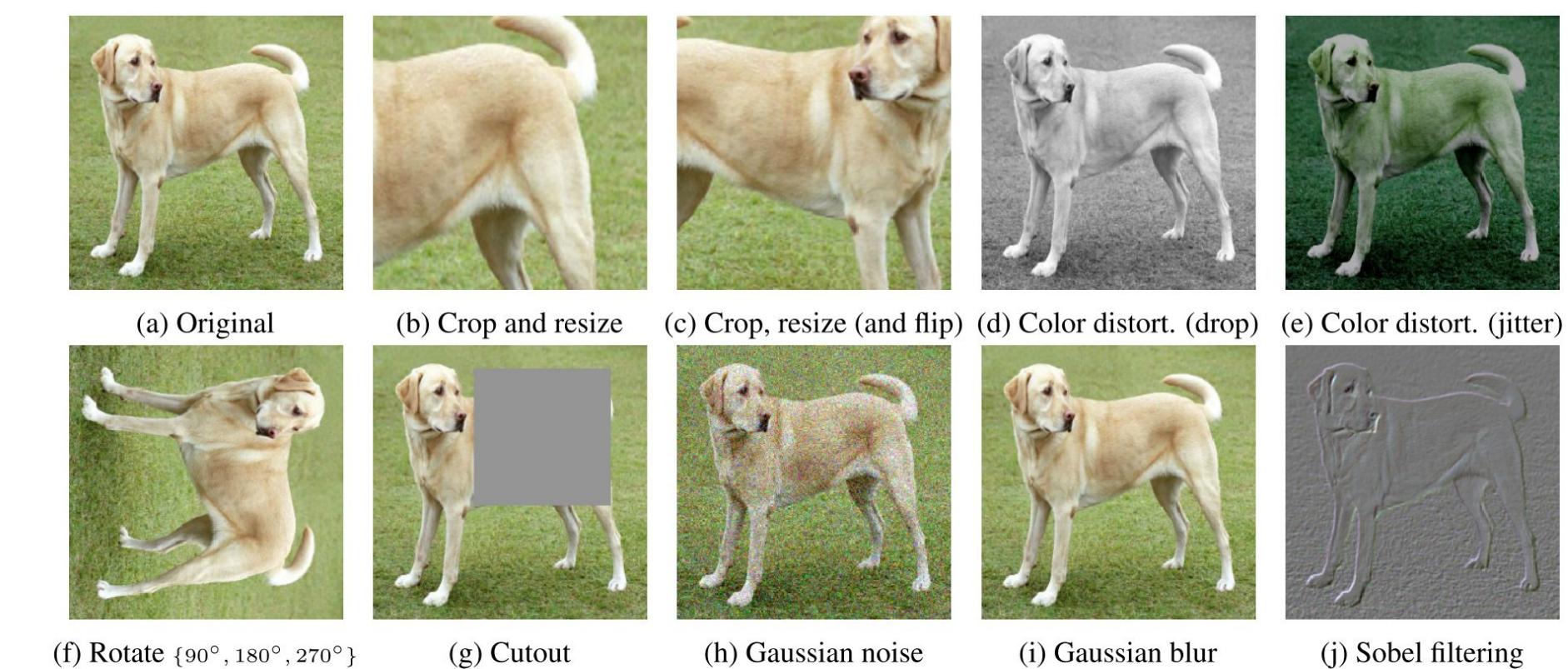
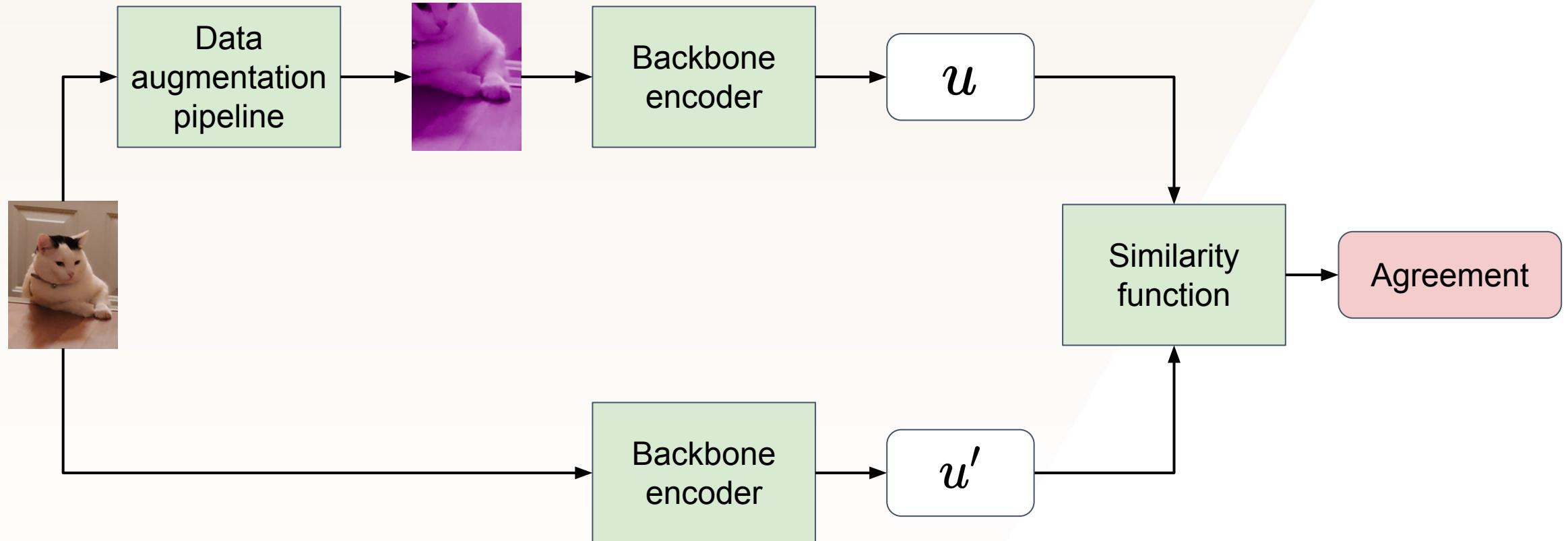
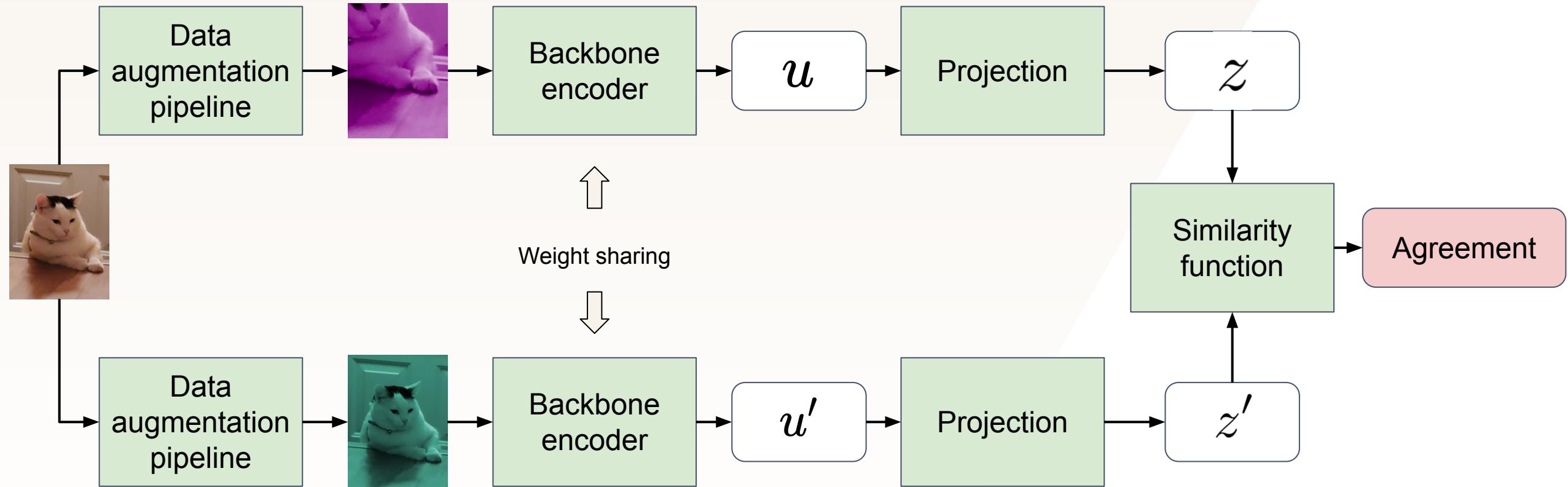
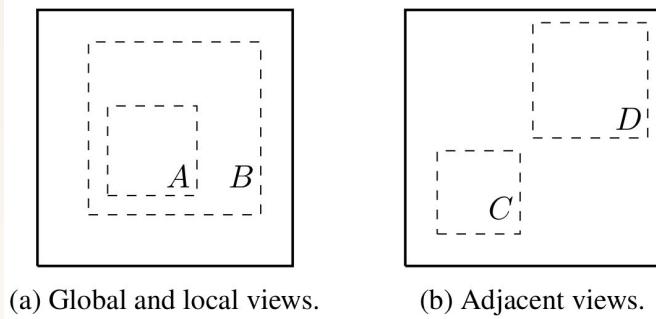


Figure 4. Illustrations of the studied data augmentation operators. Each augmentation can transform data stochastically with some internal parameters (e.g. rotation degree, noise level). Note that we *only* test these operators in ablation, the *augmentation policy used to train our models* only includes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

Putting everything together

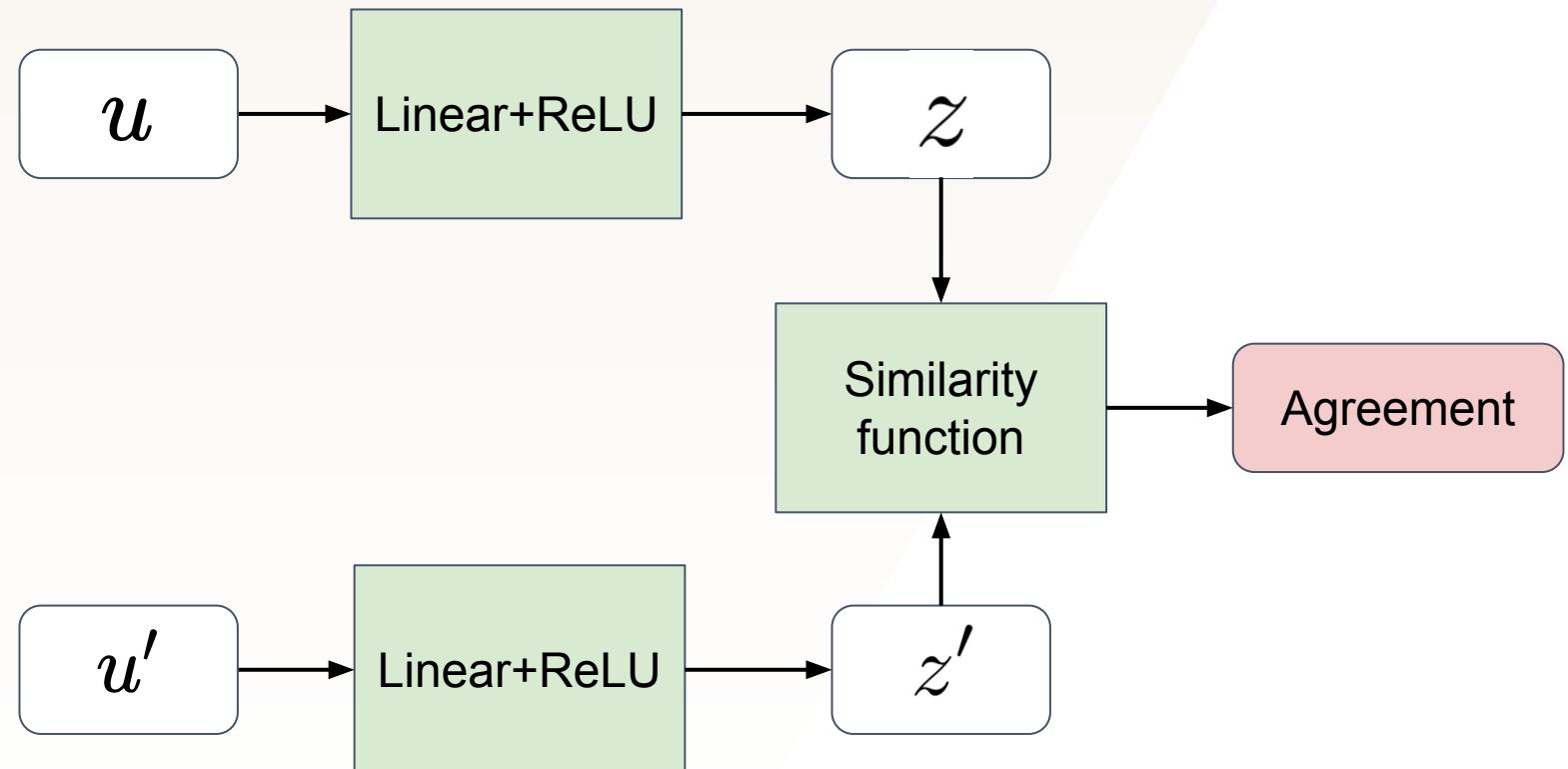


SimCLR



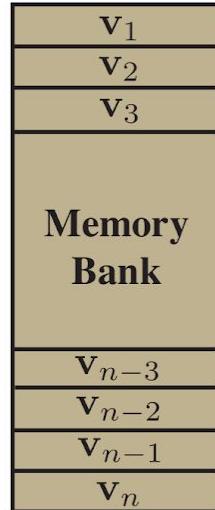
Projection trick

Add a nonlinear projection before computing the similarity



How to build the set of negative examples?

- SimCLR: large mini-batches (~8192) with LARS optimizer [1]
- Memory bank [2]
- Momentum Encoder (MoCo) [3]

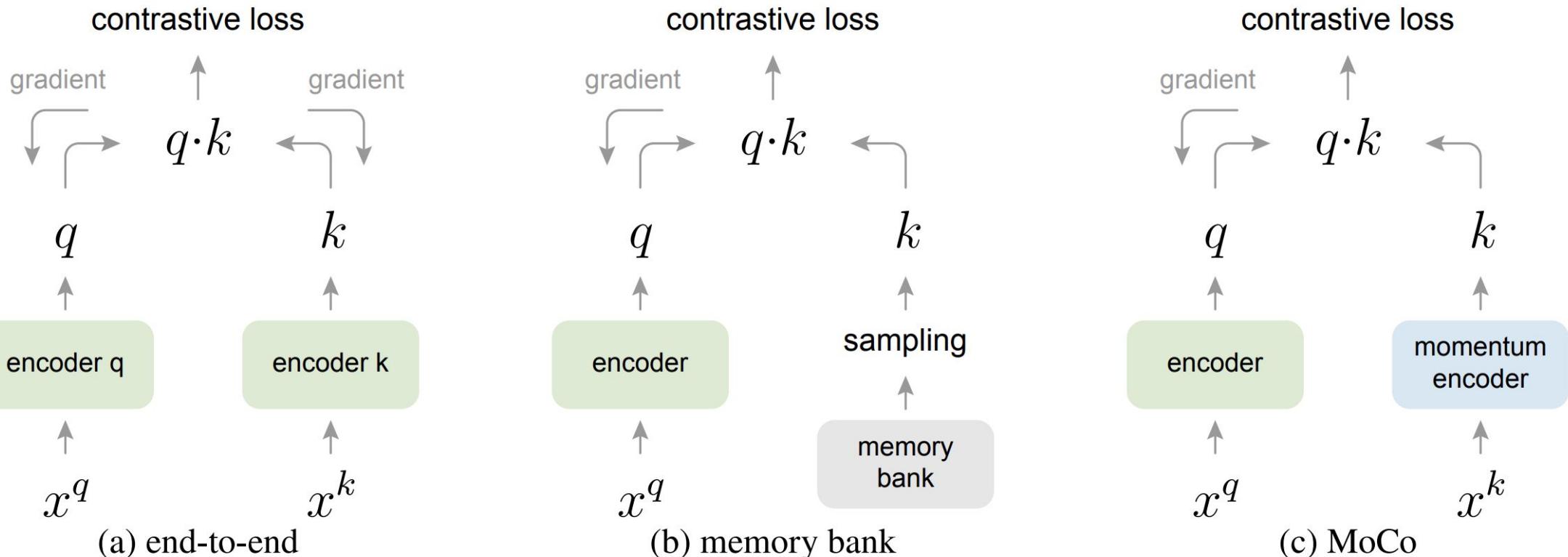


$$\mathcal{L}(z^a, z^+, Z^-) = -\log \frac{\exp(\text{sim}(z^a, z^+)/\tau)}{\sum_{z^- \in Z^-} \exp(\text{sim}(z^a, z^-)/\tau)}$$

↑
**Negative
examples**

[1] Ting Chen, Simon Kornblith, Mohammad Norouzi, Geoffrey E. Hinton: A Simple Framework for Contrastive Learning of Visual Representations. ICML 2020
[2] Zhirong Wu, Yuanjun Xiong, Stella X. Yu, Dahua Lin: Unsupervised Feature Learning via Non-Parametric Instance Discrimination. CVPR 2018: 3733-3742
[3] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, Ross B. Girshick: Momentum Contrast for Unsupervised Visual Representation Learning. CVPR 2020: 9726-9735

How to deal with negative examples?



Momentum encoder

$$\theta_k \leftarrow m\theta_k + (1 - m)\theta_q$$

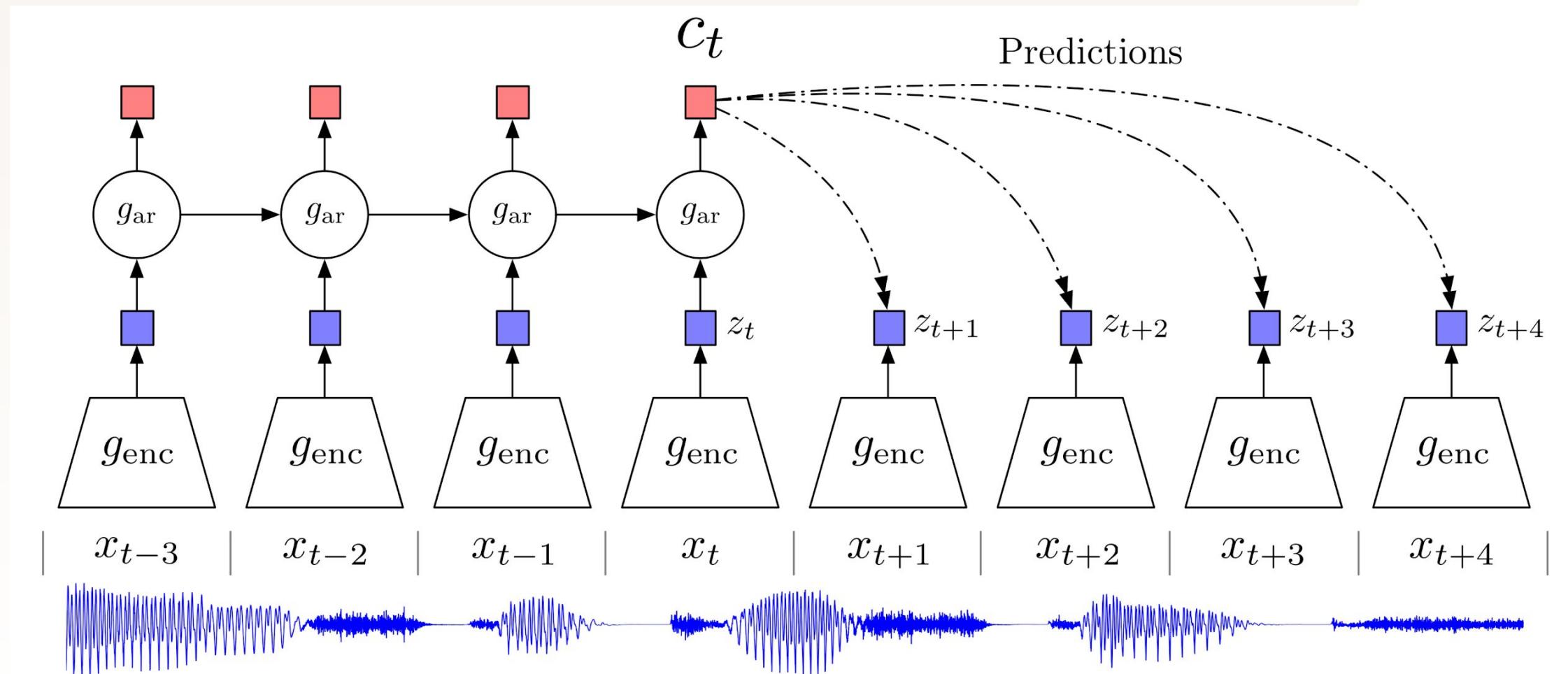
Discussion

- Handling negative examples is difficult.
- Require large batch training.
- They are computationally intensive and work best on huge unlabeled datasets.

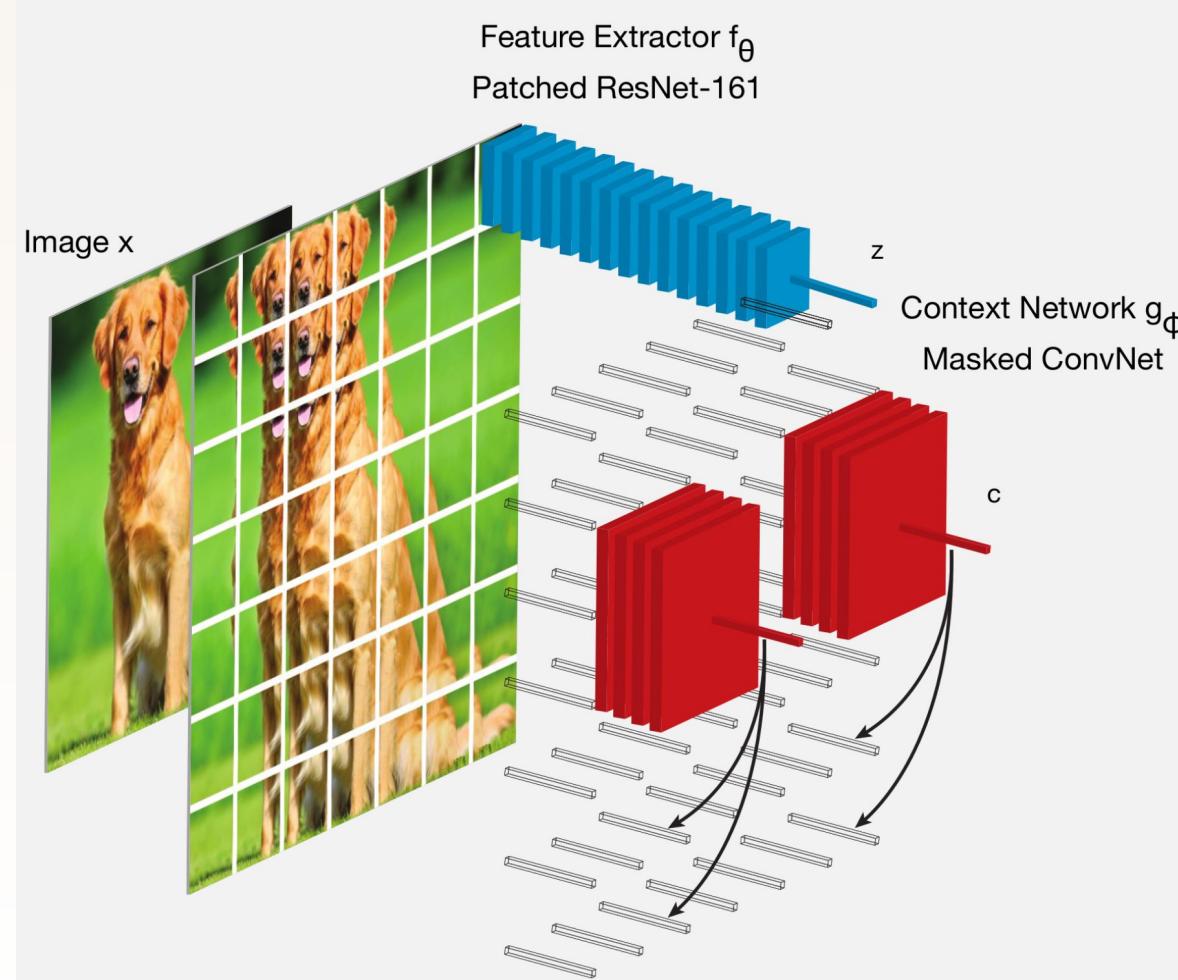
05

Autoregressive contrastive learning

Contrastive Predictive Coding (CPC)



Contrastive Predictive Coding (CPCv2)



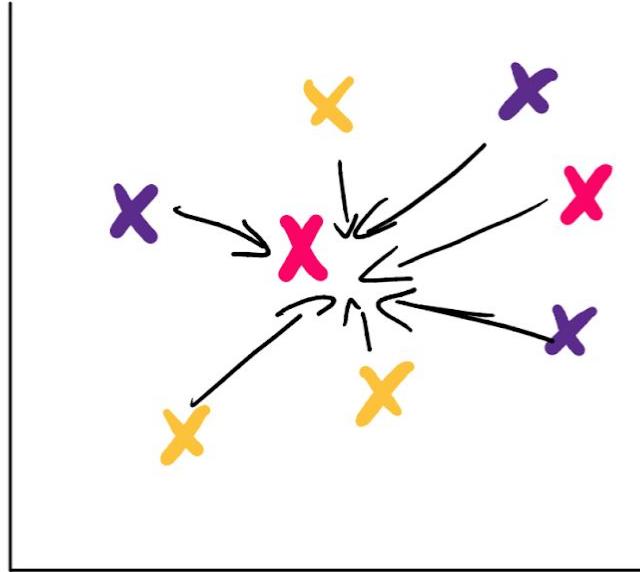
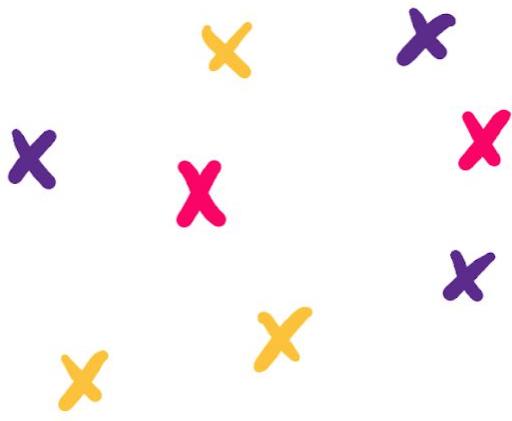
Discussion

- Require the user to fix an order, which may not be intuitive for some types of data such as images.

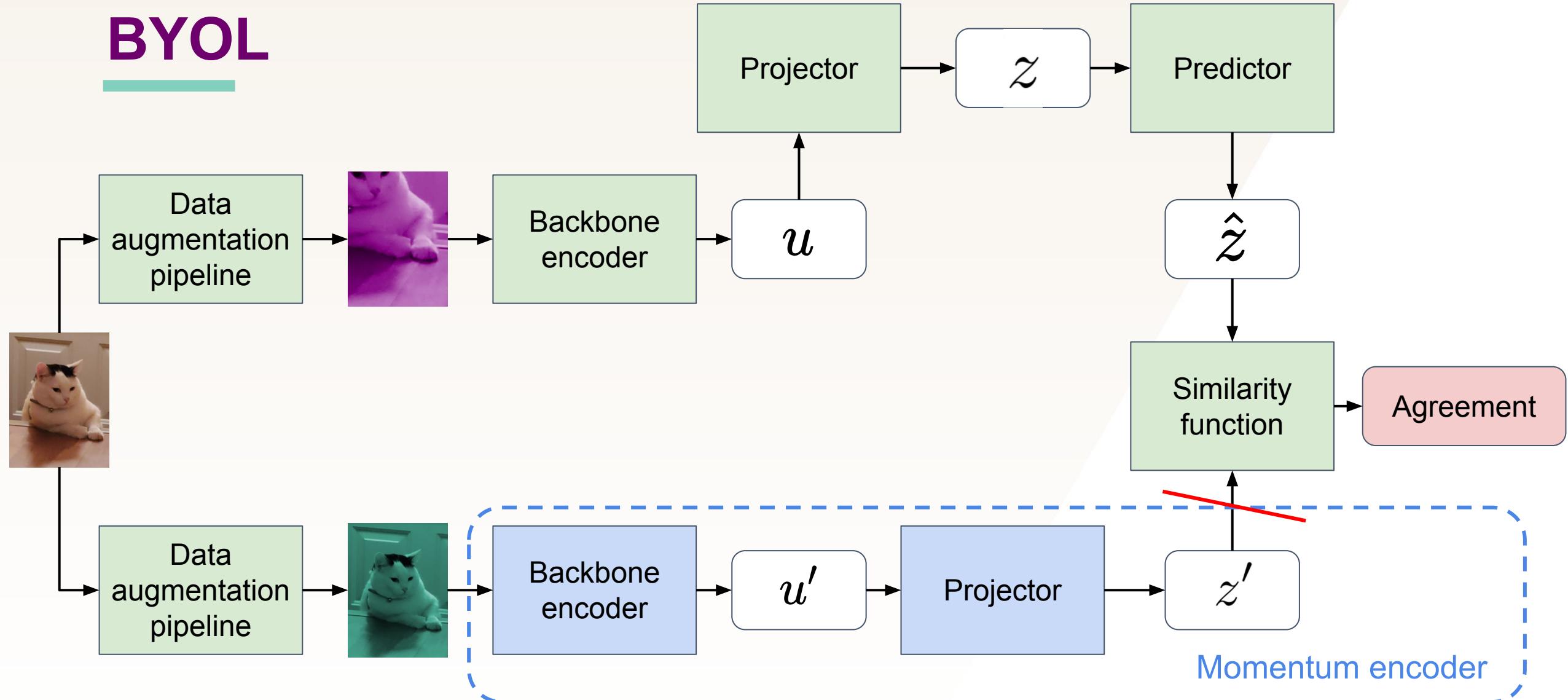
06

“Contrastive learning” without negative examples

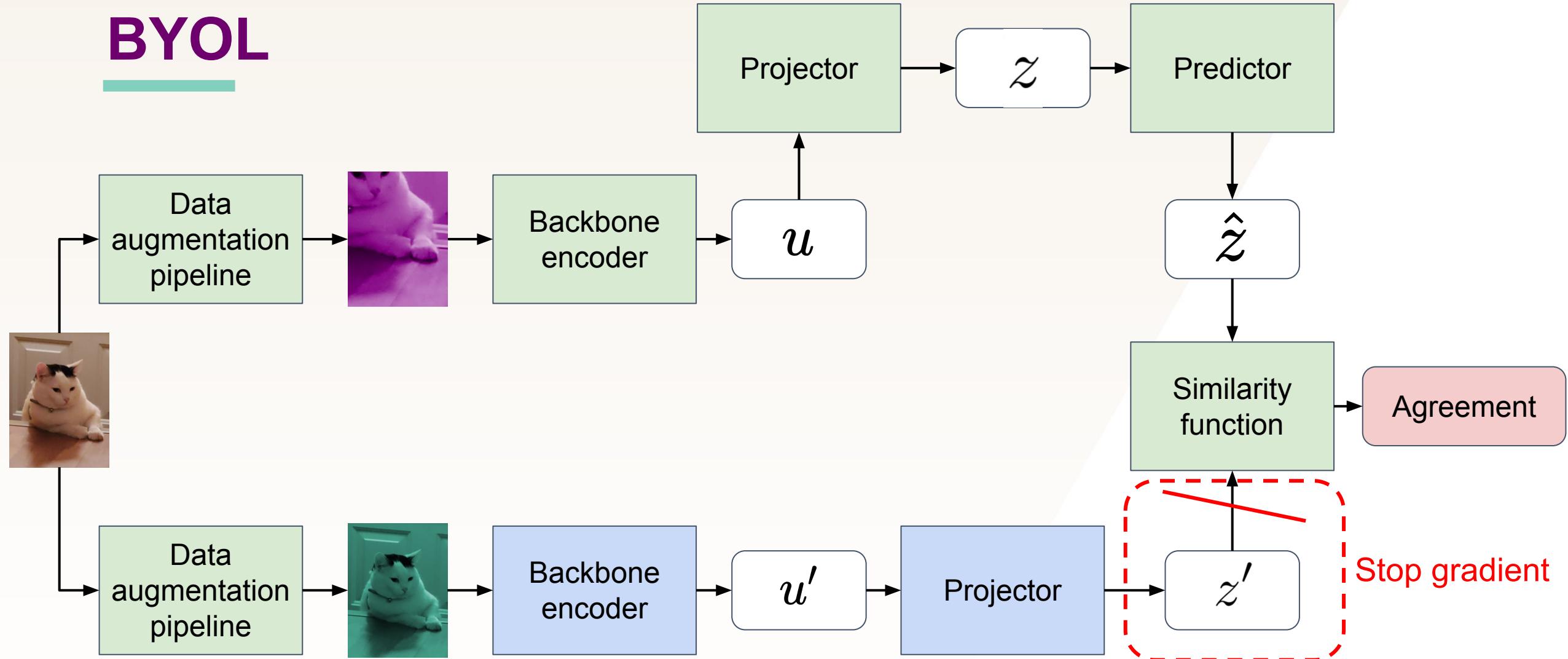
The collapse problem



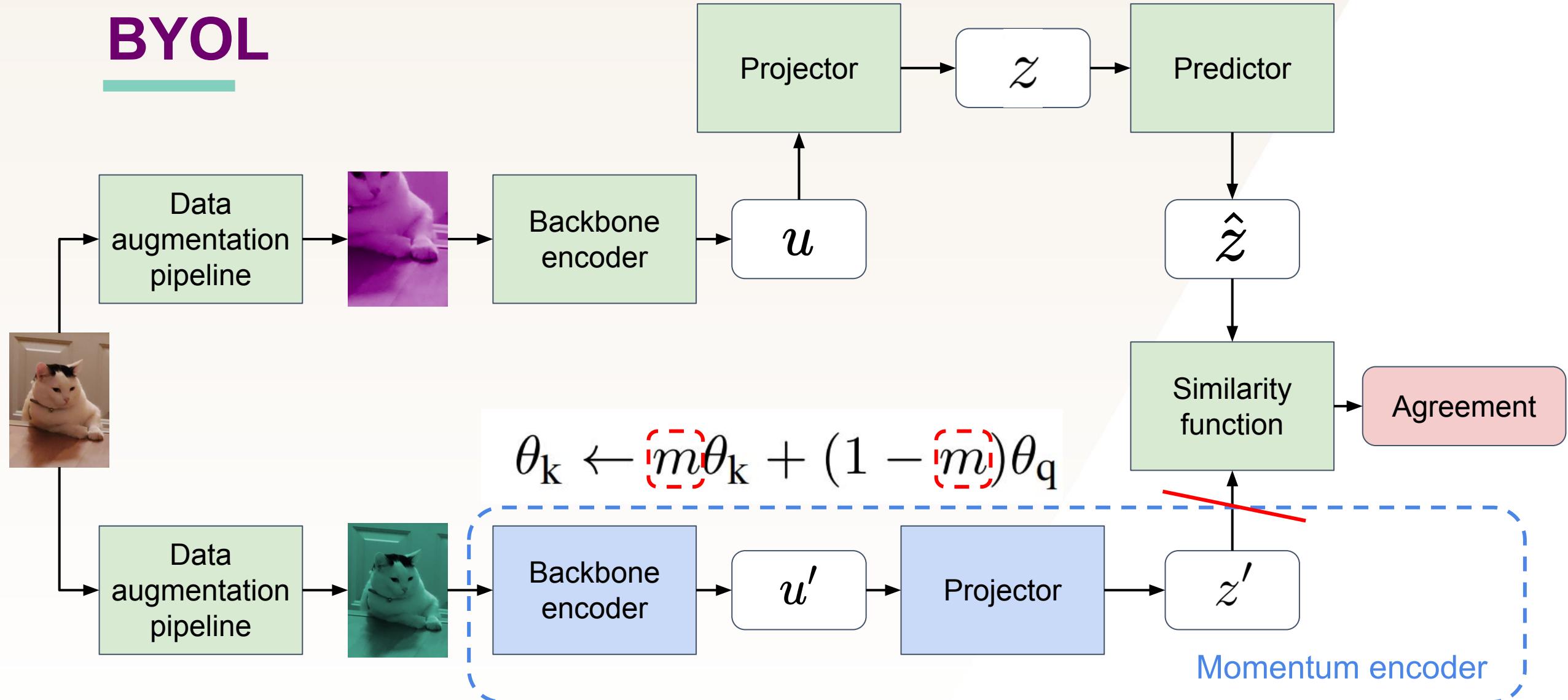
BYOL



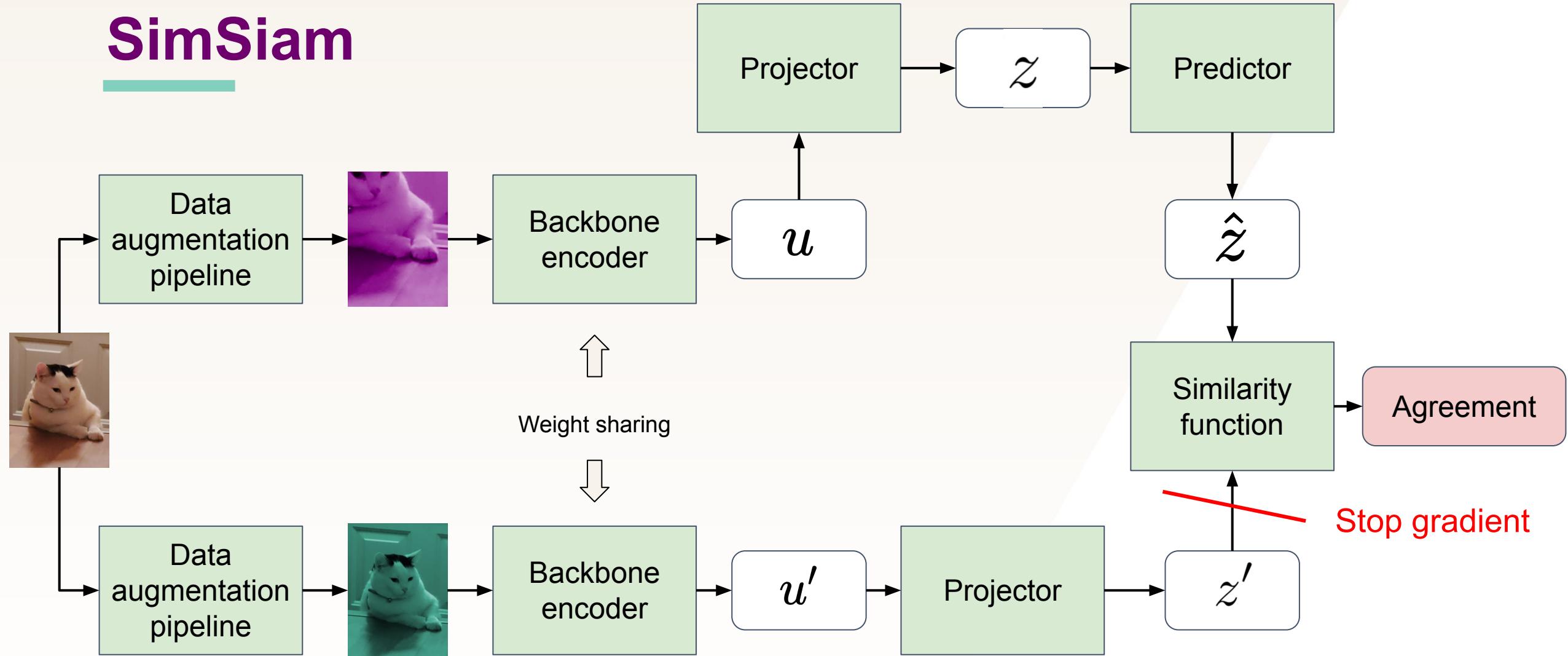
BYOL



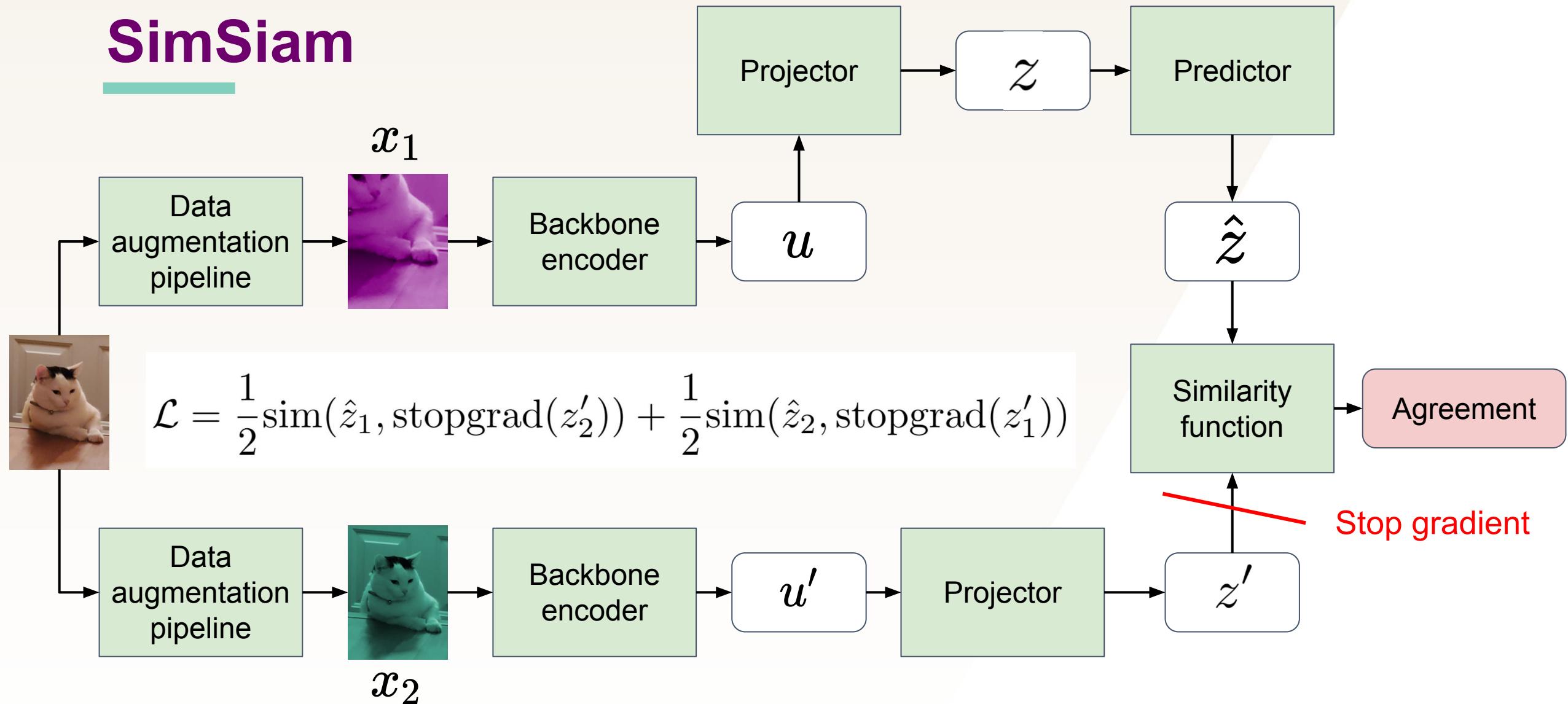
BYOL



SimSiam



SimSiam



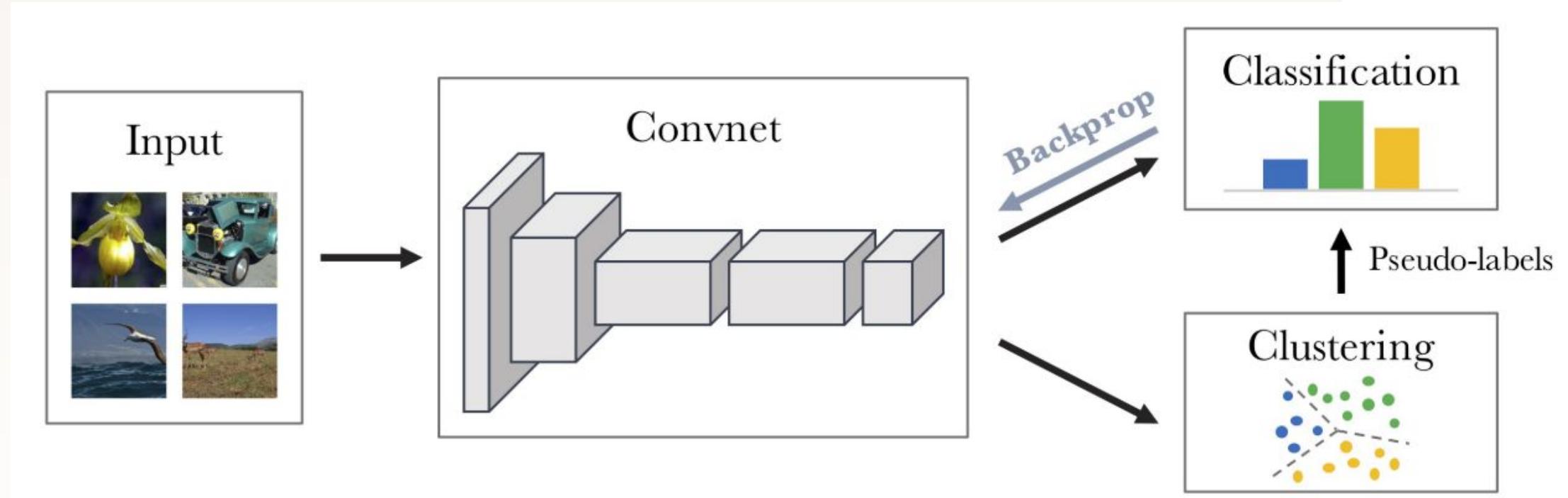
Discussion

- Different solutions to avoid the collapse problem.
- Competitive results compared with purely contrastive methods.

07

Contrastive learning with clustering

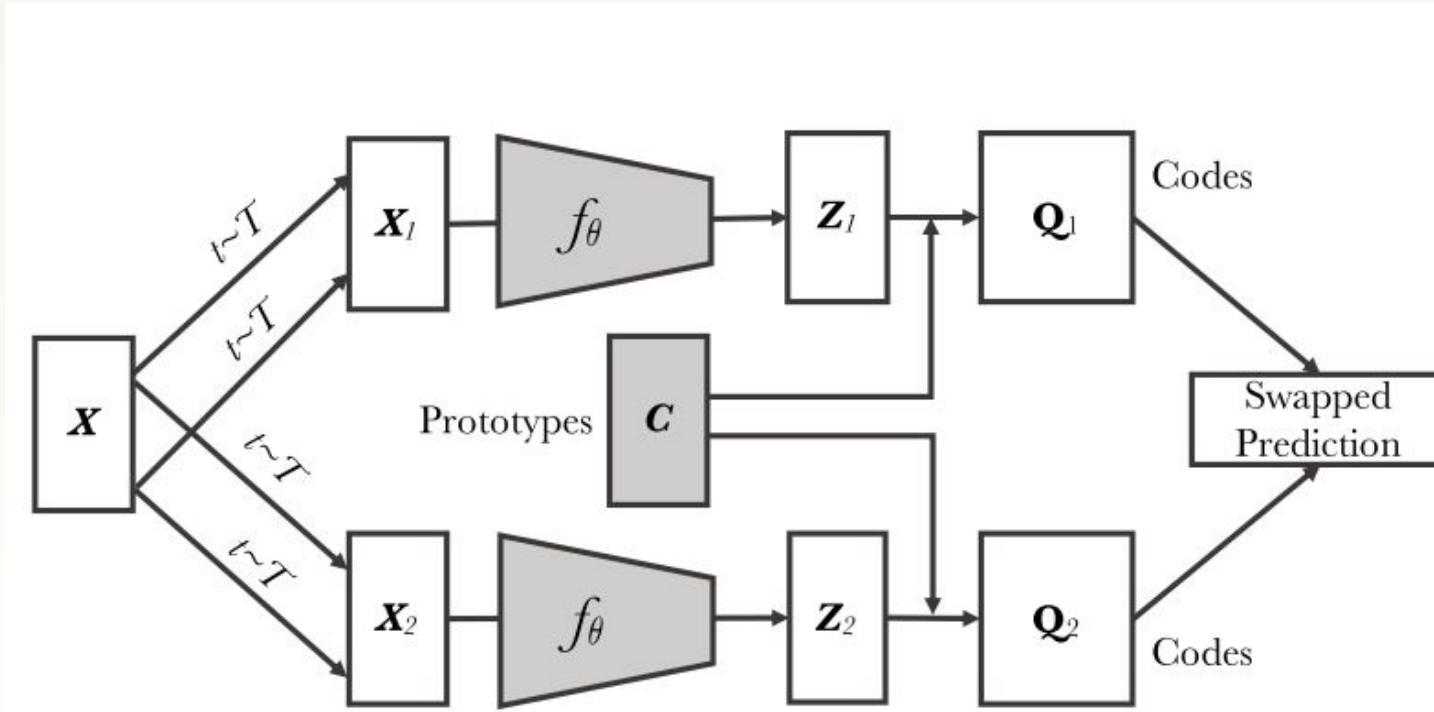
Clustering as pretext task



SwAV: Contrastive and clustering together



SwAV: Contrastive and clustering together

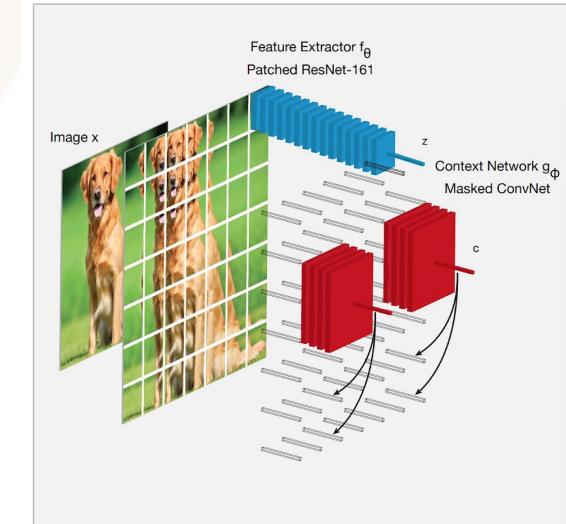
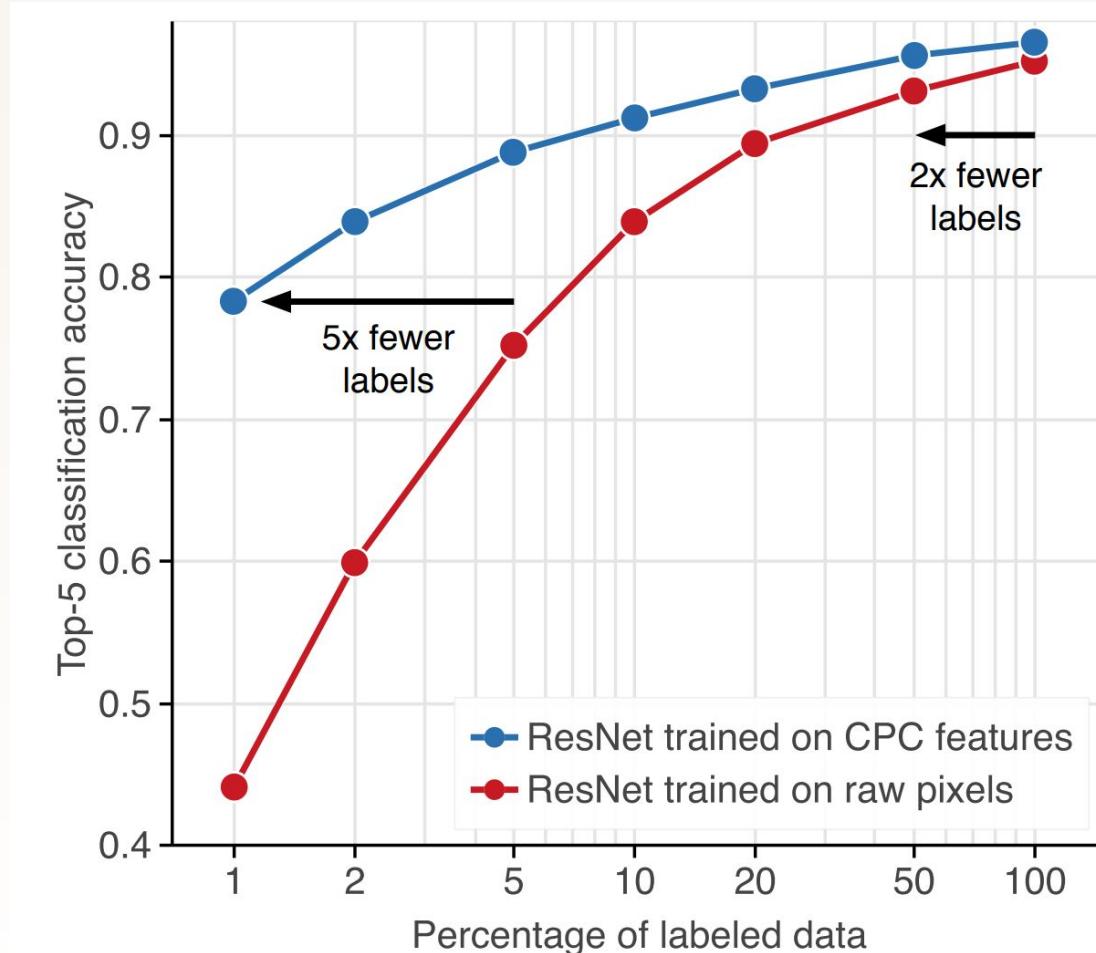


Swapping Assignments between Views (Ours)

08

Empirical results

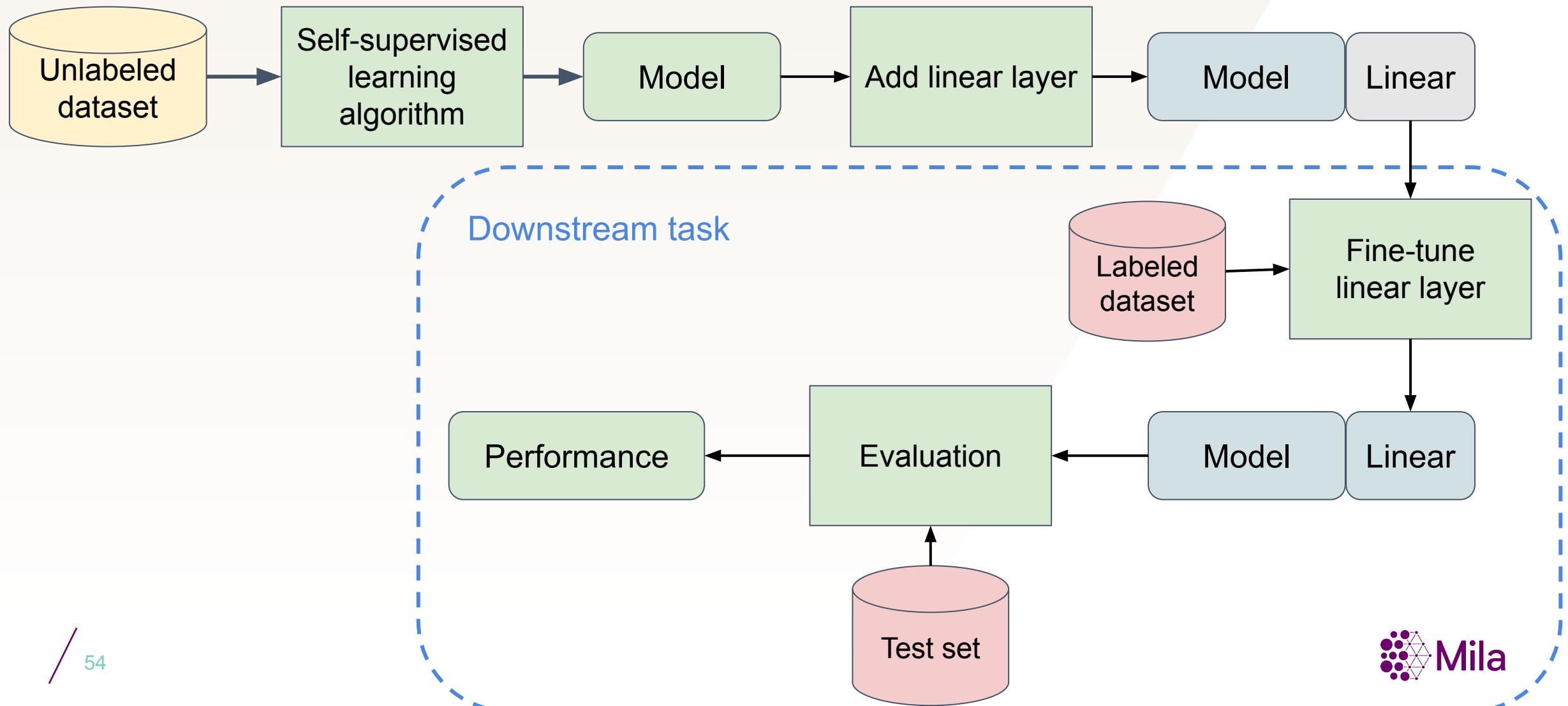
Empirical results: CPC v2



Empirical results

Method	1% labels		10% labels	
	Top-1	Top-5	Top-1	Top-5
Supervised	25.4	48.4	56.4	80.4
SimCLR [10]	48.3	75.5	65.6	87.8
SwAV	53.9	78.5	70.2	89.9

Linear evaluation protocol



Empirical results

Method	Arch.	Param.	Top1
Supervised	R50	24	76.5
Colorization [64]	R50	24	39.6
Jigsaw [45]	R50	24	45.7
NPID [57]	R50	24	54.0
BigBiGAN [15]	R50	24	56.6
LA [67]	R50	24	58.8
NPID++ [43]	R50	24	59.0
MoCo [24]	R50	24	60.6
SeLa [2]	R50	24	61.5
PIRL [43]	R50	24	63.6
CPC v2 [27]	R50	24	63.8
PCL [36]	R50	24	65.9
SimCLR [10]	R50	24	70.0
MoCov2 [11]	R50	24	71.1
SwAV	R50	24	75.3

Mathilde Caron, Ishan Misra, Julien Mairal, Priya Goyal, Piotr Bojanowski, Armand Joulin: Unsupervised Learning of Visual Features by Contrasting Cluster Assignments. NeurIPS (2020)
Grill, Jean-Bastien, et al. "Bootstrap Your Own Latent-A New Approach to Self-Supervised Learning." NeurIPS (2020).

Discussion

- Computation budget and datasets can be huge.
- SSL methods in vision depend heavily on data augmentations.
 - How to automate data augmentation discoveries?
- Pytorch lightning implements a SSL framework for fast prototyping.
- SSL has a direct impact on democratizing DL.

09

Questions

Thank you!