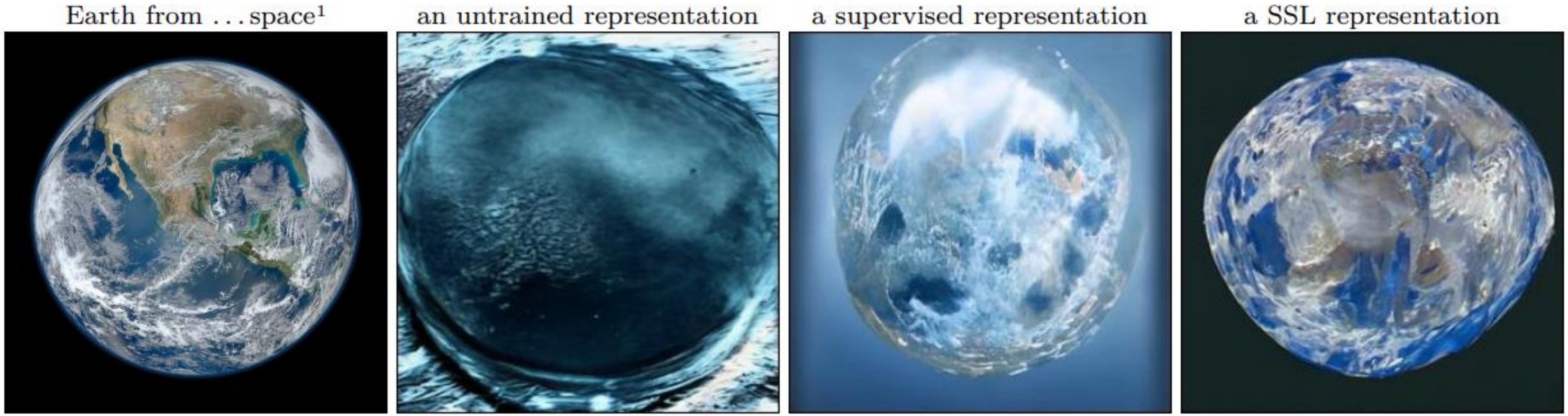
Visualizing self-supervised models' knowledges

High Fidelity Visualization of What Your Self-Supervised Representation Knows About (TMLR 2022, Meta)



Denis Shepelev, Sep 2023
https://github.com/denzist/presentations

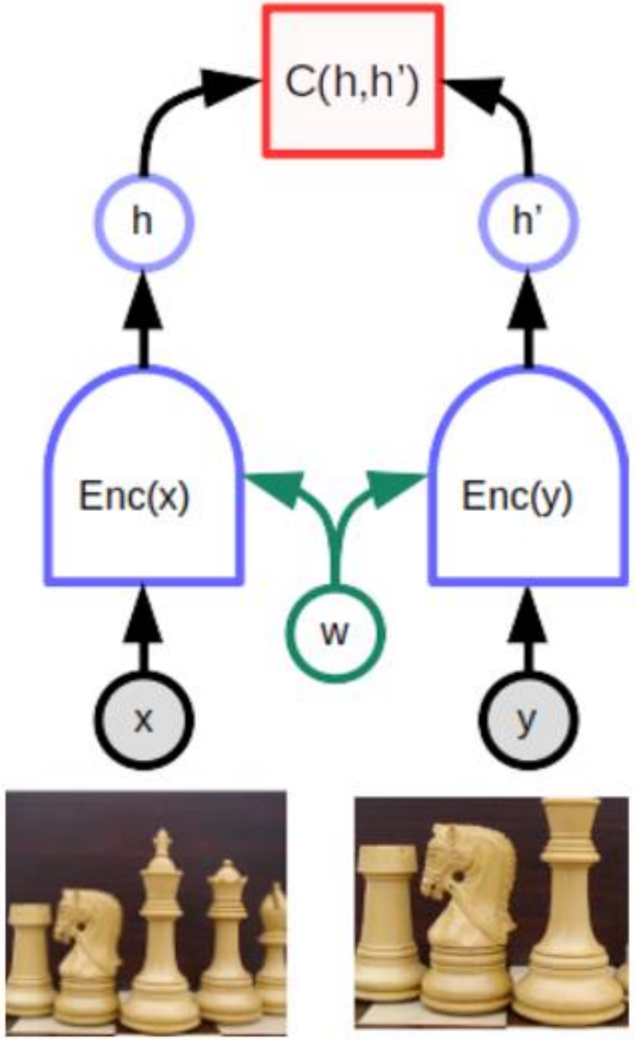
Plan

- Self-supervised learning (SSL)
- Knowledge visualization (KV) methods
- Representation Conditional Diffusion Model (RCDM)
- SSL backbone vs SSL projection vs supervised
- Conclusions

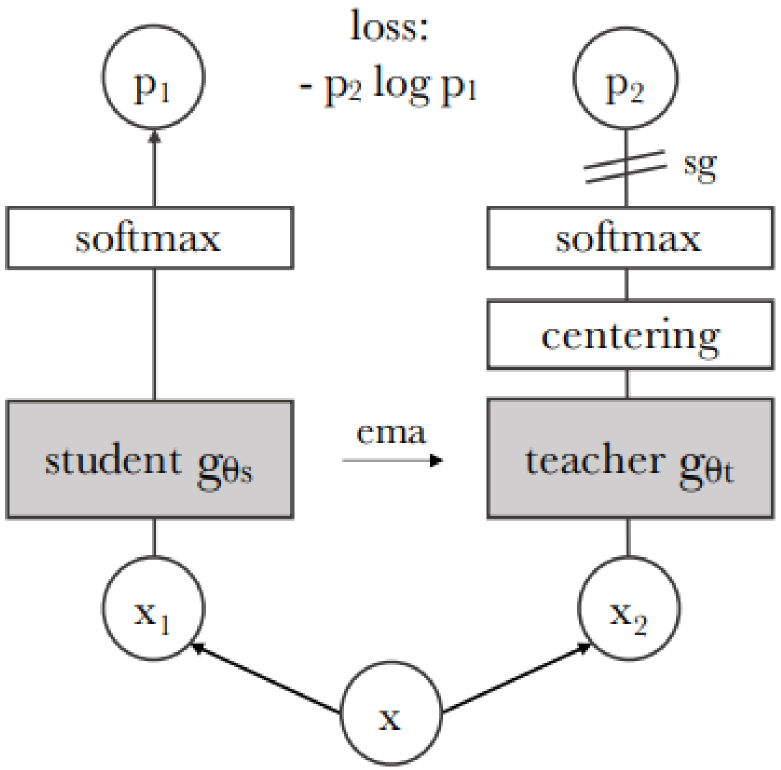
SSL

- SSL leverages the underlying structure of the data **without relying on human labels**, obtaining supervisory signals from the data itself
- **The aim of SSL is to obtain generalized features** that can be used for **other various downstream tasks,** like classification, segmentation, depth estimation, etc.
- These downstream tasks then used to quantitively evaluate the representations (knowledges) of self-supervised models (SSM) that were learned

Typical SSL scheme

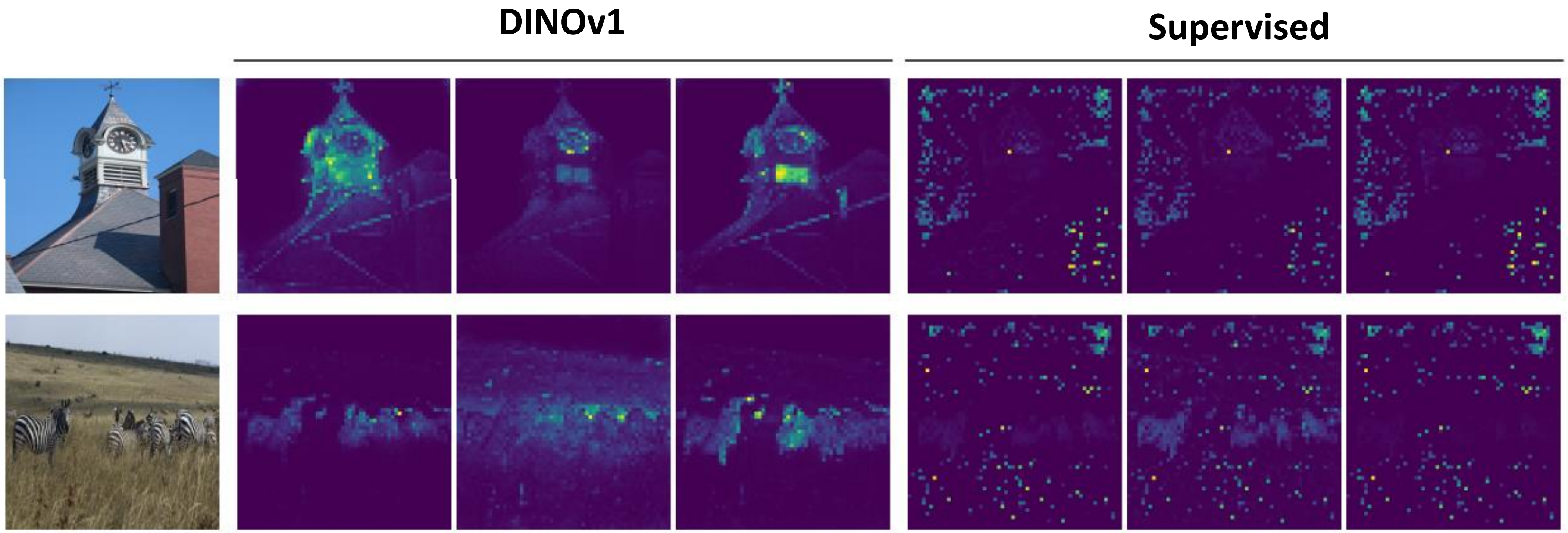


SSM example: DINOv1

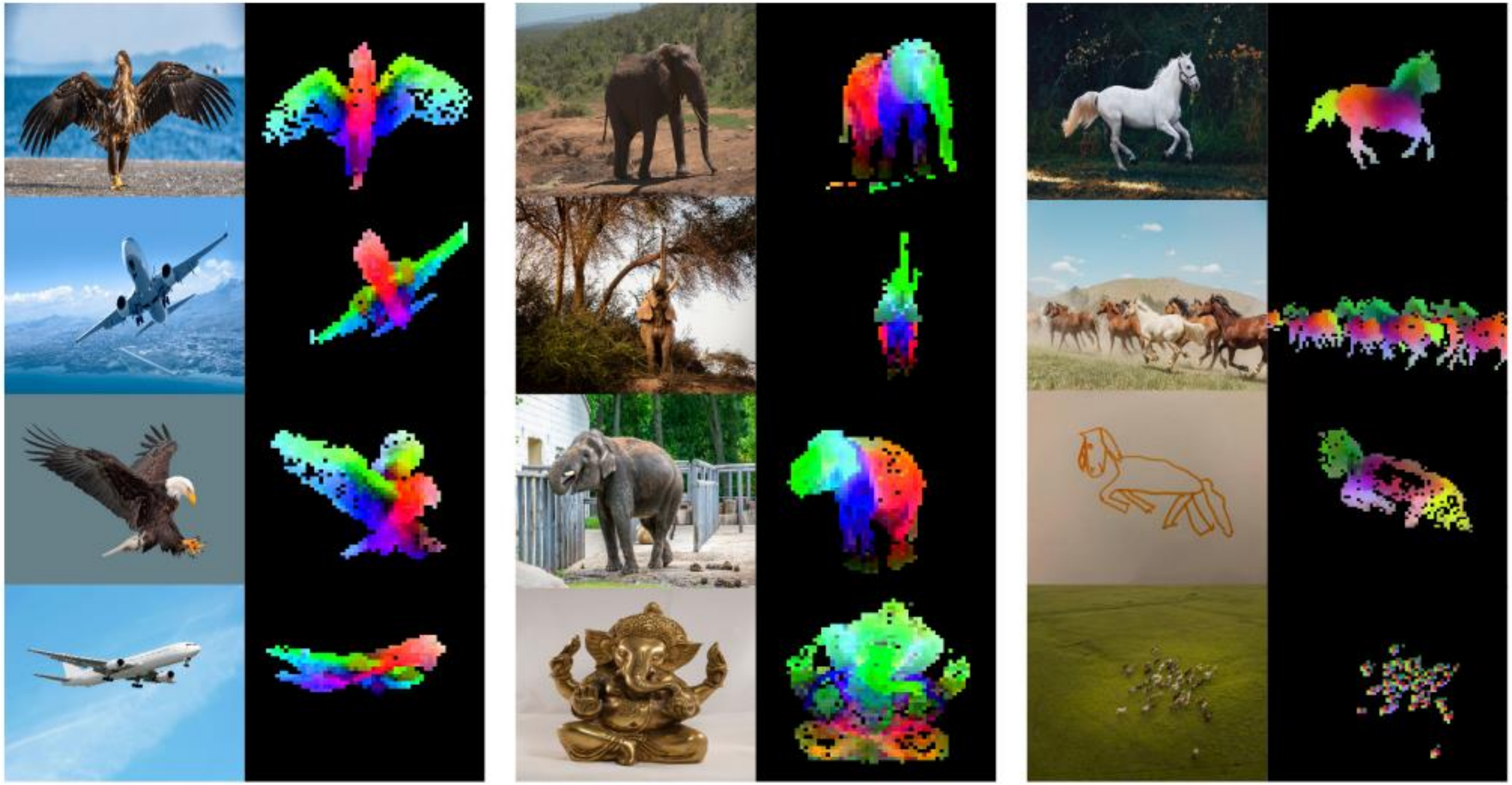


Why we need visualization?

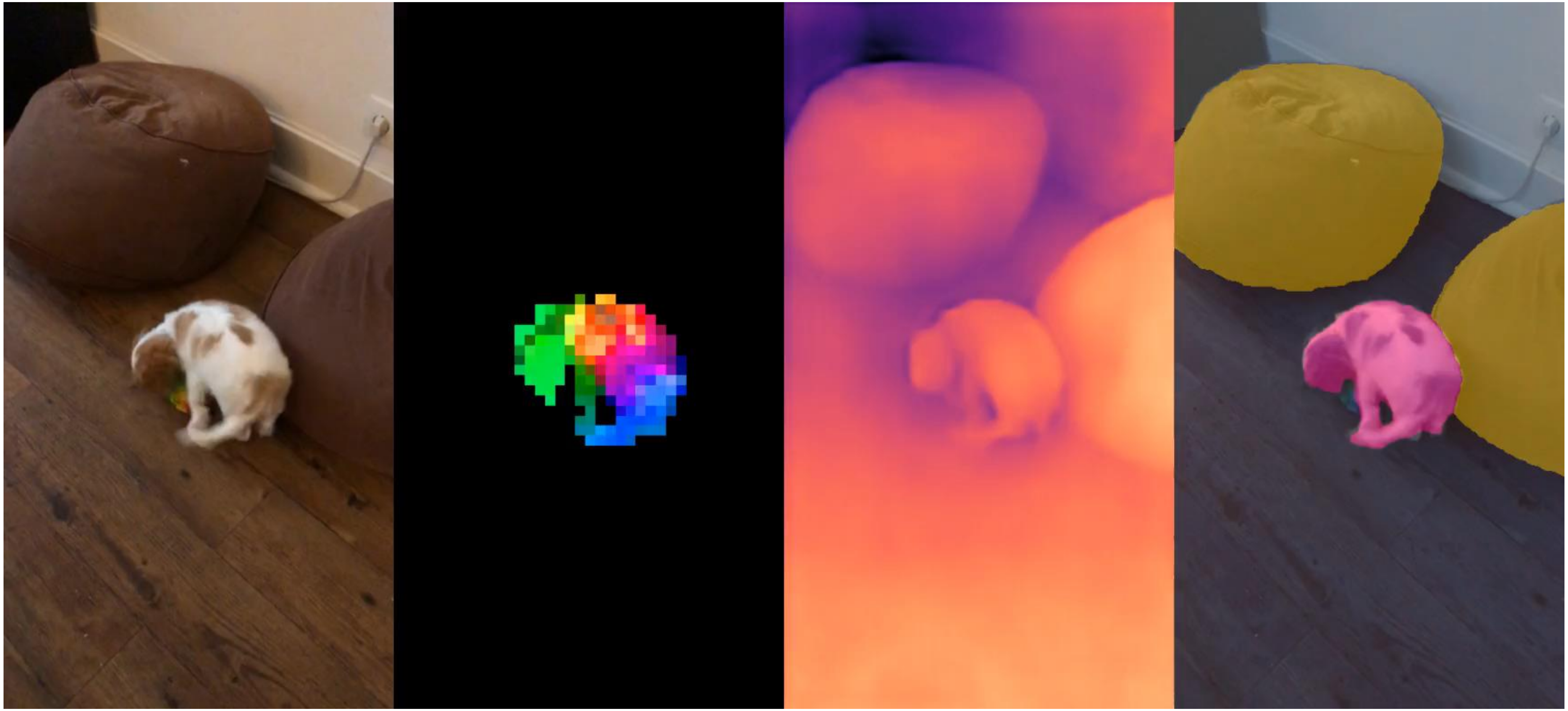
- Downstream tasks are used to quantitively evaluate the representations (knowledges) of self-supervised models (SSM) that were learned
- However, **relying only on downstream tasks only can limit our understanding of what information is learned by models**
- To analyze qualitatively retained various **knowledge visualization (KV)** methods can be used



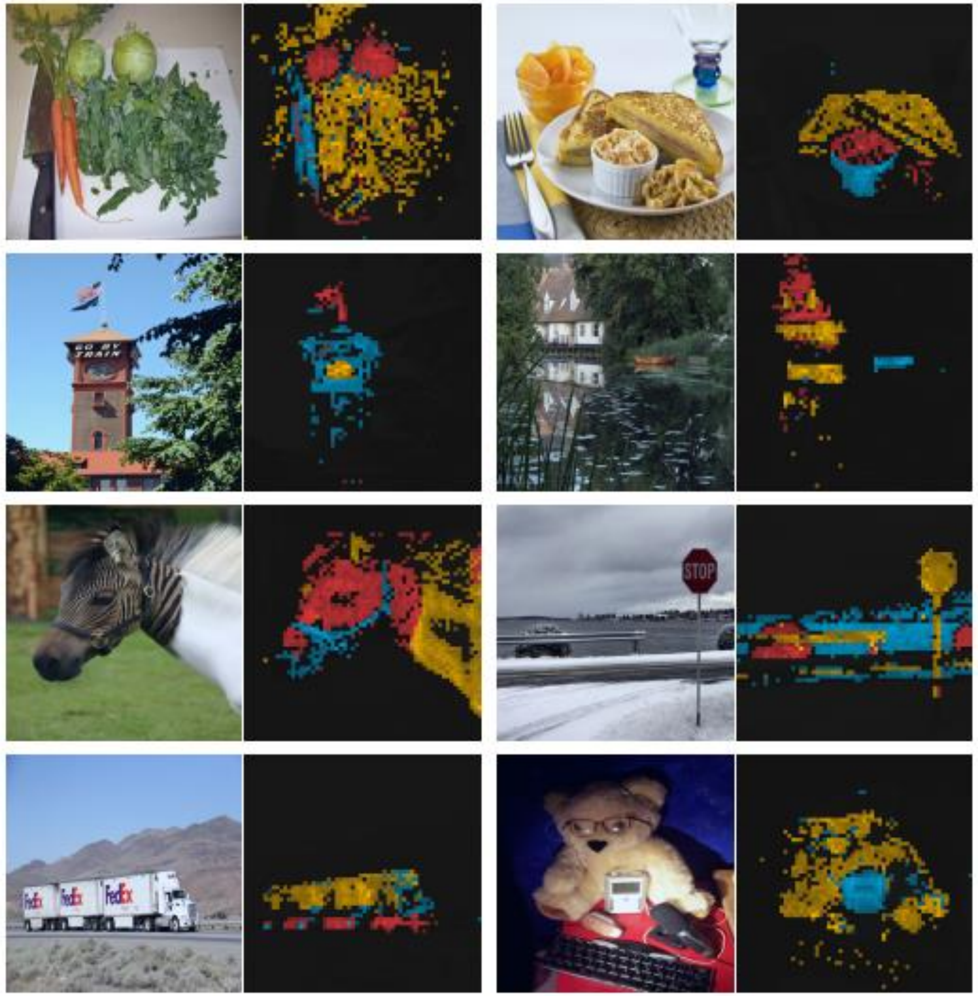
Attention maps visualization (DINOv2)



Downstream tasks visualization (DINOv2)



Segmentation on self-attention masks (DINOv1)



	Random	Supervised	DINO
ViT-S/16	22.0	27.3	45.9
ViT-S/8	21.8	23.7	44.7

Figure 4: **Segmentations from supervised versus DINO.** We visualize masks obtained by thresholding the self-attention maps to keep 60% of the mass. On top, we show the resulting masks for a ViT-S/8 trained with supervision and DINO. We show the best head for both models. The table at the bottom compares the Jaccard similarity between the ground truth and these masks on the validation images of PASCAL VOC12 dataset.

Learned patch patterns (iBOT)

What do we not understand about SSL?

- Why is SSL projector discarded when applied on downstream tasks?
  Why a backbone is better on downstream tasks than a projector?
- How do SSL augmentations affect the backbone/projector?
- What is learned on the backbone and projector levels?

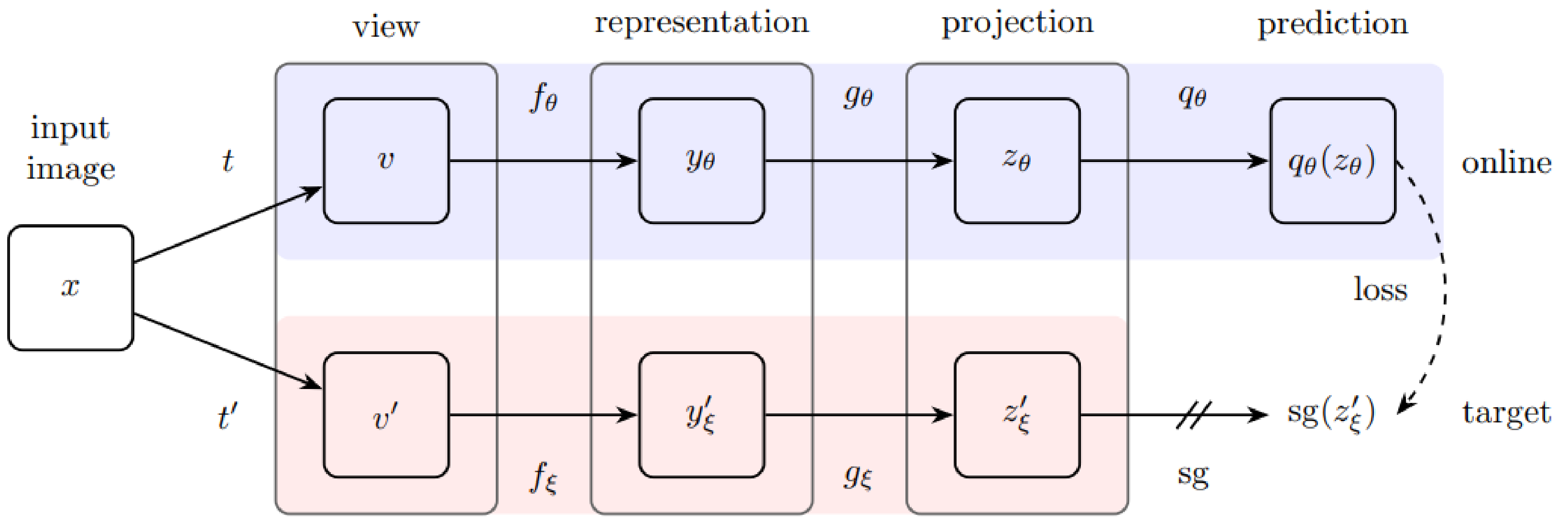
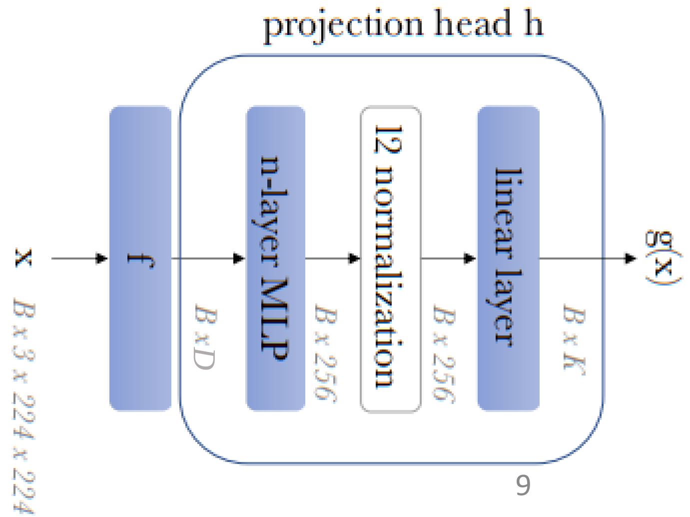
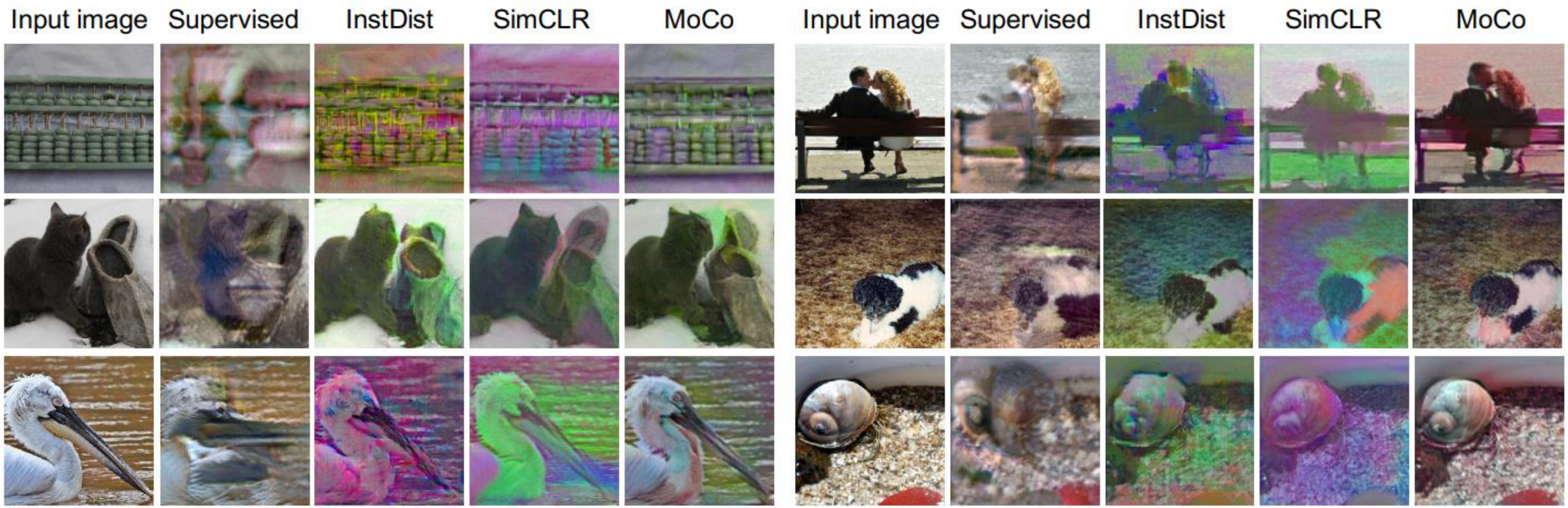


Figure 2: BYOL's architecture. BYOL minimizes a similarity loss between $q_\theta(z_\theta)$ and $\mathrm{sg}(z'_\xi)$, where θ are the trained weights, ξ are an exponential moving average of θ and sg means stop-gradient. At the end of training, everything but f_θ is discarded, and y_θ is used as the image representation.

- DINOv1 ResNet-50 backbone dim.: 2048
- DINOv1 projector bottleneck dim.: 256
- DINOv1 projector dim.: 65536 (?)



Gradient based methods

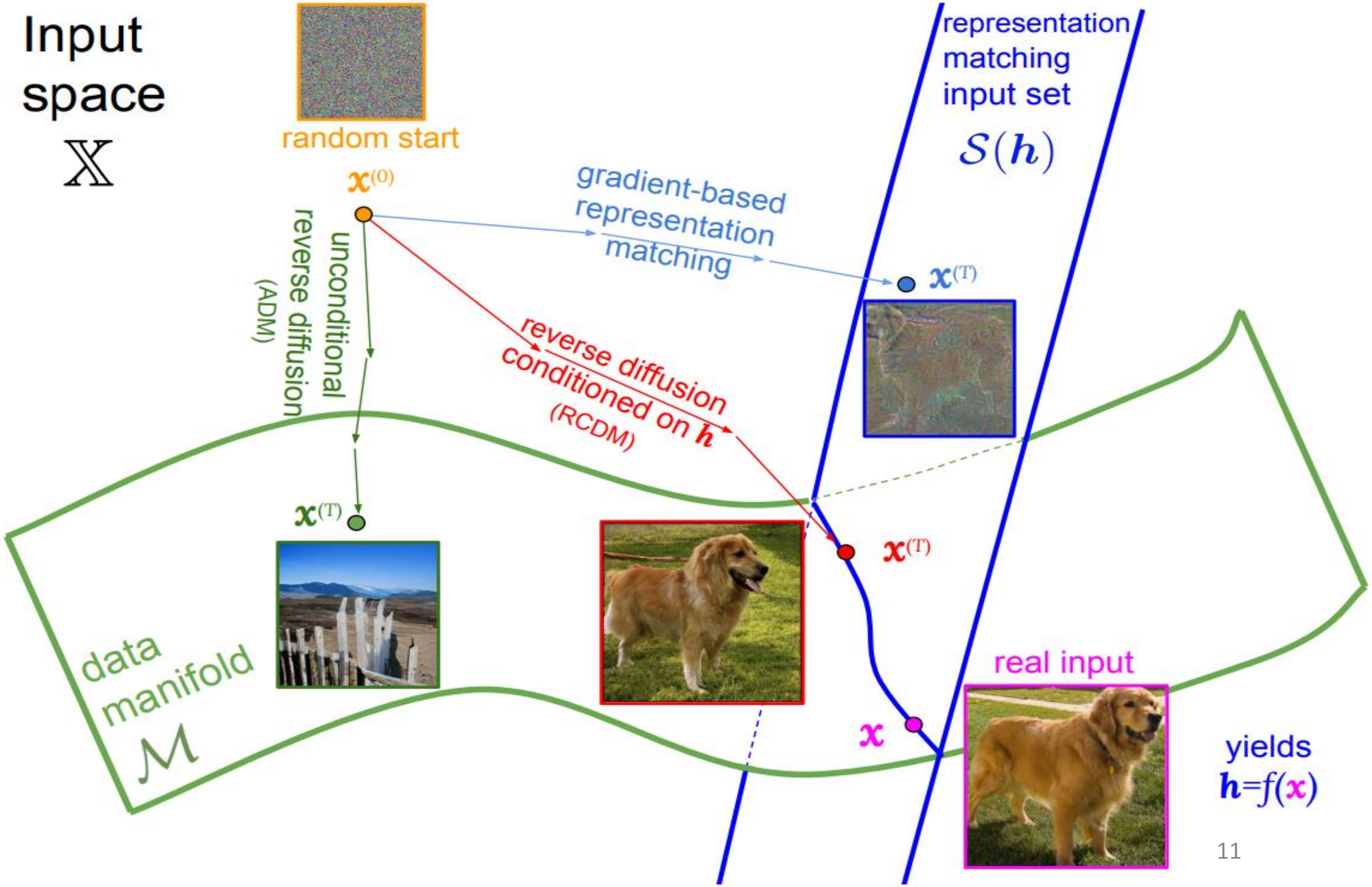


Deep Image Prior: $\min_\theta E(f(x), f(x_0)), \quad x = r_\theta(z_0)$

What makes instance discrimination good for transfer learning? (Zhao et al, ICLR 2021)
Deep Image Prior (Ulyanov et al, CVPR 2018)

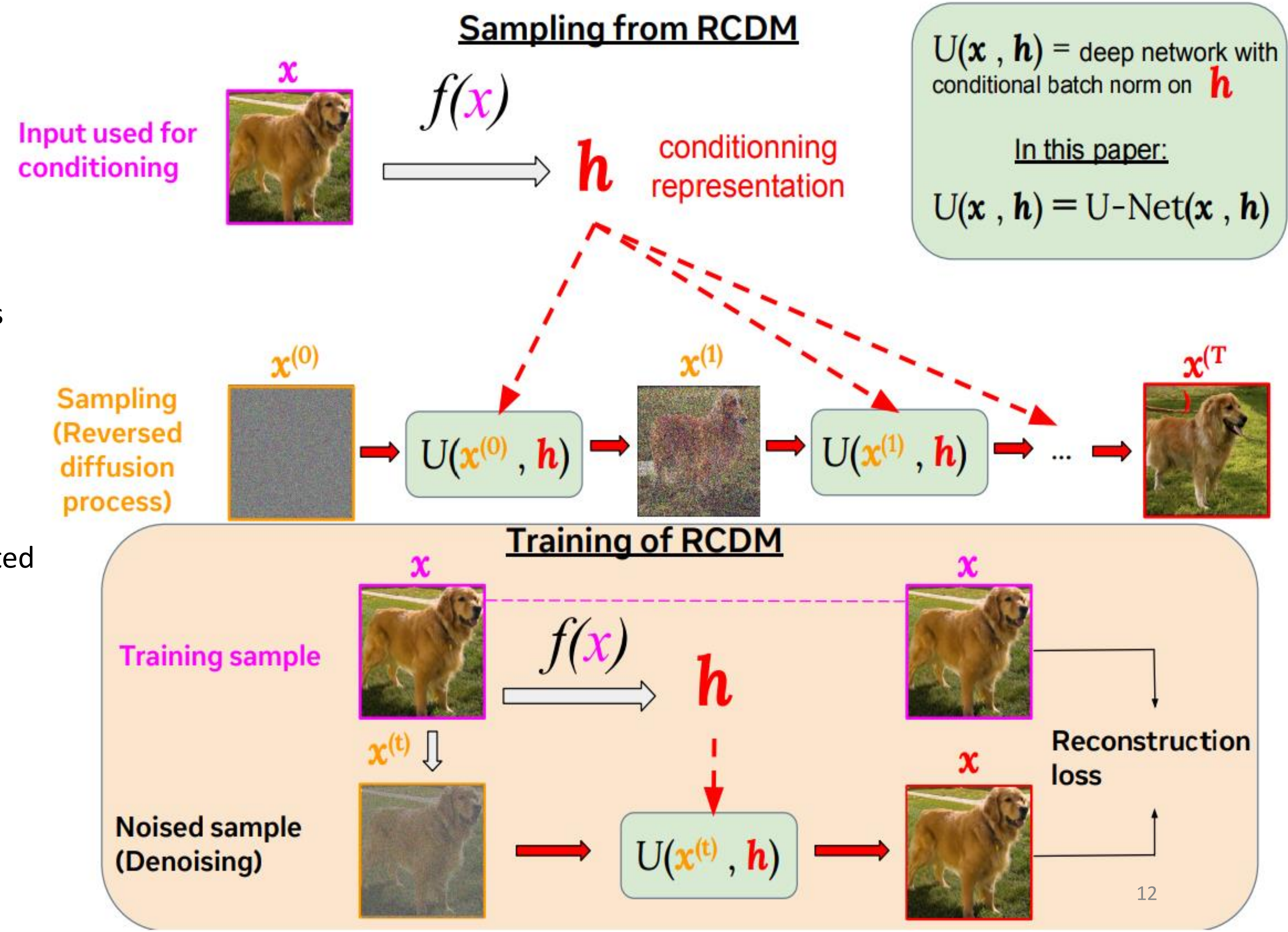
Generative methods

- Problems of gradient based methods:
  - Single output
  - Output may even lay out of images manifold if regularization is poor
- Generative methods in contrast:
  - Multiple outputs – images are generated from distribution
  - Output should be in images manifold



RCDM

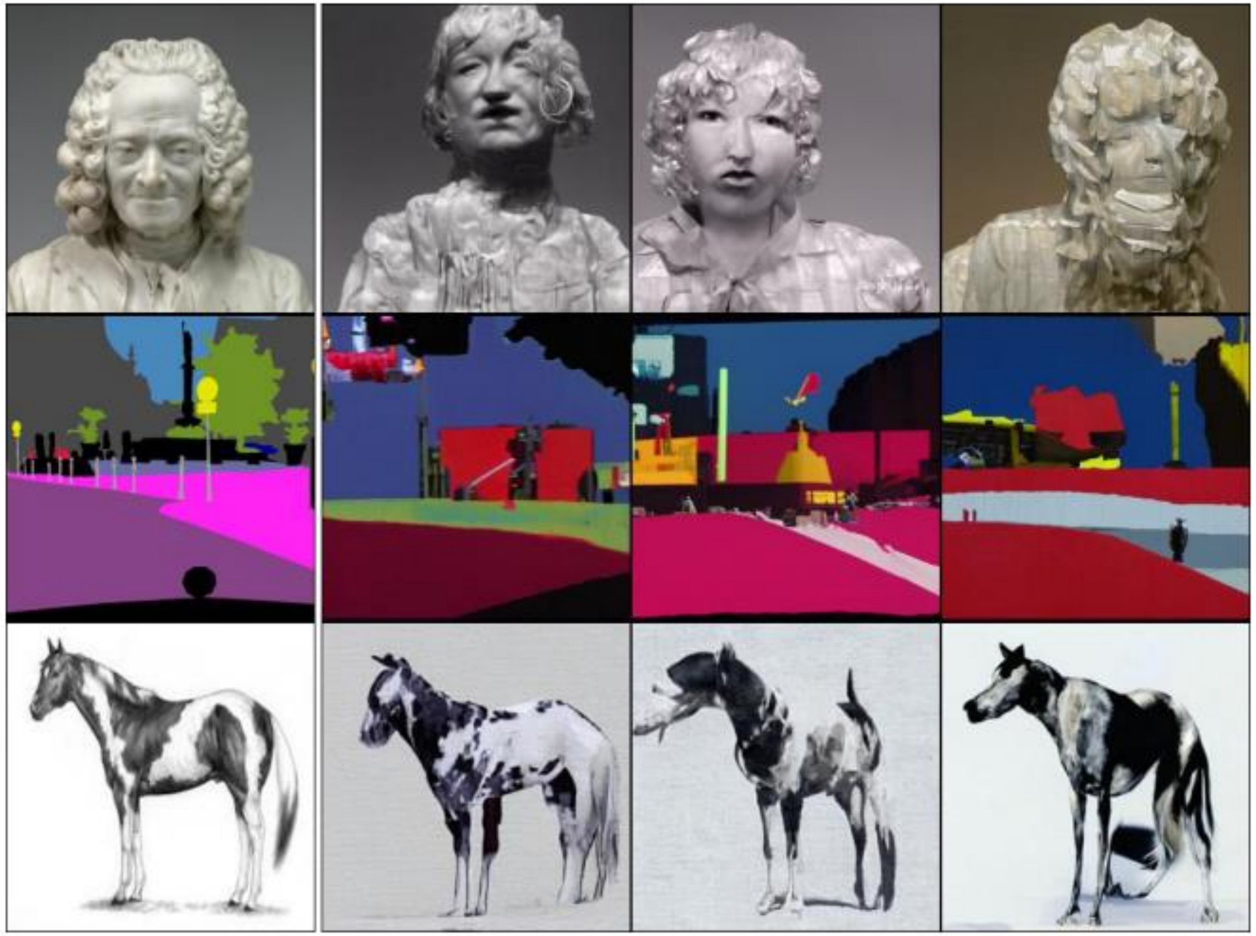
- Idea:
  Use a diffusion model to reconstruct realistic images from a representation
- Base architecture:
  Ablated Diffusion Model (ADM)
- Before conditioning representations are projected into 512 dim.



Generation examples (DINOv1 backbone)

In- distribution conditioning

Out of distribution (OOD) conditioning



Algebraic manipulations

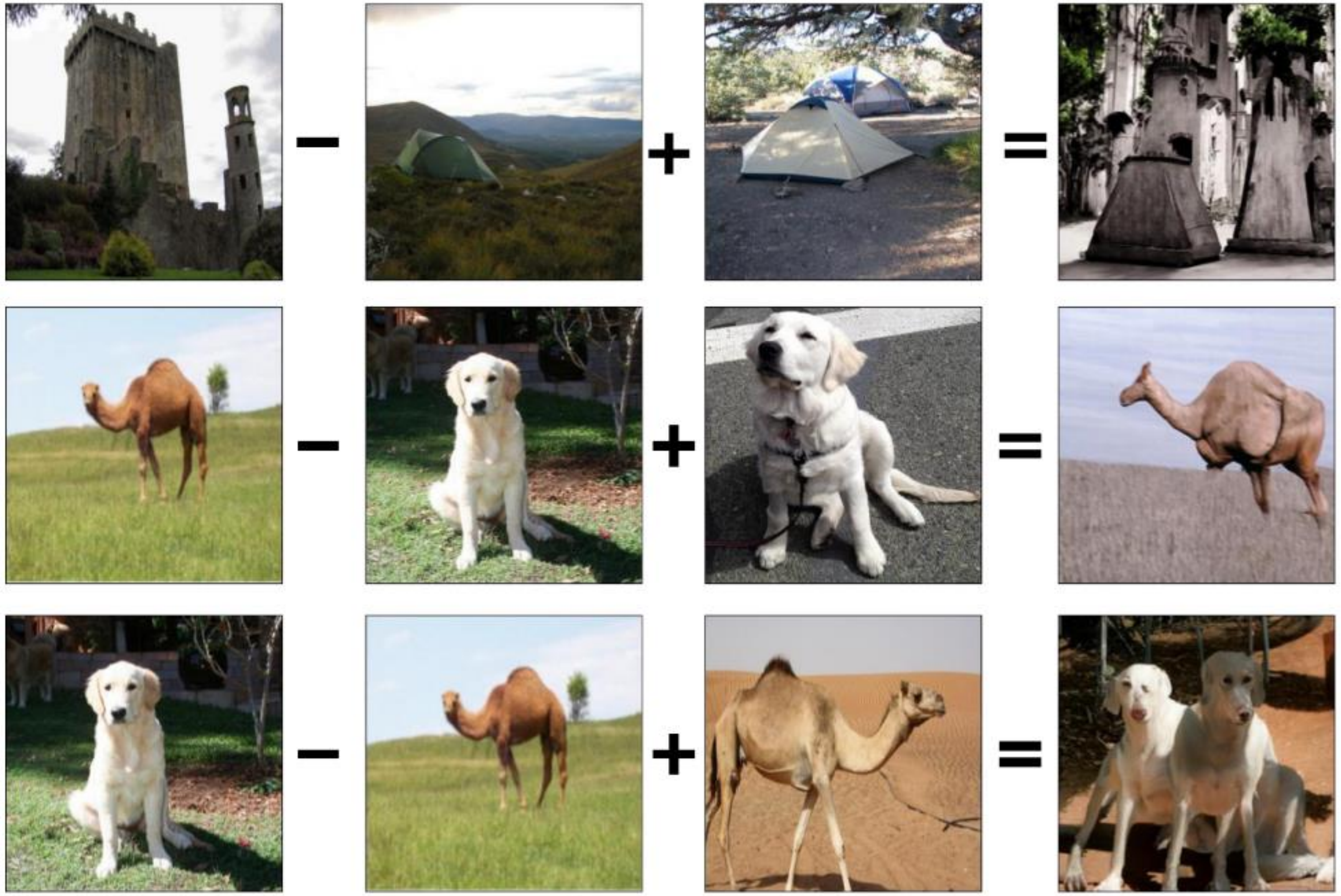
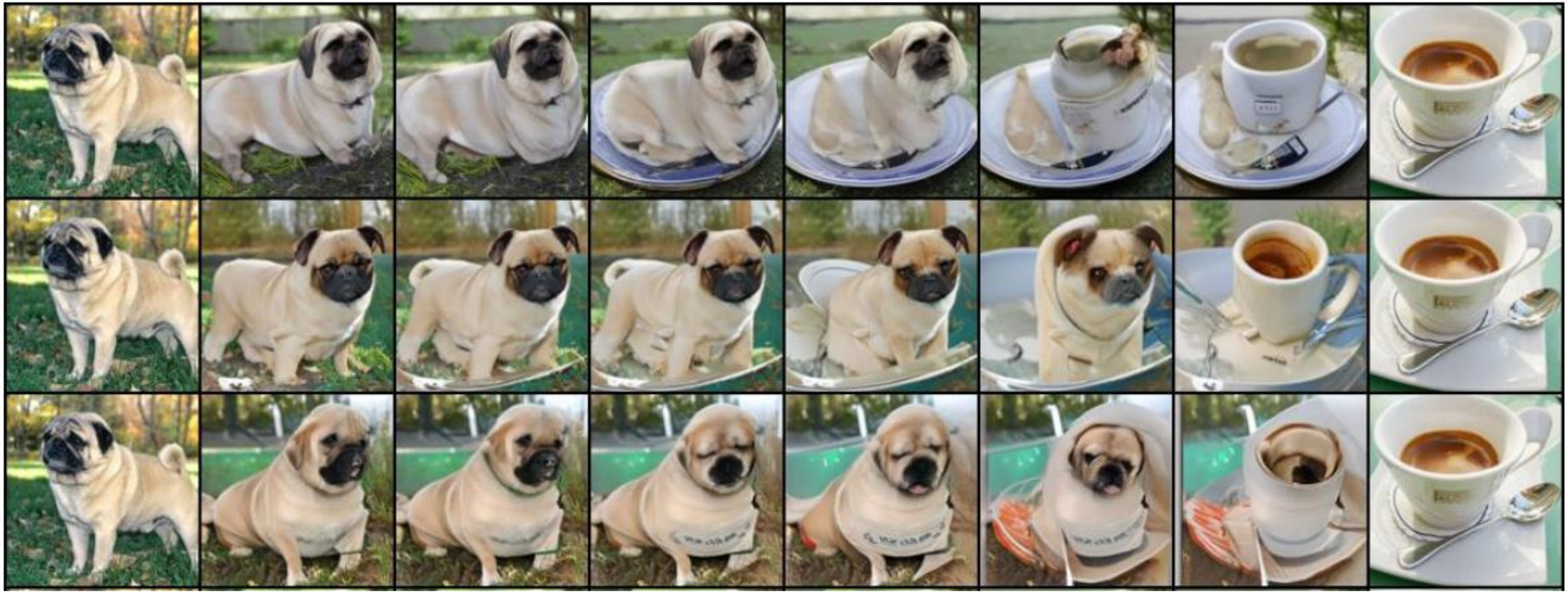


Figure 32: Algebraic manipulation of representations from real images (left-hand side of =) allows RCDM to generate new images with novel combination of factors. Here we use this technique with ImageNet images, to attempt background substitutions.

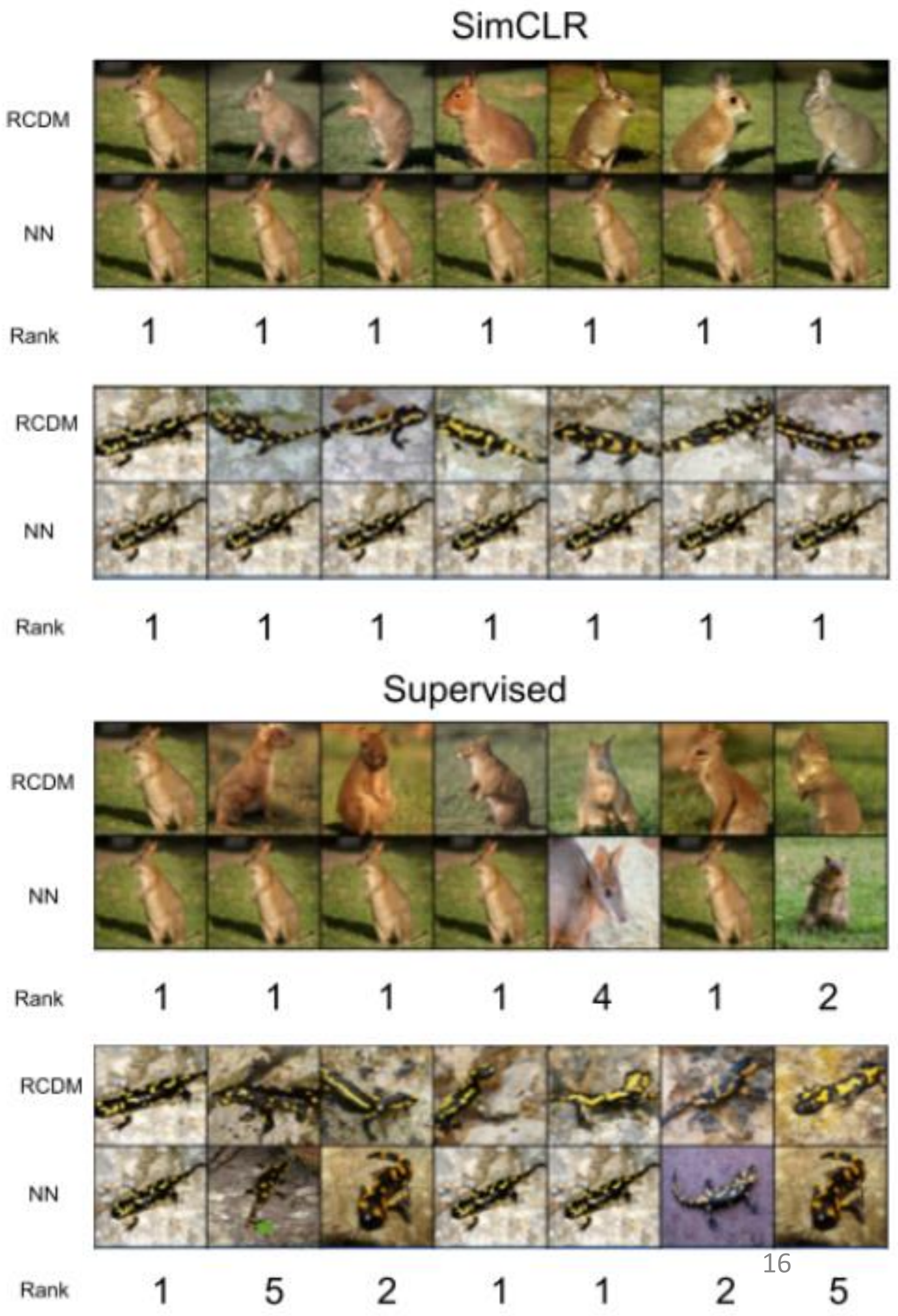
Interpolation between representations (DINOv1 backbone)



How close RCDM to conditioning?

Model	↓Mean rank	↑MRR
Dino (Caron et al., 2021)	1.00	0.99
Swav (Caron et al., 2020)	1.01	0.99
SimCLR (Chen et al., 2020)	1.16	0.97
Barlow T. (Zbontar et al., 2021))	1.00	0.99
Supervised	5.65	0.69

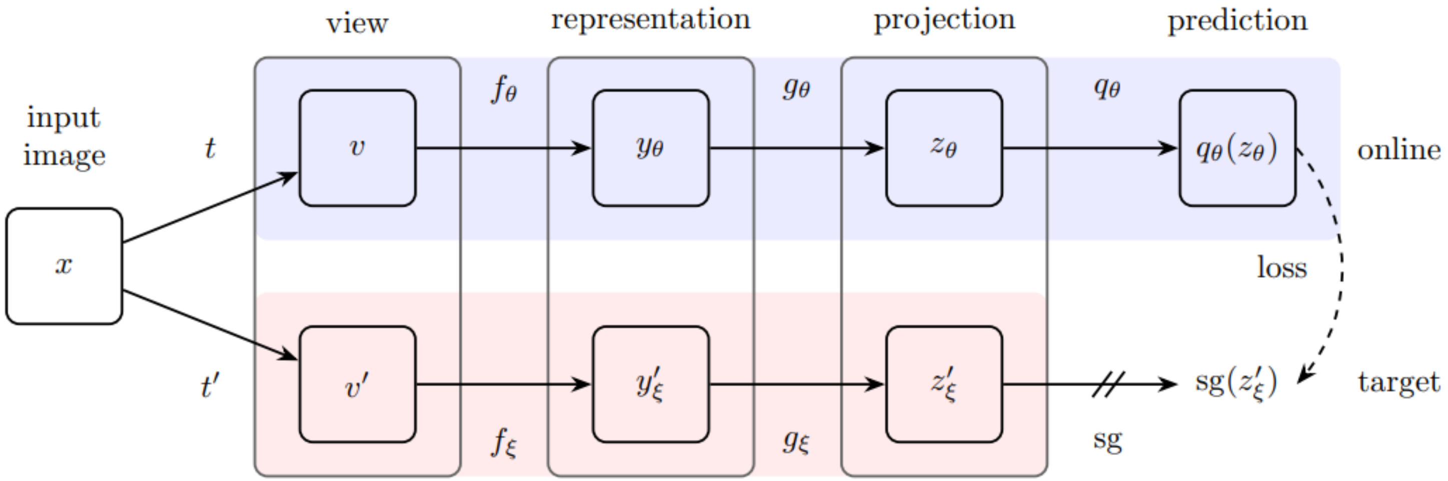
(b) For each encoder, we compute the rank and mean reciprocal rank (MRR) of the image used as conditioning within the closest set of neighbor in the representation space of the samples generated from the valid set (50K samples). A rank of one means that all of the generated samples for a given model have their representations matching the representation used as conditioning.



Representation/projection classification

Model	SimCLR Trunk	SimCLR Head	Dino Trunk	Dino Head	Barlow T. Trunk	Barlow T. Head	VicReg Trunk	VicReg Head
Val acc.	69.1 %	61.2 %	74.8 %	64.9 %	72.6 %	62.9 %	72.3 %	62.2 %

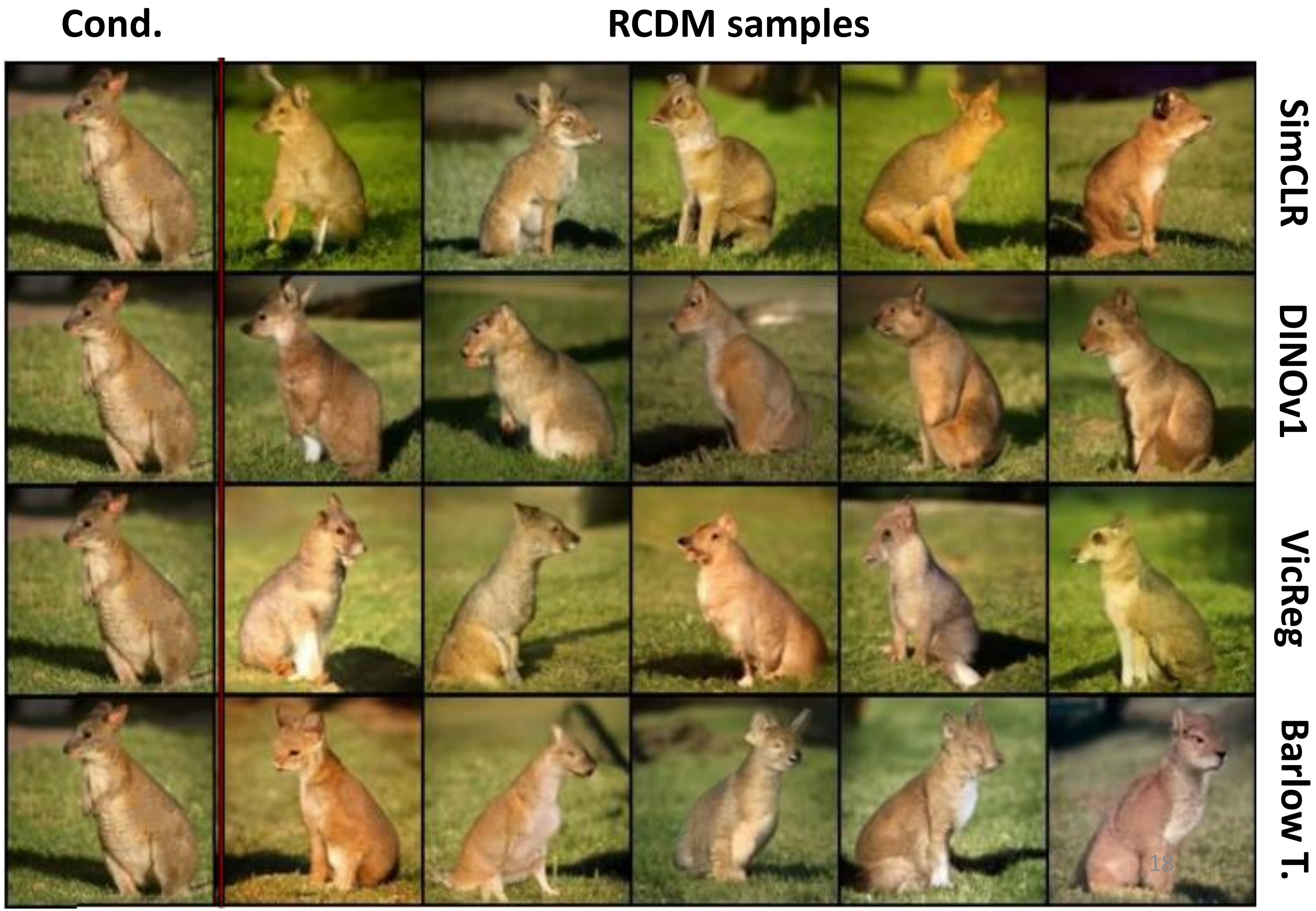
Table a): ImageNet linear probe validation accuracy on representation given by various SSL models. We observe an accuracy gap between the linear probes at the trunk level and the linear probes trained at the head level of around 10 percentage point of accuracy.



SSL backbone

- **Common/stable aspects** reveal what **is encoded** in the conditioning representation
- **Aspects that vary** show what is **not encoded**

- Backbone representations **do not allow much variance** in the generated samples
- A backbone representations preserve such information such as **pose and size of the animal, background, etc.**

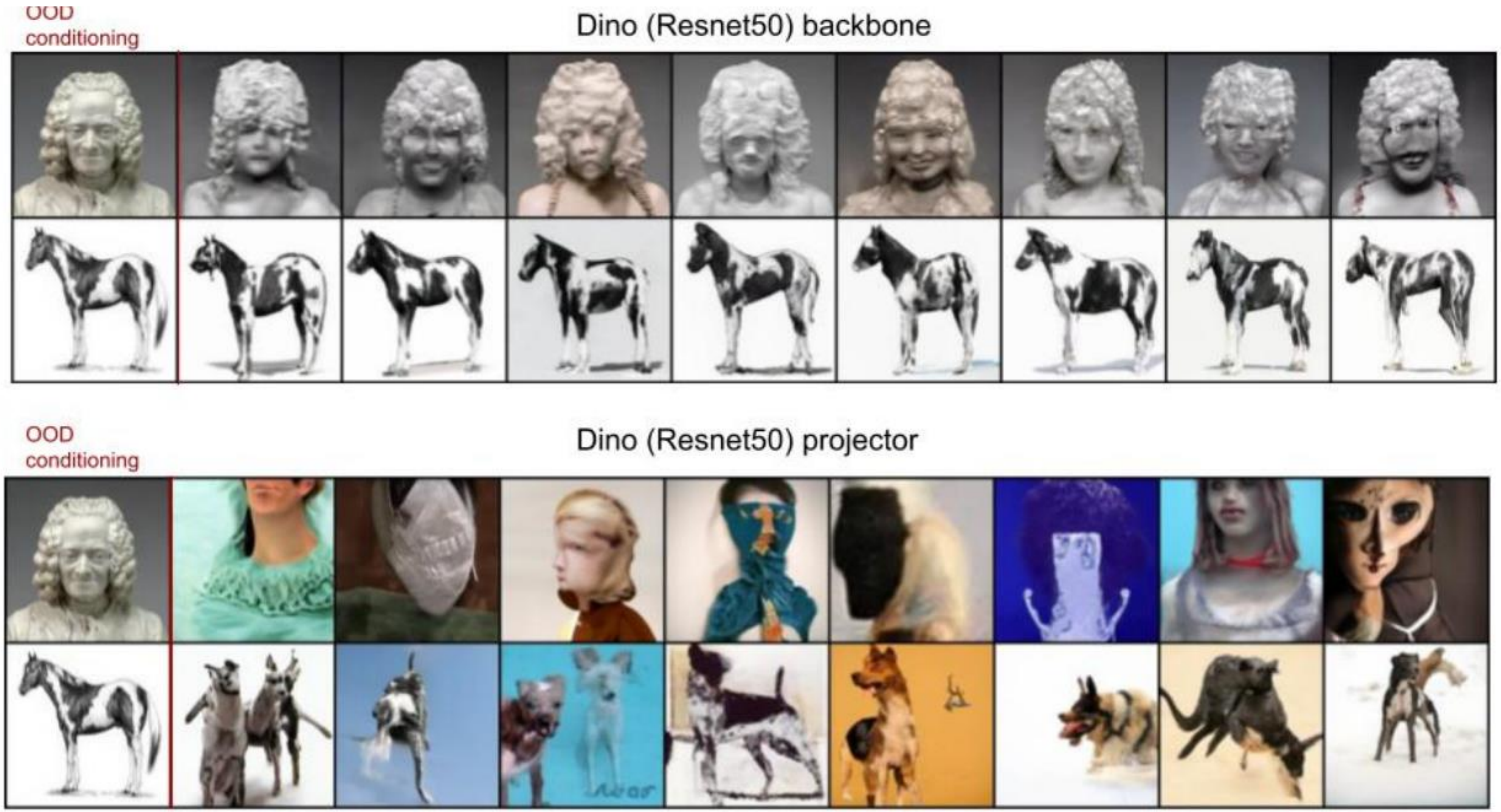


SSL projector

- **Common/stable aspects** reveal what **is encoded** in the conditioning representation
- **Aspects that vary** show what is **not encoded**

- Images sampled from **projector representations vary greatly,** which **indicates a significant loss of information**
- This indicates **that invariances in SSL models are mostly achieved in the projector representation, not the backbone**

Backbone vs projector (OOD)



SSL vs supervised

- **Common/stable aspects** reveal what **is encoded** in the conditioning representation
- **Aspects that vary** show what is **not encoded**

- We can see that supervised representations show more variance comparing to SSM backbone
- So, SSL backbone **representations are better for classifications since they contain more information about an input than the ones at the projector level**

Cond. **RCDM samples**

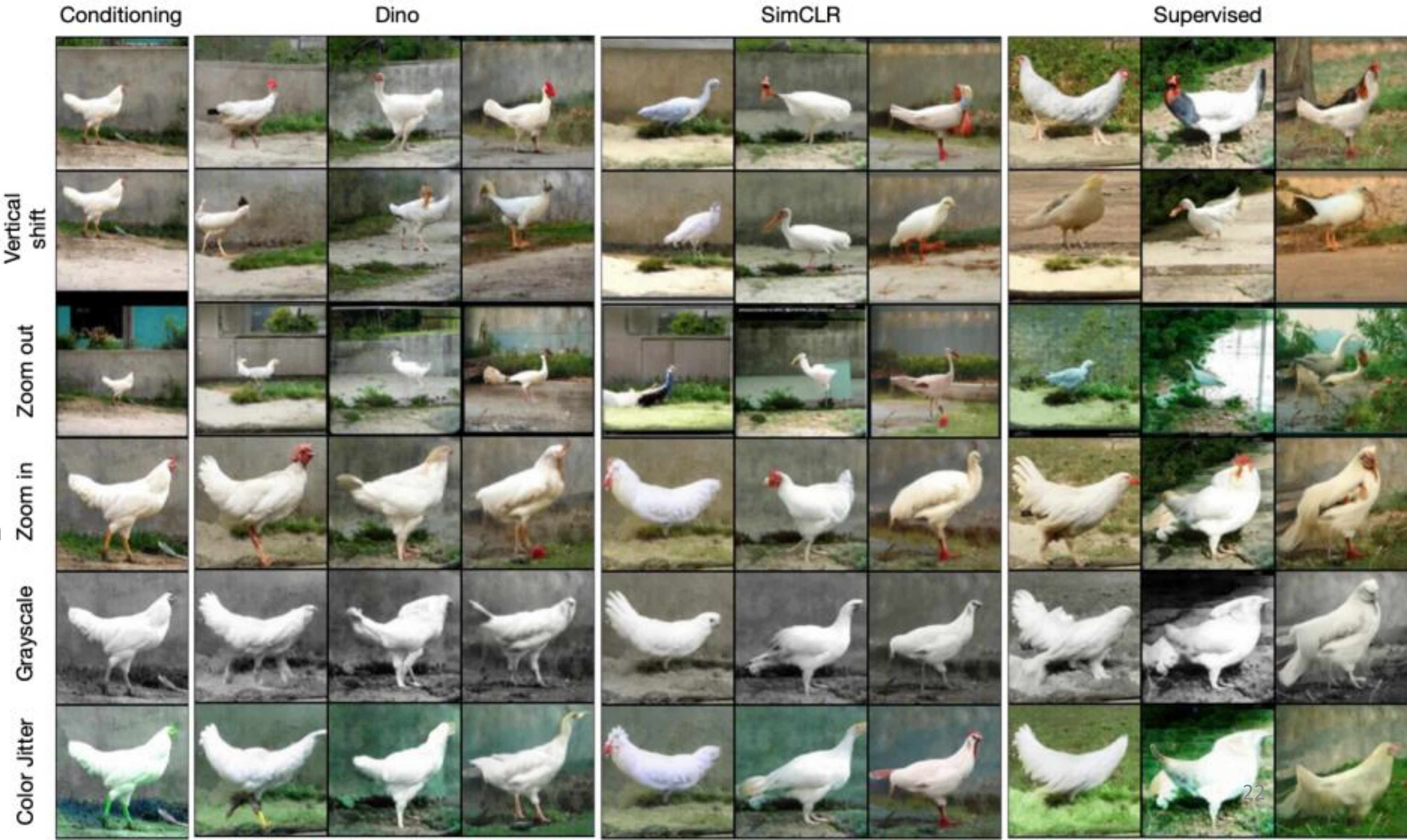
Supervised

DINOv1 Backbone

DINOv1 Projector

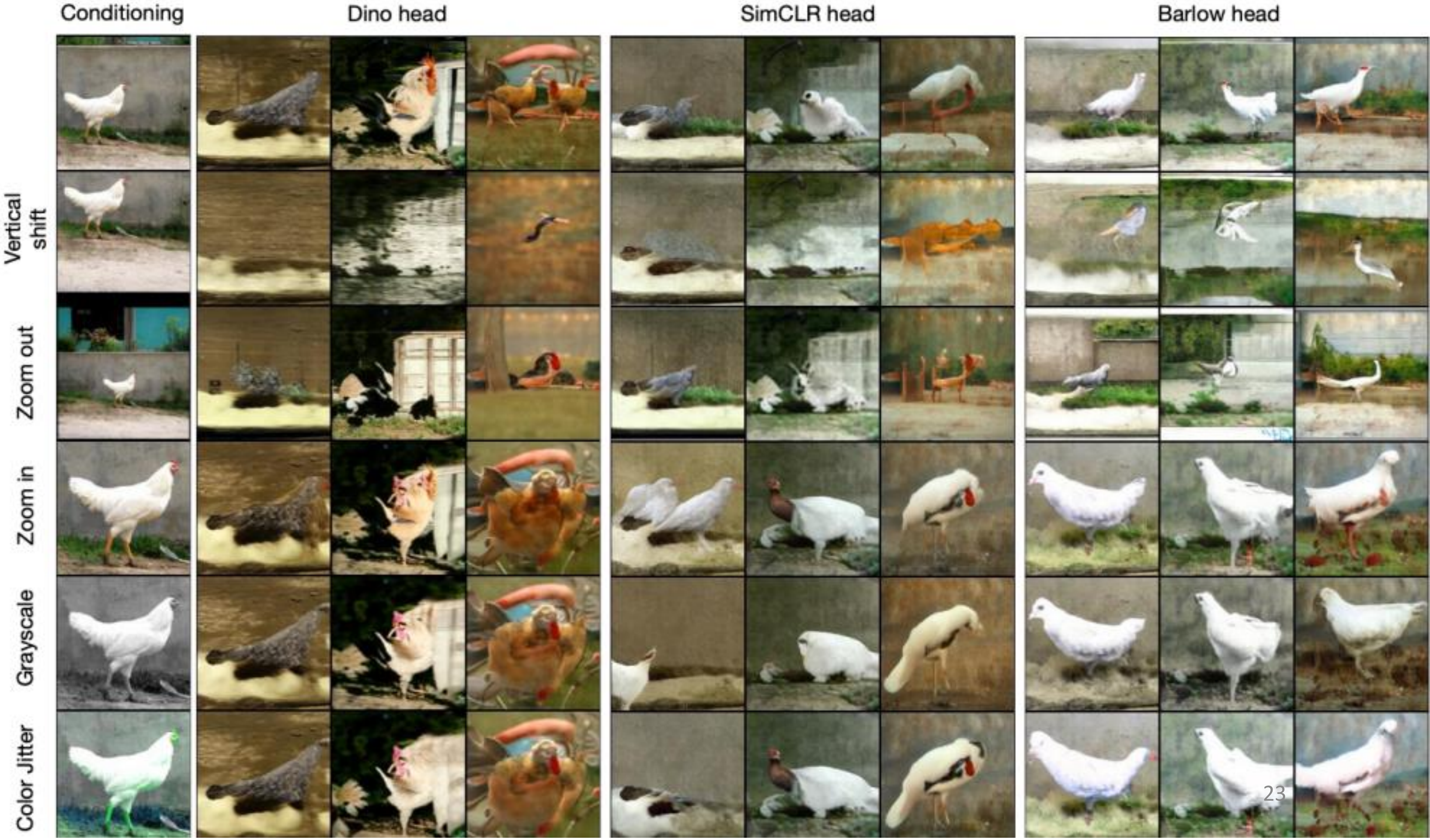
Augmentations and backbone

- SSL backbone representations do **retain information on object scale, grayscale status, and color palette of the background, much like the supervised representation**
- They do appear **invariant to vertical shifts**
- **Supervised** representation **constrain the appearance less**



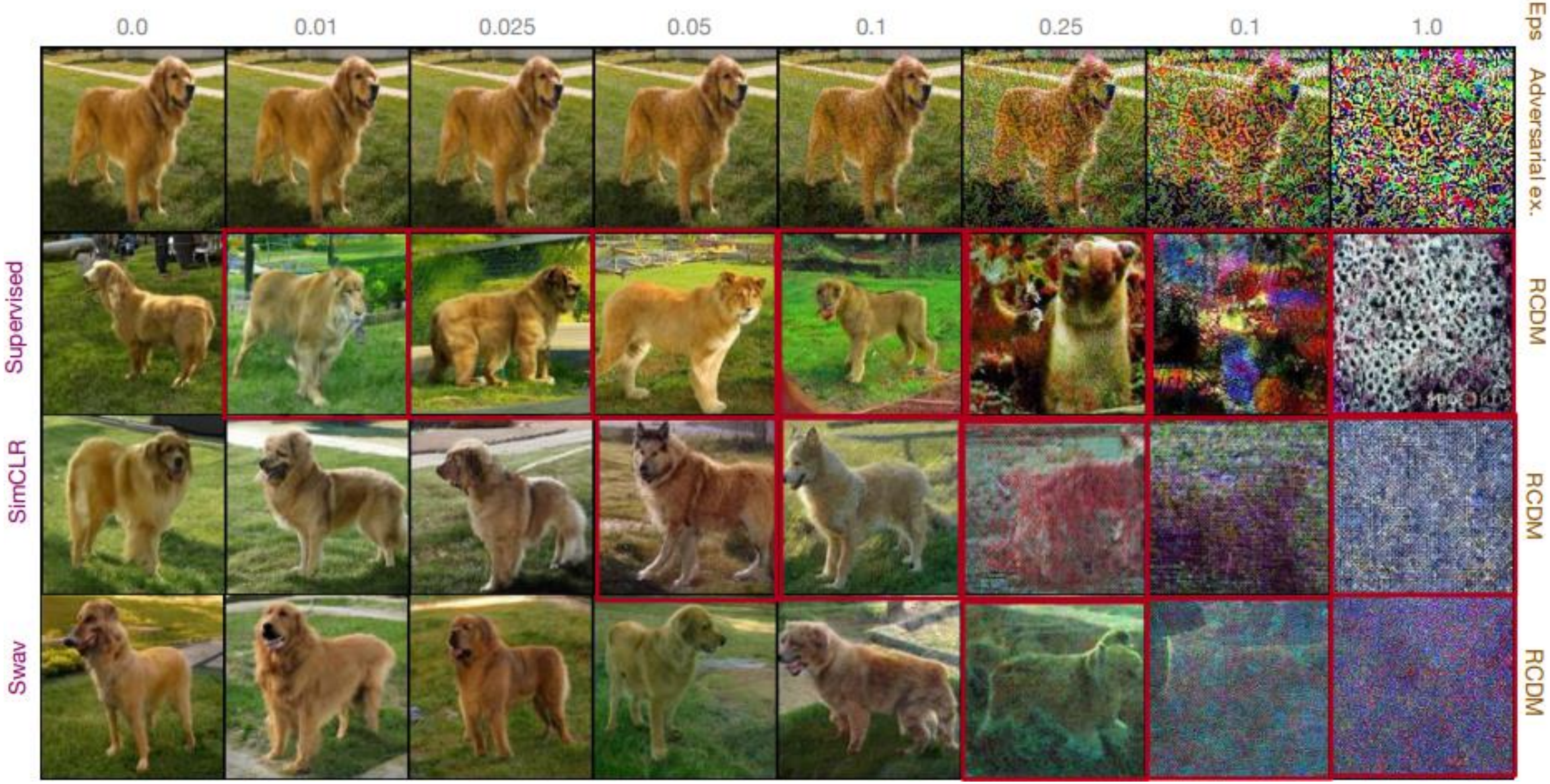
Augmentations and projector

- SSL projector representation seems to **encode object scale**
- But contrary to the backbone representation, it **does not encode grayscale-status and background color information**



Adversarial attacks

- Here RCDM conditioned on the representation of the adversarial examples to visualize if the generated images still belong to the class of the attacked image or not
- In this example attacks change the dog in the samples to a lion in the supervised setting whereas SSL methods doesn't seem to be impacted by the adversarial perturbations



Misclassified samples are in red boxes

SSM locally encode bg and fg on different dimensions

Algebraic manipulations

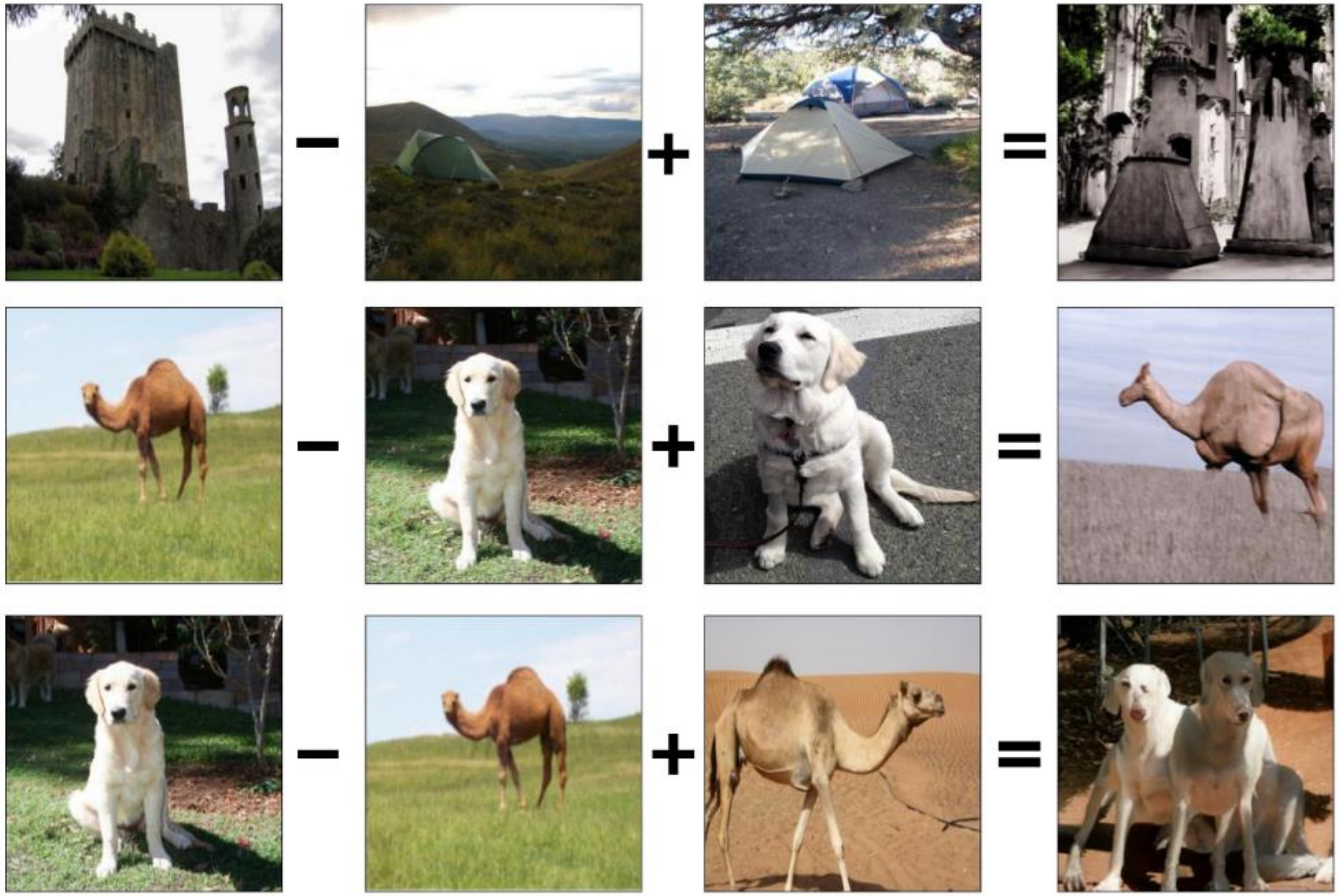


Figure 32: Algebraic manipulation of representations from real images (left-hand side of =) allows RCDM to generate new images with novel combination of factors. Here we use this technique with ImageNet images, to attempt background substitutions.

Conclusions

- **Pros:**
  - Nice method to visualize information that SSL representation (or other representations) encode
  - SSL backbone used in downstream tasks **are not invariant to data augmentations**!
  - SSL backbone **encode information about object, background, color, geometry**
  - SSL projectors discard this information, leading to poorer results in downstream tasks
  - SSL backbone **are more robust to adversarial attacks**
  - Supervised representations constrain the samples appearance much less than SSL backbone
- **Coins:**
  - **A lot of training** – need to train RCDM for every representations!
  - **Experiments results only on ResNet-50**
  - **No experiments for varying backbone/projector dimensions outputs**

Literature

- Self-supervised learning (SSL): BYOL; DINO (v1,2); iBOT; SSL models distillation
- High Fidelity Visualization of What Your Self-Supervised Representation Knows About
- What makes instance discrimination good for transfer learning?
- Deep Image Prior