b
UNIVERSITÄT
BERN

Perspectives on Unsupervised Representation Learning

Paolo Favaro

Computer Vision Group — University of Bern

UNIVERSITÄT
BERN

Perspectives on Unsupervised Representation Learning

Paolo Favaro

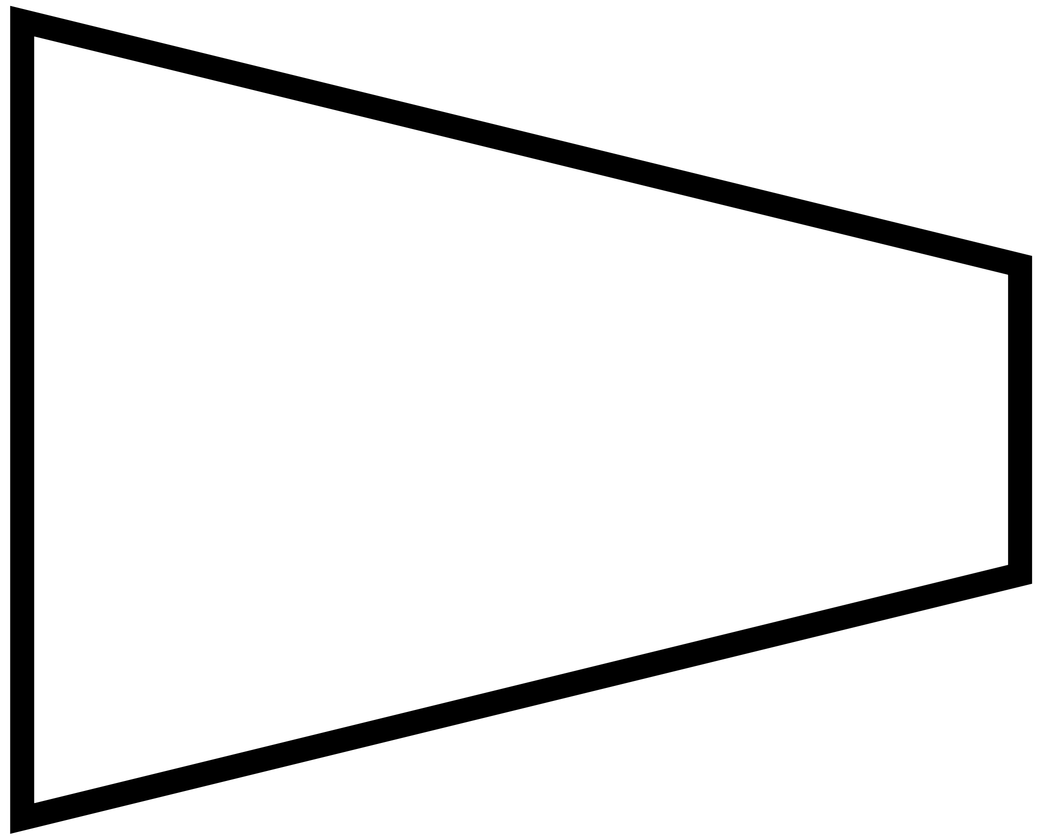
Computer Vision Group — University of Bern

Supervised Learning

data

$\boxed{x}$

neural network



task

$\boxed{p(y|x)}$

attributes

book

car

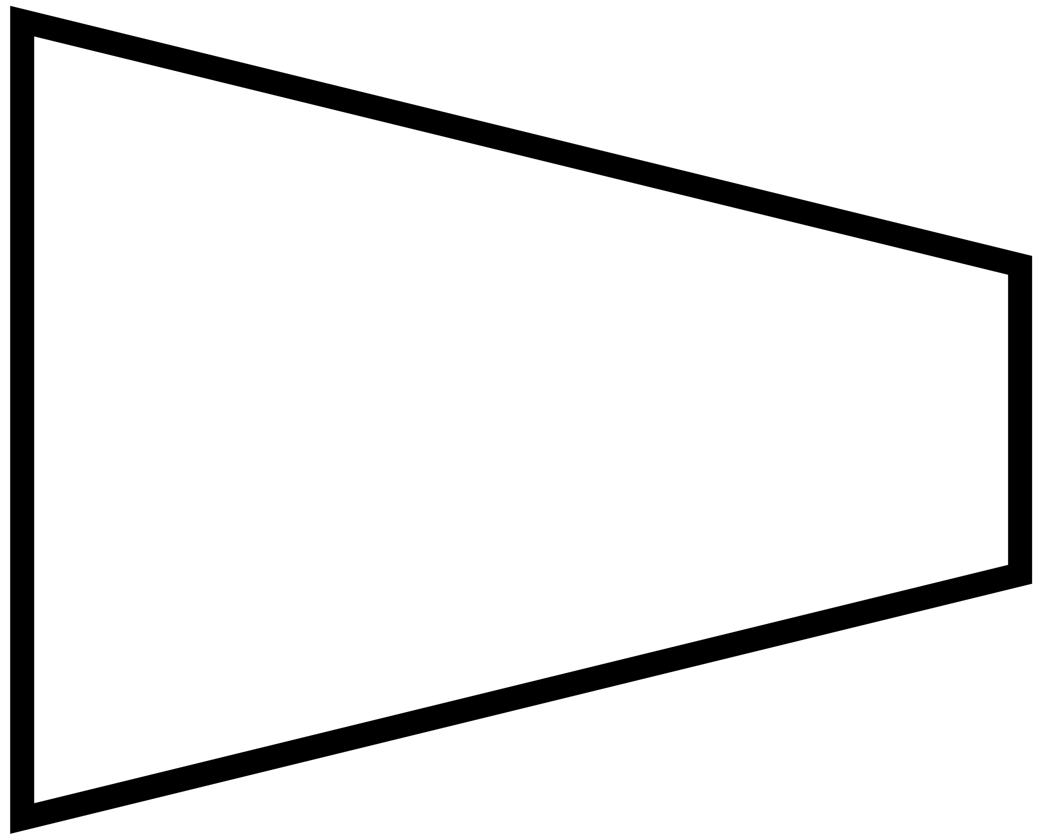
$\boxed{y}$ **dog**

shield

Supervised Learning

data | neural network | task | attributes

$\boxed{x}$



$\boxed{p(y \mid x)}$

$\boxed{y}$

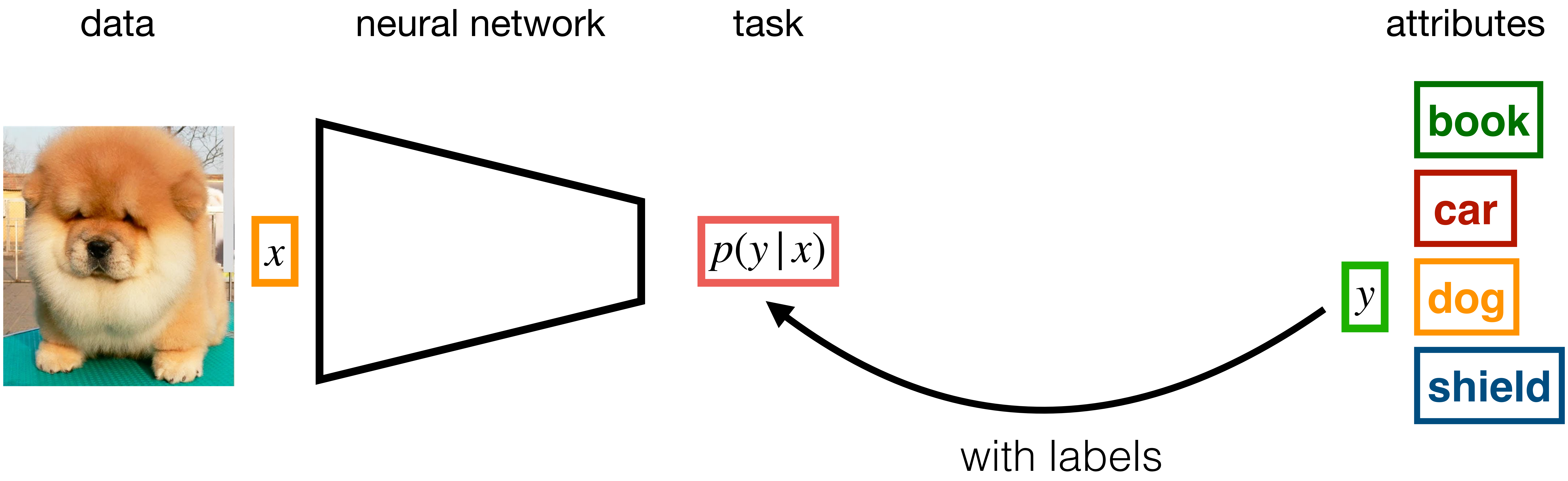
book

car

dog

shield

Supervised Learning



Unsupervised Representation Learning

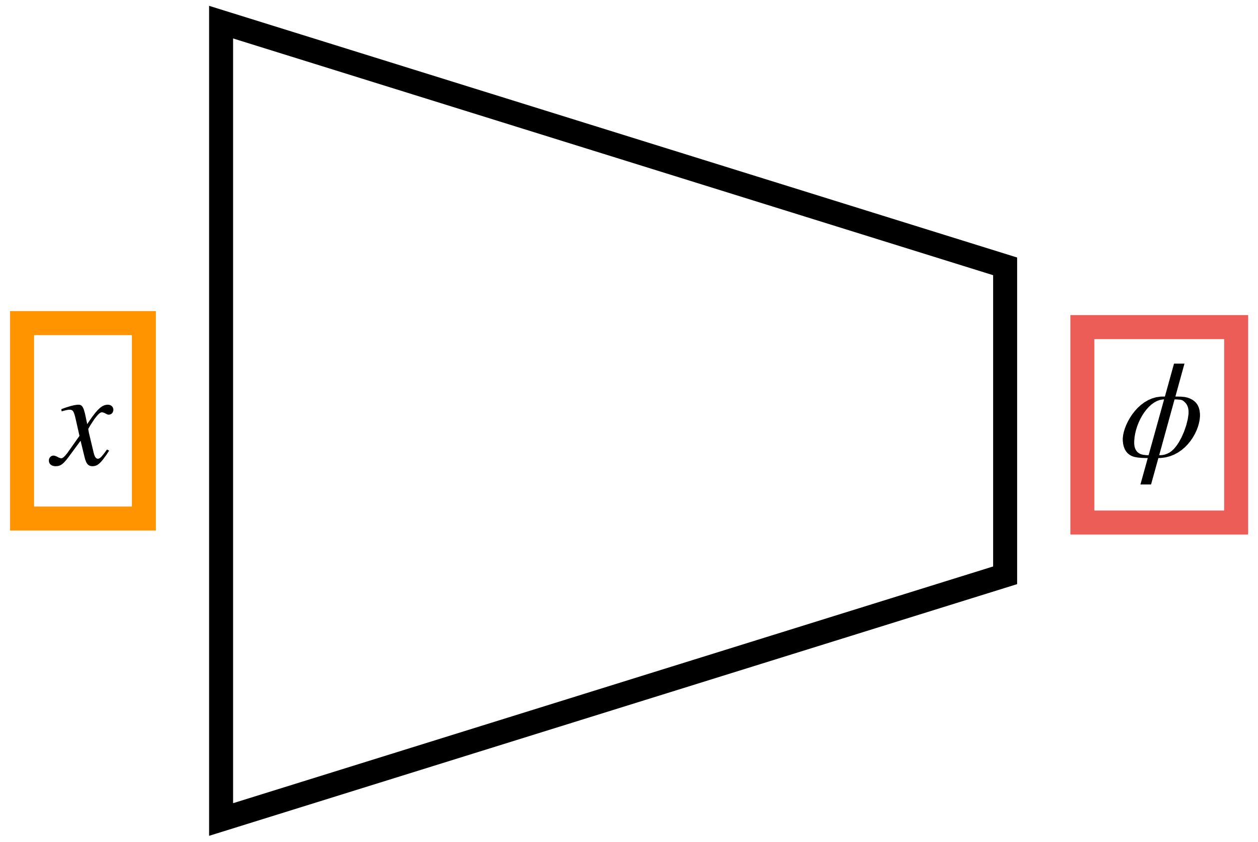
data

neural network

pretext-task

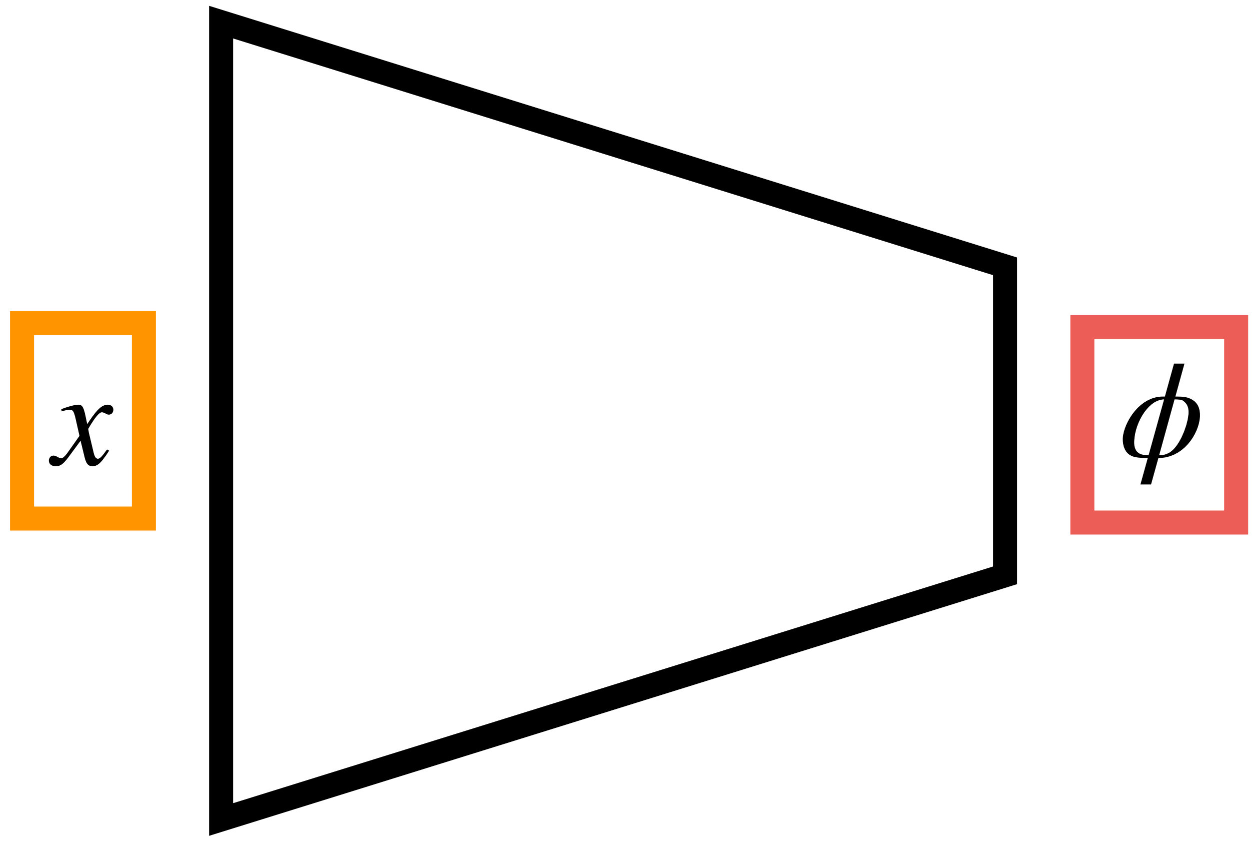
neural network

attributes



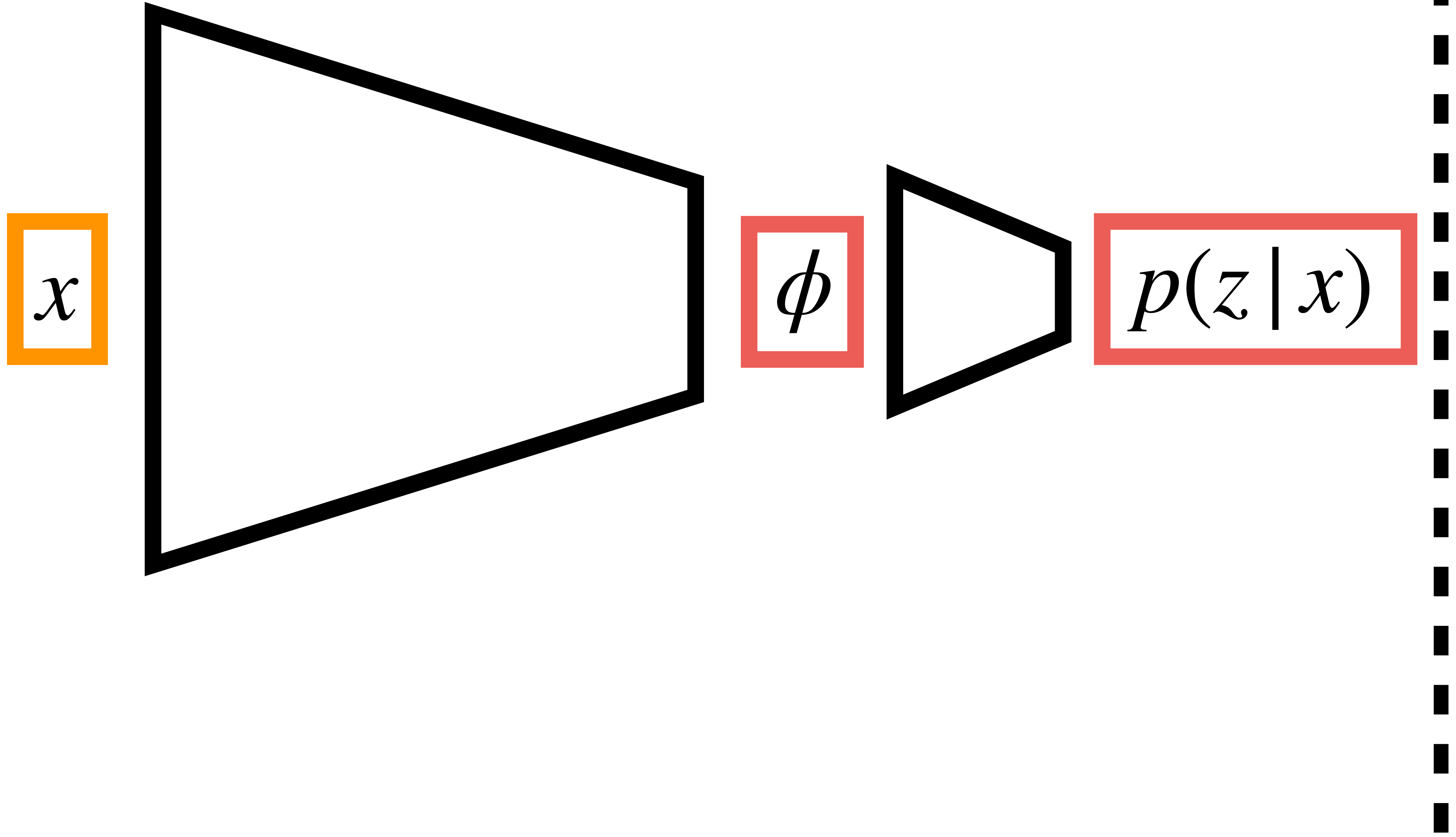
Unsupervised Representation Learning

data | neural network | pretext-task | neural network | attributes

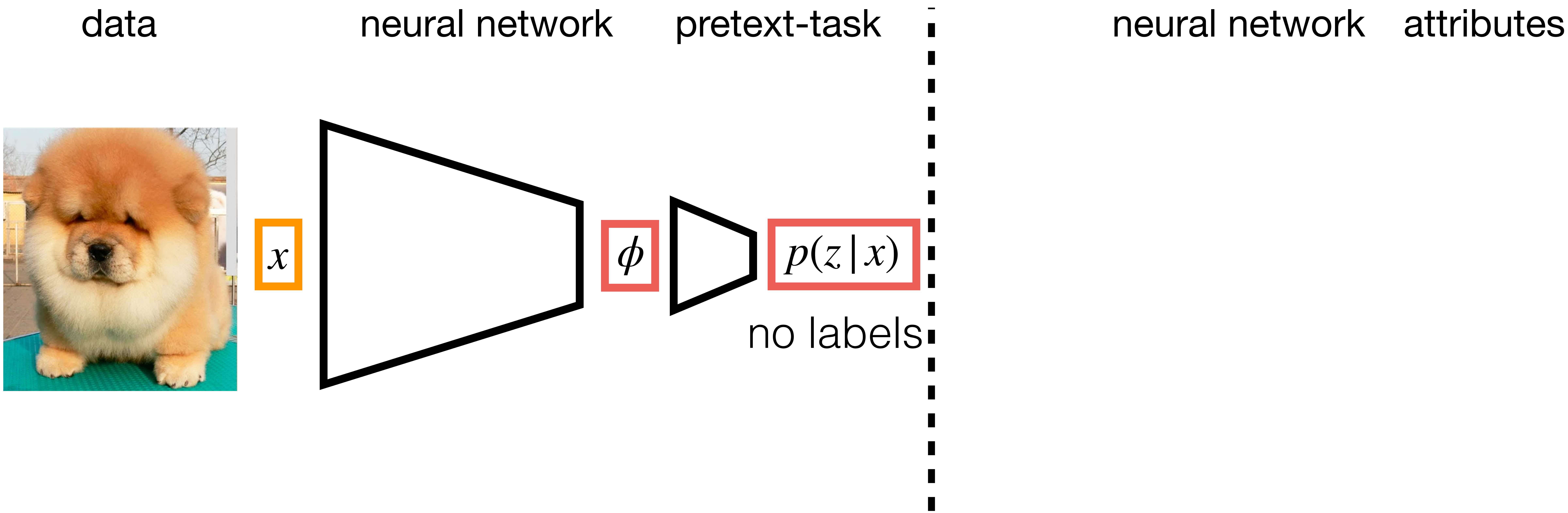


Unsupervised Representation Learning

data neural network pretext-task neural network attributes

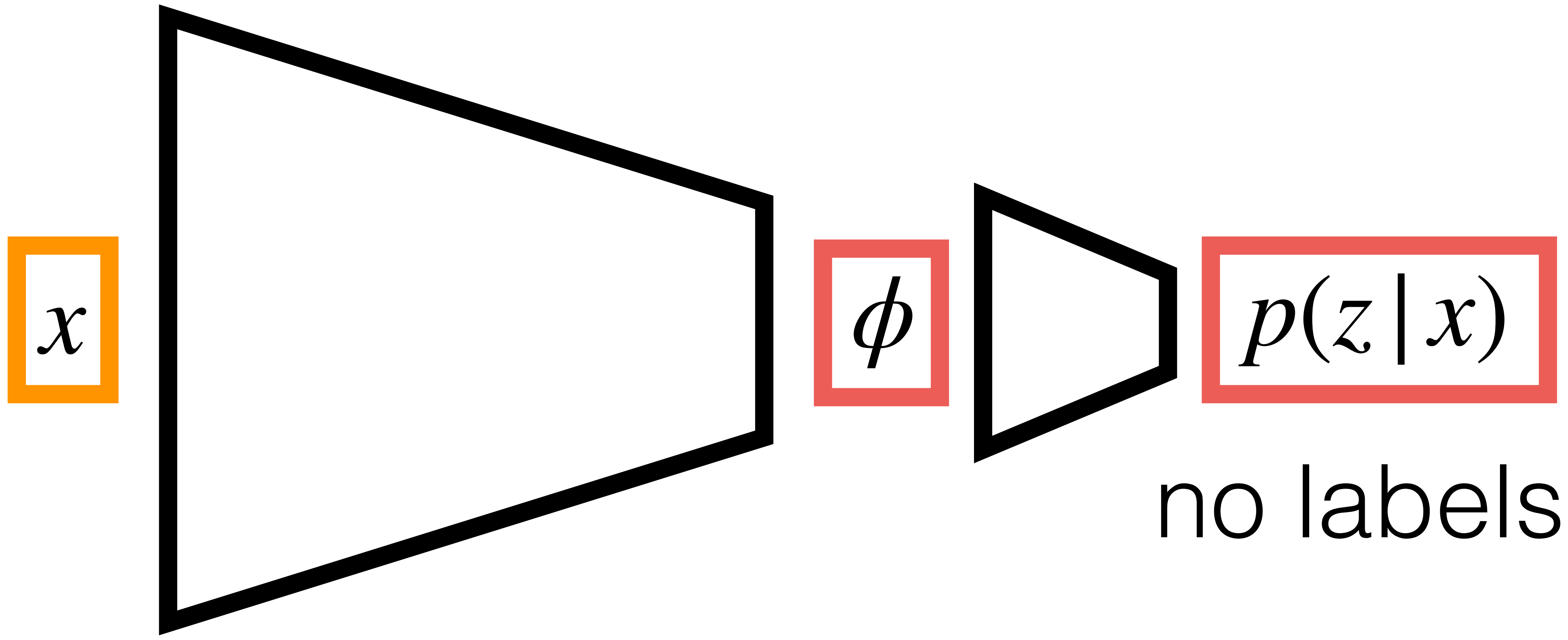


Unsupervised Representation Learning



Unsupervised Representation Learning

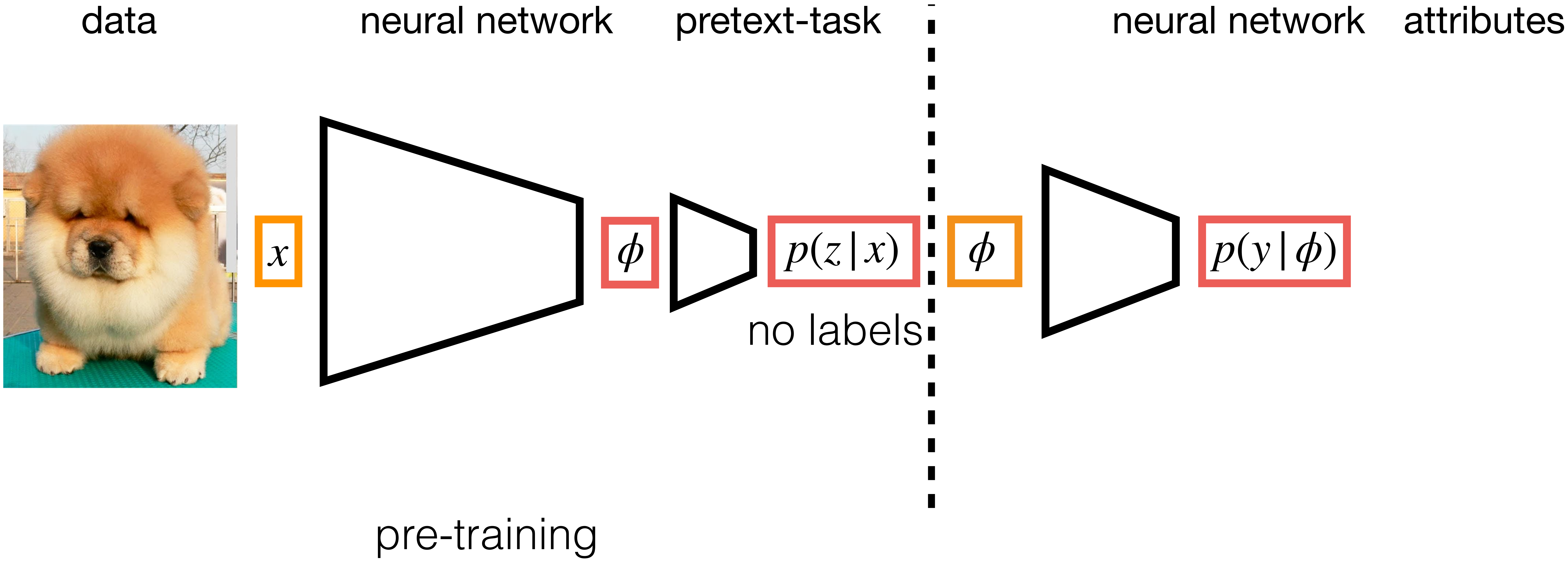
data | neural network | pretext-task | neural network | attributes



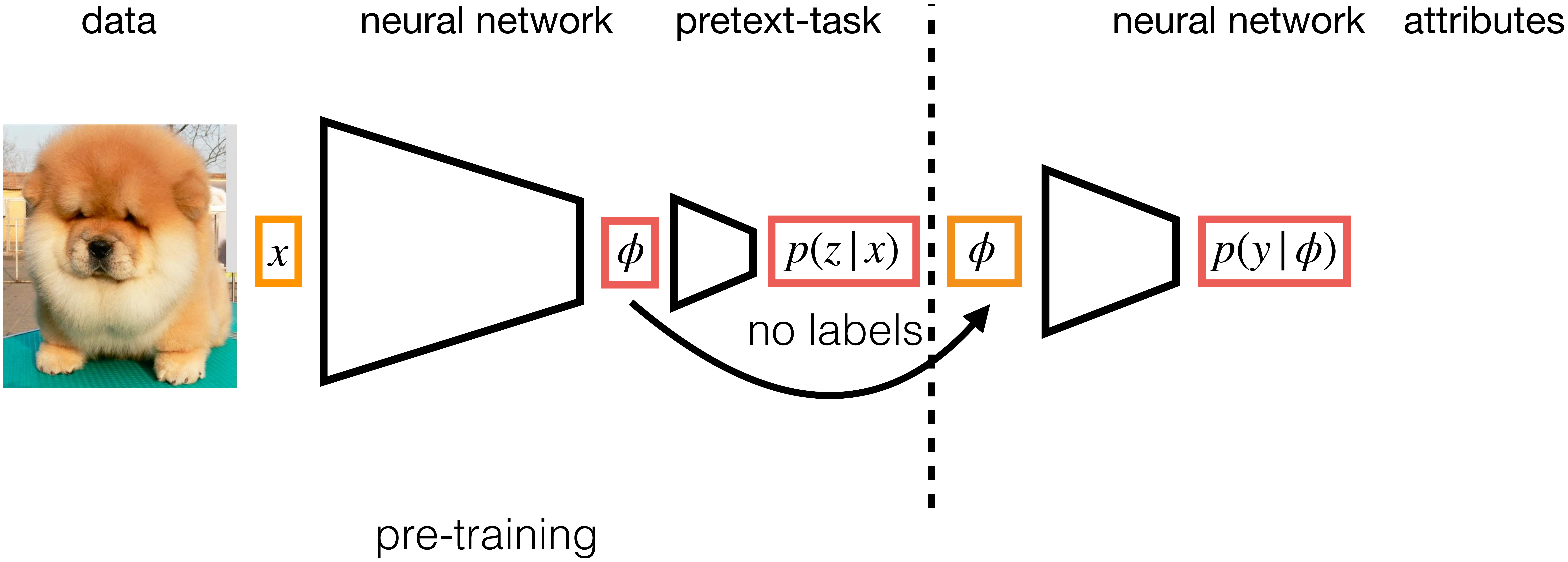
no labels

pre-training

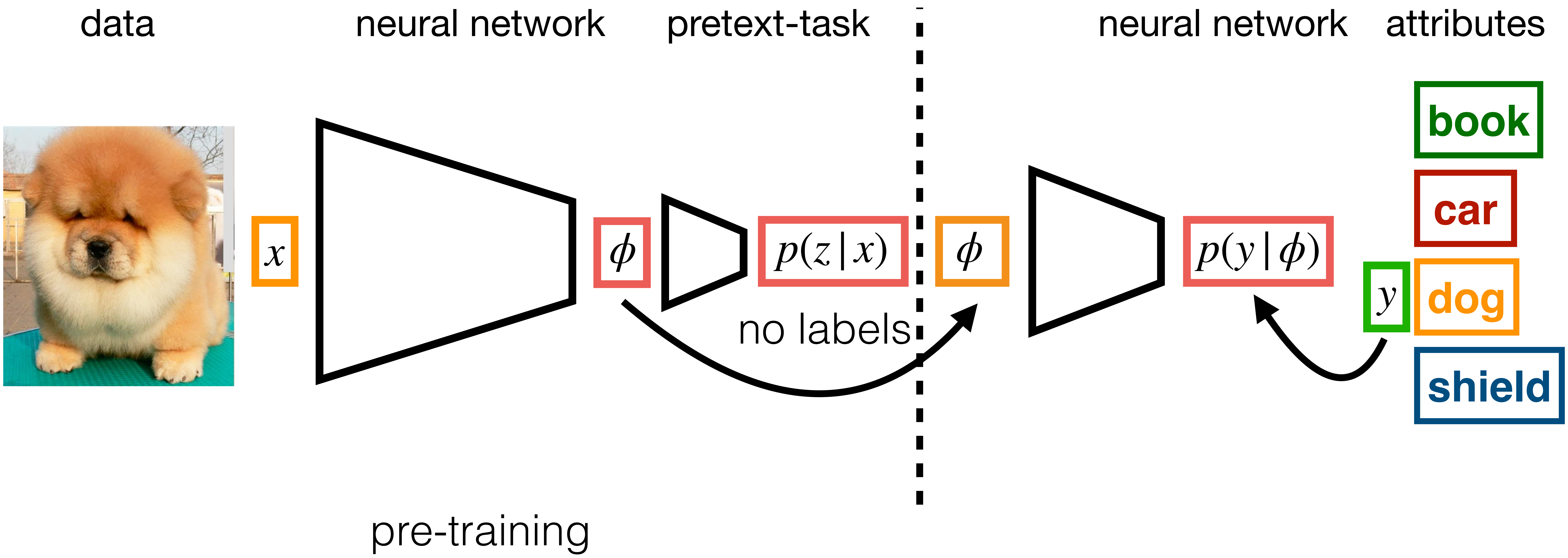
Unsupervised Representation Learning



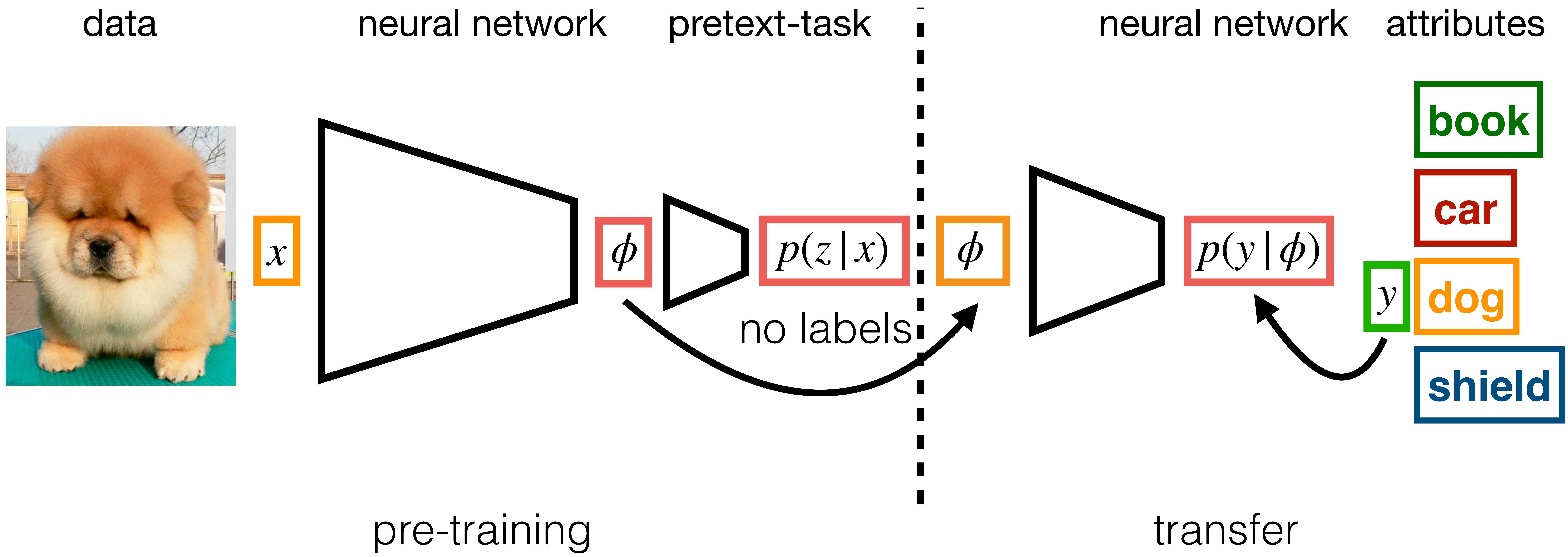
Unsupervised Representation Learning



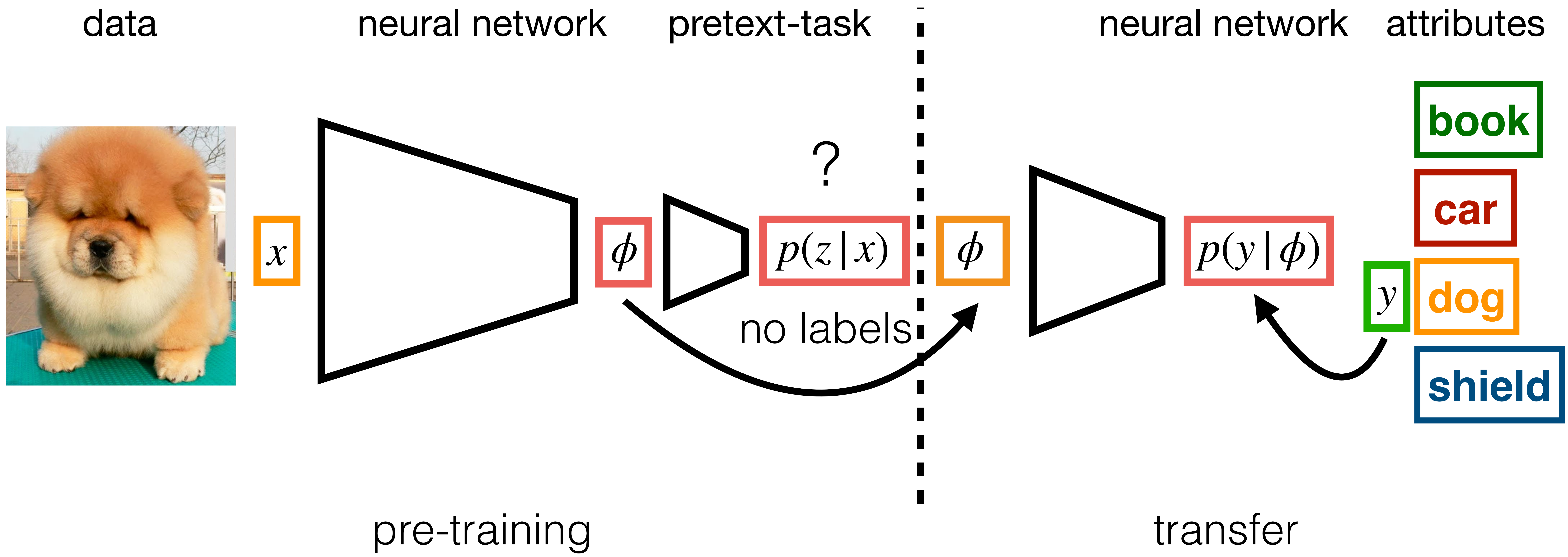
Unsupervised Representation Learning



Unsupervised Representation Learning



Unsupervised Representation Learning

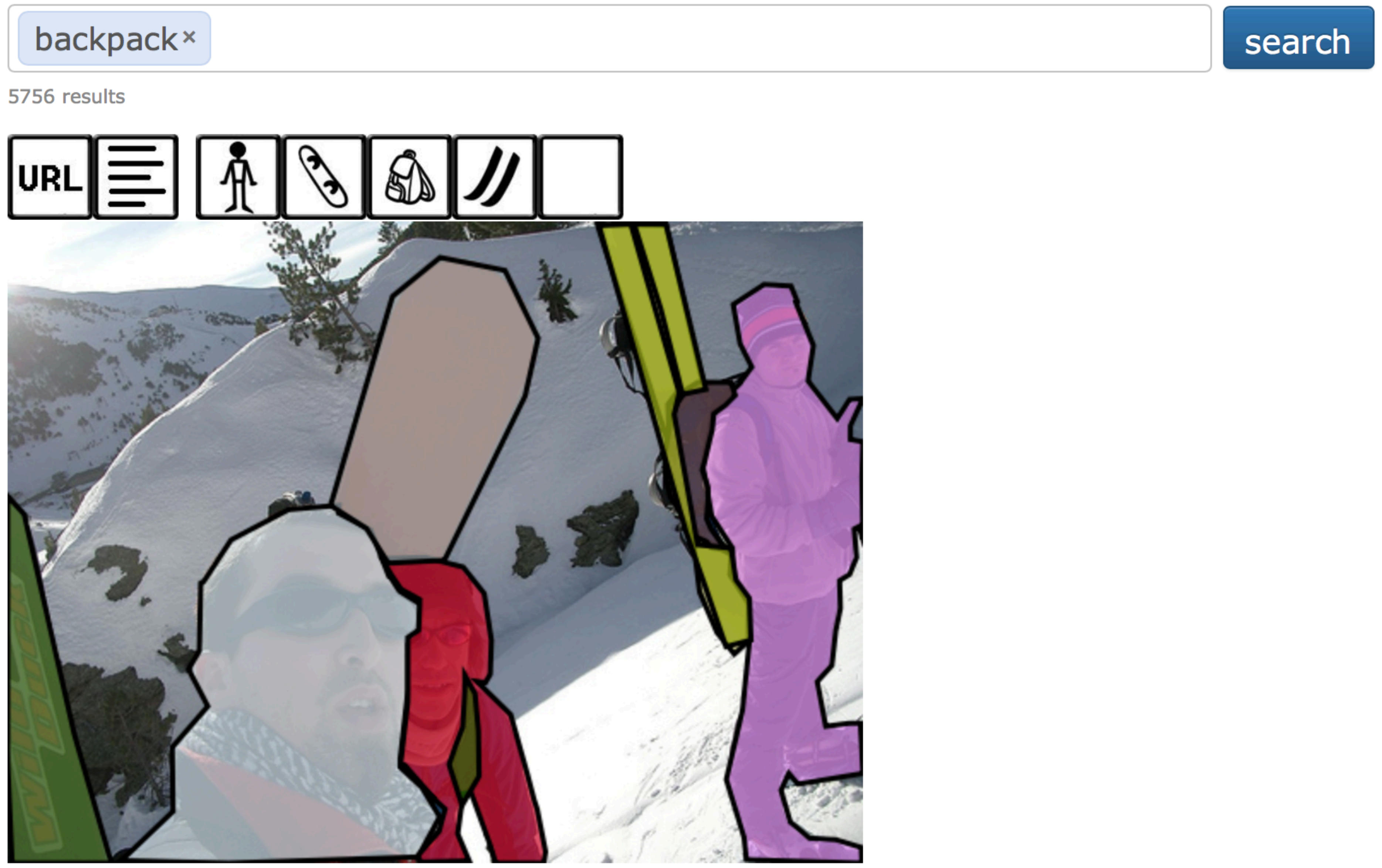


Why Unsupervised?

- Data privacy friendly
- Cheaper
- Scales better
- Labels can be unclear, noisy, limited
- Animal learning is often unsupervised

Why Unsupervised?

- Data privacy friendly
- Cheaper
- Scales better
- Labels can be unclear, noisy, limited
- Animal learning is often unsupervised



Supervision in Unsupervised Learning

- What can we learn from a random (though arbitrarily large) set of images?

Supervision in Unsupervised Learning

- What can we learn from a random (though arbitrarily large) set of images?

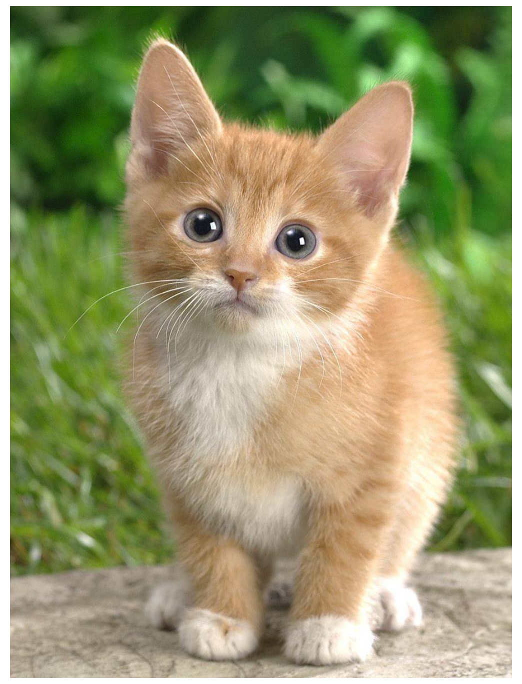
Supervision in Unsupervised Learning

- What can we learn from a random (though arbitrarily large) set of images?
  - Not much unless we make some **assumptions**

Supervision in Unsupervised Learning

- What can we learn from a random (though arbitrarily large) set of images?
  - Not much unless we make some **assumptions**
- Shifting supervision from per-sample labeling to specifying data properties

dog



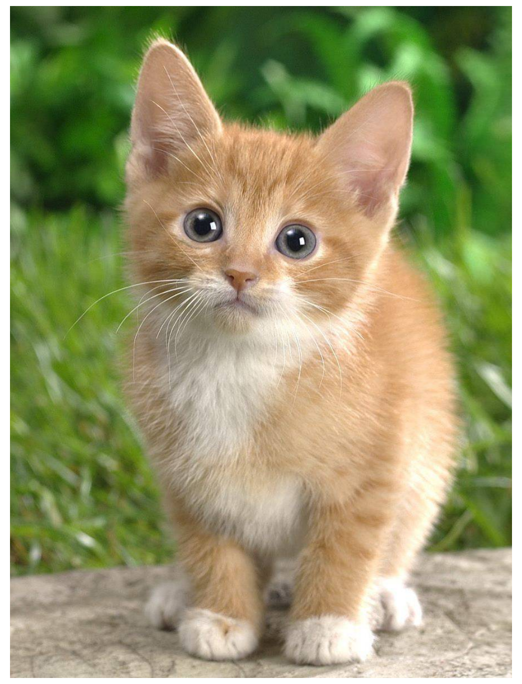
cat

horse

Supervision in Unsupervised Learning

- What can we learn from a random (though arbitrarily large) set of images?
  - Not much unless we make some **assumptions**
- Shifting supervision from per-sample labeling to specifying data properties

dog



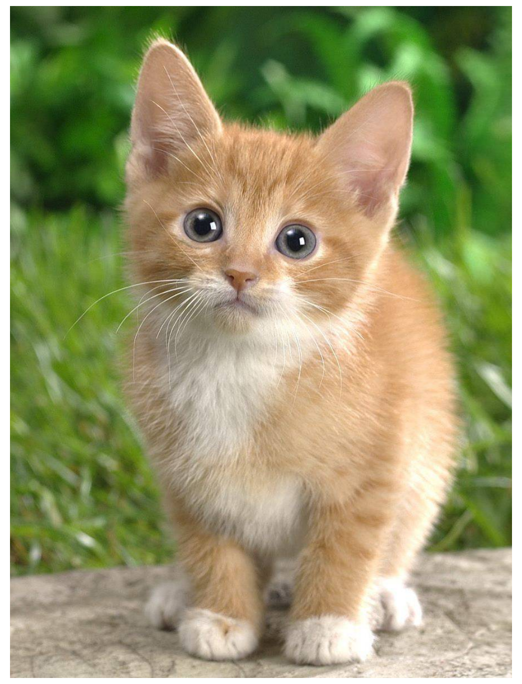
cat

horse

Supervision in Unsupervised Learning

- What can we learn from a random (though arbitrarily large) set of images?
    - Not much unless we make some **assumptions**
- Shifting supervision from per-sample labeling to specifying data properties

dog



cat

horse

similarity of data-augmented images

Supervision in Unsupervised Learning

- What can we learn from a random (though arbitrarily large) set of images?
  - Not much unless we make some **assumptions**
- Shifting supervision from per-sample labeling to specifying data properties

Supervision in Unsupervised Learning

- What can we learn from a random (though arbitrarily large) set of images?
  - Not much unless we make some **assumptions**
- Shifting supervision from per-sample labeling to specifying data properties

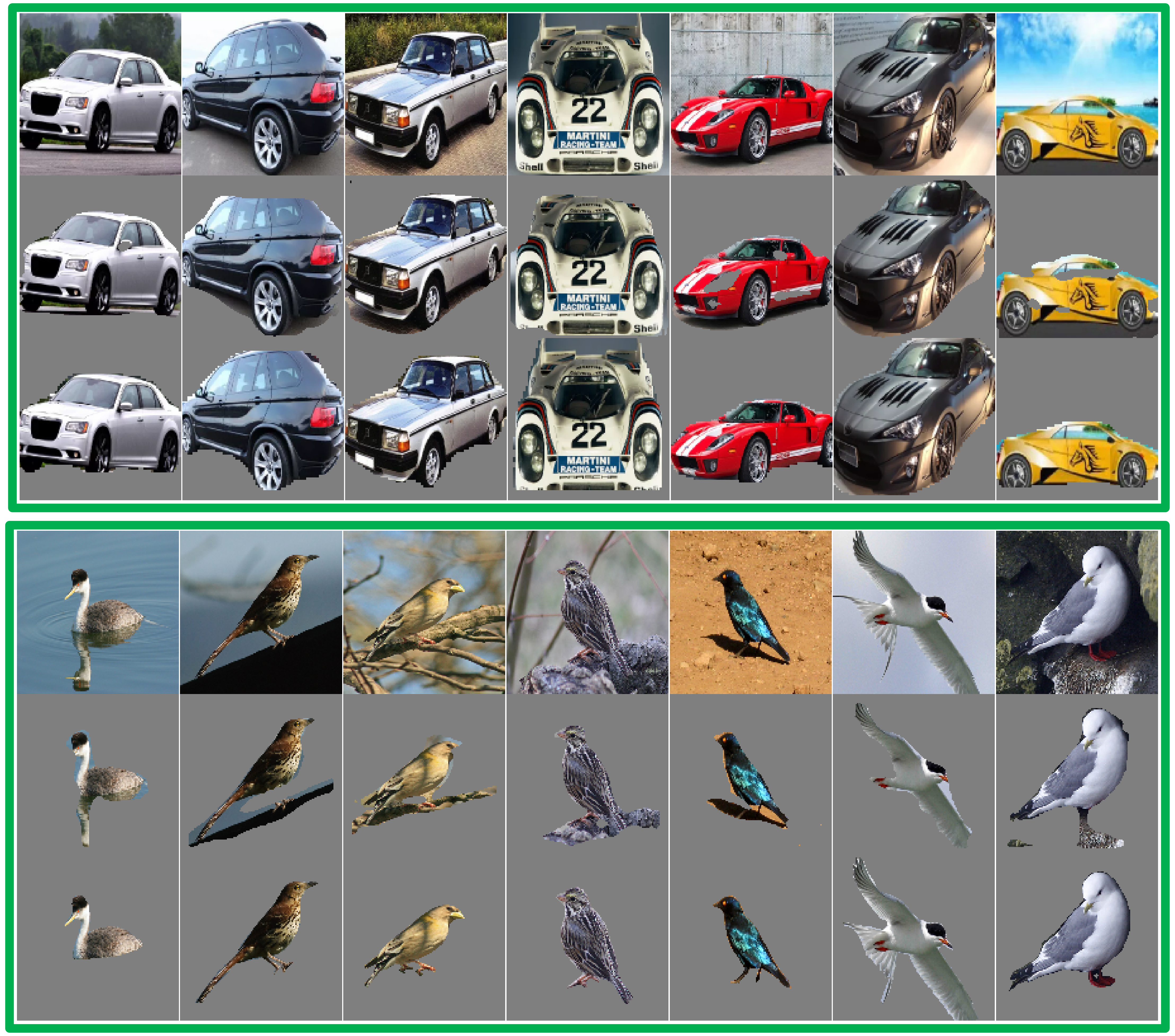
Supervision in Unsupervised Learning

- What can we learn from a random (though arbitrarily large) set of images?
  - Not much unless we make some **assumptions**
- Shifting supervision from per-sample labeling to specifying data properties
  - More powerful than expected!

Supervision in Unsupervised Learning

- What can we learn from a random (though arbitrarily large) set of images?
  - Not much unless we make some **assumptions**
- Shifting supervision from per-sample labeling to specifying data properties
  - More powerful than expected!
  - Sufficient to obtain **interpretable representations** such as: object segmentation, 3D reconstruction, viewpoint estimation, landmark detection etc

Unsupervised Segmentation

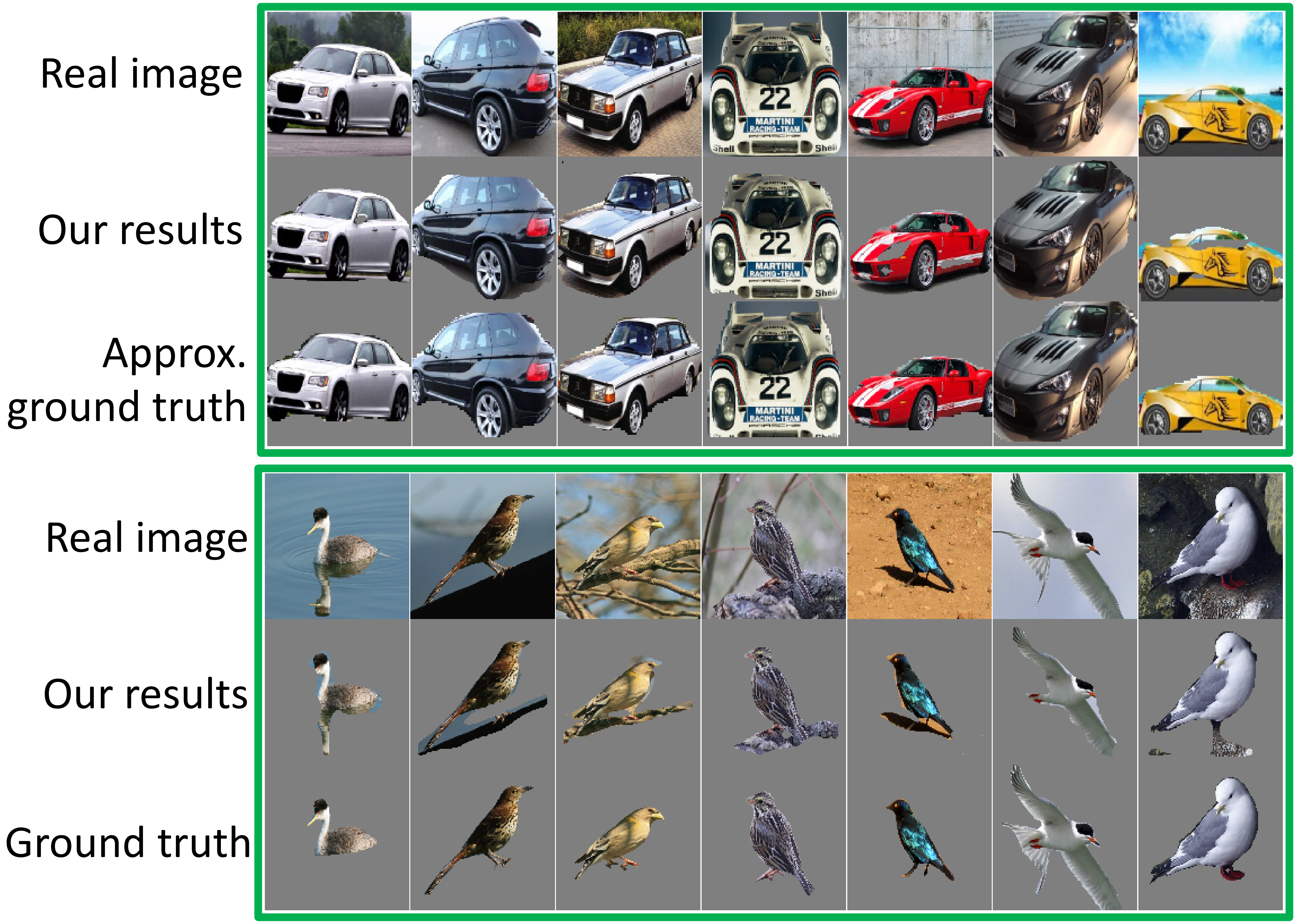


Real image
Our results
Approx. ground truth

Real image
Our results
Ground truth

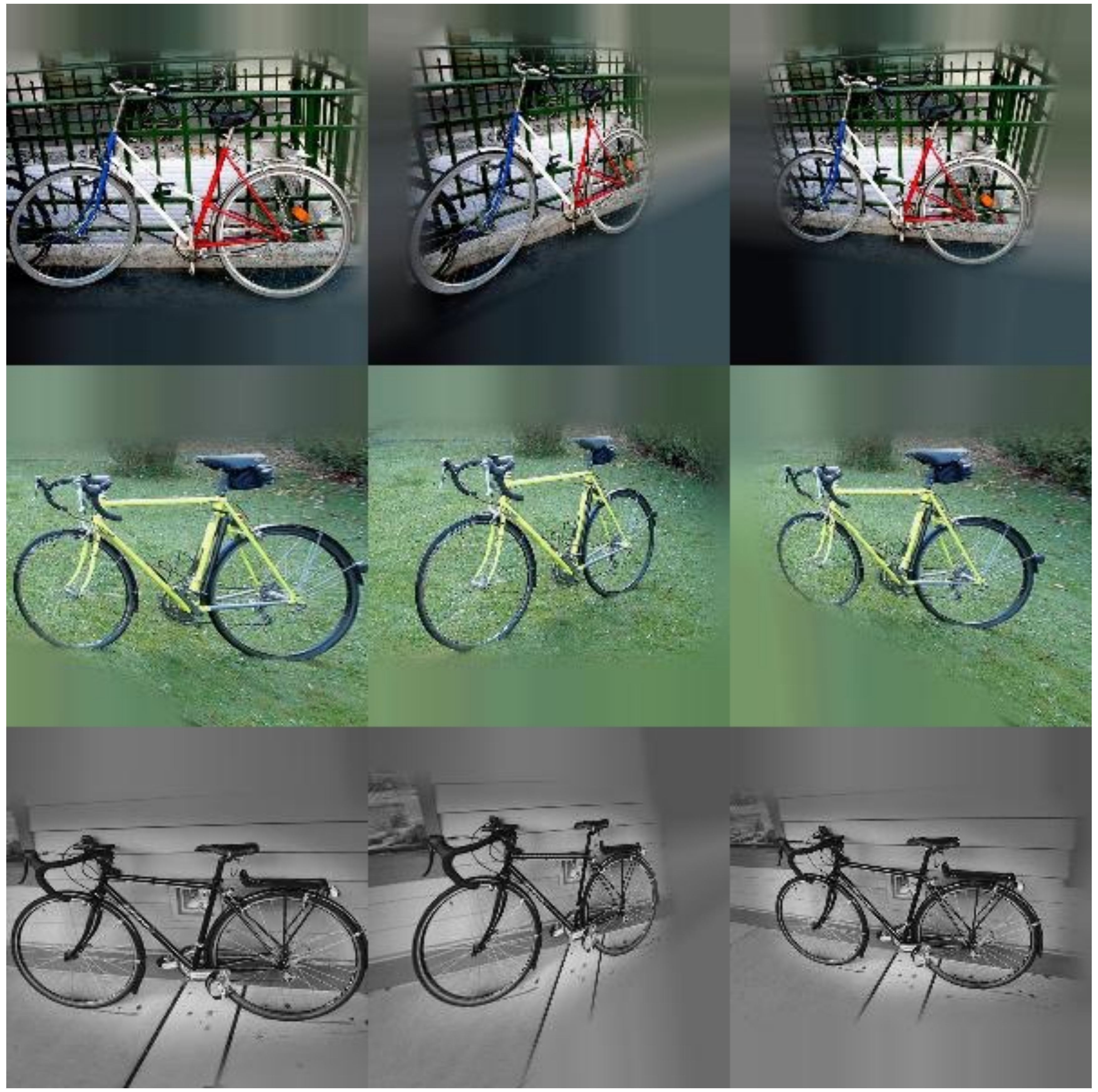
Bielski and Favaro, Emergence of Object Segmentation in Perturbed Generative Models, 2019

Unsupervised Segmentation



Bielski and Favaro, Emergence of Object Segmentation in Perturbed Generative Models, 2019

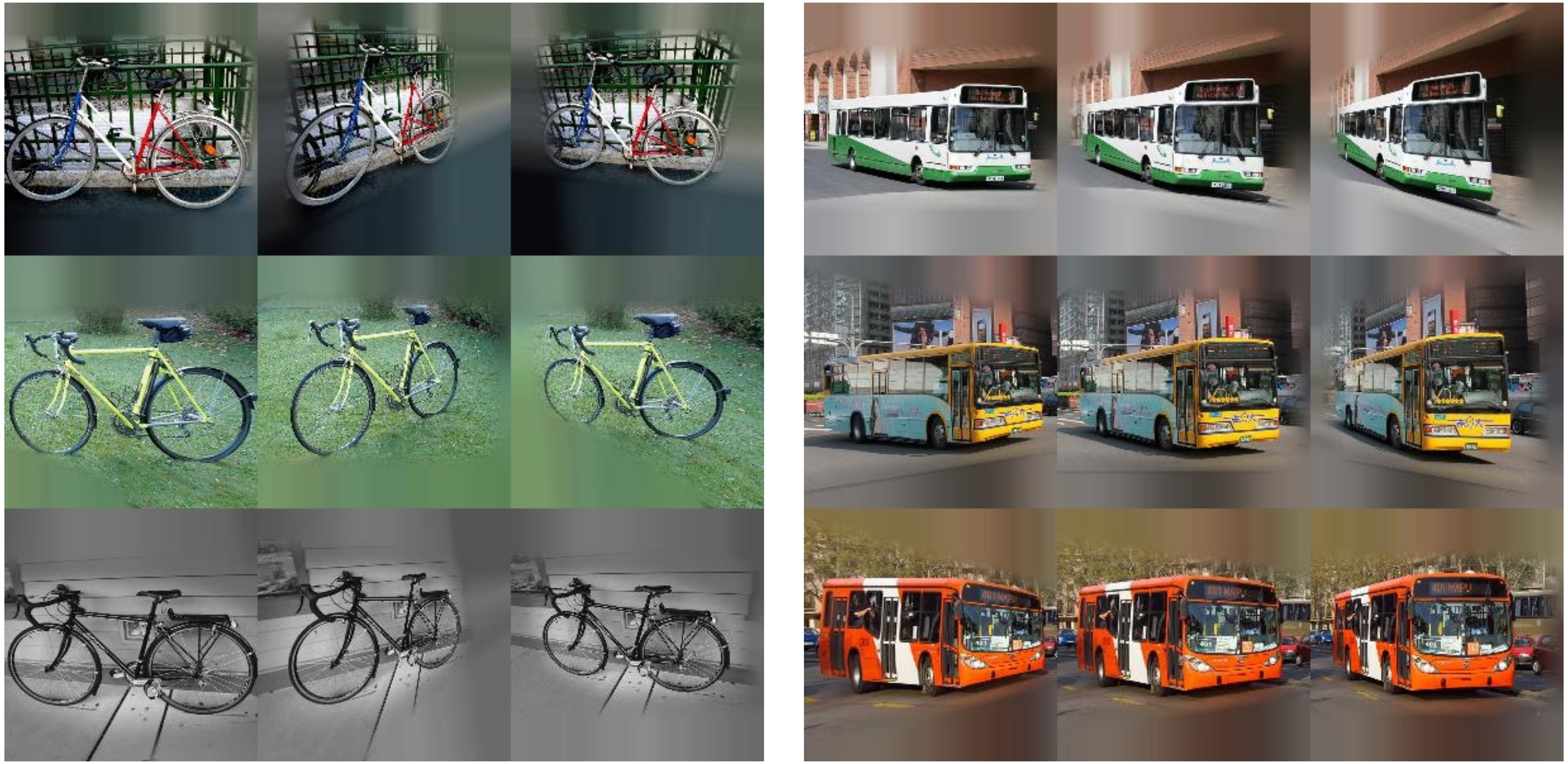
Unsupervised Viewpoint Estimation
3D Hypothesis, Intervention and Realism



original images along the diagonal

Szabó, Vedaldi and Favaro, *Building the View Graph of a Category by Exploiting Image Realism*, 3dRR, ICCV 2015

Unsupervised Viewpoint Estimation
3D Hypothesis, Intervention and Realism



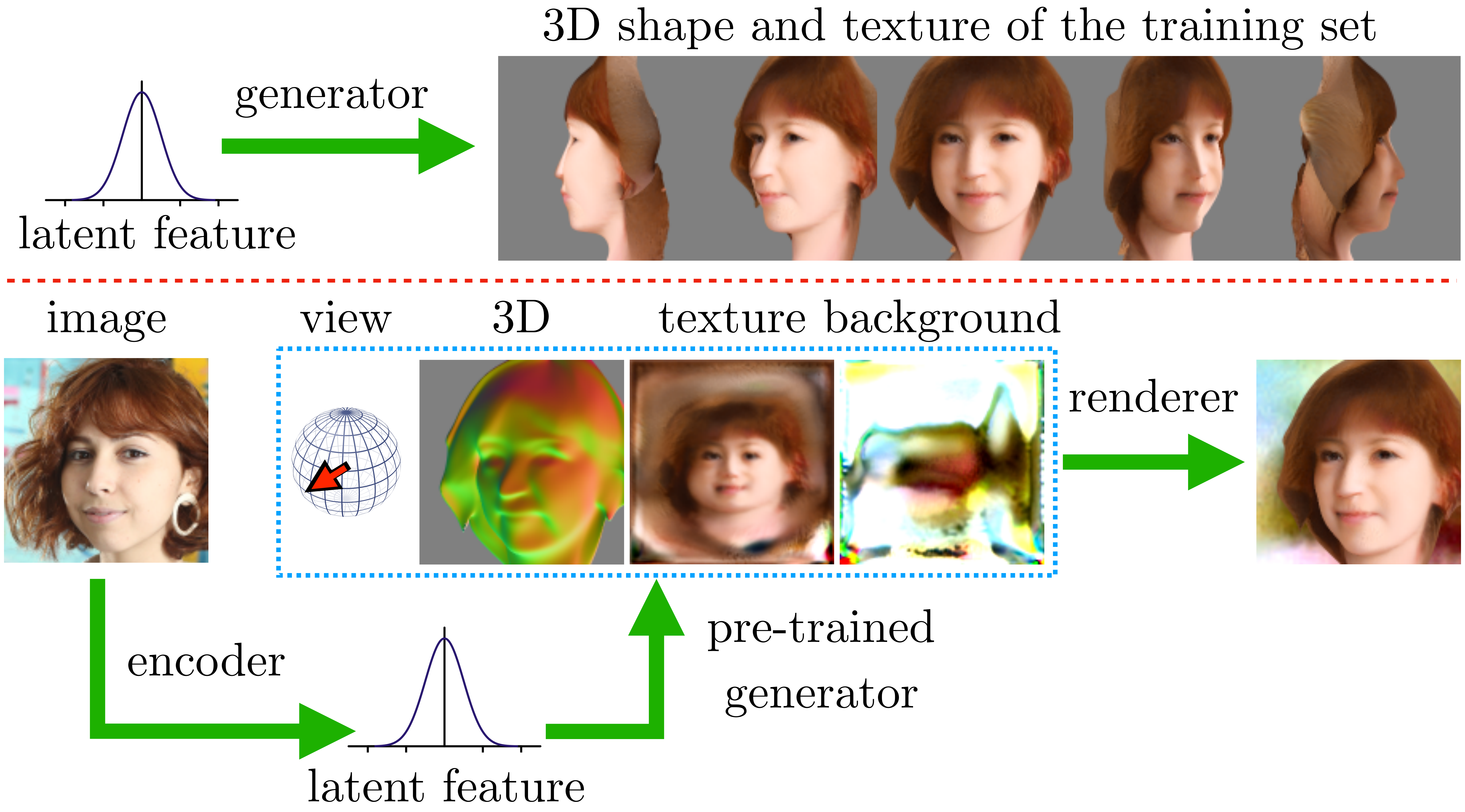
original images along the diagonal

Szabó, Vedaldi and Favaro, *Building the View Graph of a Category by Exploiting Image Realism*, 3dRR, ICCV 2015

Unsupervised Viewpoint Estimation

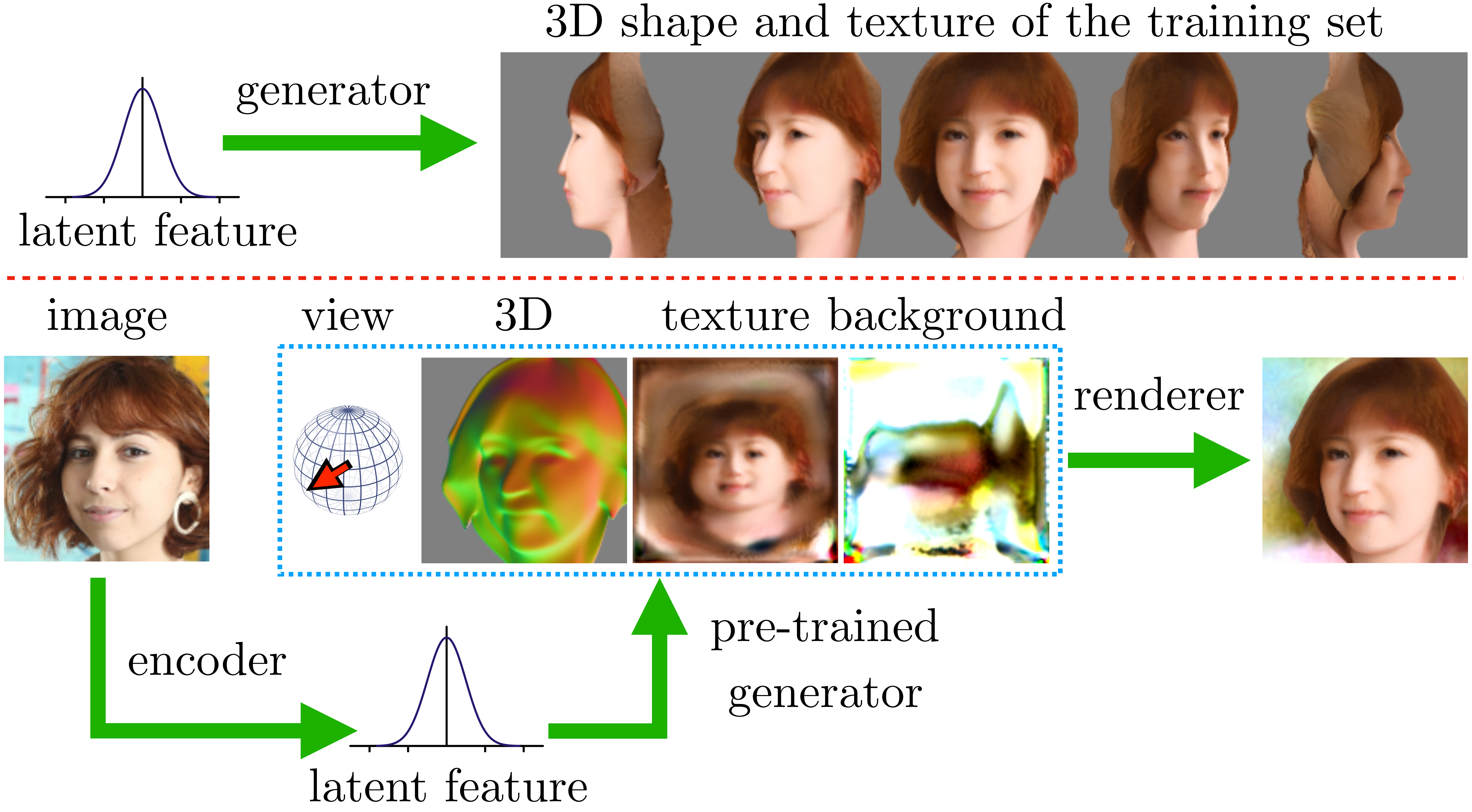
Unsupervised Viewpoint Estimation

Unsupervised 3D Estimation



Szabo and Favaro, Unsupervised 3D Shape Learning from Image Collections in the Wild, 2018
Szabo et al, Unsupervised Generative 3D Shape Learning from Natural Images, 2019

Unsupervised 3D Estimation



Szabo and Favaro, Unsupervised 3D Shape Learning from Image Collections in the Wild, 2018
Szabo et al, Unsupervised Generative 3D Shape Learning from Natural Images, 2019

Unsupervised 3D Estimation

generated image | generated 3D | generated texture | generated background | generated viewpoints

Unsupervised 3D Estimation

generated image | generated 3D | generated texture | generated background | generated viewpoints

Self-Supervised Learning

- The objective is to build features ϕ so that

$$p(y \mid \phi(x))$$

is a good approximation of $p(y \mid x)$ for several tasks (and corresponding labels)

Self-Supervised Learning

- The objective is to build features ϕ so that

$$p(y\,|\,\phi(x))$$

is a good approximation of $p(y\,|\,x)$ for several tasks (and corresponding labels)

Self-Supervised Learning

- The objective is to build features ϕ so that

$$p(y \mid \phi(x))$$

pre-training

is a good approximation of $p(y \mid x)$ for several tasks (and corresponding labels)

Self-Supervised Learning

- The objective is to build features ϕ so that

$$p(y \,|\, \phi(x))$$

pre-training

is a good approximation of $p(y \,|\, x)$ for several tasks (and corresponding labels)

- Ideally, ϕ should be such that $p(y \,|\, \phi)$ can be “simple” (otherwise $\phi = x$ would be a trivial solution), e.g., a shallow neural network

Feature Design

- One principle to design ϕ is to reduce the dimensionality of the data x while imposing the reconstructibility of x from ϕ (possible because natural images are a small subset of all images)

*Similar concept as the global structure described by Van den Oord et al, Contrastive Predictive Coding, 2018

Feature Design

- One principle to design ϕ is to reduce the dimensionality of the data x while imposing the reconstructibility of x from ϕ (possible because natural images are a small subset of all images)

*Similar concept as the global structure described by Van den Oord et al, Contrastive Predictive Coding, 2018

Feature Design

- One principle to design ϕ is to reduce the dimensionality of the data x while imposing the reconstructibility of x from ϕ (possible because natural images are a small subset of all images)
  - This leads to Autoencoders (and their variations, such as denoising AEs)

*Similar concept as the global structure described by Van den Oord et al, Contrastive Predictive Coding, 2018

Feature Design

- One principle to design ϕ is to reduce the dimensionality of the data x while imposing the reconstructibility of x from ϕ (possible because natural images are a small subset of all images)
    - This leads to Autoencoders (and their variations, such as denoising AEs)
- Another principle is to design ϕ such that it defines an ℓ_2 distance that is related to the *high-level attributes** of the data; with such features a simple classifier or regressor should suffice

*Similar concept as the global structure described by Van den Oord et al, Contrastive Predictive Coding, 2018

Attributes

- What are attributes?

Attributes

- What are attributes?

Attributes

- What are attributes?
    - We consider attributes that are statistics of random variables that are based on a hierarchy of other simpler random variables → this is what neural networks models build

Attributes

- What are attributes?
  - We consider attributes that are statistics of random variables that are based on a hierarchy of other simpler random variables → this is what neural networks models build
- The pretext-task allows to influence what attributes features should be invariant to and discriminate

Attributes

- What are attributes?
  - We consider attributes that are statistics of random variables that are based on a hierarchy of other simpler random variables → this is what neural networks models build
- The pretext-task allows to influence what attributes features should be invariant to and discriminate
- Example: A simple local attribute is the color histogram; it is the distribution of single pixels seen as independent samples

Global vs Local Attributes

- Original data

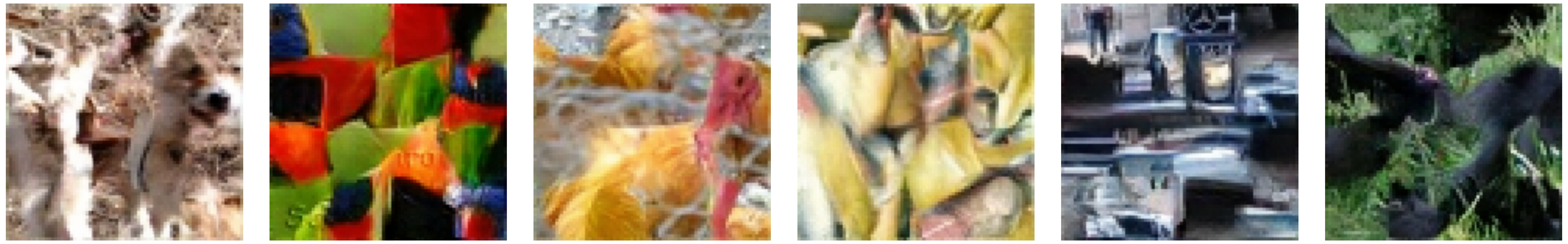
Global vs Local Attributes

- Original data

Global vs Local Attributes

- Original data

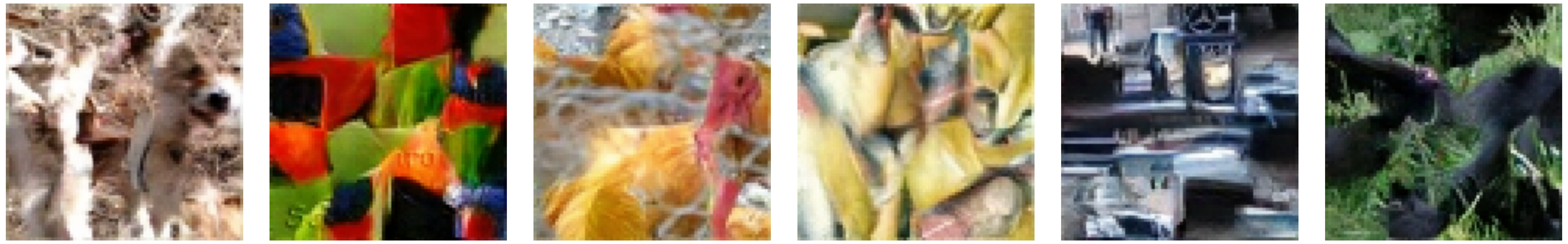
- Images where the local statistics are the same, but the global ones are not



Global vs Local Attributes

- Original data

- Images where the local statistics are the same, but the global ones are not

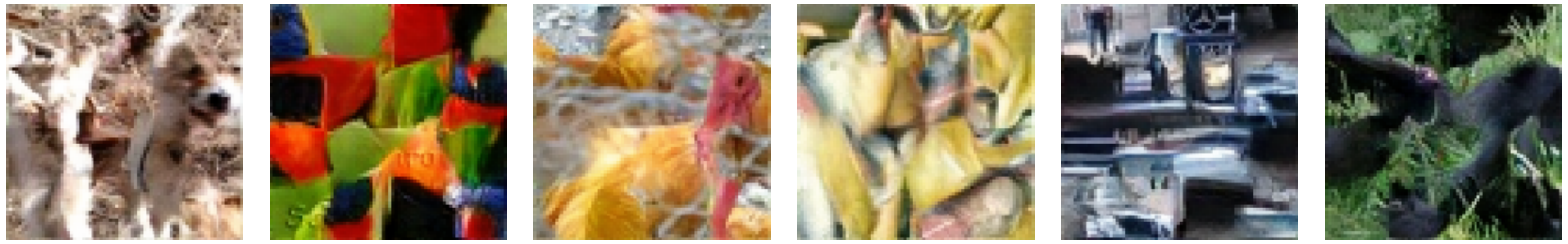


- Supervised learning features do not distinguish well between the two sets

Global vs Local Attributes

- Original data

- Images where the local statistics are the same, but the global ones are not



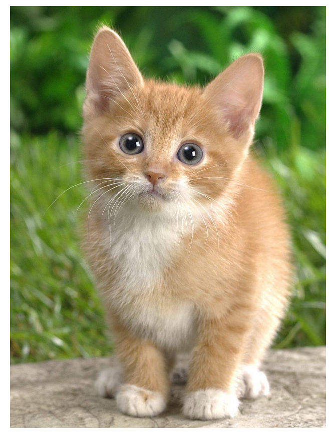
- Supervised learning features do not distinguish well between the two sets

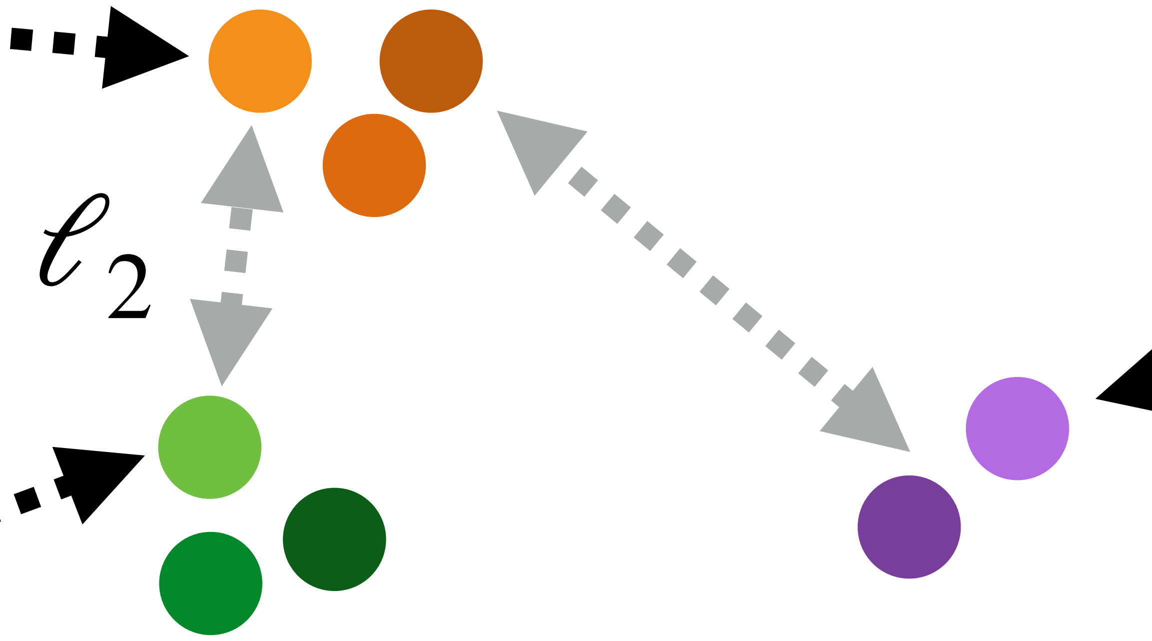
- Do we know what a model uses to solve a supervised task? → Example shows that mid-range texture* classification is sufficient to solve the supervised task

*See Jenni et al, Steering Self-Supervised Feature Learning Beyond Local Pixel Statistics, 2020 and Geirhos et al, Imagenet-trained cnns are biased towards texture; increasing shape bias improves accuracy and robustness, 2018

Pretext-Tasks and Attributes

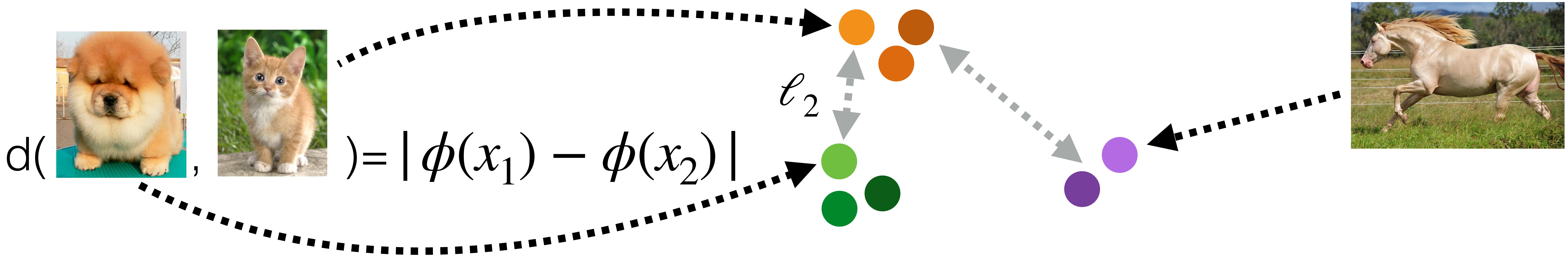
- ℓ_2 on features defines a new distance between images

d( , )= $|\phi(x_1) - \phi(x_2)|$



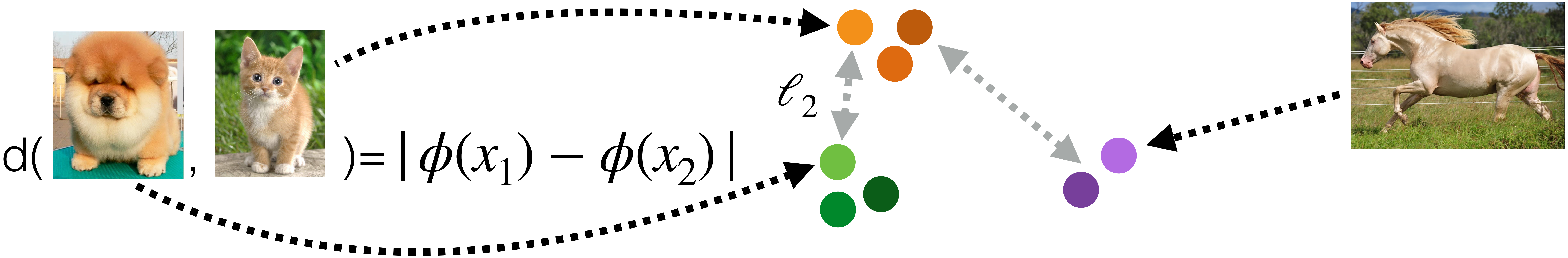
Pretext-Tasks and Attributes

- ℓ_2 on features defines a new distance between images



Pretext-Tasks and Attributes

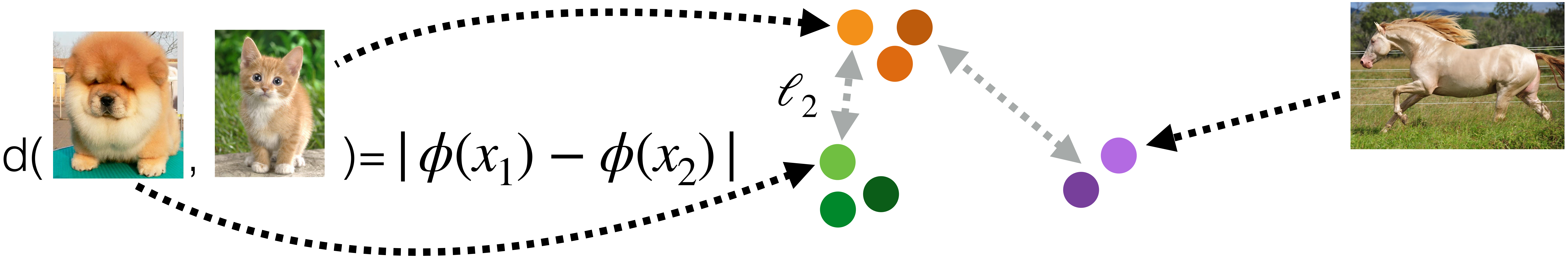
- ℓ_2 on features defines a new distance between images



- How does the pretext-task affect this distance and which attributes does this distance relate to?

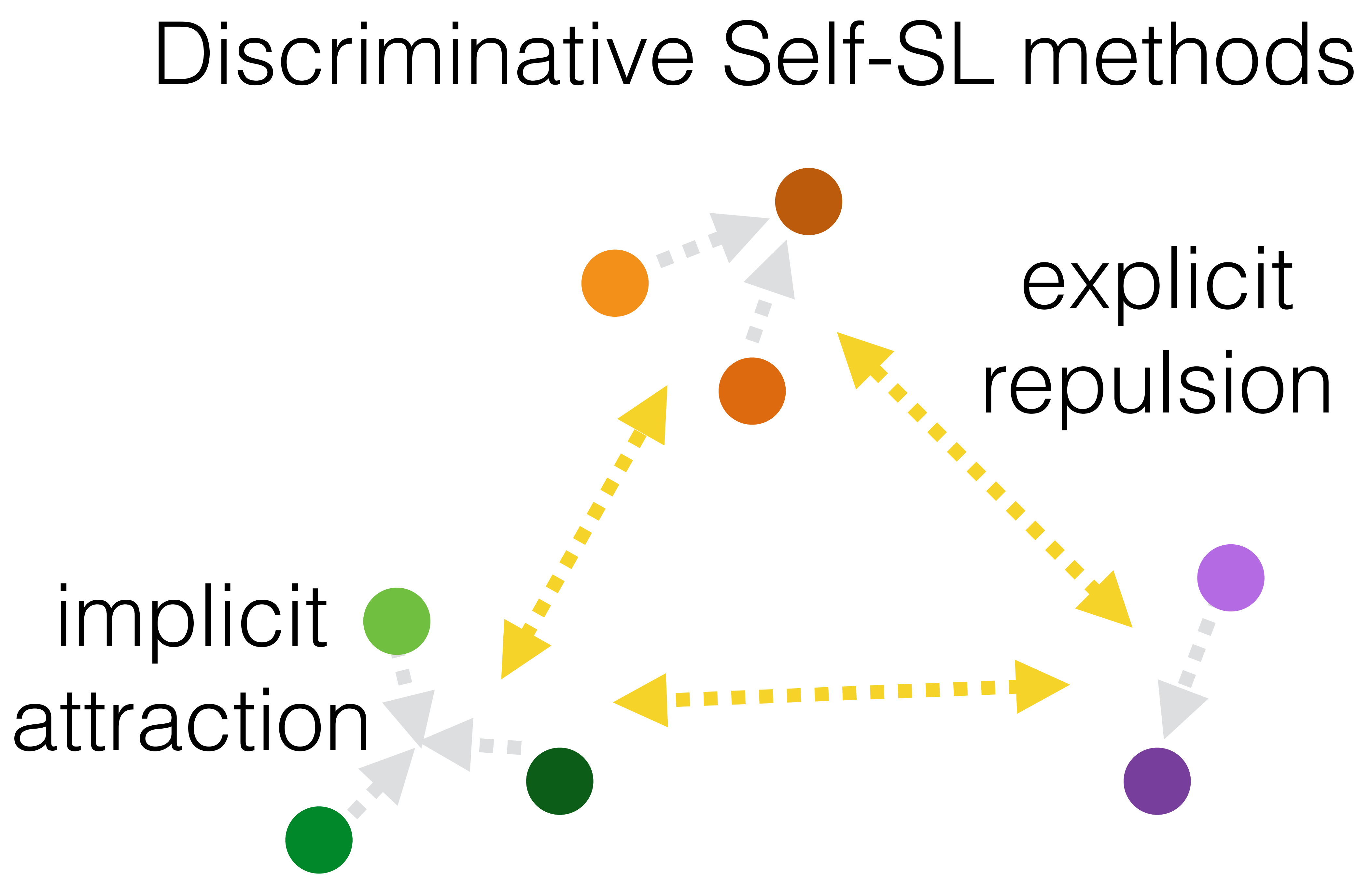
Pretext-Tasks and Attributes

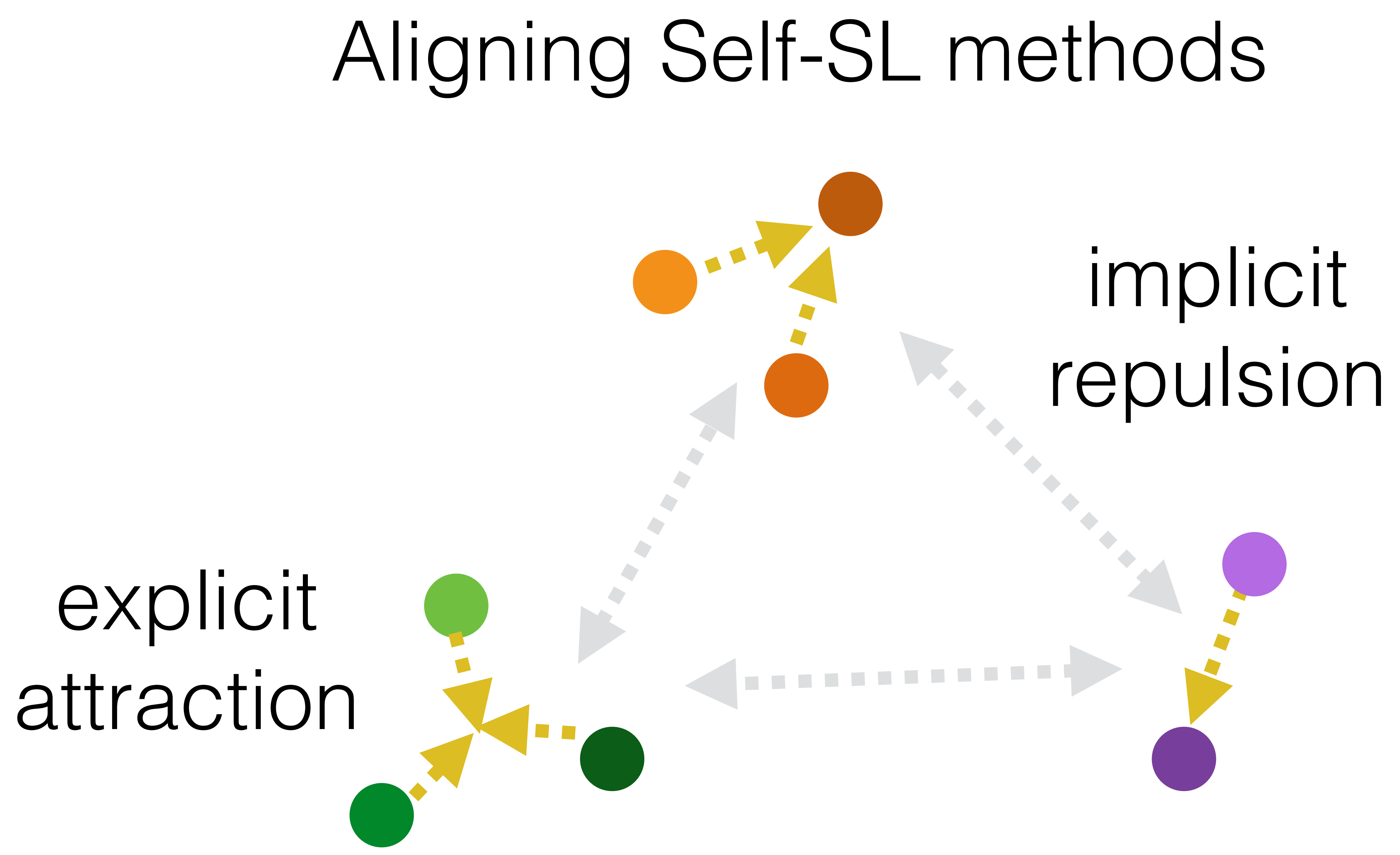
- ℓ_2 on features defines a new distance between images



- How does the pretext-task affect this distance and which attributes does this distance relate to?
  - The pretext-task builds features that can distinguish transformed versions of the data → these transformations define the attributes

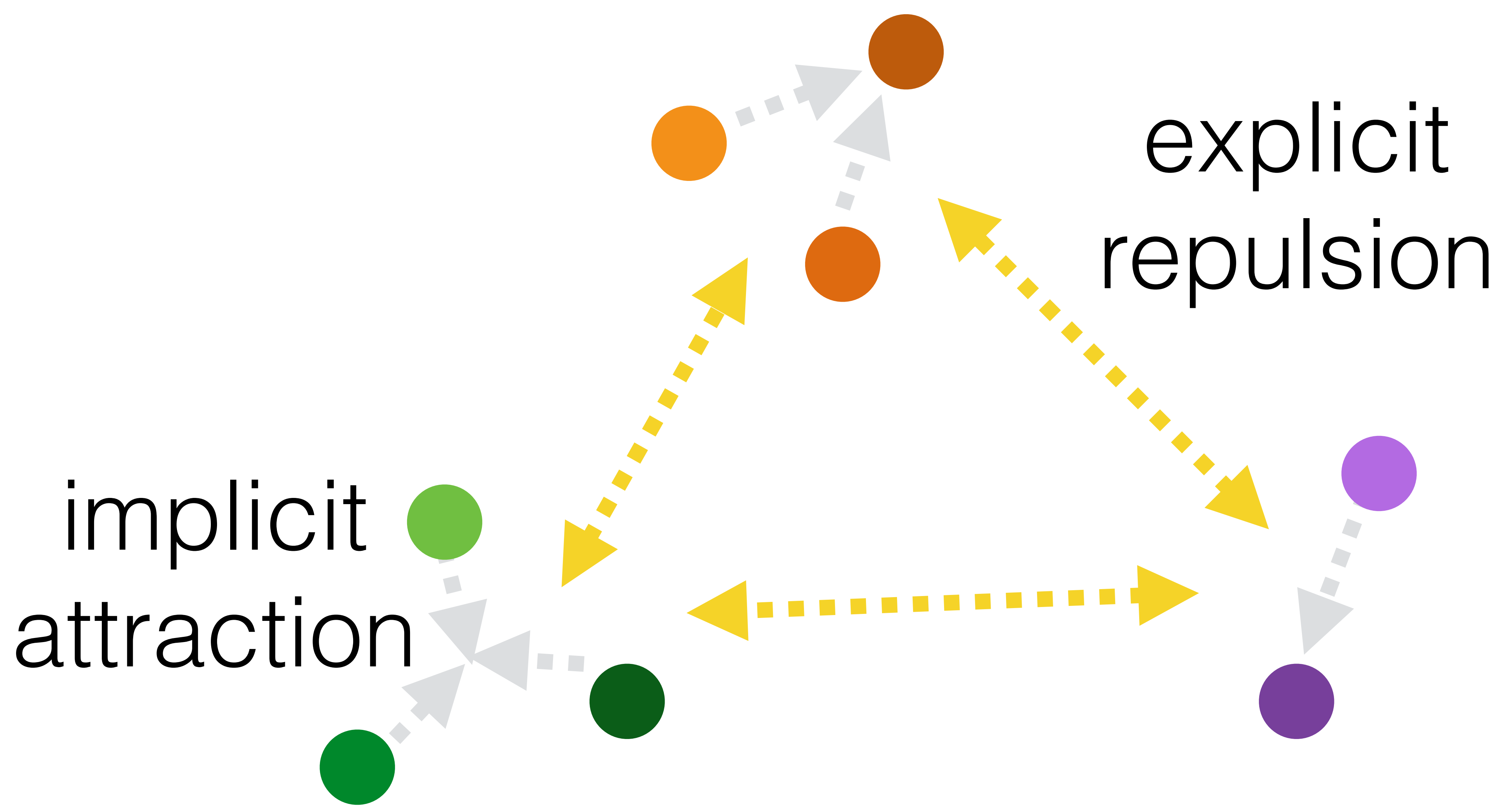
Defining the Feature Space



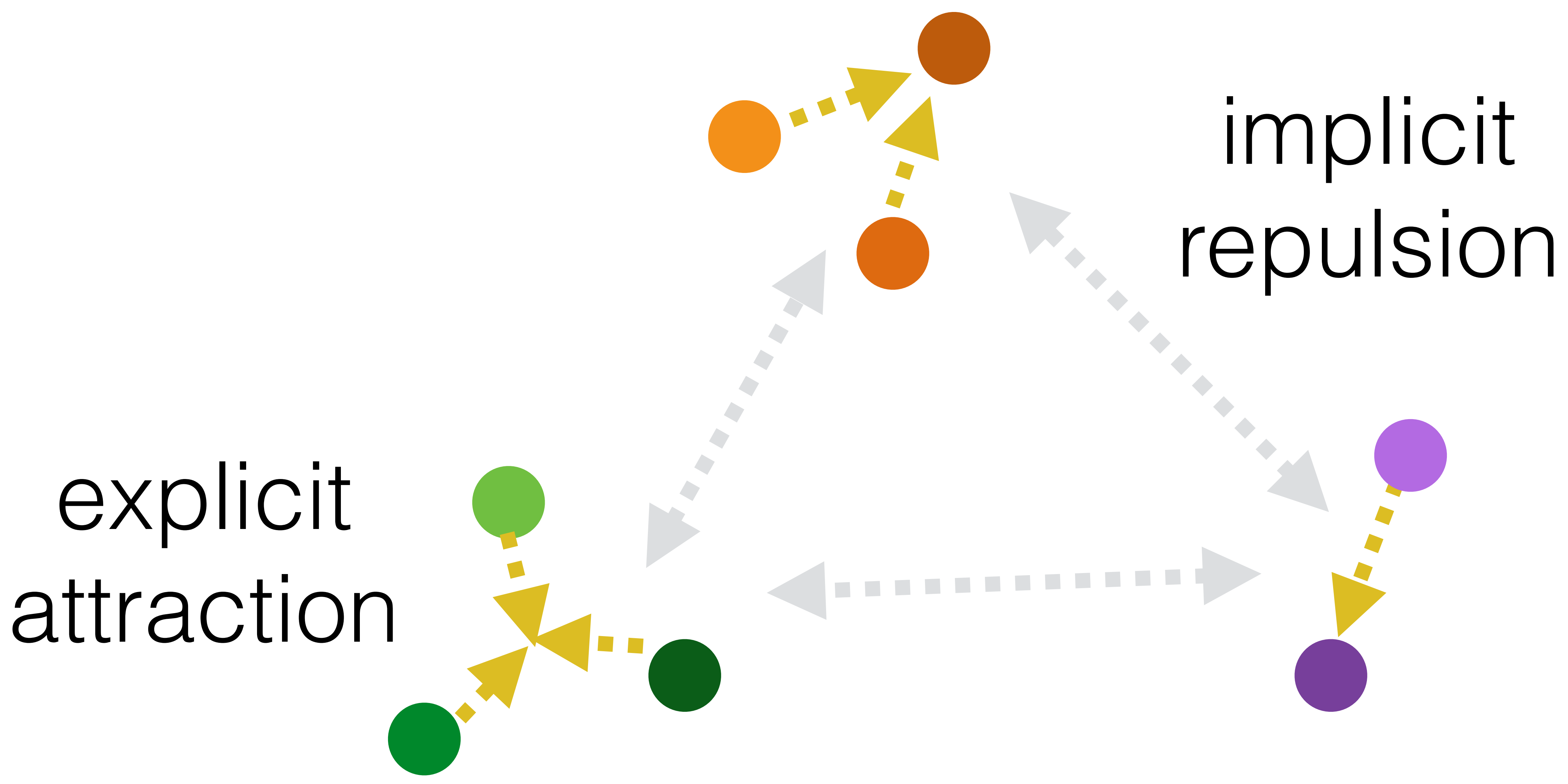


Defining the Feature Space

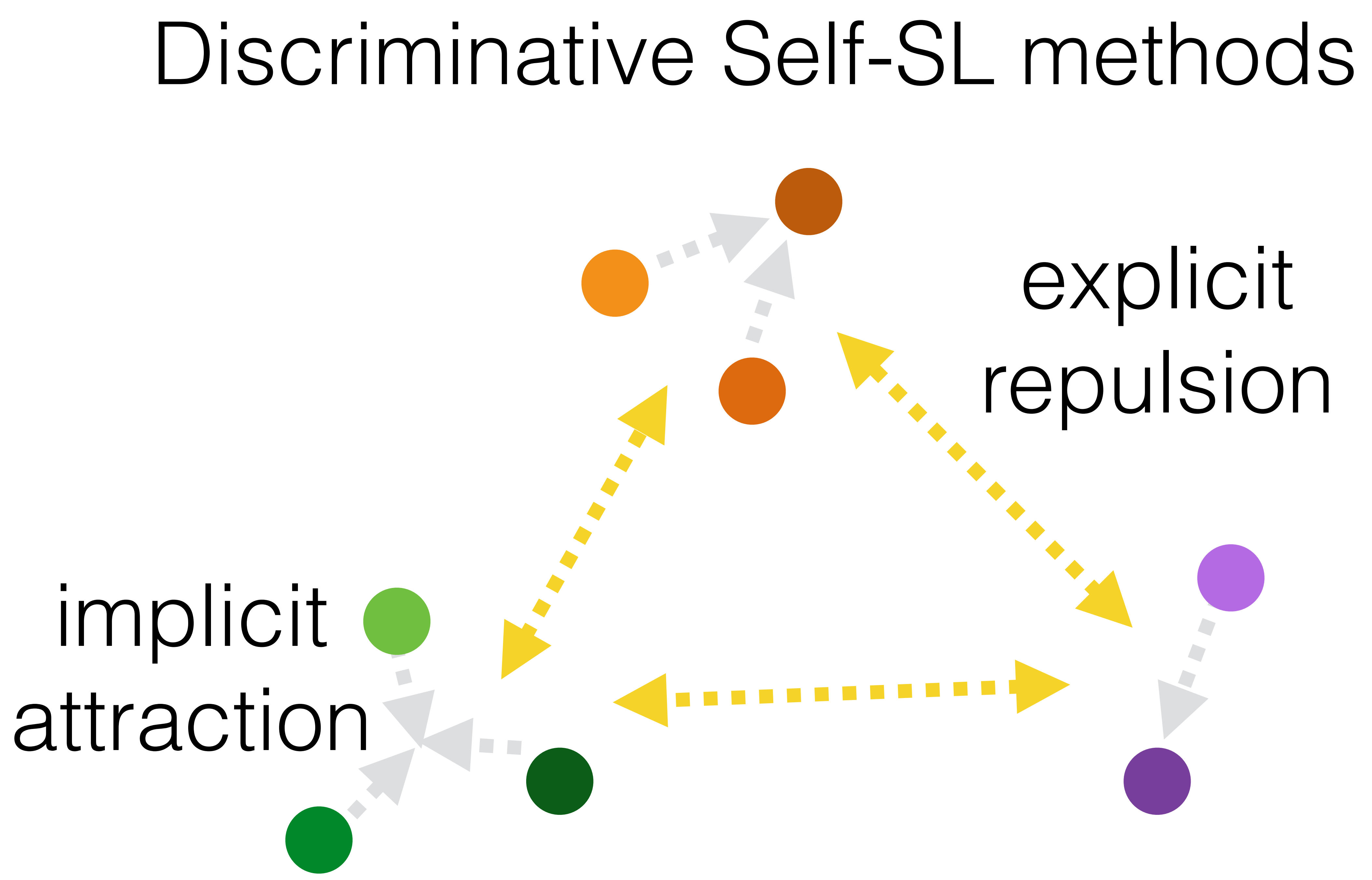
Discriminative Self-SL methods

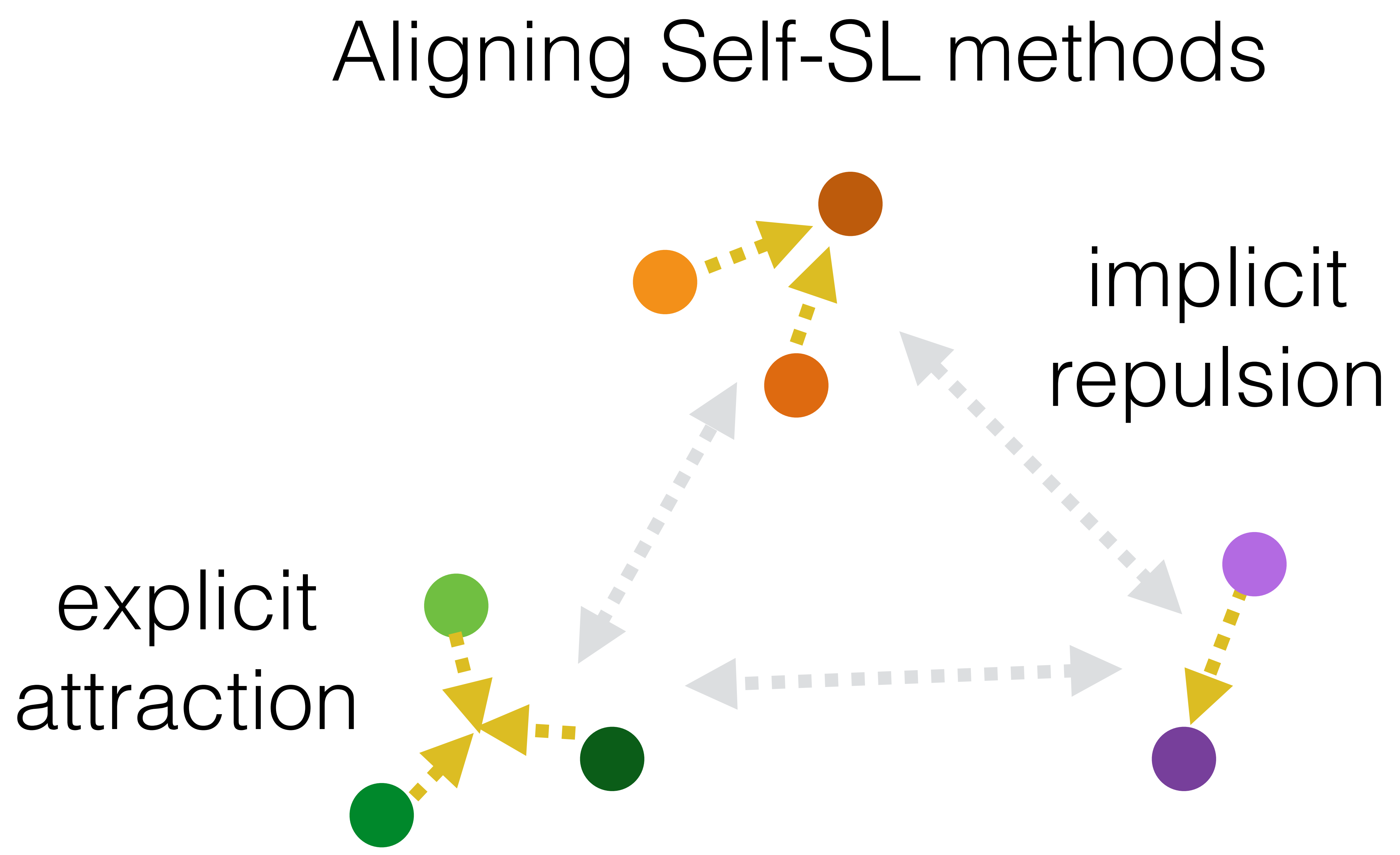


Aligning Self-SL methods



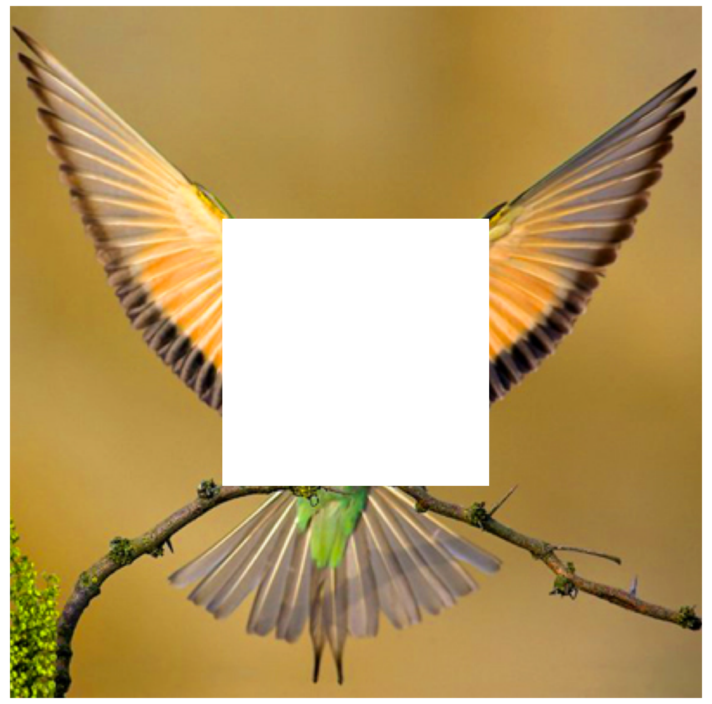
Defining the Feature Space





All Self-SL methods define some heuristic principle
This is the Unsupervised Learning alternative to labeling

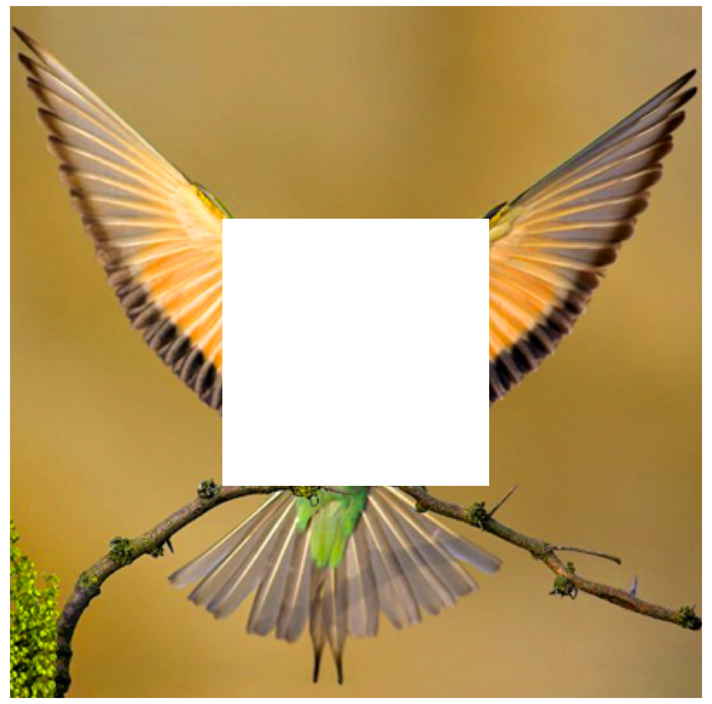
Reconstructibility



- Features should allow the reconstruction of a data sample from its context or other transformed versions of that sample

*D. Pathak et al, Context encoders: Feature learning by inpainting, 2016
G. Larsson et al, Learning representations for automatic colorization, 2016

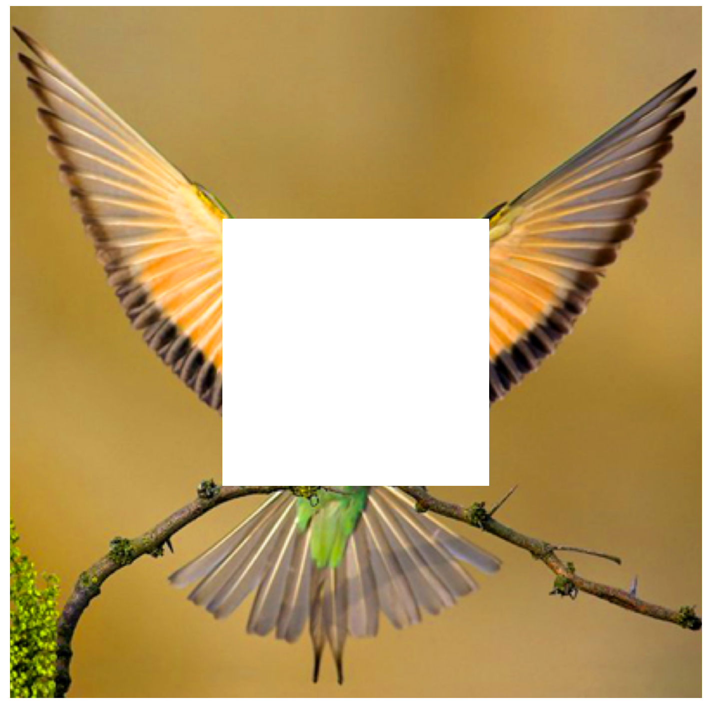
Reconstructibility



- Features should allow the reconstruction of a data sample from its context or other transformed versions of that sample

*D. Pathak et al, Context encoders: Feature learning by inpainting, 2016
G. Larsson et al, Learning representations for automatic colorization, 2016

Reconstructibility

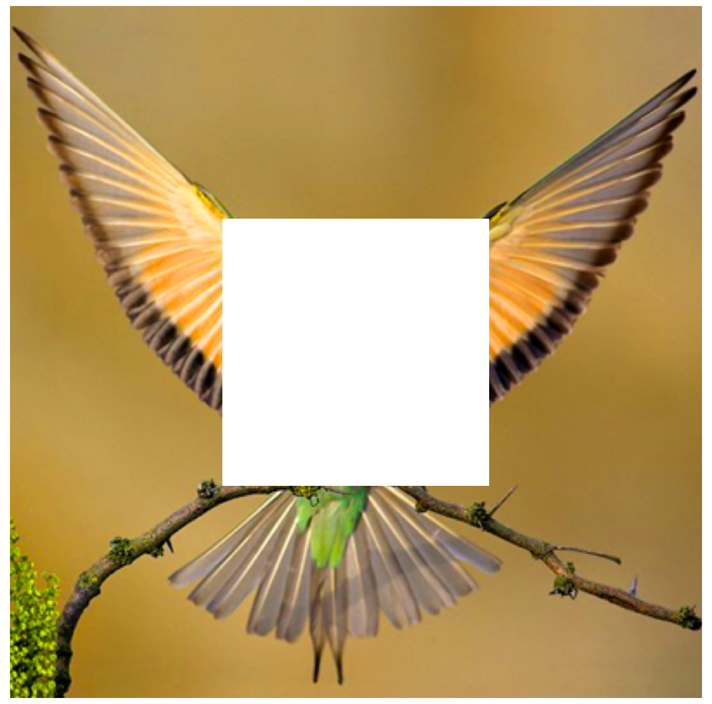


- Features should allow the reconstruction of a data sample from its context or other transformed versions of that sample

- Can be related to denoising AEs → Features are encouraged to be invariant to the added “noise”

*D. Pathak et al, Context encoders: Feature learning by inpainting, 2016
G. Larsson et al, Learning representations for automatic colorization, 2016

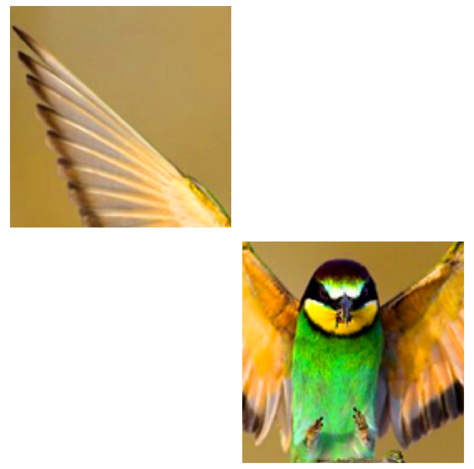
Reconstructibility



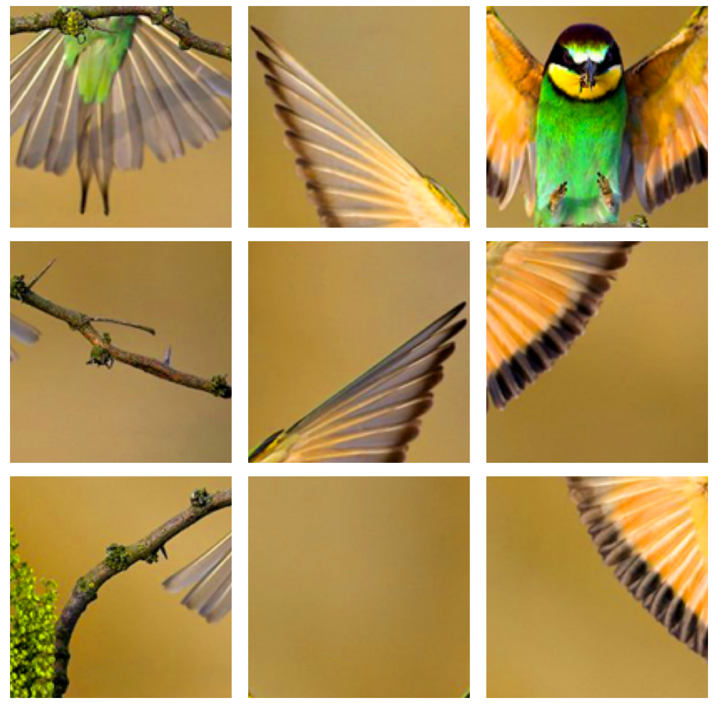
- Features should allow the reconstruction of a data sample from its context or other transformed versions of that sample
- Can be related to denoising AEs → Features are encouraged to be invariant to the added "noise"
- **Aligning Self-SL**: Images which differ by the transformation used in the pretext-task are mapped to similar features

*D. Pathak et al, Context encoders: Feature learning by inpainting, 2016
G. Larsson et al, Learning representations for automatic colorization, 2016

Spatial Configuration of Parts

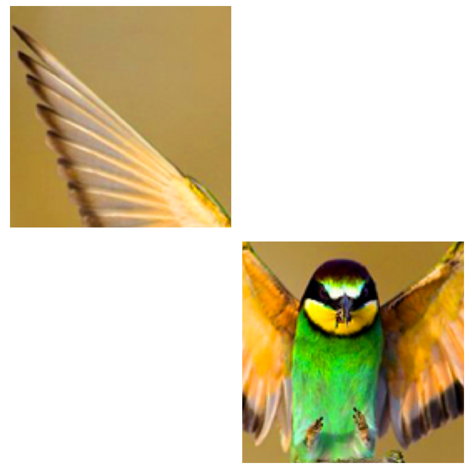


- Features of object parts must be distinguishable from those of other parts within the same image

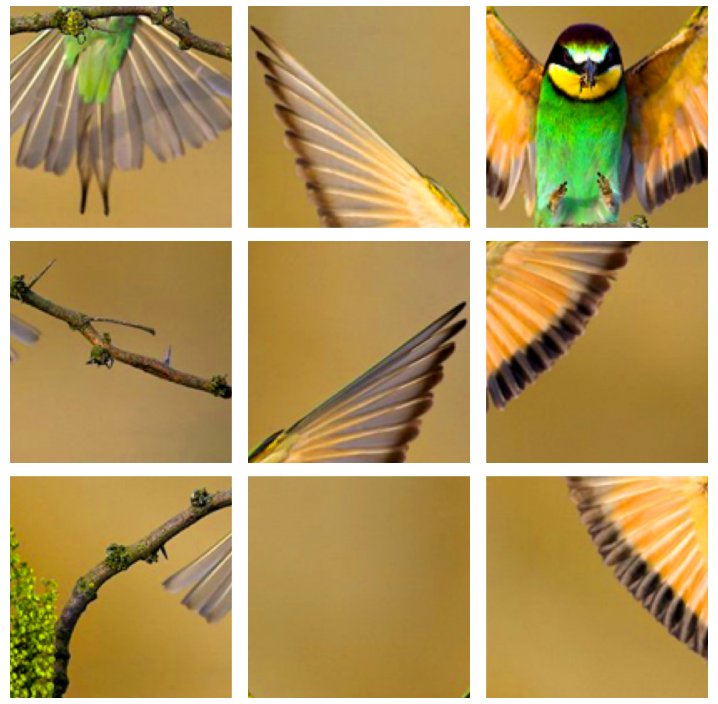


*Doersch et al 2015, Noroozi and Favaro 2016, Mundhenk et al. 2018, Noroozi et al 2018

Spatial Configuration of Parts

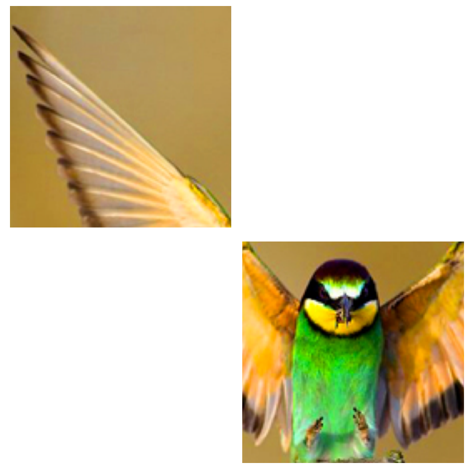


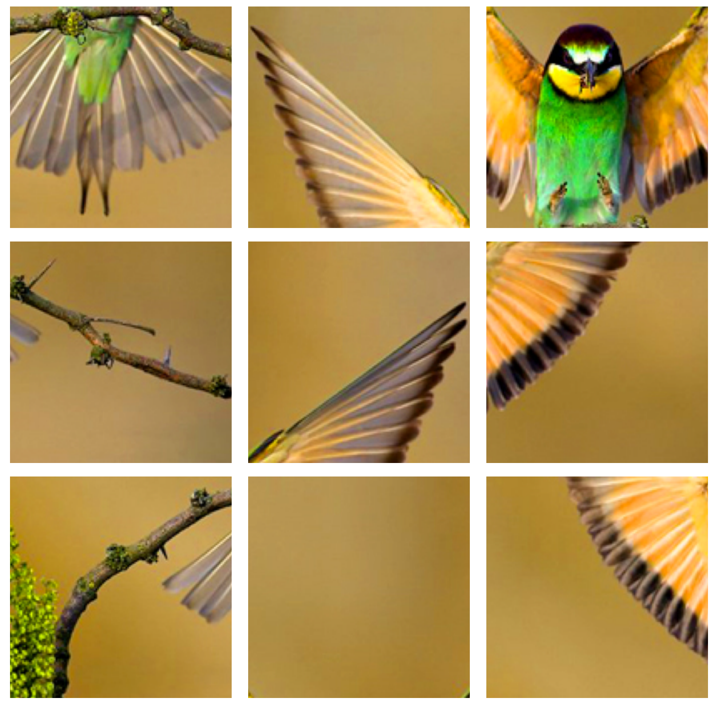
- Features of object parts must be distinguishable from those of other parts within the same image



*Doersch et al 2015, Noroozi and Favaro 2016, Mundhenk et al. 2018, Noroozi et al 2018

Spatial Configuration of Parts



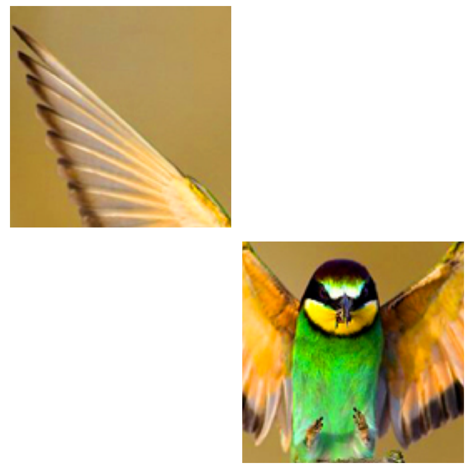


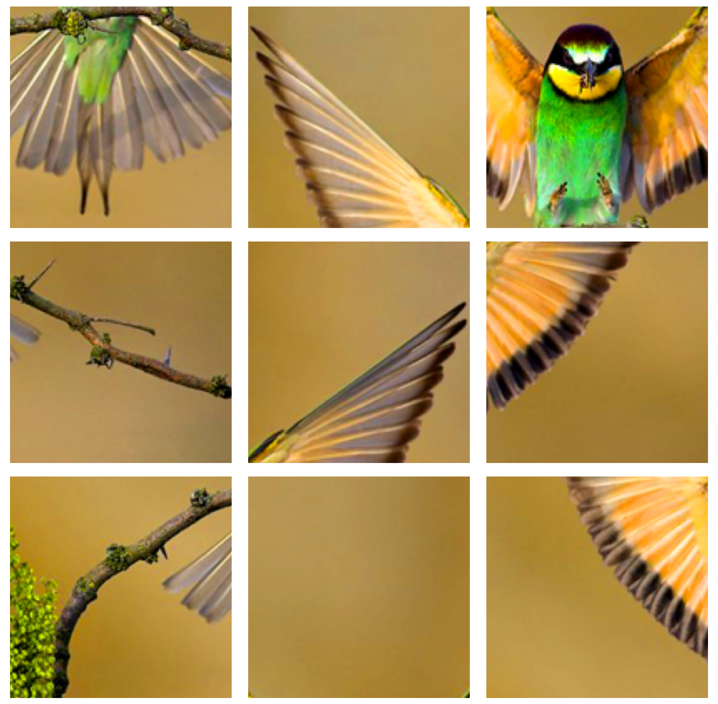
- Features of object parts must be distinguishable from those of other parts within the same image

- **Discriminative Self-SL**: No explicit constraint to group features other than the dimensionality reduction due to the neural network architecture

*Doersch et al 2015, Noroozi and Favaro 2016, Mundhenk et al. 2018, Noroozi et al 2018

Spatial Configuration of Parts





- Features of object parts must be distinguishable from those of other parts within the same image
- **Discriminative Self-SL**: No explicit constraint to group features other than the dimensionality reduction due to the neural network architecture
- Alignment might occur due to other mechanisms: E.g., the network architecture might encourage some form of alignment or the jittering used to sample the parts might facilitate some transformation invariance

*Doersch et al 2015, Noroozi and Favaro 2016, Mundhenk et al. 2018, Noroozi et al 2018

Changing Only Global Attributes

- Train a network to modify only the global attributes (e.g., missing face, disconnected limbs)

*S. Jenni and P. Favaro, Self-Supervised Feature Learning by Learning to Spot Artifacts, 2018
S. Jenni et al, Steering Self-Supervised Feature Learning Beyond Local Pixel Statistics, 2020

Changing Only Global Attributes

- Train a network to modify only the global attributes (e.g., missing face, disconnected limbs)

*S. Jenni and P. Favaro, Self-Supervised Feature Learning by Learning to Spot Artifacts, 2018
S. Jenni et al, Steering Self-Supervised Feature Learning Beyond Local Pixel Statistics, 2020

Changing Only Global Attributes

- Train a network to modify only the global attributes (e.g., missing face, disconnected limbs)
- **Discriminative Self-SL**: Features of real objects should be distinguishable from features of unrealistic ones

*S. Jenni and P. Favaro, Self-Supervised Feature Learning by Learning to Spot Artifacts, 2018
S. Jenni et al, Steering Self-Supervised Feature Learning Beyond Local Pixel Statistics, 2020

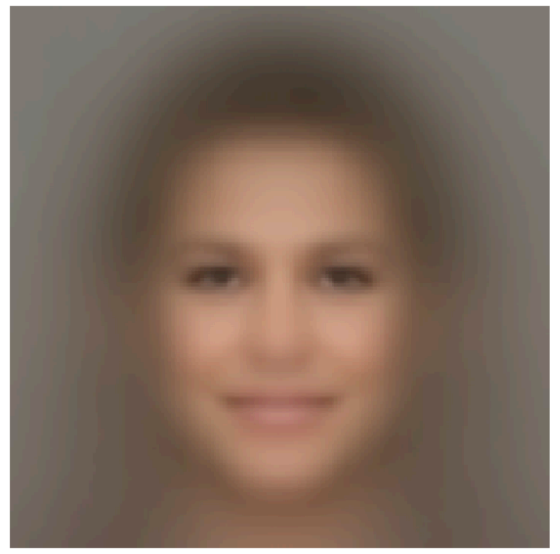
Changing Only Global Attributes

- Train a network to modify only the global attributes (e.g., missing face, disconnected limbs)
- **Discriminative Self-SL**: Features of real objects should be distinguishable from features of unrealistic ones
- Conjecture: Features of images with different global attributes are pushed away from each other; no constraint exists between images with similar global attributes

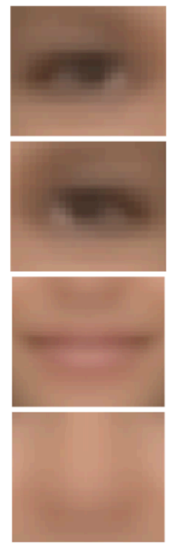
*S. Jenni and P. Favaro, Self-Supervised Feature Learning by Learning to Spot Artifacts, 2018
S. Jenni et al, Steering Self-Supervised Feature Learning Beyond Local Pixel Statistics, 2020

Image Rotations

- **Discriminative Self-SL**: Features should allow the discrimination of rotated images



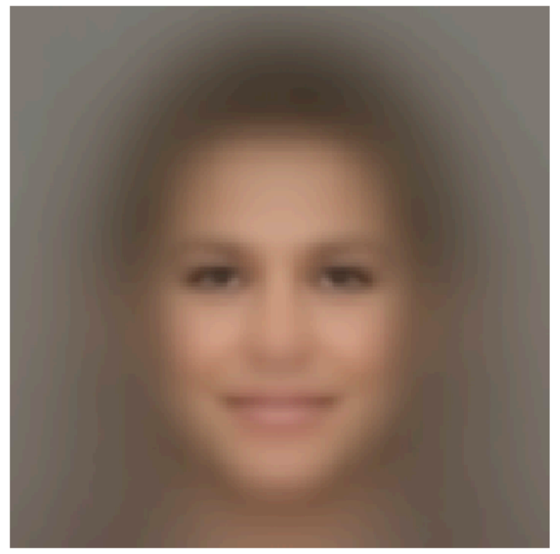
average
face



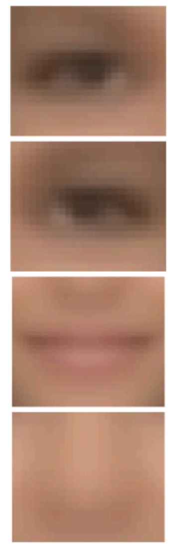
S. Gidaris et al, Unsupervised Representation Learning by Predicting Image Rotations. *2018*

Image Rotations

- **Discriminative Self-SL**: Features should allow the discrimination of rotated images



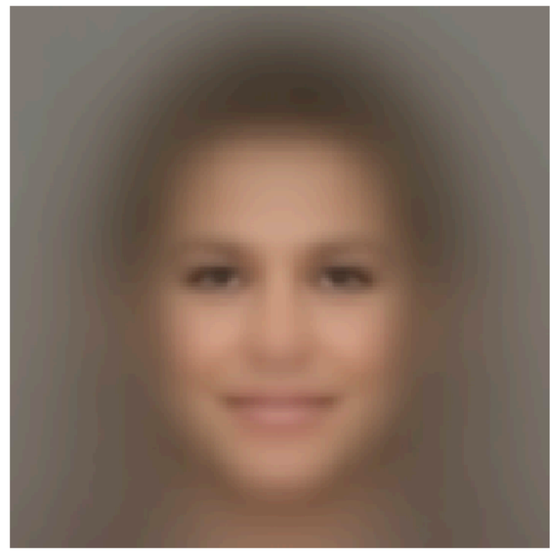
average face



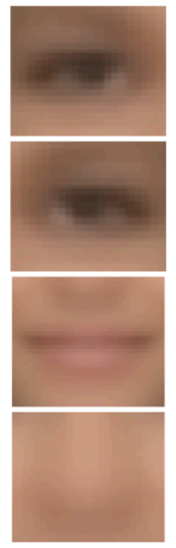
S. Gidaris et al, Unsupervised Representation Learning by Predicting Image Rotations. *2018*

Image Rotations

- **Discriminative Self-SL**: Features should allow the discrimination of rotated images
- What allows the identification of the orientation?



average face



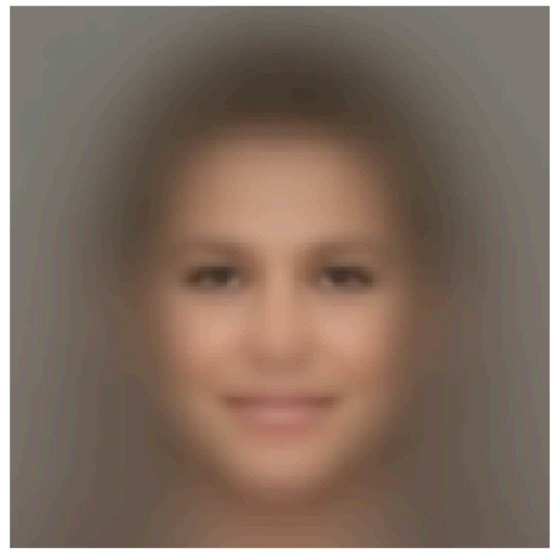
S. Gidaris et al, Unsupervised Representation Learning by Predicting Image Rotations. *2018*

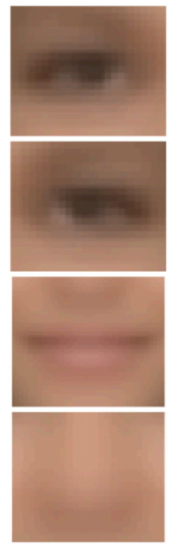
Image Rotations

- **Discriminative Self-SL**: Features should allow the discrimination of rotated images

- What allows the identification of the orientation?

- If orientation can be determined through local patterns (e.g., faces), then features only need to discriminate local patterns

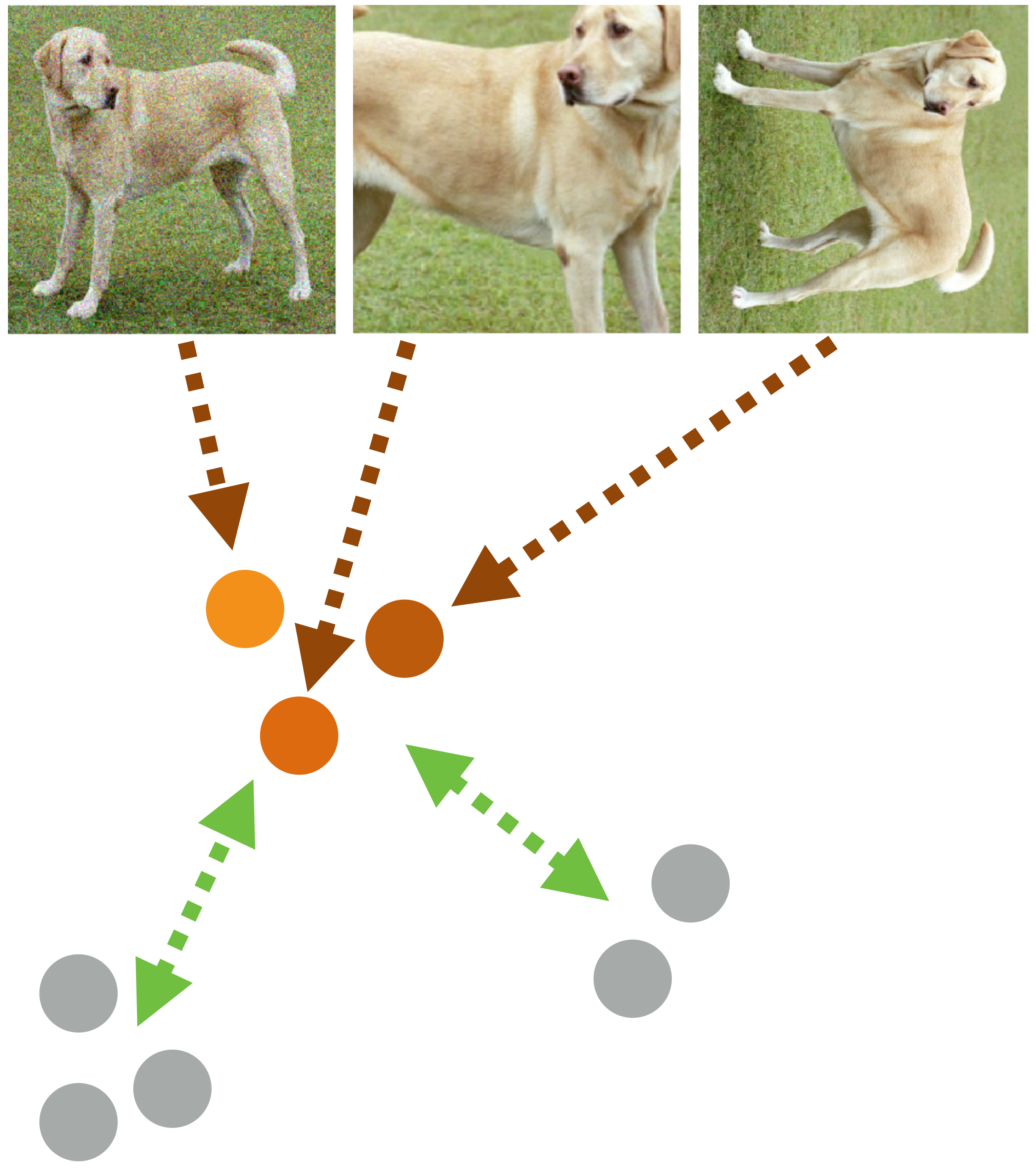




average face

S. Gidaris et al, Unsupervised Representation Learning by Predicting Image Rotations. *2018*

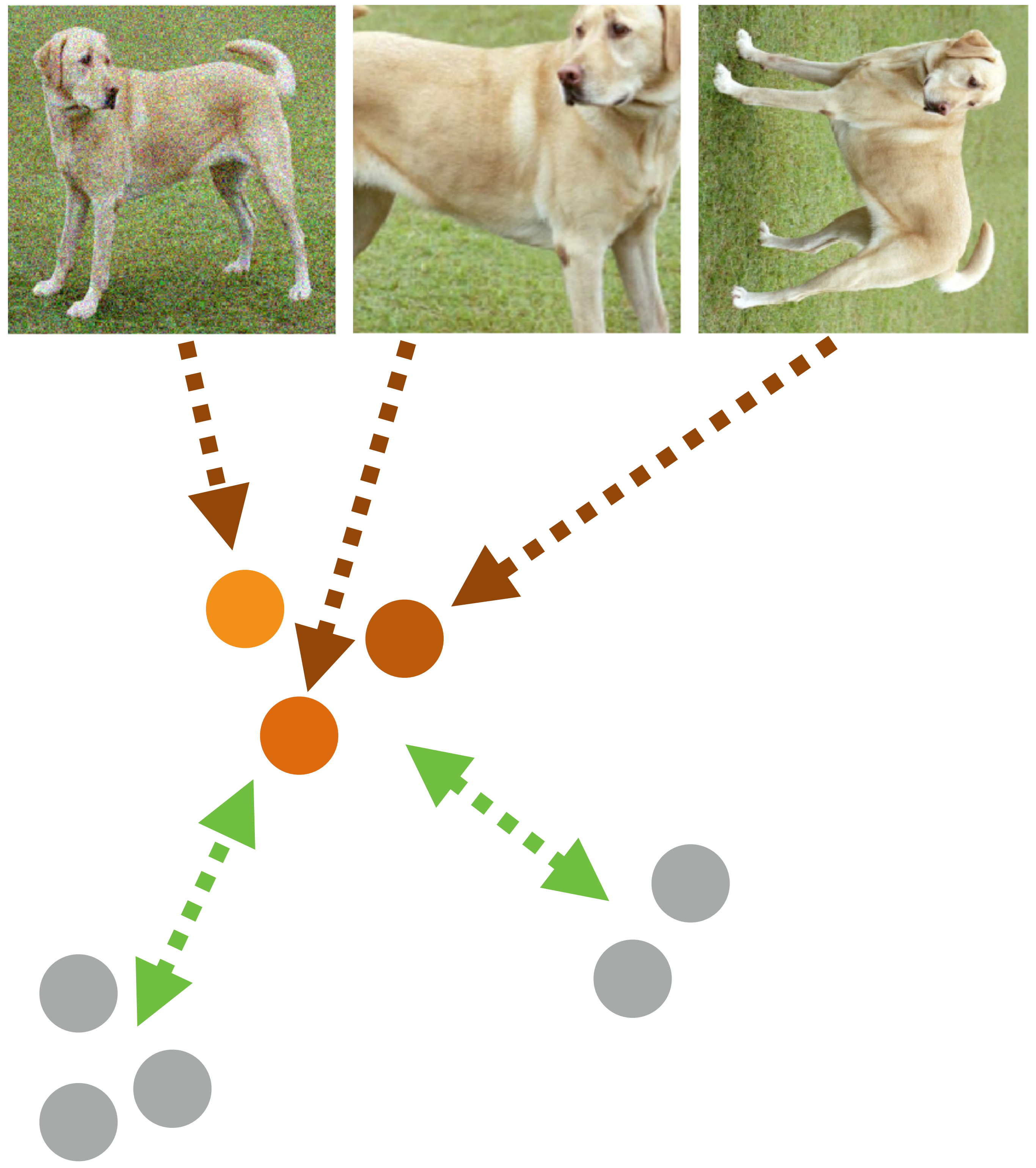
Contrastive Learning, Instance Classification and Data Augmentation



- **Aligning Self-SL**: Pretext-task explicitly defines which images are similar based on data augmentation

*Exemplar-CNN, SimCLR, MoCo, Deep Clustering, SeLa, SwAV
Noroozi et al, Representation Learning by Learning to Count, 2017
Wang and Gupta, Unsupervised Learning of Visual Representations Using Videos, 2015

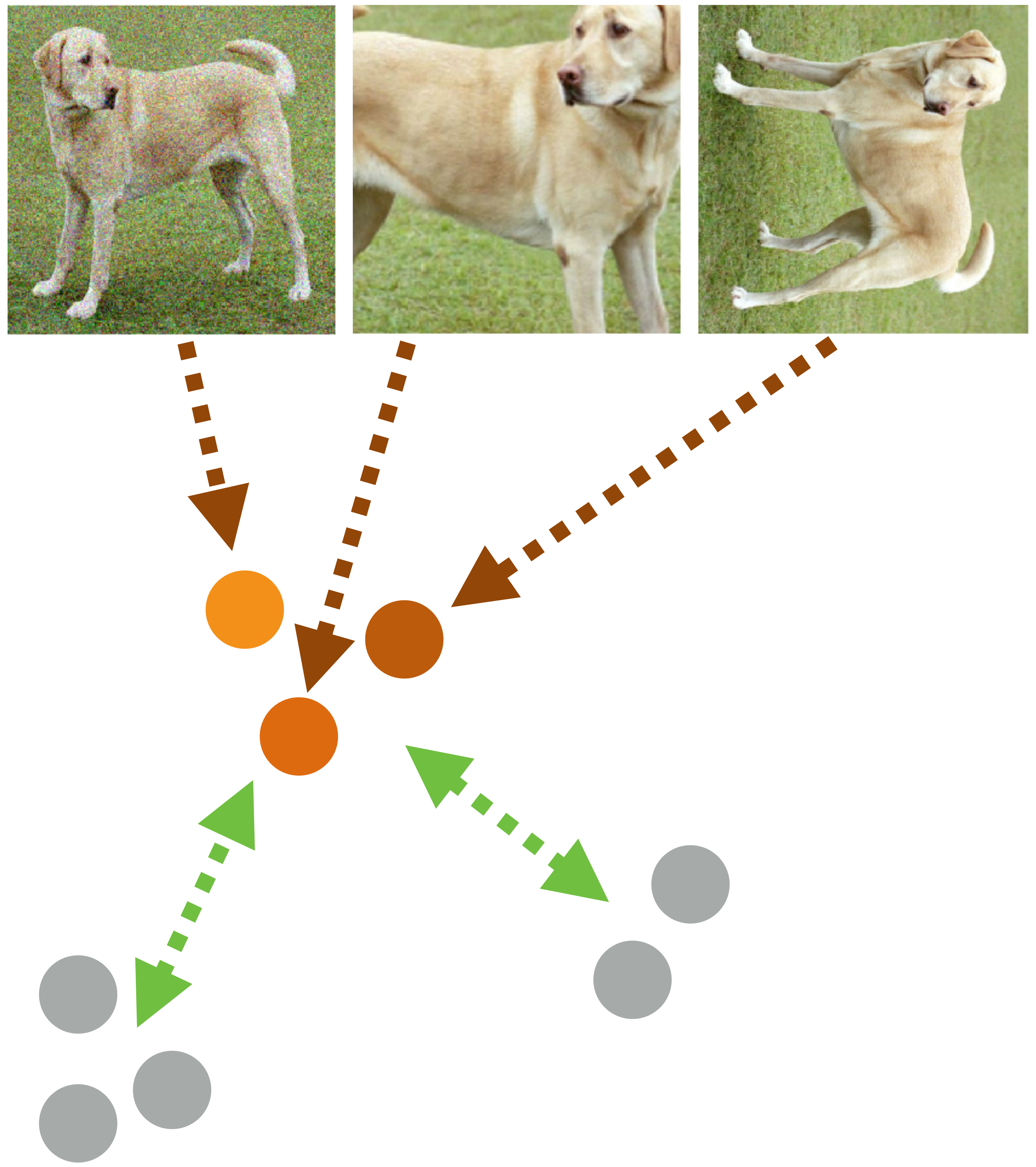
Contrastive Learning, Instance Classification and Data Augmentation



- **Aligning Self-SL**: Pretext-task explicitly defines which images are similar based on data augmentation

*Exemplar-CNN, SimCLR, MoCo, Deep Clustering, SeLa, SwAV
Noroozi et al, Representation Learning by Learning to Count, 2017
Wang and Gupta, Unsupervised Learning of Visual Representations Using Videos, 2015

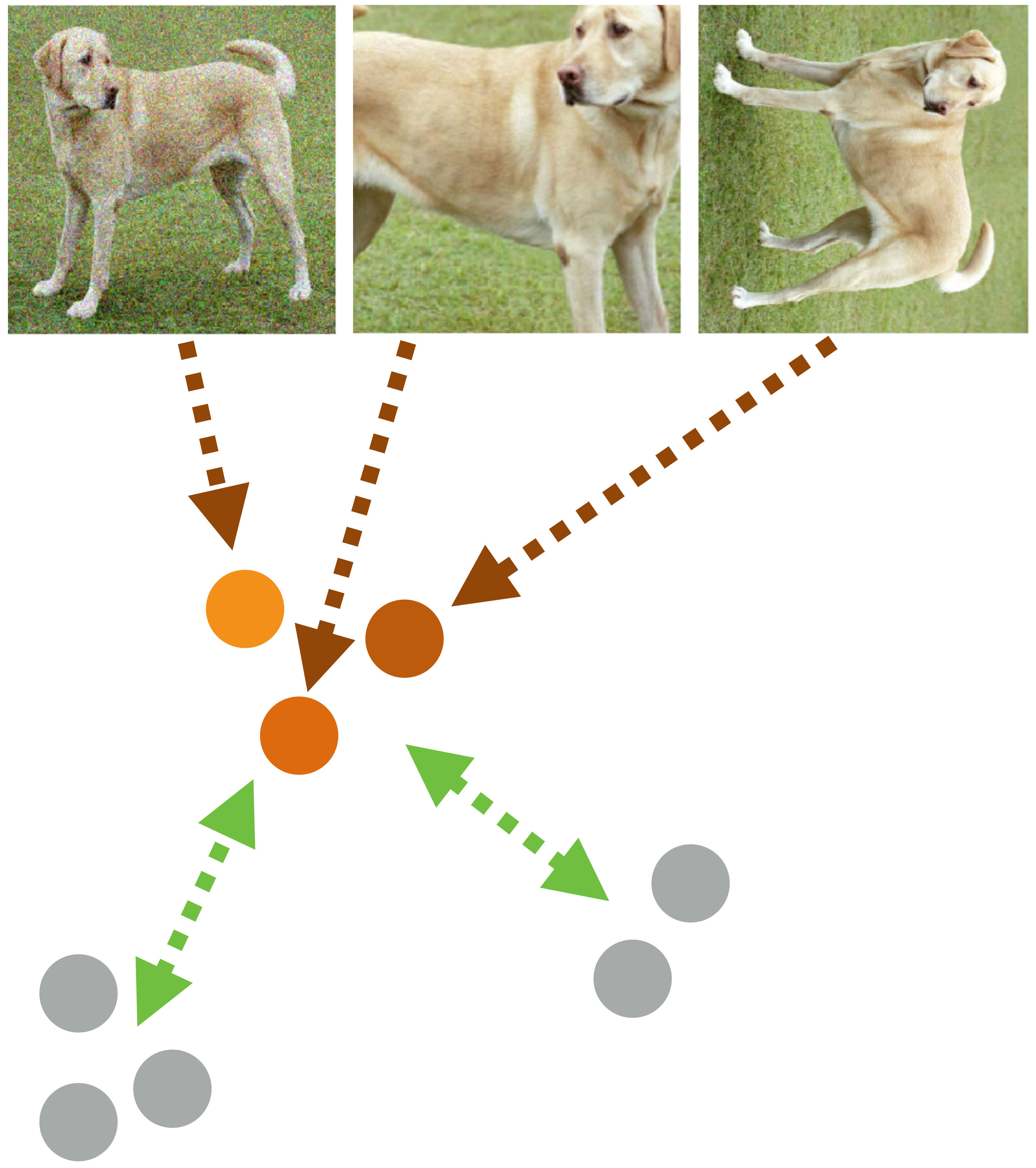
Contrastive Learning, Instance Classification and Data Augmentation



- **Aligning Self-SL**: Pretext-task explicitly defines which images are similar based on data augmentation
- Other mechanisms to separate features (e.g., entropy or simply separate each instance from all other data)

*Exemplar-CNN, SimCLR, MoCo, Deep Clustering, SeLa, SwAV
Noroozi et al, Representation Learning by Learning to Count, 2017
Wang and Gupta, Unsupervised Learning of Visual Representations Using Videos, 2015

Contrastive Learning, Instance Classification and Data Augmentation



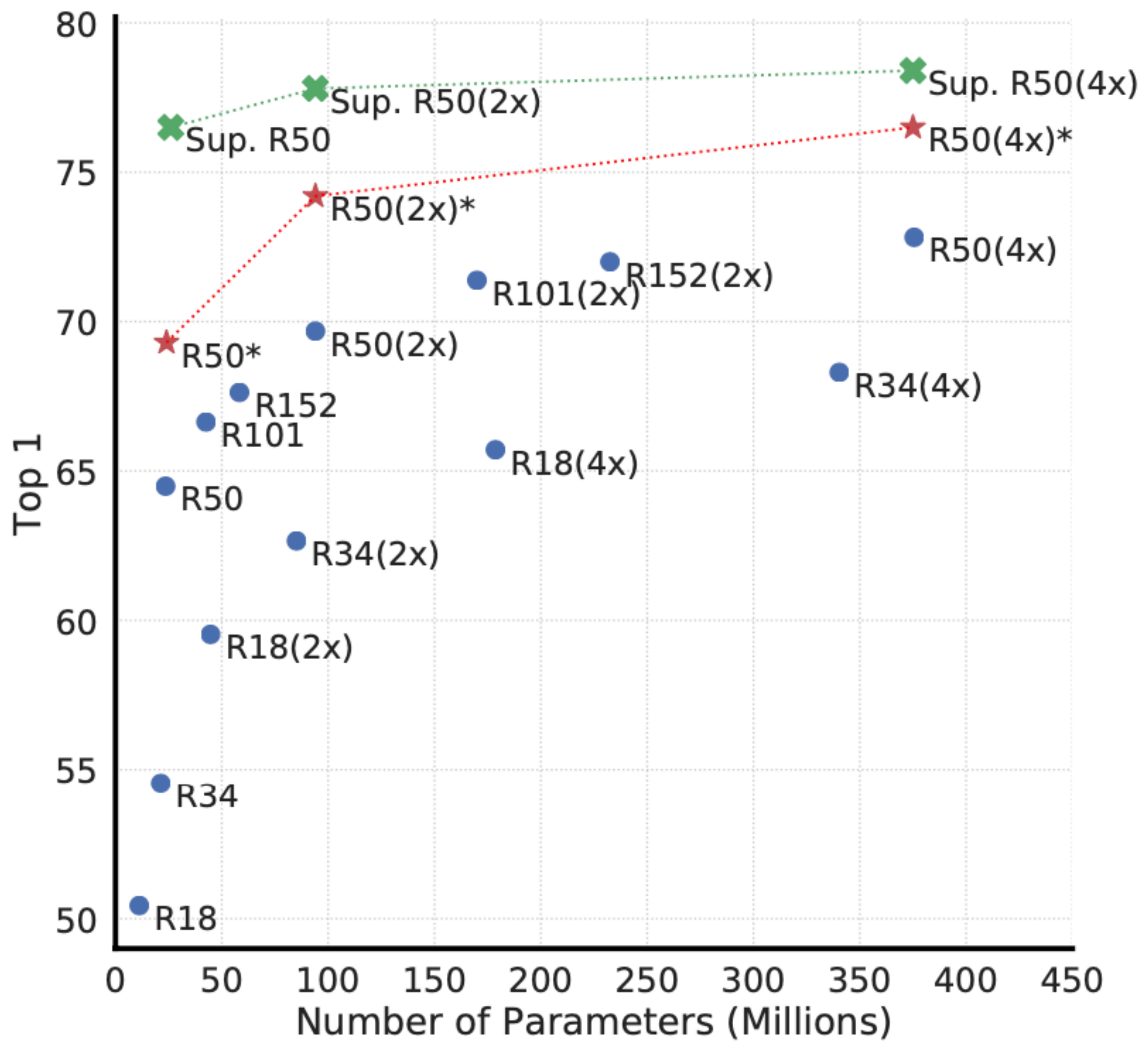
- **Aligning Self-SL**: Pretext-task explicitly defines which images are similar based on data augmentation
- Other mechanisms to separate features (e.g., entropy or simply separate each instance from all other data)
- Network and optimization design provide non trivial performance boost (e.g., large minibatches, contrastive learning, additional network "head")

*Exemplar-CNN, SimCLR, MoCo, Deep Clustering, SeLa, SwAV
Noroozi et al, Representation Learning by Learning to Count, 2017
Wang and Gupta, Unsupervised Learning of Visual Representations Using Videos, 2015

Impact of the Model Architecture

- The network architecture (hierarchy) is also important

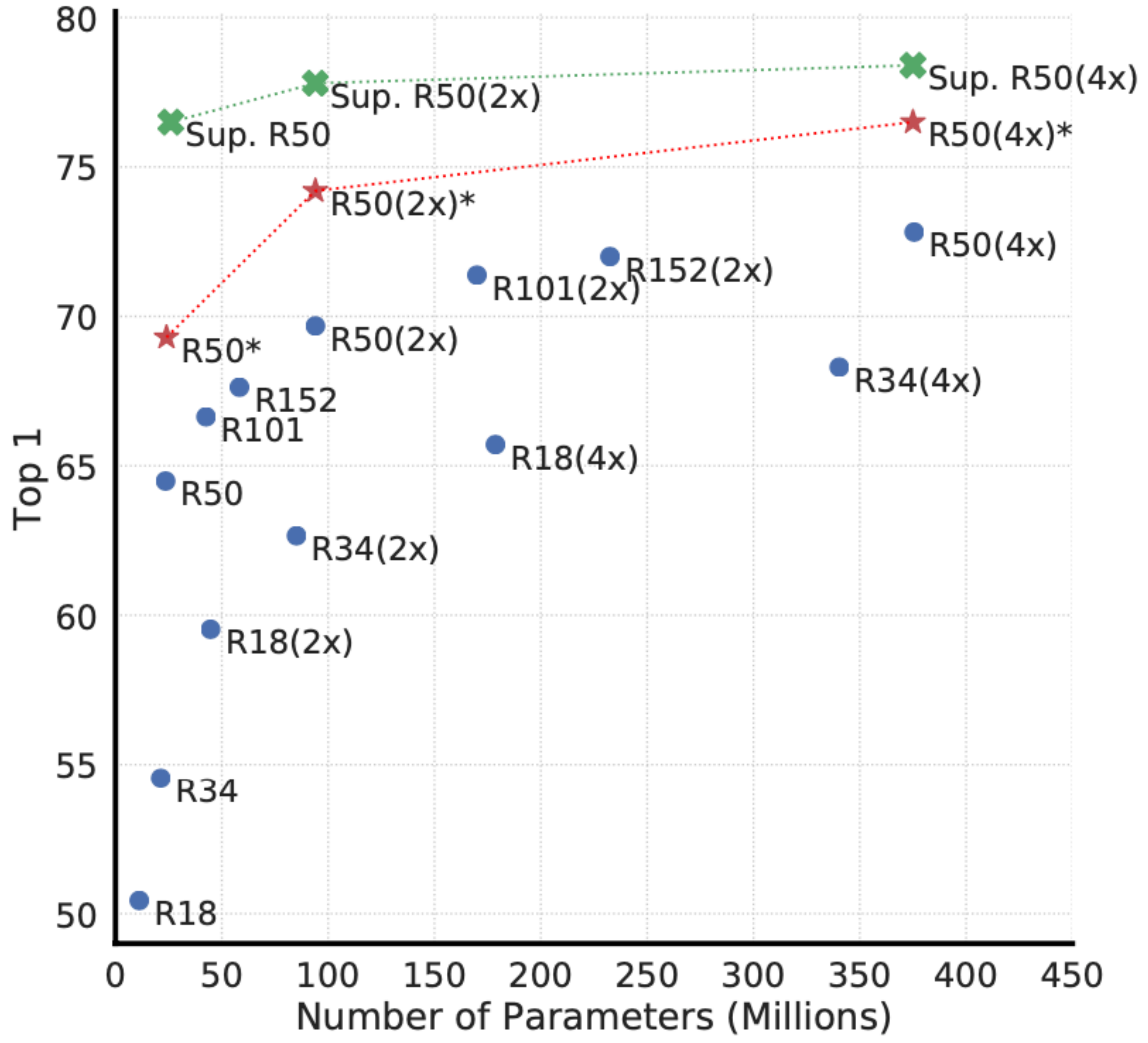
- SimCLR seems to benefit from deep models and long training more than SL



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

Impact of the Model Architecture

- The network architecture (hierarchy) is also important
- SimCLR seems to benefit from deep models and long training more than SL



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

Impact of Optimization and Cost Functions

- As shown in the lottery ticket papers* how we handle the weight initialization may play a big role in the final features

*Frankle and Carbin, The Lottery Ticket Hypothesis: Finding Sparse, Trainable Neural Networks, 2015

Impact of Optimization and Cost Functions

- As shown in the lottery ticket papers* how we handle the weight initialization may play a big role in the final features

*Frankle and Carbin, The Lottery Ticket Hypothesis: Finding Sparse, Trainable Neural Networks, 2015

Impact of Optimization and Cost Functions

- As shown in the lottery ticket papers* how we handle the weight initialization may play a big role in the final features
  - For example, SeLa or DeepCluster might favor a better optimization of the weights by adapting to the initialization of the network

*Frankle and Carbin, The Lottery Ticket Hypothesis: Finding Sparse, Trainable Neural Networks, 2015

Impact of Optimization and Cost Functions

- As shown in the lottery ticket papers* how we handle the weight initialization may play a big role in the final features
  - For example, SeLa or DeepCluster might favor a better optimization of the weights by adapting to the initialization of the network
- Losses alone may also play a role (see e.g., the comparison by Khosla et al, *Supervised Contrastive Learning*, 2020 and analysis of Wang and Isola, *Understanding Contrastive Representation Learning through Alignment and Uniformity on the Hypersphere*, 2020)

*Frankle and Carbin, The Lottery Ticket Hypothesis: Finding Sparse, Trainable Neural Networks, 2015

Conclusions

- SelfSL has made a drastic progress and now shows already better performance than SL pretraining in several transfer tasks
- There are several factors that seem to influence the quality of learned features: pretext-task, neural network model, choice of losses, and training settings
- We also should probably move away from comparing to supervised learning features as they may not be the golden standard (e.g., mid-range attributes)
- Probably a combination of both discriminative and aligning principles (through multi task learning) is a plausible direction (see also Feng et al, 2019)

Conclusions

- SelfSL has made a drastic progress and now shows already better performance than SL pretraining in several transfer tasks
- There are several factors that seem to influence the quality of learned features: pretext-task, neural network model, choice of losses, and training settings
- We also should probably move away from comparing to supervised learning features as they may not be the golden standard (e.g., mid-range attributes)
- Probably a combination of both discriminative and aligning principles (through multi task learning) is a plausible direction (see also Feng et al, 2019)