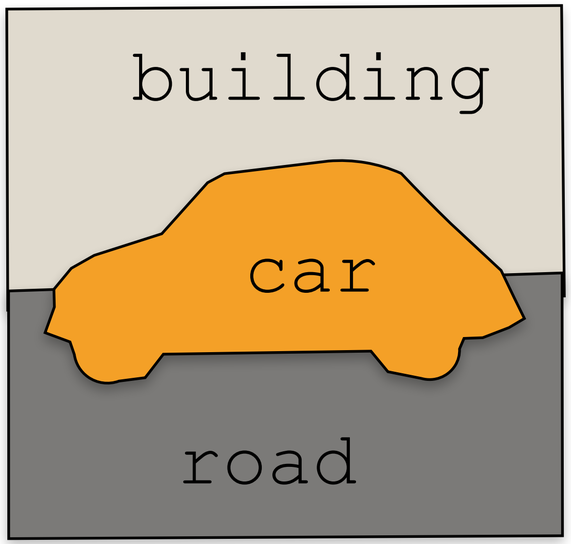
Non-Generative Representation Learning

Prof. Vincent Sitzmann

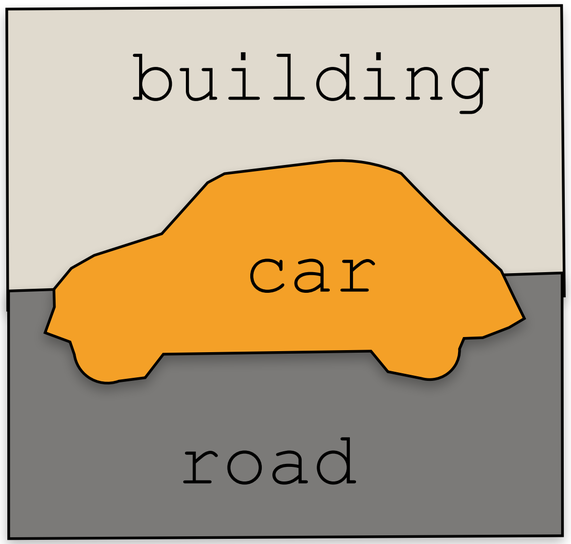
What Makes a Good Representation?



[See “Representation Learning”, Bengio 2013, for more commentary]

What Makes a Good Representation?

Good representations are:

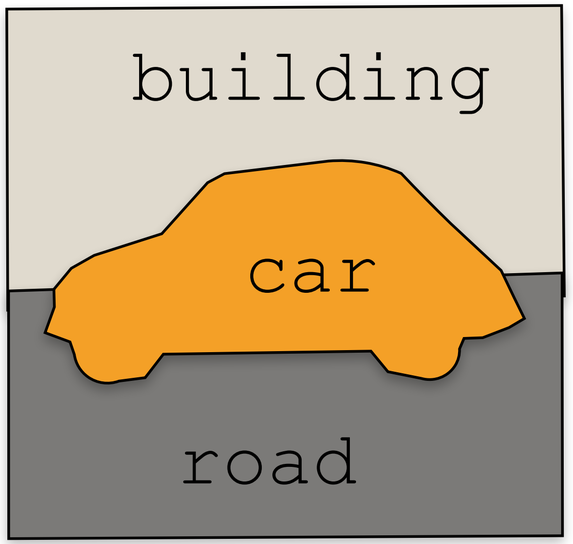


[See “Representation Learning”, Bengio 2013, for more commentary]

What Makes a Good Representation?

Good representations are:

1. Compact (*minimal*)



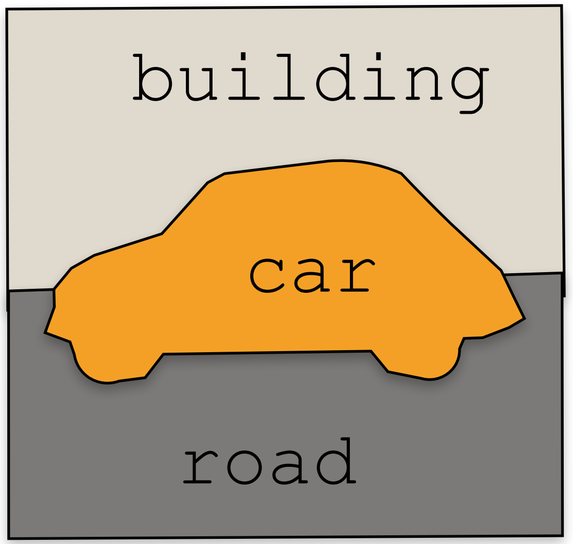
[See “Representation Learning”, Bengio 2013, for more commentary]

What Makes a Good Representation?

Good representations are:

1. Compact (*minimal*)

2. Explanatory (*sufficient*)

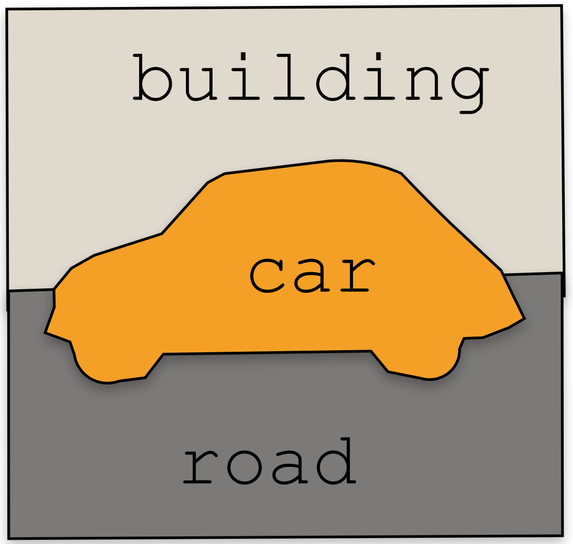


[See “Representation Learning”, Bengio 2013, for more commentary]

What Makes a Good Representation?

Good representations are:

1. Compact (*minimal*)
2. Explanatory (*sufficient*)
3. Disentangled (*independent factors*)

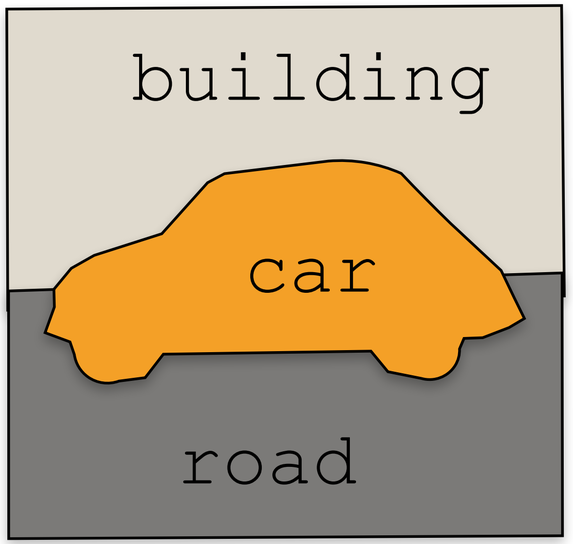


[See “Representation Learning”, Bengio 2013, for more commentary]

What Makes a Good Representation?

Good representations are:

1. Compact (*minimal*)
2. Explanatory (*sufficient*)
3. Disentangled (*independent factors*)
4. Equivariant (*transform predictably*)

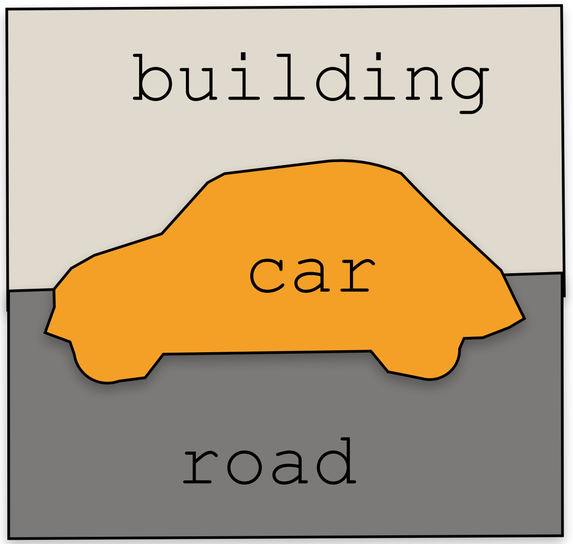


[See “Representation Learning”, Bengio 2013, for more commentary]

What Makes a Good Representation?

Good representations are:

1. Compact (*minimal*)
2. Explanatory (*sufficient*)
3. Disentangled (*independent factors*)
4. Equivariant (*transform predictably*)
5. Interpretable

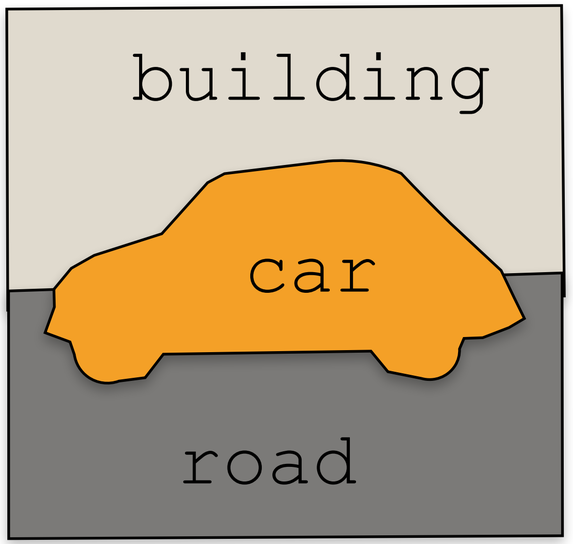


[See “Representation Learning”, Bengio 2013, for more commentary]

What Makes a Good Representation?

Good representations are:

1. Compact (*minimal*)
2. Explanatory (*sufficient*)
3. Disentangled (*independent factors*)
4. Equivariant (*transform predictably*)
5. Interpretable
6. *Make subsequent problem solving easy*

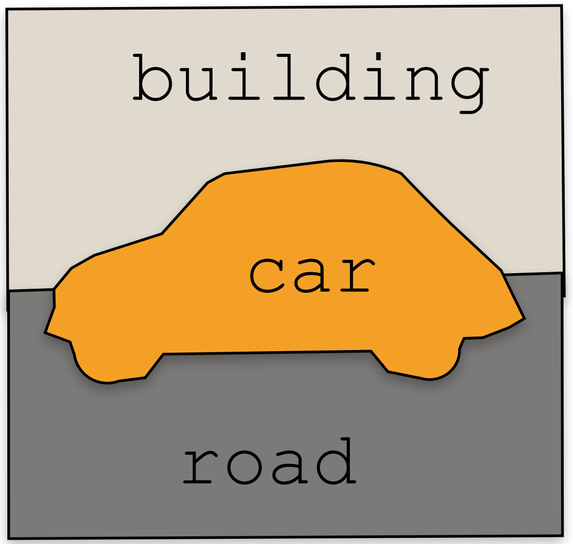


[See “Representation Learning”, Bengio 2013, for more commentary]

What Makes a Good Representation?

Good representations are:

1. Compact (*minimal*)
2. Explanatory (*sufficient*)
3. Disentangled (*independent factors*)
4. Equivariant (*transform predictably*)
5. Interpretable
6. *Make subsequent problem solving easy*
7. …?



[See “Representation Learning”, Bengio 2013, for more commentary]

What Makes a Good Representation?

Good representations are:

1. Compact (*minimal*)
2. Explanatory (*sufficient*)
3. Disentangled (*independent factors*)
4. Equivariant (*transform predictably*)
5. Interpretable

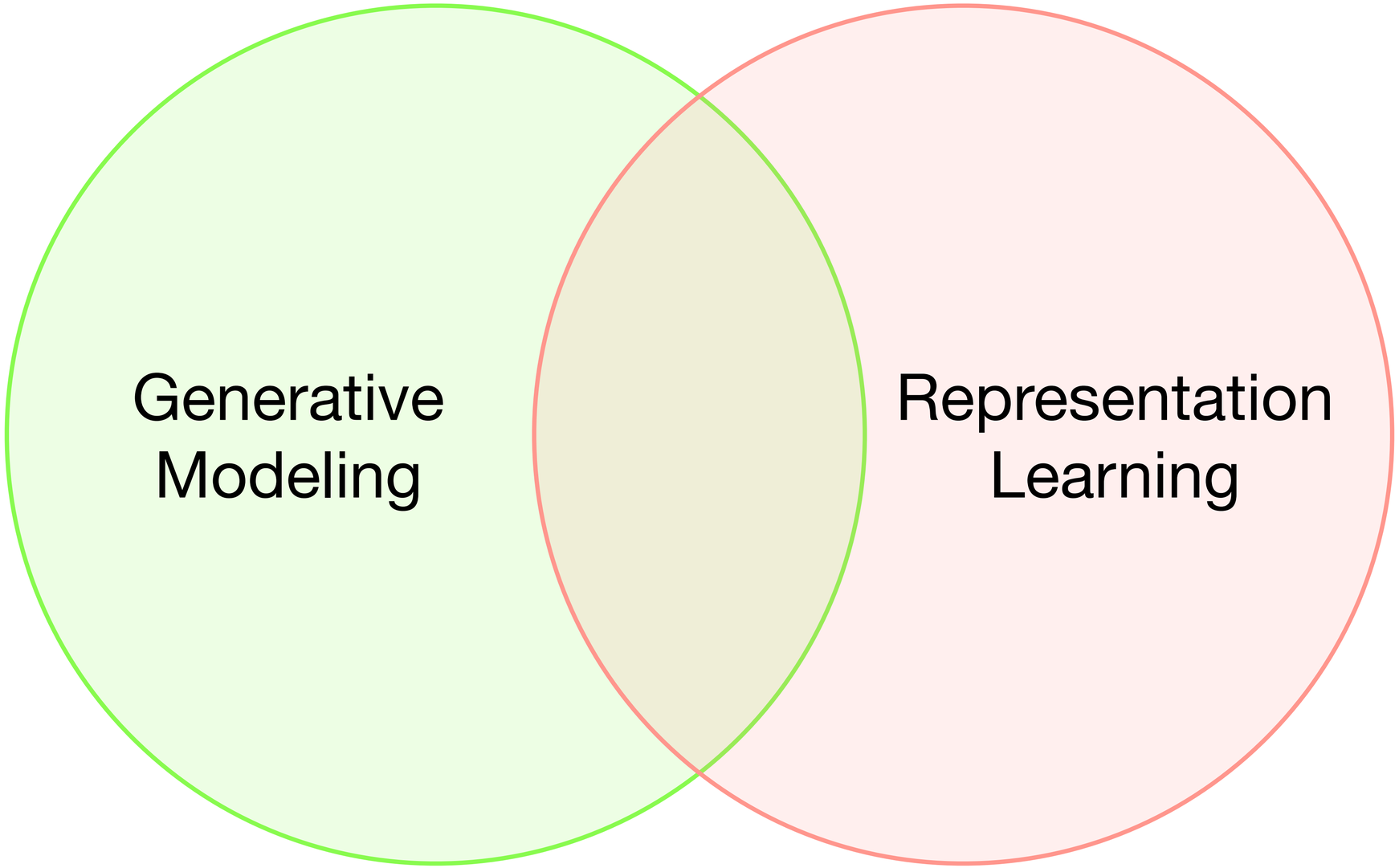
This is a tall order…

If we could do this reliably, all of 3D reconstruction would be done!

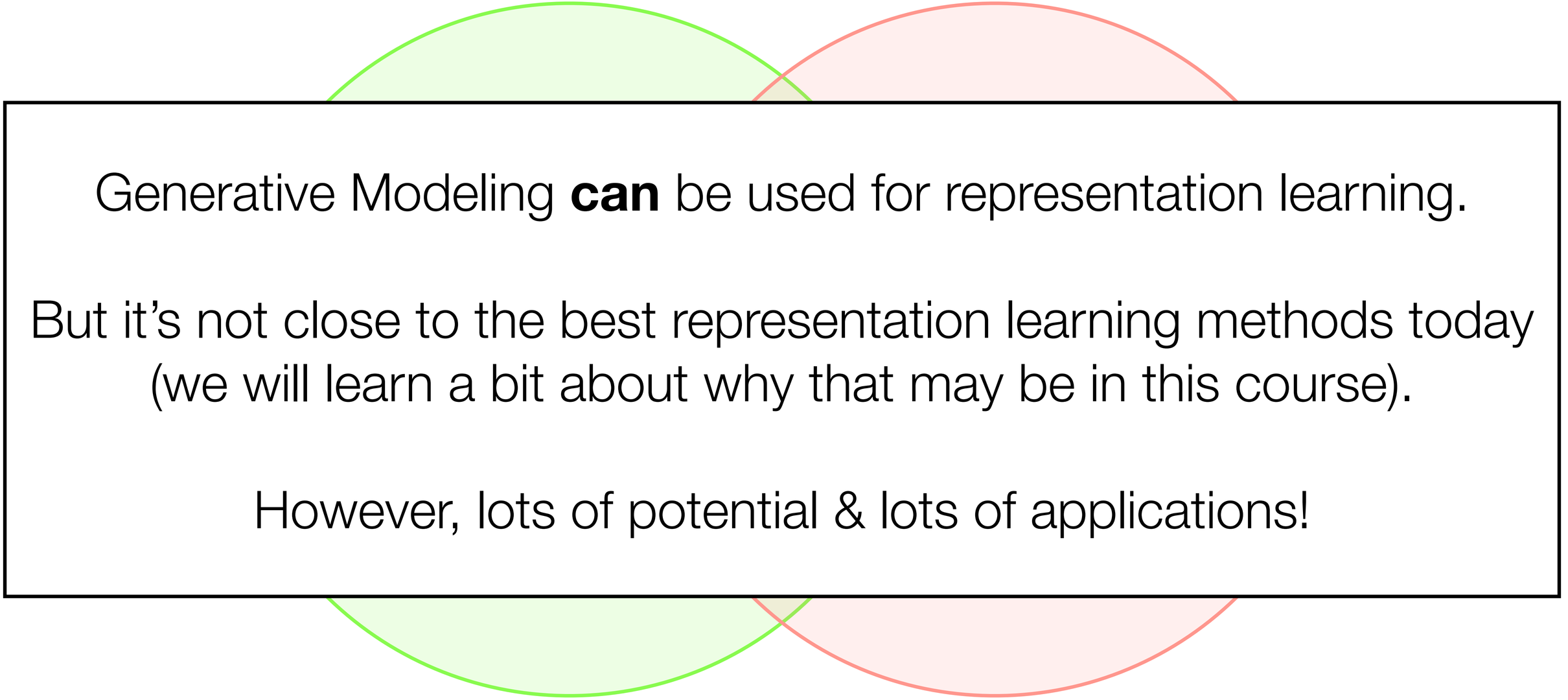
6. *Make subsequent problem solving easy*
7. …?

[See “Representation Learning”, Bengio 2013, for more commentary]

Relationship of Generative Modeling and Representation Learning



Relationship of Generative Modeling and Representation Learning

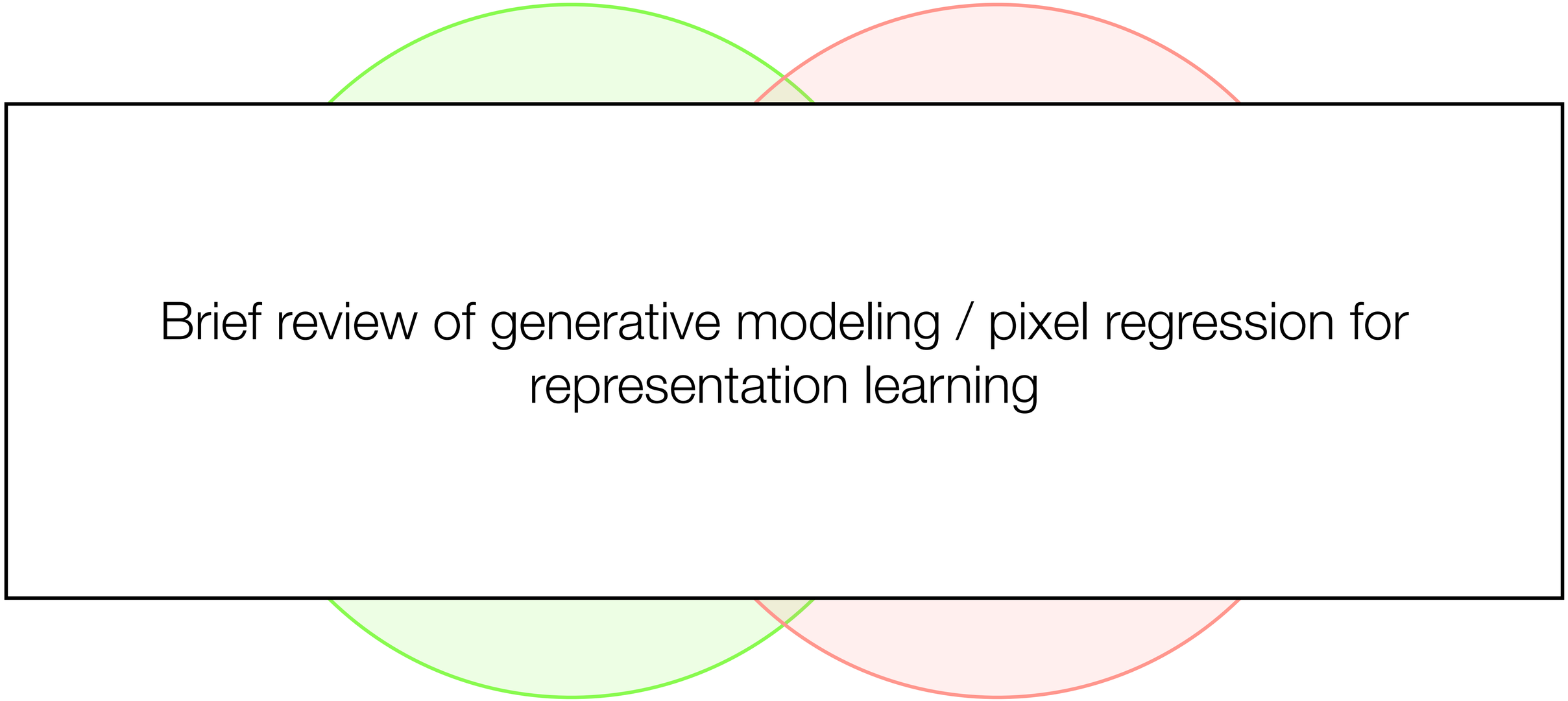


Generative Modeling **can** be used for representation learning.

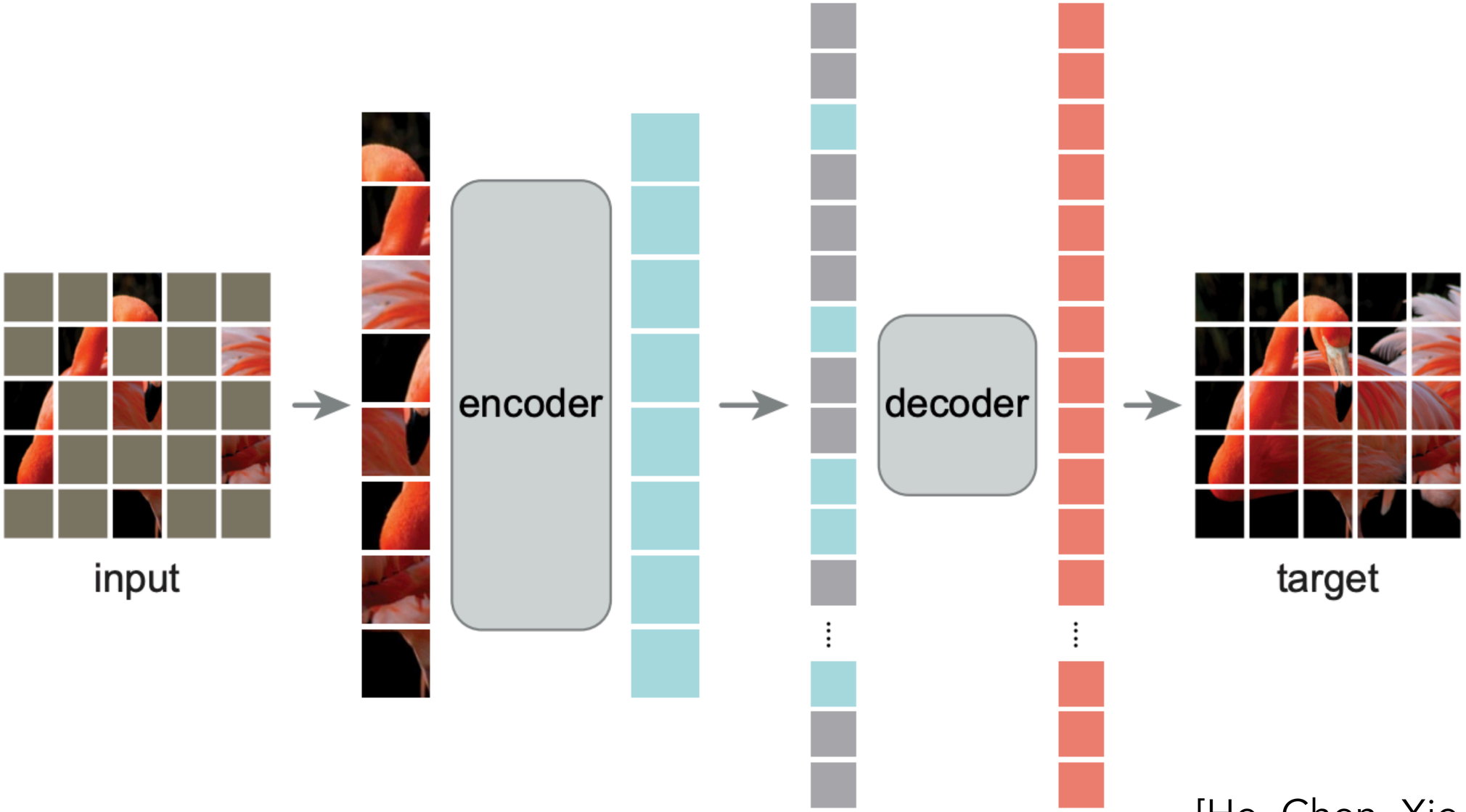
But it's not close to the best representation learning methods today (we will learn a bit about why that may be in this course).

However, lots of potential & lots of applications!

Relationship of Generative Modeling and Representation Learning

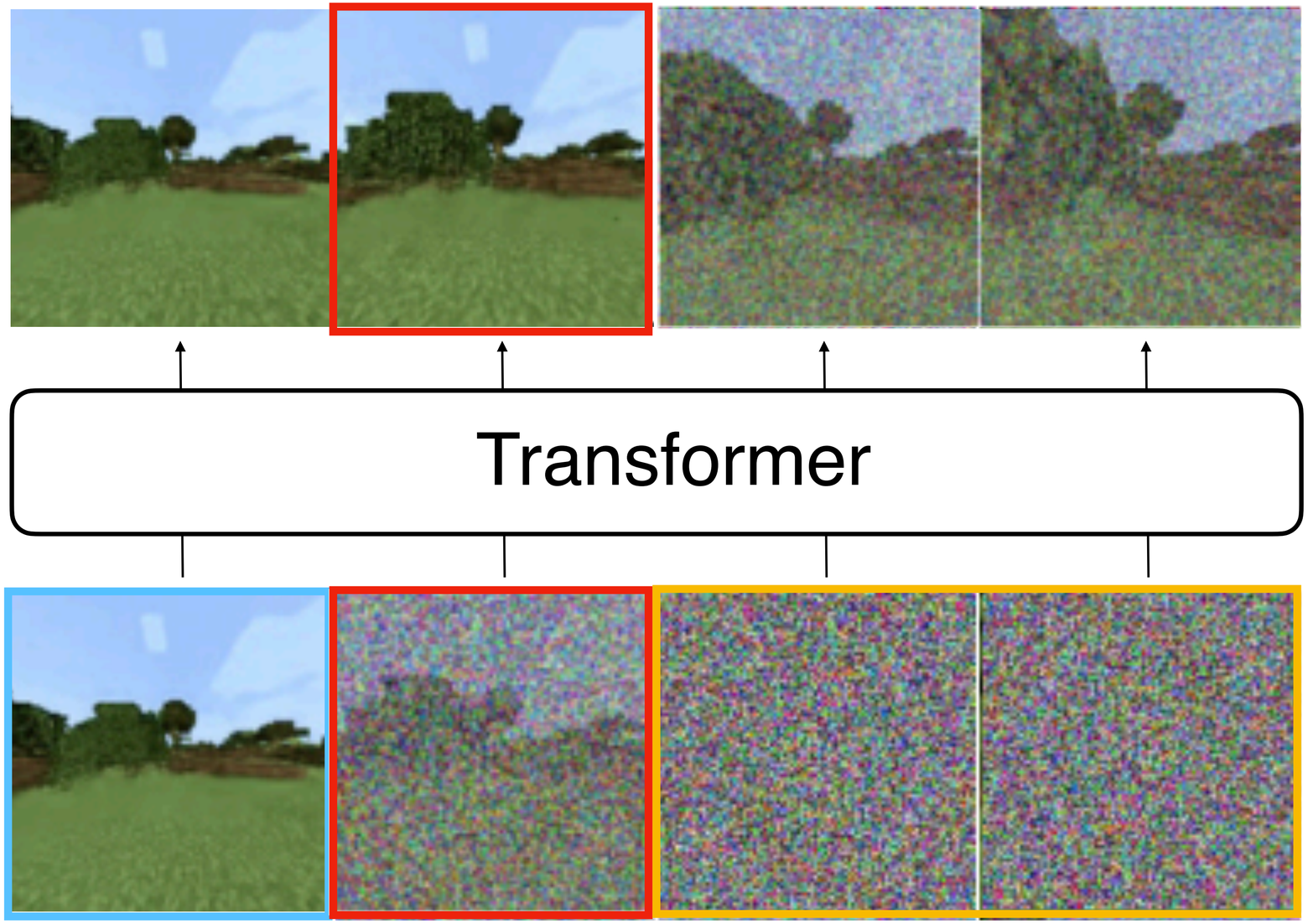


Masked Autoencoder (MAE)



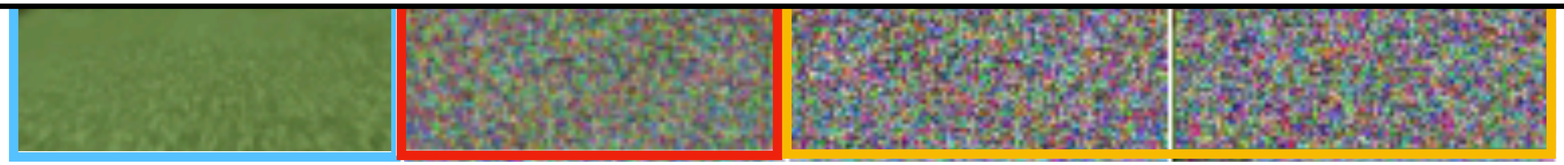
[He, Chen, Xie, et al. 2021]

Sequence Generation as a Form of Masked Auto-Encoding



Sequence Generation as a Form of Masked Auto-Encoding

Benefit of pixel regression objectives:
Straightforward & stable training, easily scalable!



Pixel Regression models do learn usable features…

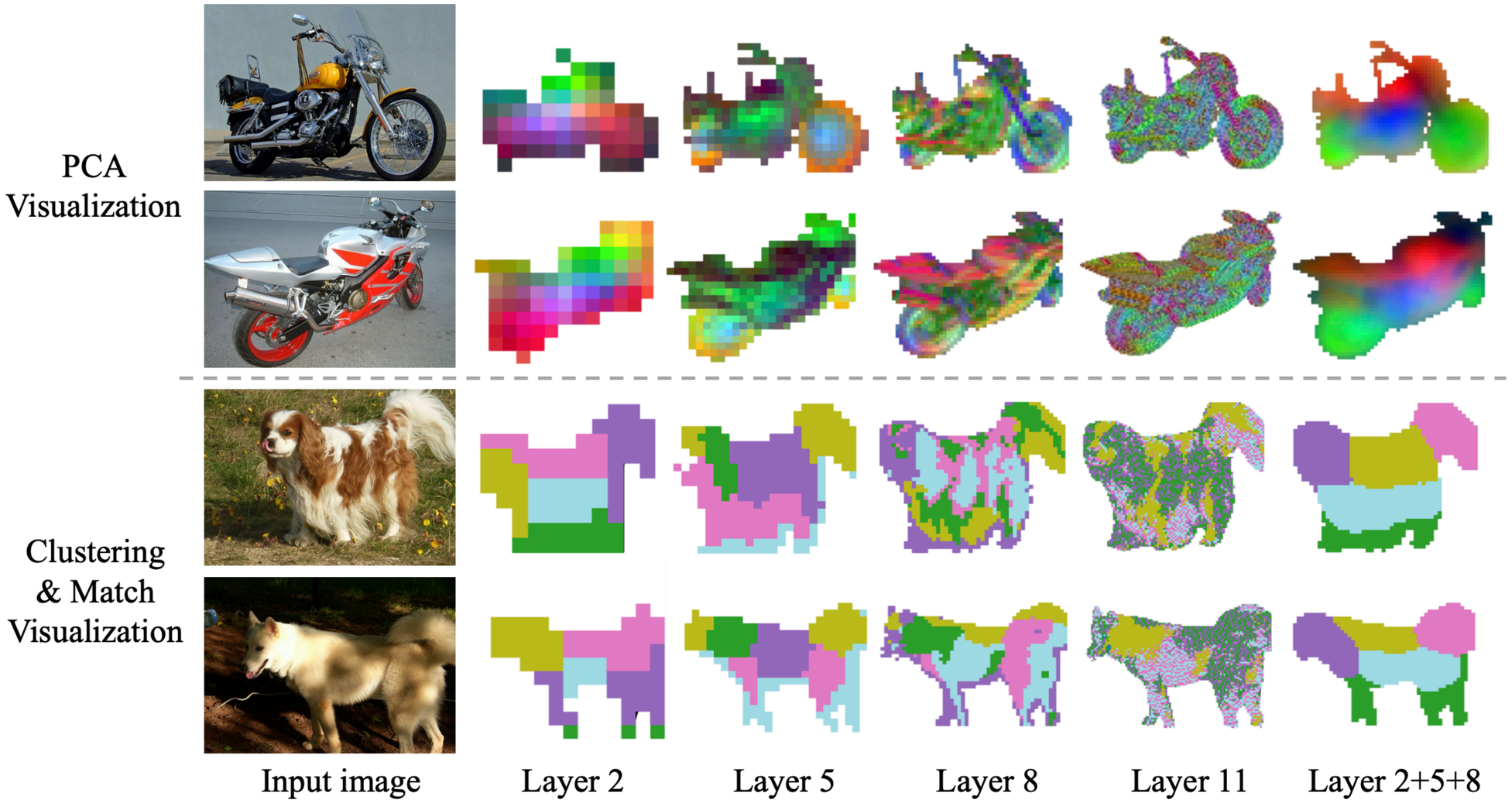


Figure 2: **Analysis of features from different decoder layers in SD.**

Zhang et al., A Tale of Two Features: Stable Diffusion Complements DINO for Zero-Shot Semantic Correspondence, 2023

…but nowhere near as good as non-“generative” counterparts

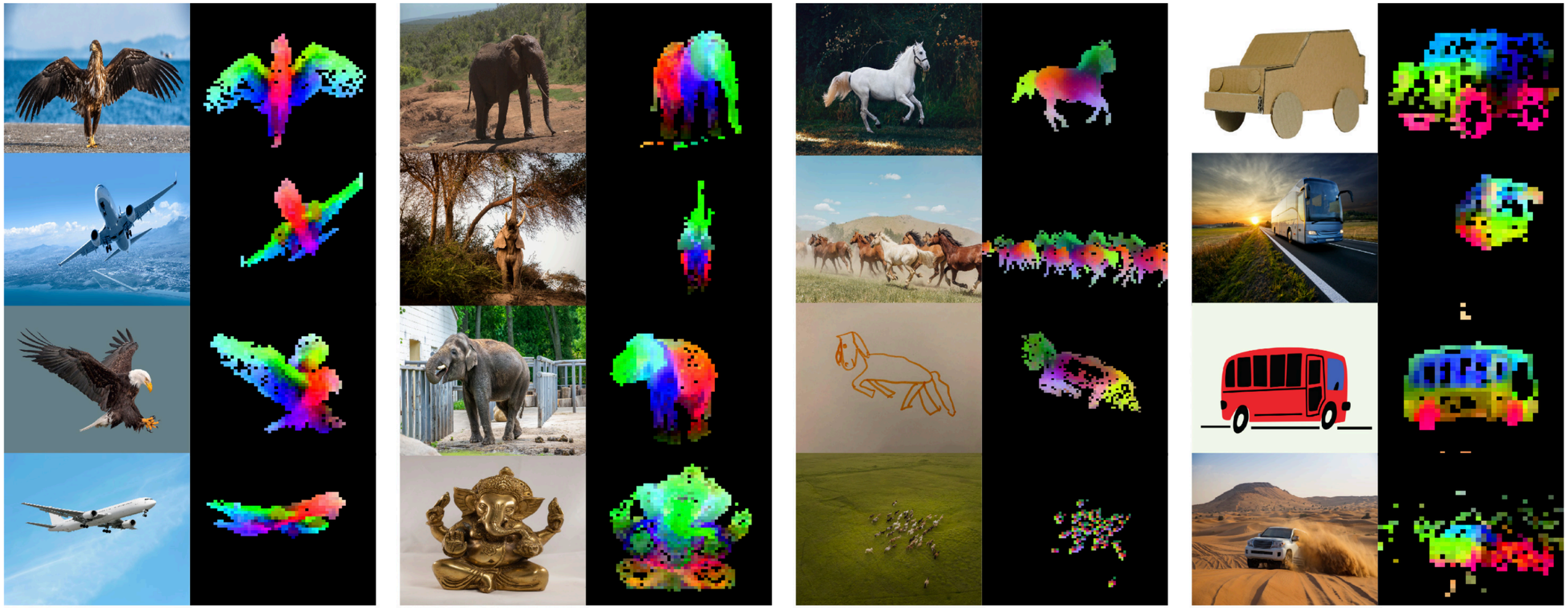
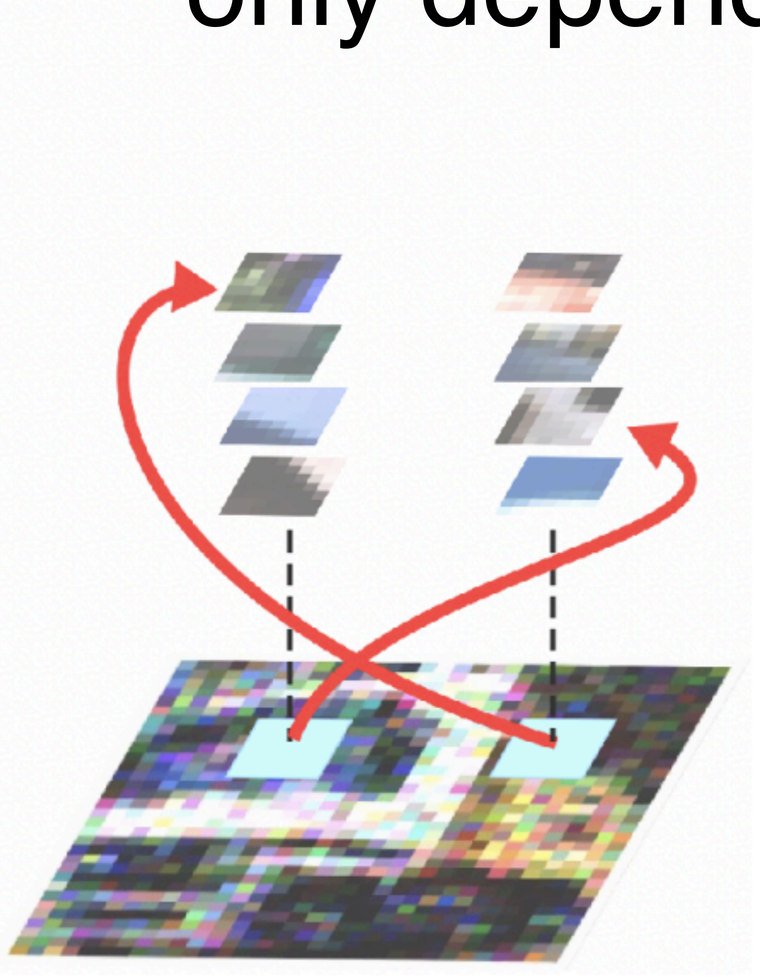


Figure 1: **Visualization of the first PCA components.** We compute a PCA between the patches of the images from the same column (a, b, c and d) and show their first 3 components. Each component is matched to a different color channel. Same parts are matched between related images despite changes of pose, style or even objects. Background is removed by thresholding the first PCA component.

From DinoV2: Learning Robust Visual Features without Supervision, Oquab et al. 2024

Why? Recall “analytical” diffusion model: Diffusion only depends on L2 similarities of patches



An analytic theory of creativity in convolutional diffusion models

Mason Kamb [1] **Surya Ganguli** [1]

Abstract

We obtain the first analytic, interpretable and predictive theory of creativity in convolutional diffusion models. Indeed, score-based diffusion models can generate highly creative images that lie far from their training data. But optimal score-matching theory suggests that these models should only be able to produce memorized training examples. To reconcile this theory-experiment gap, we identify two simple inductive biases, locality and equivariance, that: (1) induce a form of combinatorial creativity by preventing optimal score-matching; (2) result in a fully analytic, completely mechanistically interpretable, equivariant local score (ELS) machine that, (3) without *any training* can quantitatively predict the outputs of

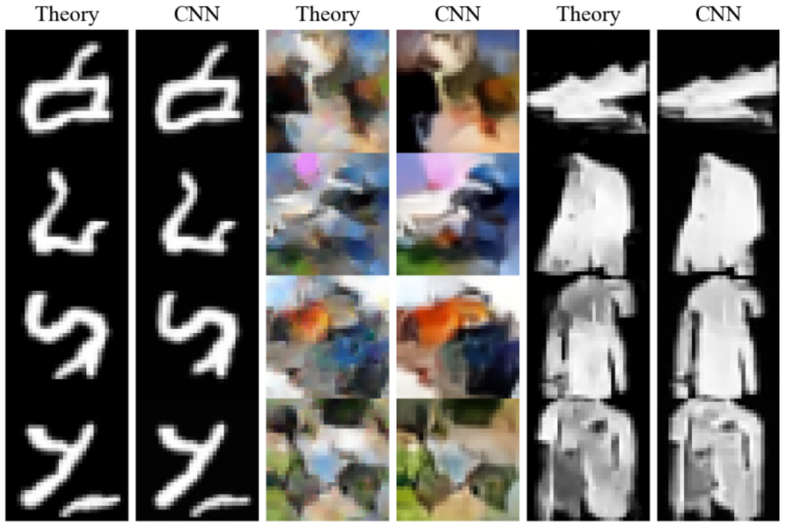
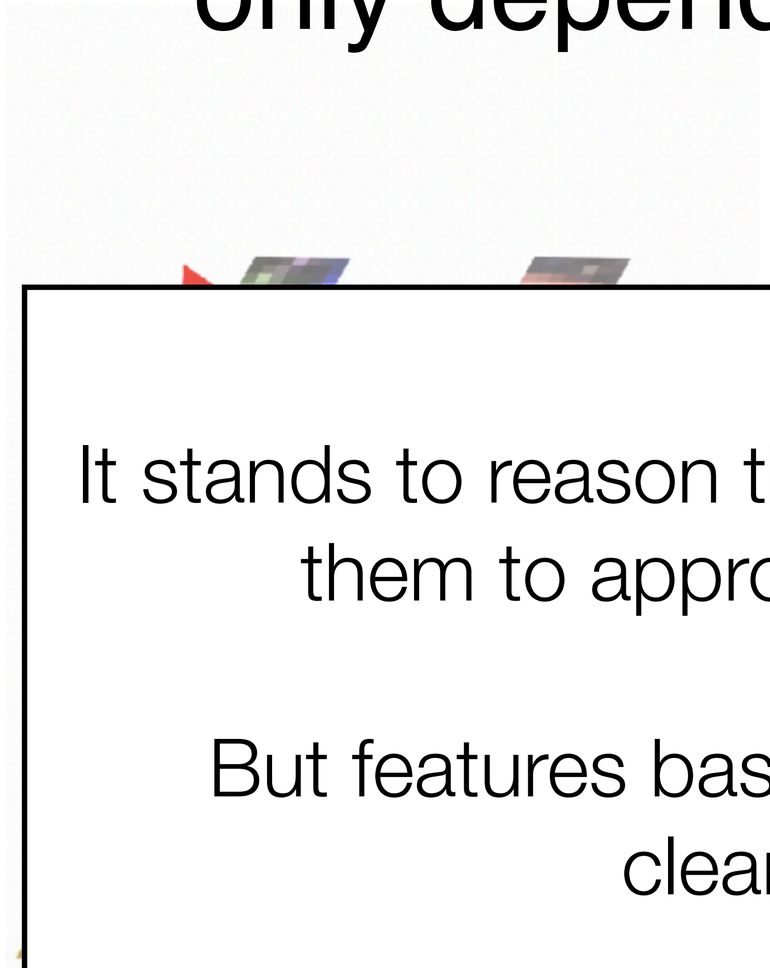


Figure 1. Our analytic theory (left columns) can accurately predict on a *case by case basis* the outputs of convolutional diffusion models (right columns), with U-Net or ResNet architectures trained

Why? Recall “analytical” diffusion model: Diffusion only depends on L2 similarities of patches



An analytic theory of creativity in convolutional diffusion models

It stands to reason that diffusion models extract features that allow them to approximate this “optimal denoiser” objective

But features based on L2 similarity of patches / images are clearly not what we generally want!

pletely mechanistically interpretable, equivariant
local score (ELS) machine that, (3) without *any training* can quantitatively predict the outputs of

Figure 1. Our analytic theory (left columns) can accurately predict on a *case by case basis* the outputs of convolutional diffusion models (right columns), with U-Net or ResNet architectures trained

I-Con: A Unifying Framework for Representation Learning

Shaden Alshammari[1] **John Hershey**[2] **Axel Feldmann**[1] **William Freeman**[1,2] **Mark Hamilton**[1,3]
[1] *MIT* [2] *Google* [3] *Microsoft*

https://aka.ms/i-con

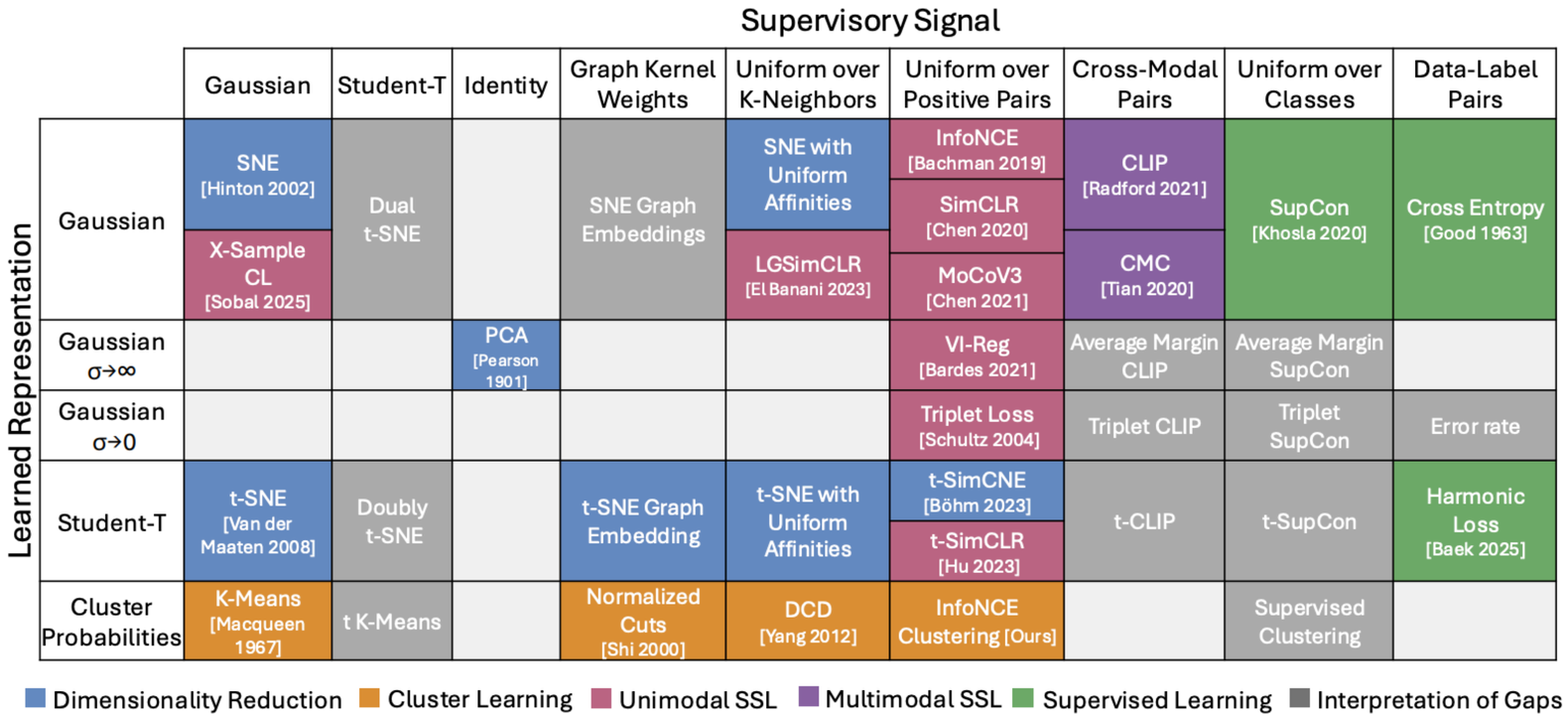


Figure 1: **A "periodic" table of representation learning methods unified by the I-Con framework.** By choosing different types of conditional probability distributions over neighbors, I-Con generalizes over 23 commonly used representation learning methods.

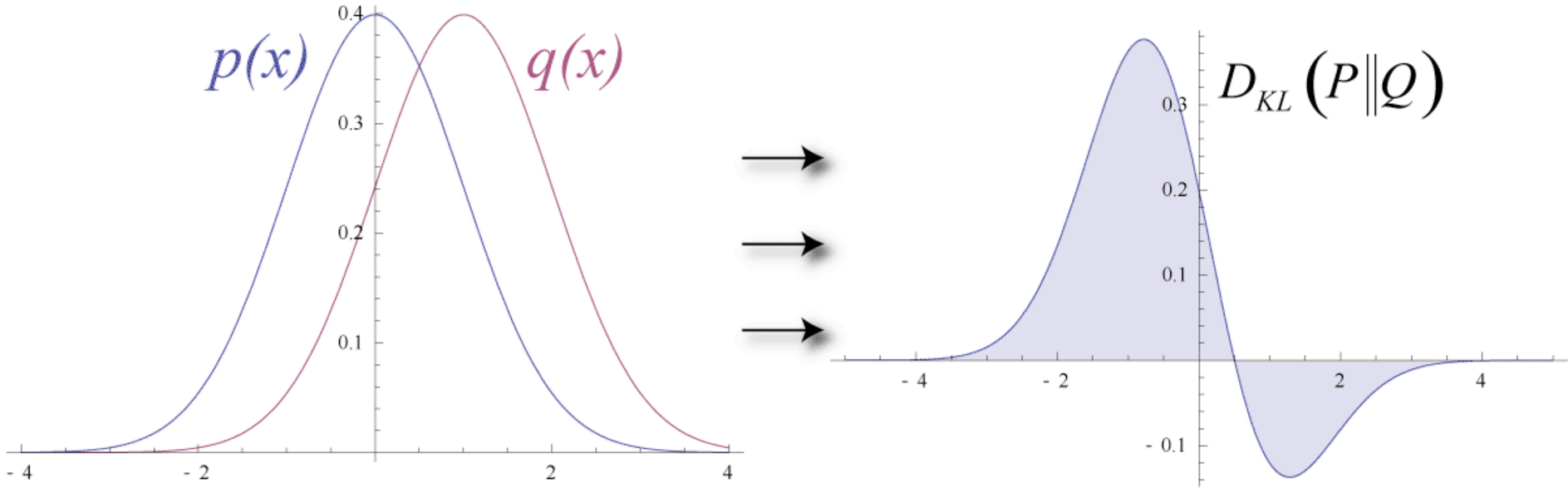
- Introduces a really cool framework that unifies a variety of joint embedding architectures
- We will be using that framework in the following slides!

Preliminaries: Kullblack-Leibler Divergence

$$D_{\mathrm{KL}}(p\|q) = \int_{\mathcal{X}} p(x)\ \log\frac{p(x)}{q(x)}\,\mathrm{d}x$$

Preliminaries: Kullblack-Leibler Divergence

$$D_{\mathrm{KL}}(p\|q) = \int_{\mathcal{X}} p(x)\,\log\frac{p(x)}{q(x)}\,\mathrm{d}x$$

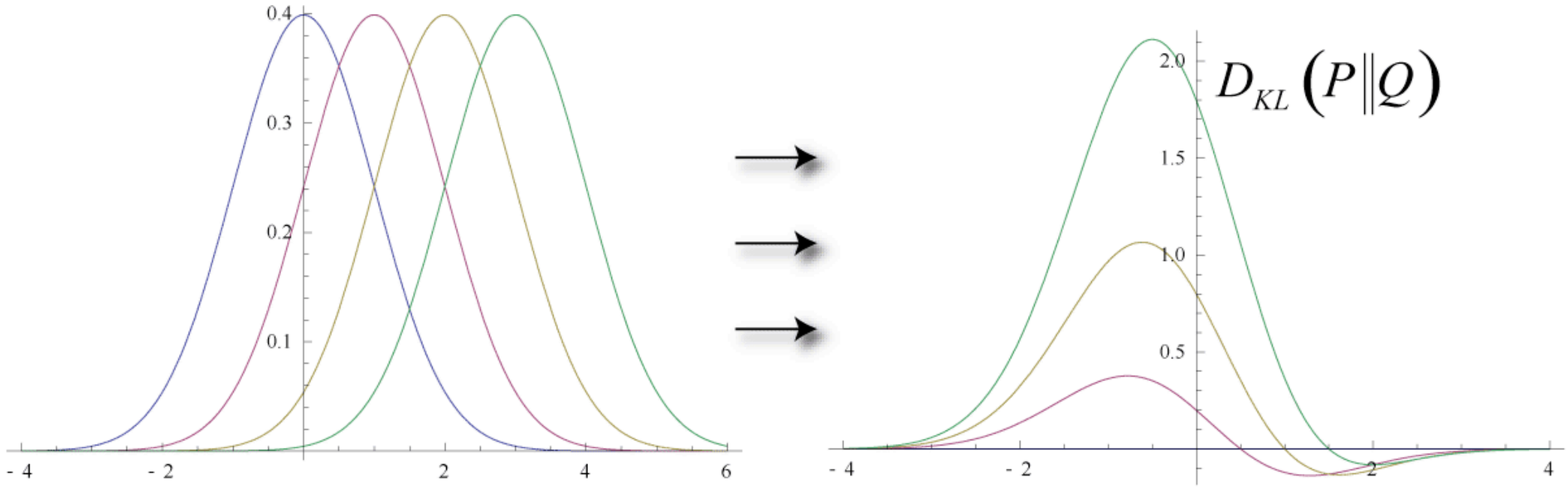


PDFs

KL-Div (Area under curve)

Preliminaries: Kullblack-Leibler Divergence

$$D_{\mathrm{KL}}(p\|q) = \int_{\mathcal{X}} p(x)\,\log\frac{p(x)}{q(x)}\,\mathrm{d}x$$

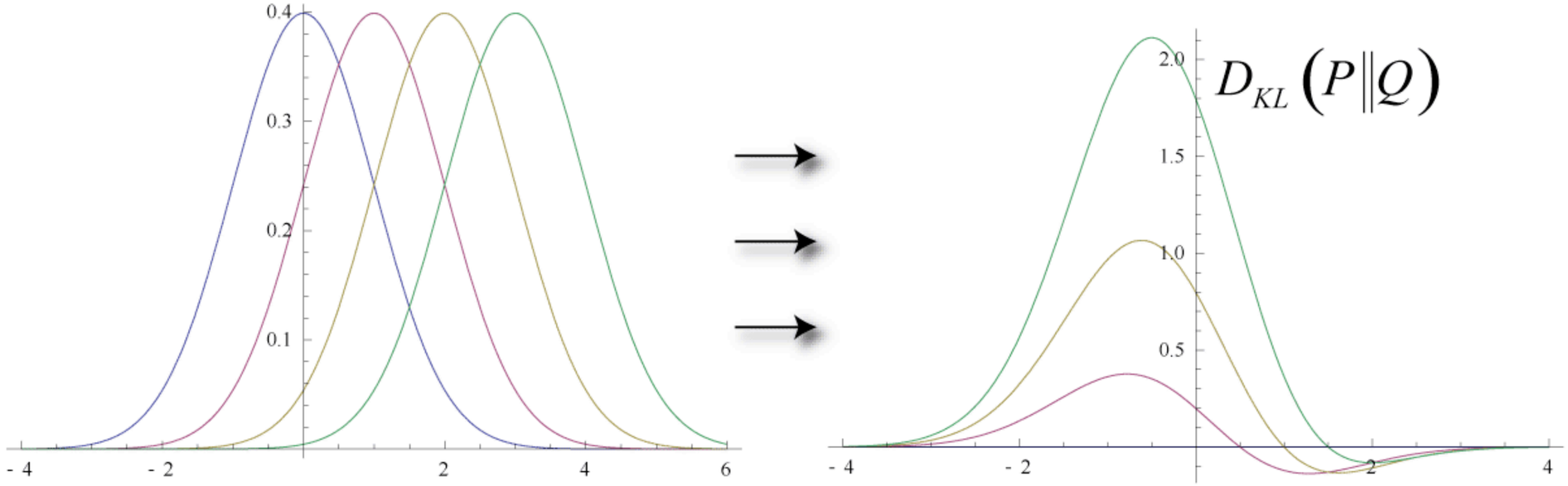


PDFs

KL-Div (Area under curve)

"Expected excess surprise" from using q as a model of p instead of p itself

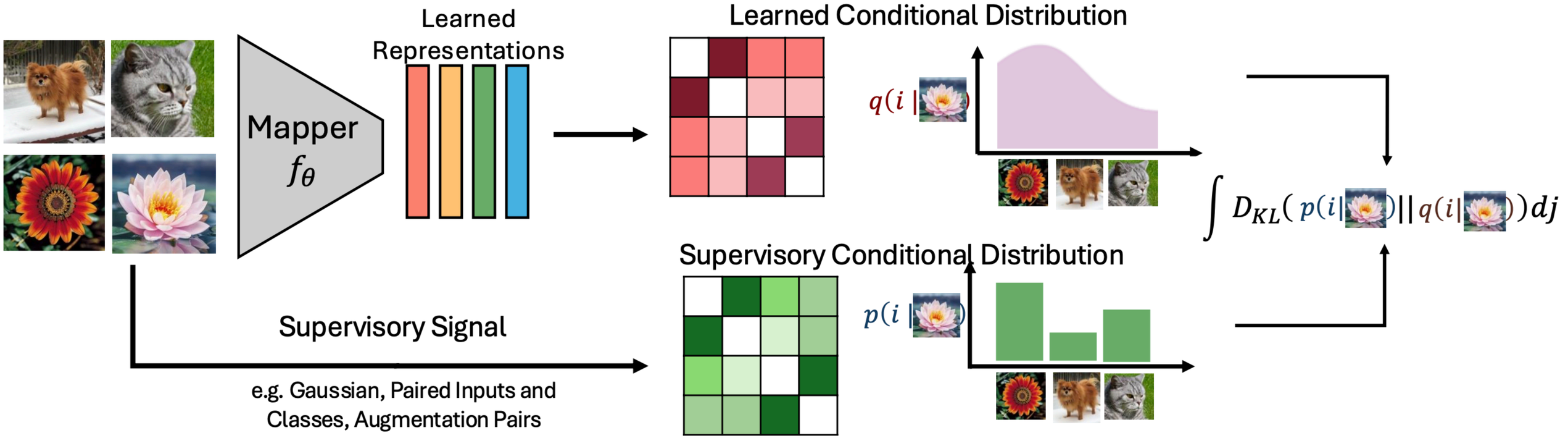
I.e., what is the expectation of drawing samples that have low likelihood under p



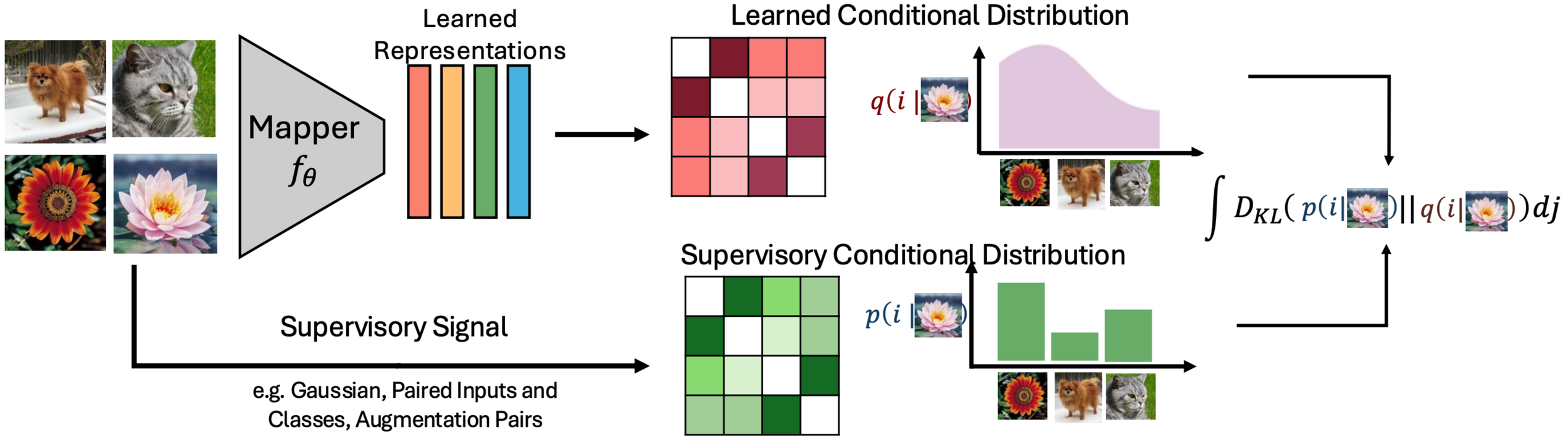
PDFs

KL-Div (Area under curve)

The I-CON Framework [Alshammari et al. 2025]



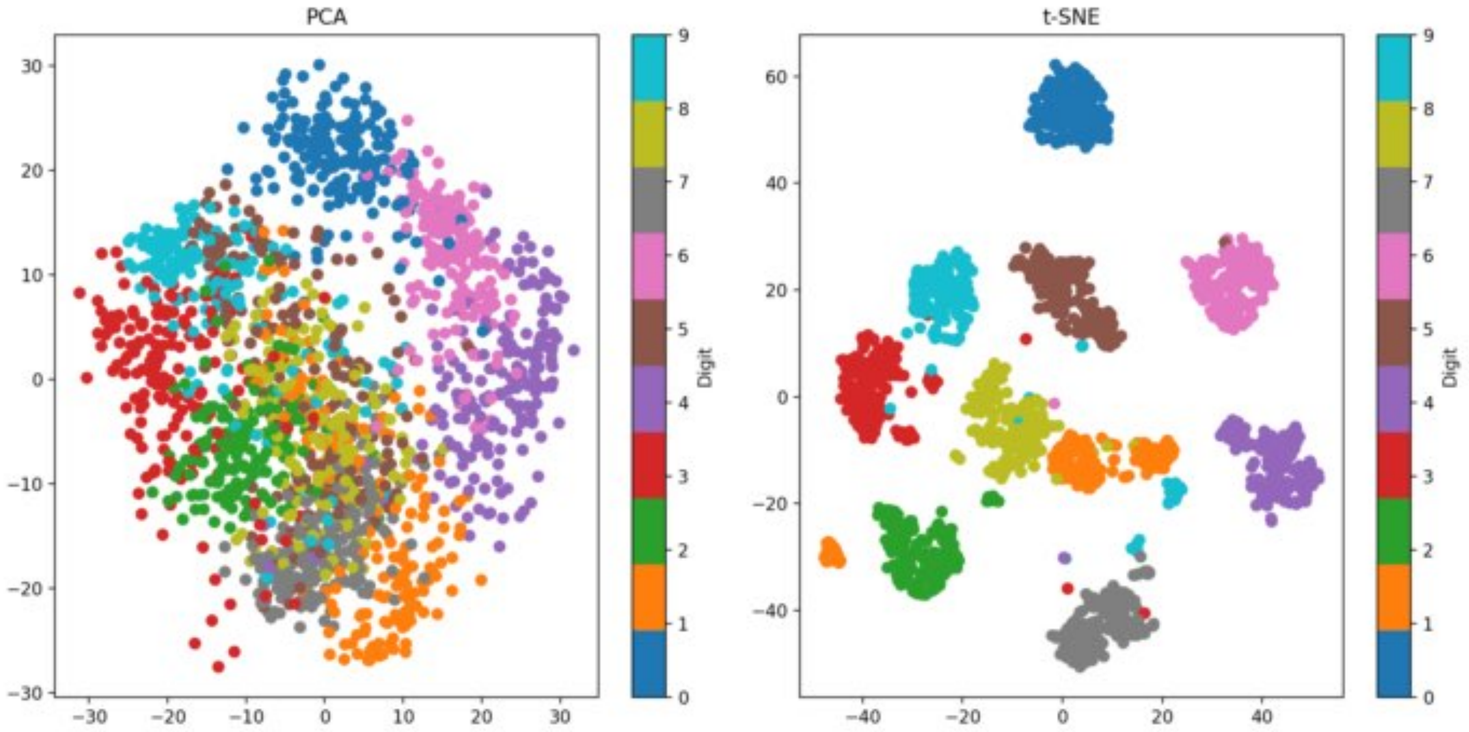
The I-CON Framework [Alshammari et al. 2025]



Casts many representation learning objectives as:

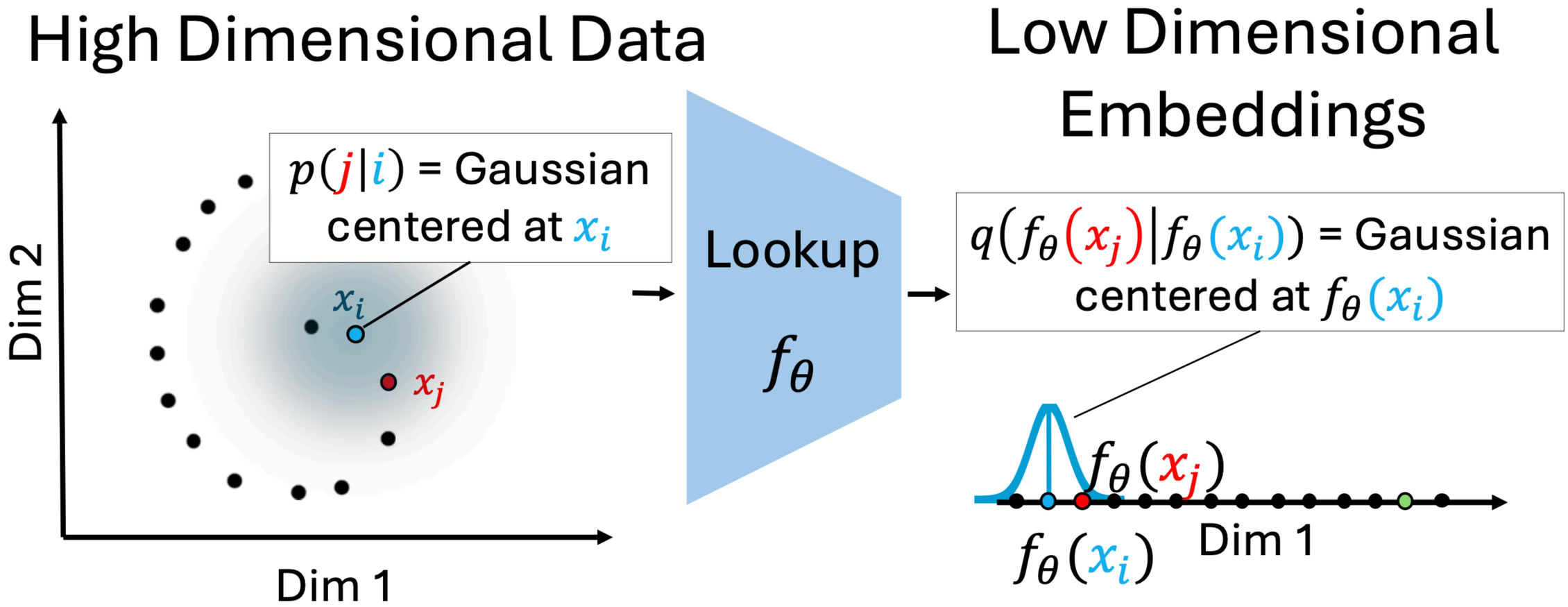
$$\mathcal{L}(\theta,\phi) = \int_{i\in\mathcal{X}} D_{\mathrm{KL}}\left(p_\theta(\cdot|i)||q_\phi(\cdot|i)\right) = \int_{i\in\mathcal{X}}\int_{j\in\mathcal{X}} p_\theta(j|i)\log\frac{p_\theta(j|i)}{q_\phi(j|i)}.$$

Dimensionality Reduction

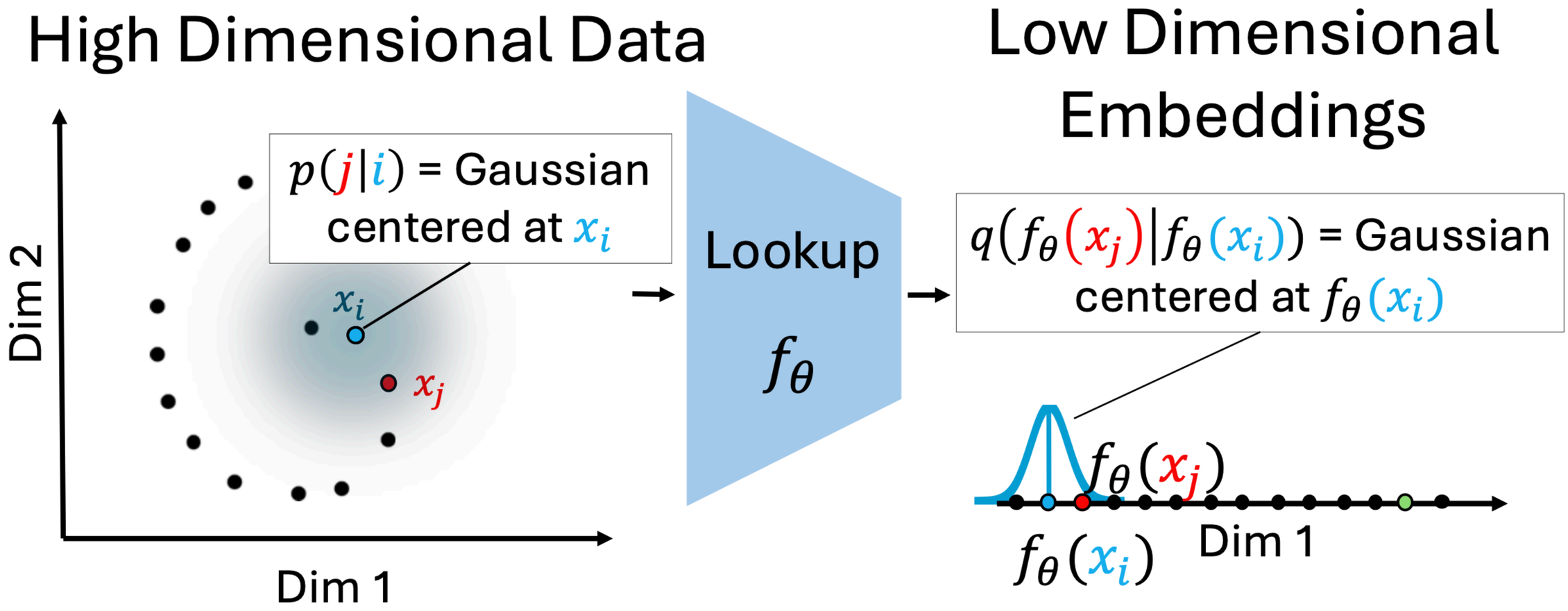


https://www.fabriziomusacchio.com/blog/2023-06-12-tsne_vs_pca/

SNE (Hinton et al. 2002)



SNE (Hinton et al. 2002)



$$\min_{\theta} D_{\mathrm{KL}}(p(j \mid i) \quad \| \quad q(f_\theta(x_i) \mid f_\theta(x_i))$$

SNE (Hinton et al. 2002)

High Dimensional Data

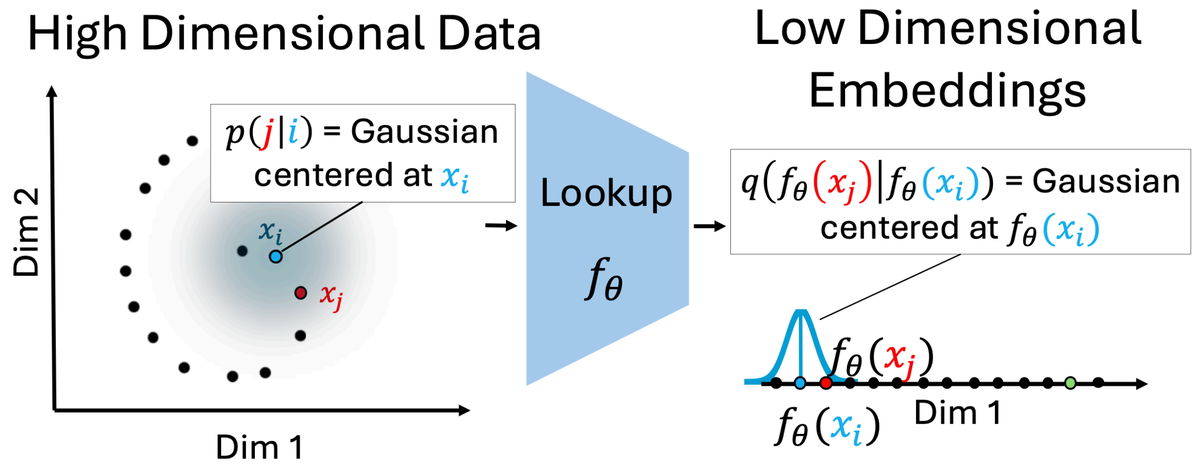
Low Dimensional

Intuition: Find embeddings such that euclidean distance in origin space matches euclidean distance in embedding space

However, this doesn't work for *all* pairs, b/c then you can't get dimensionality reduction. Instead, we use Gaussian likelihood, which quickly decays to zero for things that are far apart!

$$\min_{\theta} D_{\mathrm{KL}}(p(j \mid i) \quad \| \quad q(f_\theta(x_i) \mid f_\theta(x_i))$$

SNE (Hinton et al. 2002)



$$\min_\theta D_{\mathrm{KL}}(p(j \mid i) \quad \| \quad q(f_\theta(x_i) \mid f_\theta(x_i)))$$

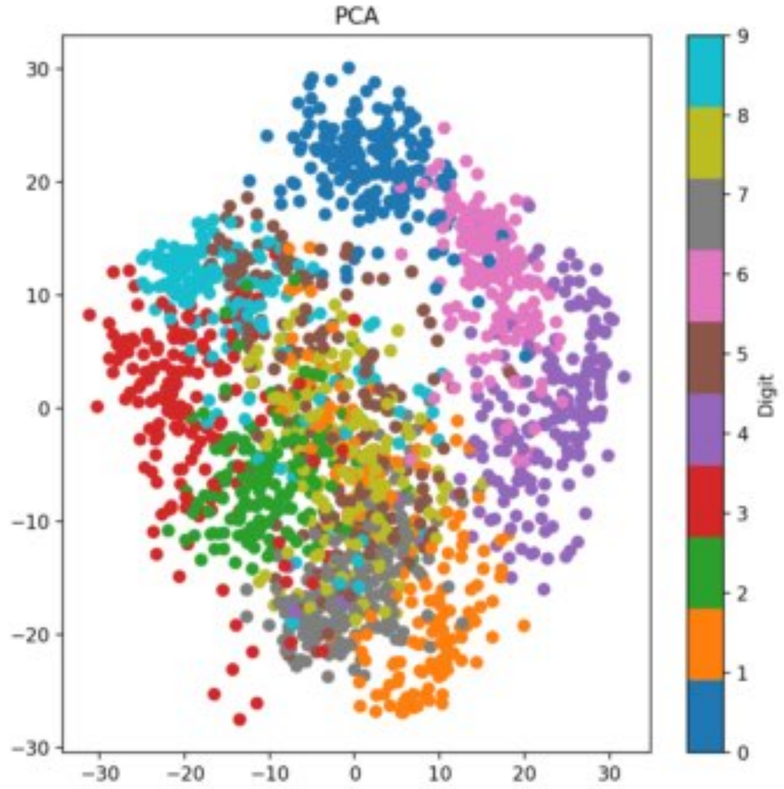
Gaussian over data points, x_i

$$\frac{\exp(-\|x_i - x_j\|^2 / 2\sigma_i^2)}{\sum_{k \neq i} \exp(-\|x_i - x_k\|^2 / 2\sigma_i^2)}$$

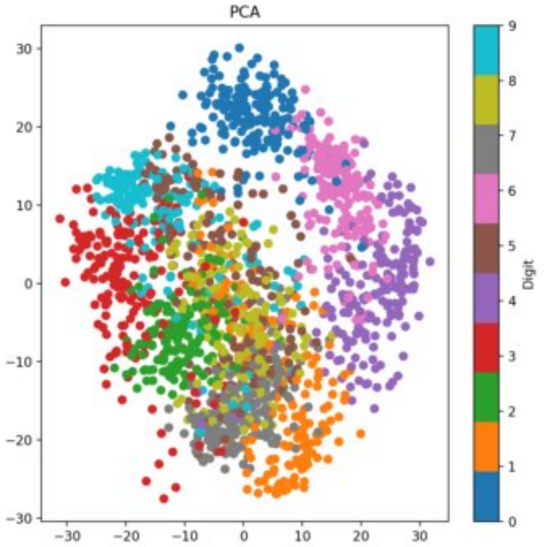
Gaussian over learned low-dimensional points, ϕ_i

$$\frac{\exp(-\|\phi_i - \phi_j\|^2)}{\sum_{k \neq i} \exp(-\|\phi_i - \phi_k\|^2)}$$

PCA (Pearson, 1901)



PCA (Pearson, 1901)



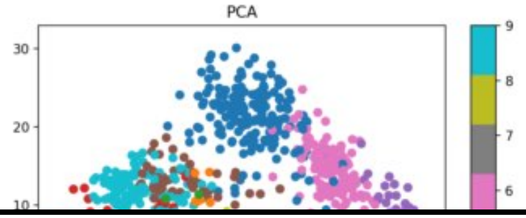
$$\min_{\theta} D_{\mathrm{KL}}(p(j \mid i) \quad \| \quad q(f_\theta(x_i) \mid f_\theta(x_i)))$$

$\mathbb{1}[\, i = j \,]$

Wide Gaussian on linear projection features, $f_\phi(x_i)$

$$\lim_{\sigma \to \infty} \frac{\exp(-\| f_\phi(x_i) - f_\phi(x_j) \|^2 / 2\sigma^2)}{\sum_{k \neq i} \exp(-\| f_\phi(x_i) - f_\phi(x_k) \|^2 / 2\sigma^2)}$$

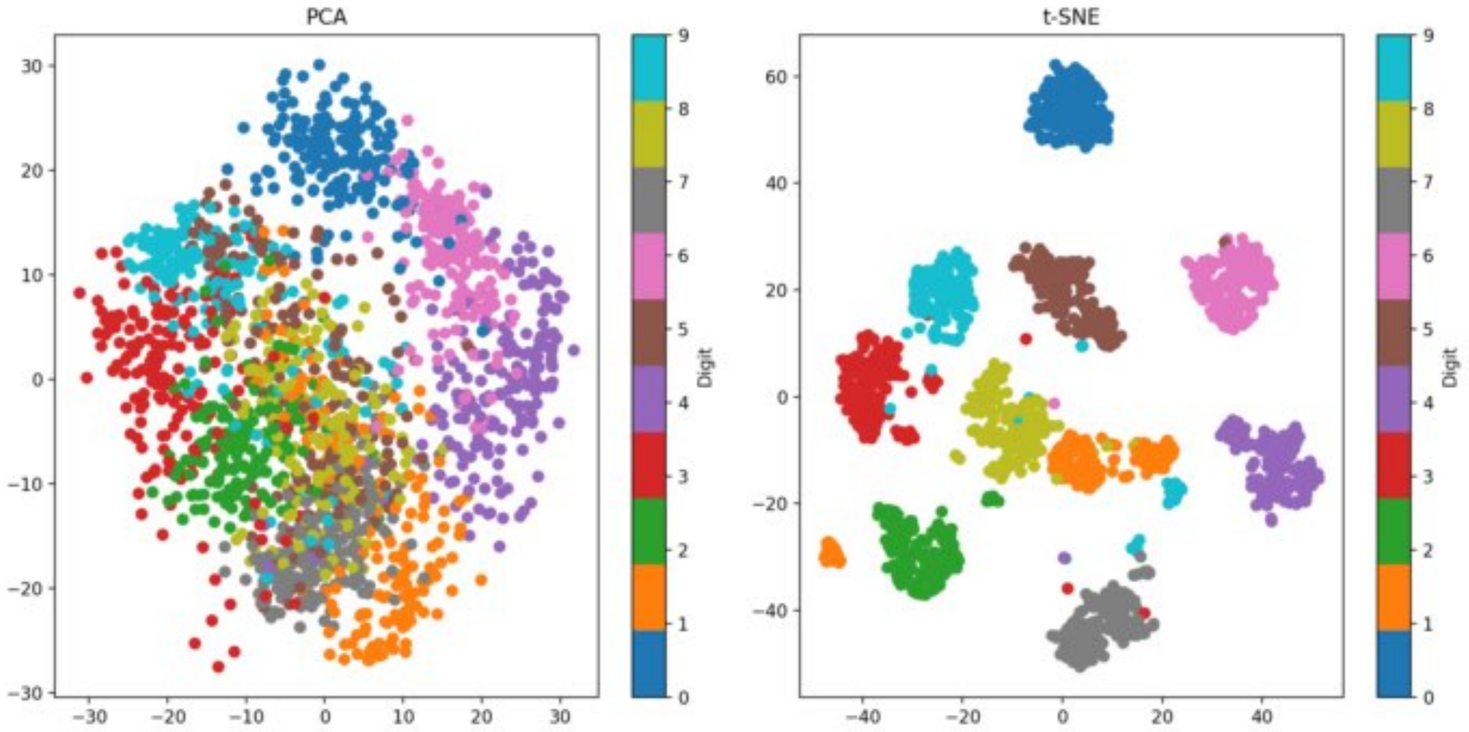
PCA (Pearson, 1901)



Can be seen as “Special Case” of SNE! We are finding f_θ (=linear function of data sample) such that $\|f_\theta(x_i) - f_\theta(x_i)\| = 0$ is *maximal* relative to all other $-\|f_\theta(x_i) - f_\theta(x_j)\|$, which means “maximize variance of the latter”.

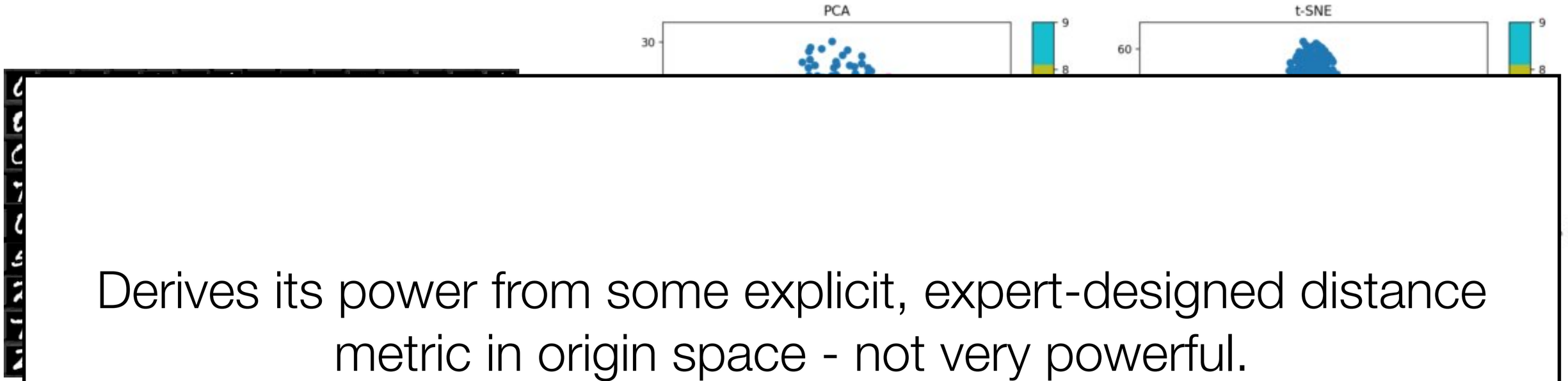
$$\lim_{\sigma \to \infty} \frac{\exp(-\|f_\phi(x_i) - f_\phi(x_j)\|^2 / 2\sigma^2)}{\sum_{k \neq i} \exp(-\|f_\phi(x_i) - f_\phi(x_k)\|^2 / 2\sigma^2)}$$

Dimensionality Reduction



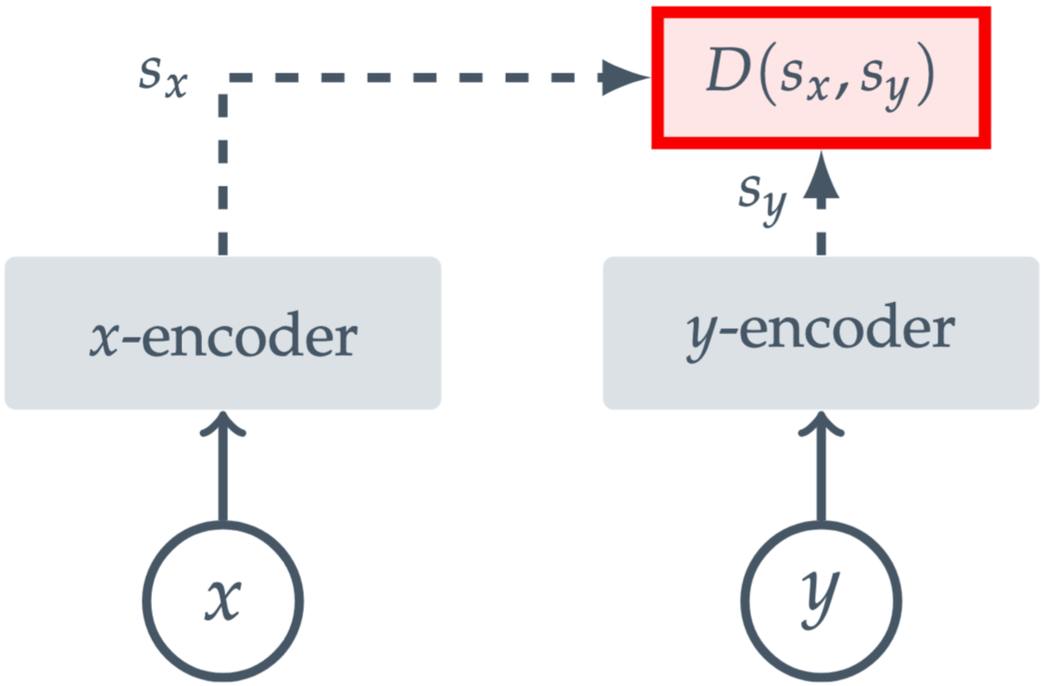
https://www.fabriziomusacchio.com/blog/2023-06-12-tsne_vs_pca/

Dimensionality Reduction



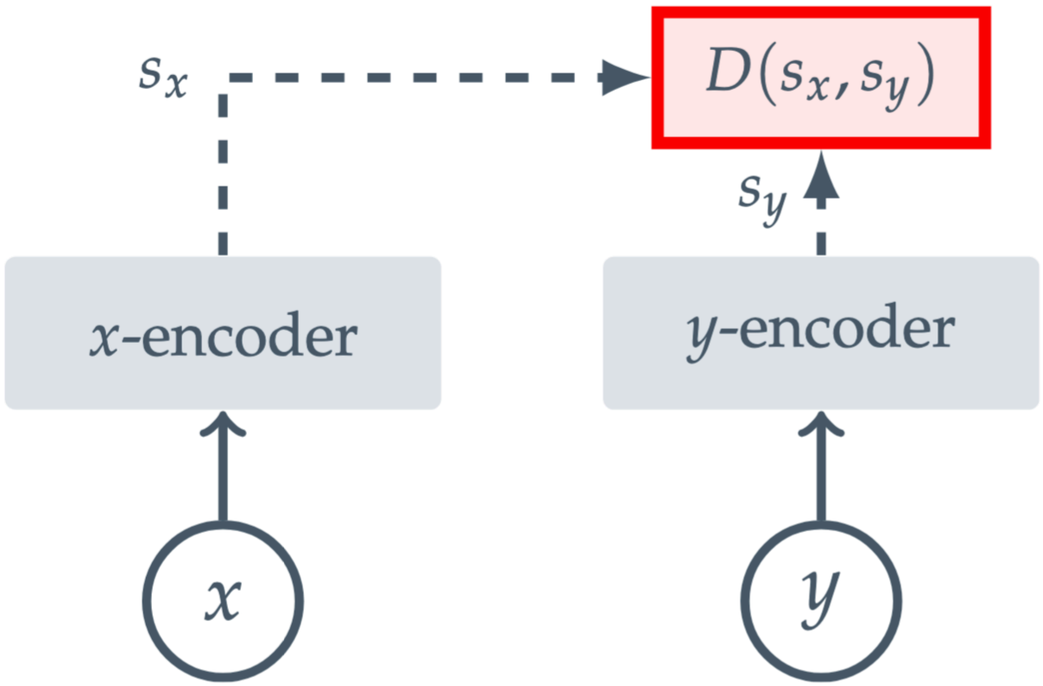
Derives its power from some explicit, expert-designed distance metric in origin space - not very powerful.

Joint Embedding Architectures



Given *compatible* x, y pairs (i.e. things that belong together, such as augmentations of an image, two similar images, etc), learn encoders that output s_x, s_y that are similar according to some distance metric D.

Problem: Collapse / Trivial Solution



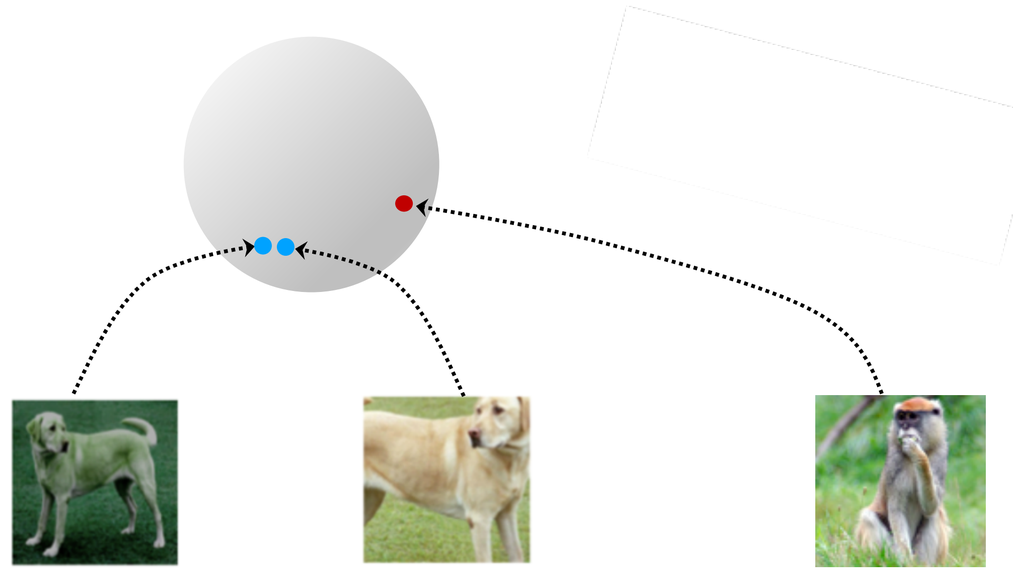
Trivial solution to

$$\min_{s_x, s_y} D(s_x, s_y)$$

Is always $s_x = s_y = c$ for some constant c (such as 0).

Contrastive Learning

- Encoder maps data onto a hypersphere: $f : \mathcal{X} \to \mathbb{S}^d$
- Cross-entropy for softmax “classifier” to discriminate “classes” defined by similarities



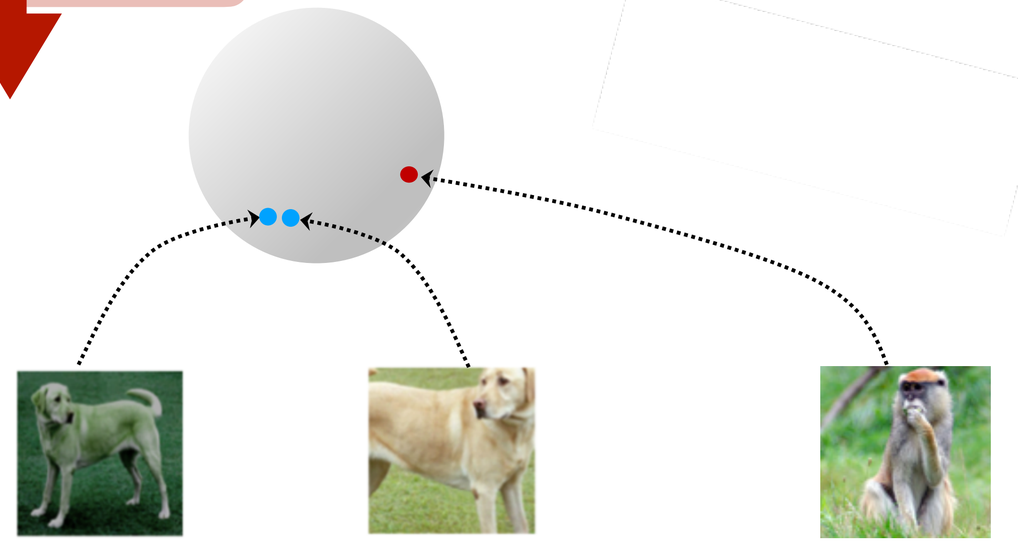
Contrastive Learning

- Encoder maps data onto a hypersphere: $f : \mathcal{X} \to \mathbb{S}^d$
- Cross-entropy for softmax “classifier” to discriminate “classes” defined by similarities

$$\min_f \quad \mathbb{E}_{(\mathbf{x},\mathbf{x}^+)\sim p_{pos},\{\mathbf{x}_i^-\}_{i=1}^N \sim p_{data}} \left[-\log \frac{e^{f(\mathbf{x})^\top f(\mathbf{x}^+)/\tau}}{e^{f(\mathbf{x})^\top f(\mathbf{x}^+)/\tau} + \sum_{i=1}^N e^{f(\mathbf{x})^\top f(\mathbf{x}_i^-)/\tau}} \right]$$

pull positive pair together

push negative pairs apart



Contrastive Learning

- Encoder maps data onto a hypersphere: $f : \mathcal{X} \to \mathbb{S}^d$
- Cross-entropy for softmax “classifier” to discriminate “classes” defined by similarities

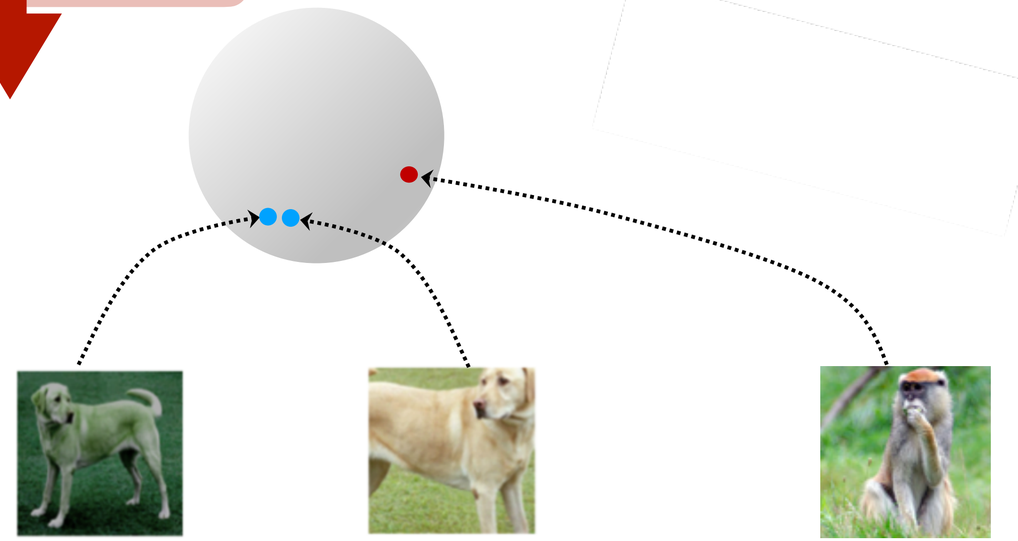
$$\min_f \; \mathbb{E}_{(\mathbf{x},\mathbf{x}^+)\sim p_{pos},\{\mathbf{x}_i^-\}_{i=1}^N \sim p_{data}} \left[-\log \frac{e^{f(\mathbf{x})^\top f(\mathbf{x}^+)/\tau}}{e^{f(\mathbf{x})^\top f(\mathbf{x}^+)/\tau} + \sum_{i=1}^N e^{f(\mathbf{x})^\top f(\mathbf{x}_i^-)/\tau}} \right]$$

pull positive pair together

push negative pairs apart

Symmetry: $\forall \mathbf{x}, \mathbf{x}^+, p_{\text{pos}}(\mathbf{x}, \mathbf{x}^+) = p_{\text{pos}}(\mathbf{x}^+, \mathbf{x})$

Matching marginal: $\forall \mathbf{x}, \int p_{\text{pos}}(\mathbf{x}, \mathbf{x}^+) d\mathbf{x}^+ = p_{\text{data}}(\mathbf{x})$



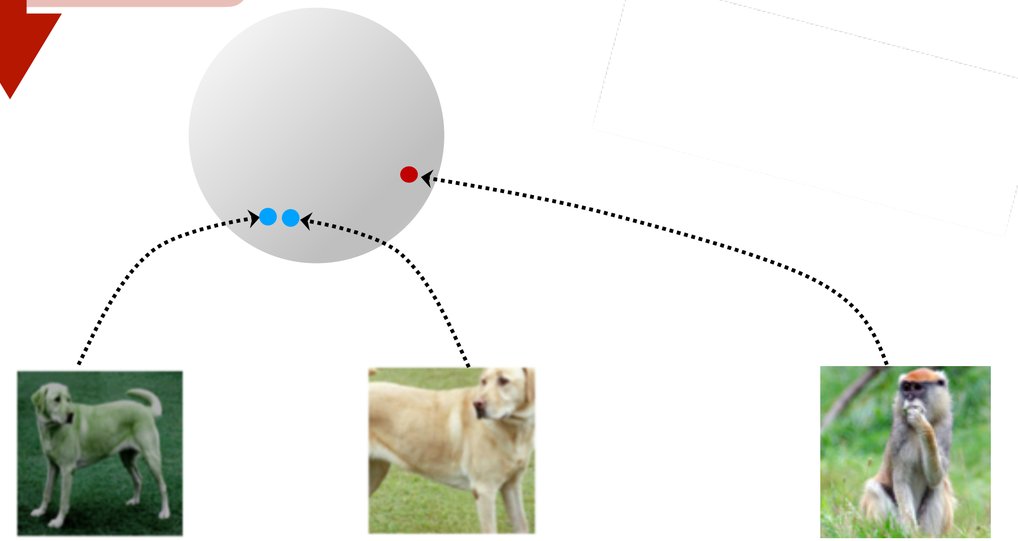
Contrastive Learning

- Encoder maps data onto a hypersphere: $f : \mathcal{X} \to \mathbb{S}^d$
- Cross-entropy for softmax "classifier" to discriminate "classes" defined by similarities

$$\min_f \quad \mathbb{E}_{(\mathbf{x},\mathbf{x}^+)\sim p_{pos},\{\mathbf{x}_i^-\}_{i=1}^N \sim p_{data}} \left[-\log \frac{e^{f(\mathbf{x})^\top f(\mathbf{x}^+)/\tau}}{e^{f(\mathbf{x})^\top f(\mathbf{x}^+)/\tau} + \sum_{i=1}^N e^{f(\mathbf{x})^\top f(\mathbf{x}_i^-)/\tau}} \right]$$

pull positive pair together

push negative pairs apart



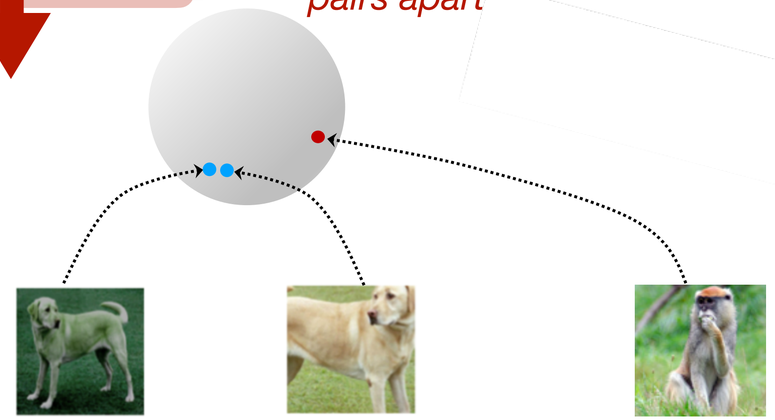
Noise-contrastive estimation (NCE) (Gutmann & Hyvärinen 2010), InfoNCE loss (van den Oord et al 2018), … similar losses also in metric learning

How can we make this “self-supervised”?

$$\min_f \; \mathbb{E}_{(\mathbf{x},\mathbf{x}^+)\sim p_{pos},\{\mathbf{x}_i^-\}_{i=1}^N \sim p_{data}} \left[-\log \frac{e^{f(\mathbf{x})^\top f(\mathbf{x}^+)/\tau}}{e^{f(\mathbf{x})^\top f(\mathbf{x}^+)/\tau} + \sum_{i=1}^N e^{f(\mathbf{x})^\top f(\mathbf{x}_i^-)/\tau}} \right]$$

pull positive pair together

push negative pairs apart



- What are the similar (positive) and dissimilar (negative) pairs?

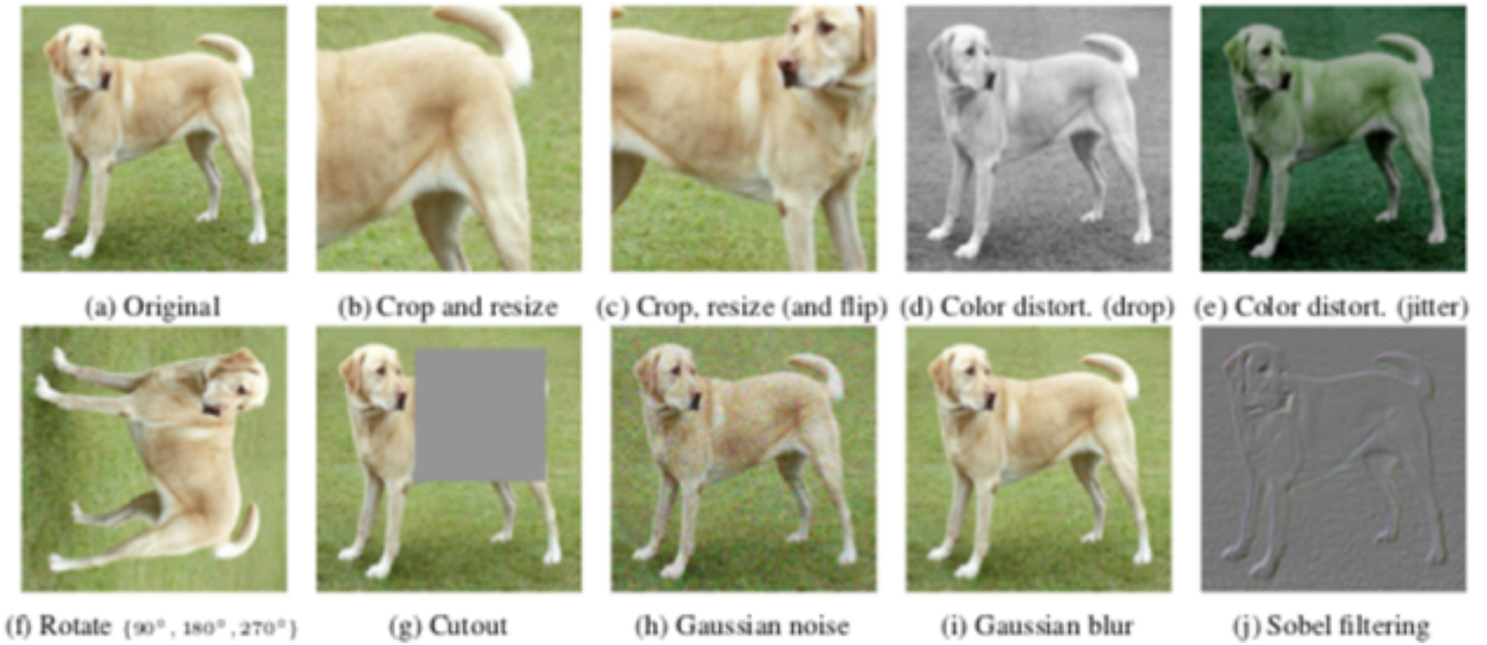
What are positive and negative examples?

Negative examples:
randomly uniformly
drawn from data

What are positive and negative examples?

Negative examples:
randomly uniformly drawn from data

Positive examples:
perturbations that keep semantic meaning, data augmentation



(Chen, Kornblith, Norouzi, Hinton 2020)

Contrastive Learning from Data Augmentation

[Slide credit: Phillip Isola]

SimCLR: [Chen, Kornblith, Norouzi, Hinton, ICML 2020]

Contrastive Learning from Data Augmentation

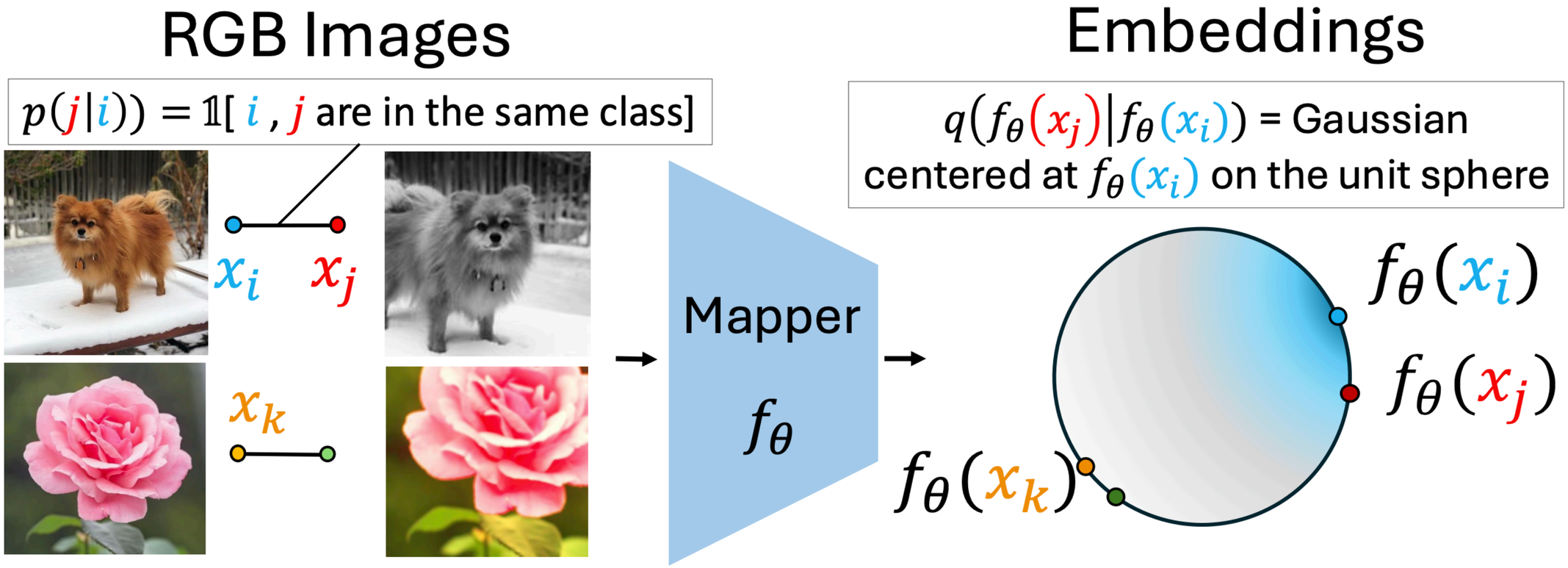
[Slide credit: Phillip Isola]

SimCLR: [Chen, Kornblith, Norouzi, Hinton, ICML 2020]

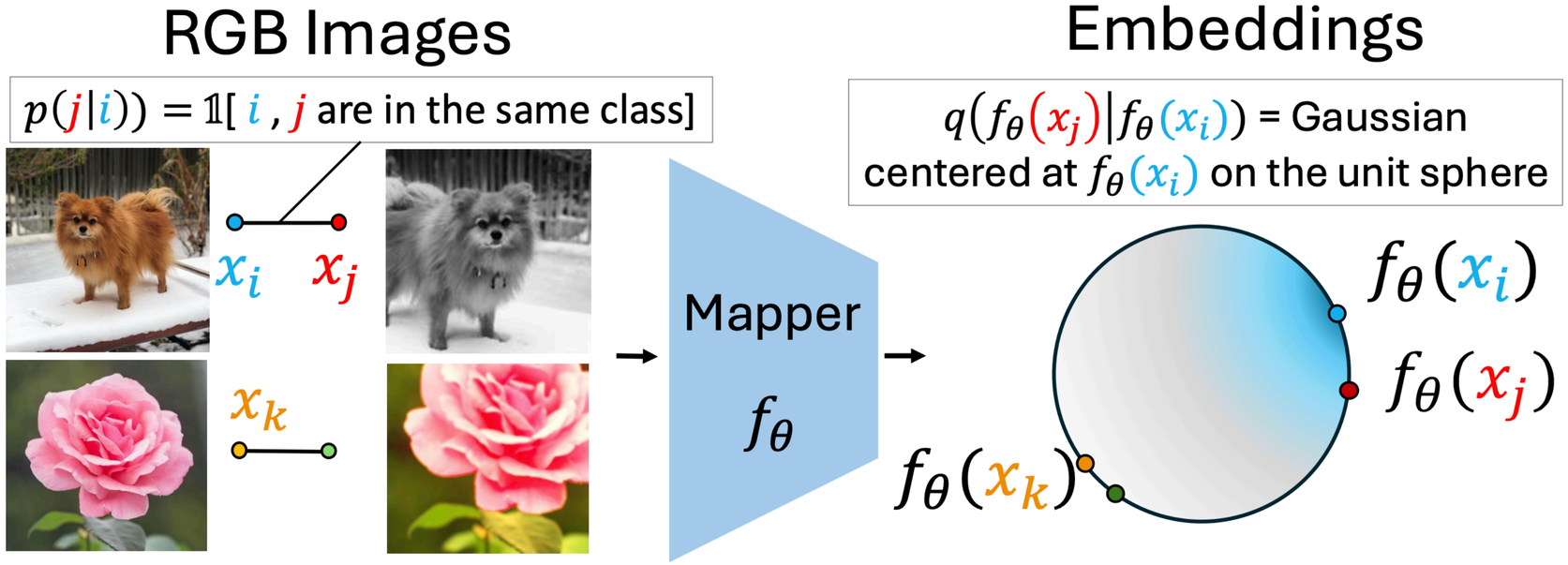
Contrastive Learning from Data Augmentation

[Slide credit: Phillip Isola]

SIMCLR in the I-Con Framework



SIMCLR in the I-Con Framework



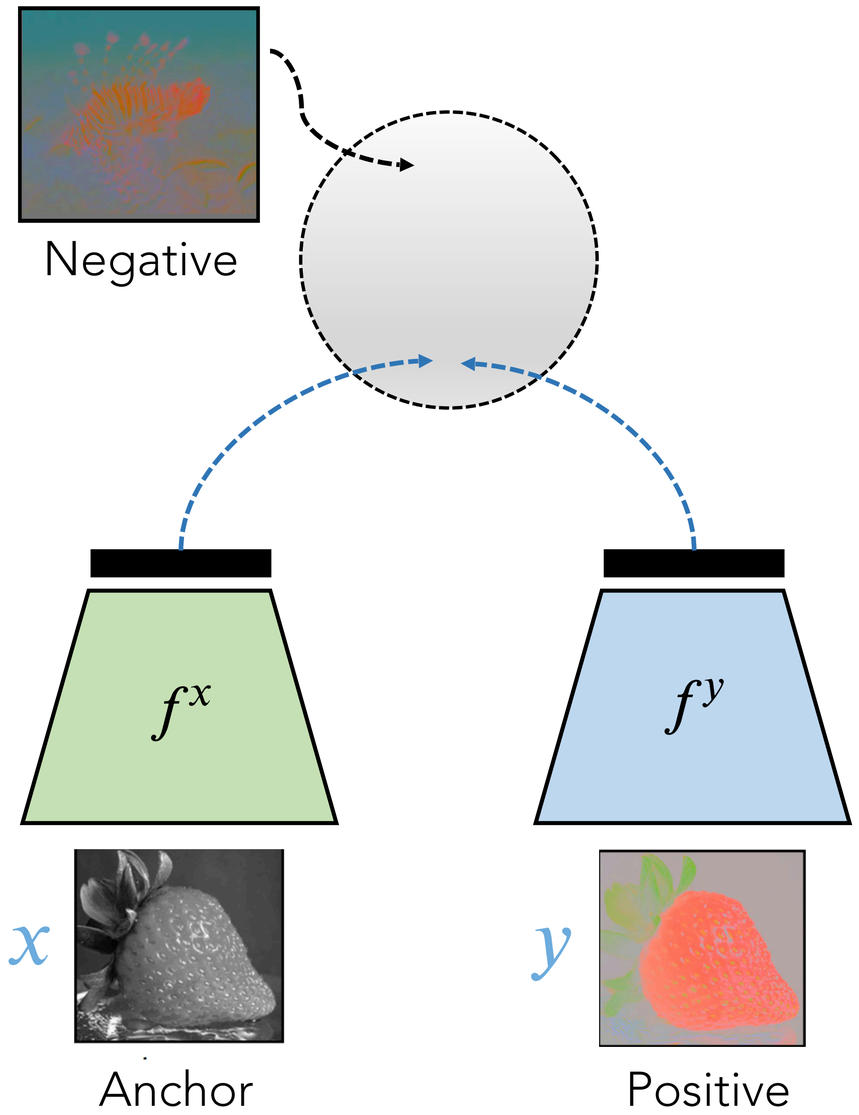
$$\min_\theta D_{\mathrm{KL}}(p(j \mid i) \quad \| \quad q(f_\theta(x_i) \mid f_\theta(x_i))$$

$\frac{1}{Z}\,\mathbb{1}[i$ and j are a positive pair$]$

Gaussian on deep normalized features

$$\frac{\exp\left(f_\phi(x_i)\cdot f_\phi(x_j)\right)}{\sum_{k\neq i}\exp\left(f_\phi(x_i)\cdot f_\phi(x_k)\right)}$$

Variations



(x, y) are two "views" of the same scene

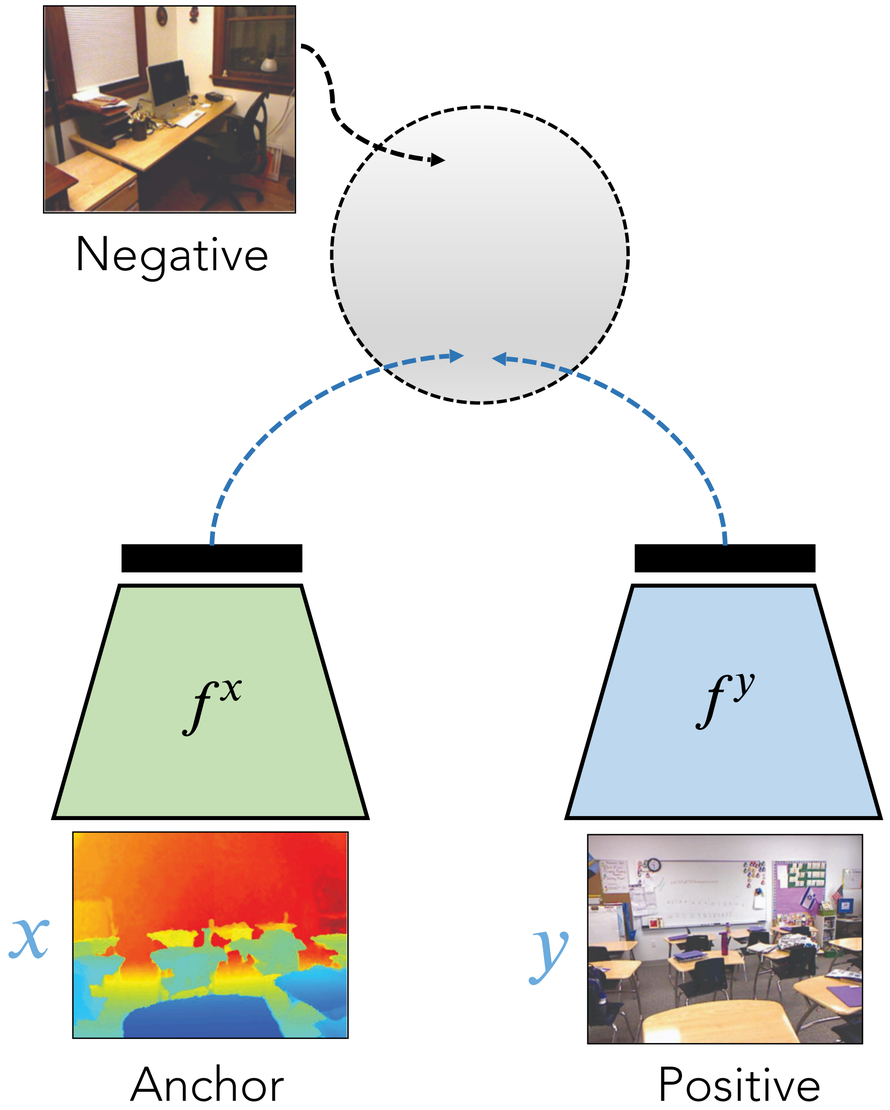
Cross-Channel Representation Learning

[CMC, Tian, Krishnan, Isola 2020]

⋮

[Slide credit: Phillip Isola]

Variations



(x, y) are two "views" of the same scene

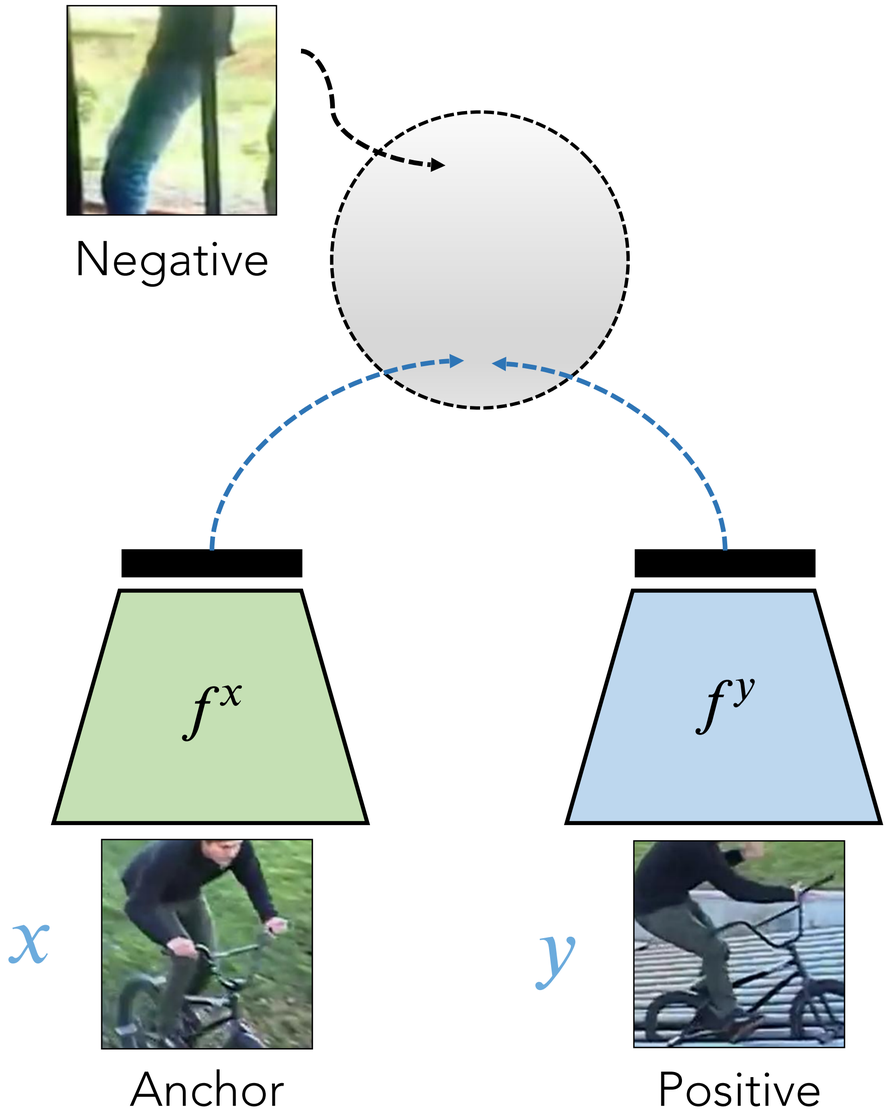
Cross-Channel Representation Learning

[CMC, Tian, Krishnan, Isola 2020]

⋮

[Slide credit: Phillip Isola]

Variations



(x, y) are two "views" of the same scene

Video Representation Learning

["Slow Feature Learning", Wiskott & Sejnowski 2002]

[Mobahi, Collobert, Weston 2009]

[Wang & Gupta 2015]

[Isola, Zoran, Krishnan, Adelson 2016]

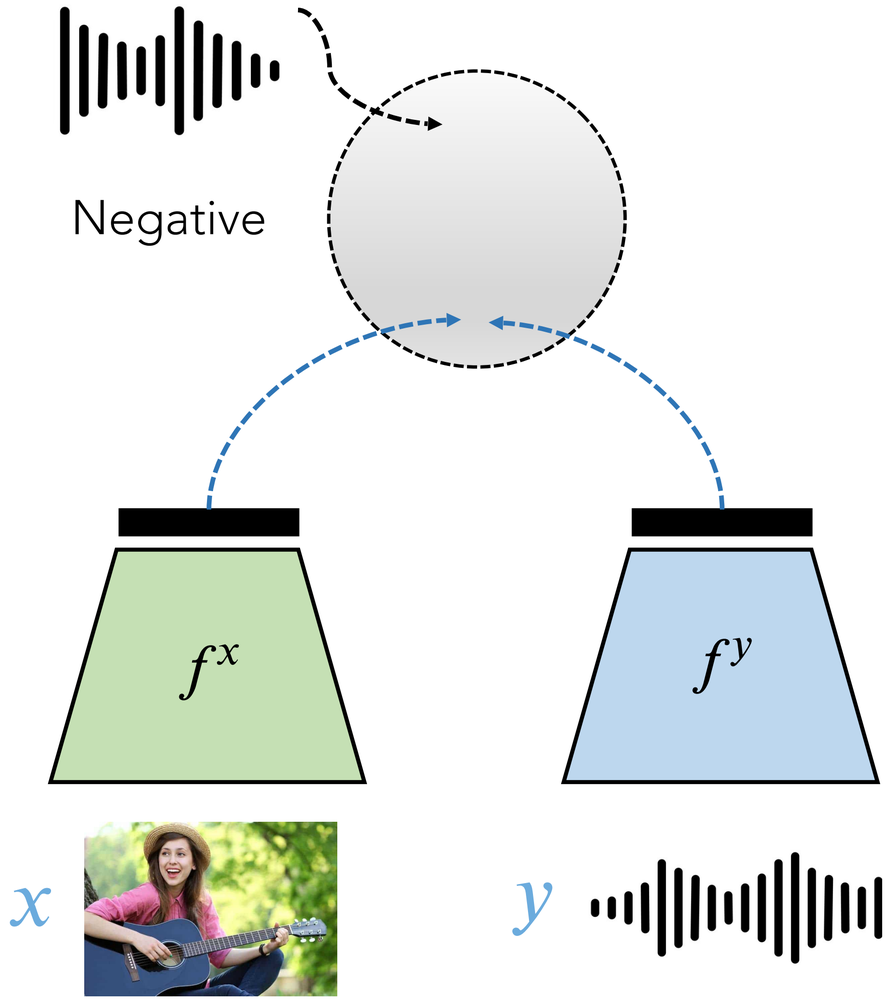
[Sermanet, Lynch, Chebotar et al. 2018]

[van den Oord, Li, Vinyals 2018]

⋮

[Slide credit: Phillip Isola]

Variations



(x, y) are two "views" of the same scene

Audio-Visual Representation Learning

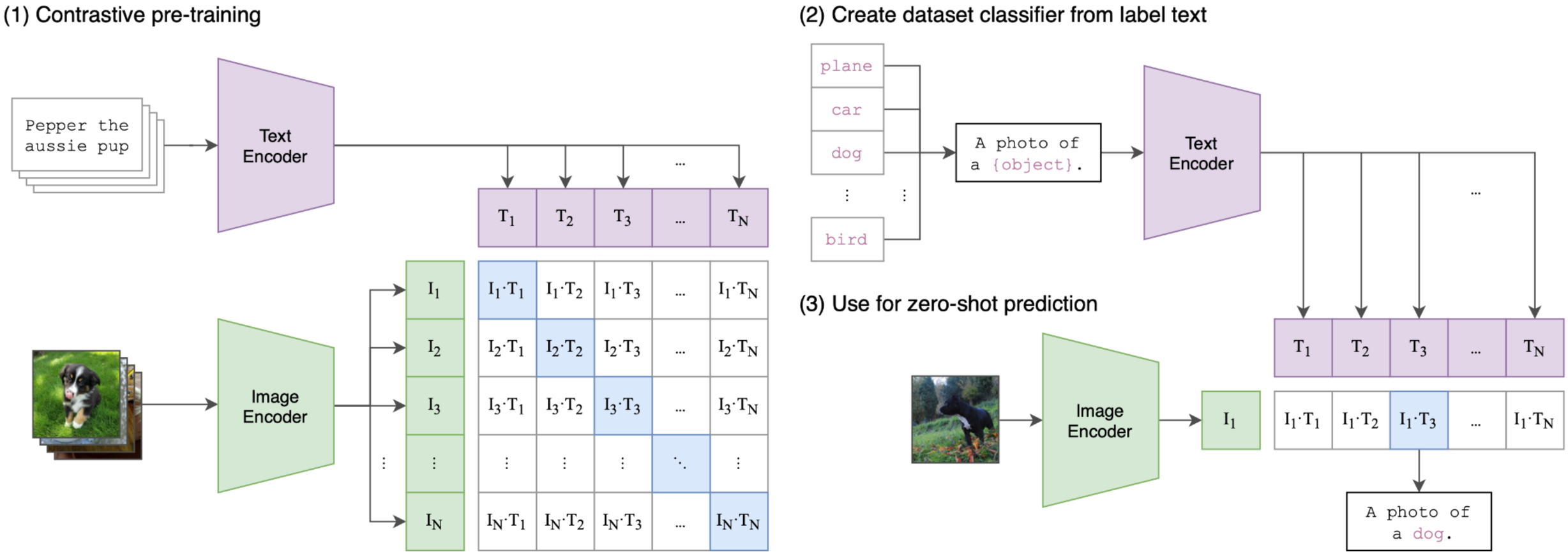
[Harwath, Recasens, Suris et al. 2018]

[Owens & Efros 2018]

⋮

[Slide credit: Phillip Isola]

Example: CLIP



What is the contrastive loss doing?

$$\mathscr{L}_{cont}(f) = \mathbb{E}_{(\mathbf{x},\mathbf{x}^+)\sim p_{pos},\{\mathbf{x}_i^-\}_{i=1}^N \sim p_{data}} \left[-\log \frac{e^{f(\mathbf{x})^\top f(\mathbf{x}^+)/\tau}}{e^{f(\mathbf{x})^\top f(\mathbf{x}^+)/\tau} + \sum_{i=1}^N e^{f(\mathbf{x})^\top f(\mathbf{x}_i^-)/\tau}} \right]$$

- cross-entropy loss to distinguish data points

What is the contrastive loss doing?

$$\mathscr{L}_{cont}(f) = \mathbb{E}_{(\mathbf{x},\mathbf{x}^+)\sim p_{pos},\{\mathbf{x}_i^-\}_{i=1}^N \sim p_{data}} \left[-\log \frac{e^{f(\mathbf{x})^\top f(\mathbf{x}^+)/\tau}}{e^{f(\mathbf{x})^\top f(\mathbf{x}^+)/\tau} + \sum_{i=1}^N e^{f(\mathbf{x})^\top f(\mathbf{x}_i^-)/\tau}} \right]$$

- cross-entropy loss to distinguish data points
- maximizes a lower bound on *mutual information* between "views" $f(\mathbf{x}), f(\mathbf{x}^+)$ *(Poole et al, 2019).*

Background: Shannon Entropy

Consider random variable X with probability density p. Then its *Shannon Entropy* is defined as:

$$H(X) = -\mathbb{E}_{p(X)}\left[\log p(X)\right]$$ Or explicitly: $$H(X) = -\int p(x)\log p(x)dx$$

$I(x) = -\log p(x)$ measures "information gain of a sample". Imagine if variable is deterministic - you don't need to draw a sample to learn what value it will have; you will not be surprised. Now imagine a uniform distribution - you are *maximally* uncertain about what value of the sample.

Entropy is hence the "expected information gain per sample"

Background: *Conditional* Shannon Entropy

For *two* random variables Z_1, Z_2, we can define their *conditional* Shannon entropy as:

$$H(Z_1 \mid Z_2) = -\iint_{\mathcal{Z}_1 \times \mathcal{Z}_2} p(z_1, z_2) \log \frac{p(z_1, z_2)}{p(z_2)}$$

Intuitively: Expected surprise of Z_1 if we are given Z_2.

For instance, if Z_1 and Z_2 are highly correlated, then $H(Z_2 \mid Z_1)$ vanishes: no more surprises!

Background: Mutual Information

Given two random variables Z_1 and Z_2, their mutual information is defined as:

$$I(Z_1; Z_2) \;=\; H(Z_2) \;-\; H(Z_2 \,|\, Z_1)\,.$$

Intuitively: How much is there to learn about Z_2 at all (imagine deterministic, there is nothing to now), minus how much of it do I learn by knowing Z_1?

What is the contrastive loss doing?

$$\mathscr{L}_{cont}(f) = \mathbb{E}_{(\mathbf{x},\mathbf{x}^+)\sim p_{pos},\{\mathbf{x}_i^-\}_{i=1}^N \sim p_{data}} \left[-\log \frac{e^{f(\mathbf{x})^\top f(\mathbf{x}^+)/\tau}}{e^{f(\mathbf{x})^\top f(\mathbf{x}^+)/\tau} + \sum_{i=1}^N e^{f(\mathbf{x})^\top f(\mathbf{x}_i^-)/\tau}} \right]$$

- cross-entropy loss to distinguish data points

What is the contrastive loss doing?

$$\mathscr{L}_{cont}(f) = \mathbb{E}_{(\mathbf{x},\mathbf{x}^+)\sim p_{pos},\{\mathbf{x}_i^-\}_{i=1}^N \sim p_{data}} \left[-\log \frac{e^{f(\mathbf{x})^\top f(\mathbf{x}^+)/\tau}}{e^{f(\mathbf{x})^\top f(\mathbf{x}^+)/\tau} + \sum_{i=1}^N e^{f(\mathbf{x})^\top f(\mathbf{x}_i^-)/\tau}} \right]$$

- cross-entropy loss to distinguish data points
- maximizes a lower bound on *mutual information* between "views" $f(\mathbf{x}), f(\mathbf{x}^+)$ *(Poole et al, 2019).*

What is the contrastive loss doing?

$$\mathscr{L}_{cont}(f) = \mathbb{E}_{(\mathbf{x},\mathbf{x}^+)\sim p_{pos},\{\mathbf{x}_i^-\}_{i=1}^N \sim p_{data}} \left[-\log \frac{e^{f(\mathbf{x})^\top f(\mathbf{x}^+)/\tau}}{e^{f(\mathbf{x})^\top f(\mathbf{x}^+)/\tau} + \sum_{i=1}^N e^{f(\mathbf{x})^\top f(\mathbf{x}_i^-)/\tau}} \right]$$

- cross-entropy loss to distinguish data points
- maximizes a lower bound on *mutual information* between "views" $f(\mathbf{x}), f(\mathbf{x}^+)$ *(Poole et al, 2019).*
  - I.e., how much is there to know about view $\mathbf{x}$, and how much does $\mathbf{x}^+$ tell me about it

On Mutual Information Maximization for Representation Learning

Michael Tschannen* **Josip Djolonga*** **Paul K. Rubenstein†** **Sylvain Gelly** **Mario Lucic**
Google Research, Brain Team

Abstract

Many recent methods for unsupervised or self-supervised representation learning train feature extractors by maximizing an estimate of the mutual information (MI) between different views of the data. This comes with several immediate problems: For example, MI is notoriously hard to estimate, and using it as an objective for representation learning may lead to highly entangled representations due to its invariance under arbitrary invertible transformations. Nevertheless, these methods have been repeatedly shown to excel in practice. In this paper we argue, and provide empirical evidence, that the success of these methods cannot be attributed to the properties of MI alone, and that they strongly depend on the inductive bias in both the choice of feature extractor architectures and the parametrization of the employed MI estimators. Finally, we establish a connection to deep metric learning and argue that this interpretation may be a plausible explanation for the success of the recently introduced methods.

Maximized MI and worsened downstream performance

Looser bounds with simpler critics can lead to better representations

What (else) is the contrastive loss doing?

- Recall: properties of “good” representations:

1. Concentration/Alignment: Data from the same class is close together, remove irrelevant information

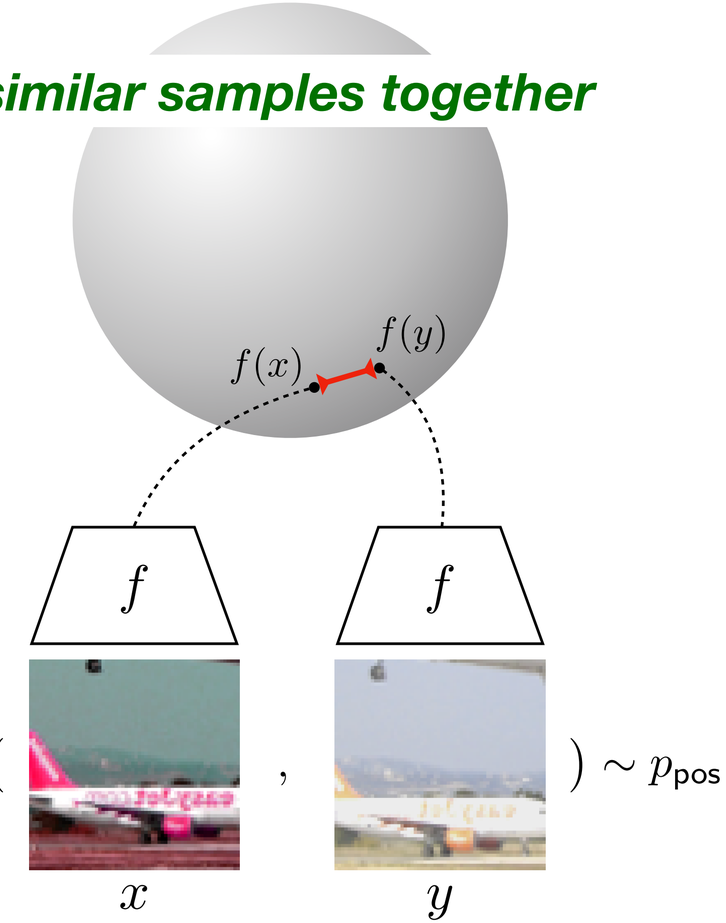
2. Separation: classes are well separated, do not lose information

3. Robustness to irrelevant perturbations

Alignment and separation

pull similar samples together

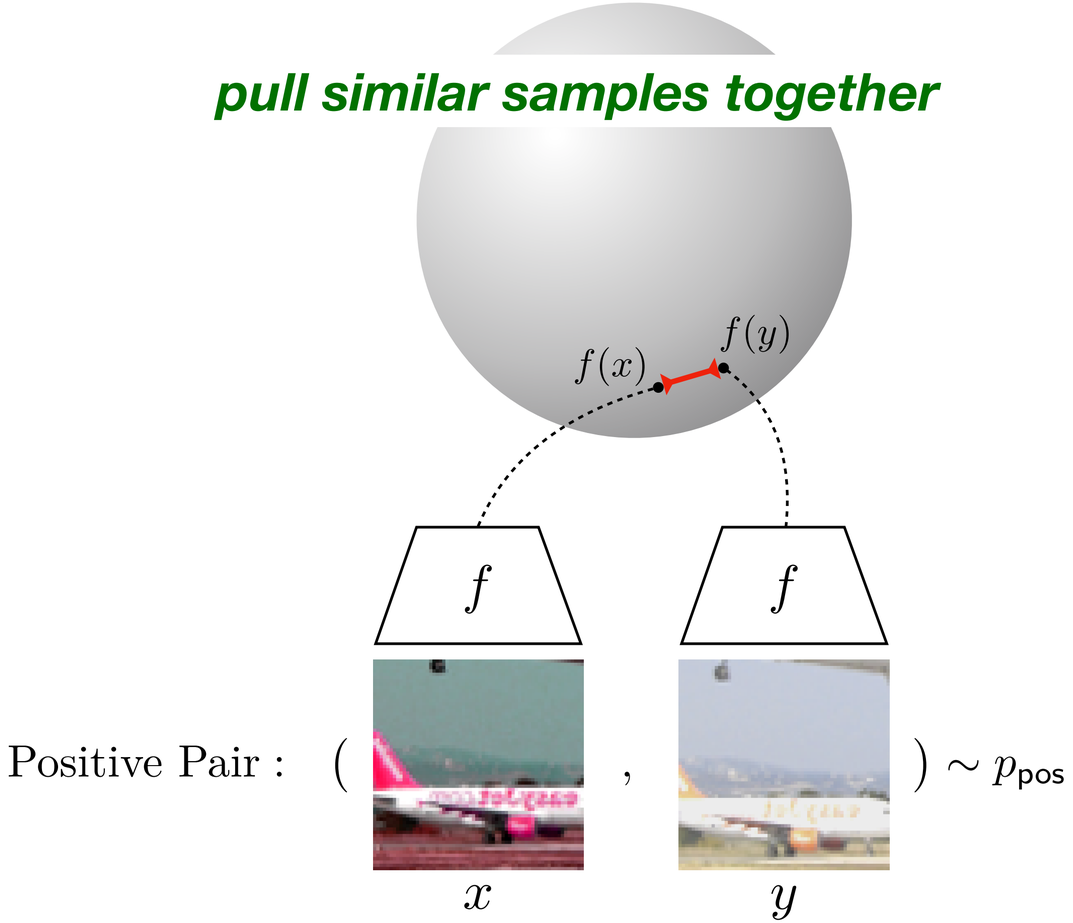
push dissimilar samples apart



Alignment: Similar samples have similar features

figures: Wang & Isola, 2020

Alignment and separation



Alignment: Similar samples have similar features

Uniformity: Preserve maximal information

figures: Wang & Isola, 2020

Feature distribution from Contrastive Learning

Toy example:

Train `CIFAR-10` encoders with $\mathcal{S}^1$ feature space (circle).

Visualize feature distributions on the validation set.

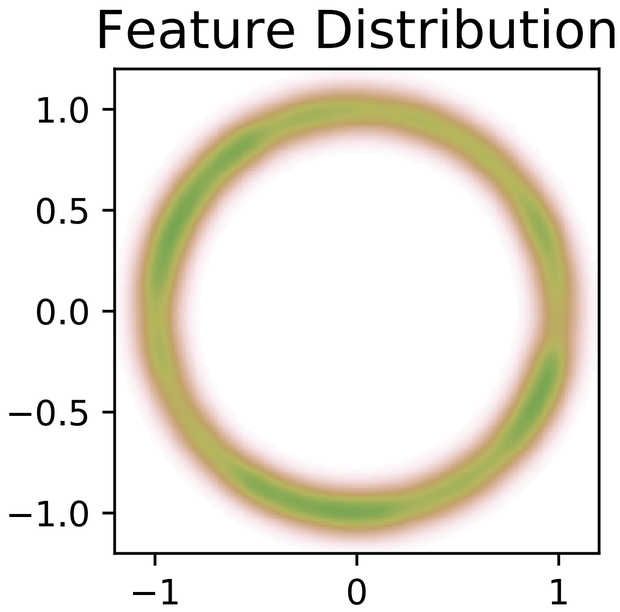
slide: Phillip Isola

Feature distribution from Contrastive Learning

Toy example:

Train `CIFAR-10` encoders with $\mathcal{S}^1$ feature space (circle).

Visualize feature distributions on the validation set.



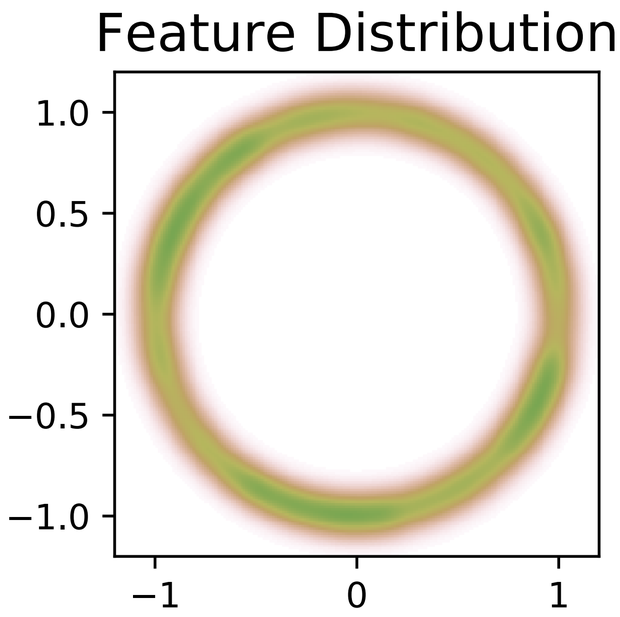
Unsupervised Contrastive Learning

slide: Phillip Isola

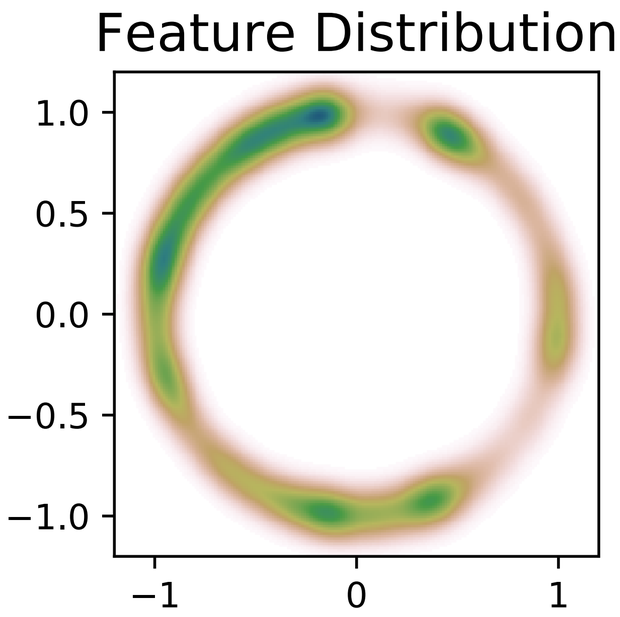
Feature distribution from Contrastive Learning

Toy example:

Train `CIFAR-10` encoders with $\mathcal{S}^1$ feature space (circle).
Visualize feature distributions on the validation set.



Unsupervised Contrastive Learning



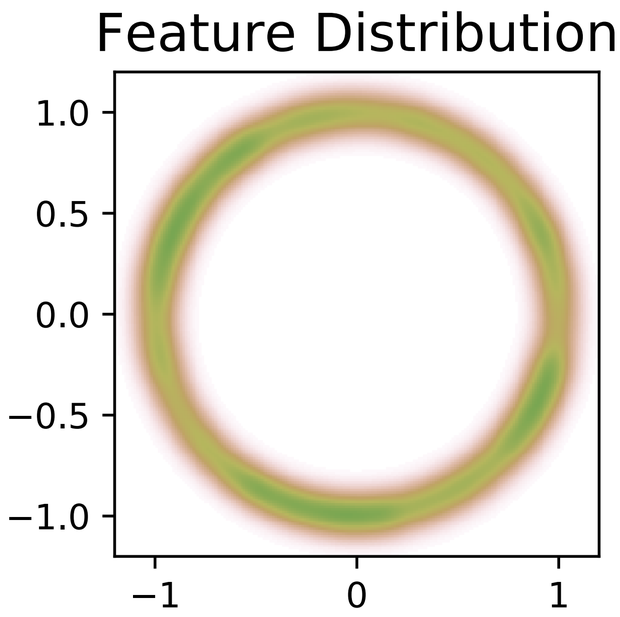
Supervised Predictive (NLL) Learning

slide: Phillip Isola

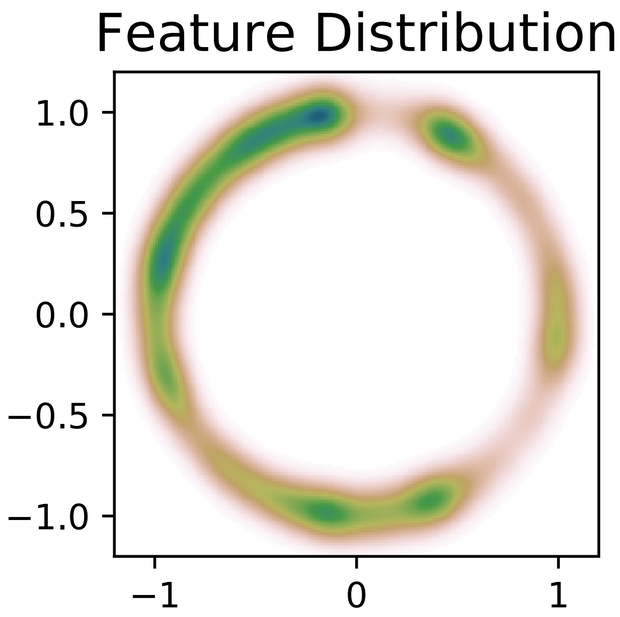
Feature distribution from Contrastive Learning

Toy example:

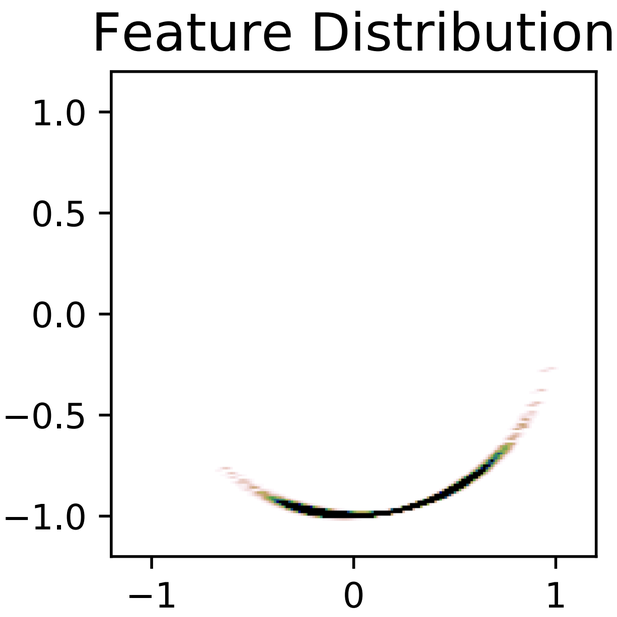
Train `CIFAR-10` encoders with $\mathcal{S}^1$ feature space (circle).
Visualize feature distributions on the validation set.



Unsupervised Contrastive Learning



Supervised Predictive (NLL) Learning



Random Network Initialization

slide: Phillip Isola

Metrics for Alignment and Uniformity

- Idea: Study alignment and uniformity via appropriate metrics

Metrics for Alignment and Uniformity

- Idea: Study alignment and uniformity via appropriate metrics

Alignment: expected positive pair feature distance

$$\mathcal{L}_{\text{align}}(f;\alpha) \triangleq \mathop{\mathbb{E}}_{(x,y)\sim p_{\text{pos}}} \left[\|f(x) - f(y)\|_2^{\alpha} \right] \qquad \alpha > 0$$

Metrics for Alignment and Uniformity

- Idea: Study alignment and uniformity via appropriate metrics

Alignment: expected positive pair feature distance

$$\mathcal{L}_{\text{align}}(f;\alpha) \triangleq \mathop{\mathbb{E}}_{(x,y)\sim p_{\text{pos}}} \left[\|f(x) - f(y)\|_2^{\alpha} \right] \qquad \alpha > 0$$

Uniformity: logarithm of expected pairwise Gaussian potential

$$\mathcal{L}_{\text{uniform}}(f;t) \triangleq \log \mathop{\mathbb{E}}_{x,y \overset{\text{i.i.d.}}{\sim} p_{\text{data}}} \left[G_t(f(x), f(y)) \right] \triangleq \log \mathop{\mathbb{E}}_{x,y \overset{\text{i.i.d.}}{\sim} p_{\text{data}}} \left[e^{-t\|f(x)-f(y)\|_2^2} \right] \qquad t > 0$$

Metrics for Alignment and Uniformity

- Idea: Study alignment and uniformity via appropriate metrics

Alignment: expected positive pair feature distance

$$\mathcal{L}_{\text{align}}(f;\alpha) \triangleq \mathop{\mathbb{E}}_{(x,y)\sim p_{\text{pos}}} \left[\|f(x) - f(y)\|_2^{\alpha} \right] \qquad \alpha > 0$$

Uniformity: logarithm of expected pairwise Gaussian potential

$$\mathcal{L}_{\text{uniform}}(f;t) \triangleq \log \mathop{\mathop{\mathbb{E}}_{x,y}}_{\text{i.i.d.}\sim} p_{\text{data}} \left[G_t(f(x), f(y)) \right] \triangleq \log \mathop{\mathop{\mathbb{E}}_{x,y}}_{\text{i.i.d.}\sim} p_{\text{data}} \left[e^{-t\|f(x)-f(y)\|_2^2} \right] \qquad t > 0$$

The uniform distribution on the hypersphere is the unique measure minimizing the expected pairwise potential

Asymptotics of the contrastive loss

Theorem 1 (Asymptotics of $\mathcal{L}_{\text{contrastive}}$)**.** *For fixed $\tau > 0$, as the number of negative samples $M \to \infty$, the (normalized) contrastive loss converges to*

$$\begin{aligned}
&\lim_{M\to\infty} \mathcal{L}_{\text{contrastive}}(f;\tau,M) - \log M \\
&\quad= \lim_{M\to\infty} \mathop{\mathbb{E}}_{\substack{(x,y)\sim p_{\text{pos}} \\ \{x_i^-\}_{i=1}^M \overset{\text{i.i.d.}}{\sim} p_{\text{data}}}} \left[-\log \frac{e^{f(x)^\top f(y)/\tau}}{e^{f(x)^\top f(y)/\tau} + \sum_i e^{f(x_i^-)^\top f(y)/\tau}} \right] - \log M \\
&\quad= -\frac{1}{\tau} \mathop{\mathbb{E}}_{(x,y)\sim p_{\text{pos}}} \left[f(x)^\top f(y) \right] + \mathop{\mathbb{E}}_{x\sim p_{\text{data}}} \left[\log \mathop{\mathbb{E}}_{x^-\sim p_{\text{data}}} \left[e^{f(x^-)^\top f(x)/\tau} \right] \right].
\end{aligned} \tag{2}$$

Asymptotics of the contrastive loss

Theorem 1 (Asymptotics of $\mathcal{L}_{\text{contrastive}}$)**.** *For fixed $\tau > 0$, as the number of negative samples $M \to \infty$, the (normalized) contrastive loss converges to*

$$\begin{aligned}
&\lim_{M\to\infty} \mathcal{L}_{\text{contrastive}}(f;\tau,M) - \log M \\
&\quad= \lim_{M\to\infty} \mathop{\mathbb{E}}_{\substack{(x,y)\sim p_{\text{pos}} \\ \{x_i^-\}_{i=1}^M \overset{\text{i.i.d.}}{\sim} p_{\text{data}}}} \left[-\log \frac{e^{f(x)^\top f(y)/\tau}}{e^{f(x)^\top f(y)/\tau} + \sum_i e^{f(x_i^-)^\top f(y)/\tau}} \right] - \log M \\
&\quad= -\frac{1}{\tau} \mathop{\mathbb{E}}_{(x,y)\sim p_{\text{pos}}} \left[f(x)^\top f(y) \right] + \mathop{\mathbb{E}}_{x\sim p_{\text{data}}} \left[\log \mathop{\mathbb{E}}_{x^-\sim p_{\text{data}}} \left[e^{f(x^-)^\top f(x)/\tau} \right] \right].
\end{aligned} \tag{2}$$

We have the following results:

1. *The first term is minimized iff f is perfectly aligned.*
2. *If perfectly uniform encoders exist, they form the exact minimizers of the second term.*

Asymptotics of the contrastive loss

Theorem 1 (Asymptotics of $\mathcal{L}_{\text{contrastive}}$)**.** *For fixed $\tau > 0$, as the number of negative samples $M \to \infty$, the (normalized) contrastive loss converges to*

$$\begin{aligned}
&\lim_{M\to\infty} \mathcal{L}_{\text{contrastive}}(f;\tau,M) - \log M \\
&\quad= \lim_{M\to\infty} \mathop{\mathbb{E}}_{\substack{(x,y)\sim p_{\text{pos}} \\ \{x_i^-\}_{i=1}^M \overset{\text{i.i.d.}}{\sim} p_{\text{data}}}} \left[-\log \frac{e^{f(x)^\top f(y)/\tau}}{e^{f(x)^\top f(y)/\tau} + \sum_i e^{f(x_i^-)^\top f(y)/\tau}} \right] - \log M \\
&\quad= -\frac{1}{\tau} \mathop{\mathbb{E}}_{(x,y)\sim p_{\text{pos}}} \left[f(x)^\top f(y) \right] + \mathop{\mathbb{E}}_{x\sim p_{\text{data}}} \left[\log \mathop{\mathbb{E}}_{x^-\sim p_{\text{data}}} \left[e^{f(x^-)^\top f(x)/\tau} \right] \right].
\end{aligned} \tag{2}$$

We have the following results:

1. *The first term is minimized iff f is perfectly aligned.*
2. *If perfectly uniform encoders exist, they form the exact minimizers of the second term.*

Perfect alignment: Optimal $\mathcal{L}_{\text{align}}$, mapping positive pairs to same features.

Asymptotics of the contrastive loss

Theorem 1 (Asymptotics of $\mathcal{L}_{\text{contrastive}}$)**.** *For fixed $\tau > 0$, as the number of negative samples $M \to \infty$, the (normalized) contrastive loss converges to*

$$\begin{aligned}
&\lim_{M\to\infty} \mathcal{L}_{\text{contrastive}}(f;\tau,M) - \log M \\
&\quad= \lim_{M\to\infty} \mathop{\mathbb{E}}_{\substack{(x,y)\sim p_{\text{pos}} \\ \{x_i^-\}_{i=1}^M \overset{\text{i.i.d.}}{\sim} p_{\text{data}}}} \left[-\log \frac{e^{f(x)^\top f(y)/\tau}}{e^{f(x)^\top f(y)/\tau} + \sum_i e^{f(x_i^-)^\top f(y)/\tau}} \right] - \log M \\
&\quad= -\frac{1}{\tau} \mathop{\mathbb{E}}_{(x,y)\sim p_{\text{pos}}} \left[f(x)^\top f(y) \right] + \mathop{\mathbb{E}}_{x\sim p_{\text{data}}} \left[\log \mathop{\mathbb{E}}_{x^-\sim p_{\text{data}}} \left[e^{f(x^-)^\top f(x)/\tau} \right] \right].
\end{aligned} \tag{2}$$

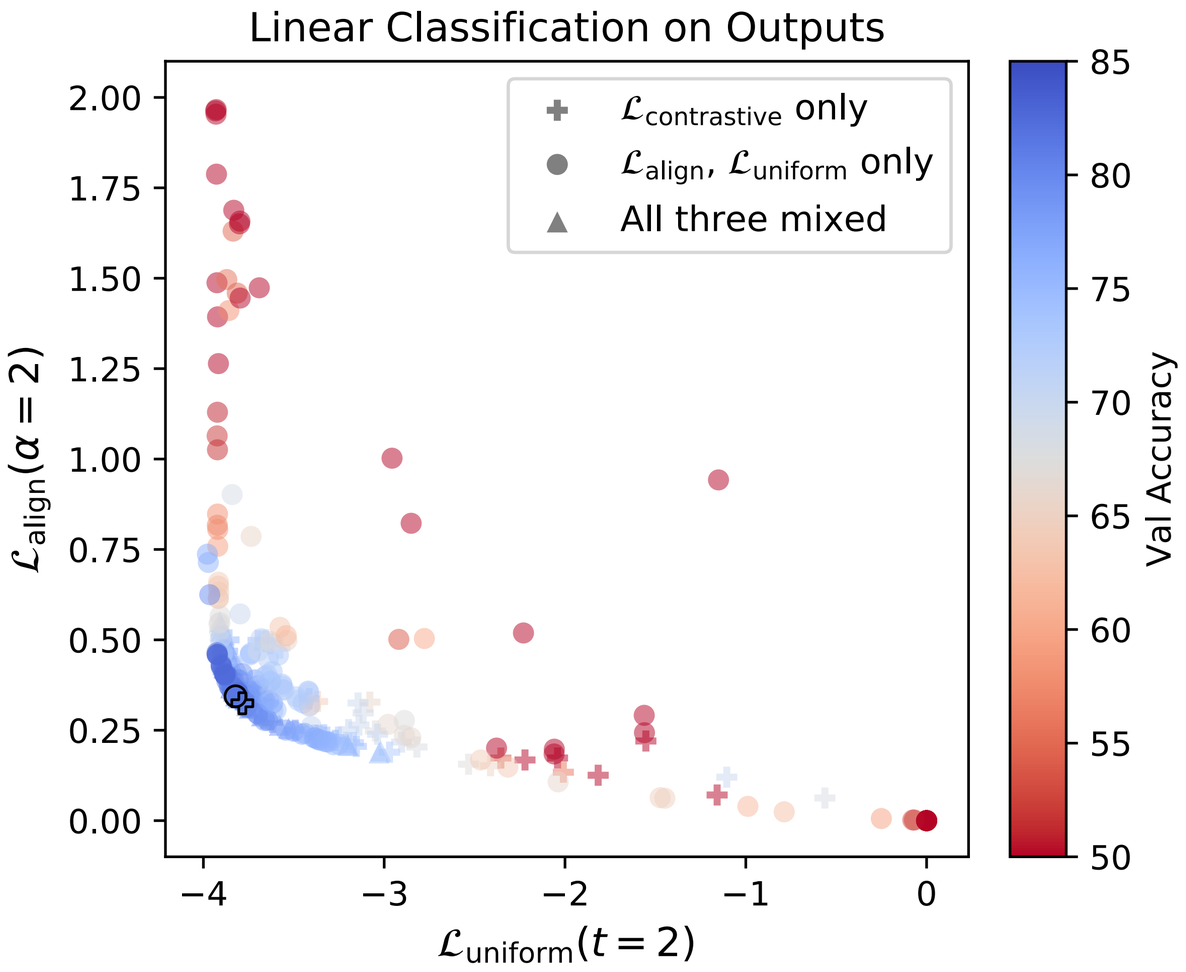
We have the following results:

1. *The first term is minimized iff f is perfectly aligned.*
2. *If perfectly uniform encoders exist, they form the exact minimizers of the second term.*

Perfect alignment: Optimal $\mathcal{L}_{\text{align}}$, mapping positive pairs to same features.

Perfect uniformity: Uniform feature distribution. (Push them as far apart as possible)

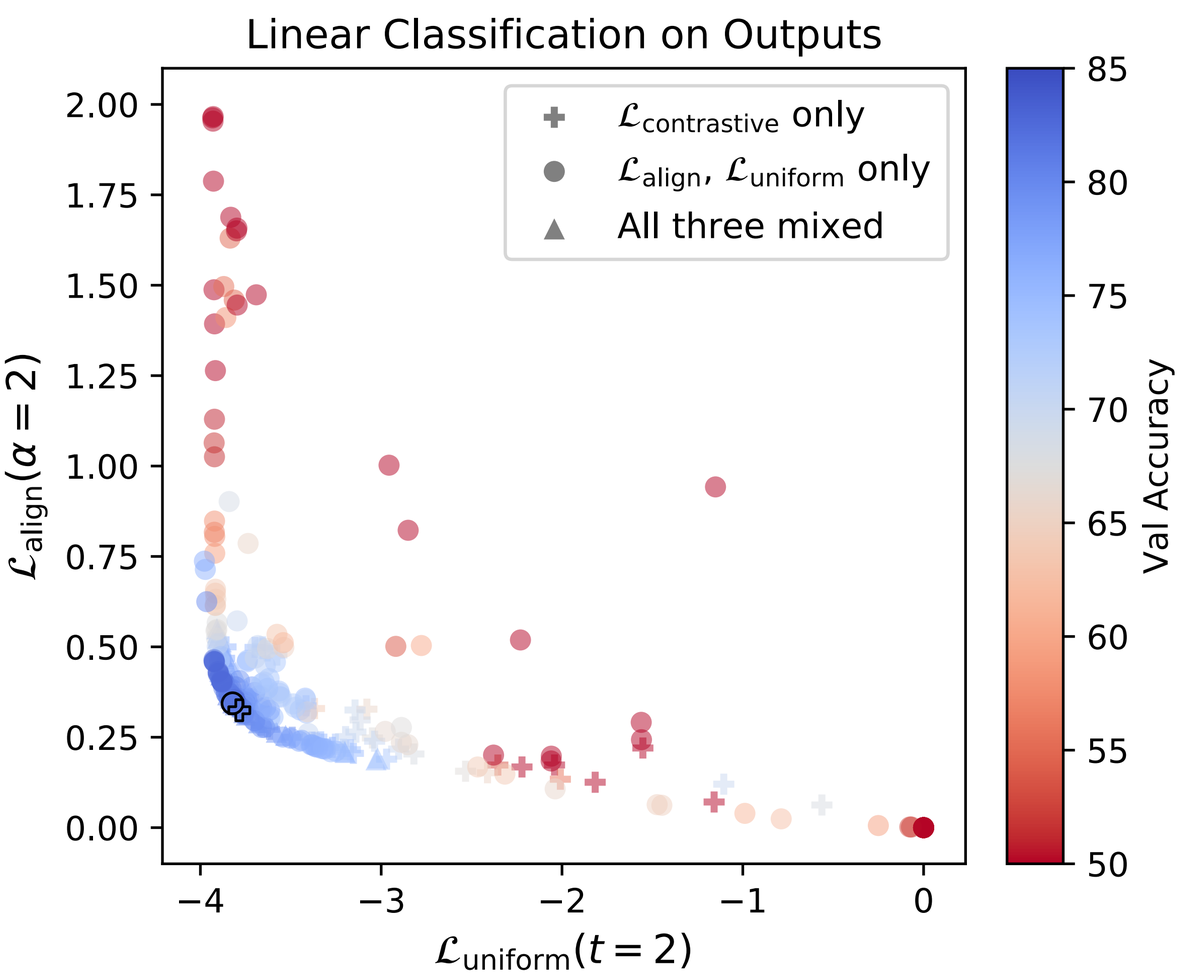
Relation Between Representation Quality and Alignment & Uniformity



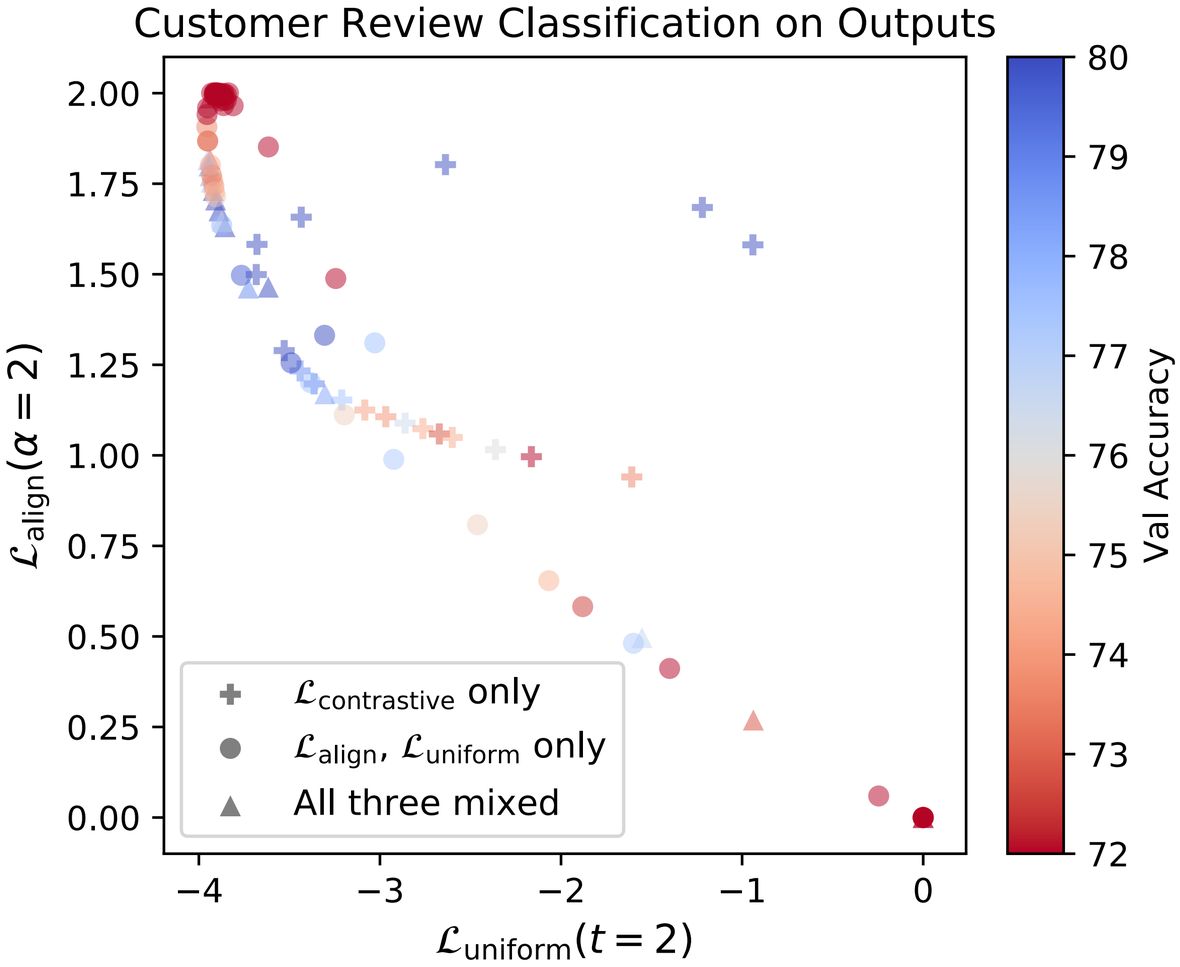
306 STL-10 Encoders

figures: Wang & Isola, 2020

Relation Between Representation Quality and Alignment & Uniformity



306 STL-10 Encoders



108 BookCorpus Encoders

figures: Wang & Isola, 2020

What is the contrastive loss doing?

- Loss function encourages:

1. Concentration/Alignment: Data from the same class is close together, remove irrelevant information

2. Separation: classes are well separated, do not lose information

- What do the selection of positive and negative pairs encourage?

What are we “teaching” the model via choice of pairs?

- positive pairs = augmentations of the same data point
  should be close

What are we “teaching” the model via choice of pairs?

- positive pairs = augmentations of the same data point
  should be close

- => learned representation is invariant to perturbations induced by data augmentations: learned invariance

What are we “teaching” the model via choice of pairs?

- positive pairs = augmentations of the same data point
  should be close

- => learned representation is invariant to perturbations induced by data augmentations: learned invariance

- Finding the “right” invariances can be challenging for different types of data

What is the contrastive loss doing?

- Loss function encourages:

1. Concentration/Alignment: Data from the same class is close together, remove irrelevant information

2. Separation: classes are well separated, do not lose information

- Data encourages:

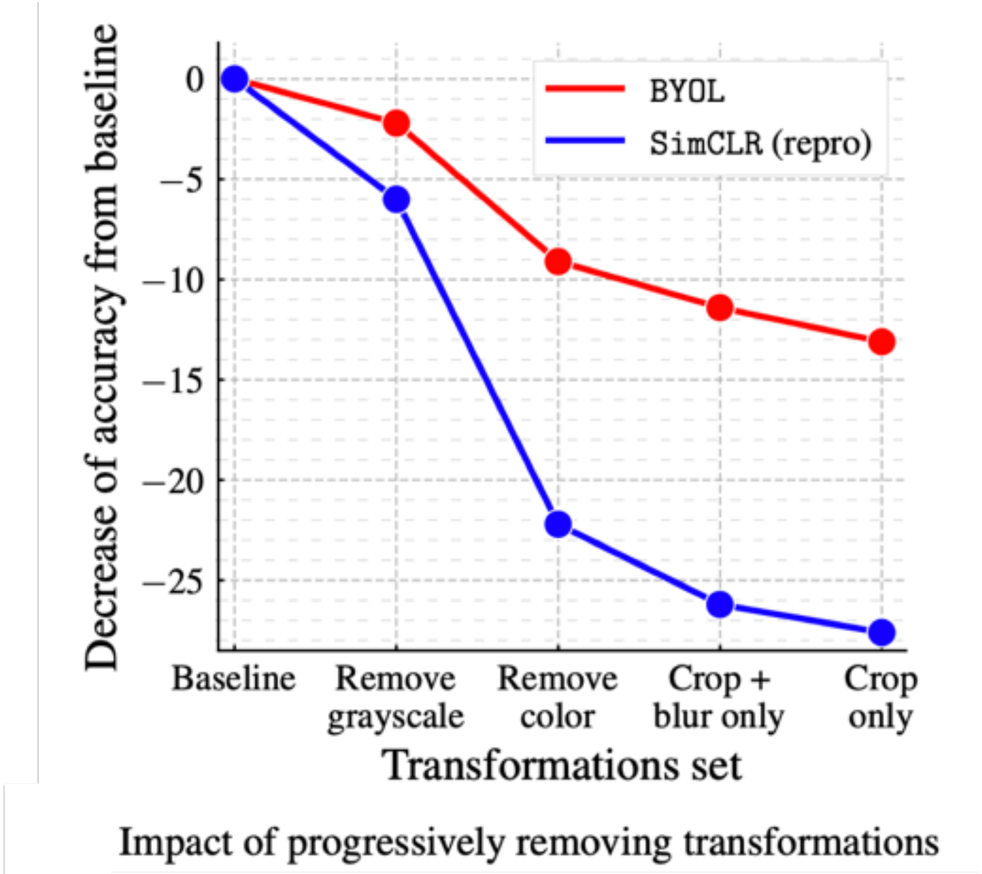
3. Robustness to **irrelevant** perturbations

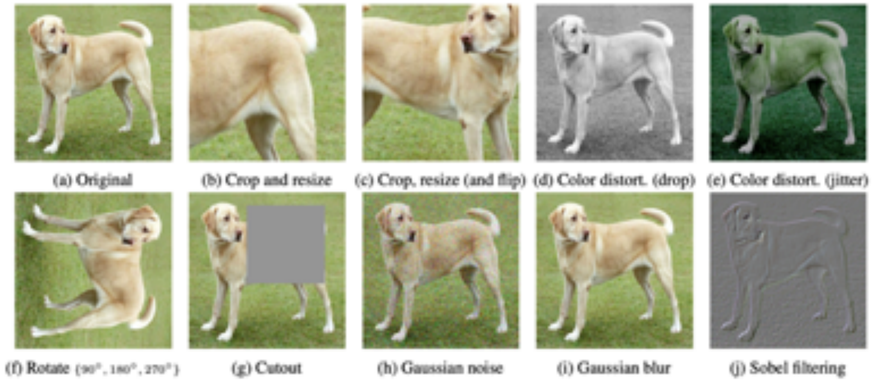
Ingredients to make self-supervised CL work (better)

- heavy data augmentation
- projection heads
- large batch size (many negative examples)
- choice of data pairs / hard negative examples

} SimCLR model

Effect of data augmentation



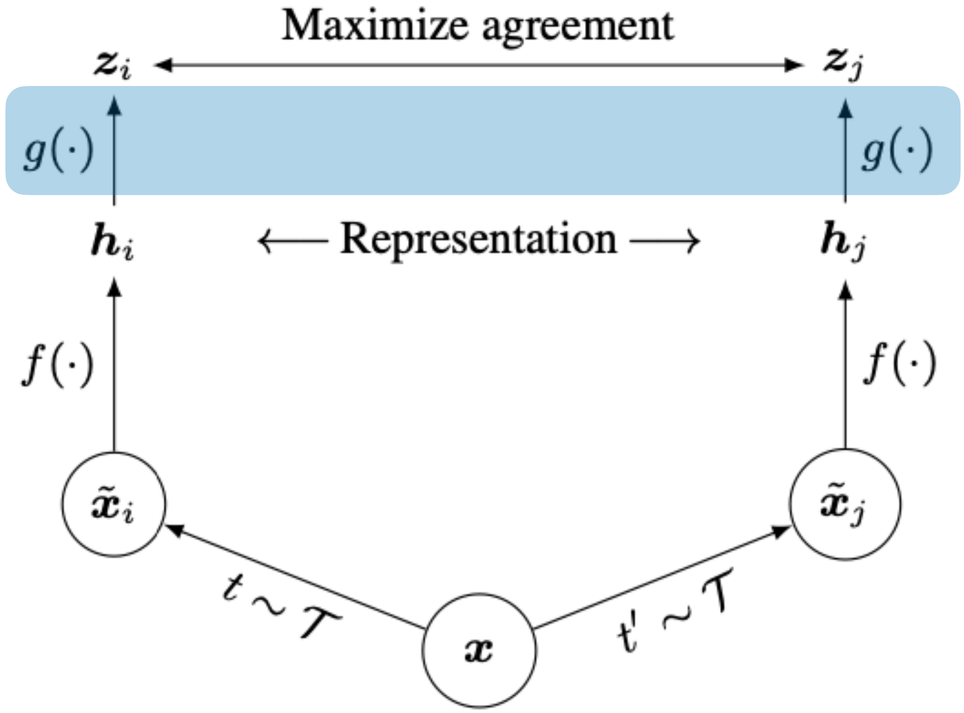


(figure: Grill et al 2020)

Projection head

- contrastive loss is applied to a transformed version $g(\mathbf{h})$ of the representation $\mathbf{h}$
- g is linear or small MLP
- use $\mathbf{h}$ for downstream task

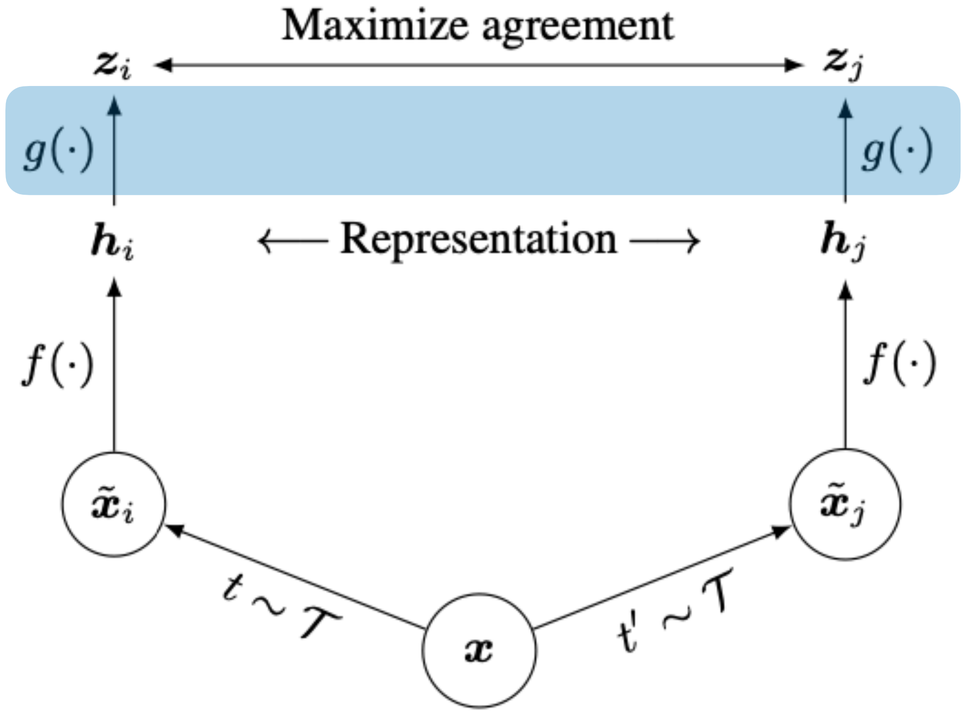
- Projection head improves performance!



Projection head

- contrastive loss is applied to a transformed version $g(\mathbf{h})$ of the representation $\mathbf{h}$
- g is linear or small MLP
- use $\mathbf{h}$ for downstream task

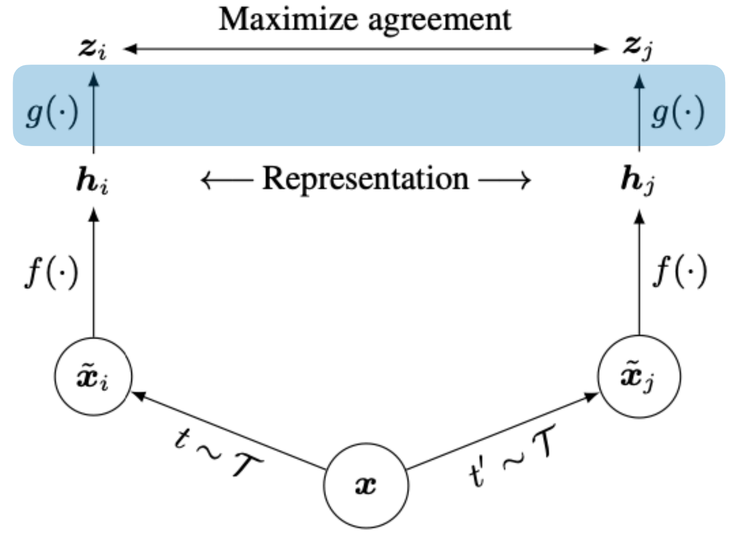
- Projection head improves performance!

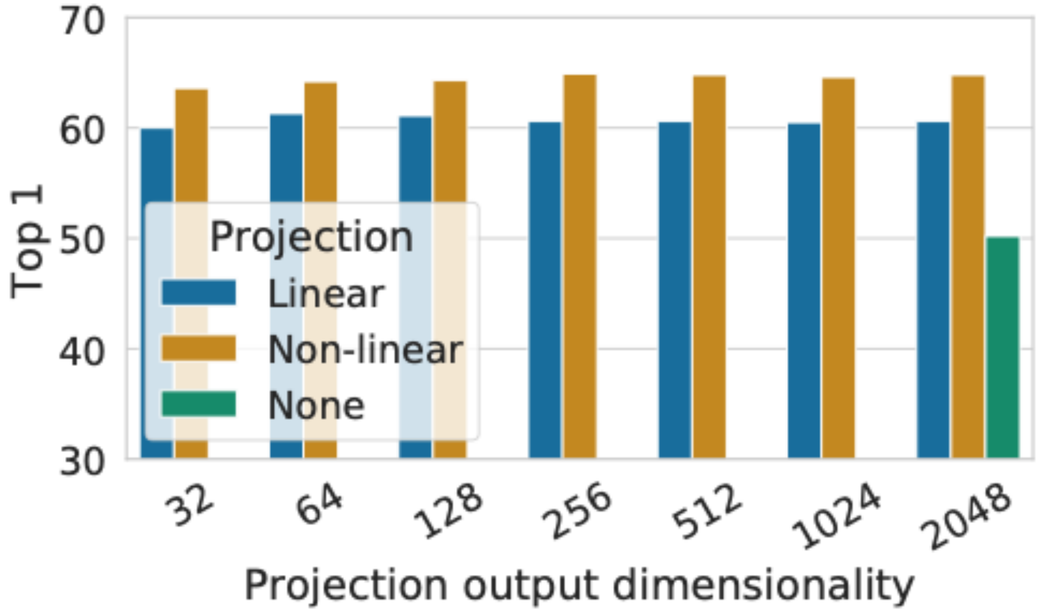


SimCLR: [Chen, Kornblith, Norouzi, Hinton, ICML 2020]

Projection head

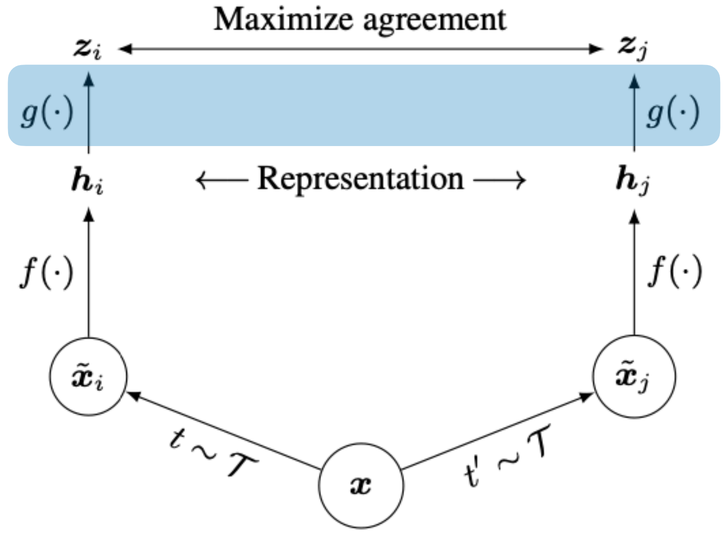
- Projection head improves performance.

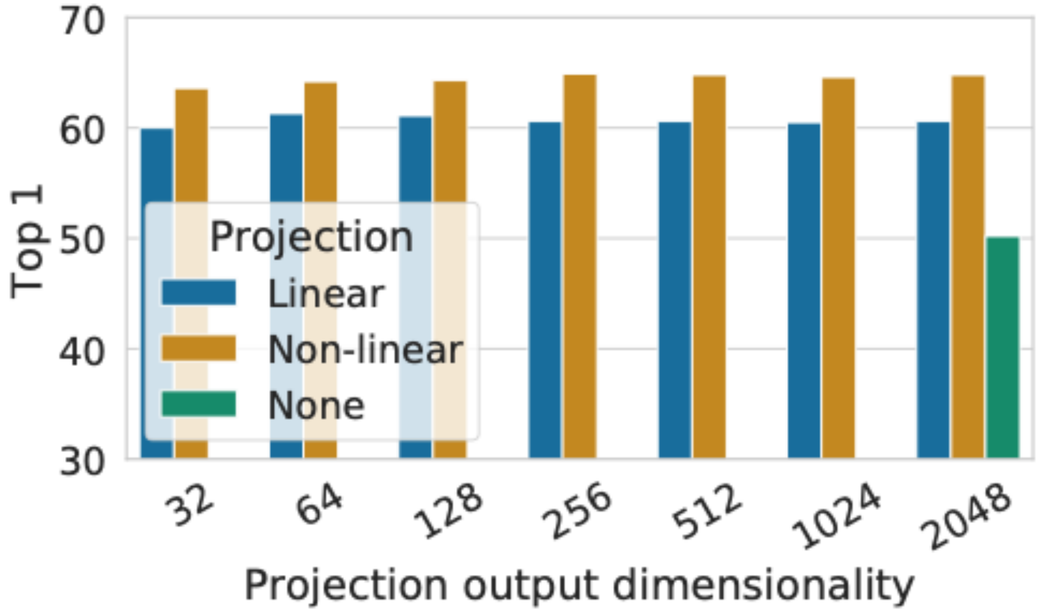




Projection head

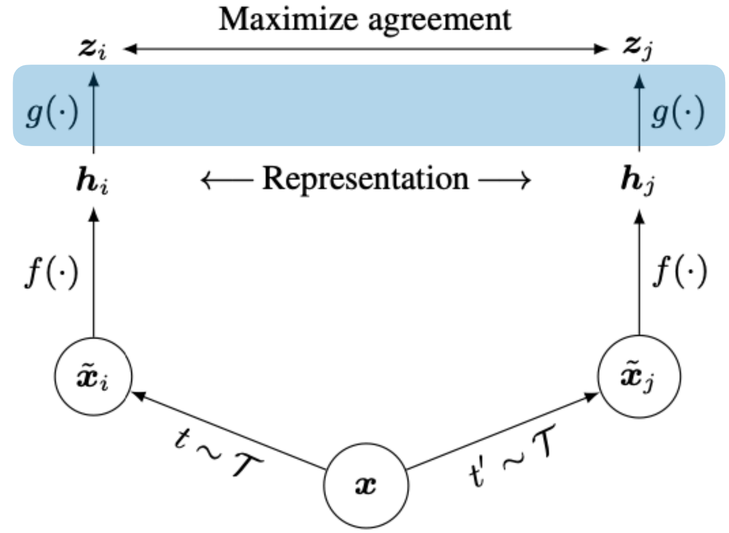
- Projection head improves performance.
- Why?
  Possibly because representation $\mathbf{h}$ then need not be completely invariant to augmentations, can retain some information

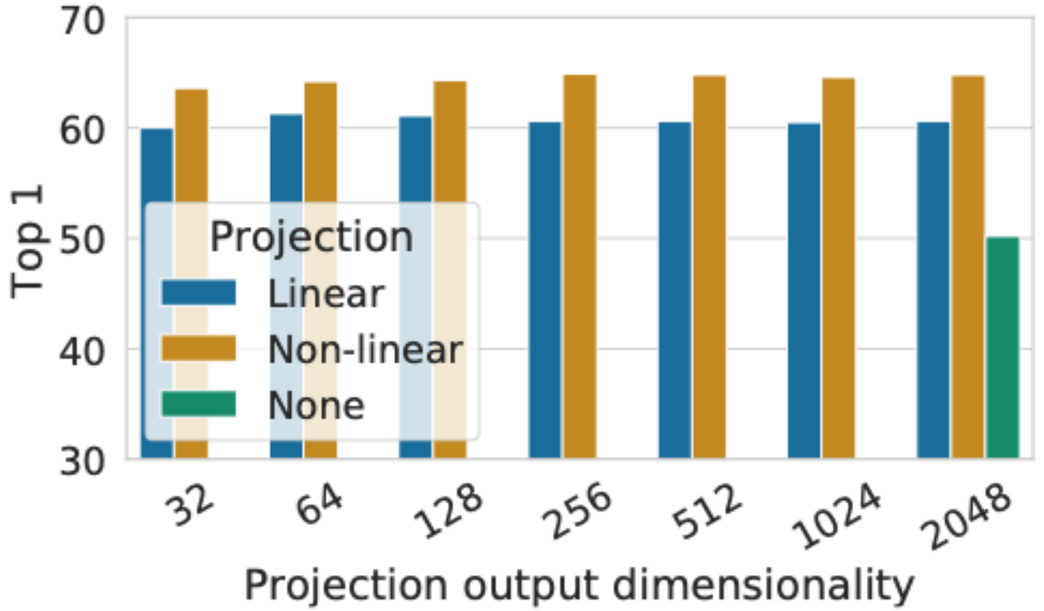




Projection head

- Projection head improves performance.
- Why?
  Possibly because representation $\mathbf{h}$ then need not be completely invariant to augmentations, can retain some information





SimCLR: [Chen, Kornblith, Norouzi, Hinton, ICML 2020]

Effect of batch size

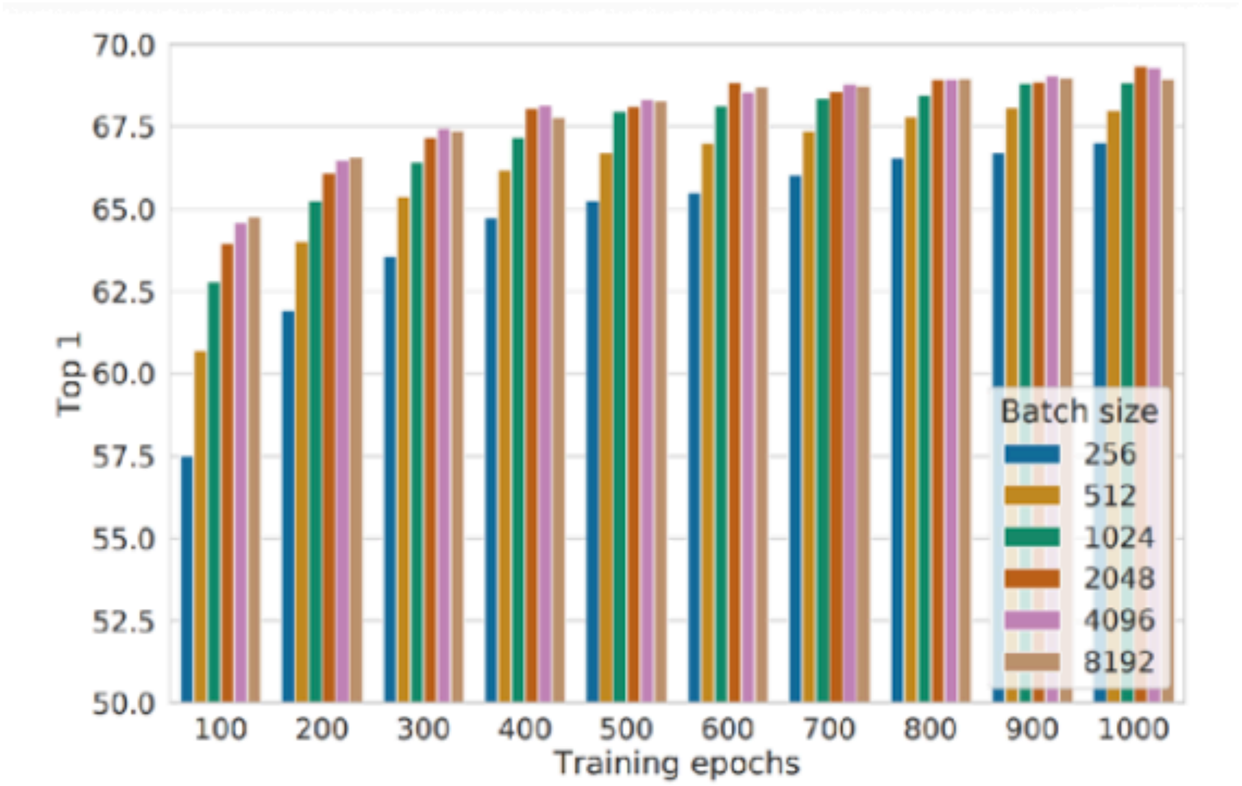


Figure 9. Linear evaluation models (ResNet-50) trained with different batch size and epochs. Each bar is a single run from scratch.[10]

(Figure from Chen et al. 2020)

- SimCLR uses all points in a batch as negative examples for a positive pair
- needs large number of negative pairs = large batch sizes
- Expensive. Newer methods make this more efficient (like MoCo, *He et al. 2020*)

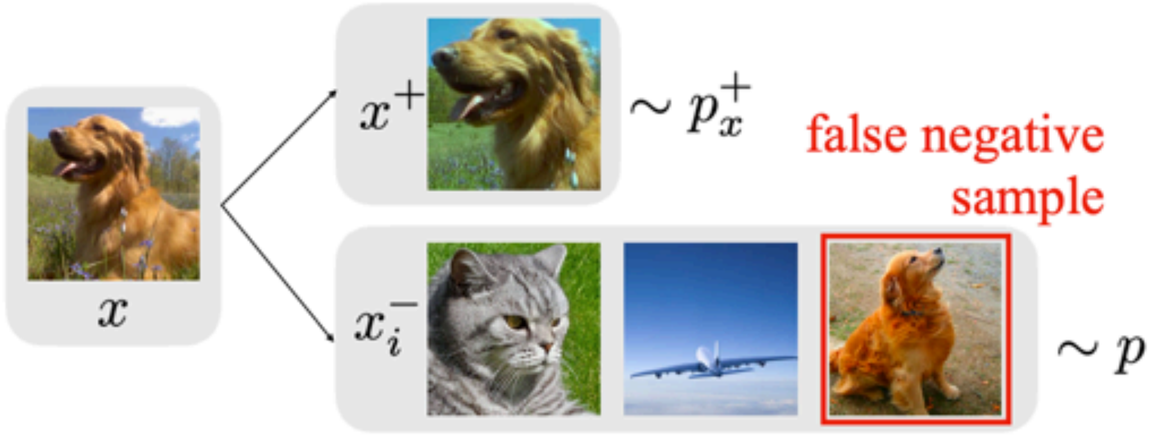
Improving negative samples

- We are pushing apart negative pairs. Negative pairs are random pairs from the data.

figure: Chuang et al, Debiased contrastive learning

Improving negative samples

- We are pushing apart negative pairs. Negative pairs are random pairs from the data.



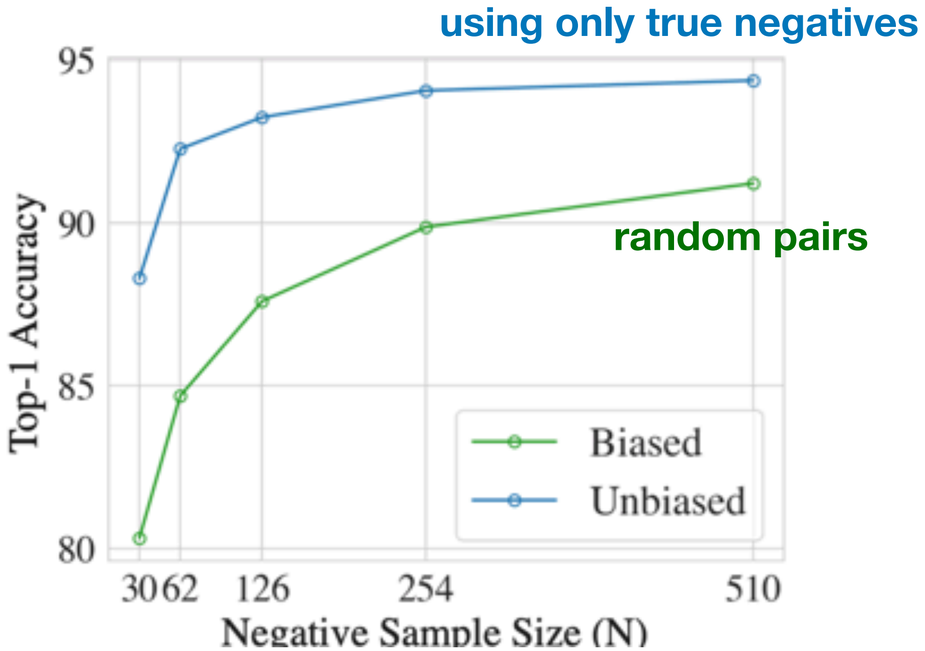


figure: Chuang et al, Debiased contrastive learning

I-Con: A Unifying Framework for Representation Learning

Shaden Alshammari[1] **John Hershey**[2] **Axel Feldmann**[1] **William Freeman**[1,2] **Mark Hamilton**[1,3]
[1] *MIT* [2] *Google* [3] *Microsoft*

https://aka.ms/i-con

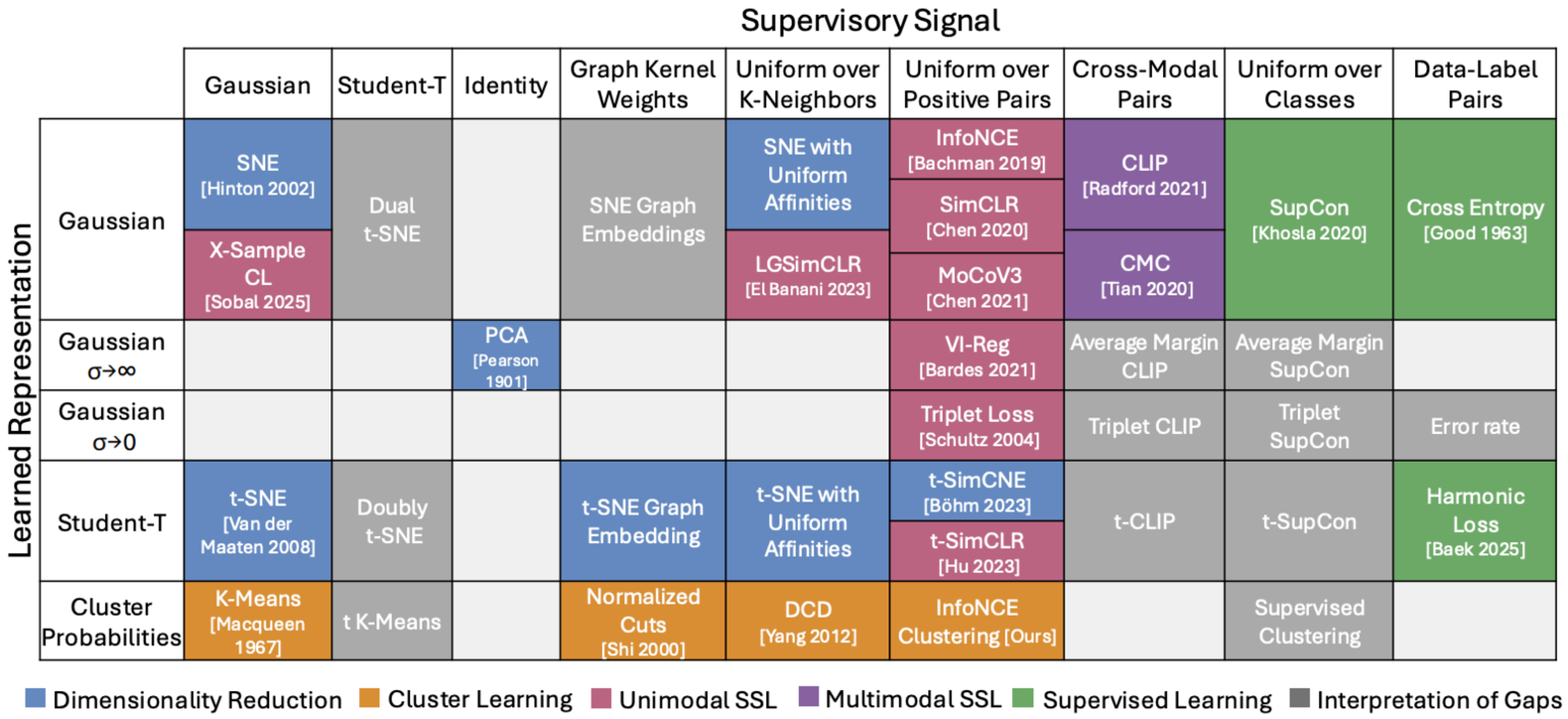
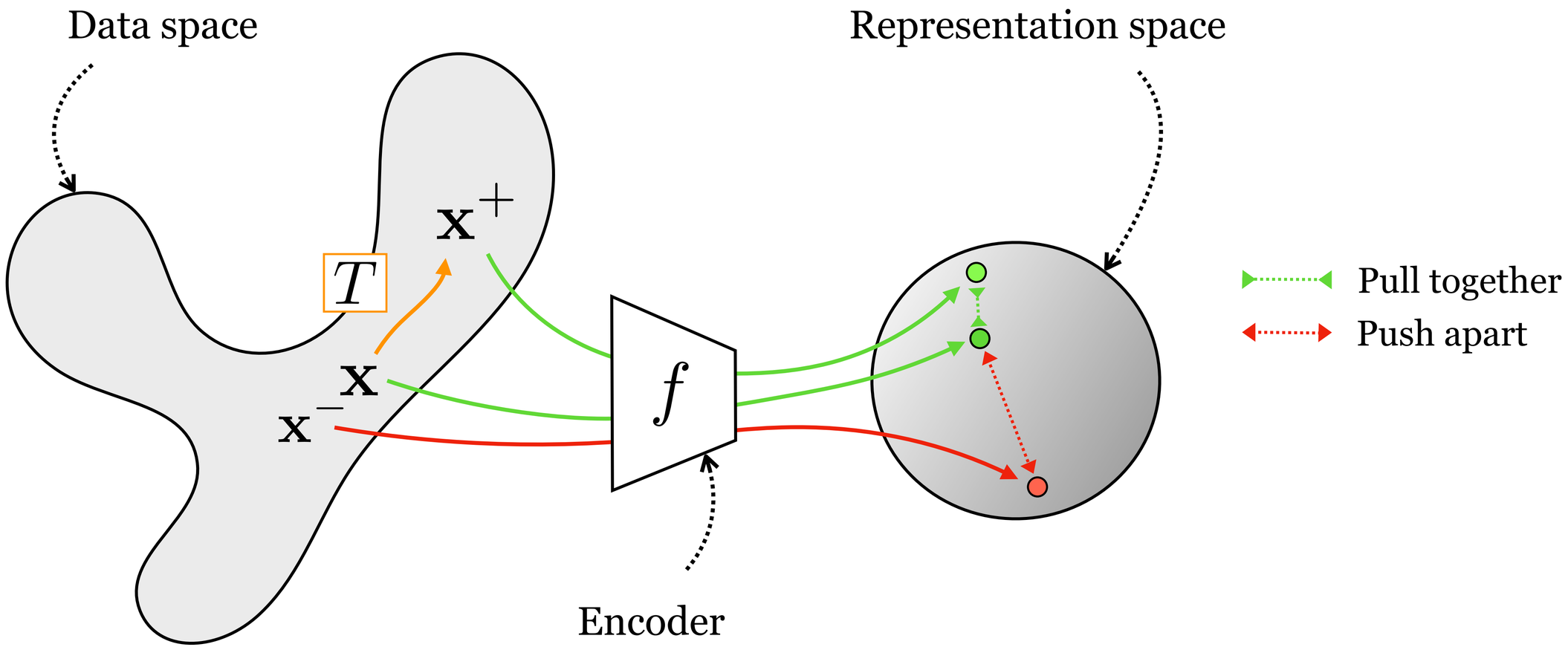


Figure 1: **A "periodic" table of representation learning methods unified by the I-Con framework.** By choosing different types of conditional probability distributions over neighbors, I-Con generalizes over 23 commonly used representation learning methods.

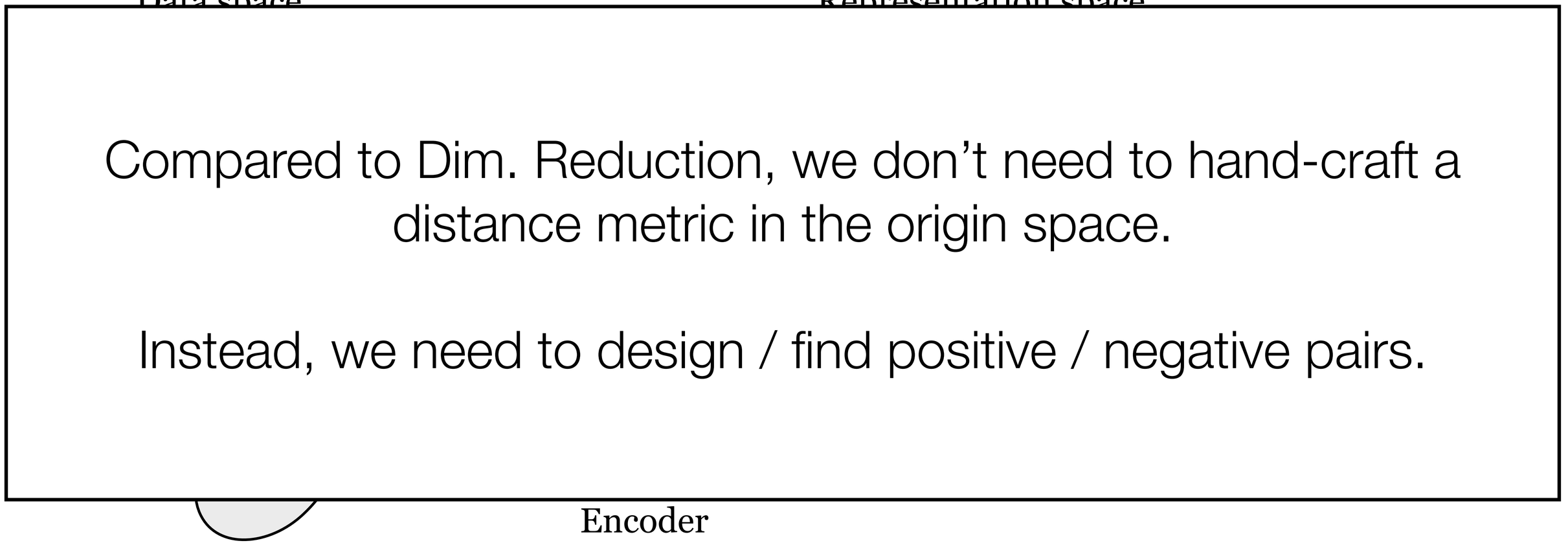
- Has some more cool analysis on the idea of "true negatives" etc

Contrastive Learning



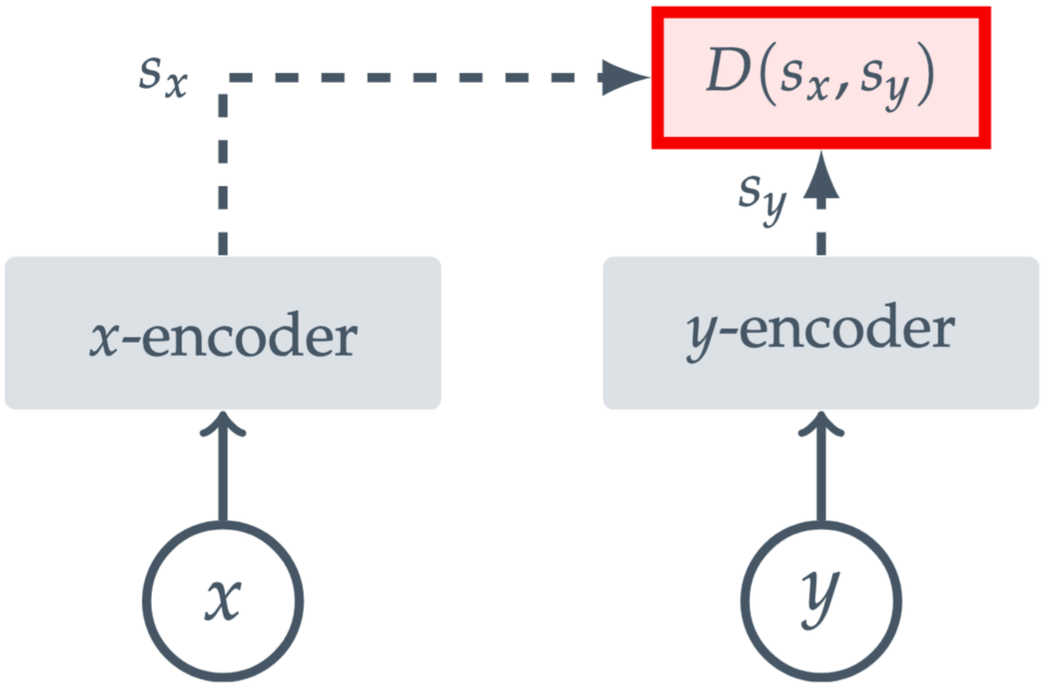
[Slide credit: Phillip Isola]

Contrastive Learning



[Slide credit: Phillip Isola]

Problem: Collapse / Trivial Solution



Trivial solution to

$$\min_{s_x, s_y} D(s_x, s_y)$$

Is always $s_x = s_y = c$ for some constant c (such as 0).

Problem: Collapse / Trivial Solution

Trivial solution to

Any other solutions?

is always $s_x = s_y = c$ for some constant c (such as 0).

DINO

DINO

DINO

Two crops:

One “global”, one “local”.

I.e, one with more info, one with less.

Heavy augmentation on both.

DINO

Student

Teacher

SOFTMAX

CENTER

SOFTMAX

DINO

Student

Teacher

SOFTMAX

CENTER

SOFTMAX

DINO

Student

Teacher

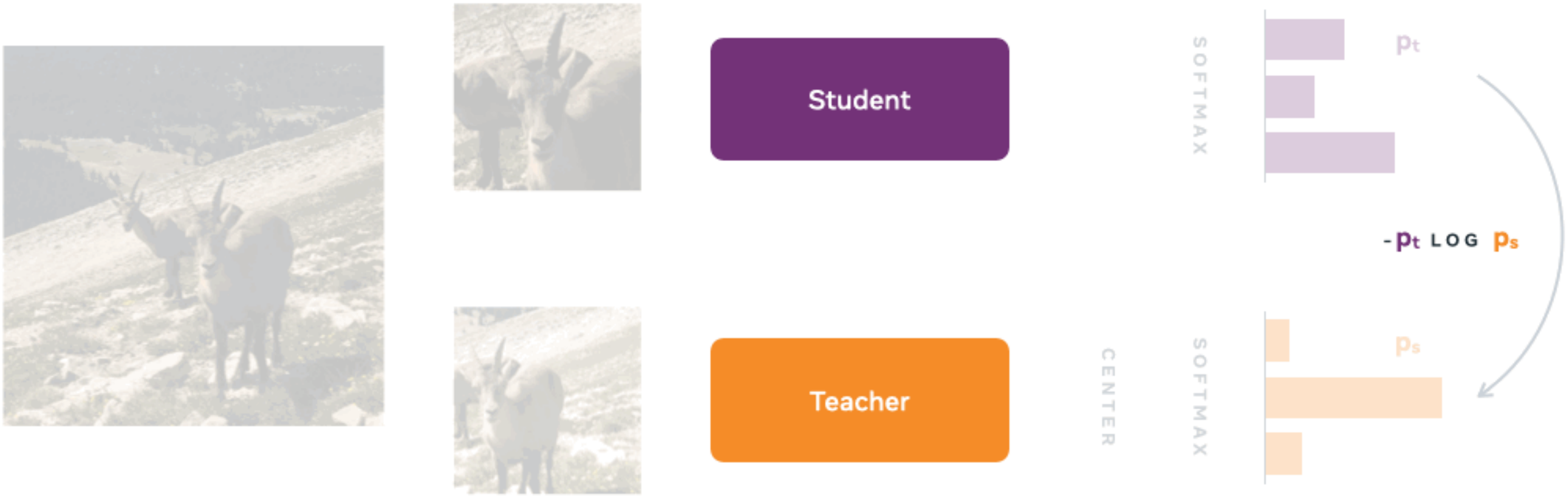
SOFTMAX

CENTER

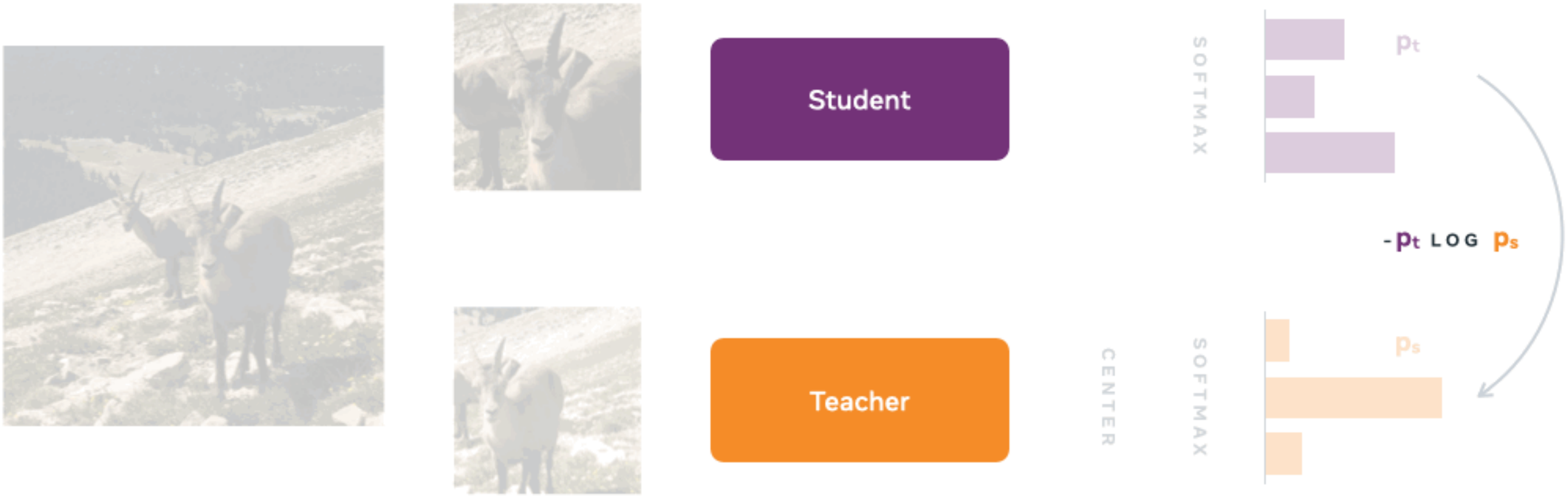
SOFTMAX

Pass through VIT. Get final global tokens. Expand into *high-dimensional 1D embedding.* Compute Softmax, compare with Cross-Entropy.

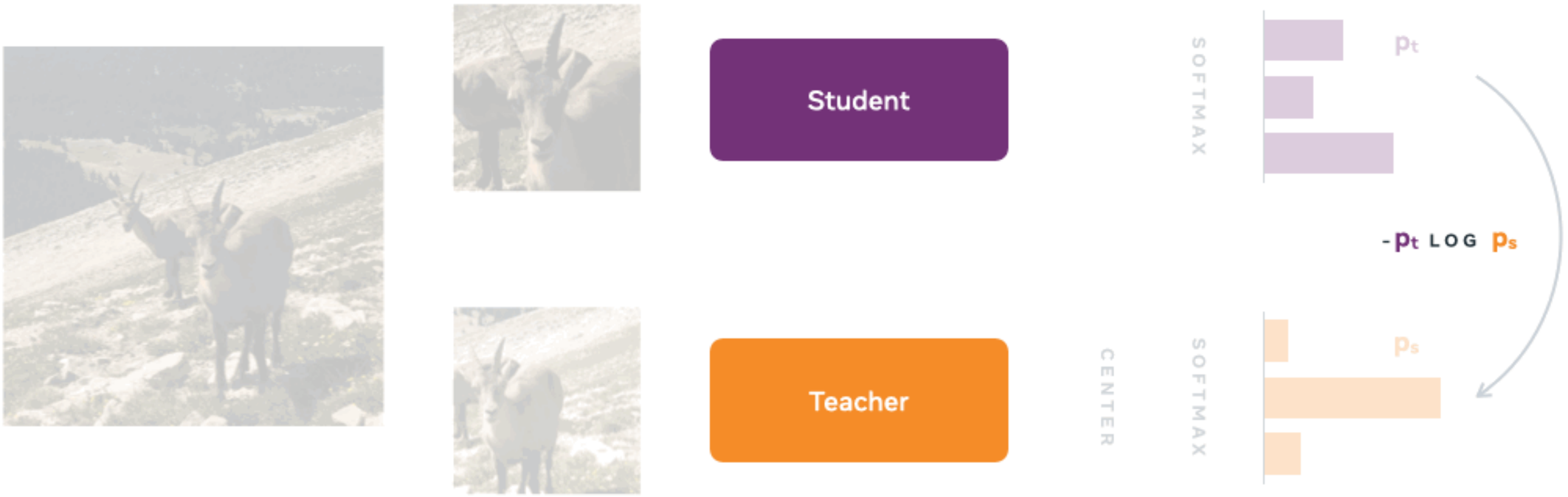
DINO



DINO

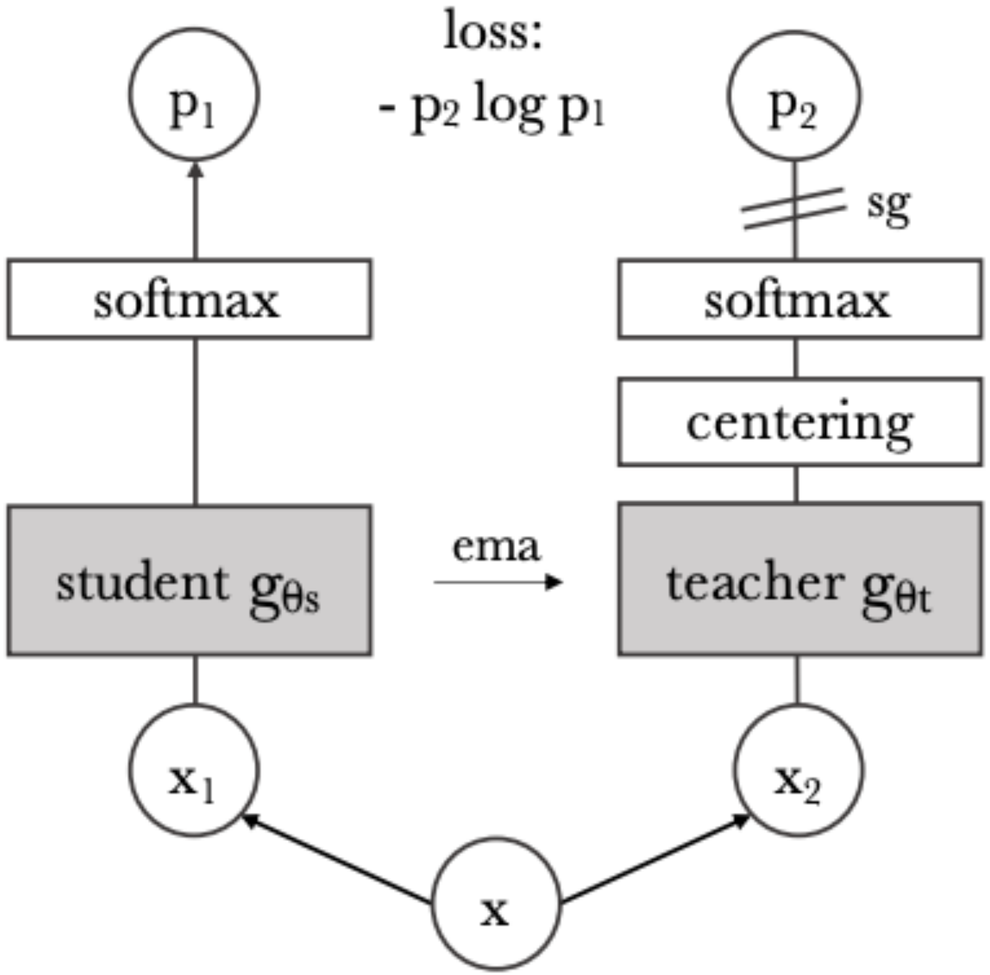


DINO



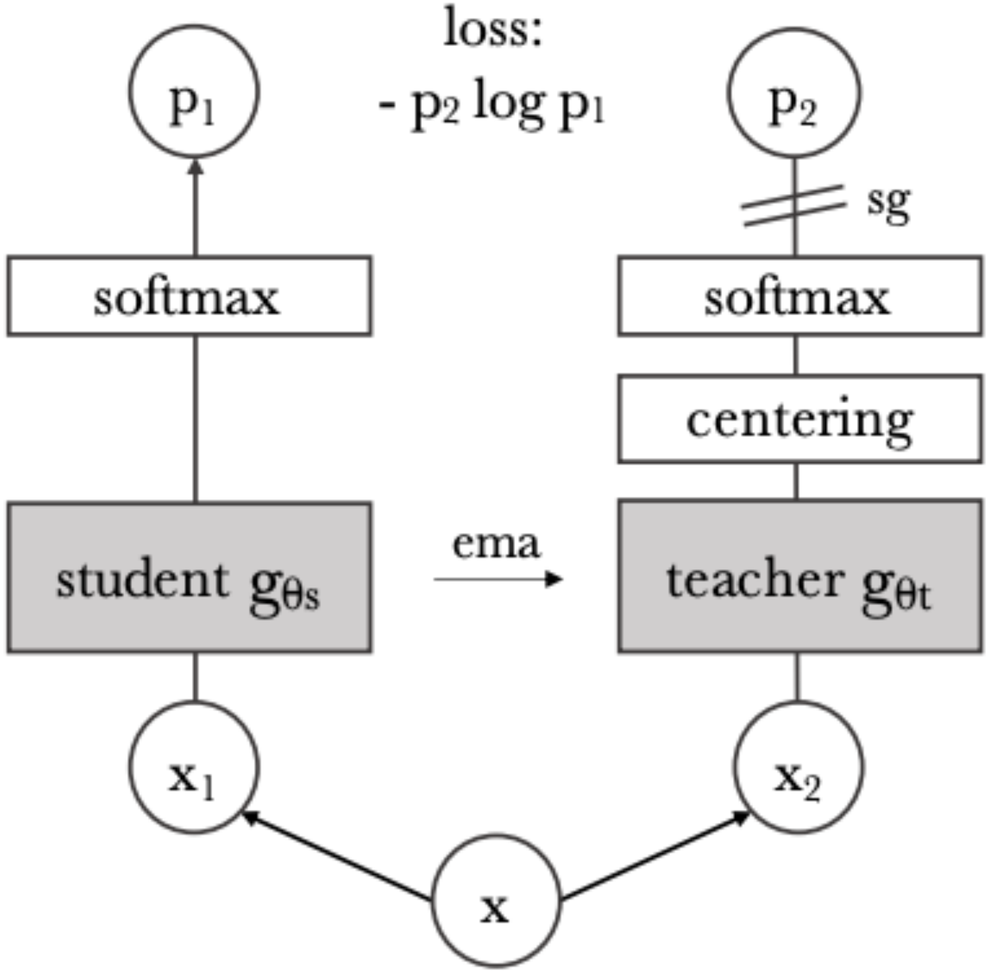
Gradients are backdropped only through student (the one who saw the “local” crop). Teacher is updated via Exponential Moving Average!

Collapse Prevention: Momentum Encoder (DINO)



[DINO, Caron et al. 2022; Momentum Contrast…, He et al., 2022]

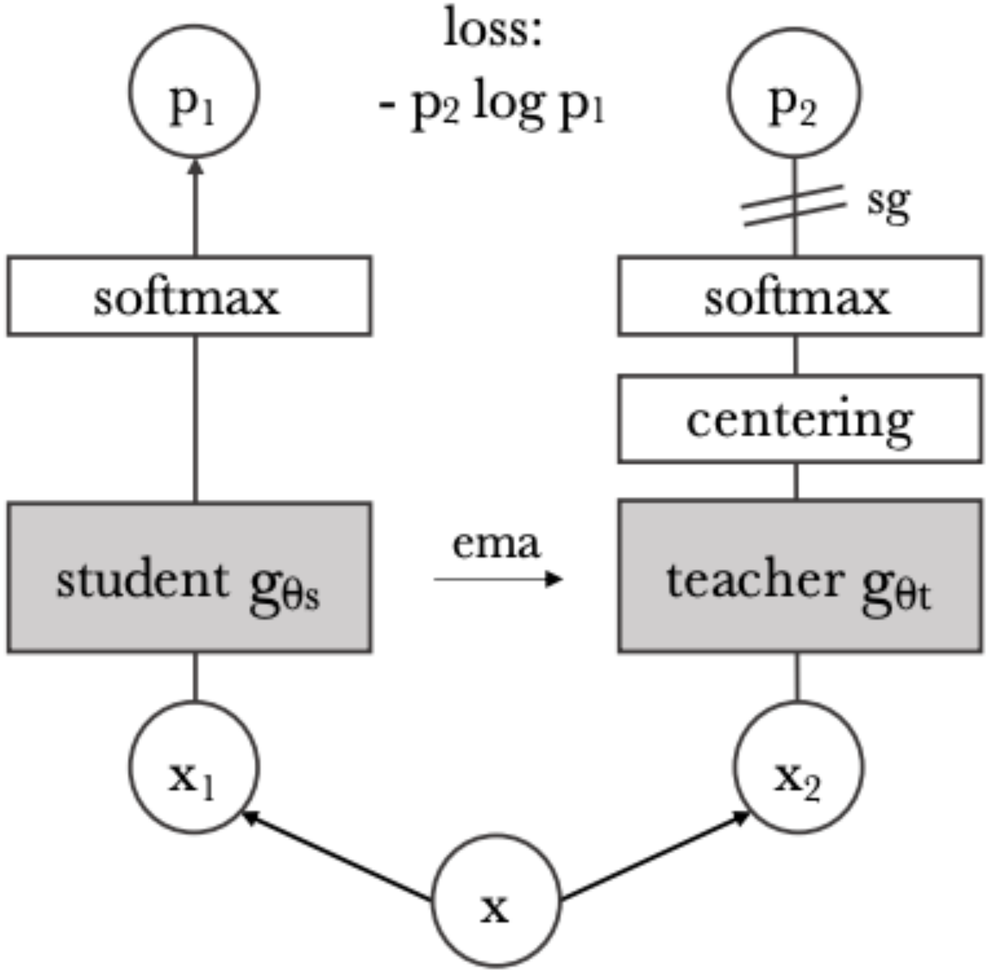
Collapse Prevention: Momentum Encoder (DINO)



DINO embeddings are produced from global tokens of MLP.

[DINO, Caron et al. 2022; Momentum Contrast…, He et al., 2022]

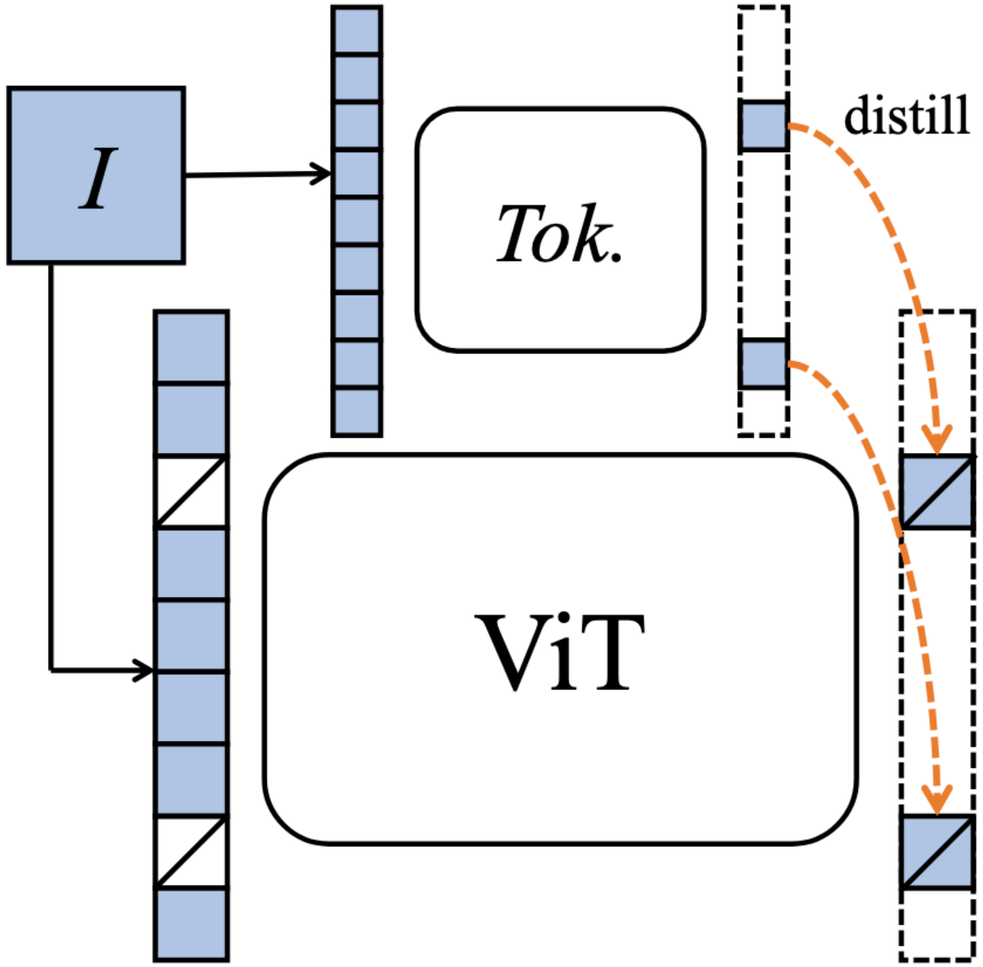
Collapse Prevention: Momentum Encoder (DINO)



I.e., only loss is on global tokens!

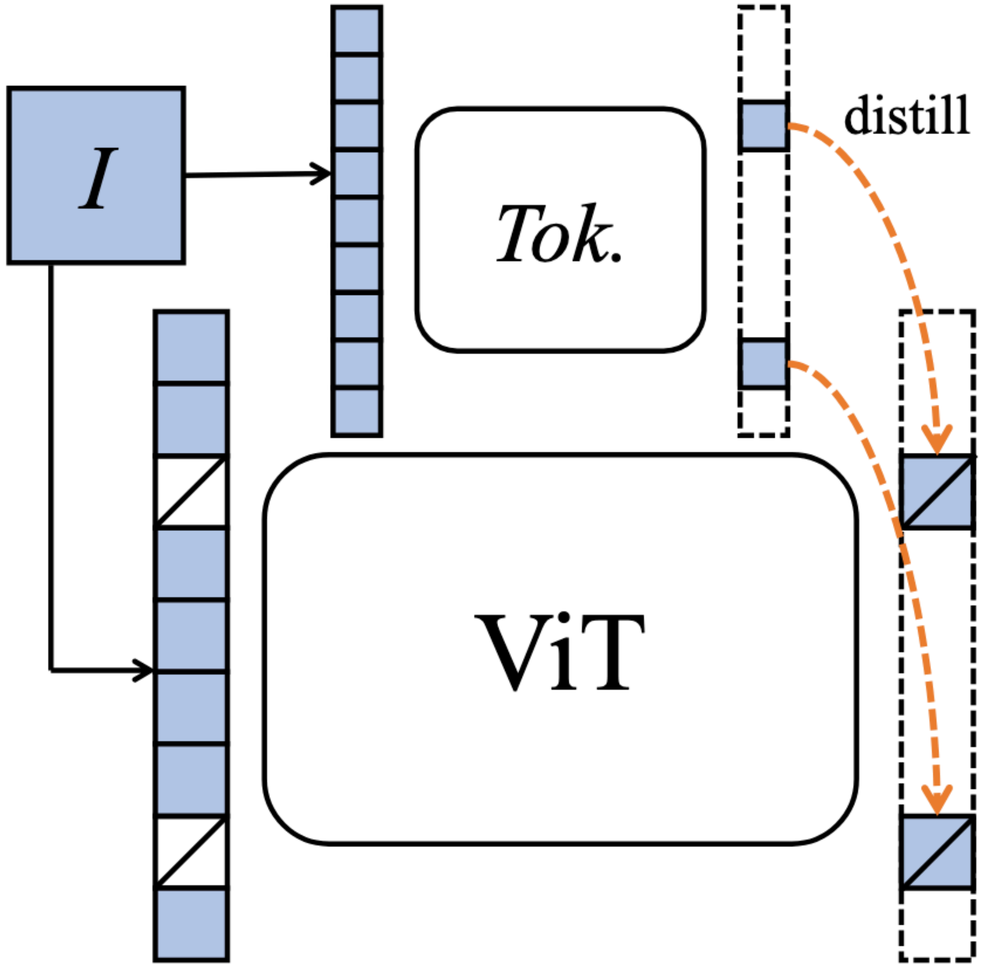
[DINO, Caron et al. 2022; Momentum Contrast…, He et al., 2022]

DINO v2: Adds another iBOT head - essentially, DINO on masked *local* tokens.



[DINOv2, Oquab et al 2024; iBOT, Zhou et al. 2024; DINO, Caron et al. 2022; Momentum Contrast…, He et al., 2022]

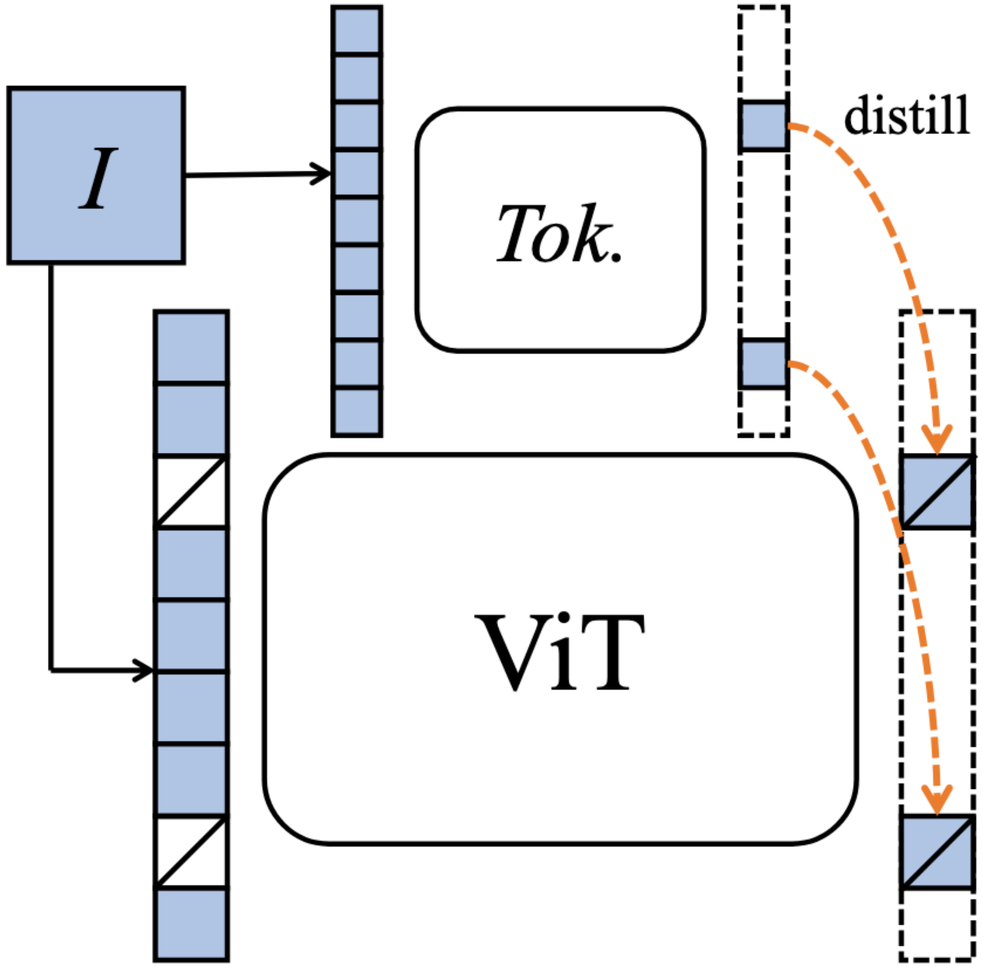
DINO v2: Adds another iBOT head - essentially, DINO on masked *local* tokens.



Randomly mask some of the input patches given to the student, but not to the teacher.

[DINOv2, Oquab et al 2024; iBOT, Zhou et al. 2024; DINO, Caron et al. 2022; Momentum Contrast…, He et al., 2022]

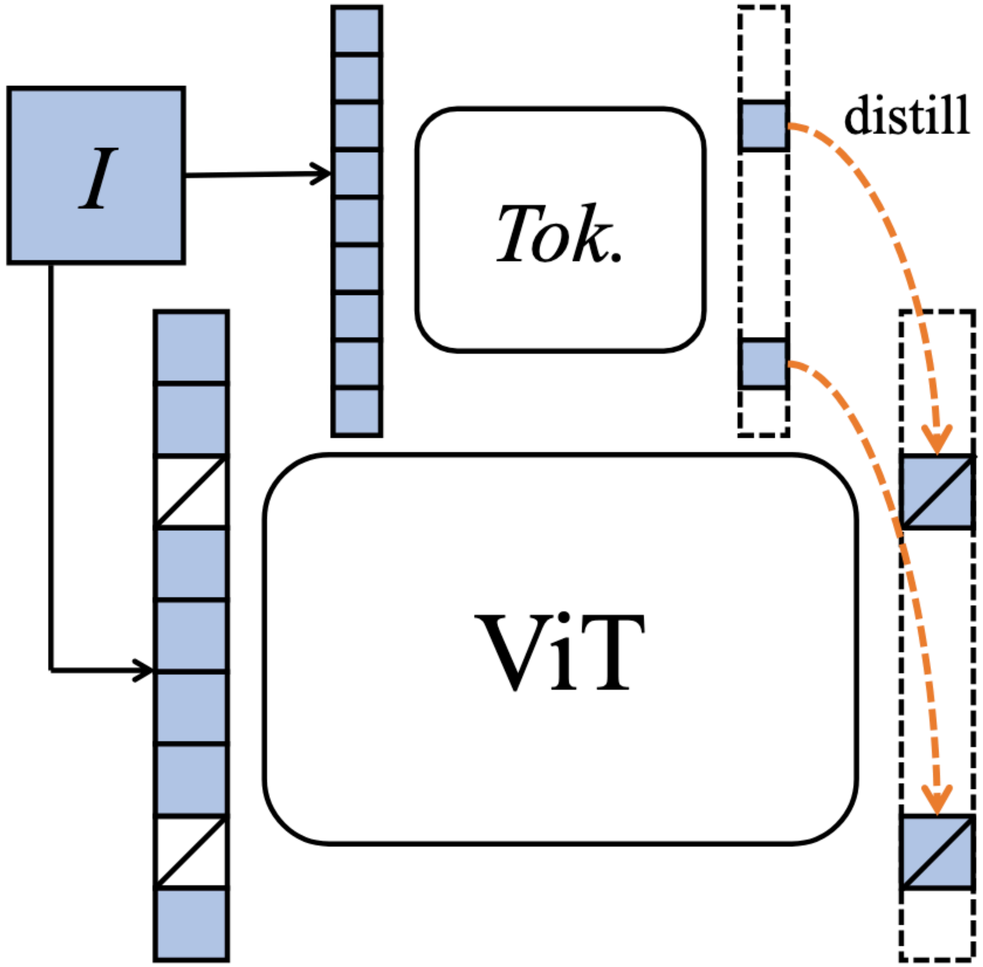
DINO v2: Adds another iBOT head - essentially, DINO on masked *local* tokens.



Generate high-dim softmax embedding from each student mask token as well as corresponding unmasked teacher tokens.

[DINOv2, Oquab et al 2024; iBOT, Zhou et al. 2024; DINO, Caron et al. 2022; Momentum Contrast…, He et al., 2022]

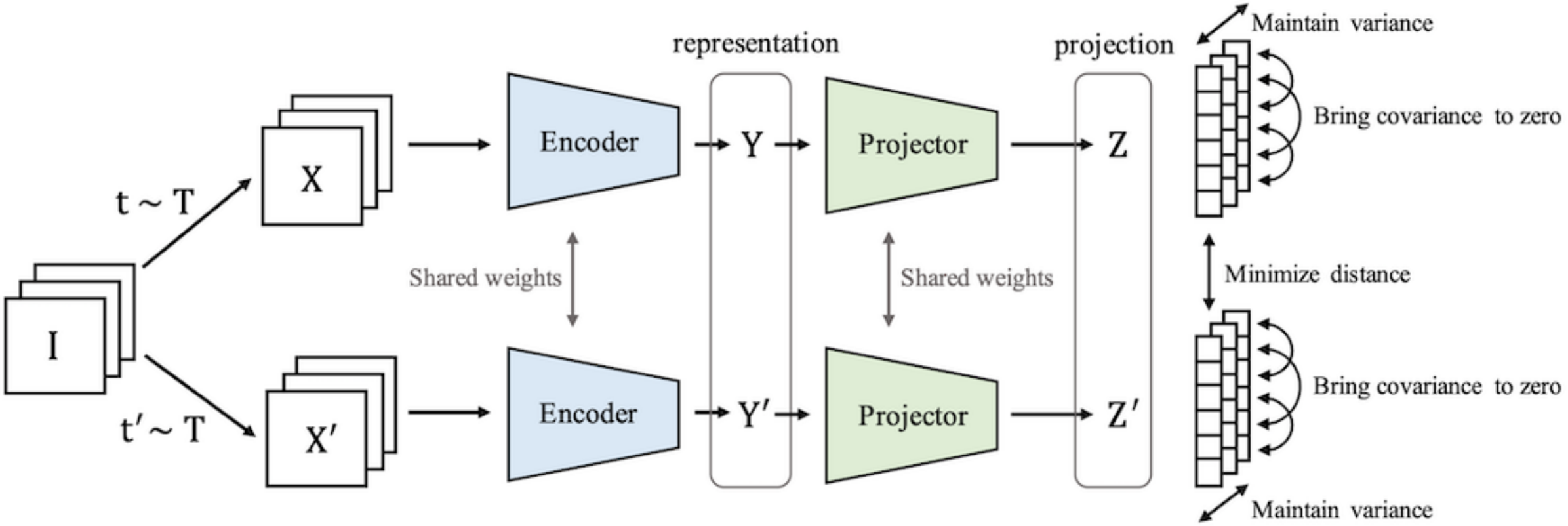
DINO v2: Adds another iBOT head - essentially, DINO on masked *local* tokens.



DINO loss between.

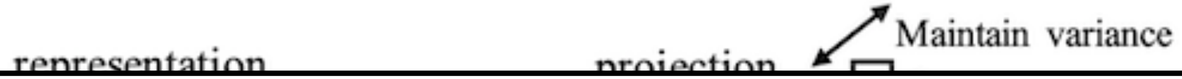
[DINOv2, Oquab et al 2024; iBOT, Zhou et al. 2024; DINO, Caron et al. 2022; Momentum Contrast…, He et al., 2022]

Collapse Prevention: Variance-Covariance Regularization



[VICReg, Bardes et al. 2022]

Collapse Prevention: Variance-Covariance Regularization



"Maintaining Variance" regularization has an interesting parallel to "uniformity" argument made by Wang & Isola!

[VICReg, Bardes et al. 2022]

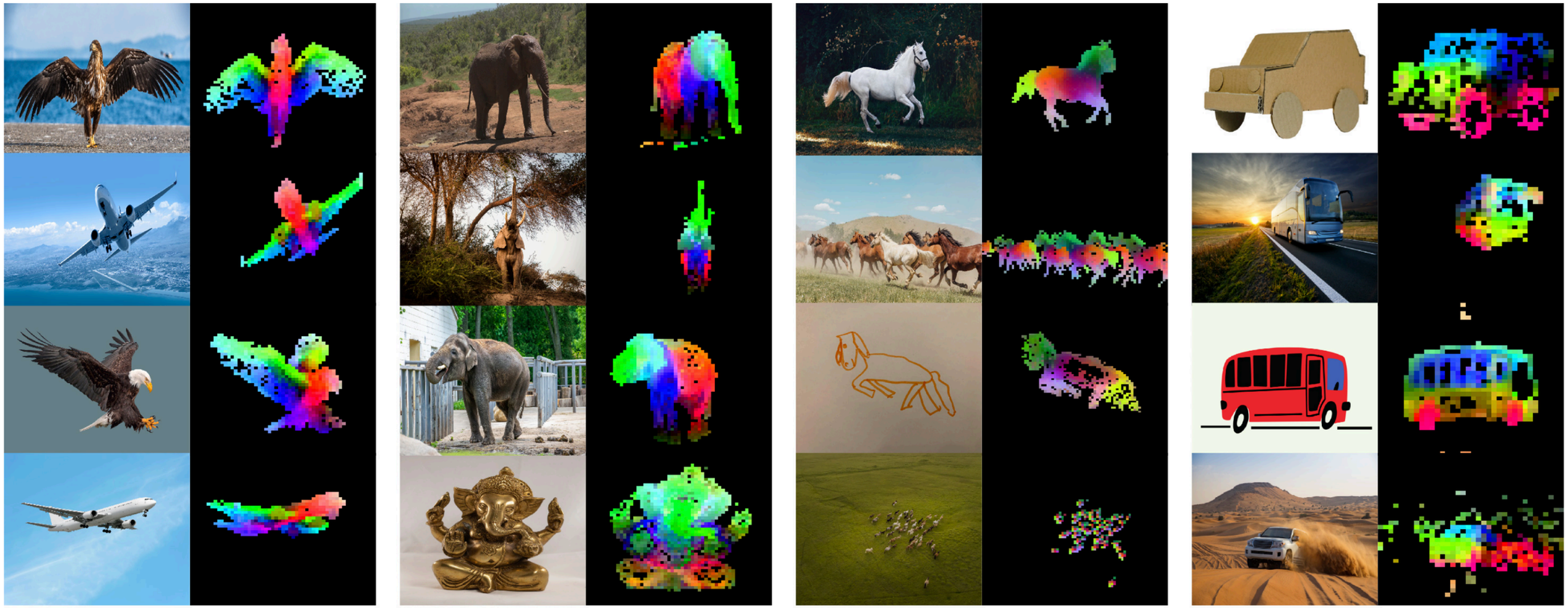


Figure 1: **Visualization of the first PCA components.** We compute a PCA between the patches of the images from the same column (a, b, c and d) and show their first 3 components. Each component is matched to a different color channel. Same parts are matched between related images despite changes of pose, style or even objects. Background is removed by thresholding the first PCA component.

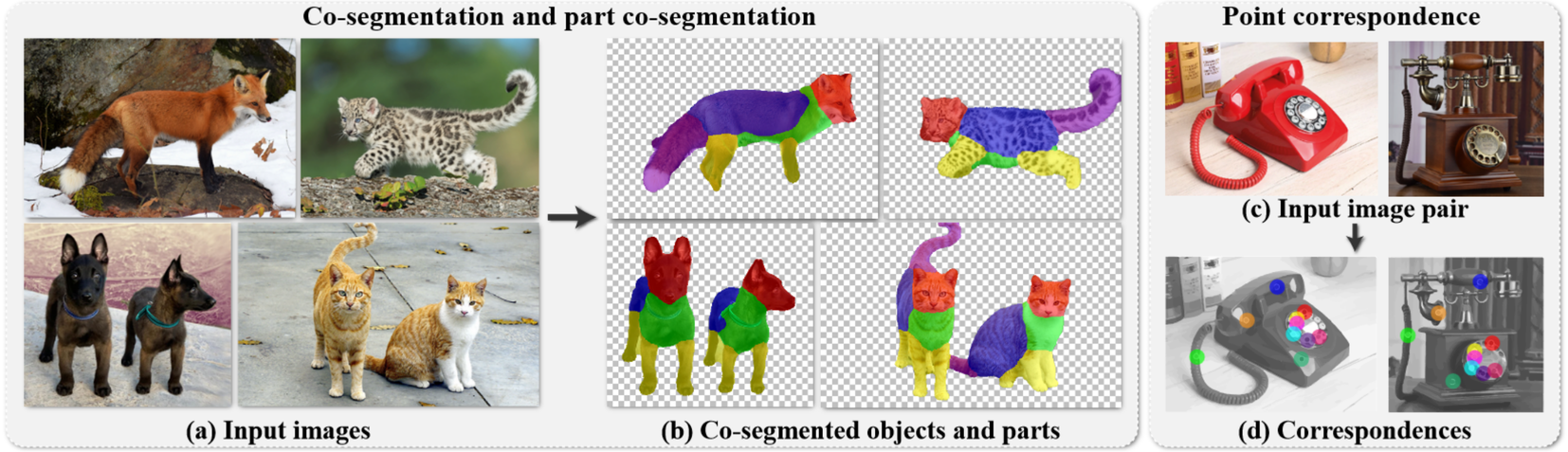
From DinoV2: Learning Robust Visual Features without Supervision, Oquab et al. 2024

DINO features are popular

[Slide credit: Christian Rupprecht]

DINO features are popular

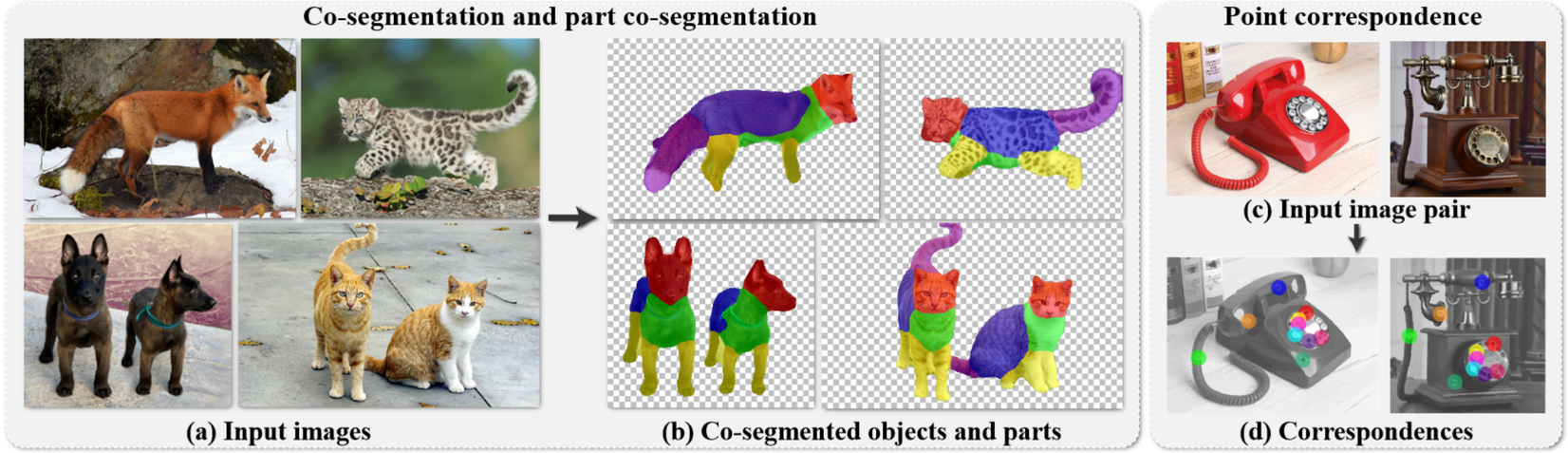
[Slide credit: Christian Rupprecht]



[Amir et al.; CVPR ‘22]

DINO features are popular

[Slide credit: Christian Rupprecht]

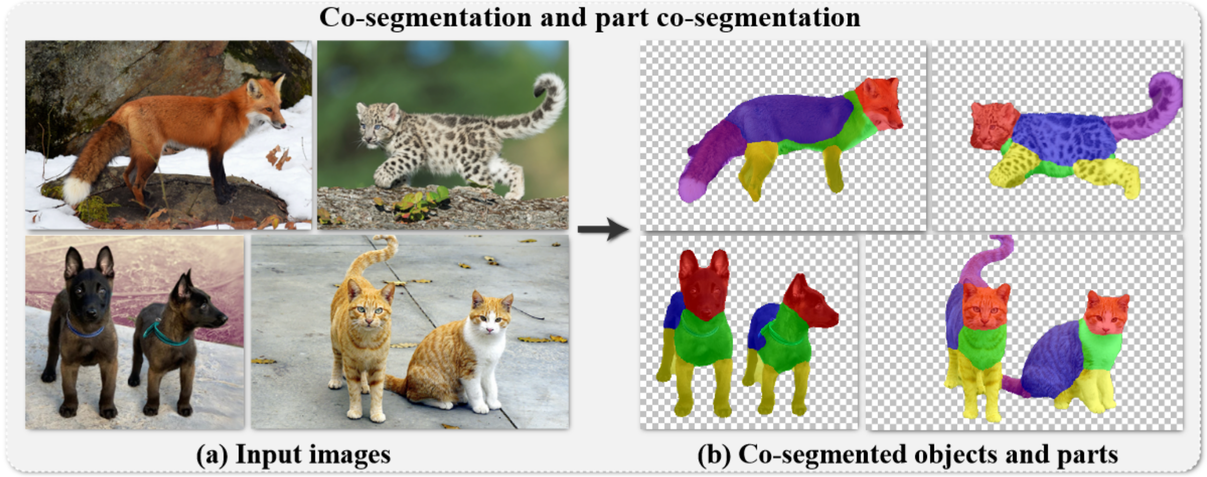


[Amir et al.; CVPR ‘22]

[Tumanyan et al.; CVPR ‘22]

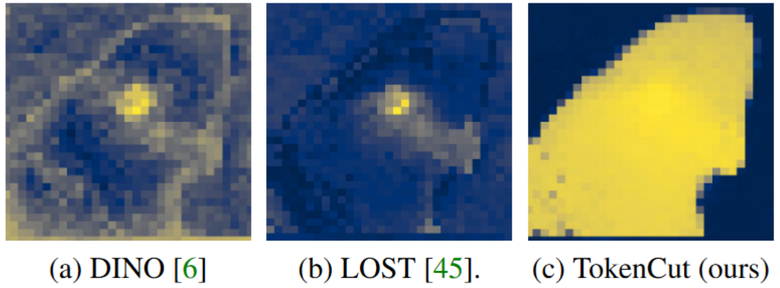
DINO features are popular

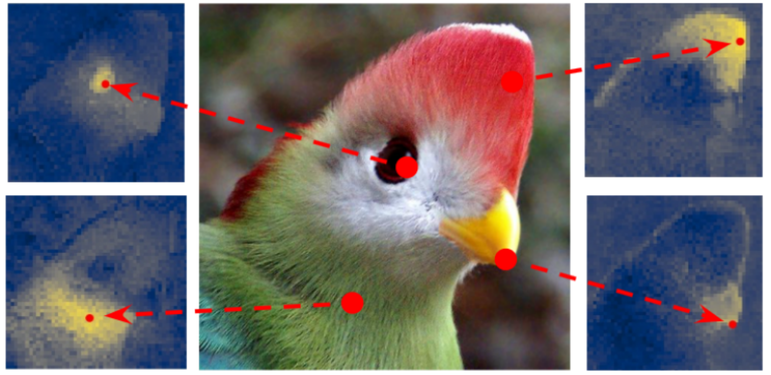
[Slide credit: Christian Rupprecht]



[Amir et al.; CVPR ‘22]

[Tumanyan et al.; CVPR ‘22]





(d) Attention maps associated to different patches

[Wang et al.; CVPR ‘22]

Novel View Synthesis via NeRF

NeRF, ECCV 2021

Plenoxels, CVPR 2022

Novel View Synthesis via NeRF

NeRF, ECCV 2021

Plenoxels, CVPR 2022

No *Understanding* of Content!

NeRF, ECCV 2021

Plenoxels, CVPR 2022

No *Understanding* of Content!

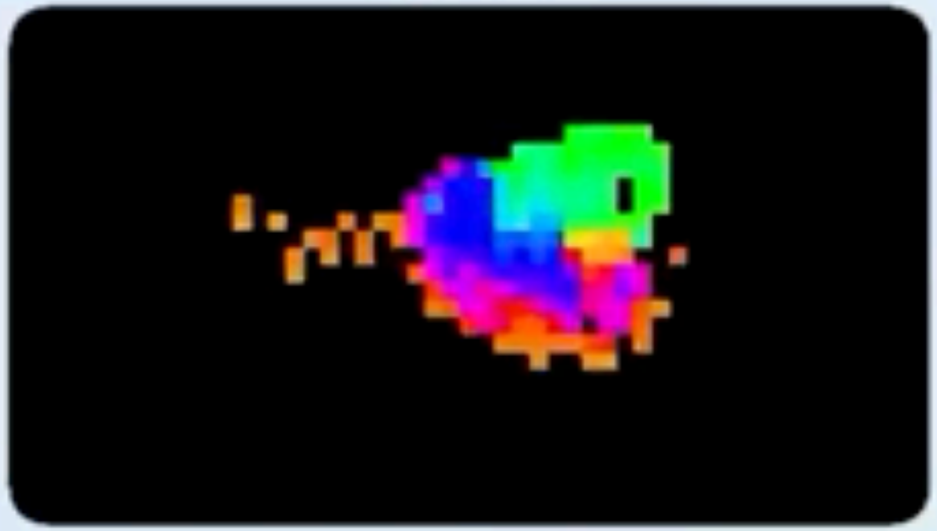
NeRF, ECCV 2021

Plenoxels, CVPR 2022

Usual formulation: Fit to RGB images.

Can we leverage **2D** vision models?

DINO

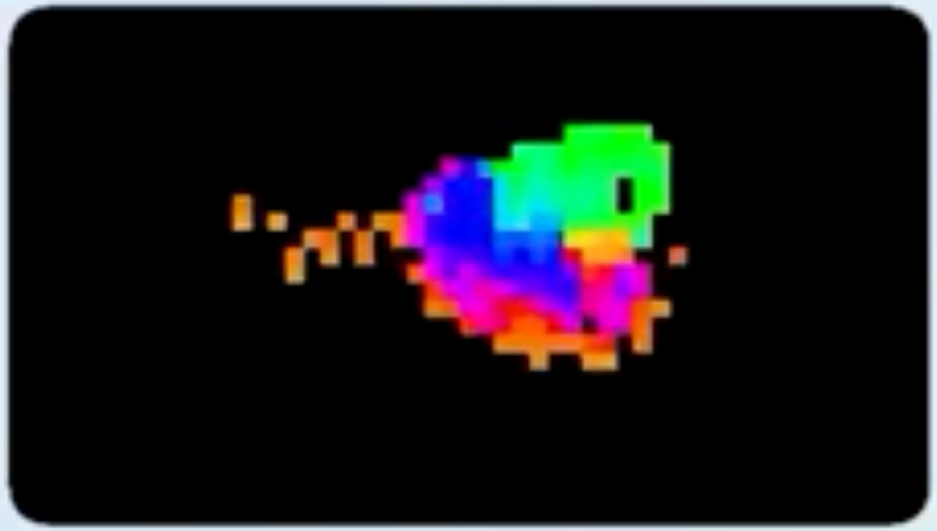


DINOv2

DINOv2, Oquab et al., arXiv 2023

Can we leverage **2D** vision models?

DINO



DINOv2

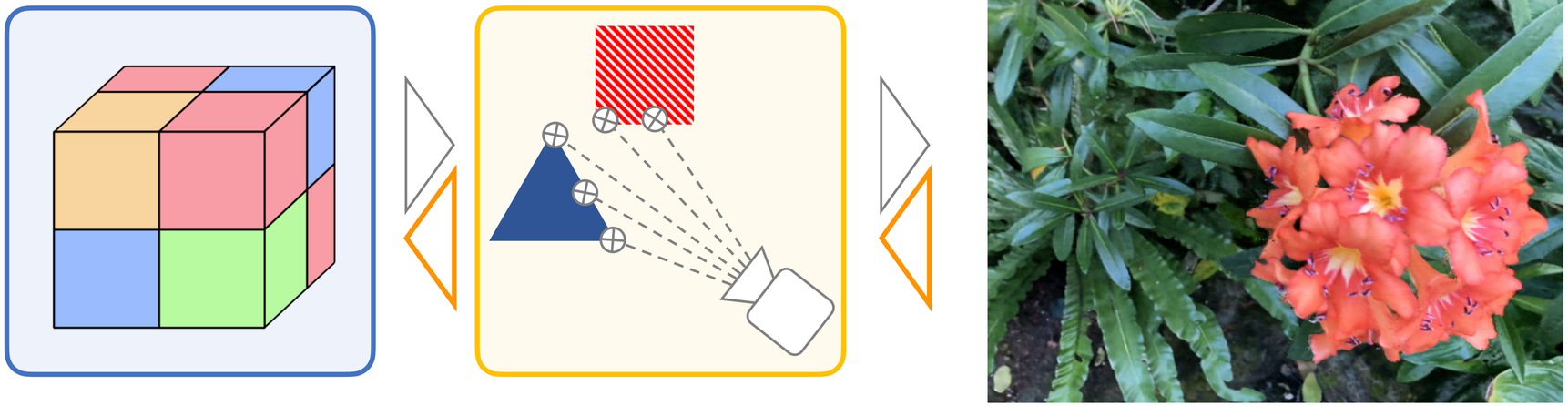
DINOv2, Oquab et al., arXiv 2023

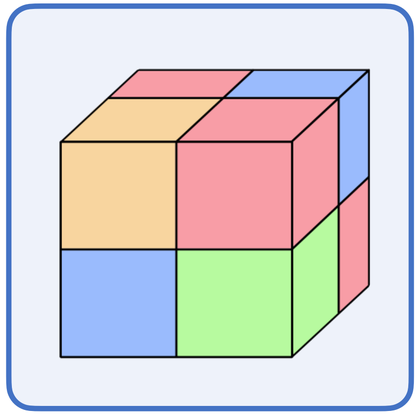
What happens if we fit on features?

DINO: Emerging properties in self-supervised vision transformers, Caron et al.

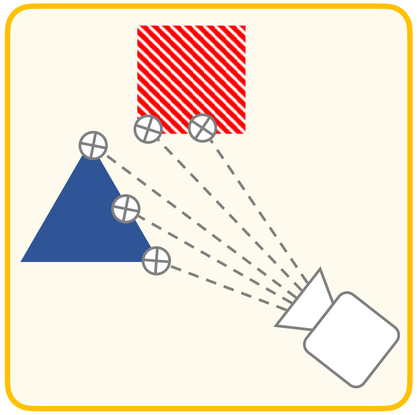
Scene Representation

Diff. Renderer





Scene Representation

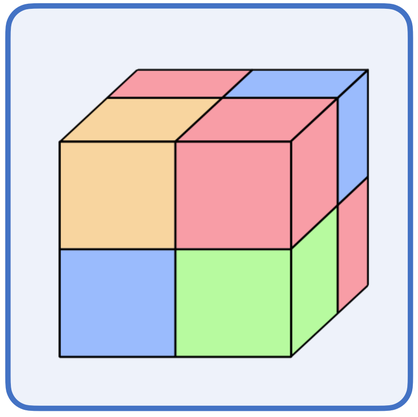


Diff. Renderer

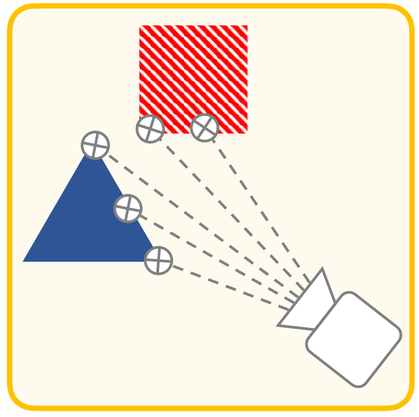
RGB

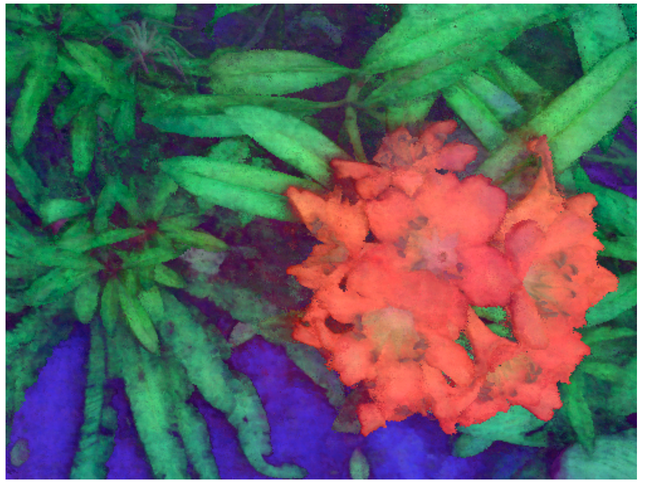
Image Feature Extractor

Scene Representation

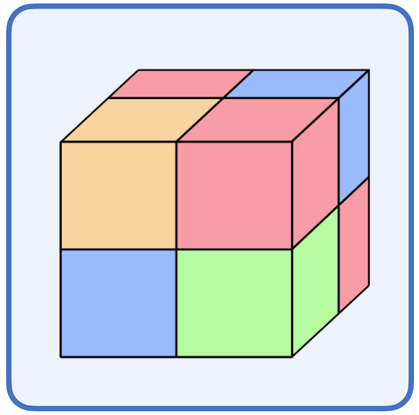


Diff. Renderer

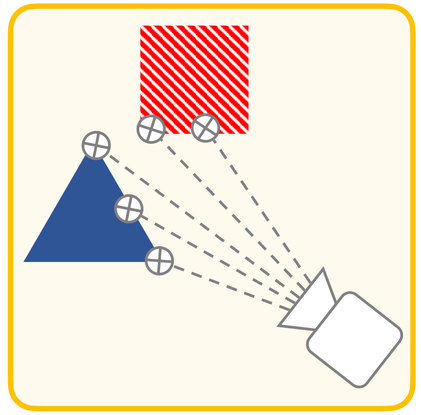




Features



Scene Representation



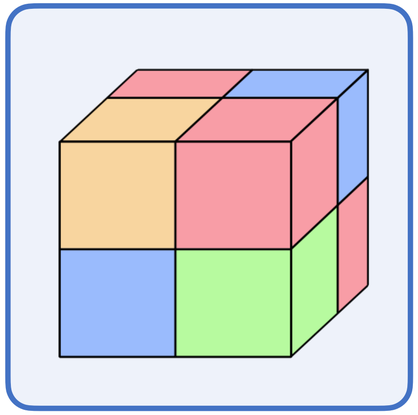
Diff. Renderer

RGB

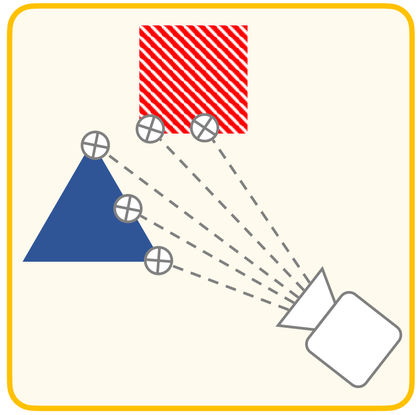
Image Feature Extractor

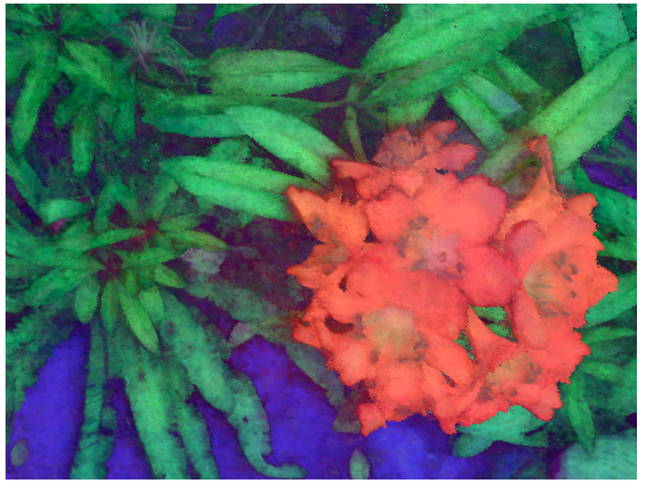
Distilled Feature Fields / 2D-to-3D Feature Fusion

Scene Representation

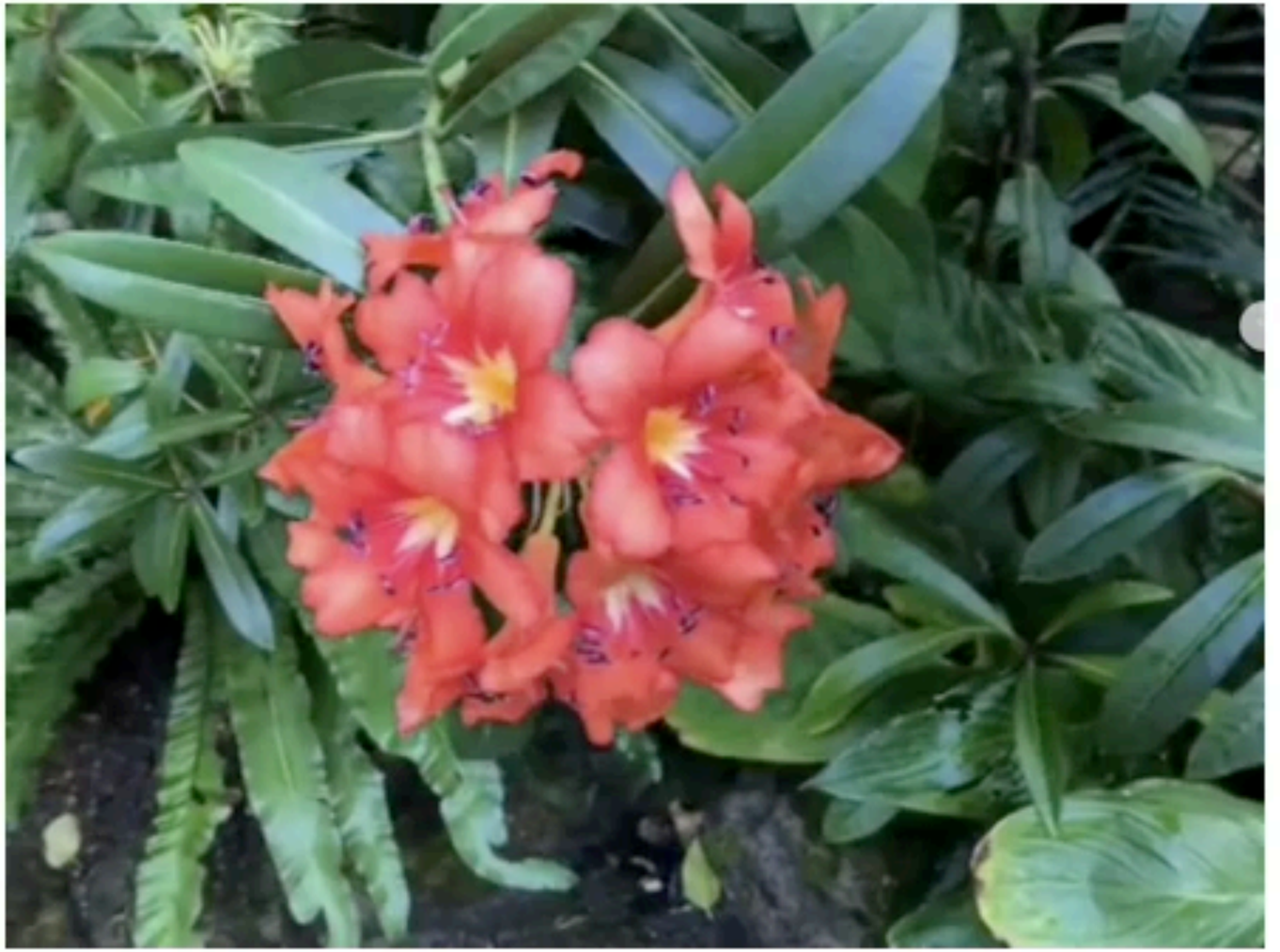


Diff. Renderer

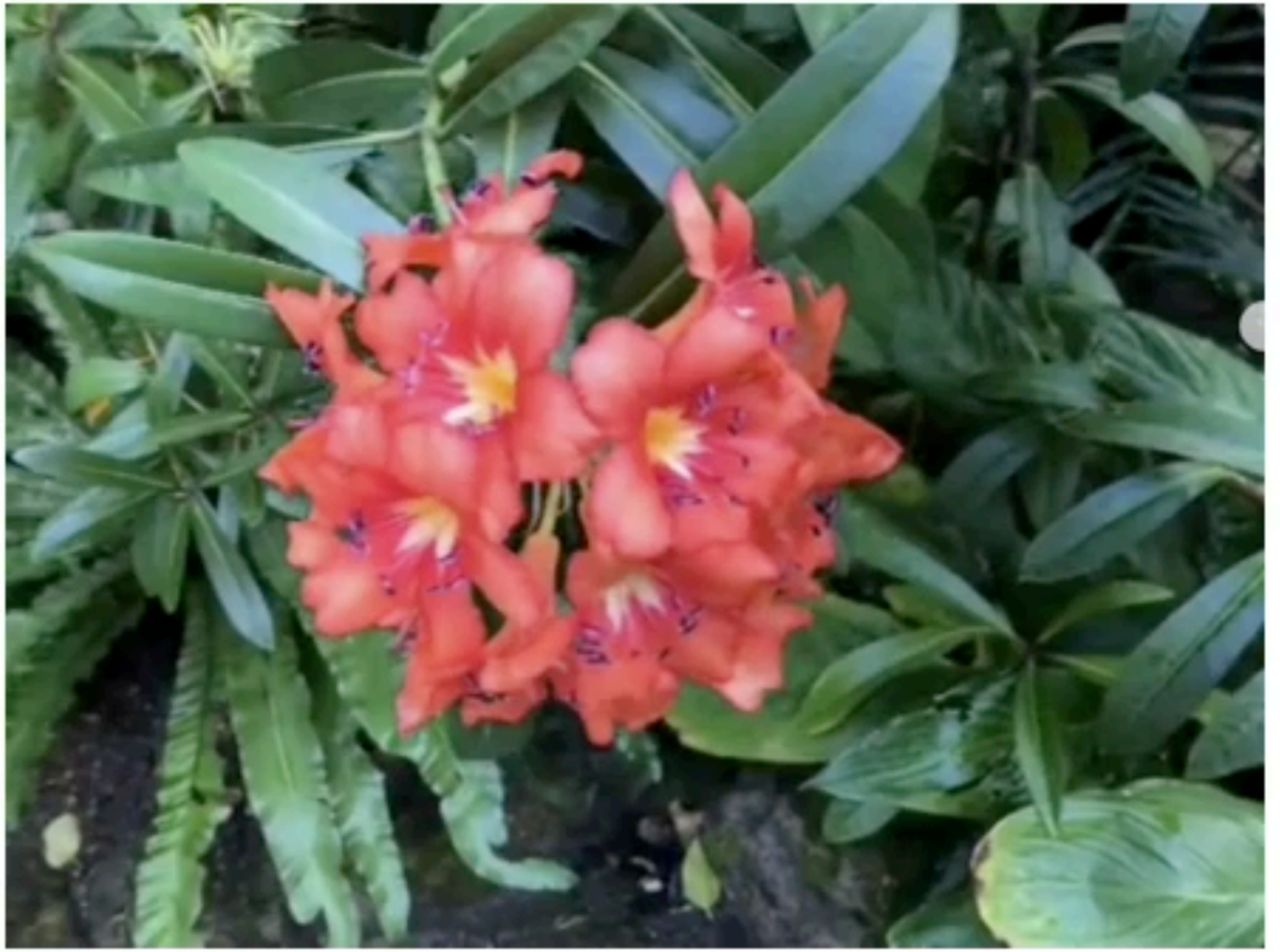




Features



Decomposing nerf for editing via feature field distillation, Kobayashi et al., NeurIPS 2022



Decomposing nerf for editing via feature field distillation, Kobayashi et al., NeurIPS 2022

Distilled Feature Fields Enable Few-Shot Manipulation

1. Scan Scene

Distilled Feature Fields Enable Few-Shot Manipulation, Shen et al., 2023

Distilled Feature Fields Enable Few-Shot Manipulation

1. Scan Scene

Distilled Feature Fields Enable Few-Shot Manipulation, Shen et al., 2023

DINO-Tracker: Taming DINO for Self-Supervised Point Tracking in a Single Video

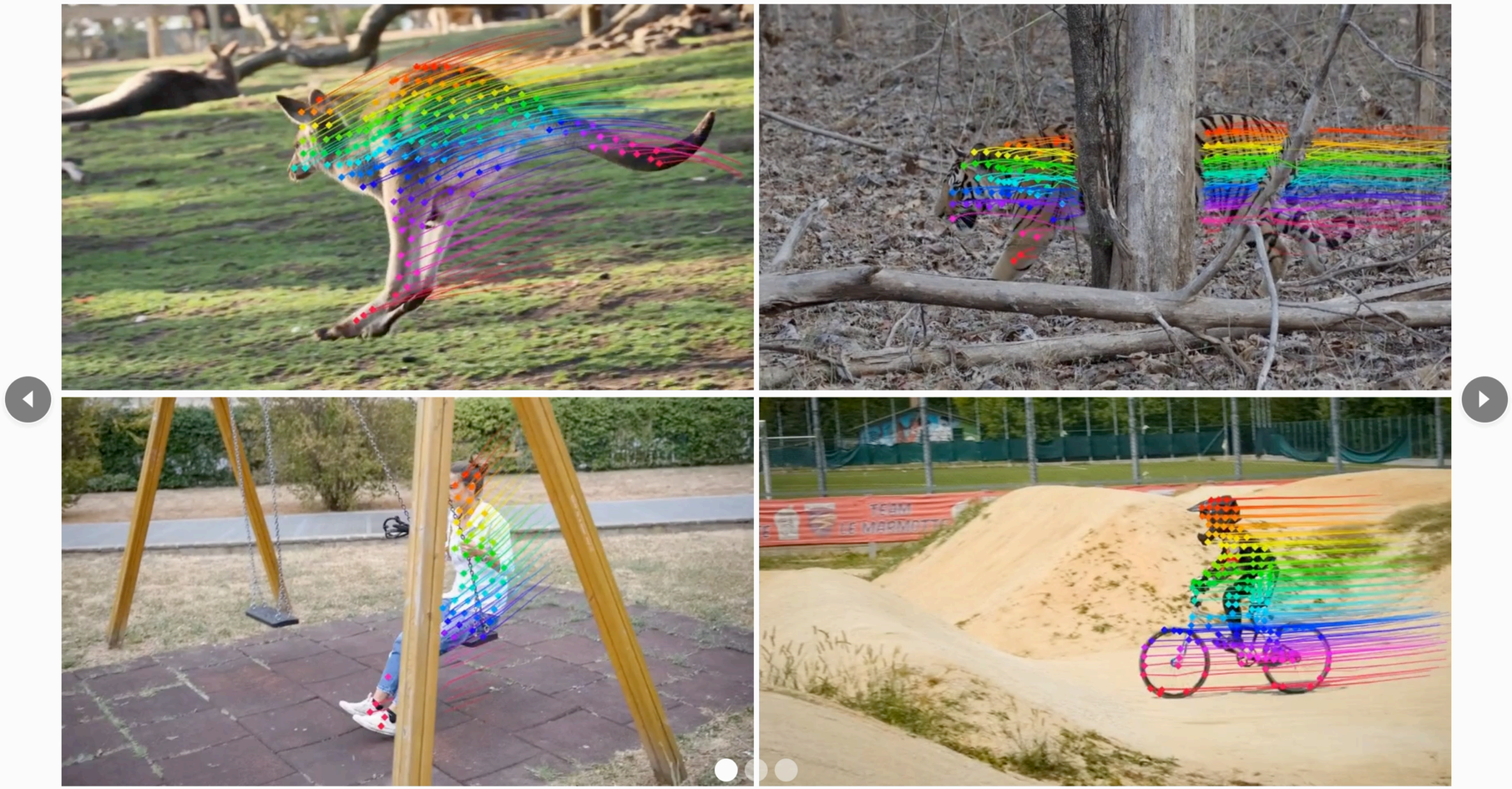
ECCV 2024

Narek Tumanyan * Assaf Singer * Shai Bagon Tali Dekel

WEIZMANN INSTITUTE OF SCIENCE

*indicates equal contribution

Paper Arxiv Code Supplementary Material



DINO-Tracker: Taming DINO for Self-Supervised Point Tracking in a Single Video

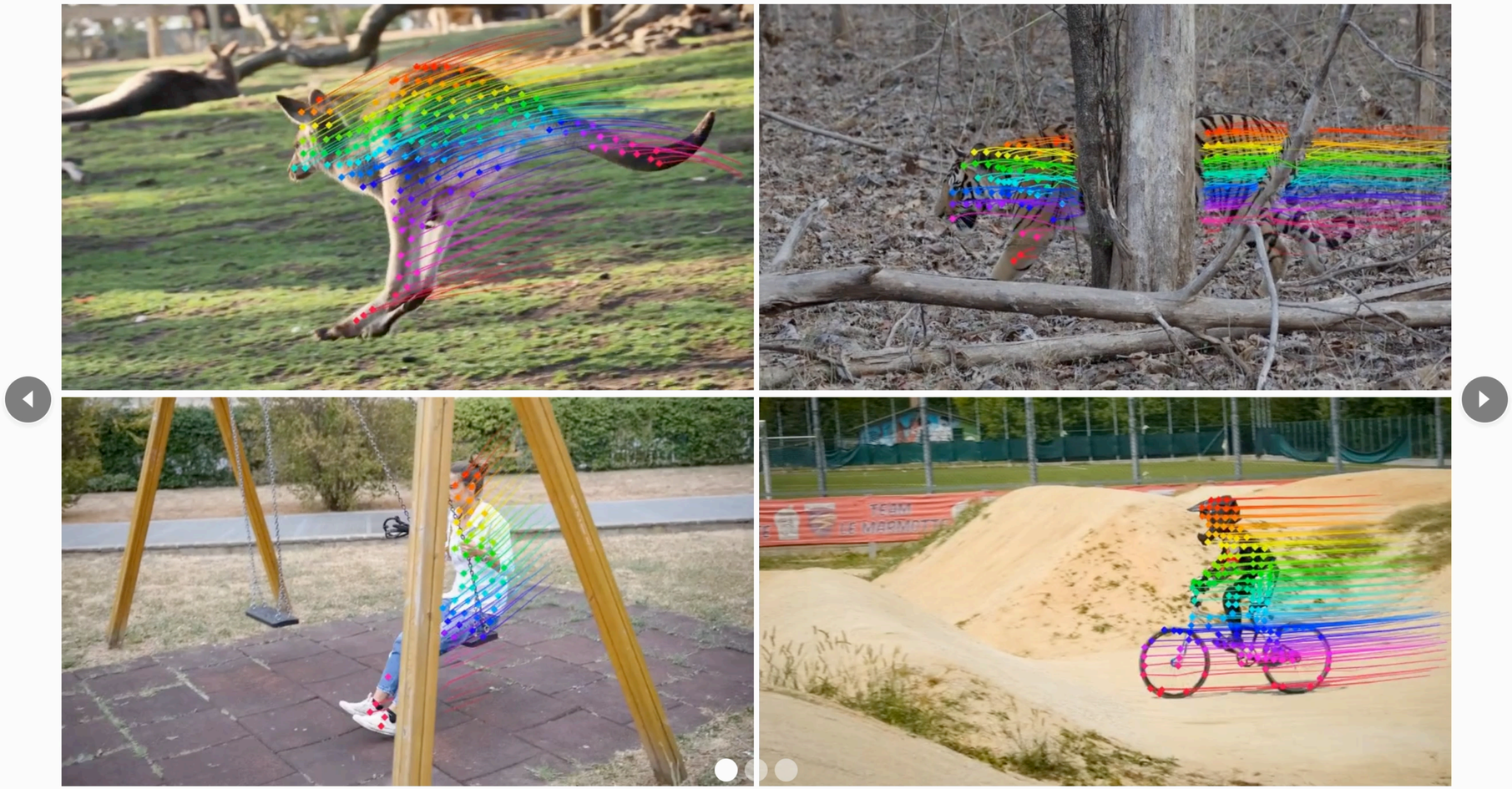
ECCV 2024

Narek Tumanyan * Assaf Singer * Shai Bagon Tali Dekel

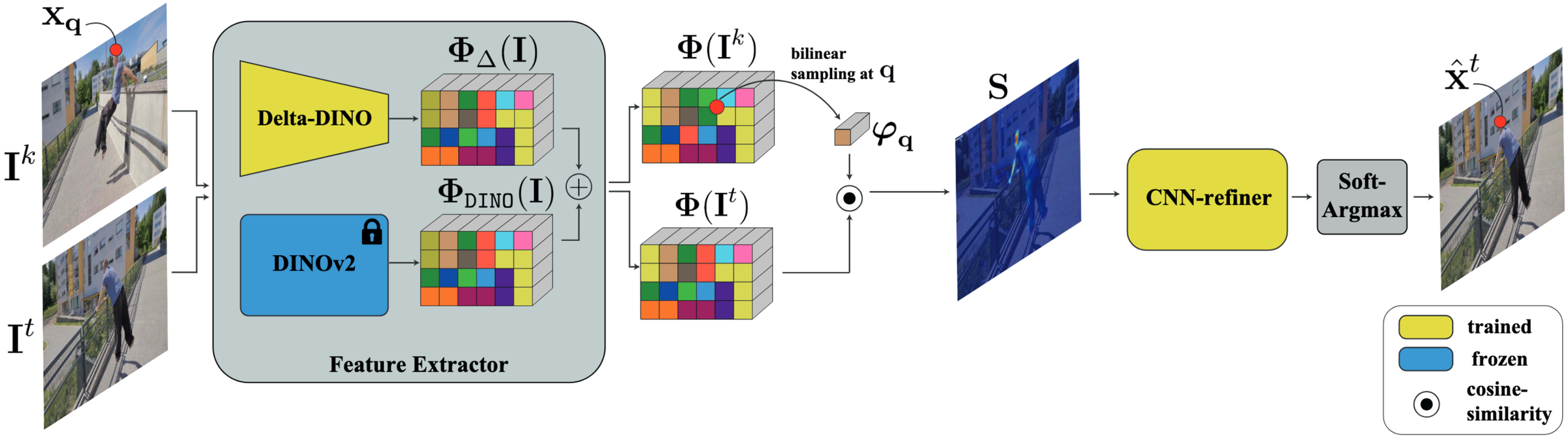
WEIZMANN INSTITUTE OF SCIENCE

*indicates equal contribution

Paper Arxiv Code Supplementary Material

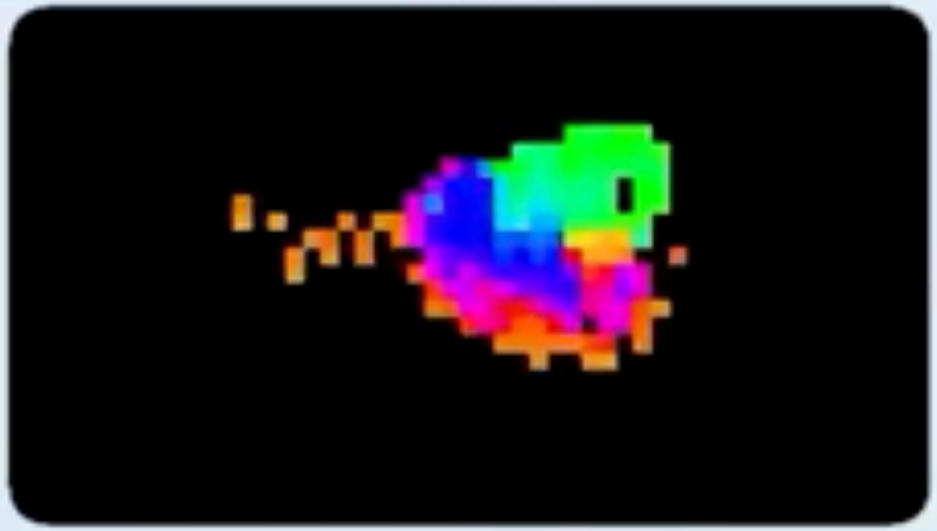


DINO-Tracker (Tumanyan, Singer et al. 2024)



Train a CNN to predict deltas of frozen DINO features per-frame, as well as a CNN that process cost volume to output softmax over next frame's pixels. Supervised with RAFT, fit to every video anew.

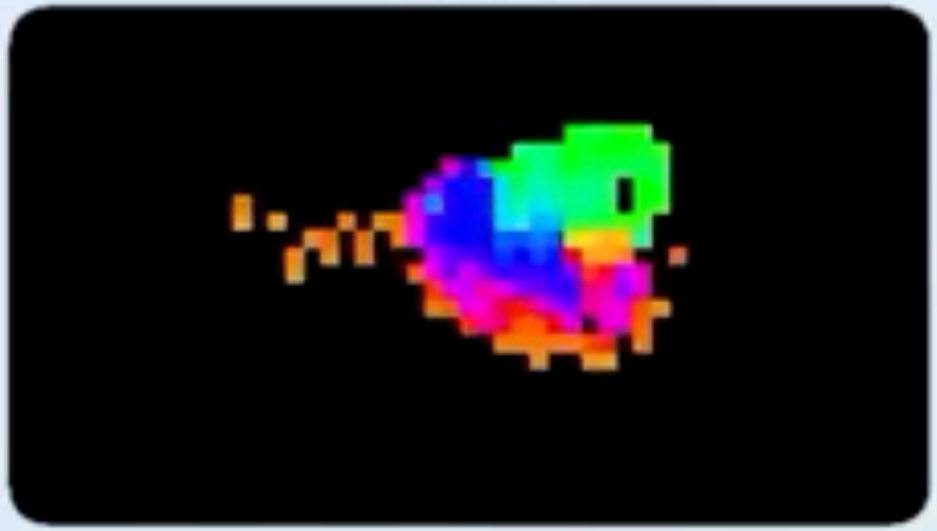
DINO



DINOv2

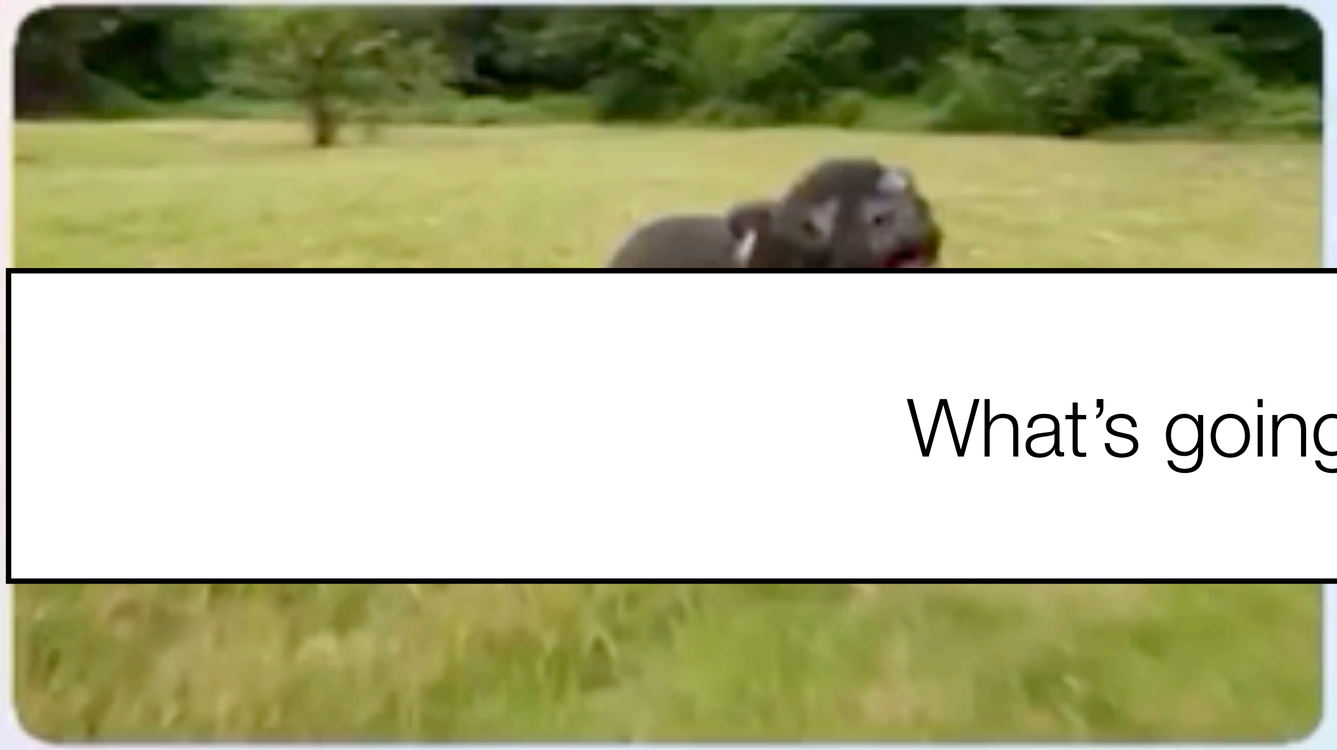
DINOv2, Oquab et al., arXiv 2023

DINO

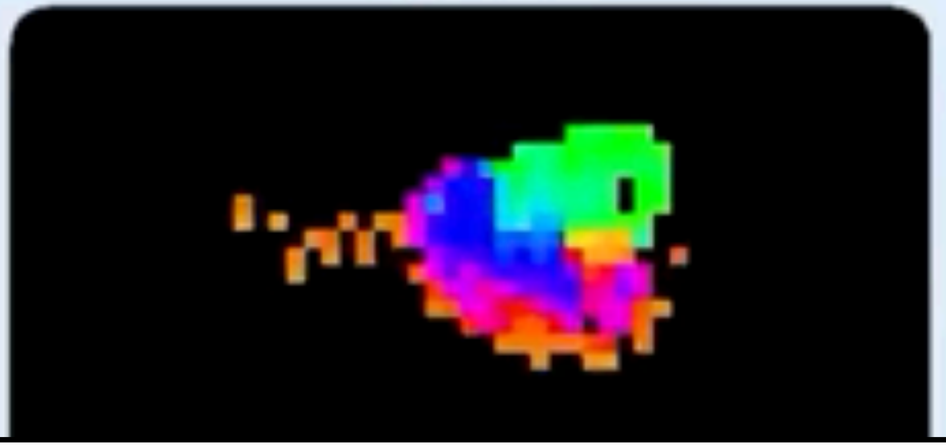


DINOv2

DINOv2, Oquab et al., arXiv 2023



DINO



What’s going on?

DINOv2, Oquab et al., arXiv 2023

ON INFORMATION MAXIMISATION IN MULTI-VIEW SELF-SUPERVISED LEARNING

Borja Rodríguez Gálvez, Arno Blaas, Xavier Suau, Jason Ramapuram,
Dan Busbridge, Luca Zappella

ABSTRACT

The strong performance of multi-view self-supervised learning (SSL) prompted the development of many different approaches (e.g. SimCLR, BYOL, and DINO). A unified understanding of how each of these methods achieves its performance has been limited by apparent differences across objectives and algorithmic details. Through the lens of information theory, we show that many of these approaches maximise an approximate lower bound on the mutual information between the representations of multiple views of the same datum. Further, we observe that this bound decomposes into a "reconstruction" term, treated identically by all SSL methods, and an "entropy" term, where existing SSL methods differ in their treatment. We prove that an exact optimisation of both terms of this lower bound encompasses and unifies current theoretical properties such as recovering the true latent variables of the underlying generative process (Zimmermann et al., 2021) or or isolating content from style in such true latent variables (Von Kügelgen et al., 2021). This theoretical analysis motivates a naive but principled objective (EntRec), that directly optimises both the reconstruction and entropy terms, thus benefiting from said theoretical properties unlike other SSL frameworks. Finally, we show EntRec achieves a downstream performance on-par with existing SSL methods on ImageNet (69.7% after 400 epochs) and on an array of transfer tasks when pre-trained on ImageNet. Furthermore, EntRec is more robust to modifying the batch size, a sensitive hyperparameter in other SSL methods.

ON INFORMATION MAXIMISATION IN MULTI-VIEW SELF-SUPERVISED LEARNING

Borja Rodríguez Gálvez, Arno Blaas, Xavier Suau, Jason Ramapuram,
Dan Busbridge, Luca Zappella

ABSTRACT

The strong performance of multi-view self-supervised learning (SSL) prompted the development of many different approaches (e.g. SimCLR, BYOL, and DINO). A unified understanding of how each of these methods achieves its performance has been limited by apparent differences across objectives and algorithmic details. Through the lens of information theory, we show that many of these approaches maximise an approximate lower bound on the mutual information between the representations of multiple views of the same datum. Further, we observe that this bound decomposes into a "reconstruction" term, treated identically by all SSL methods, and an "entropy" term, where existing SSL methods differ in their treatment. We prove that an exact optimisation of both terms of this lower bound encompasses and unifies current theoretical properties such as recovering the true latent variables of the underlying generative process (Zimmermann et al., 2021) or or isolating content from style in such true latent variables (Von Kügelgen et al., 2021). This theoretical analysis motivates a naive but principled objective (EntRec), that directly optimises both the reconstruction and entropy terms, thus benefiting from said theoretical properties unlike other SSL frameworks. Finally, we show EntRec achieves a downstream performance on-par with existing SSL methods on ImageNet (69.7% after 400 epochs) and on an array of transfer tasks when pre-trained on ImageNet. Furthermore, EntRec is more robust to modifying the batch size, a sensitive hyperparameter in other SSL methods.

Argues that multi-view representation learning methods maximize a lower bound on the *mutual information* between multi-view embeddings.

ON INFORMATION MAXIMISATION IN MULTI-VIEW SELF-SUPERVISED LEARNING

Borja Rodríguez Gálvez, Arno Blaas, Xavier Suau, Jason Ramapuram,
Dan Busbridge, Luca Zappella

ABSTRACT

The strong performance of multi-view self-supervised learning (SSL) prompted the development of many different approaches (e.g. SimCLR, BYOL, and DINO). A unified understanding of how each of these methods achieves its performance has been limited by apparent differences across objectives and algorithmic details. Through the lens of information theory, we show that many of these approaches maximise an approximate lower bound on the mutual information between the representations of multiple views of the same datum. Further, we observe that this bound decomposes into a "reconstruction" term, treated identically by all SSL methods, and an "entropy" term, where existing SSL methods differ in their treatment. We prove that an exact optimisation of both terms of this lower bound encompasses and unifies current theoretical properties such as recovering the true latent variables of the underlying generative process (Zimmermann et al., 2021) or or isolating content from style in such true latent variables (Von Kügelgen et al., 2021). This theoretical analysis motivates a naive but principled objective (EntRec), that directly optimises both the reconstruction and entropy terms, thus benefiting from said theoretical properties unlike other SSL frameworks. Finally, we show EntRec achieves a downstream performance on-par with existing SSL methods on ImageNet (69.7% after 400 epochs) and on an array of transfer tasks when pre-trained on ImageNet. Furthermore, EntRec is more robust to modifying the batch size, a sensitive hyperparameter in other SSL methods.

Intuitively, DINO local features of two patches are similar if knowing about one patch reduces uncertainty of the other.

On Mutual Information Maximization for Representation Learning

Michael Tschannen* **Josip Djolonga*** **Paul K. Rubenstein†** **Sylvain Gelly** **Mario Lucic**
Google Research, Brain Team

Abstract

Many recent methods for unsupervised or self-supervised representation learning train feature extractors by maximizing an estimate of the mutual information (MI) between different views of the data. This comes with several immediate problems: For example, MI is notoriously hard to estimate, and using it as an objective for representation learning may lead to highly entangled representations due to its invariance under arbitrary invertible transformations. Nevertheless, these methods have been repeatedly shown to excel in practice. In this paper we argue, and provide empirical evidence, that the success of these methods cannot be attributed to the properties of MI alone, and that they strongly depend on the inductive bias in both the choice of feature extractor architectures and the parametrization of the employed MI estimators. Finally, we establish a connection to deep metric learning and argue that this interpretation may be a plausible explanation for the success of the recently introduced methods.

Maximized MI and worsened downstream performance

Looser bounds with simpler critics can lead to better representations

What Makes a Good Representation?

Good representations are:

1. Compact (*minimal*)
2. Explanatory (*sufficient*)
3. Disentangled (*independent factors*)
4. Equivariant (*transform predictably*)
5. Interpretable

This is a tall order…

If we could do this reliably, all of 3D reconstruction would be done!

6. *Make subsequent problem solving easy*
7. …?

[See “Representation Learning”, Bengio 2013, for more commentary]

What Makes a Good Representation?

Good representations are:

Current (major) limitation: Contrastive learning only offers a straightforward path towards learning about what is *invariant* between two views. DINO is interesting in that it shows that local tokens also benefit from this supervision.

The hope: to eventually find a self-supervised method that learns all the latent variables that are involved in generating images.

Currently, nowhere close :)

y]