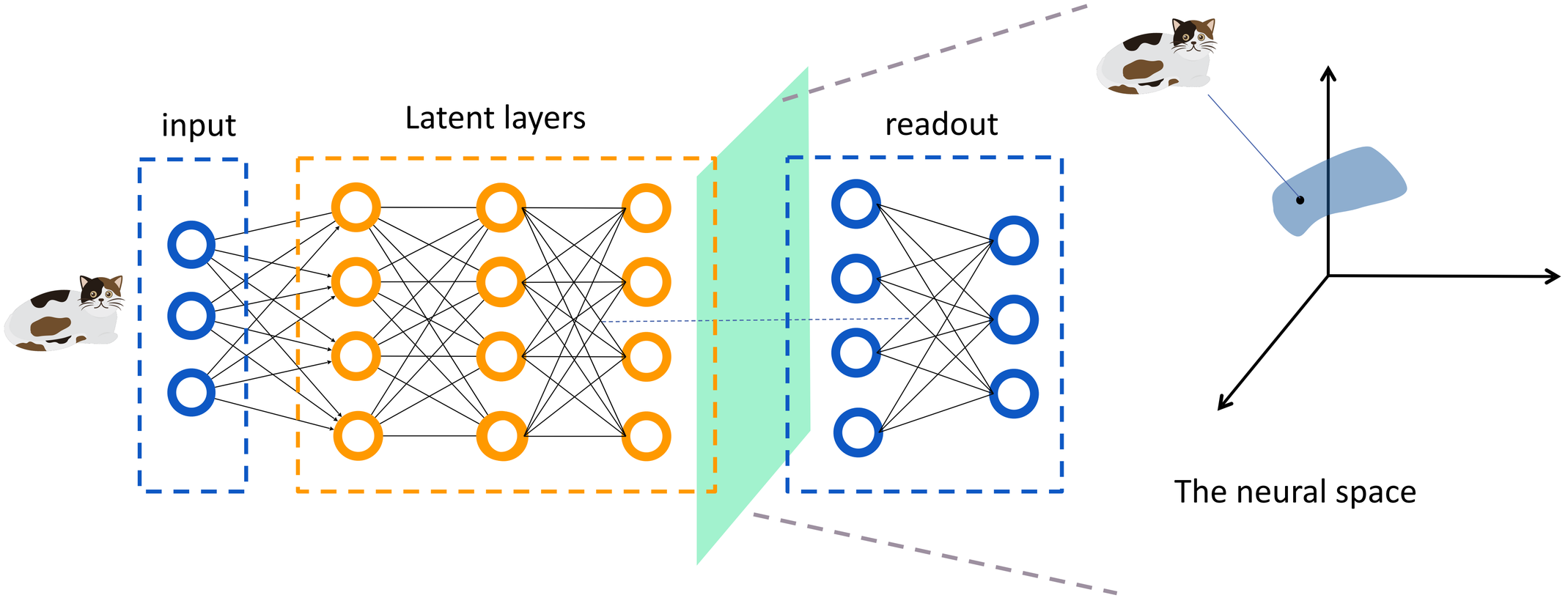
Contrastive self-supervised learning as neural manifold packing

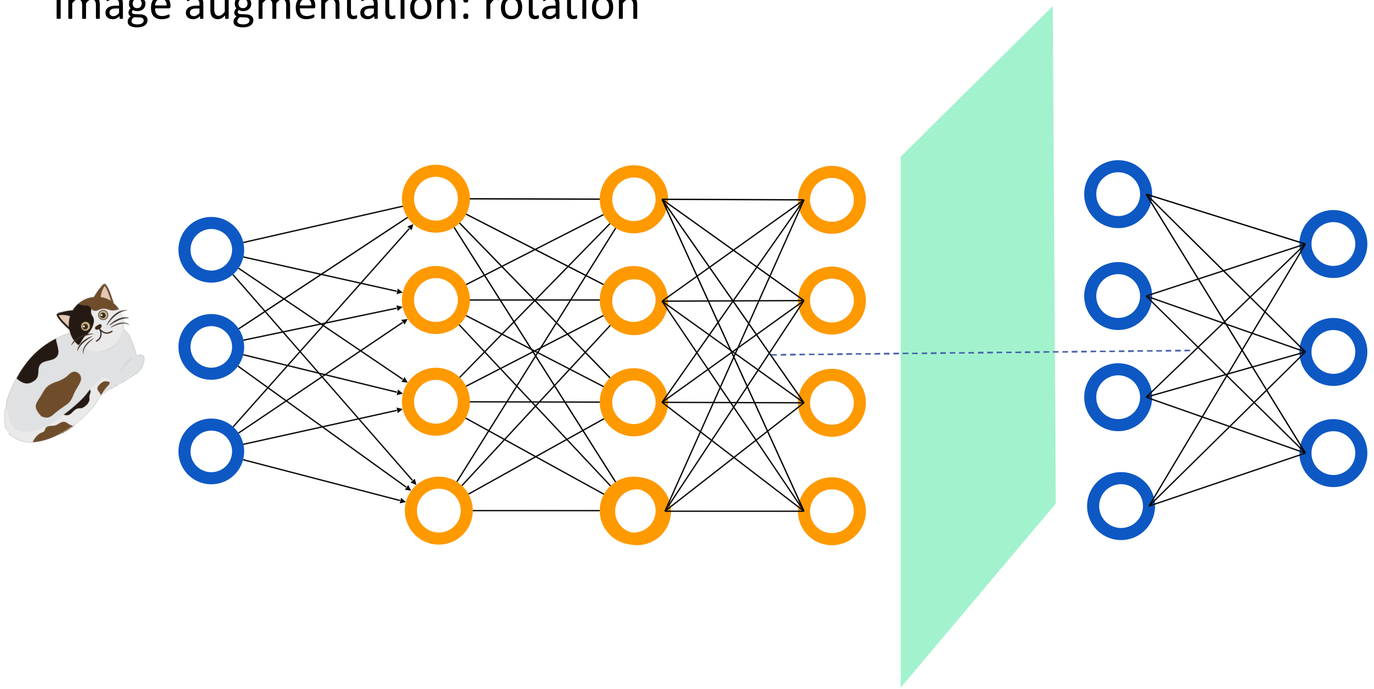
Guanming Zhang, David J. Heeger and Stefano Martiniani

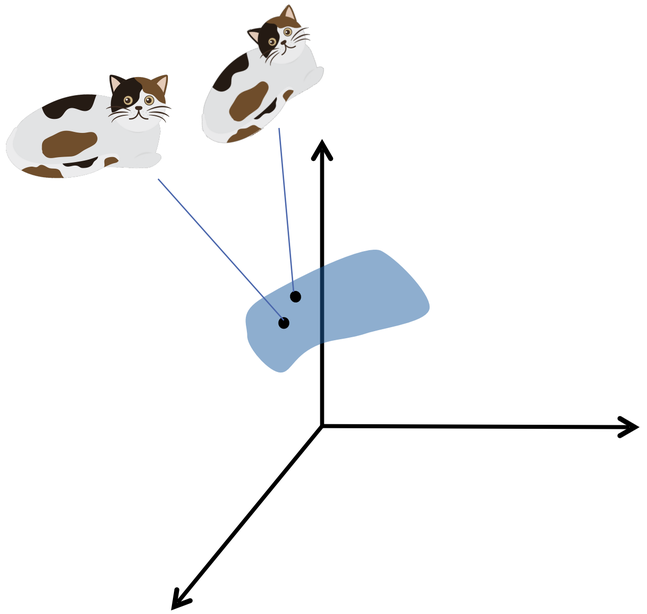
Neural manifolds in vision tasks



Neural manifolds in vision tasks

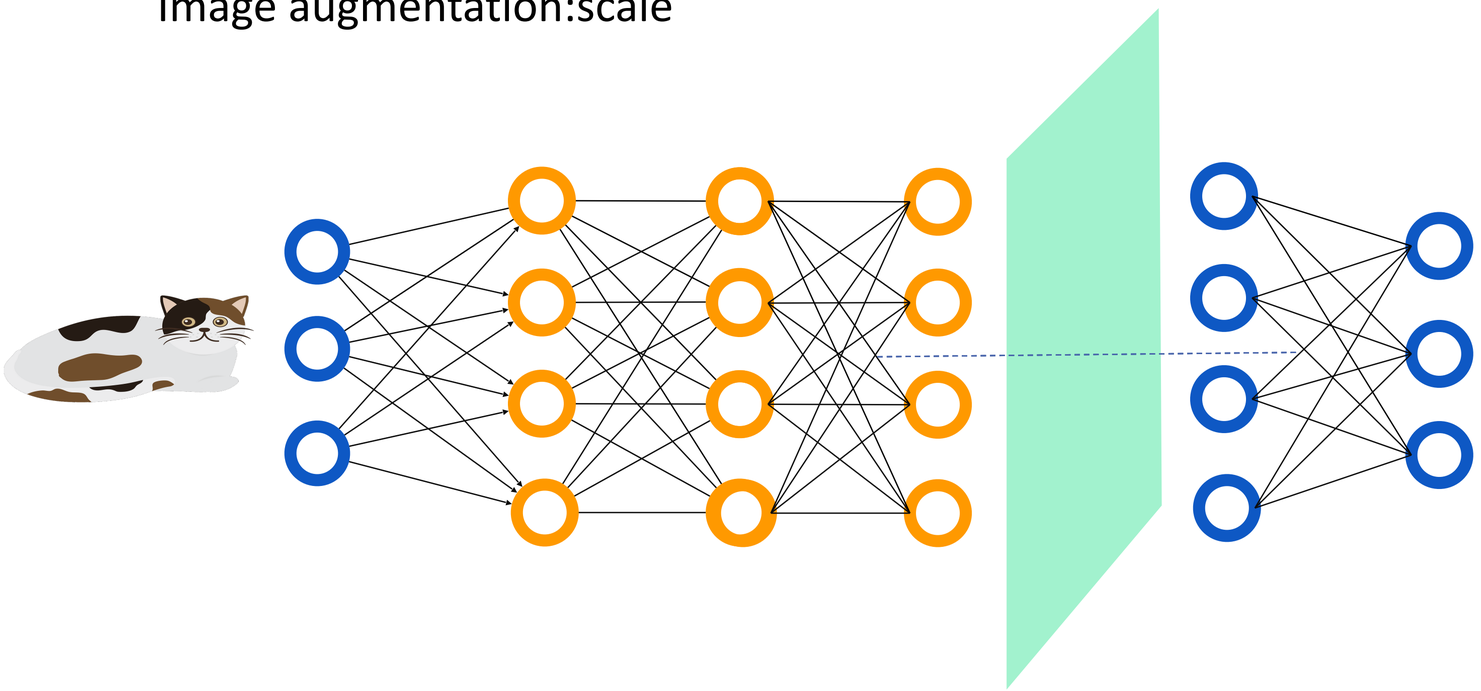
Image augmentation: rotation

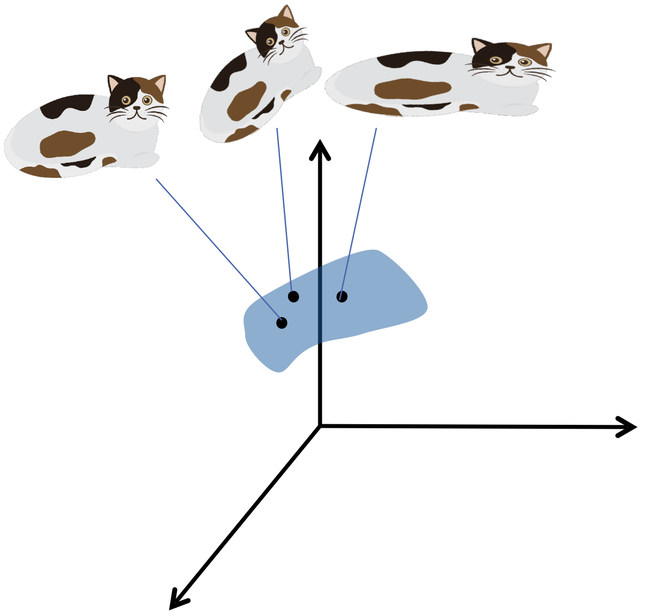




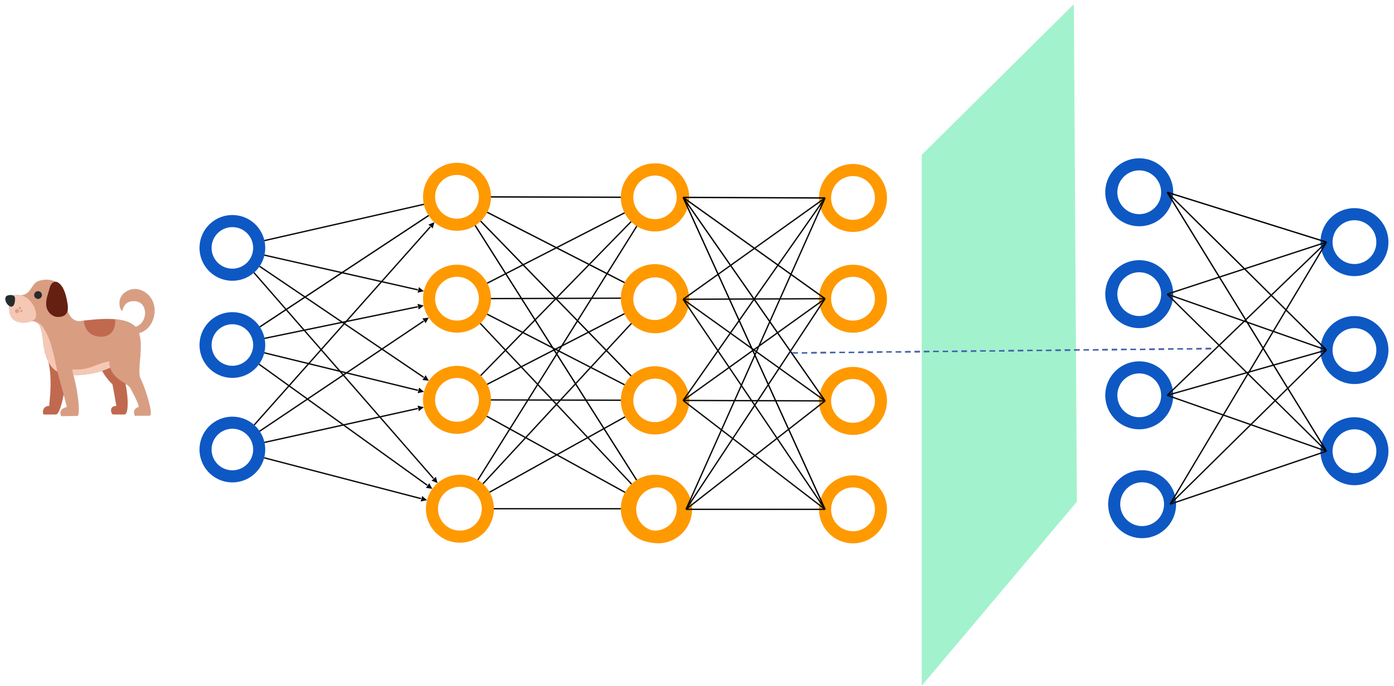
Neural manifolds in vision tasks

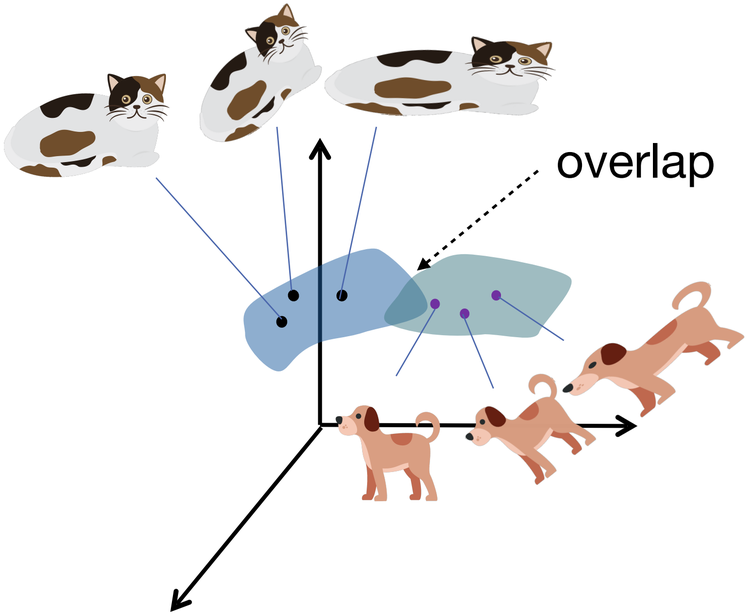
Image augmentation:scale





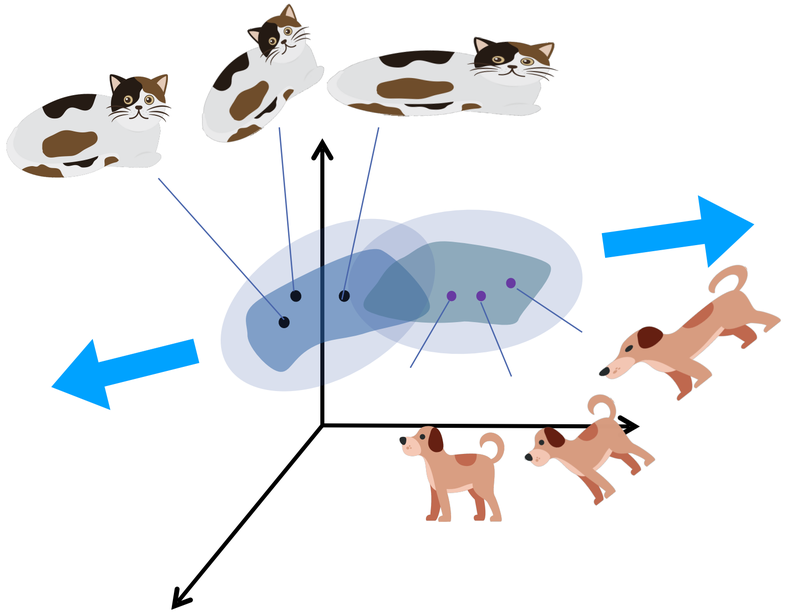
Neural manifolds in vision tasks



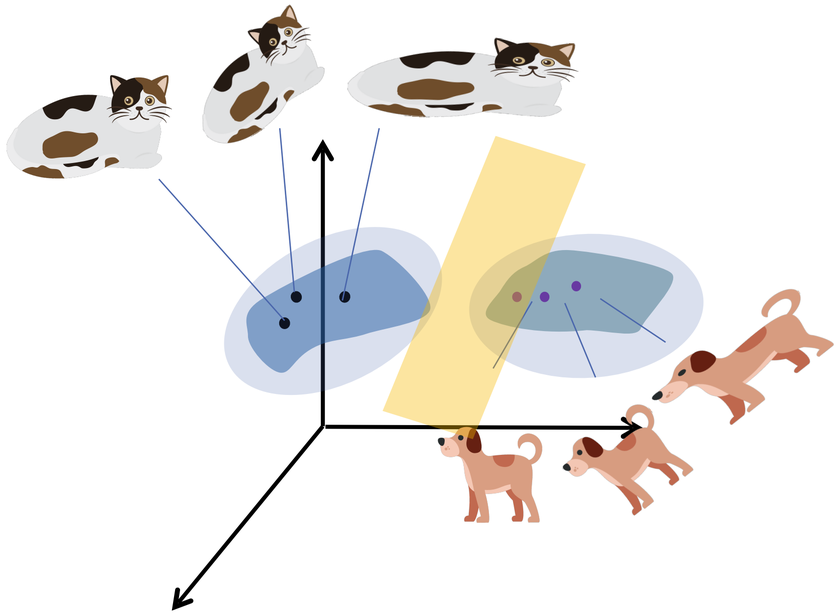


By the same token, we can generate the augmentation neural manifold for dogs.

Neural manifolds in vision tasks



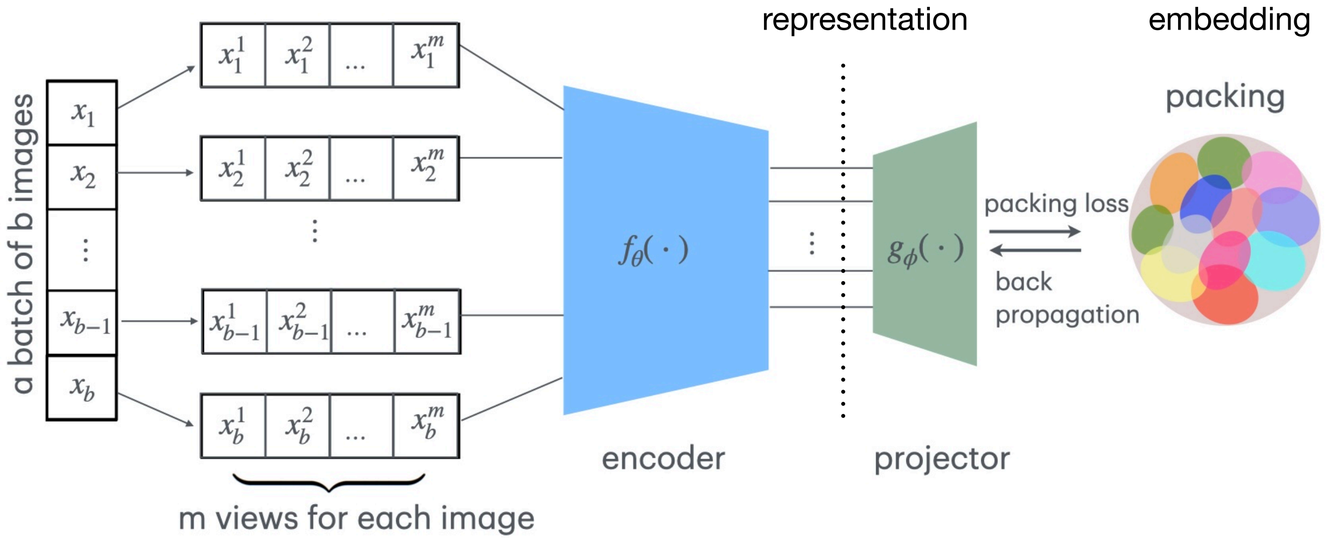
1. Enclose the sub-manifold by ellipsoids
2. Make these ellipsoids repulsive



Linearly separated sub-manifolds

Contrastive self-supervised learning as manifold packing-CLAMP

clamp



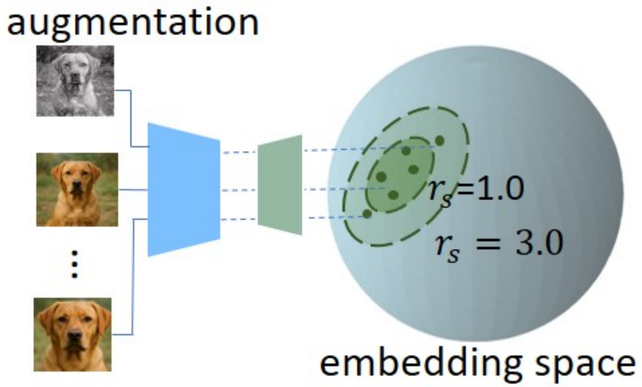
Similar to most SSL models, we treat each image as a **distinct** category. No label is needed.

We use MLP projection head as the feature projector, it is discarded after pretraining.

All embedding vector are normalized on a unit hypersphere

CLAMP

Augmentation sub-manifold and its circumscribed ellipsoids

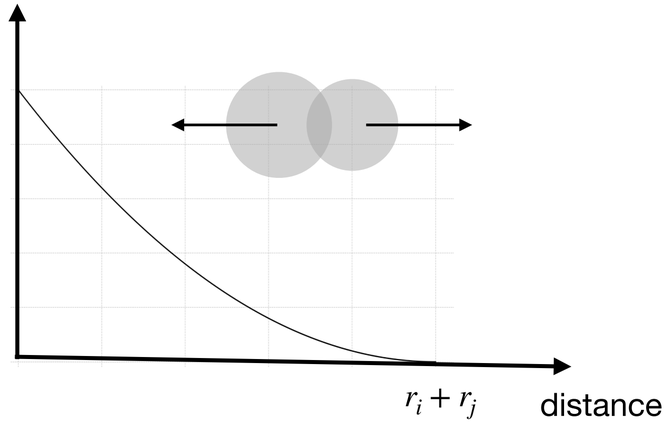


Z_i : centers of the augmentation sub-manifold

Λ_i : empirical covariance matrix of embedding vectors corresponding to the same input sample

$$(\tilde{z} - Z_i)\Lambda_i^{-1}(\tilde{z} - Z_i)^T = r_s^2$$

$$\mathcal{L}_{overlap} = \begin{cases} \sum_{i \neq j} \left(1 - \frac{\|Z_i - Z_j\|_2}{r_i + r_j}\right)^2, \text{if } \|Z_i - Z_j\|_2 < r_i + r_j \\ 0, \quad \text{otherwise} \end{cases} \quad \text{, where } r_i = r_s\sqrt{\frac{\text{Tr}(\Lambda_i)}{m}}.$$



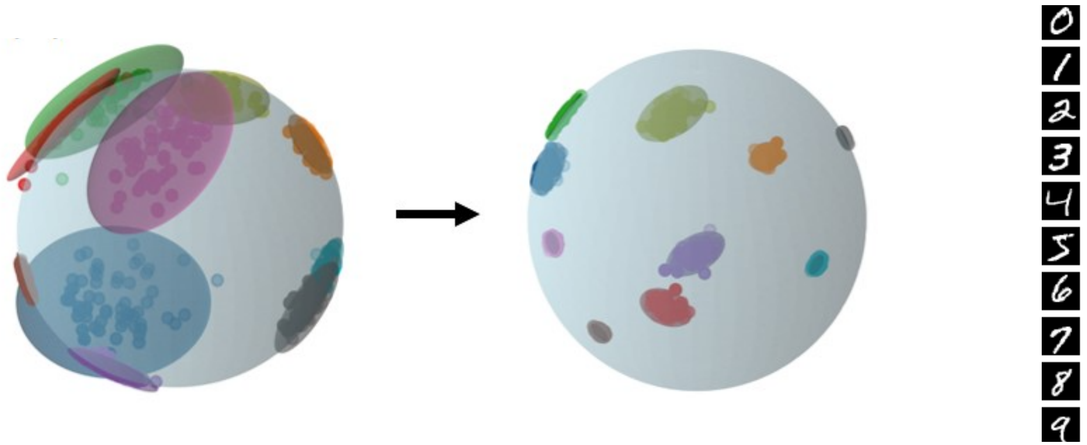
The log-normalized loss, $\log(\mathscr{L}_{overlap})$, decreases as

(1) The sizes of the sub-manifolds shrink

(2) The distances of two overlapped sub-manifolds increases

CLAMP

Toy model



We selected 10 images from the MNIST dataset, one for each digit from 0 to 9, and applied Gaussian noise augmentation.
These augmented images were then encoded into a 3-dimensional embedding space for visualization.
Solid dots: embedding points of each augmented view
Shaded regions: circumscribed ellipsoids

Evaluation results

Linear evaluation and semi-supervised learning

	Linear evaluation		Semi-supervised	
Method	ImageNet-100	ImageNet-1K	1%	10%
SimCLR[1]	79.64	66.5	42.6	61.6
SwAV[29]	-	**72.1**	**49.8**	**66.9**
Barlow Twins[42]	80.38*	68.7	45.1	61.7
BYOL[2]	80.32*	69.3	49.8	65.0
VICReg[6]	79.4	68.7	44.75	62.16
CorInfoMax[43]	80.48	69.08	44.89	64.36
MoCo-V2[3]	79.28*	67.4	43.4	63.2
SimSiam[4]	81.6	68.1	-	-
A&U [30]	74.6	67.69	-	-
MMCR (4 views+ME)[34]	82.88	71.5	49.4	66.0
CLAMP (4 views)	**85.12** $\pm$.05	69.50 $\pm$.14	47.38 $\pm$.56	65.10 $\pm$.30
CLAMP (8 views)	**85.10** $\pm$.15	70.04 $\pm$.16	47.87 $\pm$.03	65.96 $\pm$.04

Following standard linear evaluation protocols, we froze the pretrained ResNet-50 backbone encoder and trained a linear classifier on representation.Training was conducted for 100 epochs on ImageNet-1K (14,197,122 images) and 200 epochs on ImageNet-100.

Evaluation results

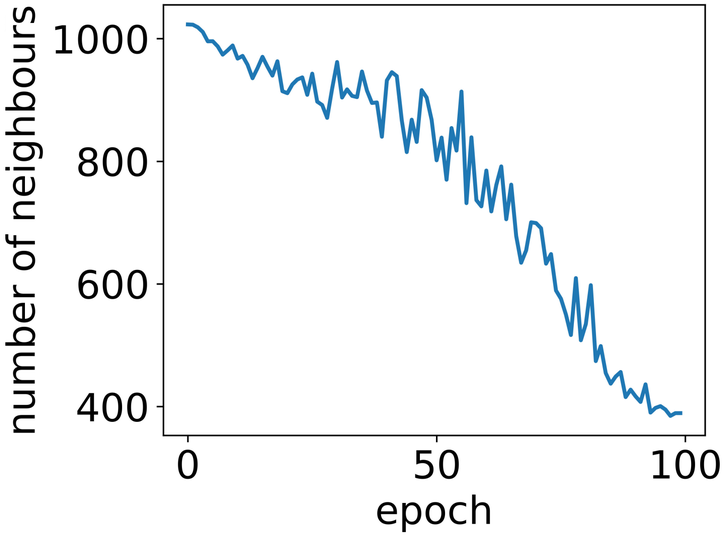
Transfer to object detection tasks

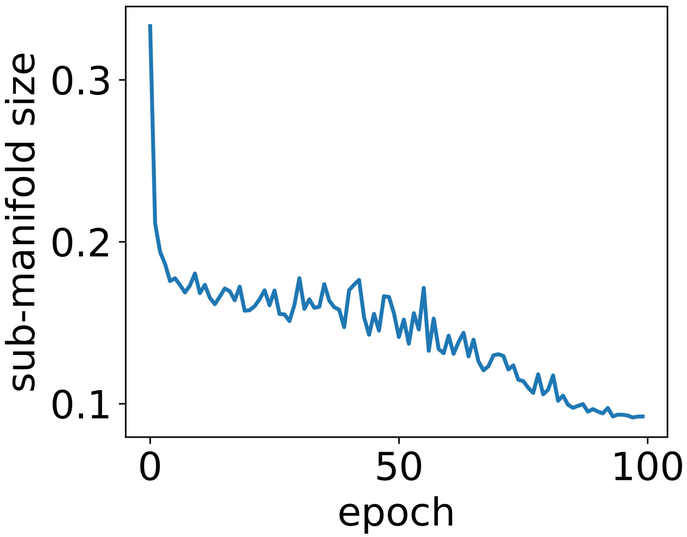
Methods	mAP ↑	AP50 ↑	AP75 ↑
SimCLR	54.4	81.6	61.0
Barlow Twins	53.1	80.9	57.7
BYOL	55.6	**82.3**	62.0
MoCo v2	54.7	81.7	60.2
MMCR	54.6	81.9	60.0
CLAMP	**55.7**	**82.3**	**62.4**

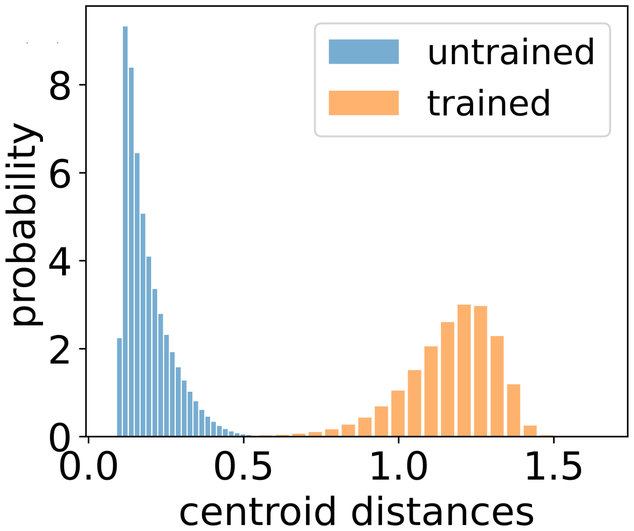
We present results on object detection using a Faster R-CNN architecture with a C4 backbone (pretrained on ImageNet-1K with 8 views and batch size 512 for 100 epochs), fine-tuned on VOC2007+2012 training dataset and tested on VOC2007 test dataset

Training dynamics-interpretability

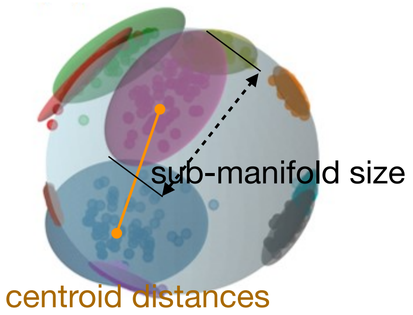
Training dynamics reflect geometry change





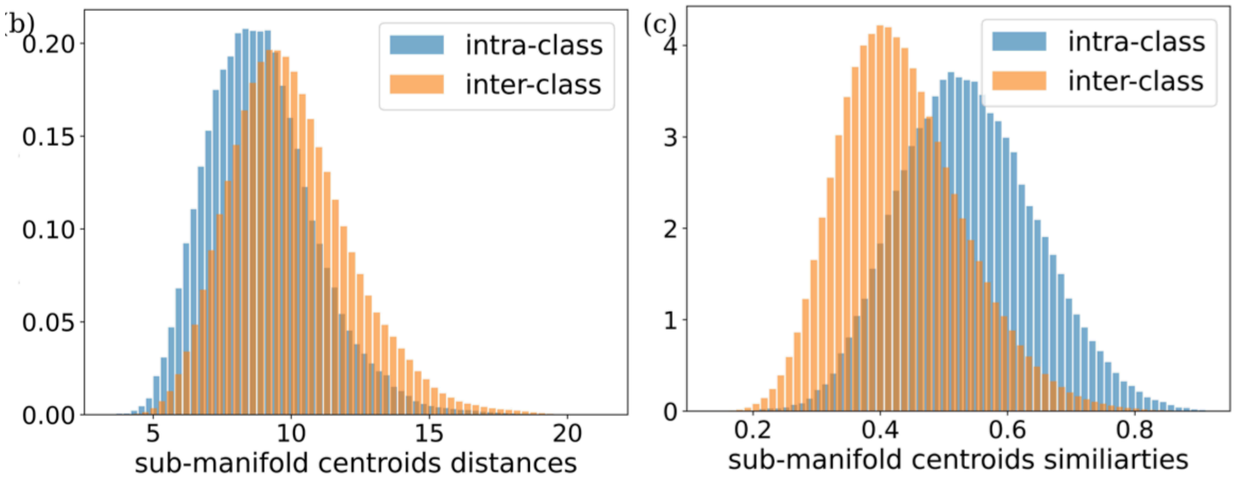


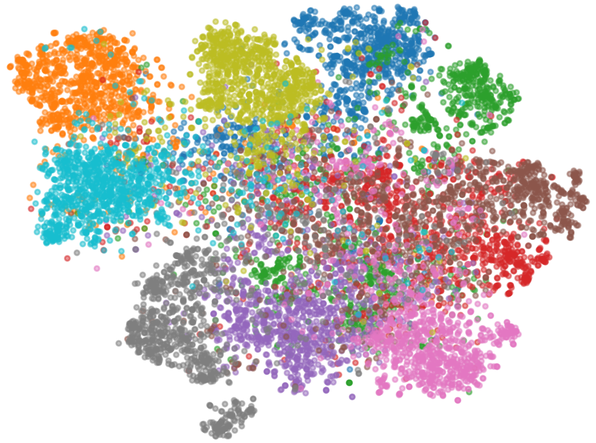
Two sub-manifolds are counted as neighbors if $||Z_i - Z_j||_2 < r_i + r_j$



Representation

Properties of the representation



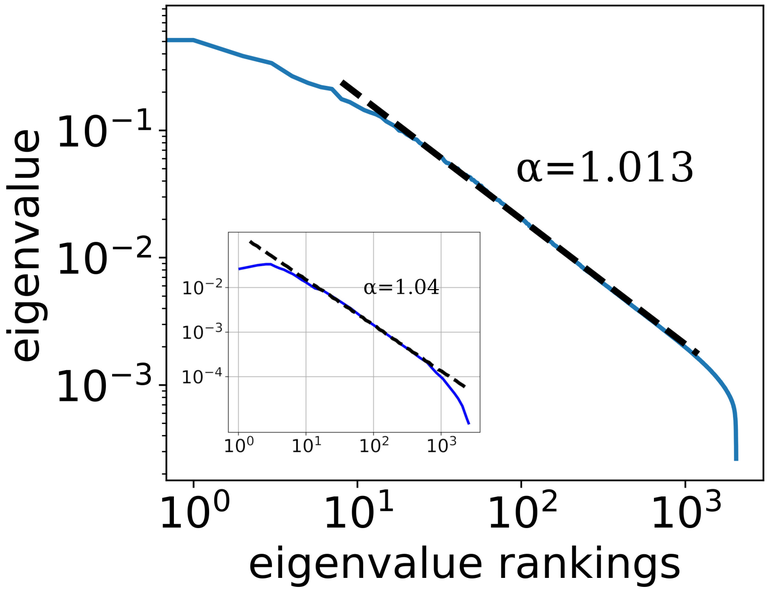


t-SNE visualization of 256-dimensional representations

We achieve 90.2% frozen-backbone linear evaluation accuracy on the CIFAR-10 dataset using ResNet-18. We find that category structures emerge as the result of SSL pretraining.

Biological implications

Eigenspectrum and Brain-score



Methods	V1	V2	V4	IT
SimCLR	0.224	0.288	0.576	0.552
SwAV	0.252	0.296	0.568	0.533
Barlow Twins	**0.276**	0.293	0.568	0.545
BYOL	0.274	0.291	**0.585**	0.55
MMCR (8 views)	0.270	0.311	0.577	0.554
CLAMP (4 views)	0.258 ±0.013	**0.336** ±0.017	0.558 ±0.004	**0.570** ±0.005

For stimulus-evoked activity in mouse V1 (primary visual cortex), the eigenspectrum of its covariance matrix follows a power-law decay, $\lambda \propto n^{-\alpha}$, with α ≈ 1.04, and α > 1 is indeed necessary to ensure differentiability of the neural code.

We then evaluated representational alignment using the Brain-Score benchmark, which measures how well model activations predict primate neural recordings via cross-validated linear regression.

Thank you!