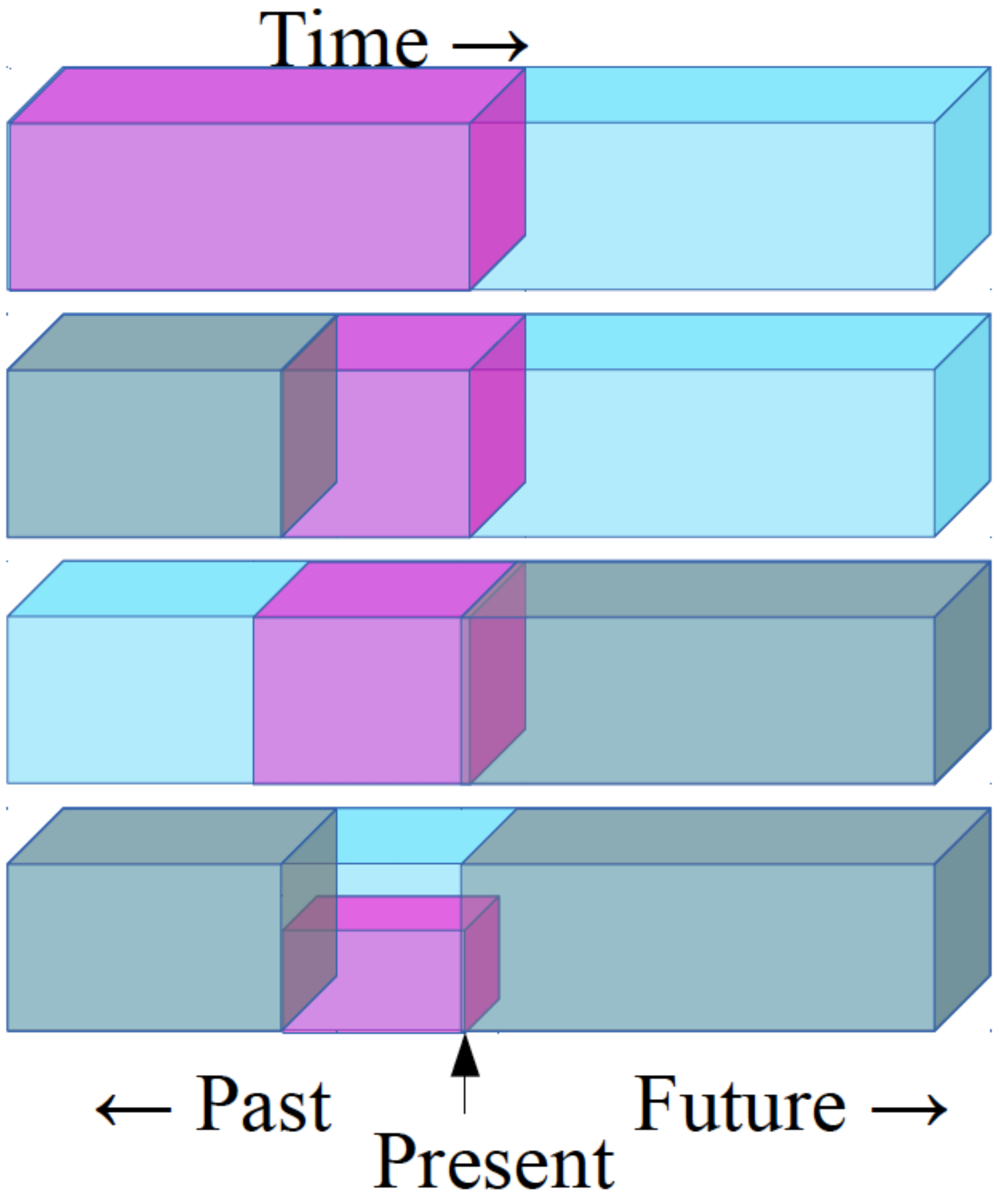
Data-Efficient Image Recognition with Contrastive Predictive Coding

Olivier J. Henaff, Ali Razavi, Carl Doersch,S. M. Ali Eslami*, Aaron van den Oord*
DeepMind

Compiled by Hongming Shan

Self-Supervised Learning: Prediction & Reconstruction

- Predict any part of the input from any other part.
- Predict the future from the past.
- Predict the future from the recent past.
- Predict the past from the present.
- Predict the top from the bottom.
- Predict the occluded from the visible
- Pretend there is a part of the input you don't know and predict that.



Slide from Yann LeCun's ACM talk

Y. LeCun

How Much Information is the Machine Given during Learning?

- **"Pure" Reinforcement Learning (cherry)**
  - The machine predicts a scalar reward given once in a while.
  - **A few bits for some samples**
- **Supervised Learning (icing)**
  - The machine predicts a category or a few numbers for each input
  - Predicting human-supplied data
  - **10→10,000 bits per sample**
- **Self-Supervised Learning (cake génoise)**
  - The machine predicts any part of its input for any observed part.
  - Predicts future frames in videos
  - **Millions of bits per sample**

Slide from Yann LeCun's ACM talk

Contrastive Predictive Coding

Representation Learning with Contrastive Predictive Coding

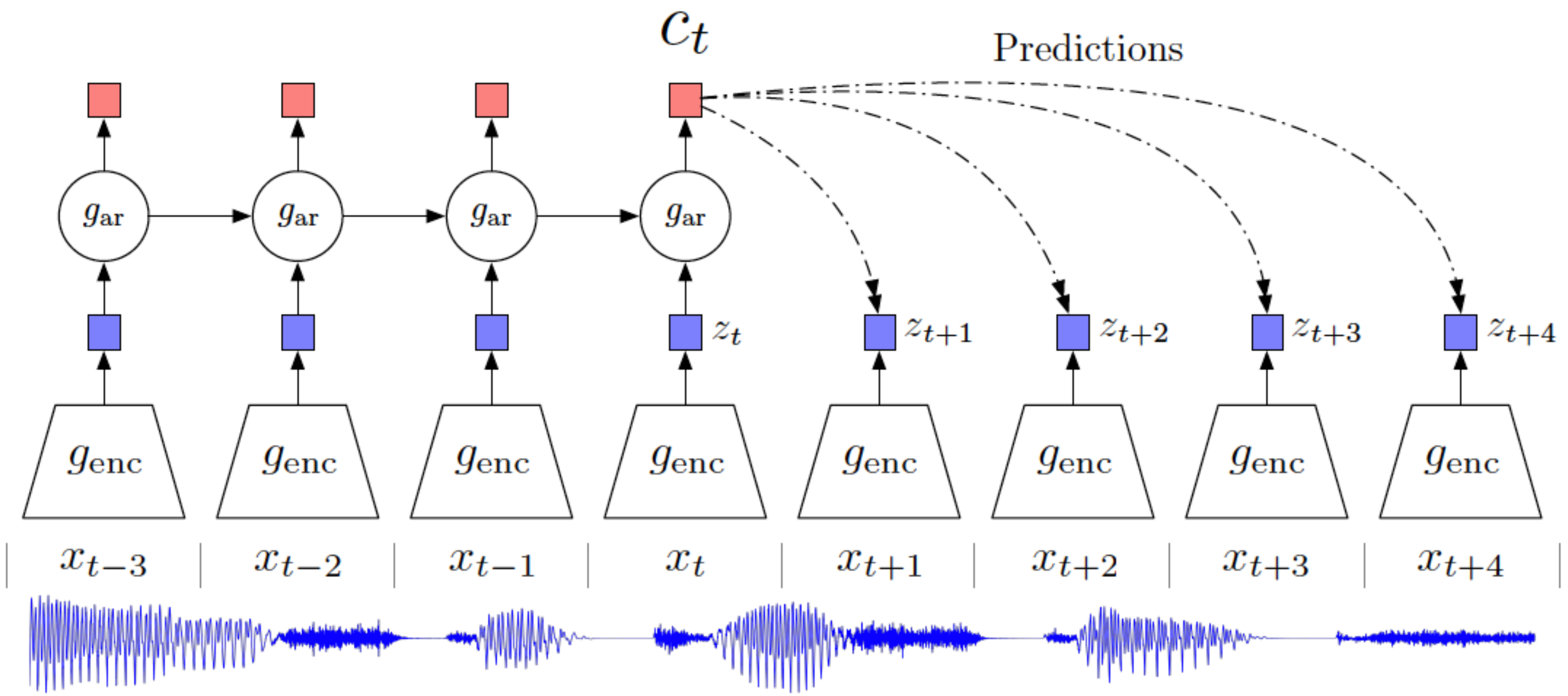
Aaron van den Oord
DeepMind
avdnoord@google.com

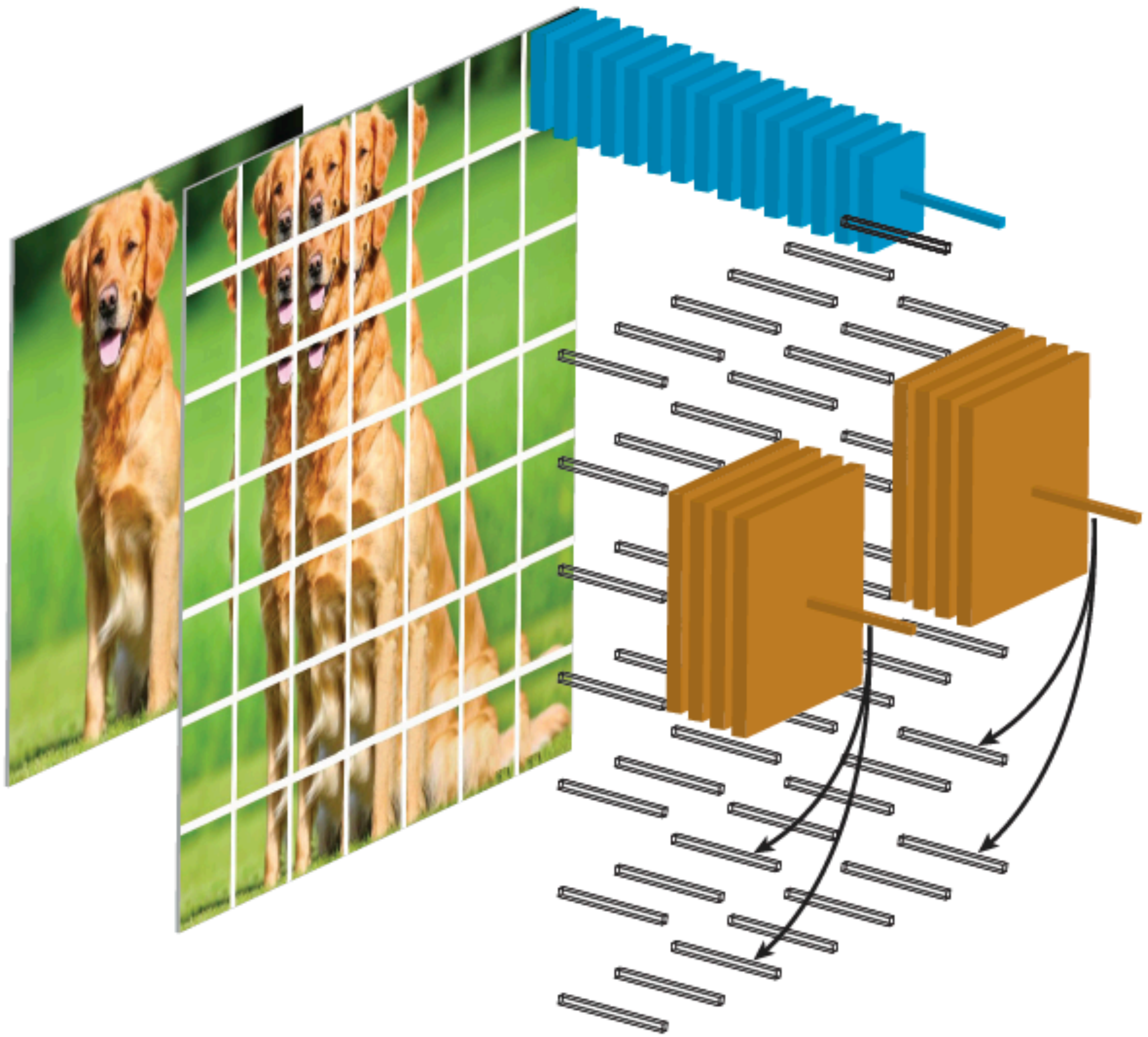
Yazhe Li
DeepMind
yazhe@google.com

Oriol Vinyals
DeepMind
vinyals@google.com

Overview of CPC

Audio input





- An image is divided into a grid of overlapping patches.
- Each patch is encoded independently from the rest with a feature extractor (blue) which terminates with a mean-pooling operation, yielding a single feature vector for that patch.
- Doing so for all patches yields a field of such feature vectors (wireframe vectors).
- Feature vectors above a certain level (in this case, the center of the image) are then aggregated with a context network (brown), yielding a row of context vectors which are used to **linearly** predict (unseen) features vectors below.

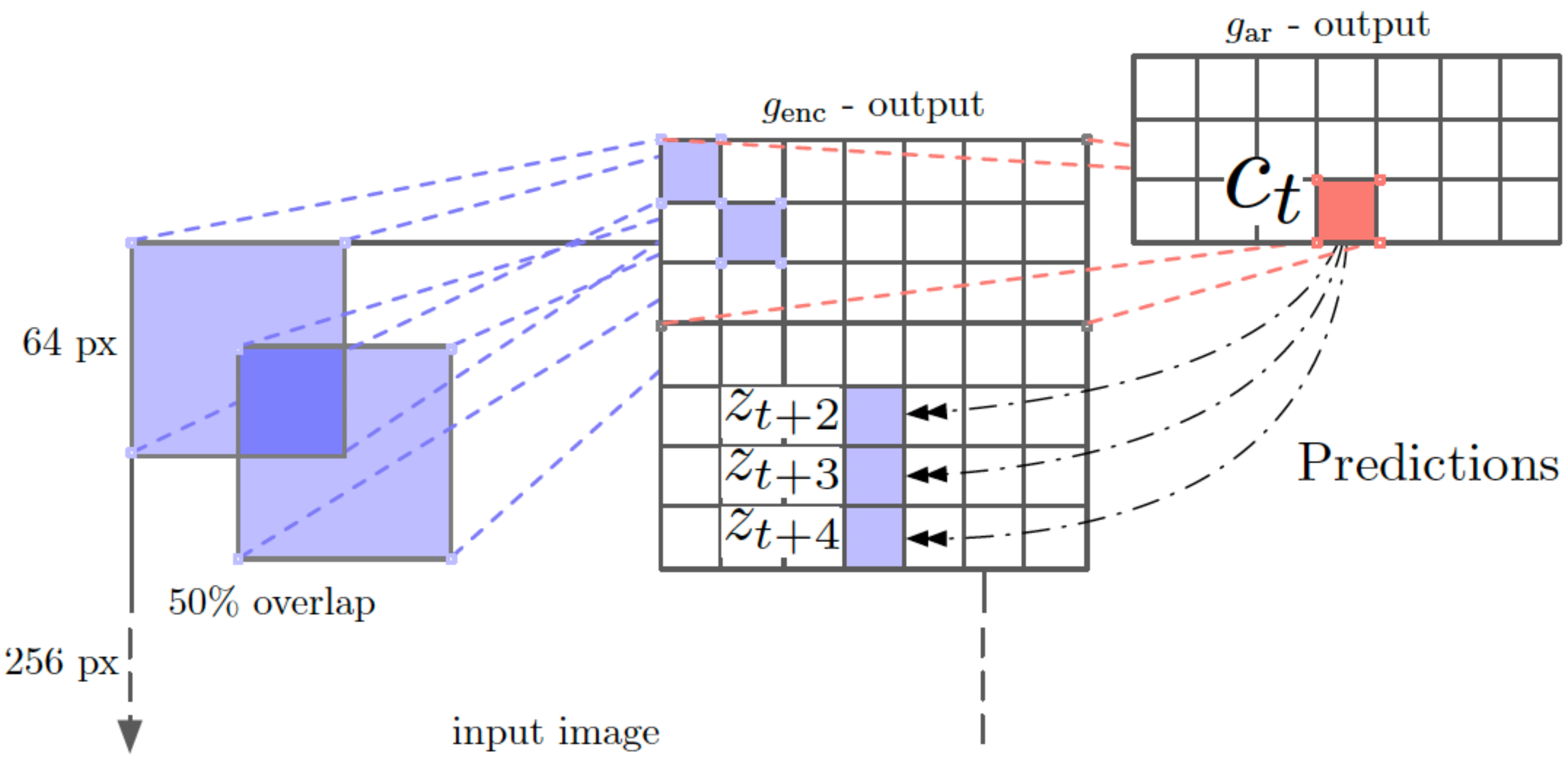


Figure 4: Visualization of Contrastive Predictive Coding for images (2D adaptation of Figure 1).

Prediction

$$\hat{z}_{i+k,j} = W_k c_{i,j}$$

Contrastive Loss

$$\begin{aligned}\mathcal{L}_{\text{CPC}} &= -\sum_{i,j,k} \log p(z_{i+k,j} | \hat{z}_{i+k,j}, \{z_l\}) \\ &= -\sum_{i,j,k} \log \frac{\exp(\hat{z}_{i+k,j}^T z_{i+k,j})}{\exp(\hat{z}_{i+k,j}^T z_{i+k,j}) + \sum_l \exp(\hat{z}_{i+k,j}^T z'_l)}\end{aligned}$$

Contributions

- With **architectural optimizations**, CPC feature encoders can be scaled to much larger networks, which can therefore absorb more useful information from unlabeled data, resulting in features that separate image categories better.

- Explore the use of this representation for classification with **a small number of labels**, as few as 1% of the entire ImageNet dataset.

- Investigate the applicability of this representation for **transfer learning**

- Explore different methods for semi-supervised learning, and find that the standard approach—end-to-end fine-tuning—is not necessarily optimal

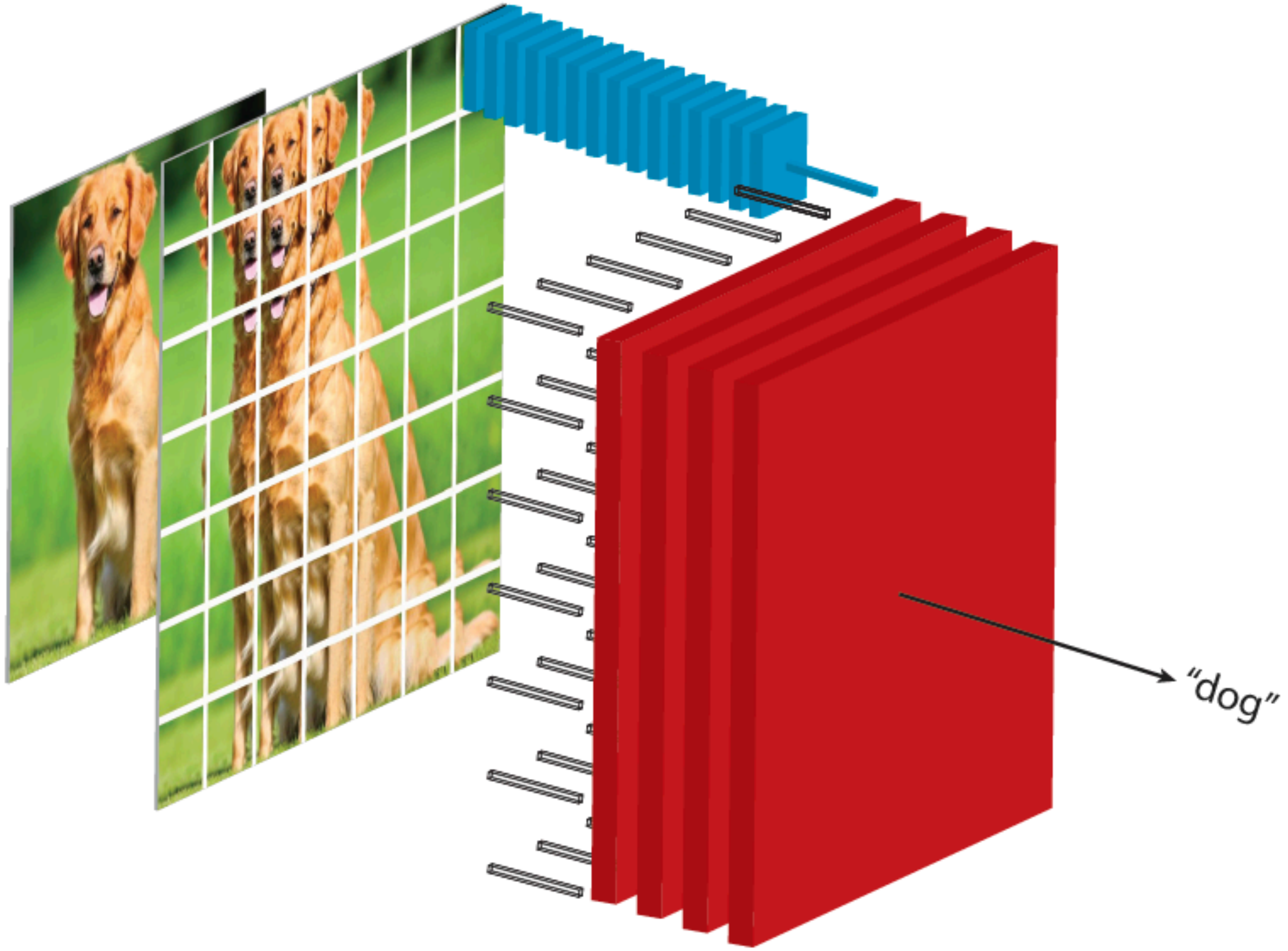
Large Network

- ResNet

layer name	output size	18-layer	34-layer	50-layer	101-layer	152-layer
conv1	112×112	7×7, 64, stride 2				
conv2_x	56×56	3×3 max pool, stride 2				
		[3×3, 64 3×3, 64] ×2	[3×3, 64 3×3, 64] ×3	[1×1, 64 3×3, 64 1×1, 256] ×3	[1×1, 64 3×3, 64 1×1, 256] ×3	[1×1, 64 3×3, 64 1×1, 256] ×3
conv3_x	28×28	[3×3, 128 3×3, 128] ×2	[3×3, 128 3×3, 128] ×4	[1×1, 128 3×3, 128 1×1, 512] ×4	[1×1, 128 3×3, 128 1×1, 512] ×4	[1×1, 128 3×3, 128 1×1, 512] ×8
conv4_x	14×14	[3×3, 256 3×3, 256] ×2	[3×3, 256 3×3, 256] ×6	[1×1, 256 3×3, 256 1×1, 1024] ×6	[1×1, 256 3×3, 256 1×1, 1024] ×23	[1×1, 256 3×3, 256 1×1, 1024] ×36
conv5_x	7×7	[3×3, 512 3×3, 512] ×2	[3×3, 512 3×3, 512] ×3	[1×1, 512 3×3, 512 1×1, 2048] ×3	[1×1, 512 3×3, 512 1×1, 2048] ×3	[1×1, 512 3×3, 512 1×1, 2048] ×3
	1×1	average pool, 1000-d fc, softmax				
FLOPs		1.8×10^9	3.6×10^9	3.8×10^9	7.6×10^9	11.3×10^9

- The third residual stack of ResNet-101 contains
  - **23** blocks with a
  - **1024**-dimensional feature maps and
  - **256**- dimensional bottleneck layers.
- We increased the model's capacity by growing this component to
  - **46** blocks with
  - **4096**-dimensional feature maps and
  - **512**-dimensional bottleneck layers, and call the resulting network ResNet-170.

Semi-supervised learning



- Having trained the encoder network, the context network is discarded and replaced by a classifier network (red) which can be trained in a supervised manner.

- For some experiments, we also fine-tune the encoder network (blue) for the classification task.

Semi-supervised learning

- **Frozen regime**: optimizing a feature extractor f solely for the CPC objective. Its parameters are then fixed and a classifier g is optimized to discriminate the output of the feature extractor

$$\theta^* = \arg\min_{\theta} \frac{1}{N} \sum_{n=1}^{N} \mathcal{L}_{\text{CPC}}[f_\theta(x_n)].$$

And given a (potentially much smaller) dataset of M labeled images $\{x_m, y_m\}$

$$\phi^* = \arg\min_{\phi} \frac{1}{M} \sum_{m=1}^{M} \mathcal{L}_{\text{Sup}}[g_\phi \circ f_{\theta^*}(x_m), y_m].$$

- **Fine-tuning regime**

Semi-supervised learning

- **Image classification**:
  - The classifier g is an 11-block ResNet architecture with 4096-dimensional feature maps and 1024-dimensional bottleneck layers.
  - The supervised loss L_Sup is the cross entropy between model predictions and image labels.

- **Objective detection**:
  - We use the Faster-RCNN architecture and loss, without any modification

Results:ImageNet classification

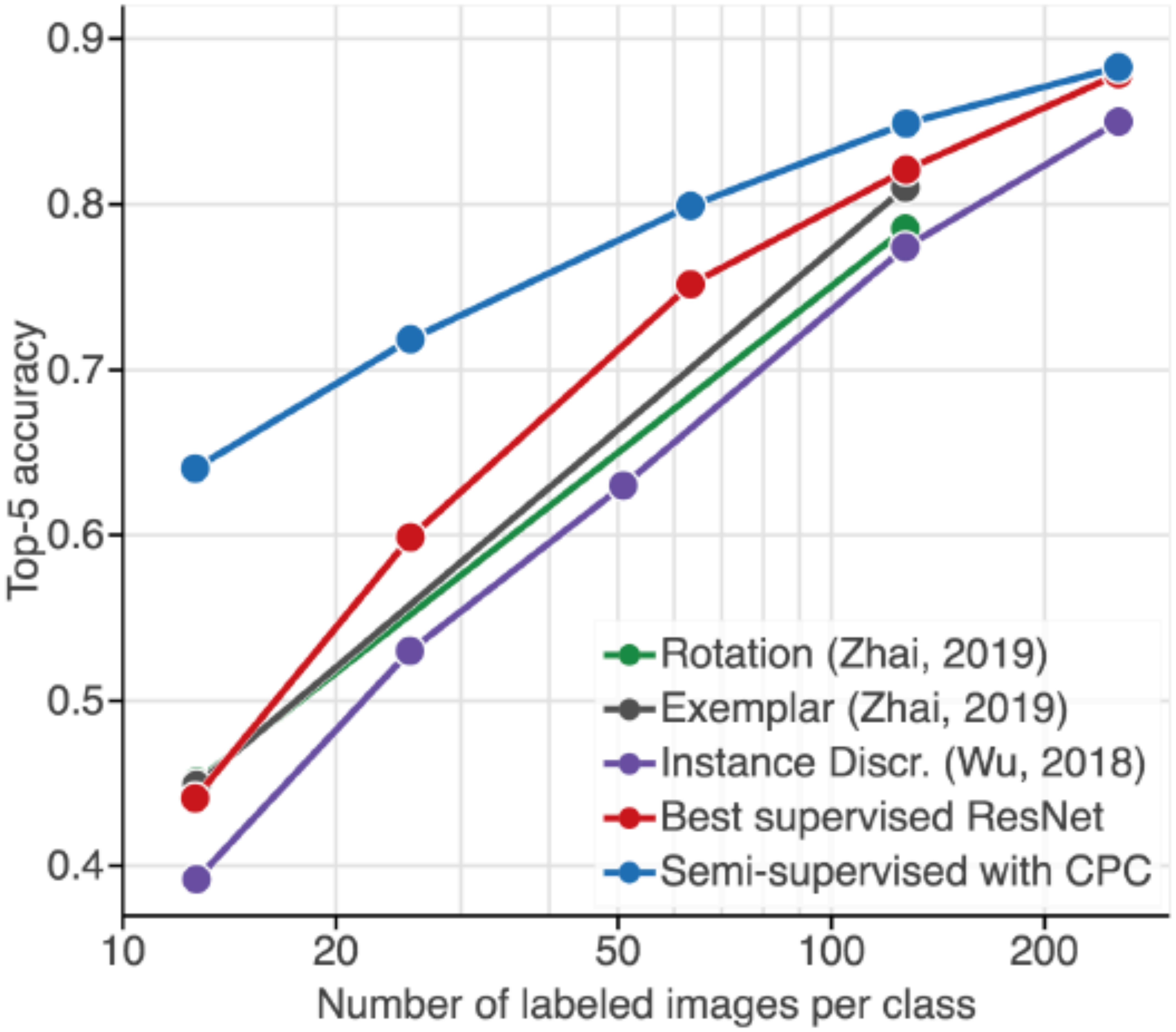
- Linear separability
- Comparison to linear separability of other self-supervised methods.
- In all cases a feature extractor is optimized in an **unsupervised** manner, and a linear classifier is trained using **all labels** in the ImageNet dataset.

Method	Top-1	Top-5
Motion Segmentation (MS) [50]	27.6	48.3
Exemplar (Ex) [17]	31.5	53.1
Relative Position (RP) [14]	36.2	59.2
Colorization (Col) [69]	39.6	62.5
Combination of MS + Ex + RP + Col [15]	-	69.3
CPC [49]	48.7	73.6
Rotation + RevNet [36]	55.4	-
CPC (ours)	**61.0**	**83.0**

Table 1. Comparison to linear separability of other self-supervised methods. In all cases a feature extractor is optimized in an unsupervised manner, and a linear classifier is trained using all labels in the ImageNet dataset.

Low-data classification

- Comparison to other methods for semi-supervised learning via self-supervised learning followed by supervised fine-tuning.
  - Blue: semi-supervised learning with CPC.
  - Purple: semisupervised learning with instance discrimination [64].
  - Green: semi-supervised learning with rotation prediction [68].
  - Grey: semi-supervised learning with exemplar learning [68].
  - Red: our supervised baseline.



Low-data classification

- Comparison to other methods for semi-supervised learning using 1% or 10% of labeled data.
- Representation learning methods learn a representation in an unsupervised manner and use it for classification.
- The classifier only considers labeled examples, and is only constrained by the supervised objective.

Labeled data	1%	10%
Method	**Top-5 accuracy**	
Supervised baseline	44.10	82.08
Methods using label-propagation:		
Pseudolabeling [68]	51.56	82.41
VAT [68]	44.05	82.78
VAT + Entropy Minimization [68]	46.96	83.39
Unsup. Data Augmentation [65]	-	88.52
Rotation + VAT + Ent. Min. [68]	-	**91.23**
Methods only using representation learning:		
Instance Discrimination [64]	39.20	77.40
Exemplar [68]	44.90	81.01
Exemplar (joint training) [68]	47.02	83.72
Rotation [68]	45.11	78.53
Rotation (joint training) [68]	53.37	83.82
CPC (ours)	**64.03**	**84.88**

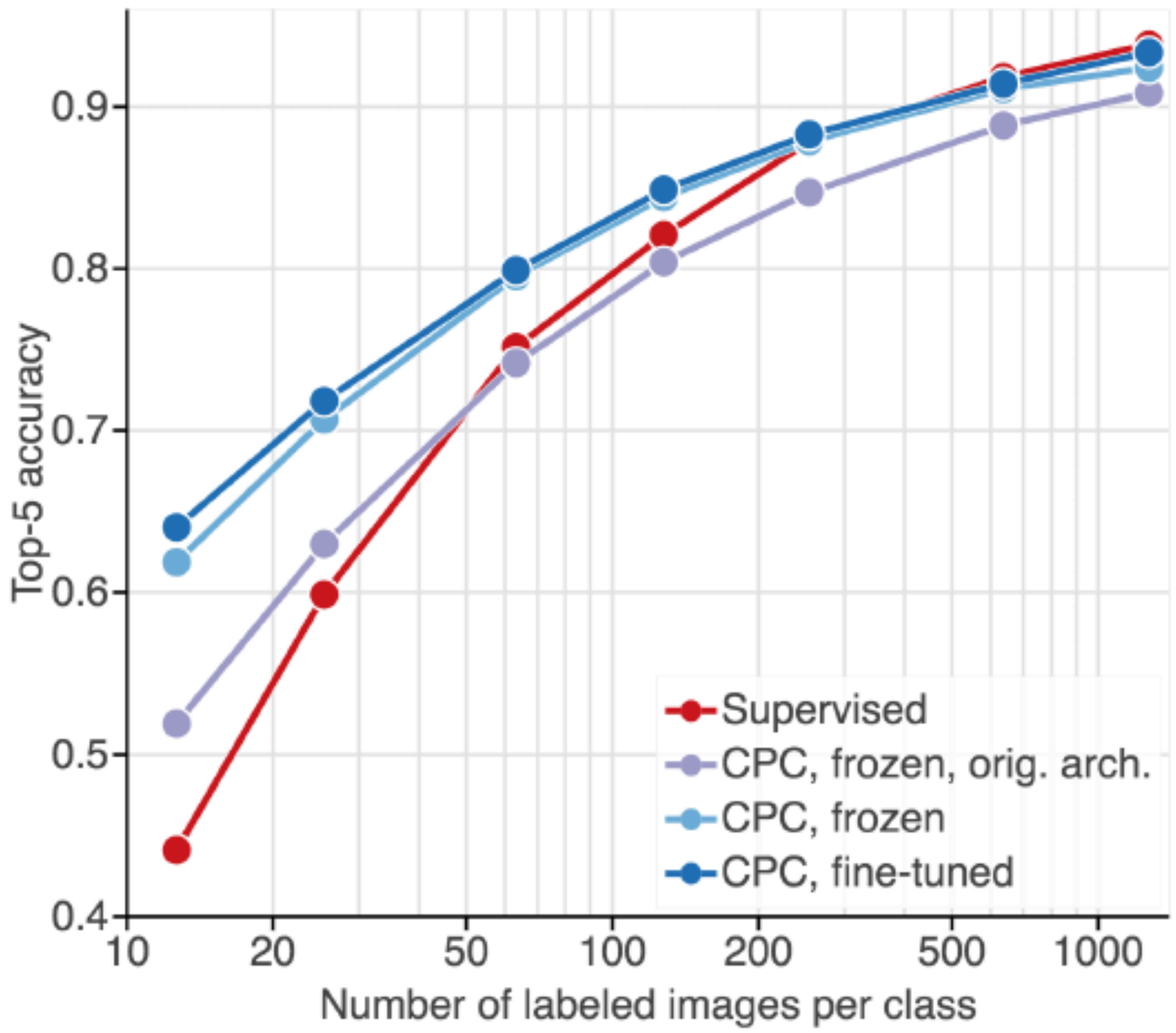
Transfer to objective detection

- Comparison of PASCAL 2007 image detection accuracy to other transfer methods.
- The first class of methods learn from unlabeled ImageNet data and fine-tune for PASCAL detection.
- The second class learns from the entire labeled ImageNet dataset before transferring.
- All results are reported in terms of mean average precision (mAP).

Method	**mAP**
Transfer from labeled ImageNet:	
Supervised - ResNet-152	74.7
Transfer from unlabeled ImageNet:	
Exemplar (Ex) [17]	60.9
Motion Segmentation (MS) [50]	61.1
Colorization (Col) [69]	65.5
Relative Position (RP) [14]	66.8
Combination of Ex + MS + Col + RP [15]	70.5
Deep Cluster [8]	65.9
Deeper Cluster [9]	67.8
CPC - ResNet-101	70.6
CPC - ResNet-170	**72.1**

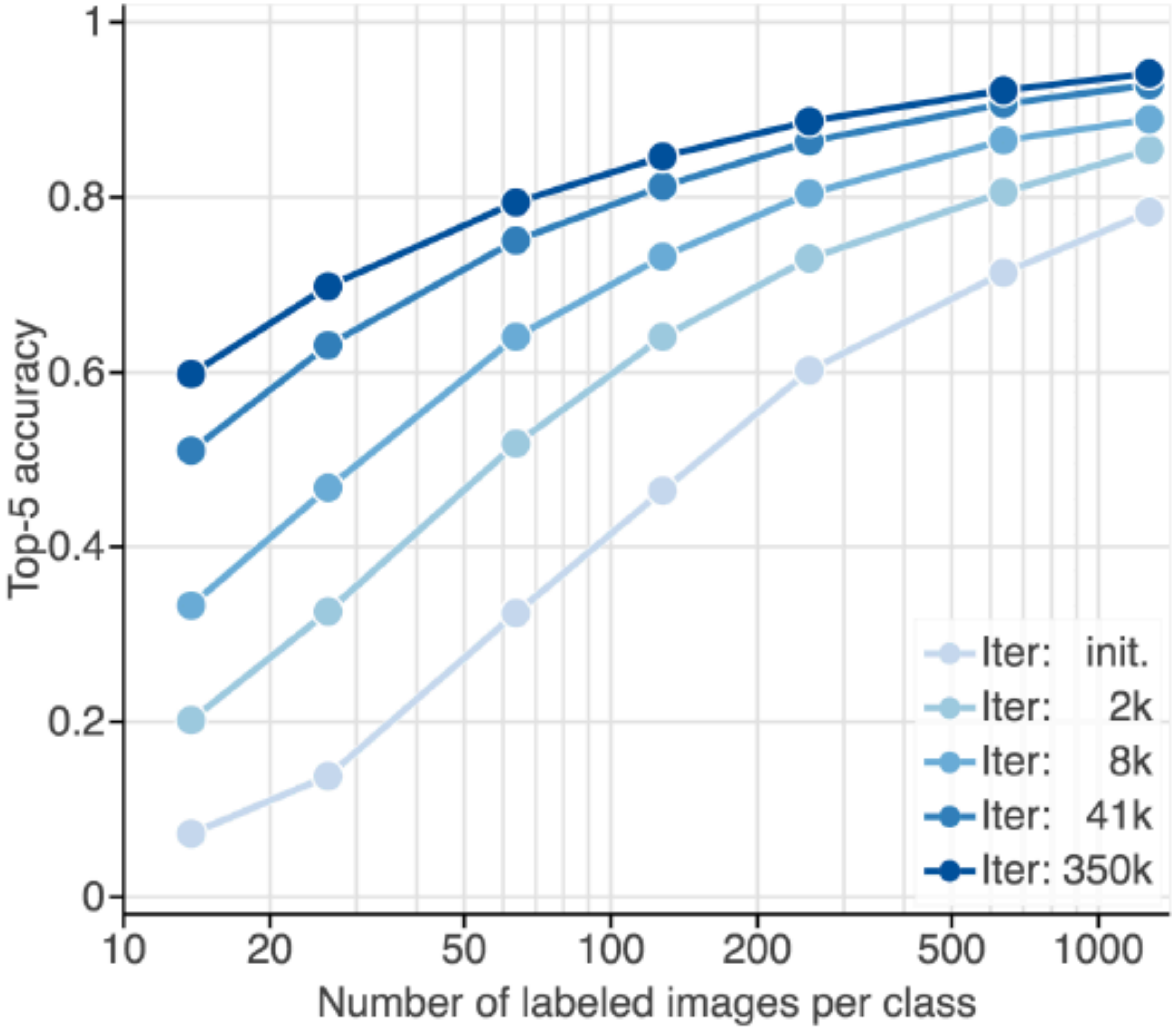
Frozen vs Fine-tuned

- Contribution of unsupervised learning and fine-tuning to recognition performance.
  - Light blue: classification performance of a frozen feature extractor followed by a supervised classifier.
  - Purple: similarly, but with the original CPC architecture.
  - Dark blue: classification performance of the fine-tuned model.
  - Red: fully supervised baseline.



Learning dynamics

- Image recognition accuracy over the course of CPC training.
- Without training, the ResNet-170 architecture achieves very low performance across data regimes.
- Over the course of training, this performance increases rapidly, reaching our final result after 350k iterations.



Conclusion

- The result is a representation which, equipped with a simple linear classifier, separates ImageNet categories better than all competing methods, and surpasses the performance of a fully-supervised AlexNet model.
- When given a small number of labeled images (as few as 13 per class), this representation retains a strong classification performance, outperforming state-of the- art semi-supervised methods by 10% Top-5 accuracy and supervised methods by 20%.
- Finally, we find our unsupervised representation to serve as a useful substrate for image detection on the PASCAL-VOC 2007 dataset, approaching the performance of representations trained with a fully annotated ImageNet dataset.
- We expect these results to open the door to pipelines that use scalable unsupervised representations as a drop-in replacement for supervised ones for real-world vision tasks where labels are scarce.