A critical analysis of self-supervision, or what we can learn from a single image

Yuki M. Asano

CDT Annual Meeting Oct 2019

work with *Christian Rupprecht* and *Andrea Vedaldi* at VGG

AIMS
EPSRC Centre for Doctoral Training in
Autonomous Intelligent Machines & Systems

VGG
University of Oxford

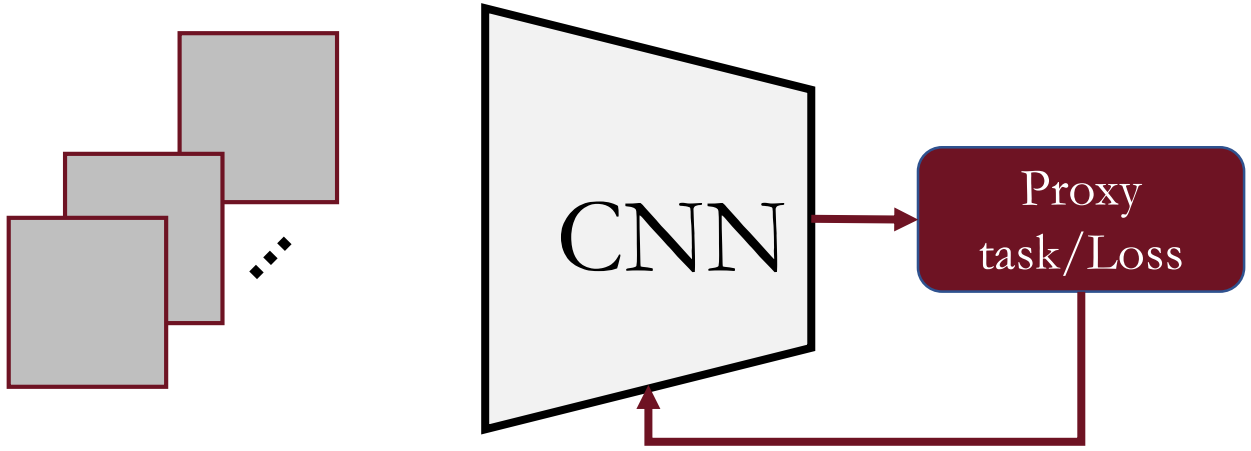
Outline

- Self-supervised learning saga
- *Or is it?*

Self-supervised learning like we do

1. Unlabelled, large collection of images
2. Train your network *without labels*
3. Use the image representations (vectors) for new tasks

Self-supervised learning like we do?



Unlabelled data
+ transformations

e.g. DeepCluster

- Run k-means on features
- Train classifier on k classes
- Repeat for 200 epochs

e.g. RotNet

- Create 4 classes based on rotations
- Exploits photographer bias
- Simple but works

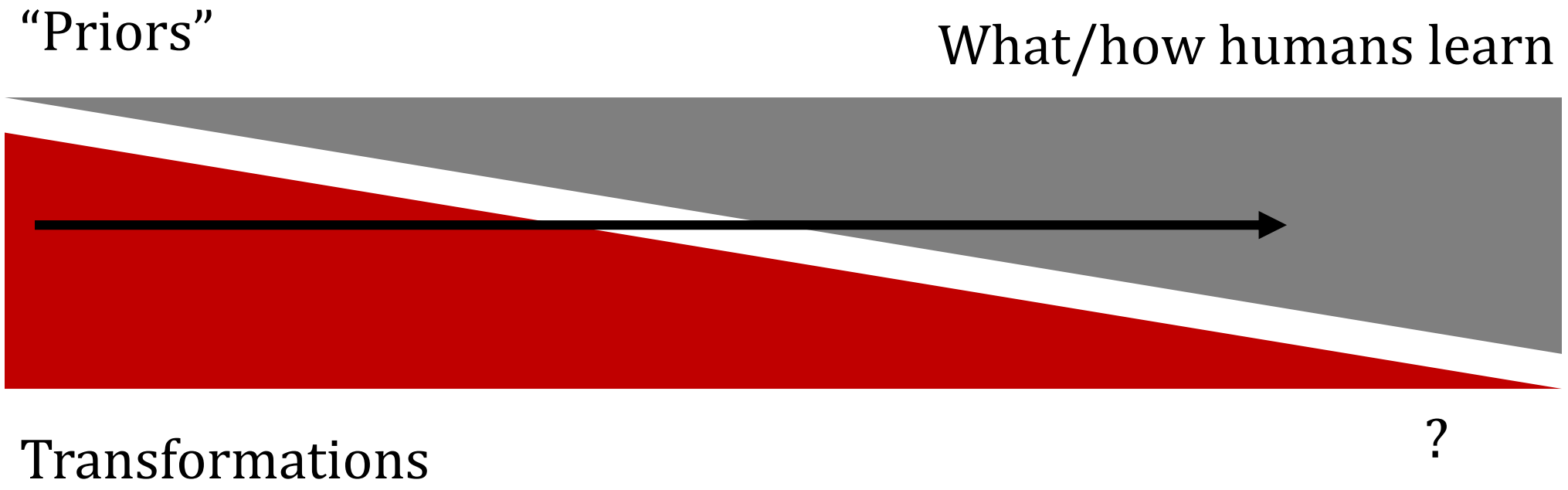
Deepclustering for unsupervised learning of visual features
M. Caron, P. Bojanowski, A. Joulin, and M. Douze
ECCV, 2018

Unsupervised representation learning by predicting image rotations
S. Gidaris, P. Singh, and N. Komodakis
ICLR, 2018

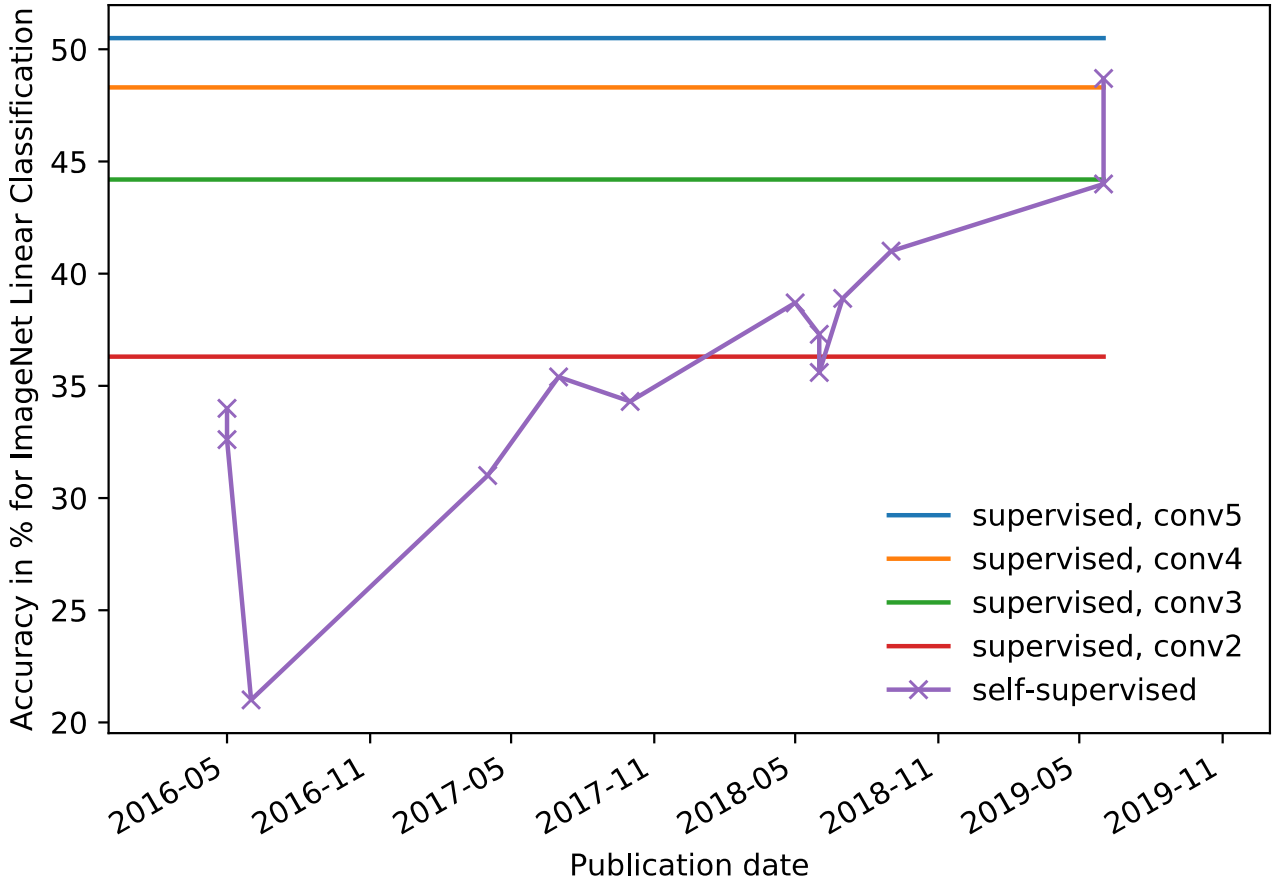
Or colorizing images

Zhang, Isola, Efros.
Colorful Image Colorization.
In ECCV, 2016

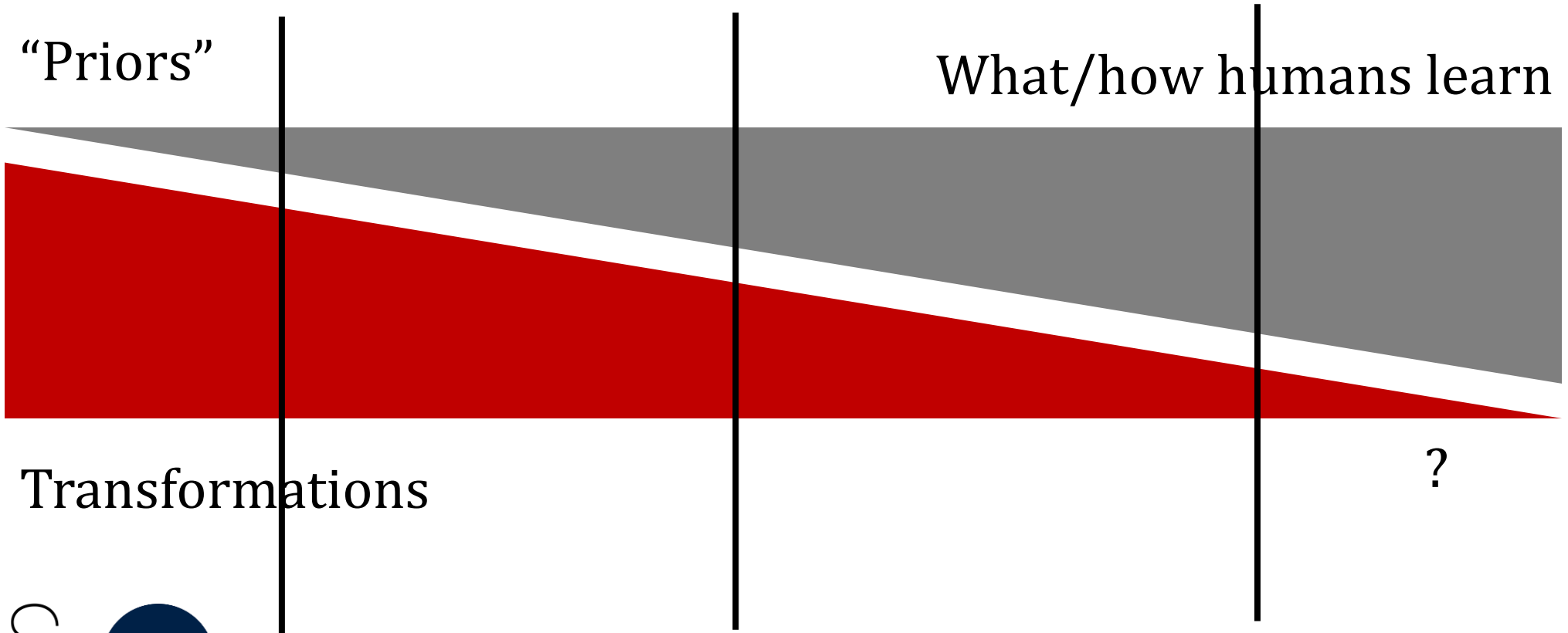
Hypothesis



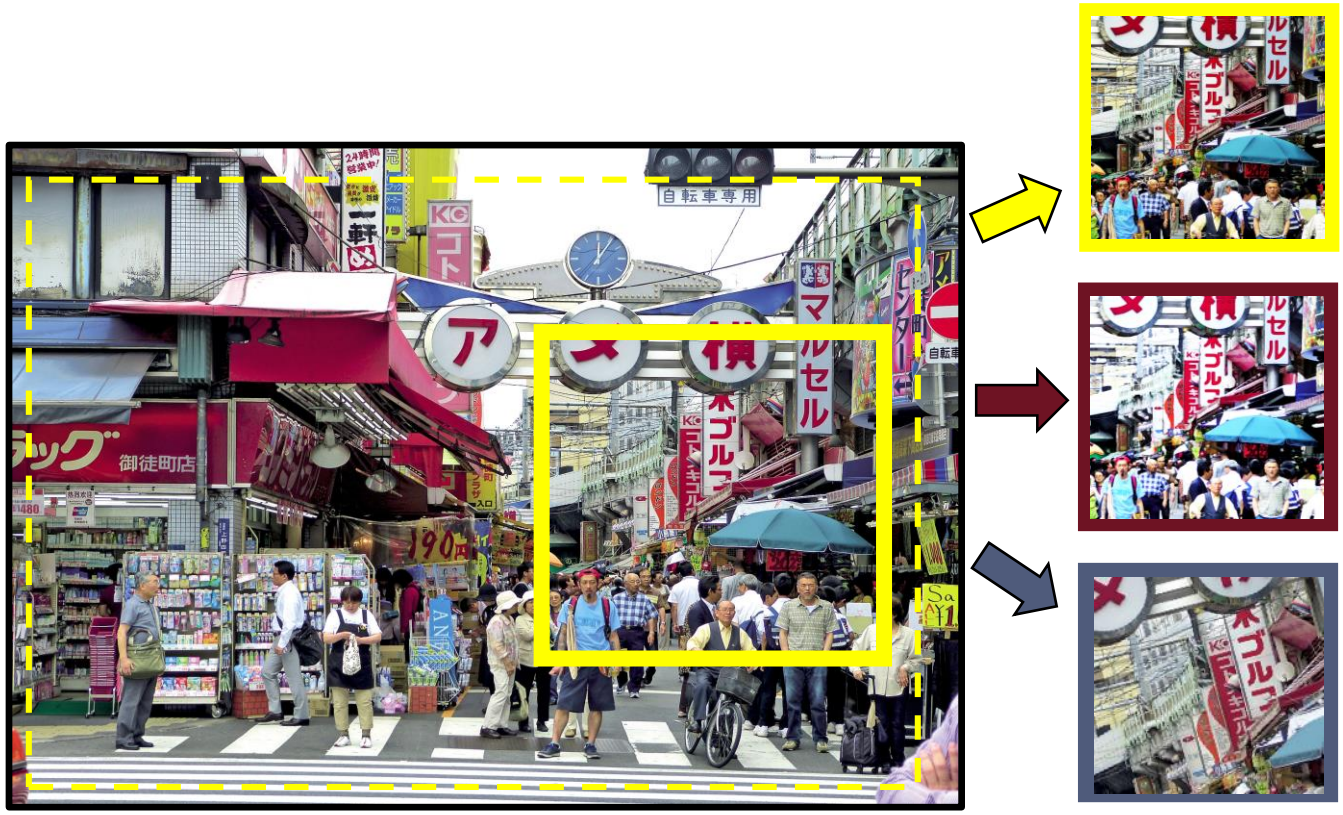
Getting there, but not quite yet



Where are we?



“Learn” from one image…
using multiple transformations

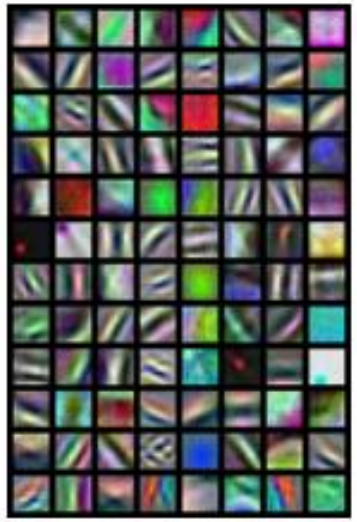
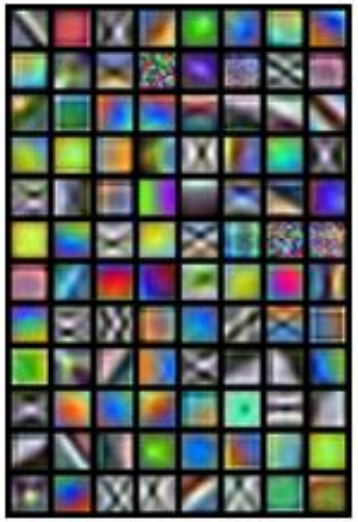
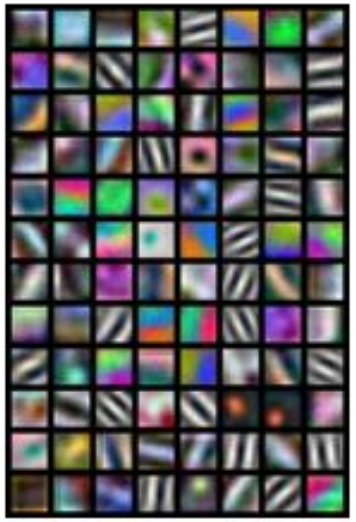


AIMS
EPSRC Centre for Doctoral Training in
Autonomous Intelligent Machines & Systems

VGG
University of Oxford

A critical analysis of self-supervision, or what we can learn from a single image
YM Asano, C Rupprecht, A Vedaldi
arXiv 1904.13132

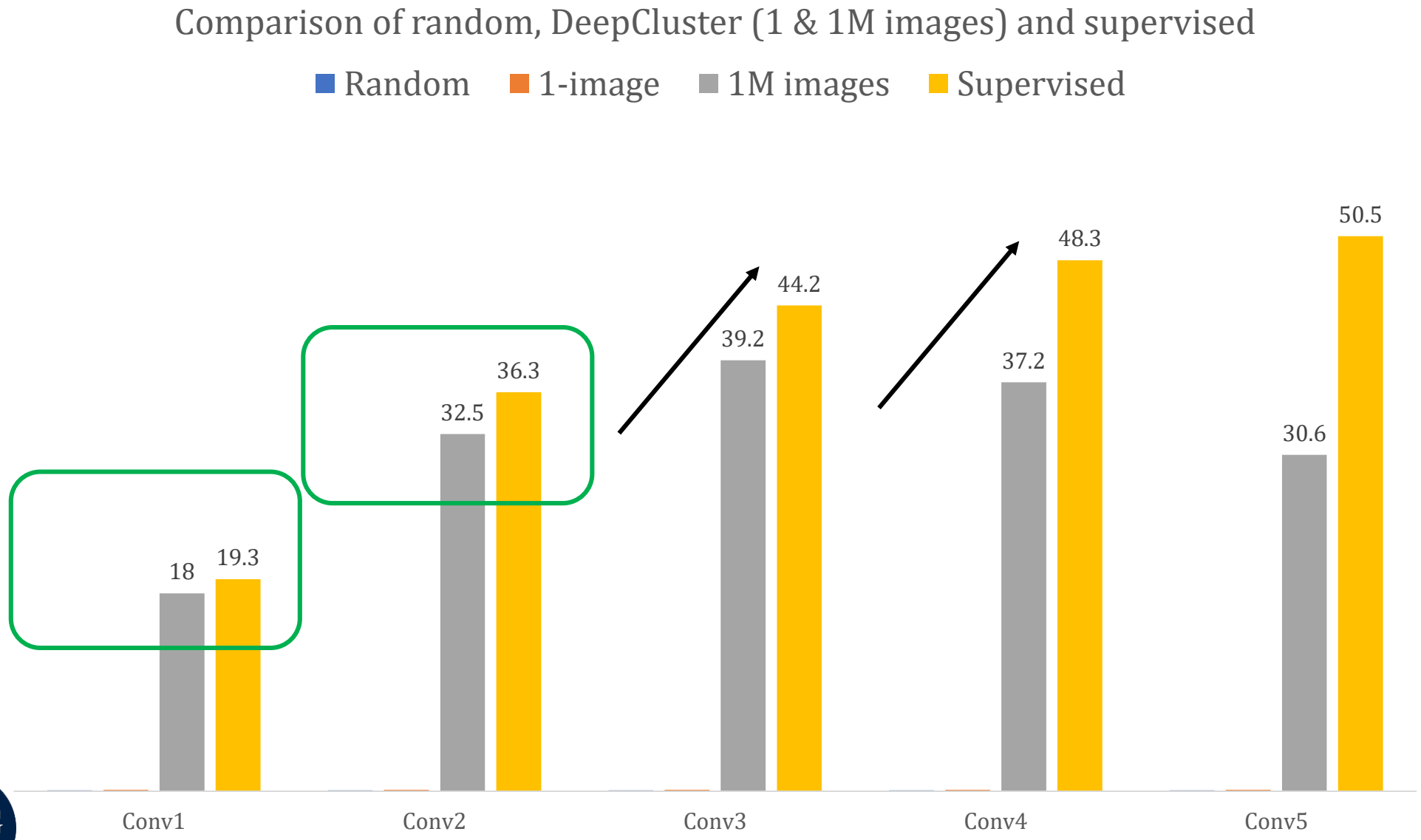
Learned first convolutional layer – from one image

Method, Image A			Method, Image B		
BiGAN	RotNet	DeepCluster	BiGAN	RotNet	DeepCluster
					

AIMS
EPSRC Centre for Doctoral Training in
Autonomous Intelligent Machines & Systems

VGG
University of Oxford

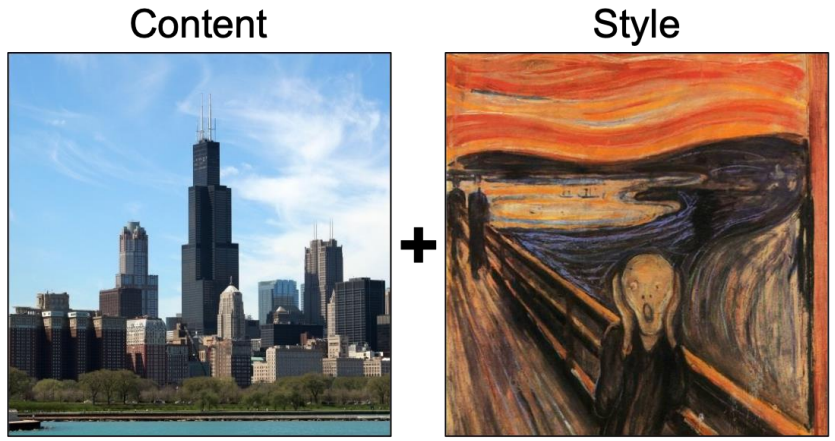
Performance



Conclusion

1. Early layers of deep networks contain limited information about natural images
2. These can be learned through self-supervision or supervised learning
3. Notably, only one **image + transformations** are necessary for this
4. Much space to go the *right* direction in self-supervised learning

Style transfer with a 1-image trained CNN



yuki@robots.ox.ac.uk
@y_m_asa

EPSRC Centre for Doctoral Training in Autonomous Intelligent Machines & Systems
VGG University of Oxford
EPSRC Engineering and Physical Sciences Research Council

Appendix

Table 4: **Finetuning experiments** The pre-trained model's conv1 and conv2 are left frozen and only the higher levels are re-trained using ImageNet LSVRC-12 training set. Accuracy is averaged over 10 crops.

	c1	c2	c3	c4	c5
Full sup.	19.3	36.3	44.2	48.3	50.5
BiGAN, A	22.5	37.6	44.2	47.6	48.3
RotNet, A	22.0	38.2	44.8	49.2	51.8
DeepCluster, A	21.8	35.9	43.6	48.8	50.4

		CIFAR-10			
		conv1	conv2	conv3	conv4
(a)	Fully sup.	66.5	70.1	72.4	75.9
(b)	Random feat.	57.8	55.5	54.2	47.3
(c)	No aug.	57.9	56.2	54.2	47.8
(d)	Jitter	58.9	58.0	57.0	49.8
(e)	Rotation	61.4	58.8	56.1	47.5
(f)	Scale	67.9	69.3	67.9	59.1
(g)	Rot. & jitter	64.9	63.6	61.0	53.4
(h)	Rot. & scale	67.6	69.9	68.0	60.7
(i)	Jitter & scale	68.1	71.3	69.5	62.4
(j)	All	**68.1**	**72.3**	**70.8**	**63.5**