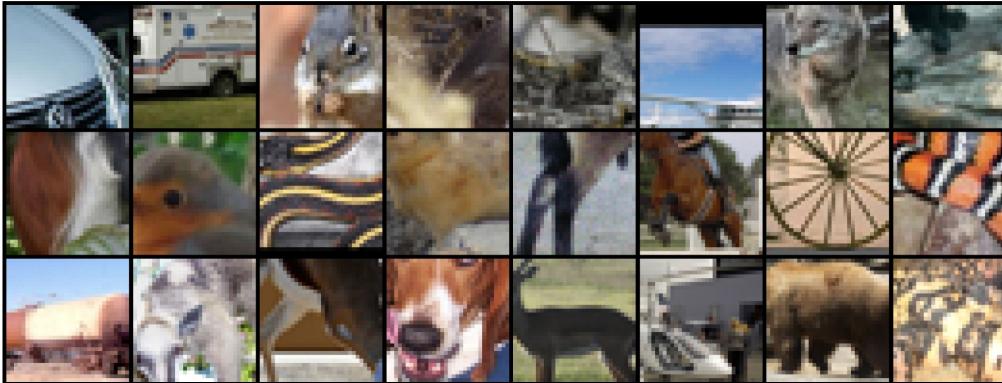


# Классические задачи предобучения

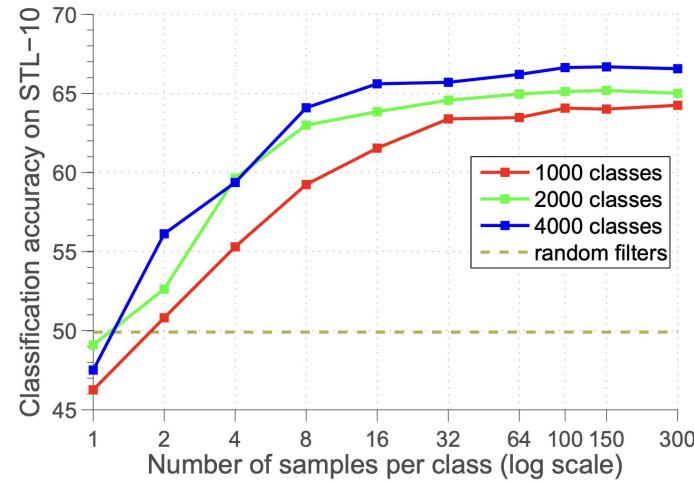
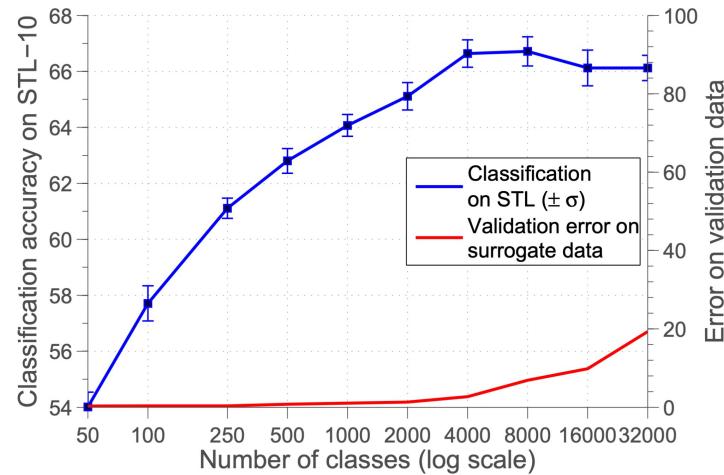
ВШЭ ФКН, Методы предобучения без учителя

Шабалин Александр

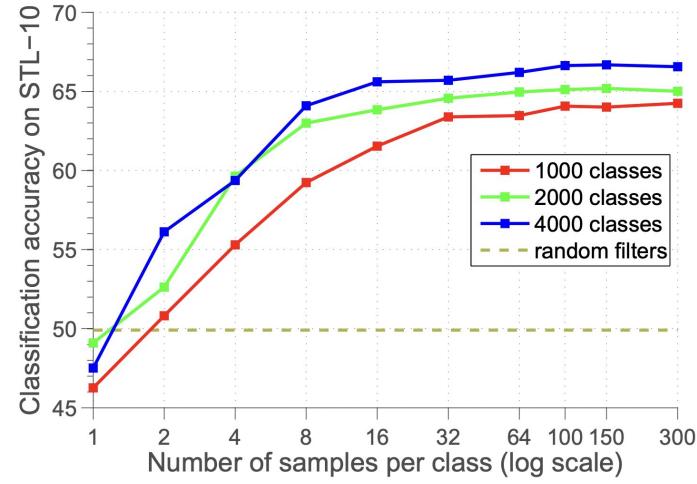
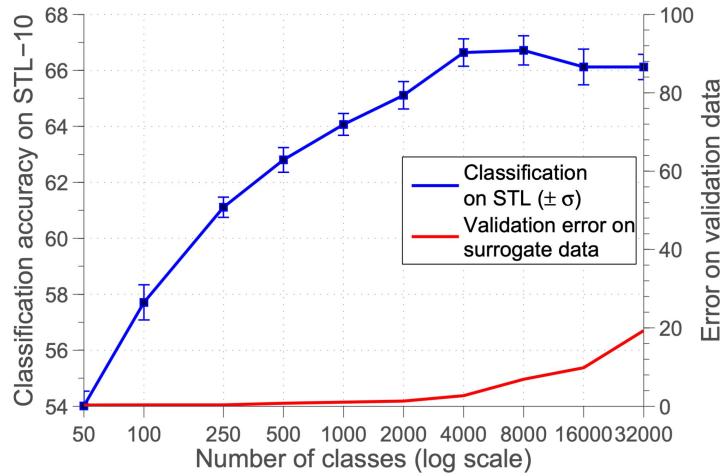
# Exemplar, 2014, Dosovitskiy et al.



# Exemplar



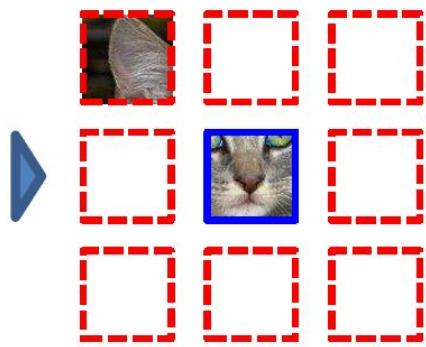
# Exemplar



Algorithm	STL-10	CIFAR-10(400)	CIFAR-10	Caltech-101	Caltech-256(30)	#features
Convolutional K-means Network [32]	$60.1 \pm 1$	$70.7 \pm 0.7$	82.0	—	—	8000
Multi-way local pooling [33]	—	—	—	$77.3 \pm 0.6$	41.7	$1024 \times 64$
Slowness on videos [14]	61.0	—	—	74.6	—	556
Hierarchical Matching Pursuit (HMP) [34]	$64.5 \pm 1$	—	—	—	—	1000
Multipath HMP [35]	—	—	—	$82.5 \pm 0.5$	50.7	5000
View-Invariant K-means [16]	63.7	$72.6 \pm 0.7$	81.9	—	—	6400
Exemplar-CNN (64c5-64c5-128f)	$67.1 \pm 0.2$	$69.7 \pm 0.3$	76.5	$79.8 \pm 0.5^*$	$42.4 \pm 0.3$	256
Exemplar-CNN (64c5-128c5-256c5-512f)	$72.8 \pm 0.4$	$75.4 \pm 0.2$	82.2	$86.1 \pm 0.5^\dagger$	$51.2 \pm 0.2$	960
Exemplar-CNN (92c5-256c5-512c5-1024f)	<b><math>74.2 \pm 0.4</math></b>	<b><math>76.6 \pm 0.2</math></b>	<b>84.3</b>	<b><math>87.1 \pm 0.7^\ddagger</math></b>	<b><math>53.6 \pm 0.2</math></b>	1884
Supervised state of the art	70.1[36]	—	92.0 [37]	91.44 [38]	70.6 [2]	—

# Context Prediction, 2015, Doersch et al.

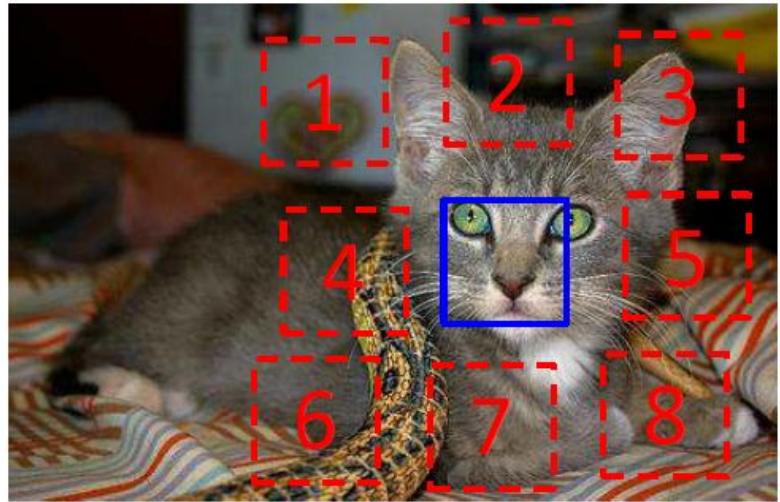
Example:



Question 1:

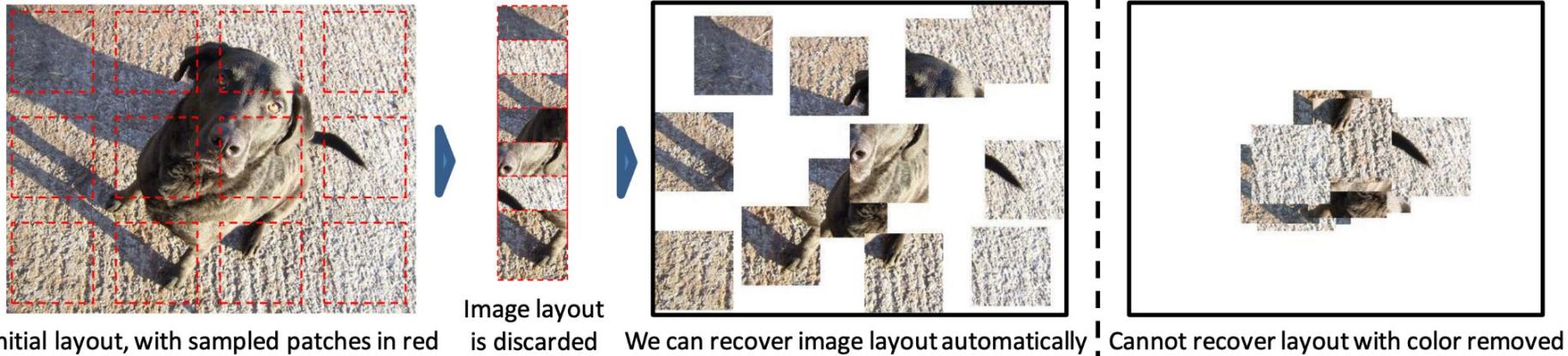


Question 2:



$$X = (\text{cat face}, \text{ear}); Y = 3$$

# Context Prediction



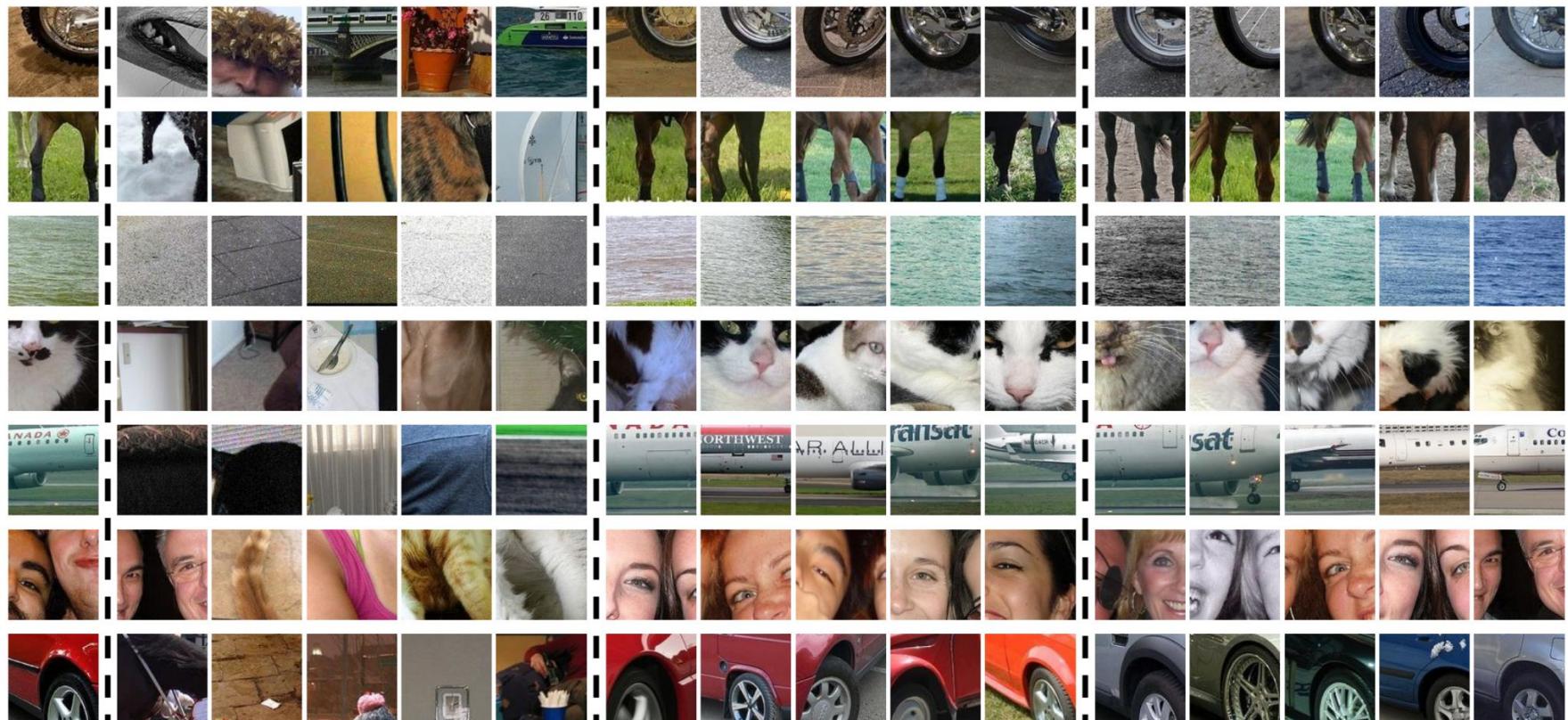
# Context Prediction

VOC-2007 Test	aero	bike	bird	boat	bottle	bus	car	cat	chair	cow	table	dog	horse	mbike	person	plant	sheep	sofa	train	tv	mAP
<b>DPM-v5[17]</b>	33.2	60.3	10.2	16.1	27.3	54.3	58.2	23.0	20.0	24.1	26.7	12.7	58.1	48.2	43.2	12.0	21.1	36.1	46.0	43.5	33.7
[8] w/o context	52.6	52.6	19.2	25.4	18.7	47.3	56.9	42.1	16.6	41.4	41.9	27.7	47.9	51.5	29.9	20.0	41.1	36.4	48.6	53.2	38.5
<b>Regionlets[58]</b>	54.2	52.0	20.3	24.0	20.1	55.5	68.7	42.6	19.2	44.2	49.1	26.6	57.0	54.5	43.4	16.4	36.6	37.7	59.4	52.3	41.7
<b>Scratch-R-CNN[2]</b>	49.9	60.6	24.7	23.7	20.3	52.5	64.8	32.9	20.4	43.5	34.2	29.9	49.0	60.4	47.5	28.0	42.3	28.6	51.2	50.0	40.7
<b>Scratch-Ours</b>	52.6	60.5	23.8	24.3	18.1	50.6	65.9	29.2	19.5	43.5	35.2	27.6	46.5	59.4	46.5	25.6	42.4	23.5	50.0	50.6	39.8
<b>Ours-projection</b>	58.4	62.8	33.5	27.7	24.4	58.5	68.5	41.2	26.3	49.5	42.6	37.3	55.7	62.5	49.4	29.0	47.5	28.4	54.7	56.8	45.7
<b>Ours-color-dropping</b>	60.5	66.5	29.6	28.5	26.3	56.1	70.4	44.8	24.6	45.5	45.4	35.1	52.2	60.2	50.0	28.1	46.7	42.6	54.8	58.6	46.3
<b>Ours-Yahoo100m</b>	56.2	63.9	29.8	27.8	23.9	57.4	69.8	35.6	23.7	47.4	43.0	29.5	52.9	62.0	48.7	28.4	45.1	33.6	49.0	55.5	44.2
<b>ImageNet-R-CNN[21]</b>	64.2	69.7	50	41.9	32.0	62.6	71.0	60.7	32.7	58.5	46.5	56.1	60.6	66.8	54.2	31.5	52.8	48.9	57.9	64.7	54.2
<b>K-means-rescale [31]</b>	55.7	60.9	27.9	30.9	12.0	59.1	63.7	47.0	21.4	45.2	55.8	40.3	67.5	61.2	48.3	21.9	32.8	46.9	61.6	51.7	45.6
<b>Ours-rescale [31]</b>	61.9	63.3	35.8	32.6	17.2	68.0	67.9	54.8	29.6	52.4	62.9	51.3	67.1	64.3	50.5	24.4	43.7	54.9	67.1	52.7	51.1
<b>ImageNet-rescale [31]</b>	64.0	69.6	53.2	44.4	24.9	65.7	69.6	69.2	28.9	63.6	62.8	63.9	73.3	64.6	55.8	25.7	50.5	55.4	69.3	56.4	56.5
<b>VGG-K-means-rescale</b>	56.1	58.6	23.3	25.7	12.8	57.8	61.2	45.2	21.4	47.1	39.5	35.6	60.1	61.4	44.9	17.3	37.7	33.2	57.9	51.2	42.4
<b>VGG-Ours-rescale</b>	71.1	72.4	54.1	48.2	29.9	75.2	78.0	71.9	38.3	60.5	62.3	68.1	74.3	74.2	64.8	32.6	56.5	66.4	74.0	60.3	61.7
<b>VGG-ImageNet-rescale</b>	76.6	79.6	68.5	57.4	40.8	79.9	78.4	85.4	41.7	77.0	69.3	80.1	78.6	74.6	70.1	37.5	66.0	67.5	77.4	64.9	68.6

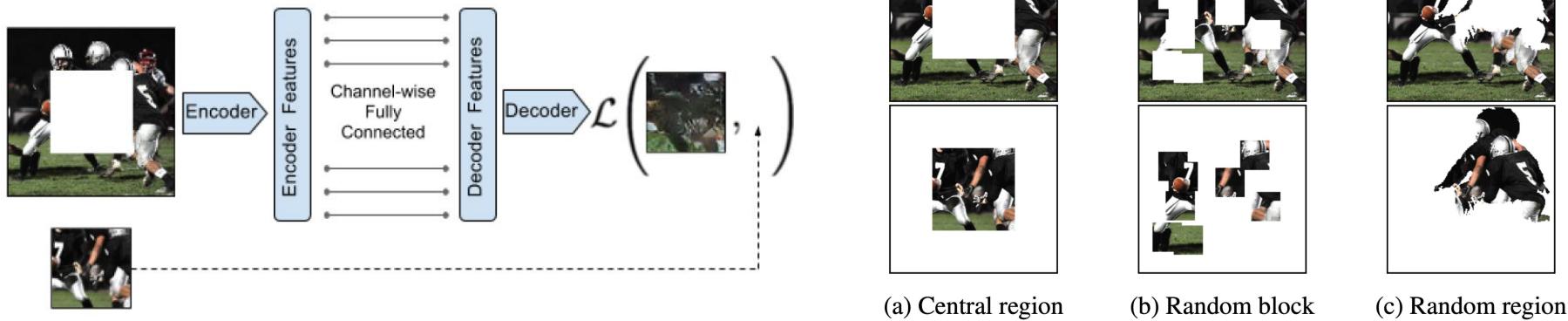
Table 1. Mean Average Precision on VOC-2007.

# Context Prediction

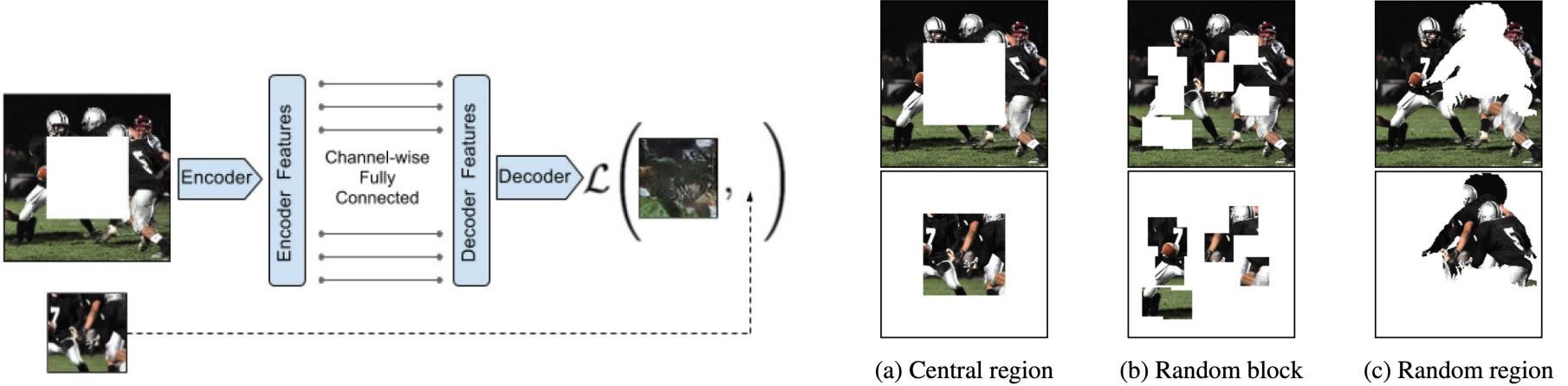
Input Random Initialization ImageNet AlexNet Ours



# Inpainting, 2016, Pathak et al.



# Inpainting, 2016, Pathak et al.



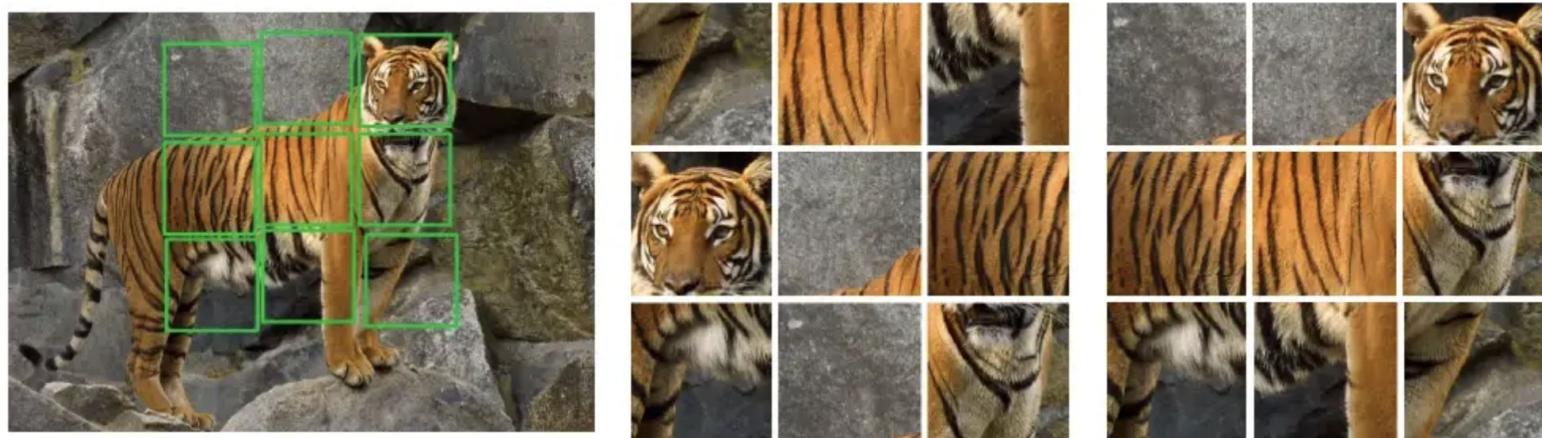
$$\mathcal{L}_{rec}(x) = \|\hat{M} \odot (x - F((1 - \hat{M}) \odot x))\|_2^2$$

$$\begin{aligned} \mathcal{L}_{adv} = \max_D \quad & \mathbb{E}_{x \in \mathcal{X}} [\log(D(x))] \\ & + \log(1 - D(F((1 - \hat{M}) \odot x))) \end{aligned}$$

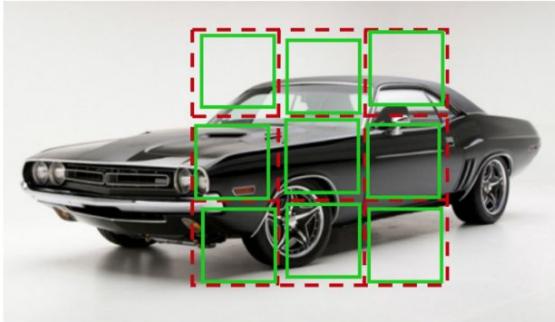
# Inpainting, 2016

Pretraining Method	Supervision	Pretraining time	Classification	Detection	Segmentation
ImageNet [26]	1000 class labels	3 days	78.2%	56.8%	48.0%
Random Gaussian	initialization	< 1 minute	53.3%	43.4%	19.8%
Autoencoder	-	14 hours	53.8%	41.9%	25.2%
Agrawal <i>et al.</i> [1]	egomotion	10 hours	52.9%	41.8%	-
Wang <i>et al.</i> [39]	motion	1 week	58.7%	47.4%	-
Doersch <i>et al.</i> [7]	relative context	4 weeks	55.3%	46.6%	-
Ours	context	14 hours	56.5%	44.5%	30.0%

# Jigsaw Puzzles, 2016, Noroozi et al.



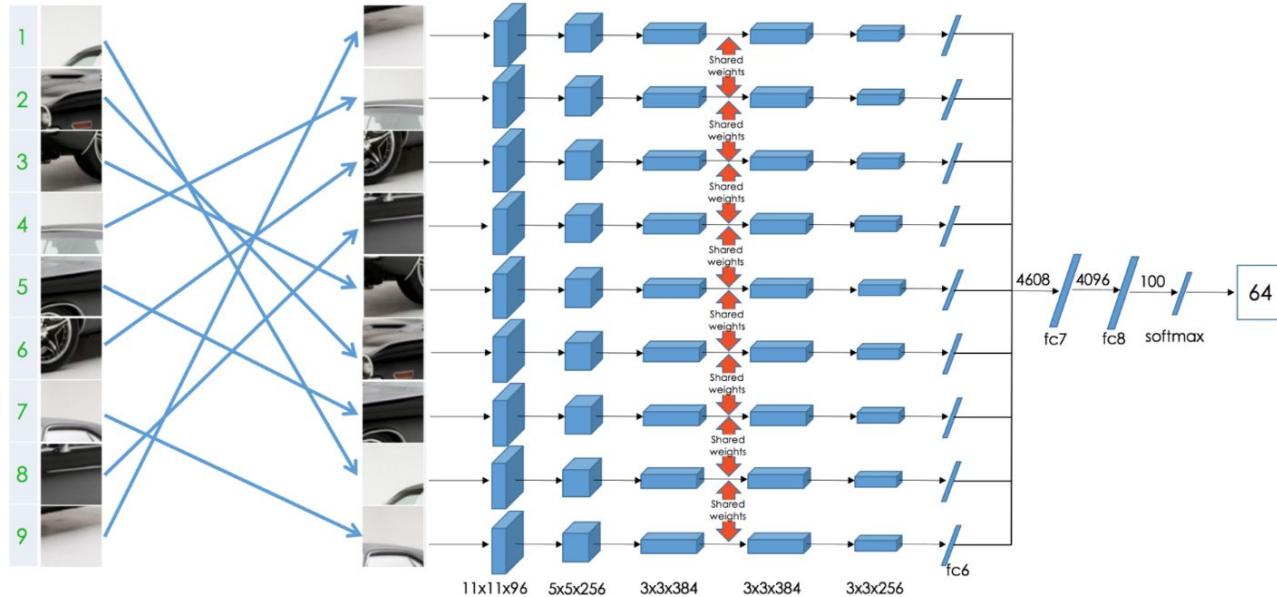
# Jigsaw Puzzles



Permutation Set

index	permutation
64	9,4,6,8,3,2,5,1,7

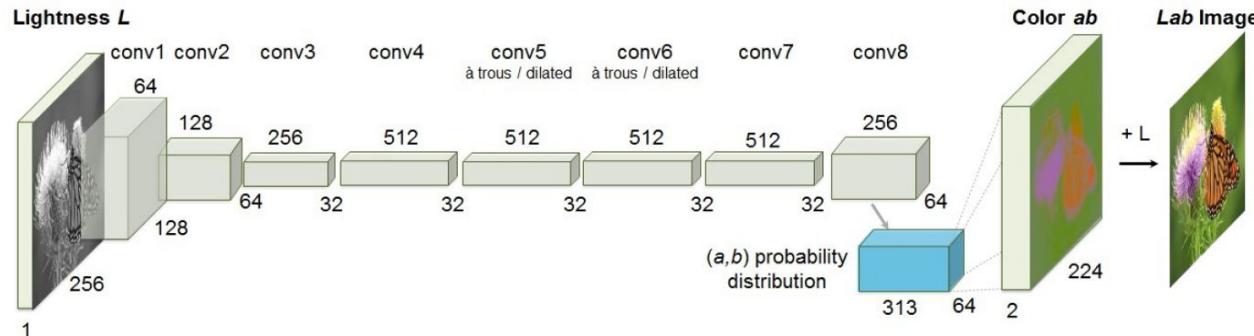
Reorder patches according to the selected permutation



# Jigsaw Puzzles

Method	Pretraining time	Supervision	Classification	Detection	Segmentation
Krizhevsky <i>et al.</i> [25]	3 days	1000 class labels	<b>78.2%</b>	<b>56.8%</b>	<b>48.0%</b>
Wang and Gupta[39]	1 week	motion	58.4%	44.0%	-
Doersch <i>et al.</i> [10]	4 weeks	context	55.3%	46.6%	-
Pathak <i>et al.</i> [30]	14 hours	context	56.5%	44.5%	29.7%
Ours	2.5 days	context	<b>67.6%</b>	<b>53.2%</b>	<b>37.6%</b>

# Colorization, 2016, Zhang et. al.



# Colorization

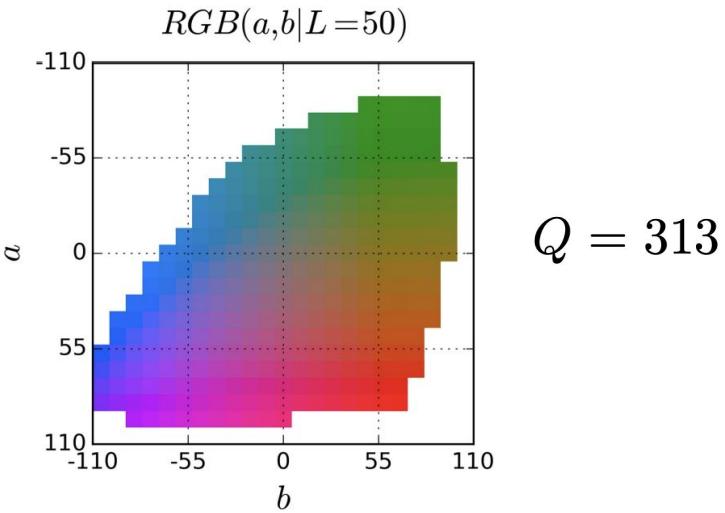
L2 Loss is **NOT** Robust!

$$L_2(\hat{\mathbf{Y}}, \mathbf{Y}) = \frac{1}{2} \sum_{h,w} \|\mathbf{Y}_{h,w} - \hat{\mathbf{Y}}_{h,w}\|_2^2$$

# Colorization

L2 Loss is **NOT** Robust!

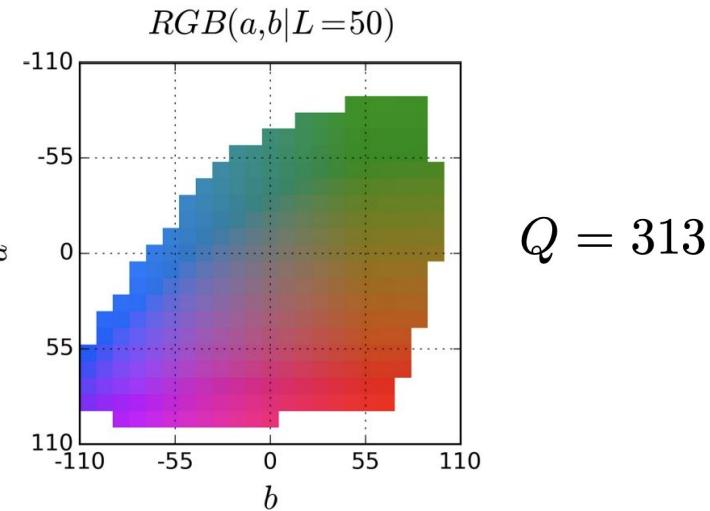
$$L_2(\hat{\mathbf{Y}}, \mathbf{Y}) = \frac{1}{2} \sum_{h,w} \|\mathbf{Y}_{h,w} - \hat{\mathbf{Y}}_{h,w}\|_2^2$$



# Colorization

L2 Loss is **NOT** Robust!

$$\text{L}_2(\hat{\mathbf{Y}}, \mathbf{Y}) = \frac{1}{2} \sum_{h,w} \|\mathbf{Y}_{h,w} - \hat{\mathbf{Y}}_{h,w}\|_2^2$$



$$\hat{\mathbf{Z}} \in [0, 1]^{H \times W \times Q}$$

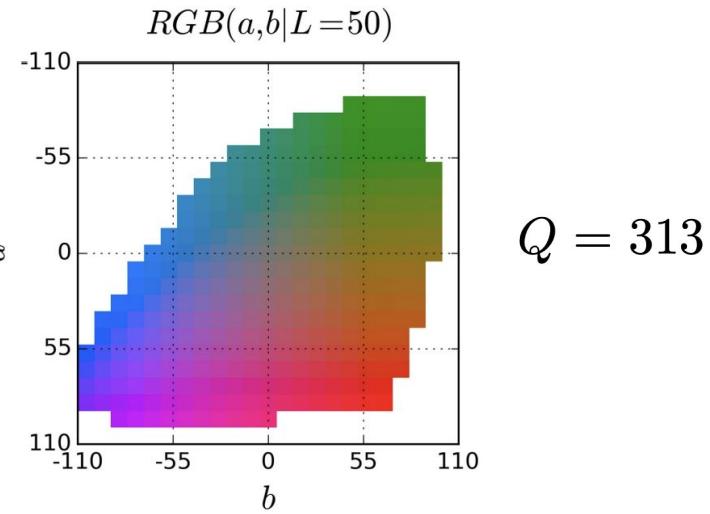
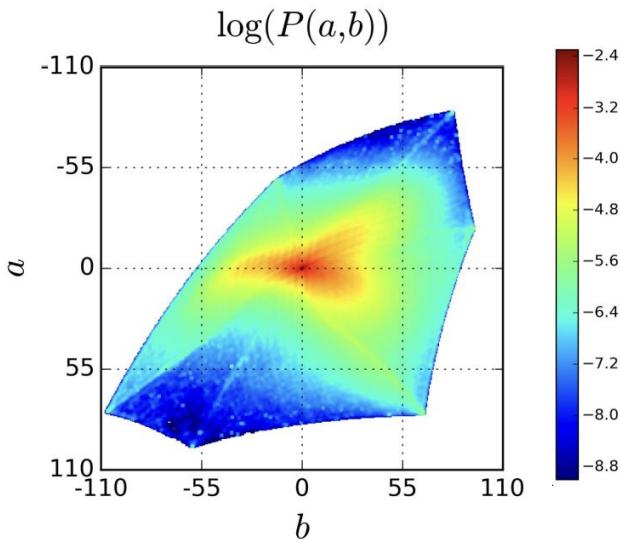
$$\mathbf{Z} = \mathcal{H}_{gt}^{-1}(\mathbf{Y})$$

$$\text{L}_{cl}(\hat{\mathbf{Z}}, \mathbf{Z}) = - \sum_{h,w} v(\mathbf{Z}_{h,w}) \sum_q \mathbf{Z}_{h,w,q} \log(\hat{\mathbf{Z}}_{h,w,q})$$

# Colorization

L2 Loss is **NOT** Robust!

$$L_2(\hat{\mathbf{Y}}, \mathbf{Y}) = \frac{1}{2} \sum_{h,w} \|\mathbf{Y}_{h,w} - \hat{\mathbf{Y}}_{h,w}\|_2^2$$



$$\hat{\mathbf{Z}} \in [0, 1]^{H \times W \times Q}$$

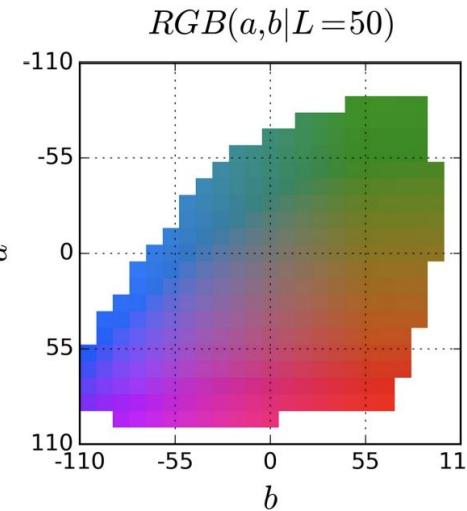
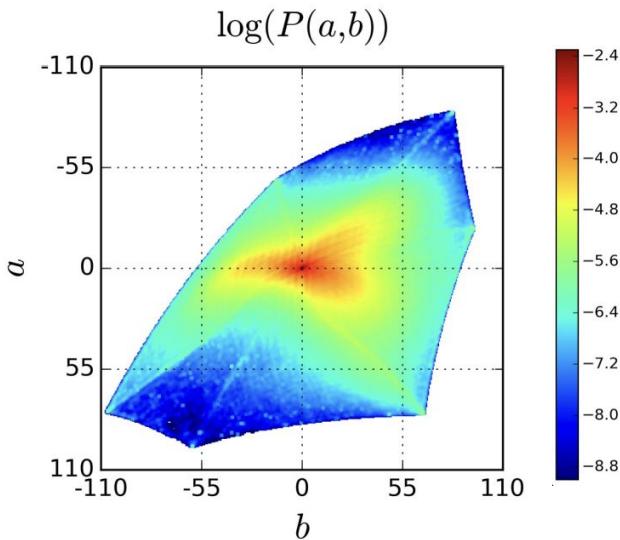
$$\mathbf{Z} = \mathcal{H}_{gt}^{-1}(\mathbf{Y})$$

$$L_{cl}(\hat{\mathbf{Z}}, \mathbf{Z}) = - \sum_{h,w} v(\mathbf{Z}_{h,w}) \sum_q \mathbf{Z}_{h,w,q} \log(\hat{\mathbf{Z}}_{h,w,q})$$

# Colorization

L2 Loss is **NOT** Robust!

$$L_2(\hat{\mathbf{Y}}, \mathbf{Y}) = \frac{1}{2} \sum_{h,w} \|\mathbf{Y}_{h,w} - \hat{\mathbf{Y}}_{h,w}\|_2^2$$



$$\hat{\mathbf{Z}} \in [0, 1]^{H \times W \times Q}$$

$$\mathbf{Z} = \mathcal{H}_{gt}^{-1}(\mathbf{Y})$$

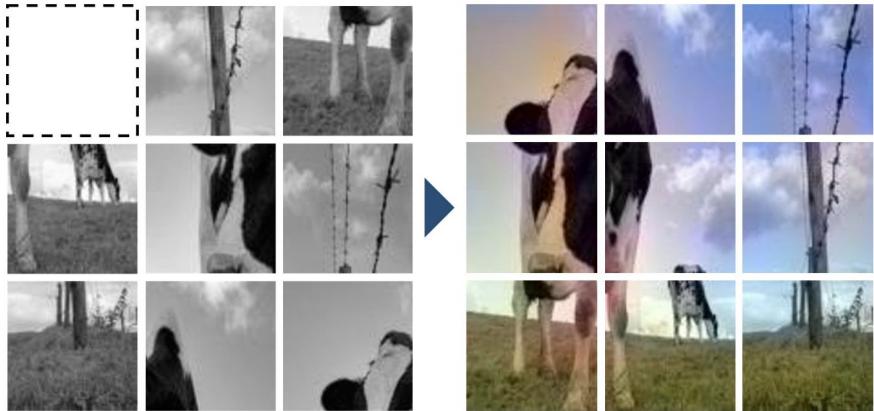
$$L_{cl}(\hat{\mathbf{Z}}, \mathbf{Z}) = - \sum_{h,w} v(\mathbf{Z}_{h,w}) \sum_q \mathbf{Z}_{h,w,q} \log(\hat{\mathbf{Z}}_{h,w,q})$$

$$\mathcal{H}(\mathbf{Z}_{h,w}) = \mathbb{E}[f_T(\mathbf{Z}_{h,w})], \quad f_T(\mathbf{z}) = \frac{\exp(\log(\mathbf{z})/T)}{\sum_q \exp(\log(\mathbf{z}_q)/T)}$$

# Colorization

fine-tune layers	[Ref]	Class. (%mAP)			Det. (%mAP)		Seg. (%mIU)	
		fc8	fc6-8	all	[Ref]	all	[Ref]	all
ImageNet [38]	-	76.8	78.9	79.9	[36]	56.8	[42]	48.0
Gaussian	[10]	—	—	53.3	[10]	43.4	[10]	19.8
Autoencoder	[16]	24.8	16.0	53.8	[10]	41.9	[10]	25.2
k-means [36]	[16]	32.0	39.2	56.6	[36]	45.6	[16]	32.6
Agrawal et al. [8]	[16]	31.2	31.0	54.2	[36]	43.9	—	—
Wang & Gupta [15]	—	28.1	52.2	58.7	[36]	47.4	—	—
*Doersch et al. [14]	[16]	44.7	55.1	<b>65.3</b>	[36]	<b>51.1</b>	—	—
*Pathak et al. [10]	[10]	—	—	56.5	[10]	44.5	[10]	29.7
*Donahue et al. [16]	—	38.2	50.2	58.6	[16]	46.2	[16]	34.9
Ours (gray)	—	<b>52.4</b>	<b>61.5</b>	<b>65.9</b>	—	46.1	—	35.0
Ours (color)	—	<b>52.4</b>	<b>61.5</b>	<b>65.6</b>	—	46.9	—	<b>35.6</b>

# Damaged Jigsaw Puzzles, 2018



Method	Class.	Det.	Segm.
ImageNet [20]	79.9	56.8	48.0
Random	53.3	43.4	19.8
RelativePosition [4]	65.3	51.1	-
Jigsaw [25]	67.6	<b>53.2</b>	37.6
Ego-motion [36]	54.2	43.9	-
Adversarial [6]	58.6	46.2	34.9
Inpainting [30]	56.5	44.5	29.7
Colorization [38]	65.9	46.9	35.6
Split-Brain [39]	67.1	46.7	36.0
ColorProxy [21]	65.9	-	<u>38.4</u>
WatchingObjectMove [29]	61.0	52.2	-
Counting [26]	<u>67.7</u>	51.4	36.6
CDJP	<b>69.2</b>	<u>52.4</u>	<b>39.3</b>

# Rotations, 2018, Gidaris et. al.



90° rotation



270° rotation



180° rotation

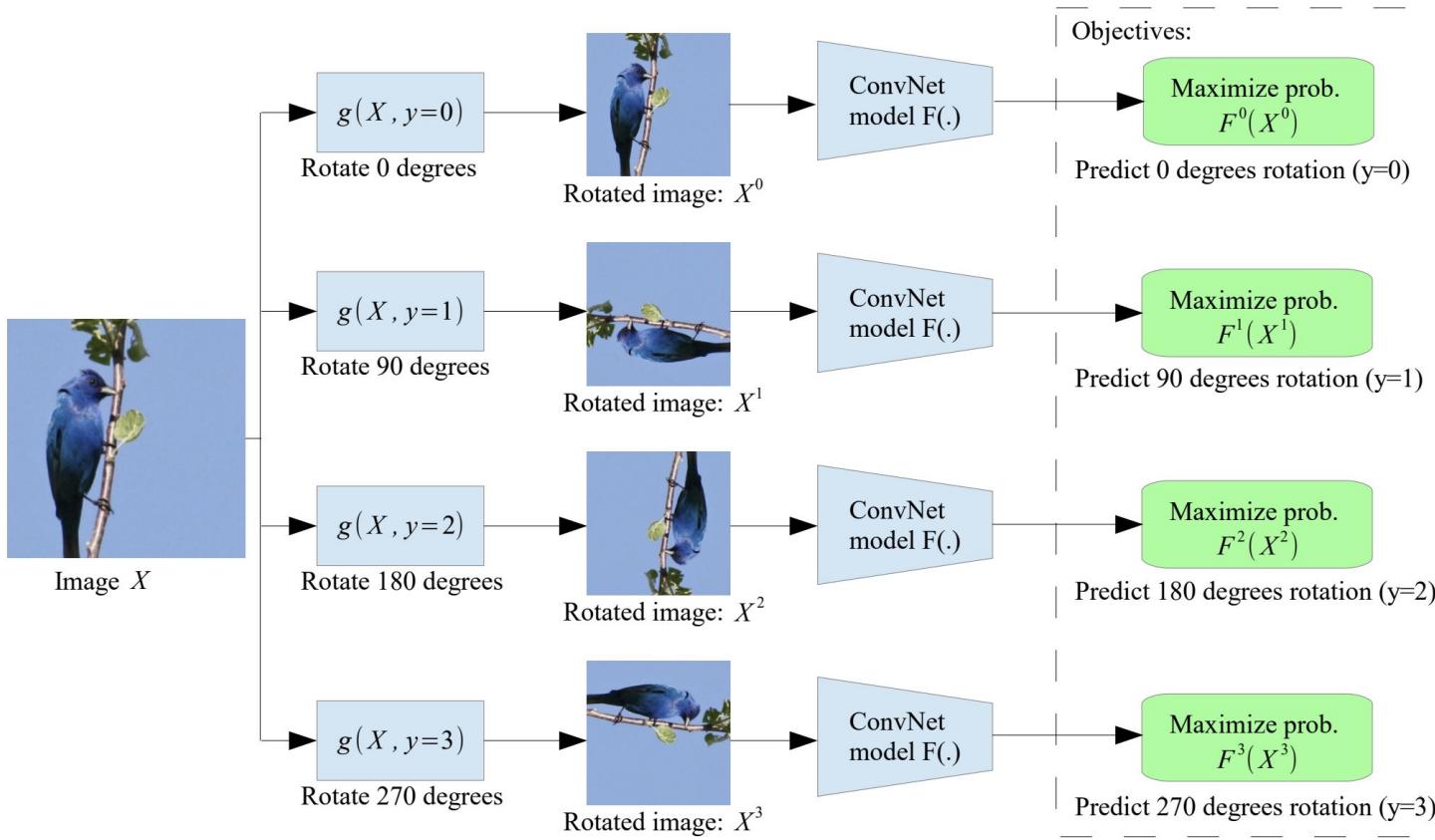


0° rotation



270° rotation

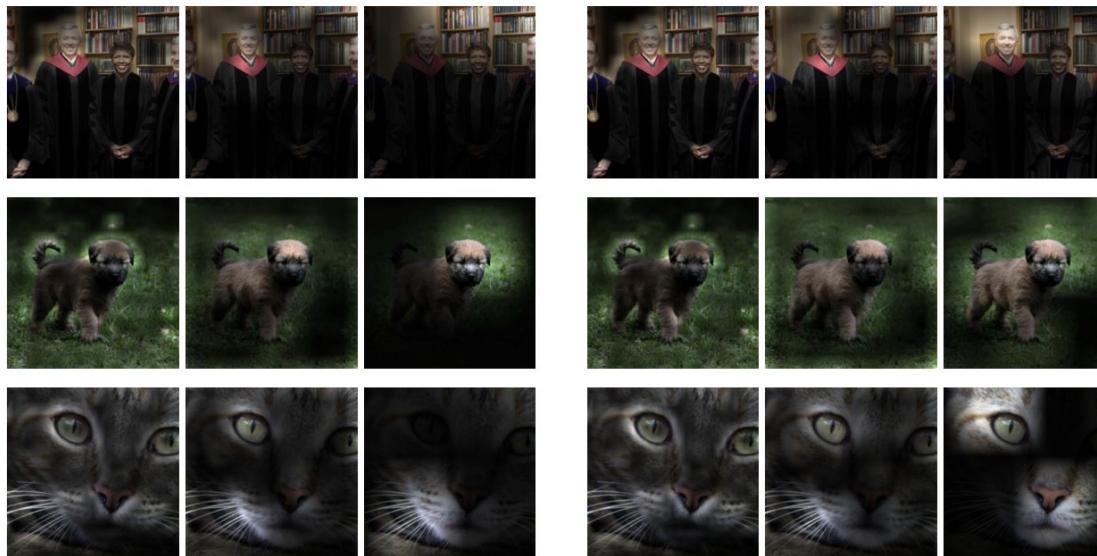
# Rotations, 2018, Gidaris et. al.



# Rotations



Input images on the models



Conv1  $27 \times 27$  Conv3  $13 \times 13$  Conv5  $6 \times 6$

Conv1  $27 \times 27$  Conv3  $13 \times 13$  Conv5  $6 \times 6$

(a) Attention maps of supervised model

(b) Attention maps of our self-supervised model

# Rotations

Model	ConvB1	ConvB2	ConvB3	ConvB4	ConvB5
RotNet with 3 conv. blocks	85.45	88.26	62.09	-	-
RotNet with 4 conv. blocks	85.07	89.06	86.21	61.73	-
RotNet with 5 conv. blocks	85.04	<b>89.76</b>	86.82	74.50	50.37

# Rotations

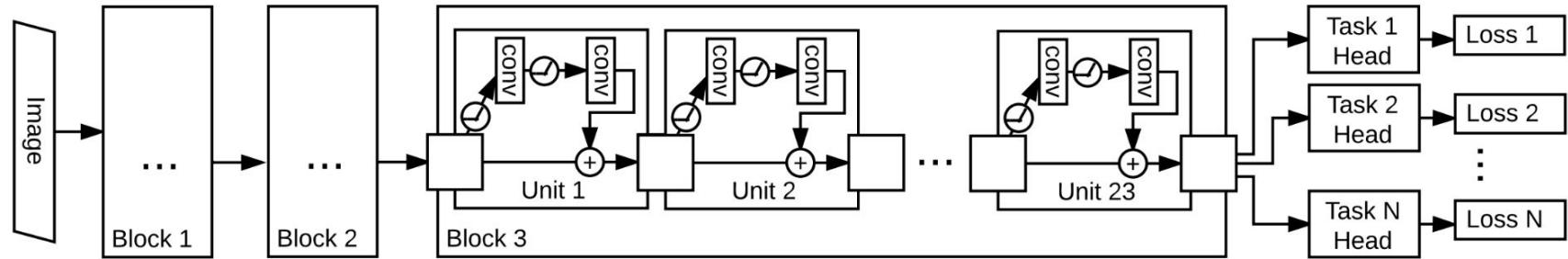
Model	ConvB1	ConvB2	ConvB3	ConvB4	ConvB5
RotNet with 3 conv. blocks	85.45	88.26	62.09	-	-
RotNet with 4 conv. blocks	85.07	89.06	86.21	61.73	-
RotNet with 5 conv. blocks	85.04	<b>89.76</b>	86.82	74.50	50.37
Method	Conv1	Conv2	Conv3	Conv4	Conv5
ImageNet labels	19.3	36.3	44.2	48.3	50.5
Random	11.6	17.1	16.9	16.3	14.1
Random rescaled Krähenbühl et al. (2015)	17.5	23.0	24.5	23.2	20.6
Context (Doersch et al., 2015)	16.2	23.3	30.2	31.7	29.6
Context Encoders (Pathak et al., 2016b)	14.1	20.7	21.0	19.8	15.5
Colorization (Zhang et al., 2016a)	12.5	24.5	30.4	31.5	30.3
Jigsaw Puzzles (Noroozi & Favaro, 2016)	18.2	28.8	34.0	33.9	27.1
BIGAN (Donahue et al., 2016)	17.7	24.5	31.0	29.9	28.0
Split-Brain (Zhang et al., 2016b)	17.7	29.3	35.4	35.2	32.8
Counting (Noroozi et al., 2017)	18.0	30.6	34.3	32.5	25.7
(Ours) RotNet	<b>18.8</b>	<b>31.7</b>	<b>38.7</b>	<b>38.2</b>	<b>36.5</b>

# Rotations

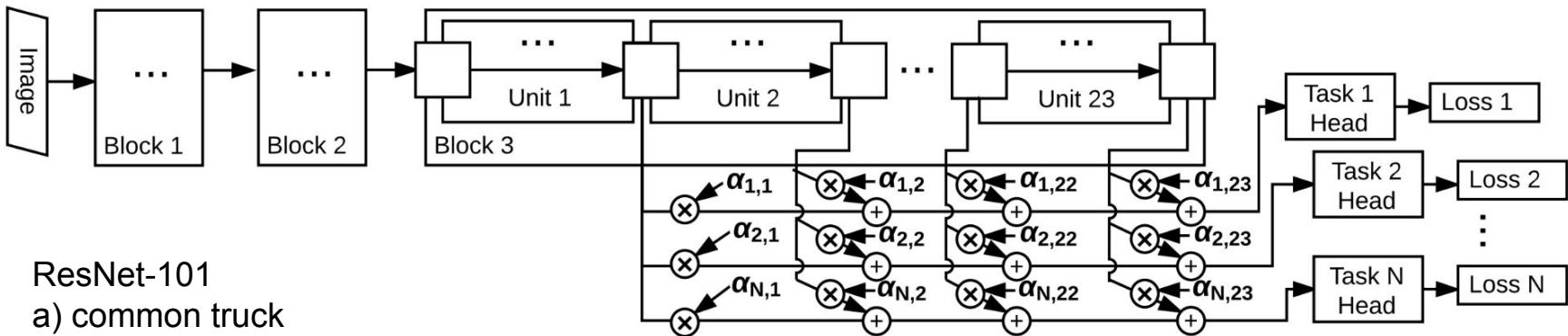
	Classification (%mAP)	Detection (%mAP)	Segmentation (%mIoU)
Trained layers	fc6-8 all	all	all
ImageNet labels	78.9	79.9	56.8 48.0
Random		53.3	43.4 19.8
Random rescaled Krähenbühl et al. (2015)	39.2	56.6	45.6 32.6
Egomotion (Agrawal et al., 2015)	31.0	54.2	43.9
Context Encoders (Pathak et al., 2016b)	34.6	56.5	44.5 29.7
Tracking (Wang & Gupta, 2015)	55.6	63.1	47.4
Context (Doersch et al., 2015)	55.1	65.3	51.1
Colorization (Zhang et al., 2016a)	61.5	65.6	46.9 35.6
BIGAN (Donahue et al., 2016)	52.3	60.1	46.9 34.9
Jigsaw Puzzles (Noroozi & Favaro, 2016)	-	67.6	53.2 37.6
NAT (Bojanowski & Joulin, 2017)	56.7	65.3	49.4
Split-Brain (Zhang et al., 2016b)	63.0	67.1	46.7 36.0
ColorProxy (Larsson et al., 2017)		65.9	38.4
Counting (Noroozi et al., 2017)	-	67.7	51.4 36.6
(Ours) RotNet	<b>70.87</b>	<b>72.97</b>	<b>54.4</b> <b>39.1</b>

# Task Ensemble, 2017, Doersch et al.

a)



b)



ResNet-101

a) common truck

b) lasso architecture

# Task Ensemble

Net structure	ImageNet	PASCAL	NYU
No Lasso	69.30	<b>70.53</b>	79.25
Eval Only Lasso	<b>70.18</b>	68.86	79.41
Pre-train Only Lasso	68.09	68.49	78.96
Pre-train & Eval Lasso	69.44	68.98	79.45

Pre-training	ImageNet	PASCAL	NYU
RP	59.21	66.75	<b>80.54</b>
RP+Col	66.64	68.75	79.87
RP+Ex	65.24	69.44	78.70
RP+MS	63.73	68.81	78.72
RP+Col+Ex	68.65	69.48	80.17
RP+Col+Ex+MS	<b>69.30</b>	<b>70.53</b>	79.25
INet Labels	85.10	74.17	80.06