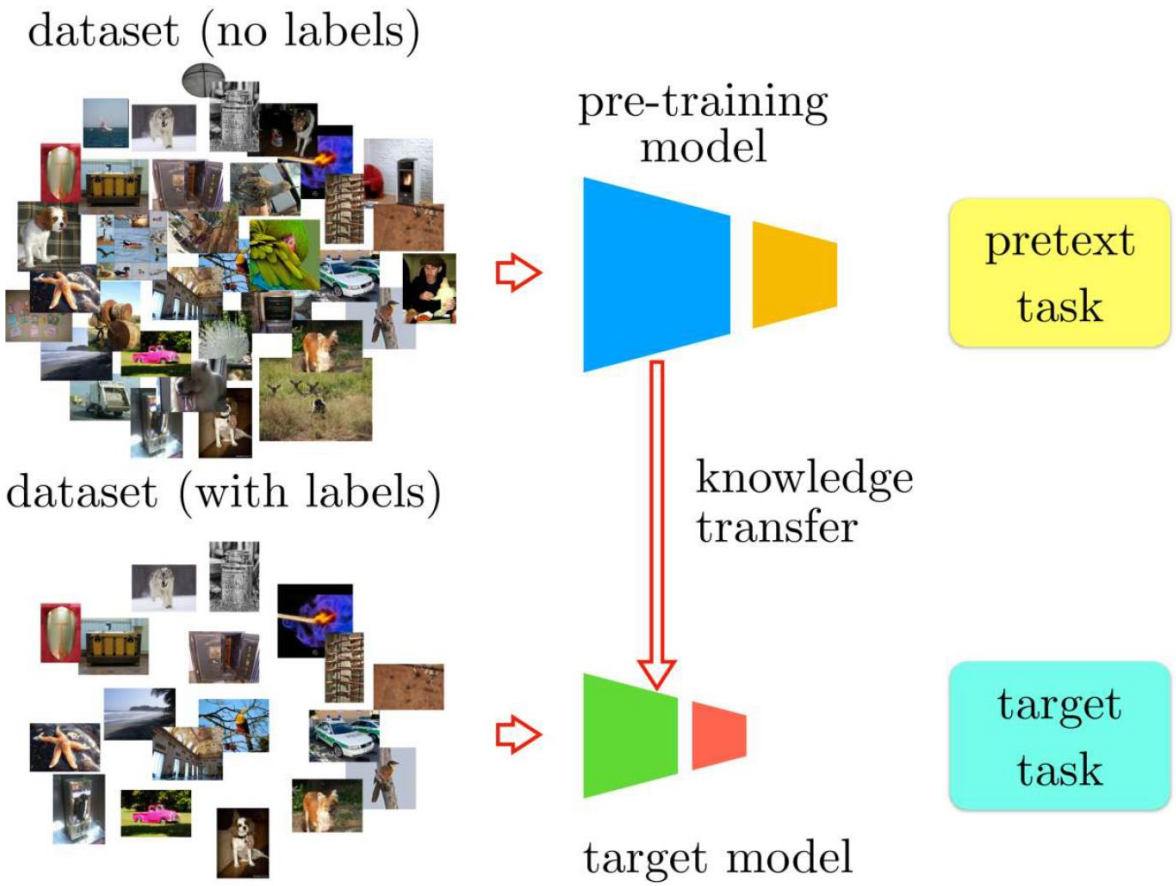
自监督视觉学习

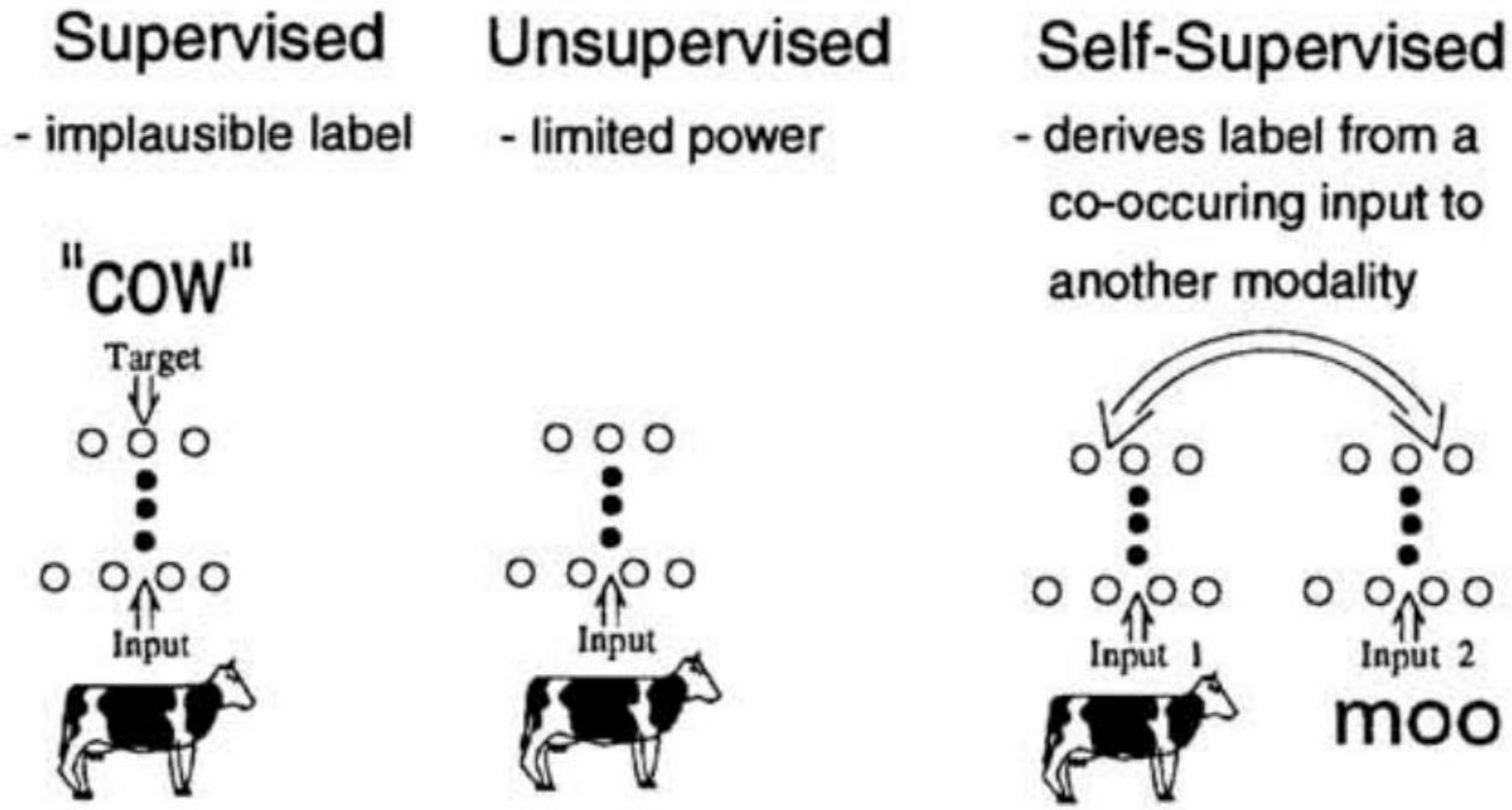
Wangmeng Zuo

Centre on Machine Learning Research
Harbin Institute of Technology

Self-Supervised Learning



Self-Supervised Learning in Vision

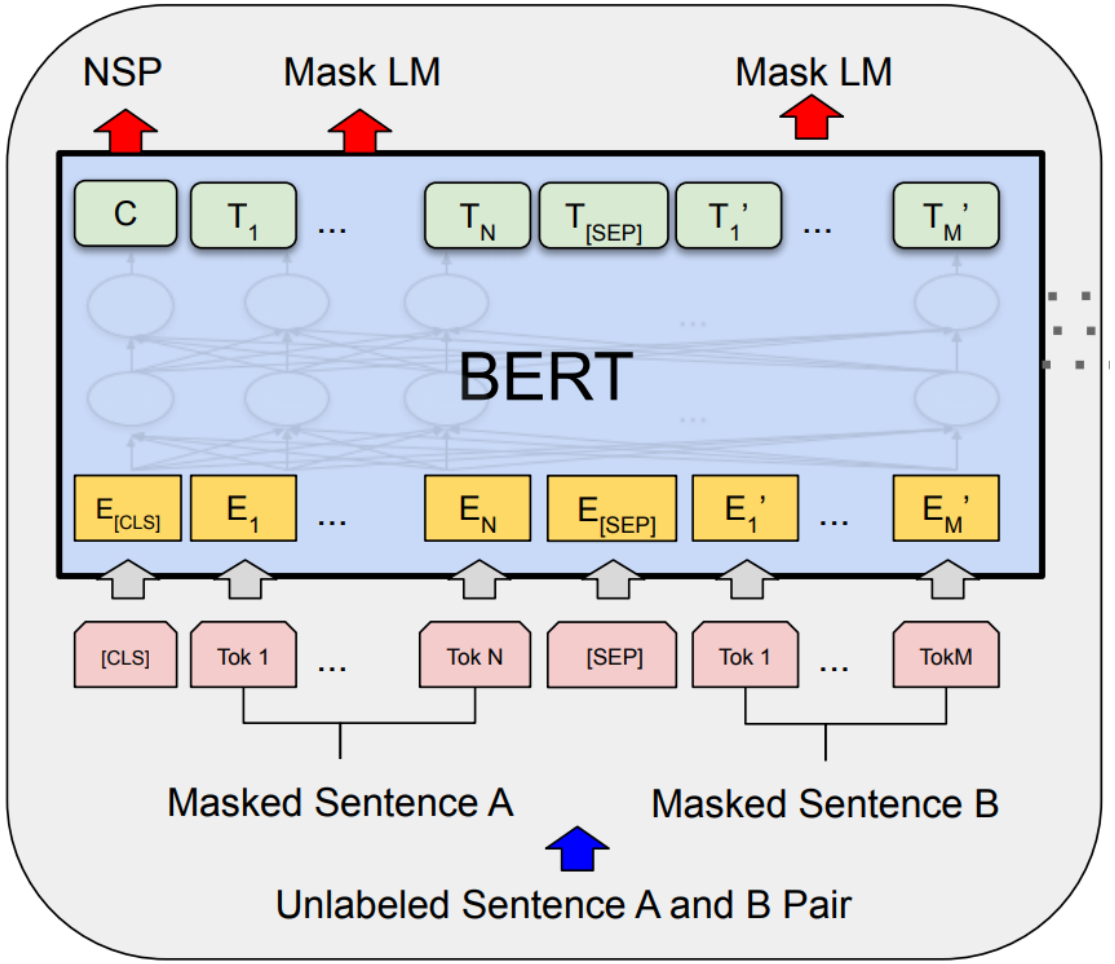


Virginia de Sa, 1994, Image: Learning classification with Unlabeled Data

BERT for Language Understanding (Devlin et al., Arxiv2018)

- Citations: 61701 (2018.10 —)

- Randomly masks some of the tokens from the input, and the objective is to predict the original vocabulary id of the masked word based only on its context.



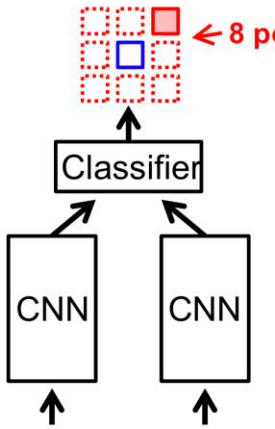
Pre-training

Content

- Self-supervised Vision Learning
  - 1. Self-supervised Contextual Modeling
  - 2. Contrastive learning-based self-supervised learning
  - 3. Returning of Self-supervised Contextual Modeling

Deep Context Modeling

- Jigsaw Puzzles
- Image Rotation
- Image Colorization
- Image Inpainting



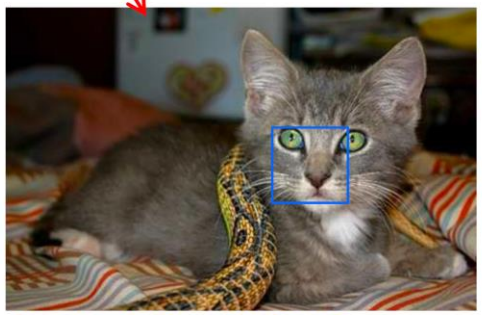
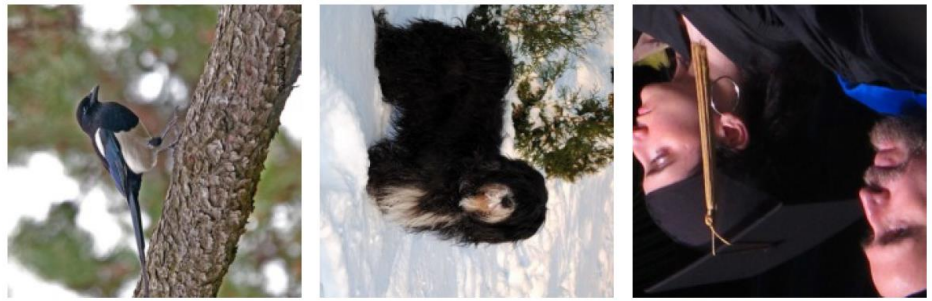
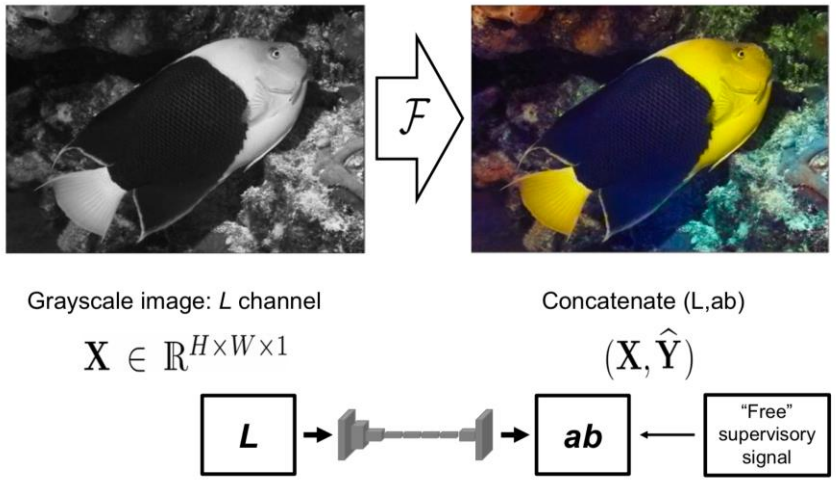
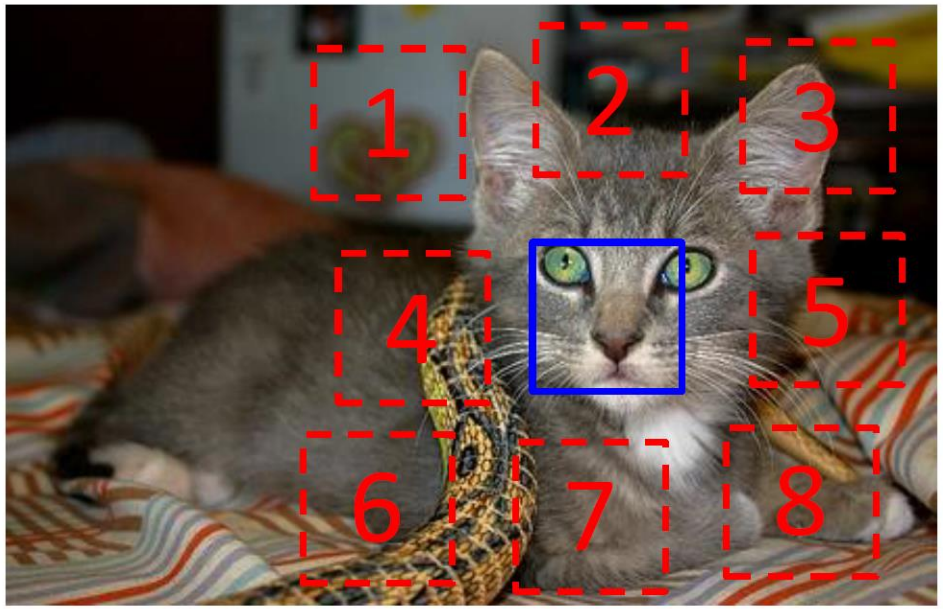


Image Rotation

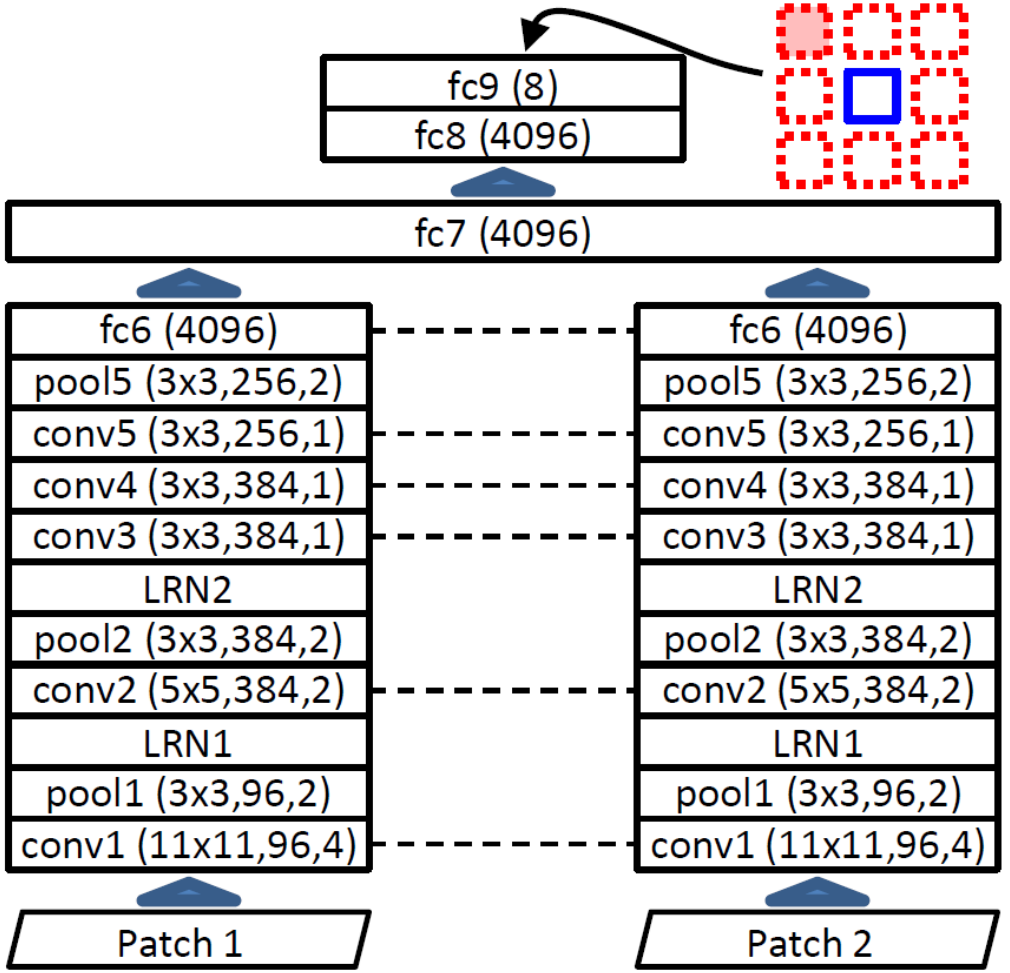




Jigsaw Puzzles (1)



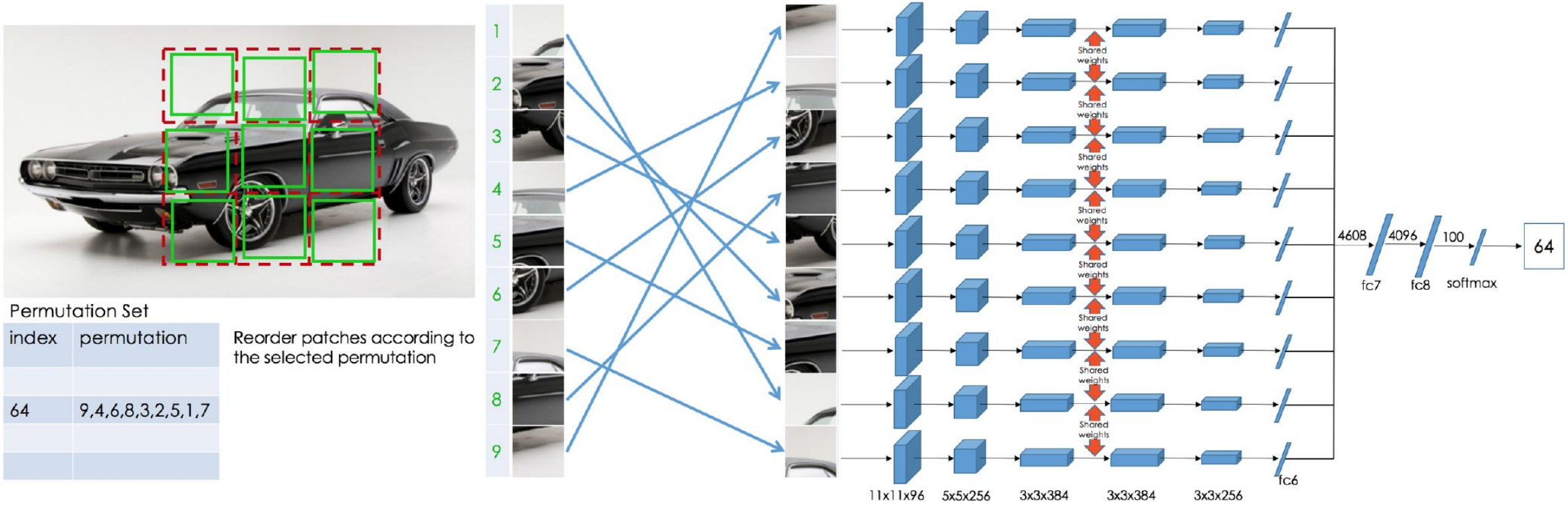
X = (,); Y = 3



C. Doersch, A. Gupta, and A. A. Efros. Unsupervised Visual Representation Learning by Context Prediction. In ICCV 2015.

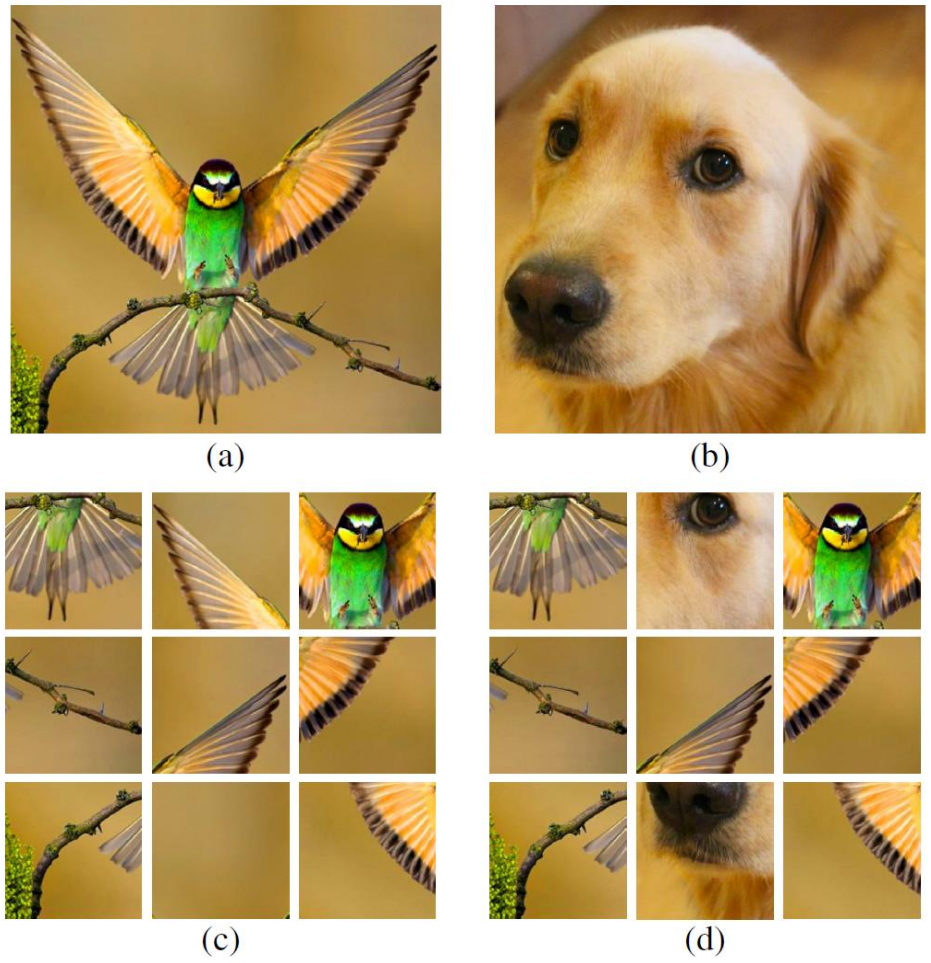
Jigsaw Puzzles (2)

- Permutation set: 1000, Average Hamming distance: 8.00



M. Noroozi and P. Favaro, Unsupervised Learning of Visual Representations by Solving Jigsaw Puzzles, ECCV 2016.

Jigsaw++



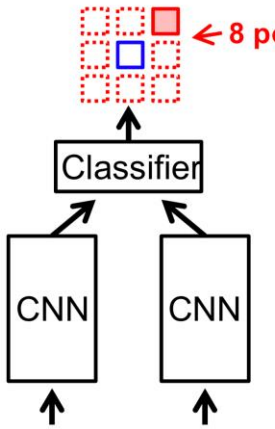
Method	Ref	Class.	Det.		Segm.
			SS	MS	
Supervised [17]		79.9	59.1	59.8	48.0
CC+HOG [6]		70.2	53.2	53.5	39.2
Random	[27]	53.3	43.4	-	19.8
ego-motion [1]	[1]	54.2	43.9	-	-
BiGAN [9]	[9]	58.6	46.2	-	34.9
ContextEncoder [27]	[27]	56.5	44.5	-	29.7
Video [32]	[16]	63.1	47.2	-	-
Colorization [39]	[39]	65.9	46.9	-	35.6
Split-Brain [40]	[40]	67.1	46.7	-	36.0
Context [7]	[16]	55.3	46.6	-	-
Context [7]*	[16]	65.3	51.1	-	-
Counting [23]	[23]	67.7	51.4	-	36.6
WatchingObjectsMove [26]	[26]	61.0	-	52.2	-
Jigsaw [21]	[21]	67.7	53.2	-	-
Jigsaw++		69.8	55.5	55.7	38.1
CC+Context-ColorDrop [7]		67.9	52.8	53.4	-
CC+Context-ColorProjection [7]		66.7	51.5	51.8	-
CC+Jigsaw++		69.9	55.0	55.8	40.0
[37]+vgg-Jigsaw++		70.6	54.8	55.2	38.0
CC+vgg-Context [7]		68.0	53.0	53.5	-
CC+vgg-Jigsaw++		**72.5**	**56.5**	**57.2**	**42.6**

VOC 2007

Noroozi et al., Boosting Self-Supervised Learning via Knowledge Transfer, CVPR 2018.

Deep Context Modeling

- Jigsaw Puzzles
- Image Rotation
- Image Colorization
- Image Inpainting



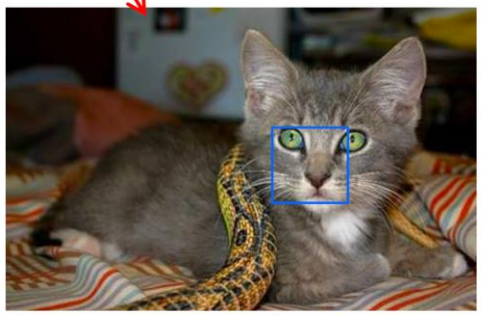
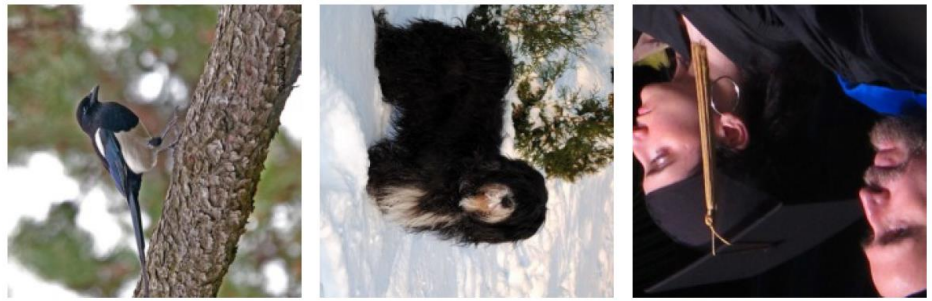
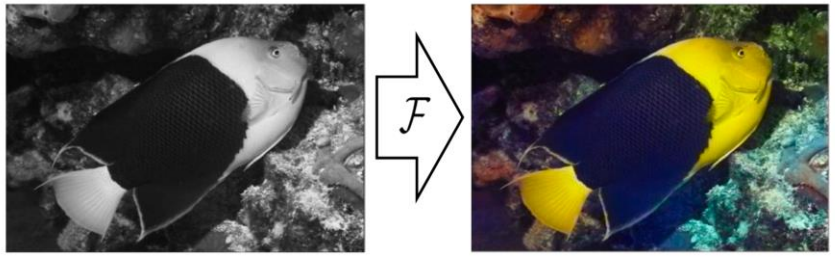
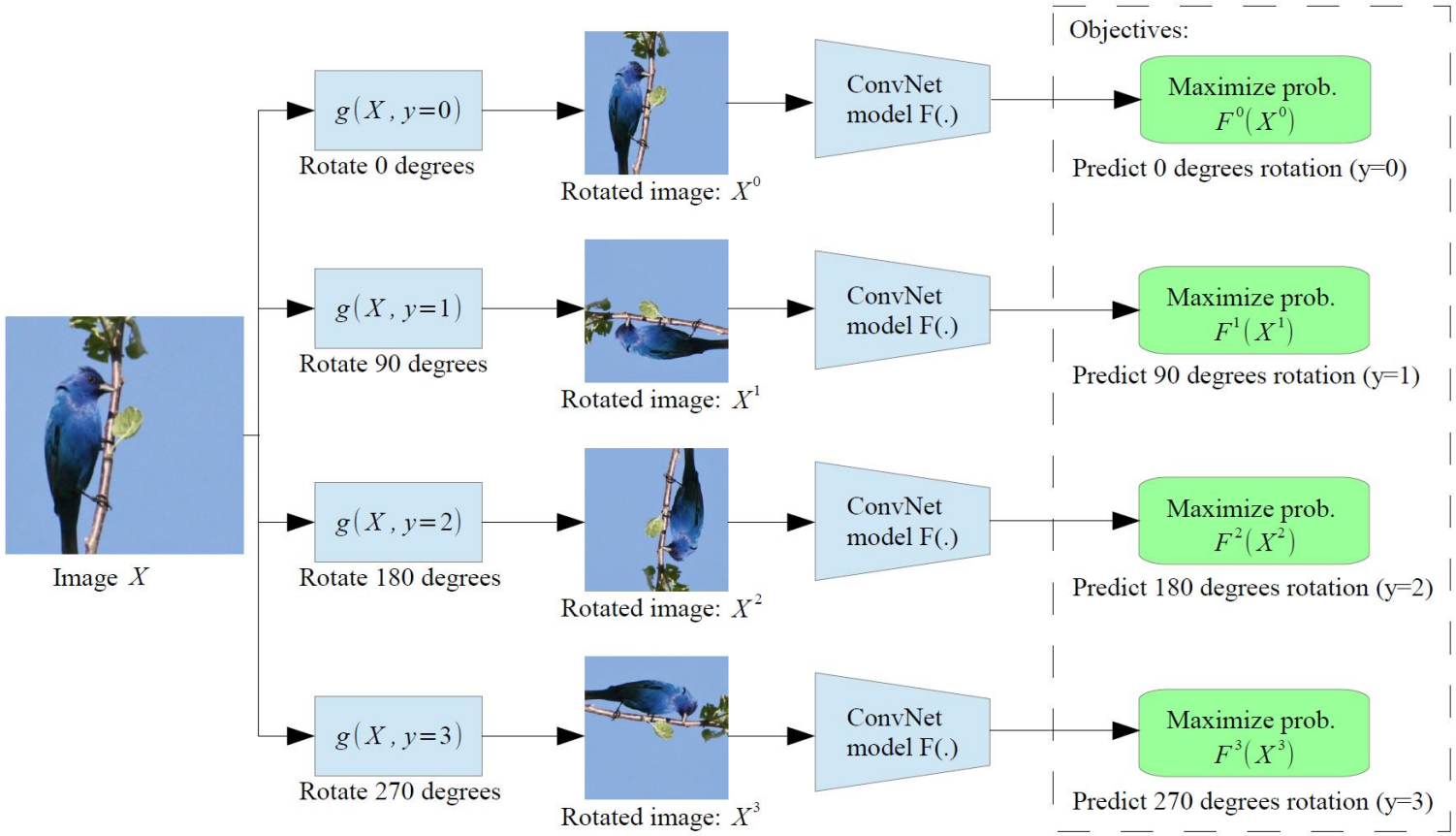


Image Rotation





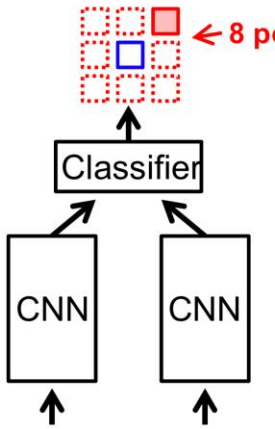
Predicting Image Rotations



Spyros Gidaris, Praveer Singh, Nikos Komodakis, Unsupervised Representation Learning by Predicting Image Rotations, ICLR 2018.

Deep Context Modeling

- Jigsaw Puzzles
- Image Rotation
- Image Colorization
- Image Inpainting



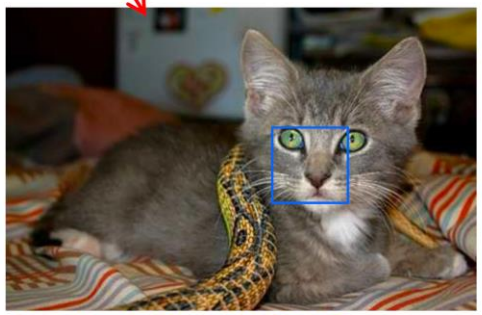
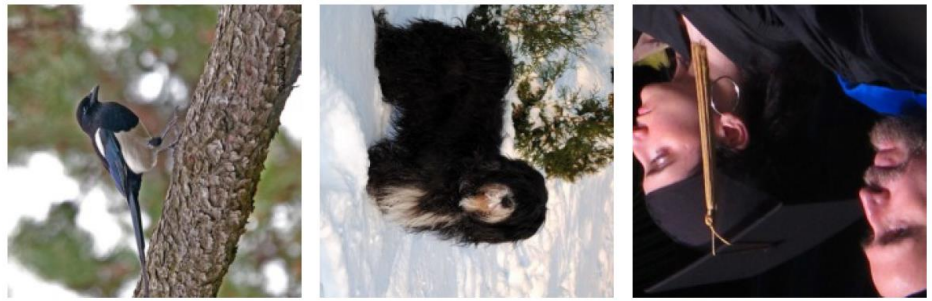


Image Rotation



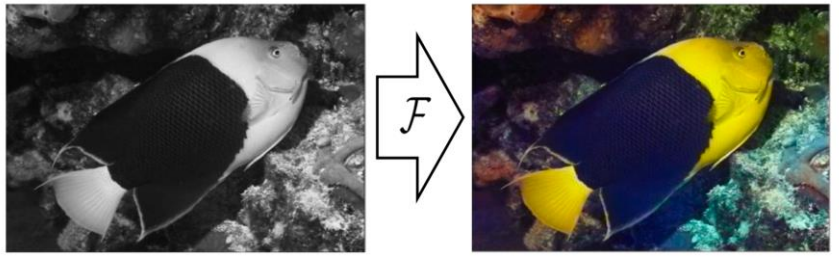
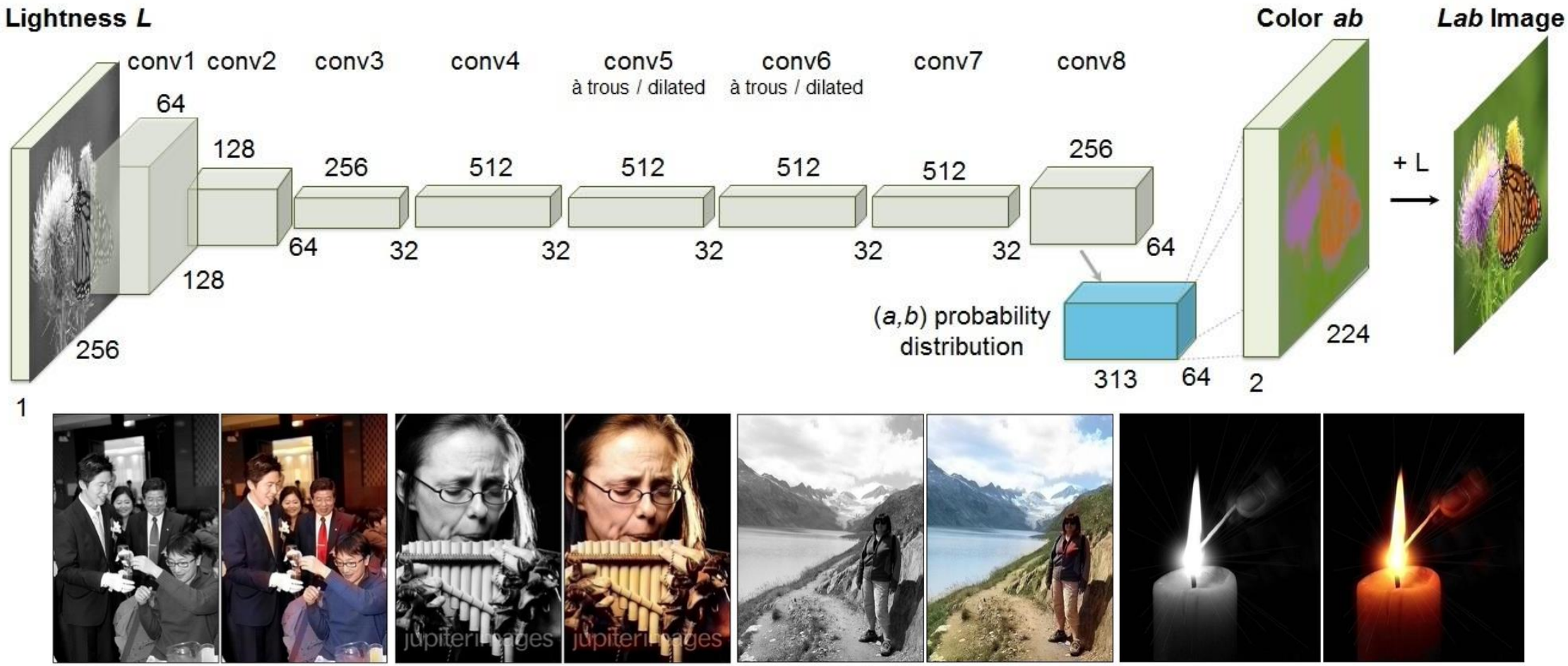
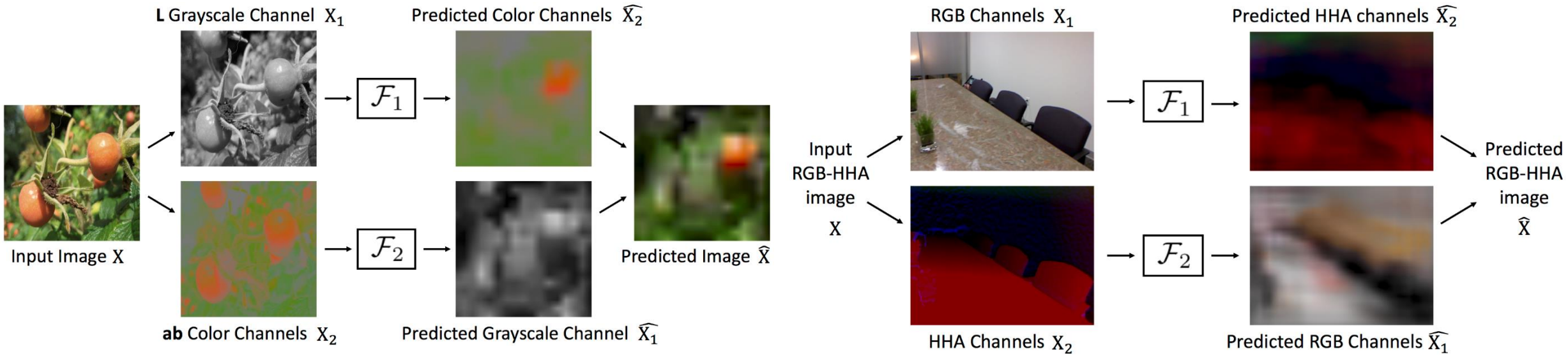


Image Colorization



R. Zhang, P. Isola, and A.A. Efros. Colorful image colorization, ECCV 2016

Split-Brain Autoencoders



R. Zhang, P. Isola, A. A. Efros, Split-Brain Autoencoders: Unsupervised Learning by Cross-Channel Prediction, CVPR 2017.

Deep Context Modeling

- Jigsaw Puzzles
- Image Rotation
- Image Colorization
- Image Inpainting

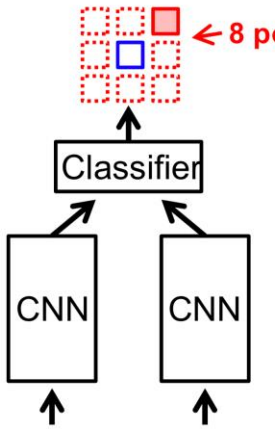
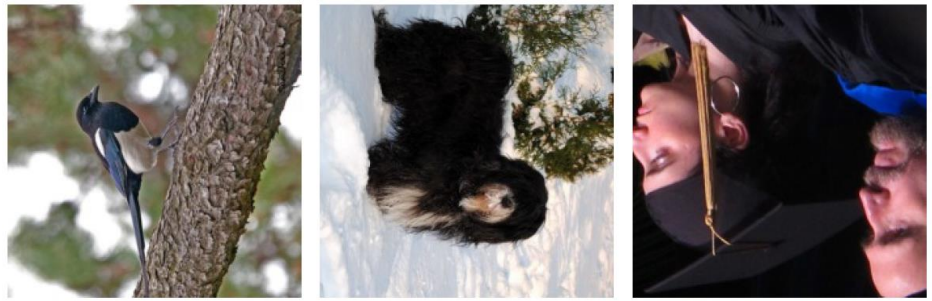
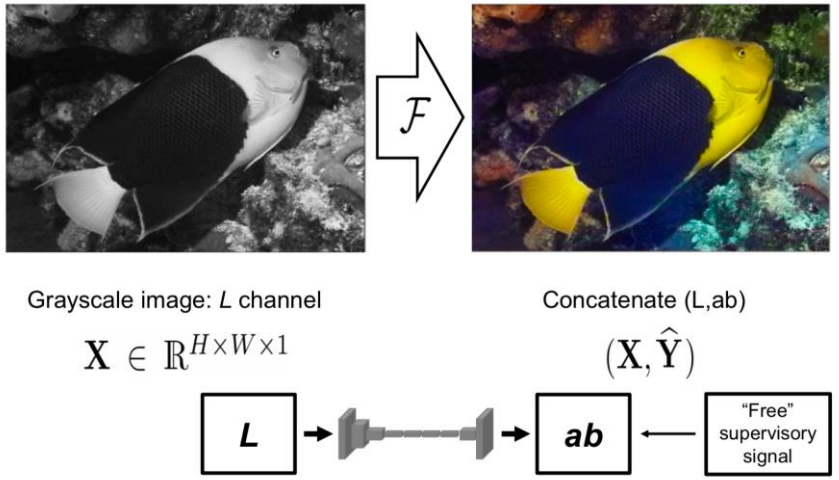
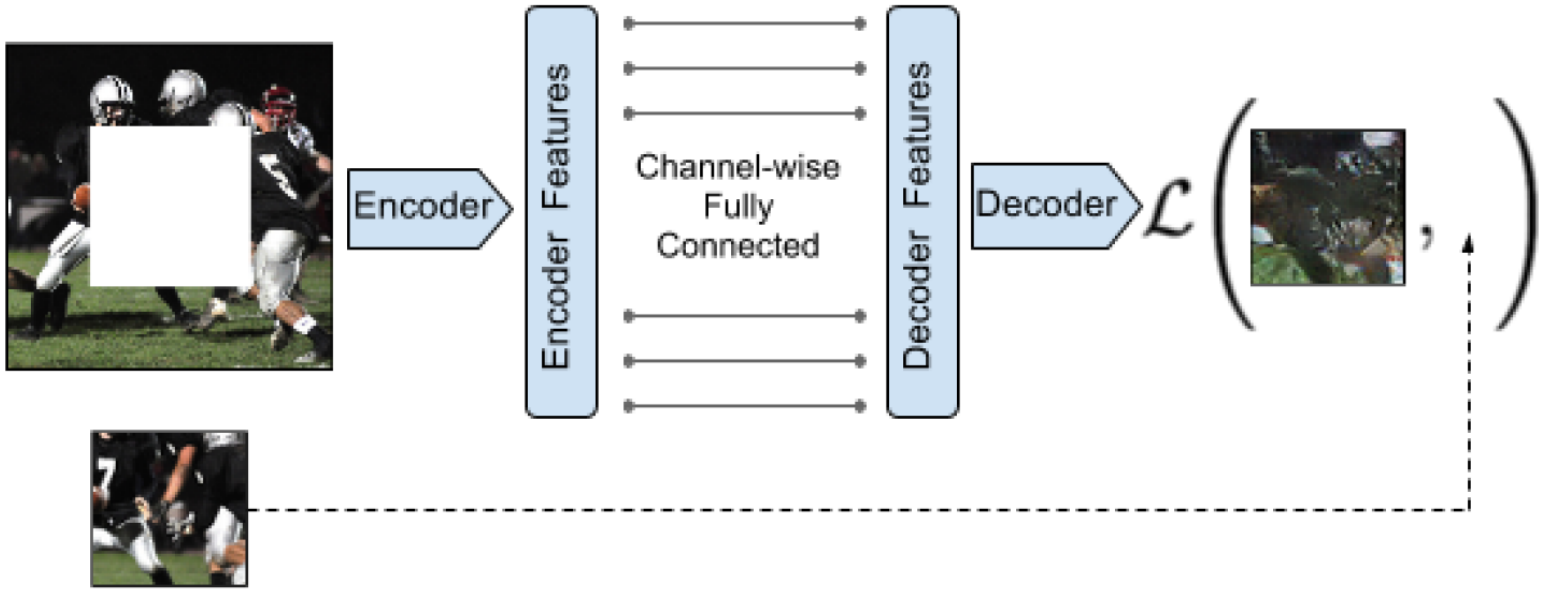


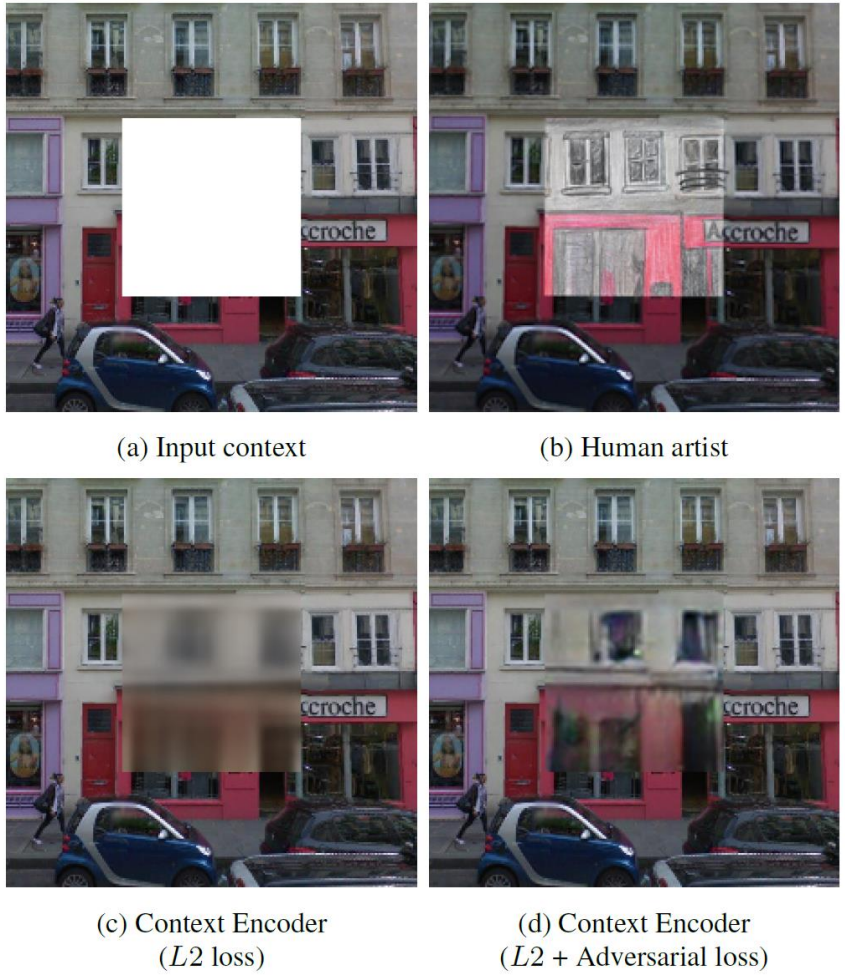
Image Rotation





Context Encoders



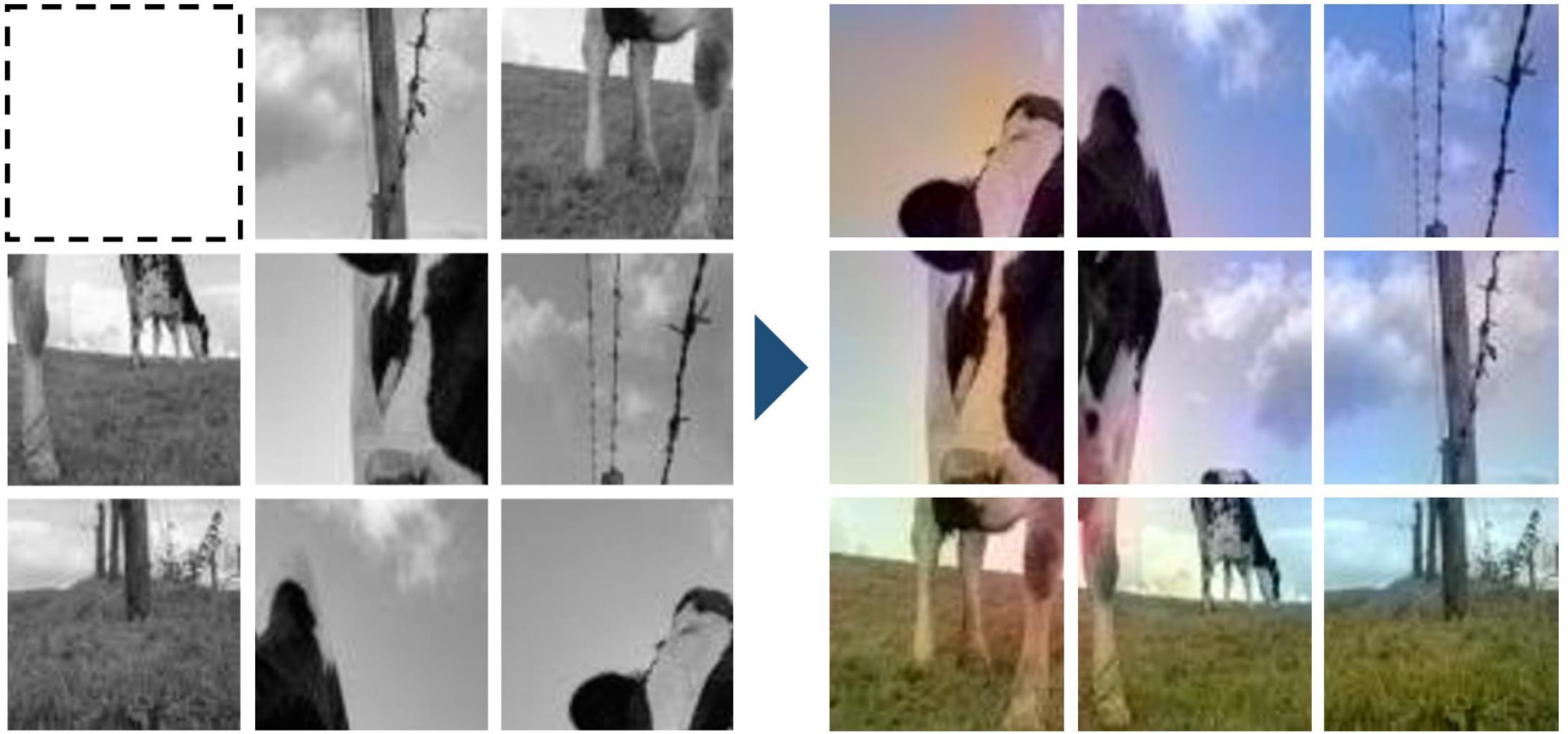


(a) Input context (b) Human artist

(c) Context Encoder ($L2$ loss) (d) Context Encoder ($L2$ + Adversarial loss)

D. Pathak, P. Krahenbuhl, J. Donahue, T. Darrell, A. A. Efros, Context Encoders: Feature Learning by Inpainting, CVPR 2016

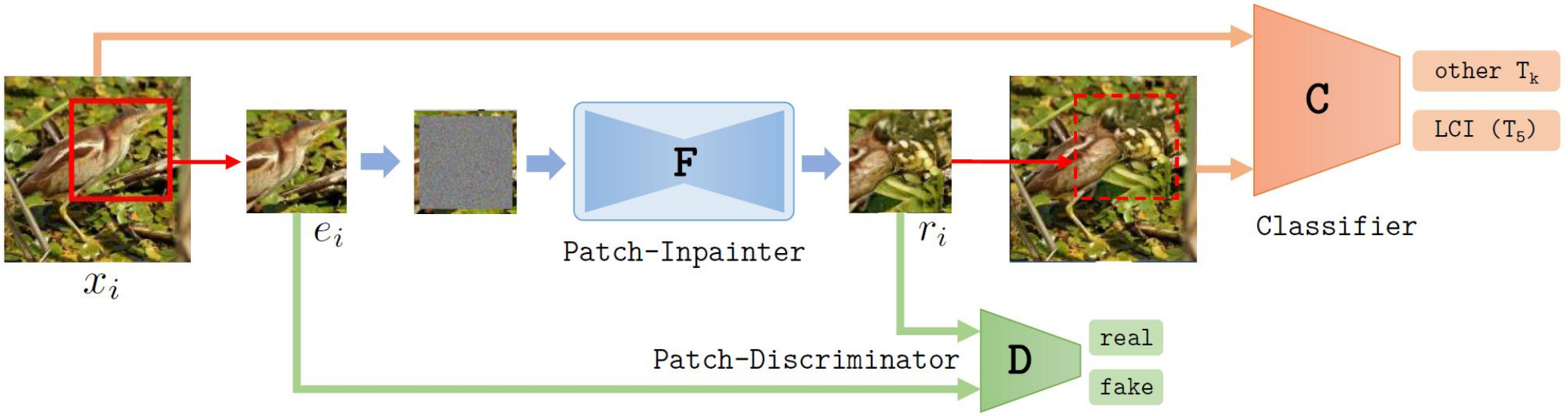
Jigsaw+Colorization+Inpainting



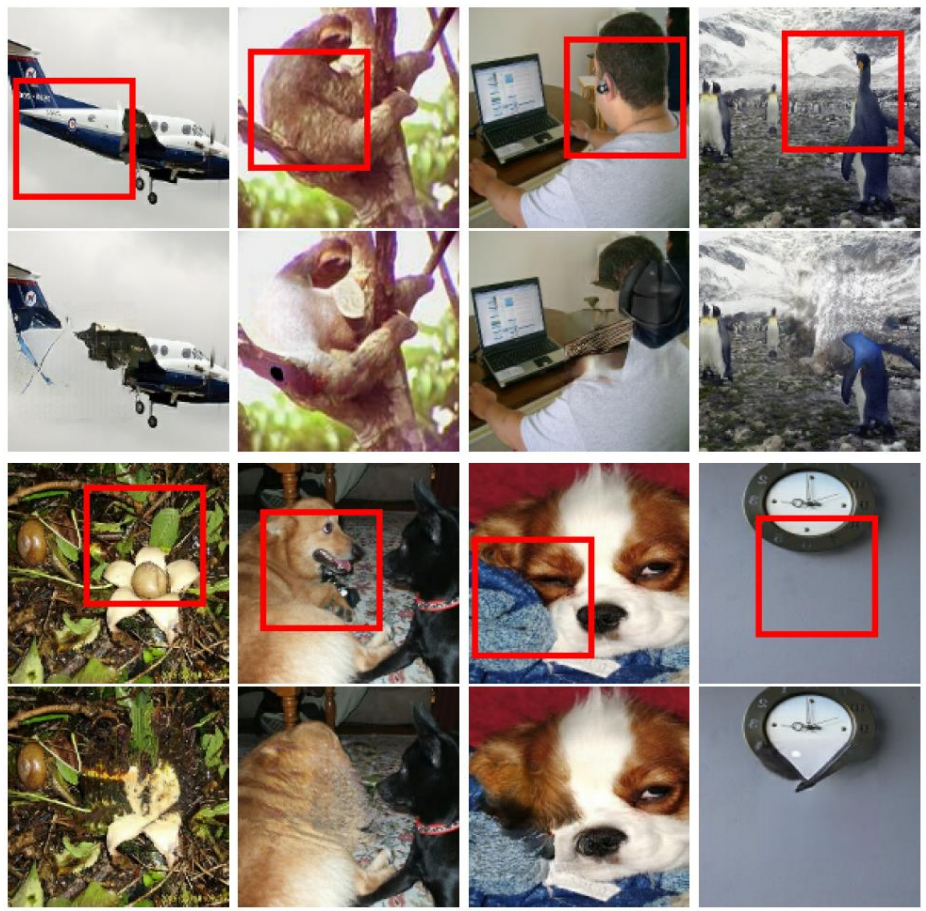
Learning Image Representations by Completing Damaged Jigsaw Puzzles, WACV 2018.

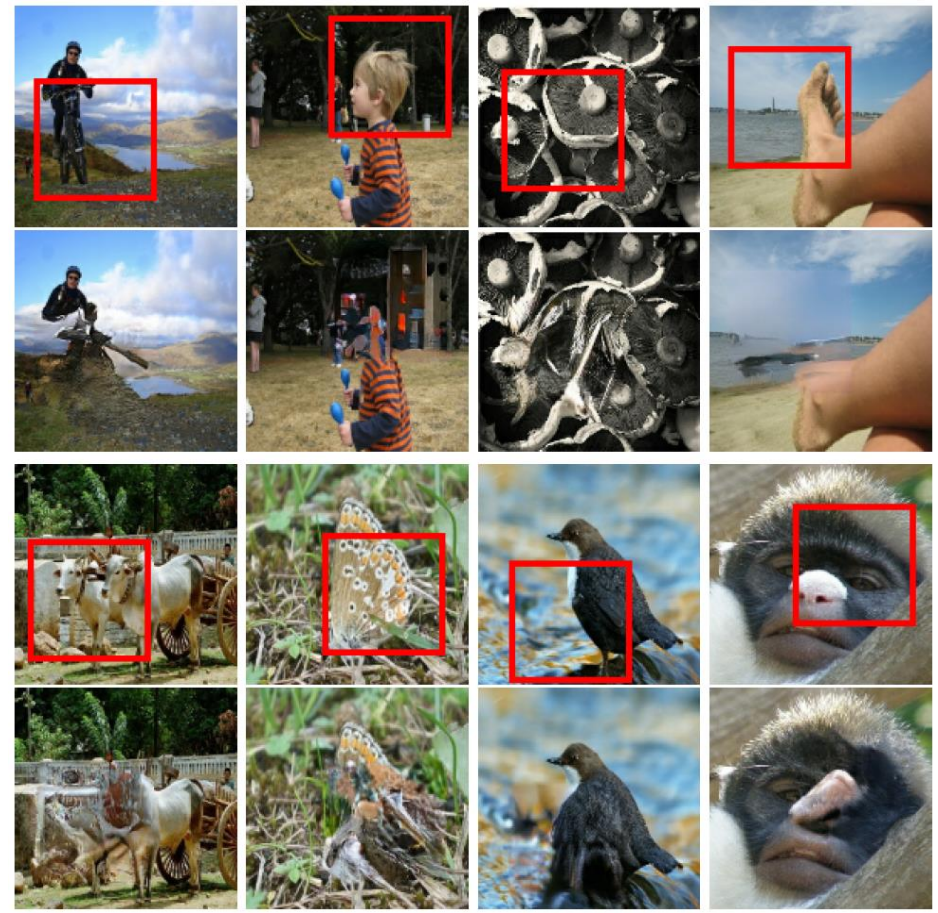
Inpainting+Rotation

- Limited context inpainting

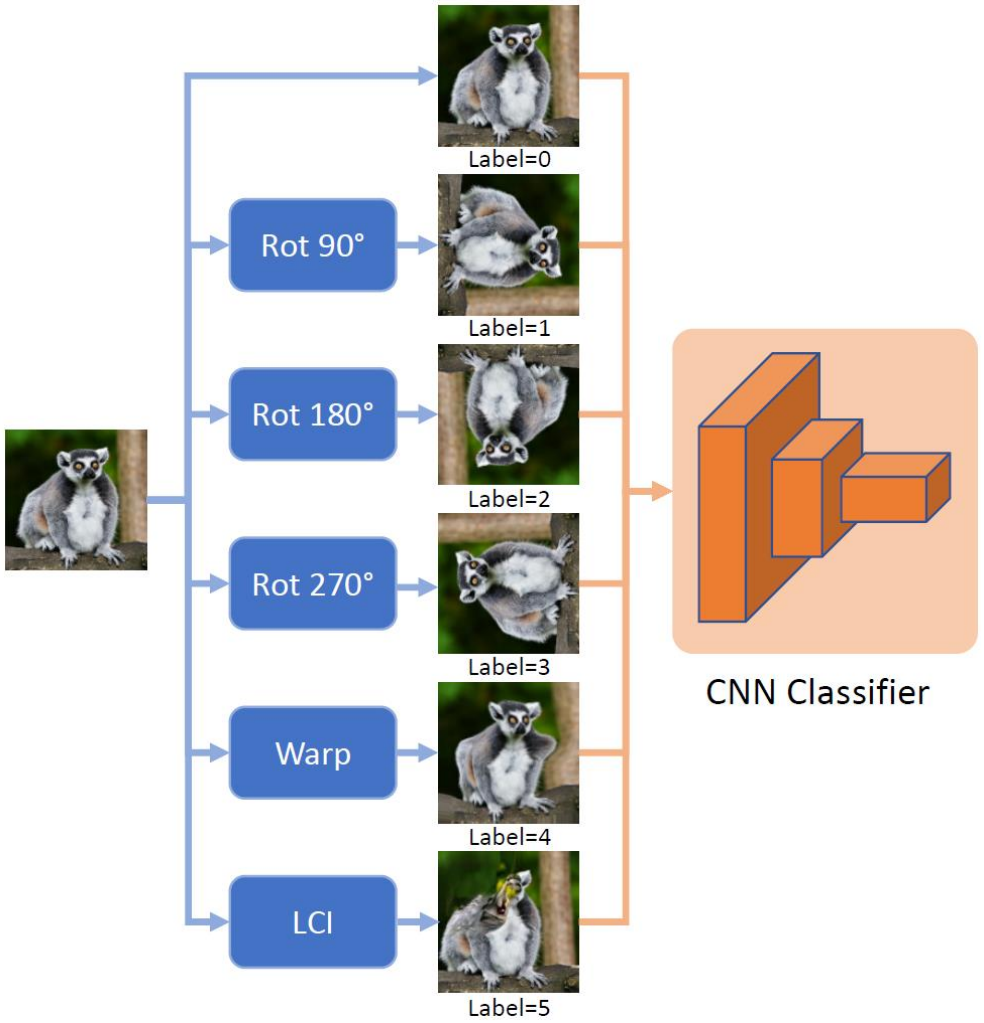


Simon Jenni, Hailin Jin, Paolo Favaro , Steering Self-Supervised Feature Learning Beyond Local Pixel Statistics, CVPR 2020.





Learning Global Statistics



Model\Layer	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%
Donahue *et al.* [10]	17.7%	24.5%	31.0%	29.9%	28.0%
Feng *et al.* [15]	19.3%	33.3%	**40.8%**	41.8%	**44.3%**
Gidaris *et al.* [18]	18.8%	31.7%	38.7%	38.2%	36.5%
Huang *et al.* [25]	15.6%	27.0%	35.9%	39.7%	37.9%
Jenni & Favaro [29]	19.5%	33.3%	37.9%	38.9%	34.9%
Noroozi & Favaro [46]	18.2%	28.8%	34.0%	33.9%	27.1%
Noroozi *et al.* [47]	18.0%	30.6%	34.3%	32.5%	25.7%
Noroozi *et al.* [48]	19.2%	32.0%	37.3%	37.1%	34.6%
Tian *et al.* [57]	18.4%	33.5%	38.1%	40.4%	42.6%
Wu *et al.* [61]	16.8%	26.5%	31.8%	34.1%	35.6%
Zhang *et al.* [65]	13.1%	24.8%	31.0%	32.6%	31.8%
Zhang *et al.* [66]	17.7%	29.3%	35.4%	35.2%	32.8%
Zhang *et al.* [64]	19.2%	32.8%	40.6%	39.7%	37.7%
Doersch *et al.* [8]*	16.2%	23.3%	30.2%	31.7%	29.6%
Caron *et al.* [4]*	12.9%	29.2%	38.2%	39.8%	36.1%
Zhuang *et al.* [71]*†	18.7%	32.7%	38.1%	42.3%	42.4%
Ours	**20.8%**	**34.5%**	40.2%	**43.1%**	41.4%
Ours†	22.0%	36.4%	42.4%	45.4%	44.4%

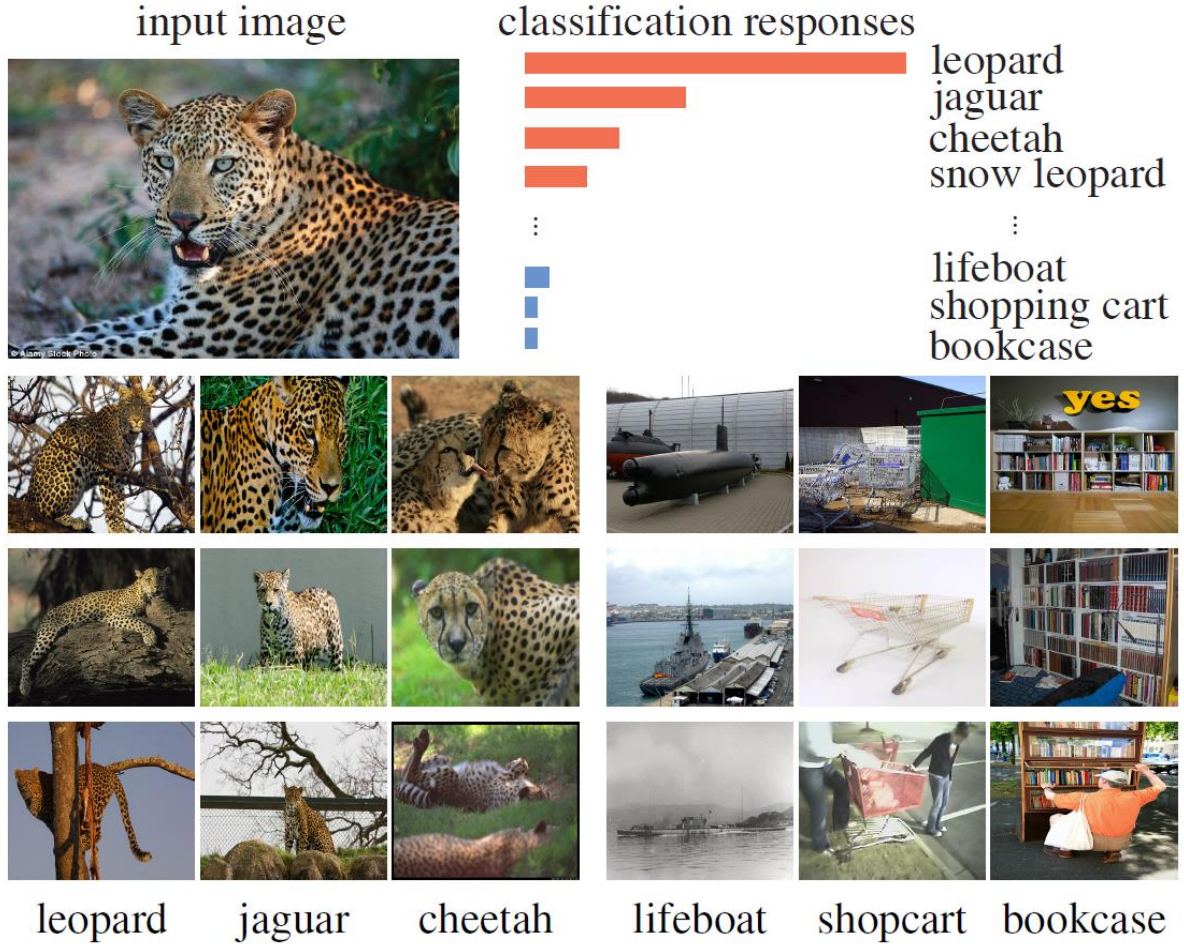
ImageNet (AlexNet)

Content

- Self-supervised Vision Learning
  - 1. Self-supervised Contextual Modeling
  - 2. Contrastive learning-based self-supervised learning
  - 3. Returning of Self-supervised Contextual Modeling

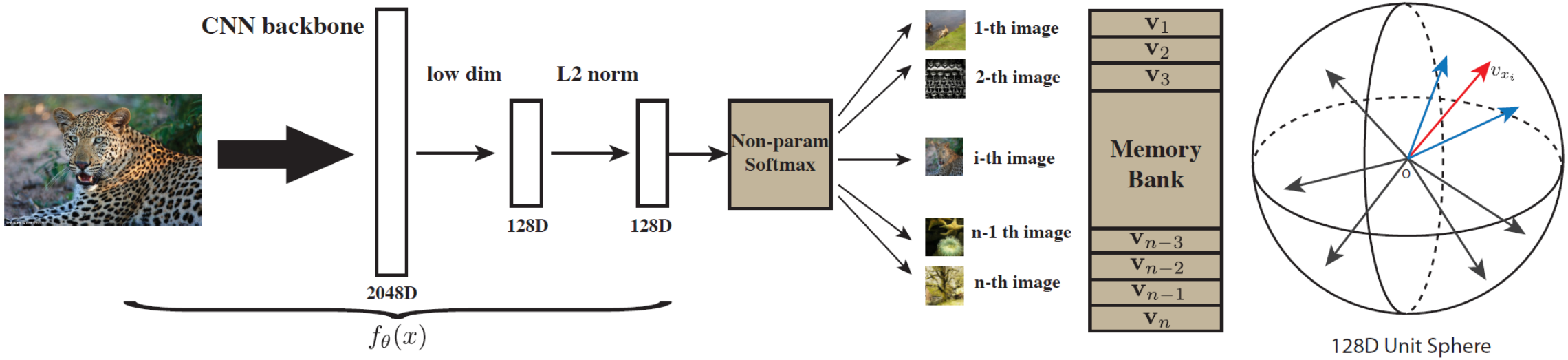
Instance Discrimination

- Takes the class-wise supervision to the extreme
- Learns a feature representation that discriminates among individual instances



Zhirong Wu, Yuanjun Xiong, Stella Yu, and Dahua Lin. Unsupervised feature learning via non-parametric instance discrimination. In CVPR, 2018.

Instance Discrimination



- Memory bank $V = \{\mathbf{v}_j\}$
- Forward feature $\mathbf{f}_i = f_{\boldsymbol{\theta}}(x_i)$
- $\mathbf{f}_i \rightarrow \mathbf{v}_i$

Memory Bank $\mathbf{v}_j^T \mathbf{v} \quad \|\mathbf{v}\| = 1$

Noise Contrastive Learning

- Softmax classifier

$$P(i|\mathbf{v}) = \frac{\exp(\mathbf{v}^T \mathbf{f}_i/\tau)}{Z_i}$$

$$Z_i = \sum_{j=1}^{n} \exp\left(\mathbf{v}_j^T \mathbf{f}_i/\tau\right)$$

- Posterior probability

$$P_n = 1/n \qquad h(i, \mathbf{v}) := P(D = 1|i, \mathbf{v}) = \frac{P(i|\mathbf{v})}{P(i|\mathbf{v}) + mP_n(i)}$$

- Monte Carlo approximation

$$Z \simeq Z_i \simeq nE_j \left[\exp(\mathbf{v}_j^T \mathbf{f}_i/\tau)\right] = \frac{n}{m} \sum_{k=1}^{m} \exp(\mathbf{v}_{j_k}^T \mathbf{f}_i/\tau)$$

- Negative log-posterior

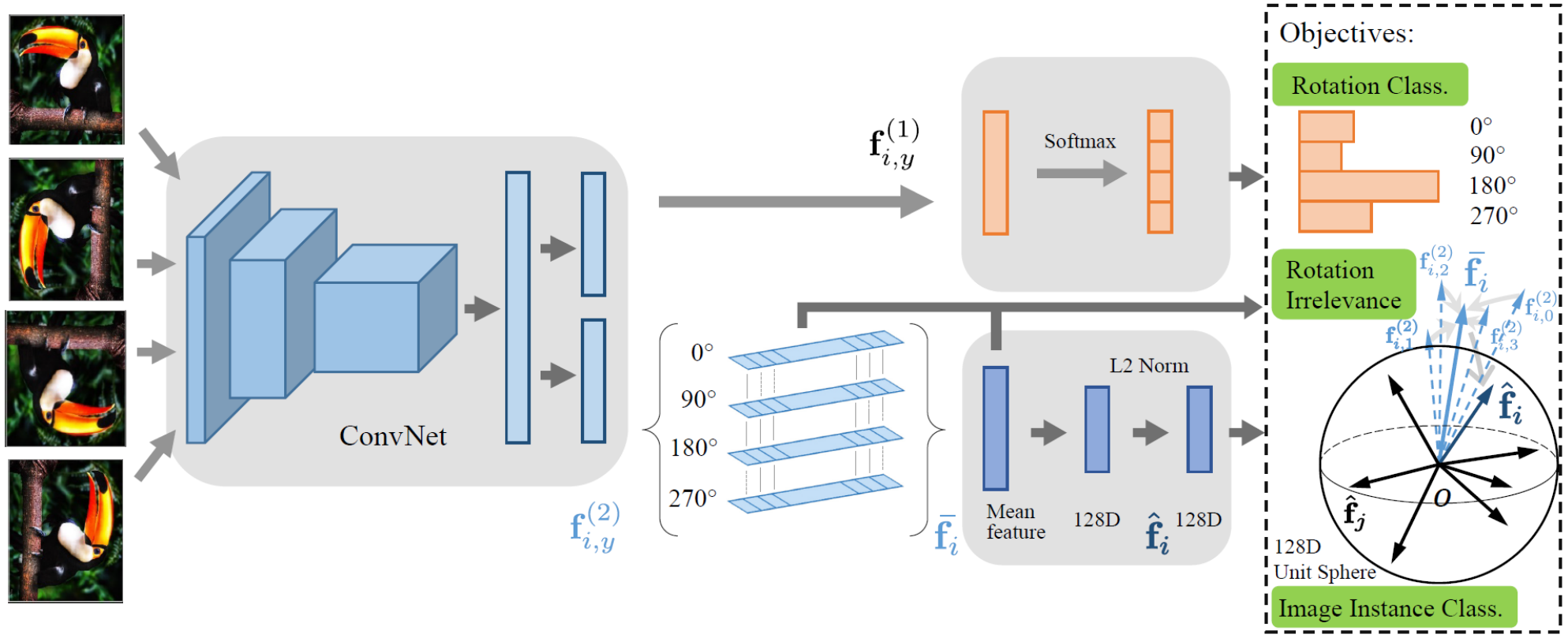
$$J_{NCE}(\boldsymbol{\theta}) = -E_{P_d}\left[\log h(i, \mathbf{v}_i^{(t-1)}) - \lambda \|\mathbf{v}_i^{(t)} - \mathbf{v}_i^{(t-1)}\|_2^2\right]$$
$$-m \cdot E_{P_n}\left[\log(1 - h(i, \mathbf{v}'^{(t-1)}))\right]. \qquad (10)$$

Training / Testing	Linear SVM	Nearest Neighbor
Param Softmax	60.3	63.0
Non-Param Softmax	75.4	**80.8**
NCE $m = 1$	44.3	42.5
NCE $m = 10$	60.2	63.4
NCE $m = 512$	64.3	78.4
NCE $m = 4096$	70.2	**80.4**

Image Classification Accuracy on ImageNet							
method	conv1	conv2	conv3	conv4	conv5	kNN	#dim
Random	11.6	17.1	16.9	16.3	14.1	3.5	10K
Data-Init [16]	17.5	23.0	24.5	23.2	20.6	-	10K
Context [2]	16.2	23.3	30.2	31.7	29.6	-	10K
Adversarial [4]	17.7	24.5	31.0	29.9	28.0	-	10K
Color [47]	13.1	24.8	31.0	32.6	31.8	-	10K
Jigsaw [27]	19.2	30.1	34.7	33.9	28.3	-	10K
Count [28]	18.0	30.6	34.3	32.5	25.7	-	10K
SplitBrain [48]	17.7	29.3	35.4	35.2	32.8	11.8	10K
Exemplar[3]			31.5			-	4.5K
Ours Alexnet	16.8	26.5	31.8	34.1	**35.6**	31.3	128
Ours VGG16	16.5	21.4	27.6	35.1	**39.2**	33.9	128
Ours Resnet18	16.0	19.9	29.8	39.0	**44.5**	**41.0**	128
Ours Resnet50	15.3	18.8	24.9	40.6	**54.0**	**46.5**	128

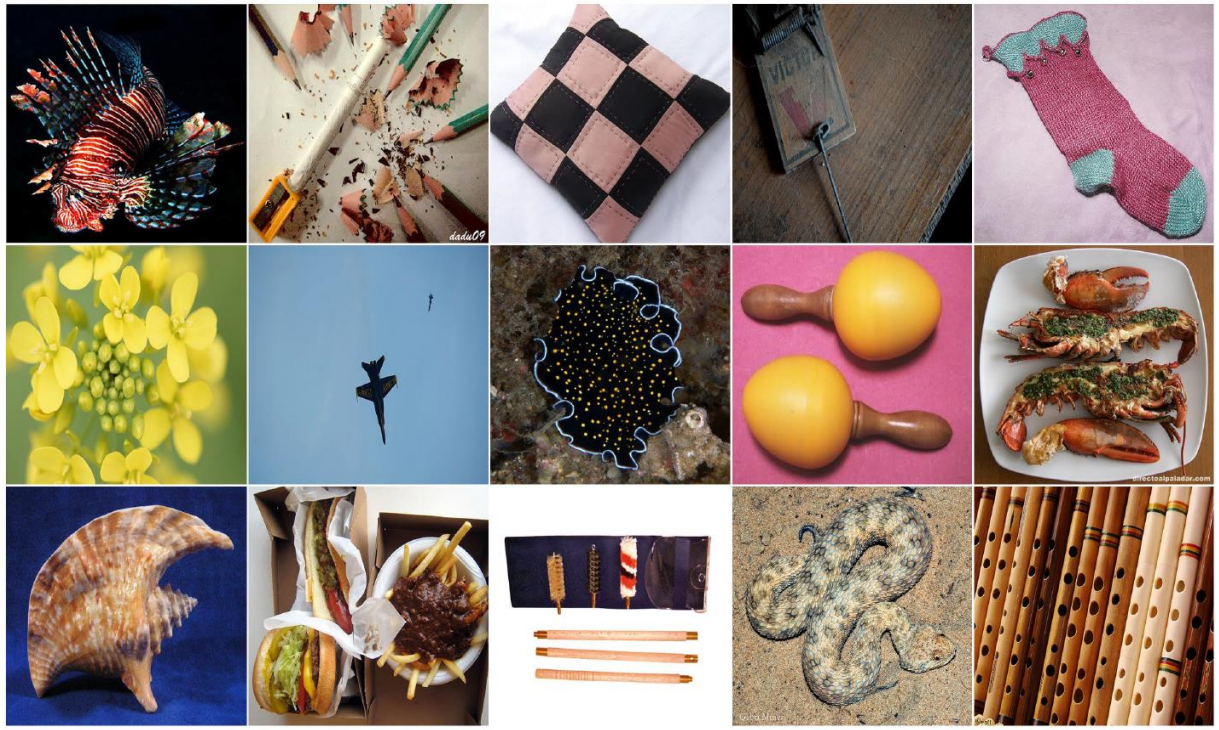
Table 2: Top-1 classification accuracy on ImageNet.

Rotation + Instance Discrimination



Z. Feng, C. Xu, D. Tao, Self-Supervised Representation Learning by Rotation Feature Decoupling, CVPR 2019

Rotation + Instance Discrimination

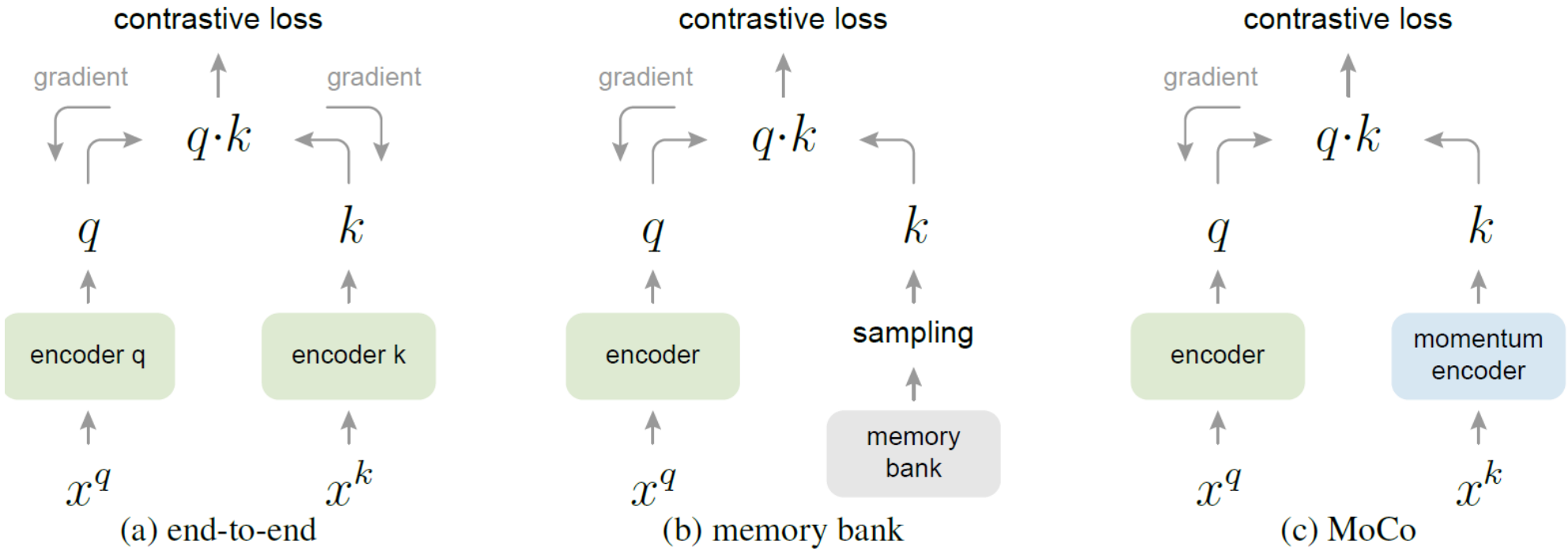


Rotation agnostic

Method\Layer	conv1	conv2	conv3	conv4	conv5
ImageNet-labels [25, 52]	19.3	36.3	44.2	48.3	50.5
Random [53]	11.6	17.1	16.9	16.3	14.1
Krähenbühl *et al.* [22]	17.5	23.0	24.5	23.2	20.6
Pathak *et al.* (Inpainting) [43]	14.1	20.7	21.0	19.8	15.5
Noroozi & Favaro (Jigsaw) [37]	18.2	28.8	34.0	33.9	27.1
Zhang *et al.* (Colorization) [52]	13.1	24.8	31.0	32.6	31.8
Donahue *et al.* (BiGANs) [12]	17.7	24.5	31.0	29.9	28.0
Zhang *et al.* (Split-Brain) [53]	17.7	29.3	35.4	35.2	32.8
Noroozi *et al.* (Counting) [38]	18.0	30.6	34.3	32.5	25.7
Gidaris *et al.* (RotNet) [17]	18.8	31.7	38.7	38.2	36.5
Jenni & Favaro [21]	19.5	**33.3**	37.9	38.9	34.9
Mundhenk *et al.* [36]	**19.6**	31.8	37.6	37.8	33.7
Noroozi *et al.* (CC+) [39]	18.9	30.5	35.7	35.4	32.2
Noroozi *et al.* (CC+vgg-) [39]	19.2	32.0	37.3	37.1	34.6
Wu *et al.* [48]	16.8	26.5	31.8	34.1	35.6
Doersch *et al.* (Context) [10]*	16.2	23.3	30.2	31.7	29.6
Ren & Lee [44]*	16.5	27.0	30.5	30.1	26.5
Caron *et al.* (DeepCluster) [6]*†	13.4	32.3	41.0	39.6	38.2
Ours	19.3	**33.3**	**40.8**	**41.8**	**44.3**
Ours (*bigger* AlexNet)*	20.8	35.2	41.8	44.3	44.4
Ours (*bigger* AlexNet)*†	22.2	38.2	45.7	48.7	48.3

ImageNet (AlexNet)

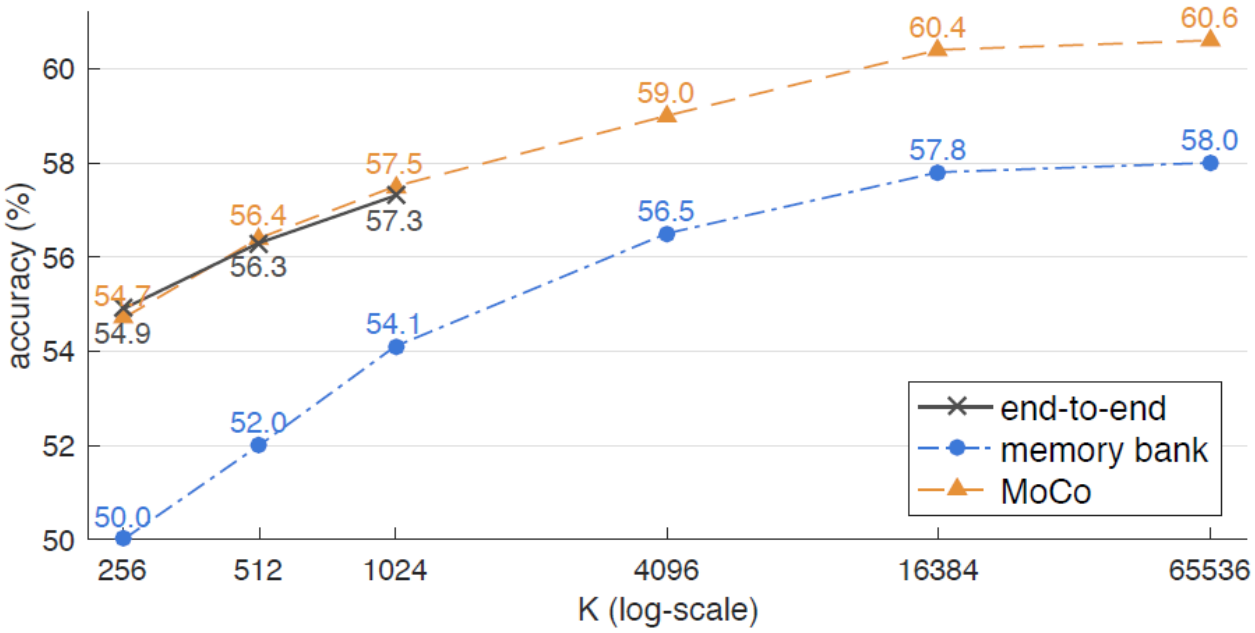
MoCo: Momentum Contrast



- Dynamic dictionary as a queue of data samples
- Momentum update $\theta_{\mathrm{k}} \leftarrow m\theta_{\mathrm{k}} + (1-m)\theta_{\mathrm{q}} \quad m = 0.999$

K. He, H. Fan, Y. Wu, S. Xie R. Girshick, Momentum Contrast for Unsupervised Visual Representation Learning, CVPR 2020

Ablation studies



momentum m	0	0.9	0.99	0.999	0.9999
accuracy (%)	*fail*	55.2	57.8	59.0	58.9

	R50-dilated-C5			R50-C4		
pre-train	AP_{50}	AP	AP_{75}	AP_{50}	AP	AP_{75}
end-to-end	77.8	50.1	53.8	79.7	53.0	57.9
memory bank	79.6	51.9	56.3	80.3	53.9	58.9
MoCo	**81.1**	**53.8**	**58.6**	**81.4**	**55.2**	**61.2**

method	architecture	#params (M)	accuracy (%)
Exemplar [15]	R50w3×	211	46.0 [36]
RelativePosition [11]	R50w2×	94	51.4 [36]
Jigsaw [43]	R50w2×	94	44.6 [36]
Rotation [17]	Rv50w4×	86	55.4 [36]
Colorization [62]	R101*	28	39.6 [12]
DeepCluster [3]	VGG [51]	15	48.4 [4]
BigBiGAN [14]	R50	24	56.6
	Rv50w4×	86	61.3
methods based on contrastive learning follow:			
InstDisc [59]	R50	24	54.0
LocalAgg [64]	R50	24	58.8
CPC v1 [44]	R101*	28	48.7
CPC v2 [33]	$R170^*_{wider}$	303	65.9
CMC [54]	$R50_{L+ab}$	47	64.1†
	$R50w2\times_{L+ab}$	188	68.4†
AMDIM [2]	$AMDIM_{small}$	194	63.5†
	$AMDIM_{large}$	626	68.1†
MoCo	R50	24	60.6
	RX50	46	63.9
	R50w2×	94	65.4
	R50w4×	375	**68.6**

pre-train	AP_{50}	AP	AP_{75}
random init.	58.0	32.8	32.5
super. IN-1M	81.5	53.6	58.9
MoCo IN-1M	81.1 (−0.4)	53.8 (+0.2)	58.6 (−0.3)
MoCo IG-1B	81.6 (+0.1)	54.8 (+**1.2**)	60.3 (+**1.4**)

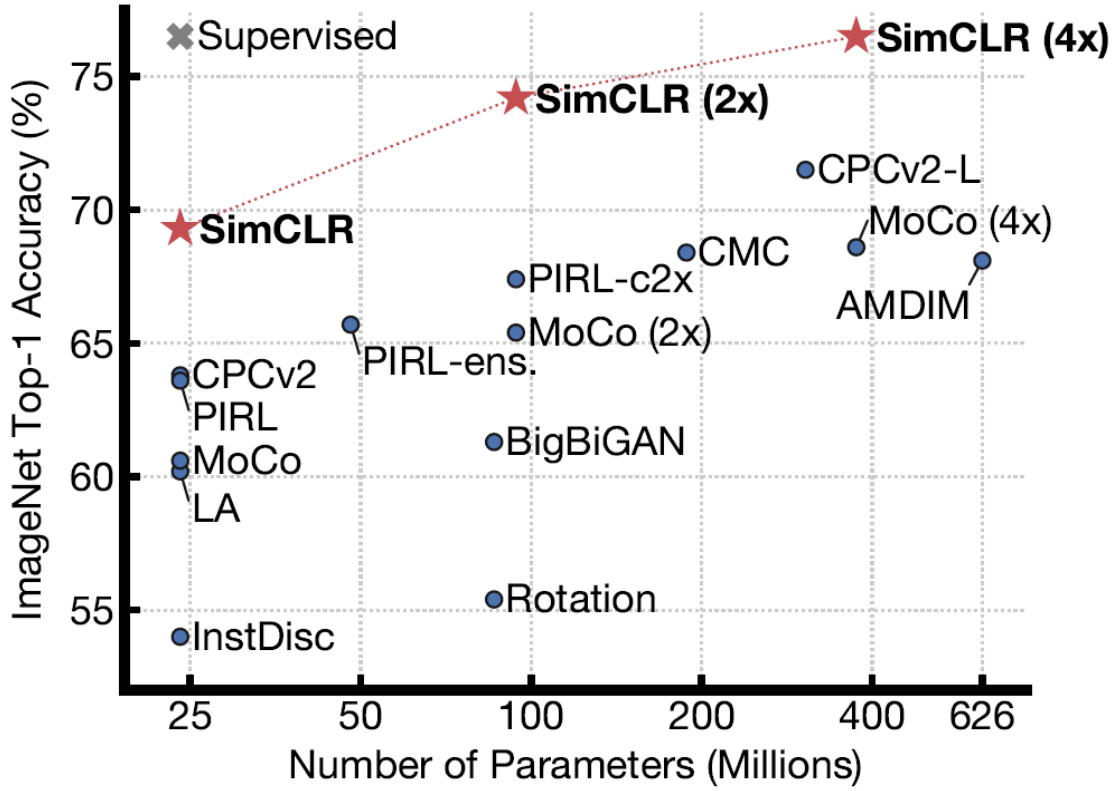
(a) Faster R-CNN, R50-**dilated-C5**

pre-train	AP_{50}	AP	AP_{75}
random init.	52.5	28.1	26.2
super. IN-1M	80.8	52.0	56.5
MoCo IN-1M	81.4 (+**0.6**)	55.2 (+**3.2**)	61.2 (+**4.7**)
MoCo IG-1B	82.1 (+**1.3**)	56.2 (+**4.2**)	62.3 (+**5.8**)

(b) Faster R-CNN, R50-**C4**

SimCLR

- Data augmentations
- Adding nonlinear transformation
  - Projection head
- Larger batch sizes and more training steps



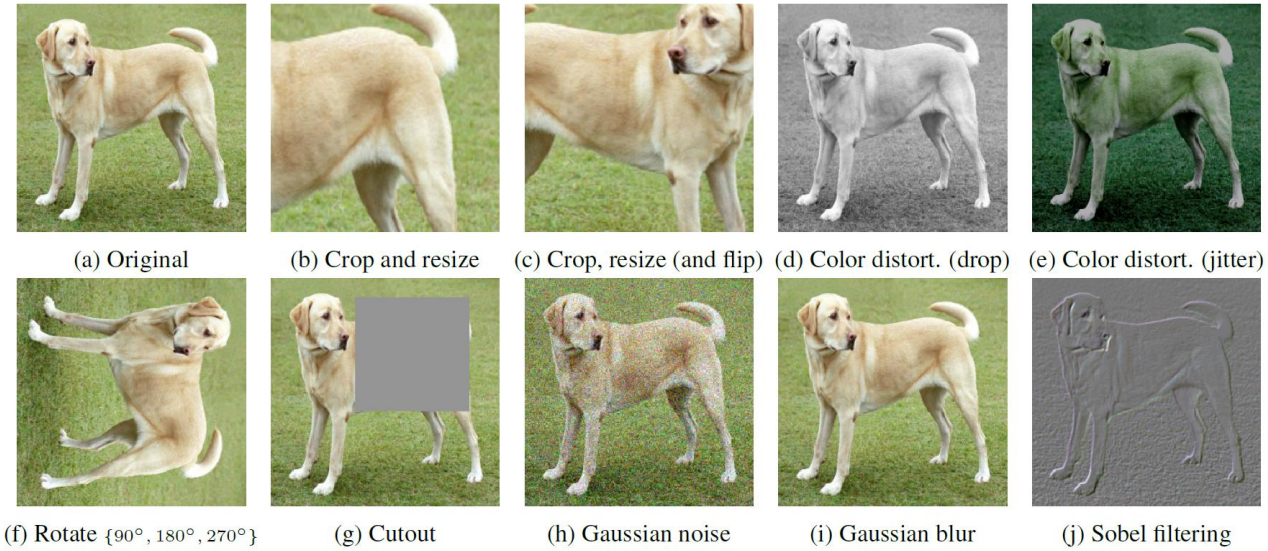
Ting Chen, Simon Kornblith, Mohammad Norouzi, Geoffrey Hinton, A Simple Framework for Contrastive Learning of Visual Representations, ICML 2020.

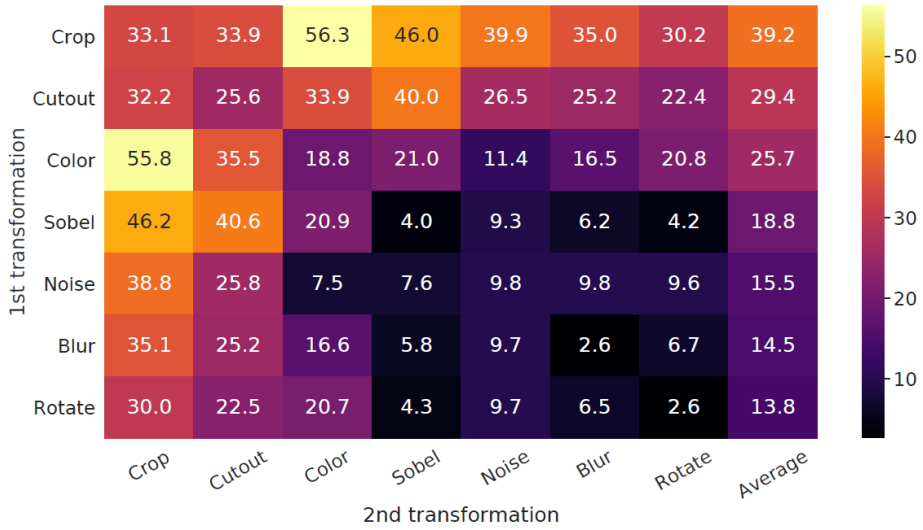
SimCLR

- Loss function

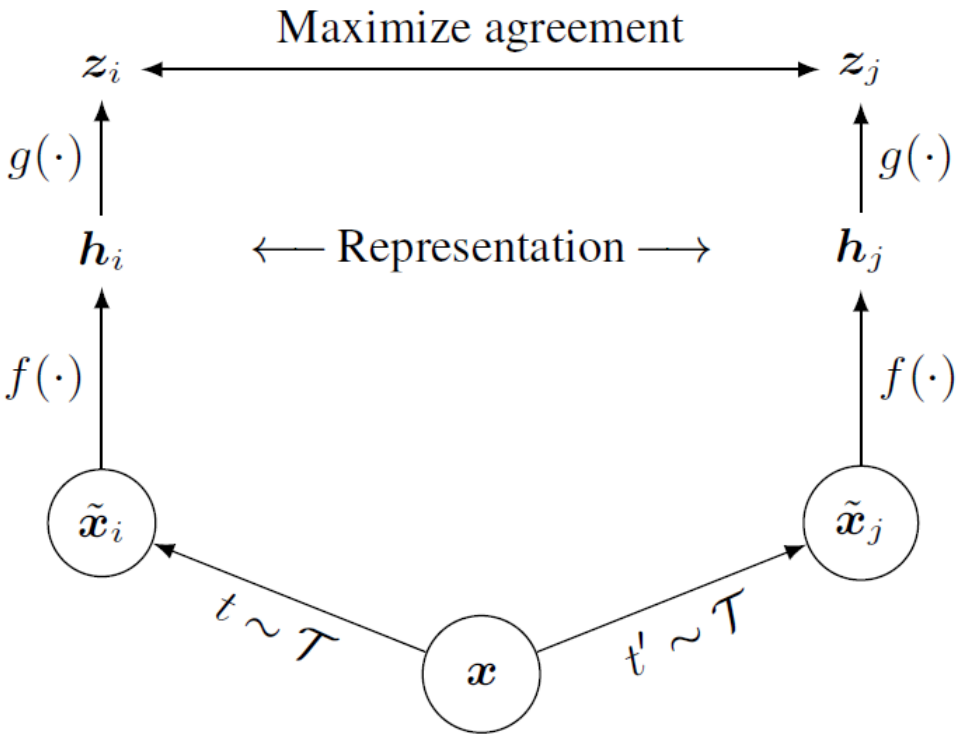
$$\ell_{i,j} = -\log \frac{\exp(\text{sim}(\boldsymbol{z}_i, \boldsymbol{z}_j)/\tau)}{\sum_{k=1}^{2N} \mathbb{1}_{[k \neq i]} \exp(\text{sim}(\boldsymbol{z}_i, \boldsymbol{z}_k)/\tau)}$$

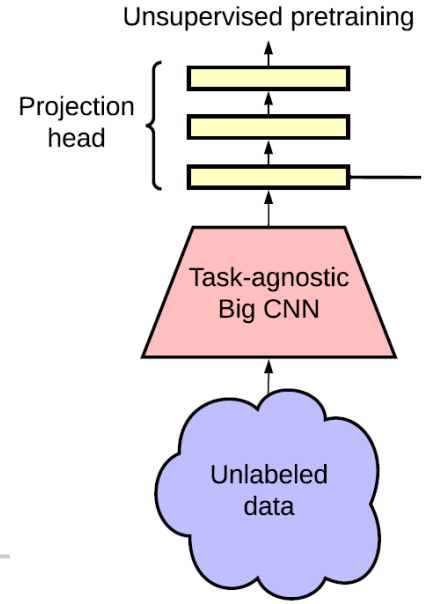
- Data Augmentation

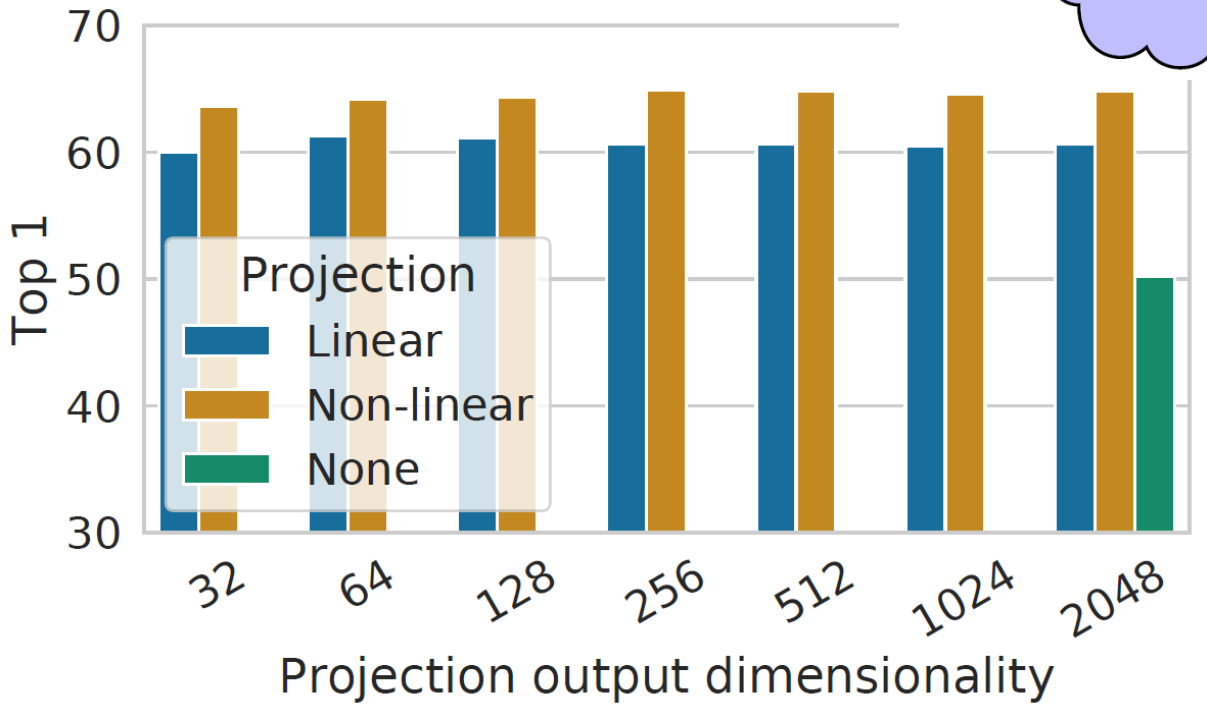




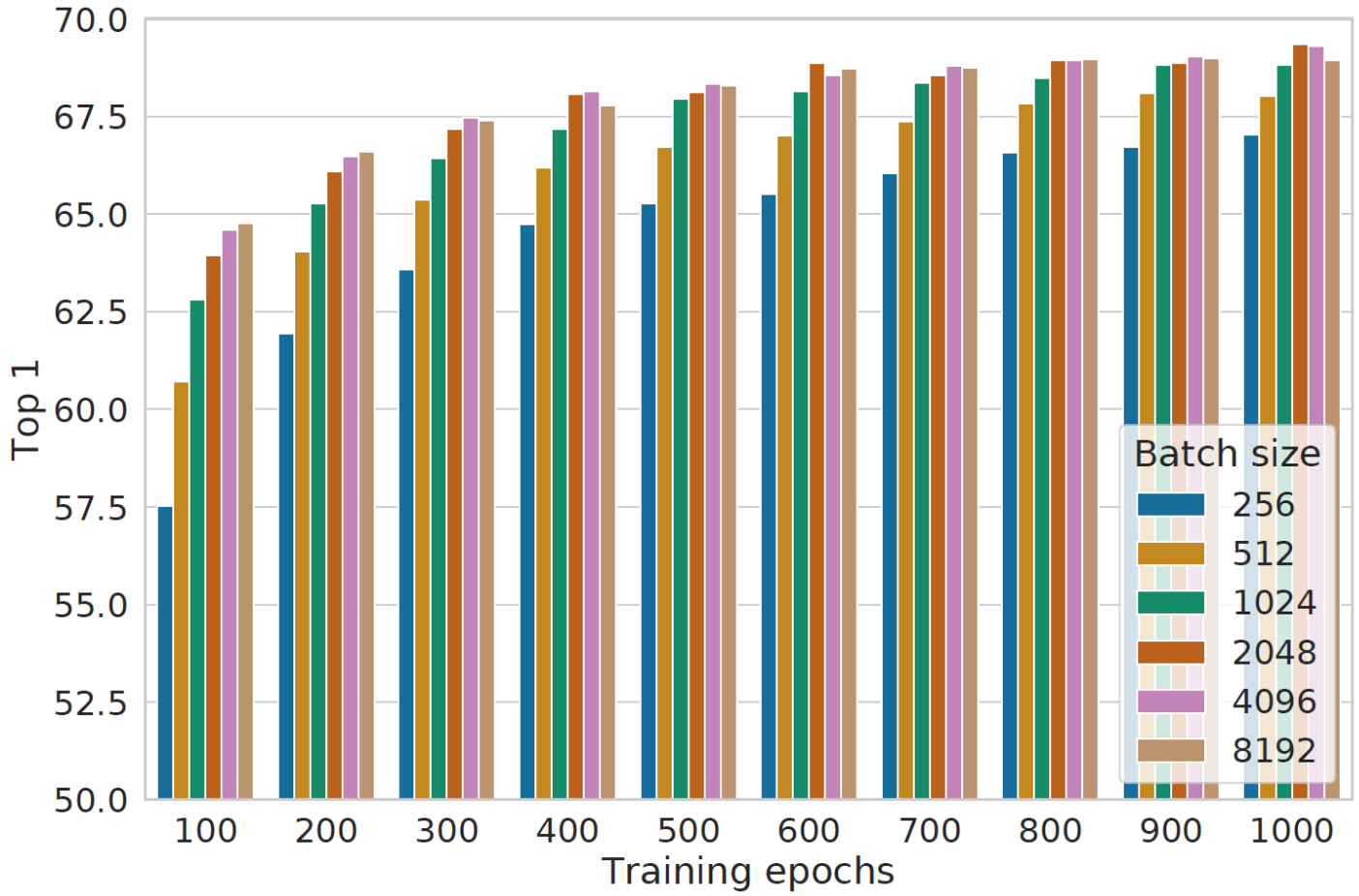
Adding nonlinear transformation







Larger batch sizes and longer training



Method	Architecture	Param.	Top 1	Top 5
Methods using ResNet-50:				
Local Agg.	ResNet-50	24	60.2	-
MoCo	ResNet-50	24	60.6	-
PIRL	ResNet-50	24	63.6	-
CPC v2	ResNet-50	24	63.8	85.3
SimCLR (ours)	ResNet-50	24	**69.3**	**89.0**
Methods using other architectures:				
Rotation	RevNet-50 (4×)	86	55.4	-
BigBiGAN	RevNet-50 (4×)	86	61.3	81.9
AMDIM	Custom-ResNet	626	68.1	-
CMC	ResNet-50 (2×)	188	68.4	88.2
MoCo	ResNet-50 (4×)	375	68.6	-
CPC v2	ResNet-161 (∗)	305	71.5	90.1
SimCLR (ours)	ResNet-50 (2×)	94	74.2	92.0
SimCLR (ours)	ResNet-50 (4×)	375	**76.5**	**93.2**

Method	Architecture	Label fraction (Top 5) 1%	10%
Methods using other label-propagation:			
Pseudo-label	ResNet50	51.6	82.4
VAT+Entropy Min.	ResNet50	47.0	83.4
UDA (w. RandAug)	ResNet50	-	88.5
FixMatch (w. RandAug)	ResNet50	-	89.1
S4L (Rot+VAT+En. M.)	ResNet50 (4×)	-	91.2
Methods using representation learning only:			
InstDisc	ResNet50	39.2	77.4
BigBiGAN	RevNet-50 (4×)	55.2	78.8
PIRL	ResNet-50	57.2	83.8
CPC v2	ResNet-161(∗)	77.9	91.2
SimCLR (ours)	ResNet-50	75.5	87.8
SimCLR (ours)	ResNet-50 (2×)	83.0	91.2
SimCLR (ours)	ResNet-50 (4×)	**85.8**	**92.6**

	Food	CIFAR10	CIFAR100	Birdsnap	SUN397	Cars	Aircraft	VOC2007	DTD	Pets	Caltech-101	Flowers
Linear evaluation:												
SimCLR (ours)	**76.9**	**95.3**	80.2	48.4	**65.9**	60.0	61.2	**84.2**	**78.9**	89.2	**93.9**	**95.0**
Supervised	75.2	**95.7**	**81.2**	**56.4**	64.9	**68.8**	**63.8**	83.8	**78.7**	**92.3**	**94.1**	94.2
Fine-tuned:												
SimCLR (ours)	**89.4**	**98.6**	**89.0**	**78.2**	**68.1**	**92.1**	**87.0**	**86.6**	**77.8**	92.1	**94.1**	97.6
Supervised	88.7	98.3	**88.7**	**77.8**	67.0	91.4	**88.0**	86.5	**78.8**	**93.2**	**94.2**	**98.0**
Random init	88.3	96.0	81.9	**77.0**	53.7	91.3	84.8	69.4	64.1	82.7	72.5	92.5

MoCo v2

- MLP projection head (in SimCLR)
- More data augmentation (in SimCLR): blur augmentation

case	unsup. pre-train MLP	aug+	cos	epochs	ImageNet acc.	VOC detection AP_{50}	AP	AP_{75}
supervised					76.5	81.3	53.5	58.8
MoCo v1				200	60.6	81.5	55.9	62.6
(a)	✓			200	66.2	82.0	56.4	62.6
(b)		✓		200	63.4	82.2	56.8	63.2
(c)	✓	✓		200	67.3	**82.5**	57.2	63.9
(d)	✓	✓	✓	200	67.5	82.4	57.0	63.6
(e)	✓	✓	✓	**800**	**71.1**	**82.5**	**57.4**	**64.0**

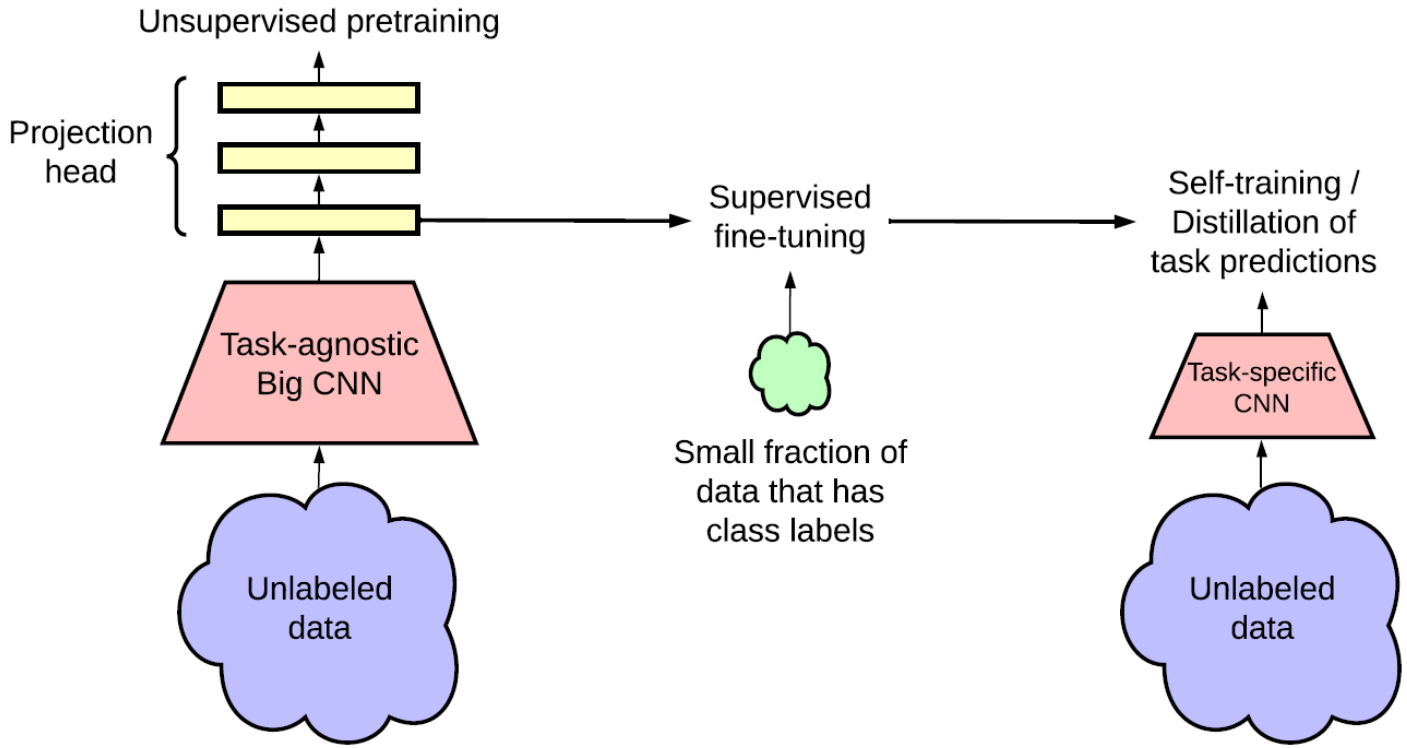
τ	0.07	0.1	0.2	0.3	0.4	0.5
w/o MLP	60.6	**60.7**	59.0	58.2	57.2	56.4
w/ MLP	62.9	64.9	**66.2**	65.7	65.0	64.3

Improved Baselines with Momentum Contrastive Learning, Arxiv 2020.

MoCo v2

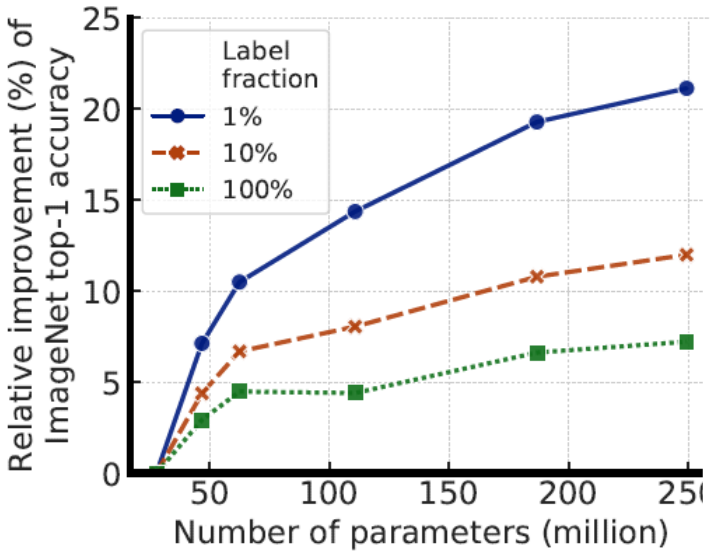
case	unsup. pre-train MLP	aug+	cos	epochs	batch	ImageNet acc.
MoCo v1 [6]				200	256	60.6
SimCLR [2]	✓	✓	✓	200	256	61.9
SimCLR [2]	✓	✓	✓	200	8192	66.6
MoCo v2	✓	✓	✓	200	256	**67.5**
*results of **longer** unsupervised training follow:*						
SimCLR [2]	✓	✓	✓	1000	4096	69.3
MoCo v2	✓	✓	✓	800	256	**71.1**

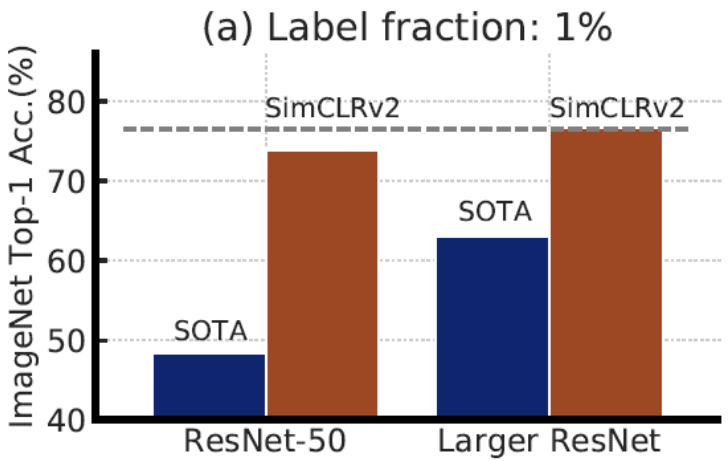
SimCLR v2 (NeurIPS 2020)

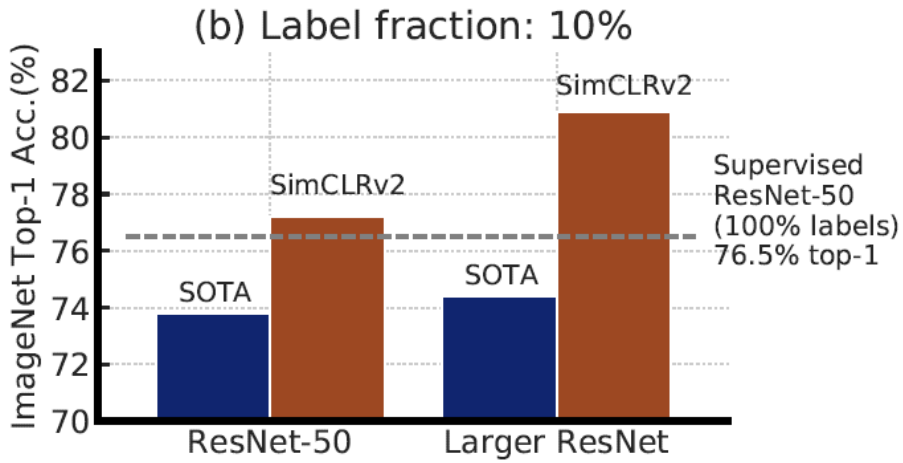


$$\mathcal{L}^{\text{distill}} = -\sum_{\boldsymbol{x}_i \in \mathcal{D}} \left[\sum_y P^T(y|\boldsymbol{x}_i;\tau) \log P^S(y|\boldsymbol{x}_i;\tau) \right]$$

SimCLRv2

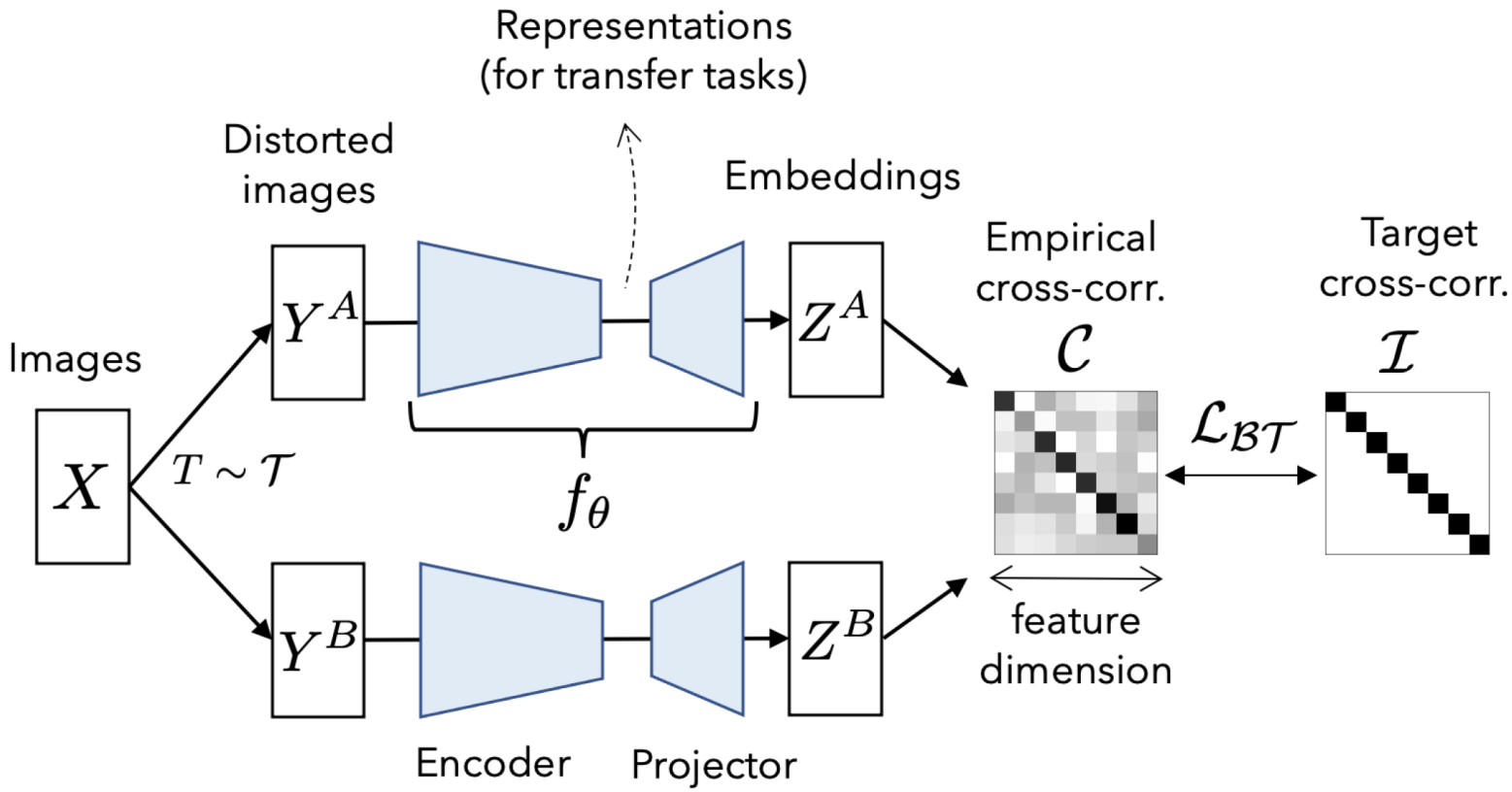

Relative improvement (%) of ImageNet top-1 accuracy
Number of parameters (million)
Label fraction
1%
10%
100%
0
5
10
15
20
25
50
100
150
200


(a) Label fraction: 1%
ImageNet Top-1 Acc.(%)
SimCLRv2
SimCLRv2
SOTA
SOTA
ResNet-50
Larger ResNet
40
50
60
70
80


(b) Label fraction: 10%
ImageNet Top-1 Acc.(%)
SimCLRv2
SimCLRv2
SOTA
SOTA
Supervised ResNet-50 (100% labels) 76.5% top-1
ResNet-50
Larger ResNet
70
72
74
76
78
80
82

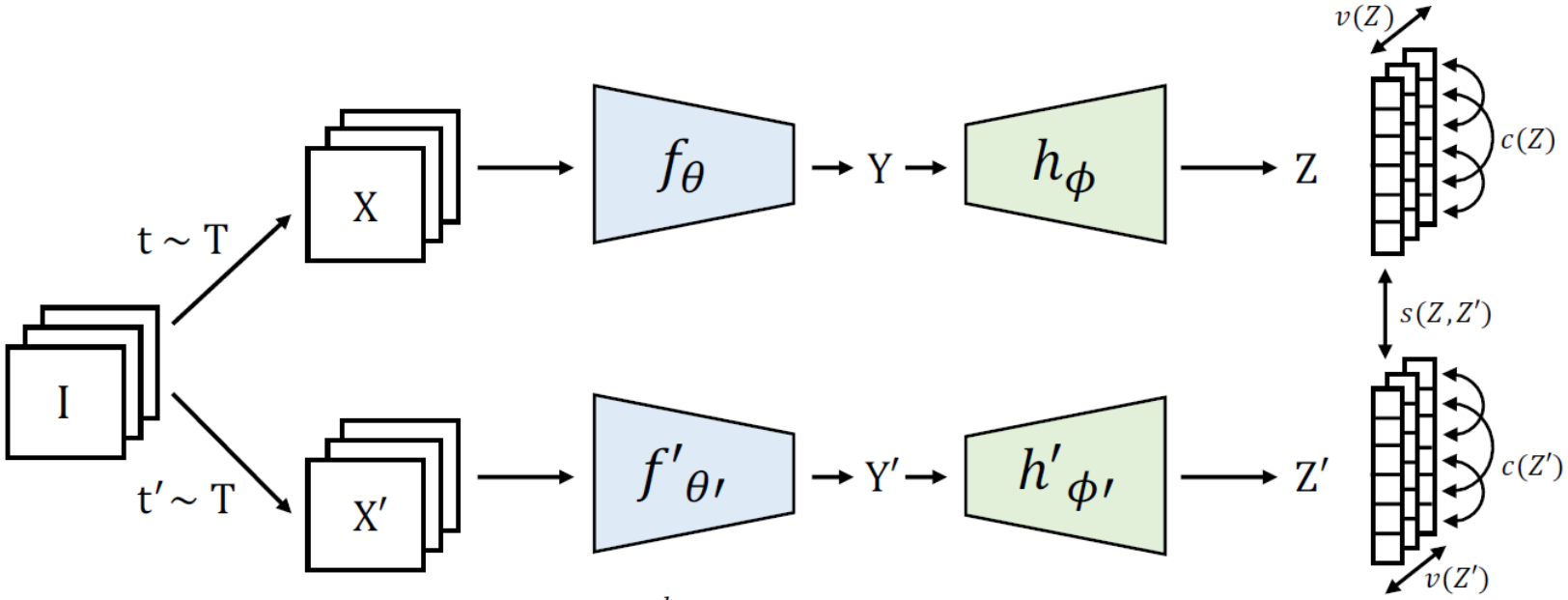
Barlow Twins (ICML 2021)

$$\mathcal{L}_{\mathcal{BT}} \triangleq \underbrace{\sum_i (1 - \mathcal{C}_{ii})^2}_{\text{invariance term}} + \lambda \underbrace{\sum_i \sum_{j \neq i} \mathcal{C}_{ij}{}^2}_{\text{redundancy reduction term}}$$



Barlow Twins: Self-Supervised Learning via Redundancy Reduction, ICML 2021.

VICReg (ICLR 2022)



v : maintain variance
c : bring covariance to zero
s : minimize distance

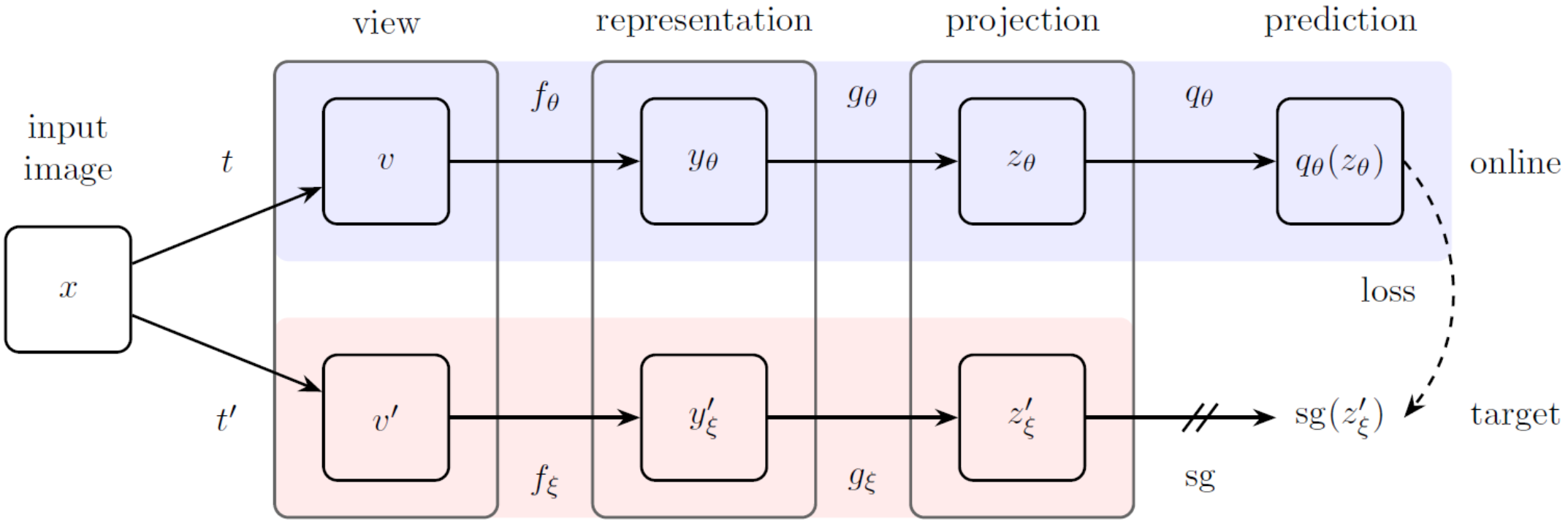
T : distribution of transformations
t,t' : random transformations

$f_\theta, f'_{\theta'}$: encoders
$h_\phi, h'_{\phi'}$: expanders

I : batch of images
X, X' : batches of views
Y, Y' : batches of representations
Z, Z' : batches of embeddings

- Variance $v(Z) = \frac{1}{d}\sum_{j=1}^{d} \max(0, \gamma - S(z^j, \epsilon))$
- Invariance $s(Z, Z') = \frac{1}{n}\sum_{i} \|z_i - z'_i\|_2^2$
- Covariance $c(Z) = \frac{1}{d}\sum_{i \neq j} [C(Z)]_{i,j}^2$

BYOL (NeurIPS 2020)



Exponential moving average

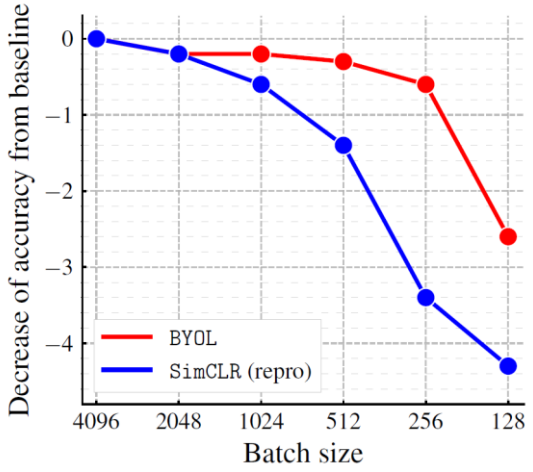
$$\xi \leftarrow \tau\xi + (1-\tau)\theta$$

$$\mathcal{L}_{\theta,\xi} \triangleq \left\|\overline{q_\theta}(z_\theta) - \overline{z}'_\xi\right\|_2^2 = 2 - 2 \cdot \frac{\langle q_\theta(z_\theta), z'_\xi \rangle}{\left\|q_\theta(z_\theta)\right\|_2 \cdot \left\|z'_\xi\right\|_2}$$

$$\theta \leftarrow \mathrm{optimizer}\left(\theta, \nabla_\theta \mathcal{L}^{\texttt{BYOL}}_{\theta,\xi}, \eta\right),$$
$$\xi \leftarrow \tau\xi + (1-\tau)\theta,$$

Bootstrap Your Own Latent: A New Approach to Self-Supervised Learning, NeurIPS 2020.

BYOL (NeurIPS 2020)



Local Agg.	60.2	-
PIRL [35]	63.6	-
CPC v2 [32]	63.8	85.3
CMC [11]	66.2	87.0
SimCLR [8]	69.3	89.0
MoCo v2 [37]	71.1	-
InfoMin Aug. [12]	73.0	91.1
BYOL (ours)	**74.3**	**91.6**

(a) ResNet-50 encoder.

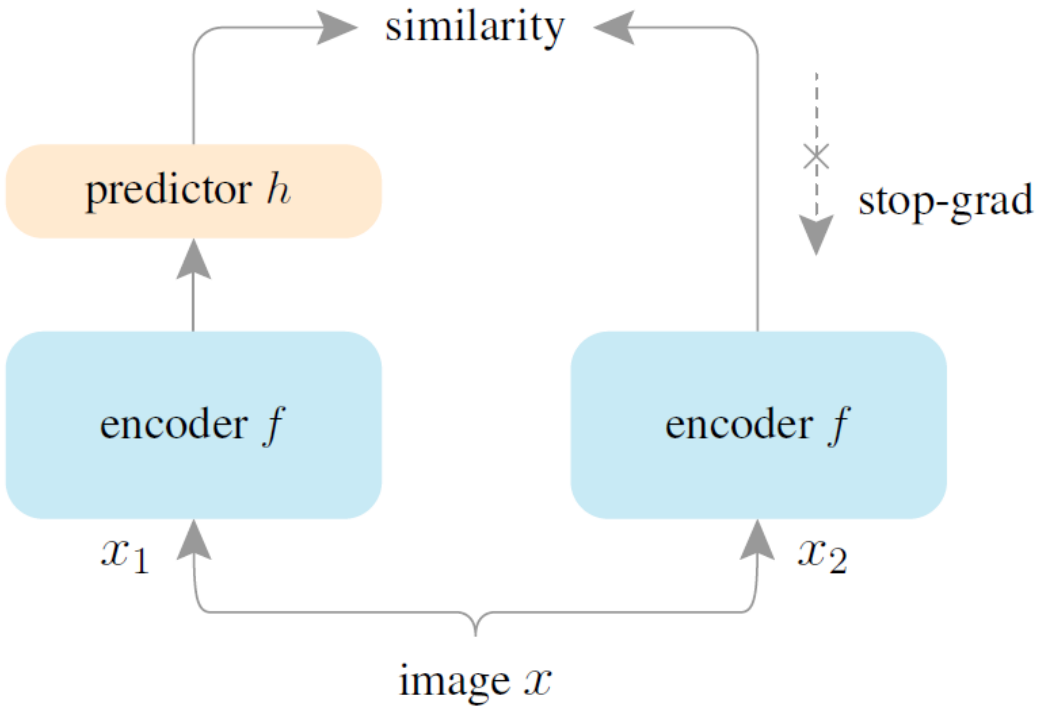
Method	Architecture	Param.	Top-1	Top-5
SimCLR [8]	ResNet-50 (2×)	94M	74.2	92.0
CMC [11]	ResNet-50 (2×)	94M	70.6	89.7
BYOL (ours)	ResNet-50 (2×)	94M	77.4	93.6
CPC v2 [32]	ResNet-161	305M	71.5	90.1
MoCo [9]	ResNet-50 (4×)	375M	68.6	-
SimCLR [8]	ResNet-50 (4×)	375M	76.5	93.2
BYOL (ours)	ResNet-50 (4×)	375M	78.6	94.2
BYOL (ours)	ResNet-200 (2×)	250M	**79.6**	**94.8**

(b) Other ResNet encoder architectures.

- Large batch size is still required.

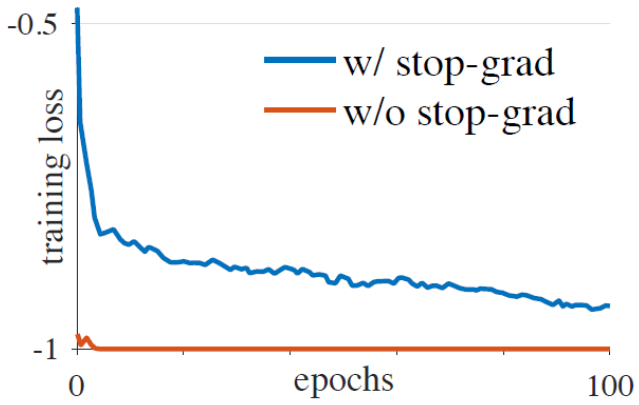
SimSiam (CVPR 2021)

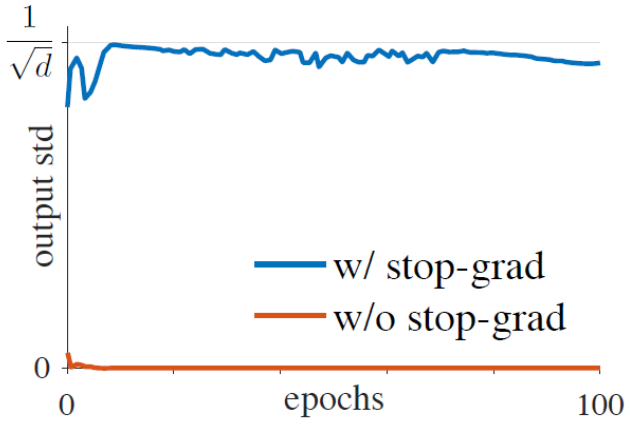
- Can learn meaningful representations even using none of the following:
  - Negative sample pairs
  - Large batches
  - Momentum encoders
- A stop-gradient operation plays an essential role in preventing collapsing

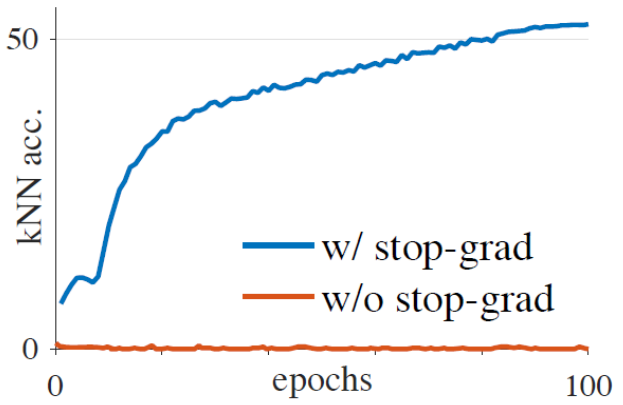


Exploring Simple Siamese Representation Learning, CVPR 2021.

SimSiam (CVPR 2021)







	acc. (%)
w/ stop-grad	67.7±0.1
w/o stop-grad	0.1

MoCo v3

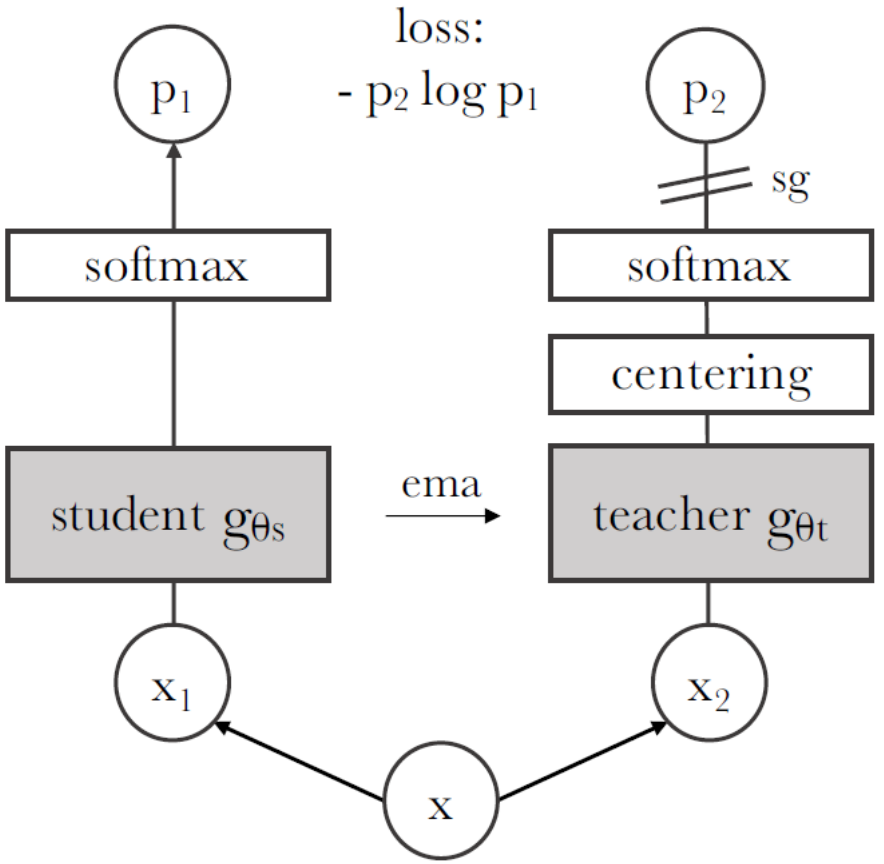
- Self-supervised learning for Vision Transformers
  - Batch size: 1k and 2k
  - Learning rate: lr x BatchSize / 256
  - Optimizer: LAMB, AdamW-counterpart of LARS.
  - Freezing the patch projection layer with a fixed random patch projection

framework	model	params	acc. (%)
linear probing:			
iGPT [9]	iGPT-L	1362M	69.0
iGPT [9]	iGPT-XL	6801M	72.0
MoCo v3	ViT-B	86M	76.7
MoCo v3	ViT-L	304M	77.6
MoCo v3	ViT-H	632M	78.1
MoCo v3	ViT-BN-H	632M	79.1
MoCo v3	ViT-BN-L/7	304M	**81.0**
end-to-end fine-tuning:			
masked patch pred. [16]	ViT-B	86M	79.9†
MoCo v3	ViT-B	86M	83.2
MoCo v3	ViT-L	304M	**84.1**

An Empirical Study of Training Self-Supervised Vision Transformers, ICCV 2021

DINO (ICCV 2021)

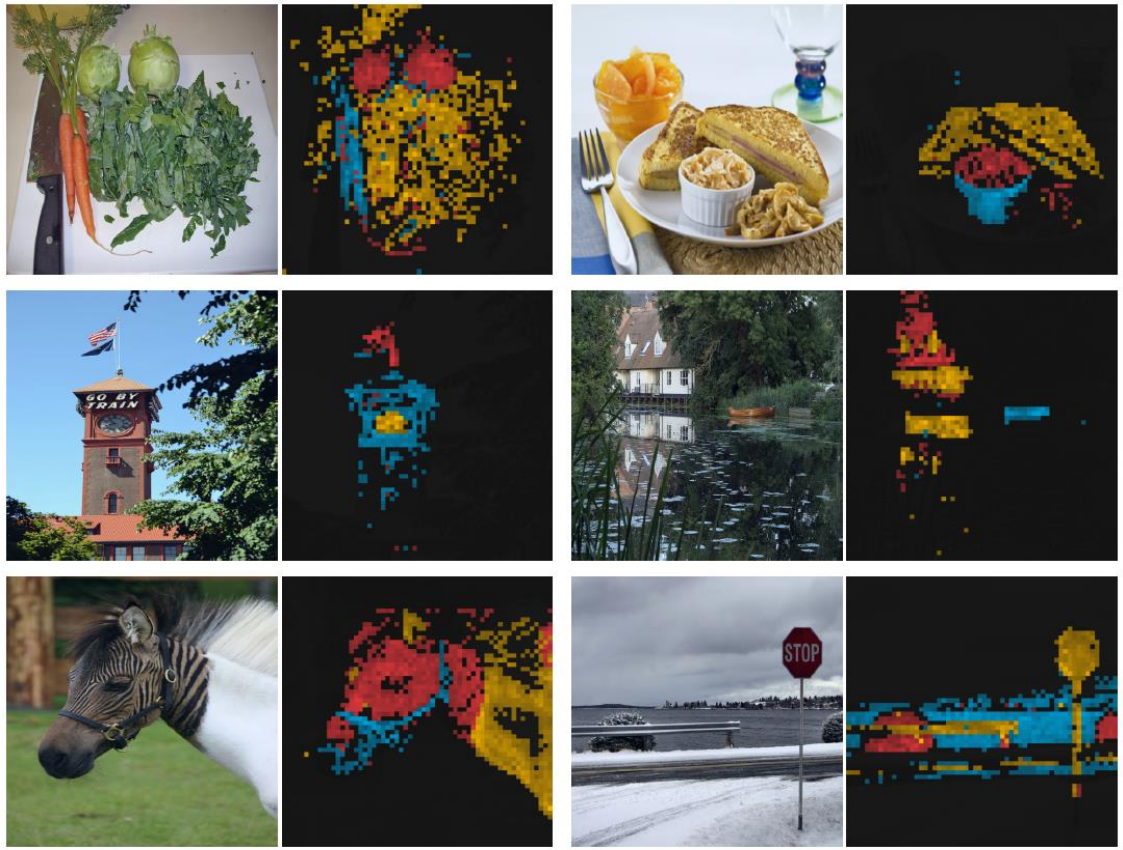
- Self-**di**stillation with **no** labels
- Self-supervised ViT
- Containing explicit information about the semantic segmentation of an image
- 78.3% top-1 on ImageNet with a small ViT



Emerging Properties in Self-Supervised Vision Transformers, ICCV 2021

DINO (ICCV 2021)

- Attention map



Comparison across architectures					
SCLR [11]	RN50w4	375	117	76.8	69.3
SwAV [9]	RN50w2	93	384	77.3	67.3
BYOL [23]	RN50w2	93	384	77.4	–
DINO	ViT-B/16	85	312	78.2	76.1
SwAV [9]	RN50w5	586	76	78.5	67.1
BYOL [23]	RN50w4	375	117	78.6	–
BYOL [23]	RN200w2	250	123	79.6	73.9
DINO	ViT-S/8	21	180	79.7	**78.3**
SCLRv2 [12]	RN152w3+SK	794	46	79.8	73.1
DINO	ViT-B/8	85	63	**80.1**	77.4

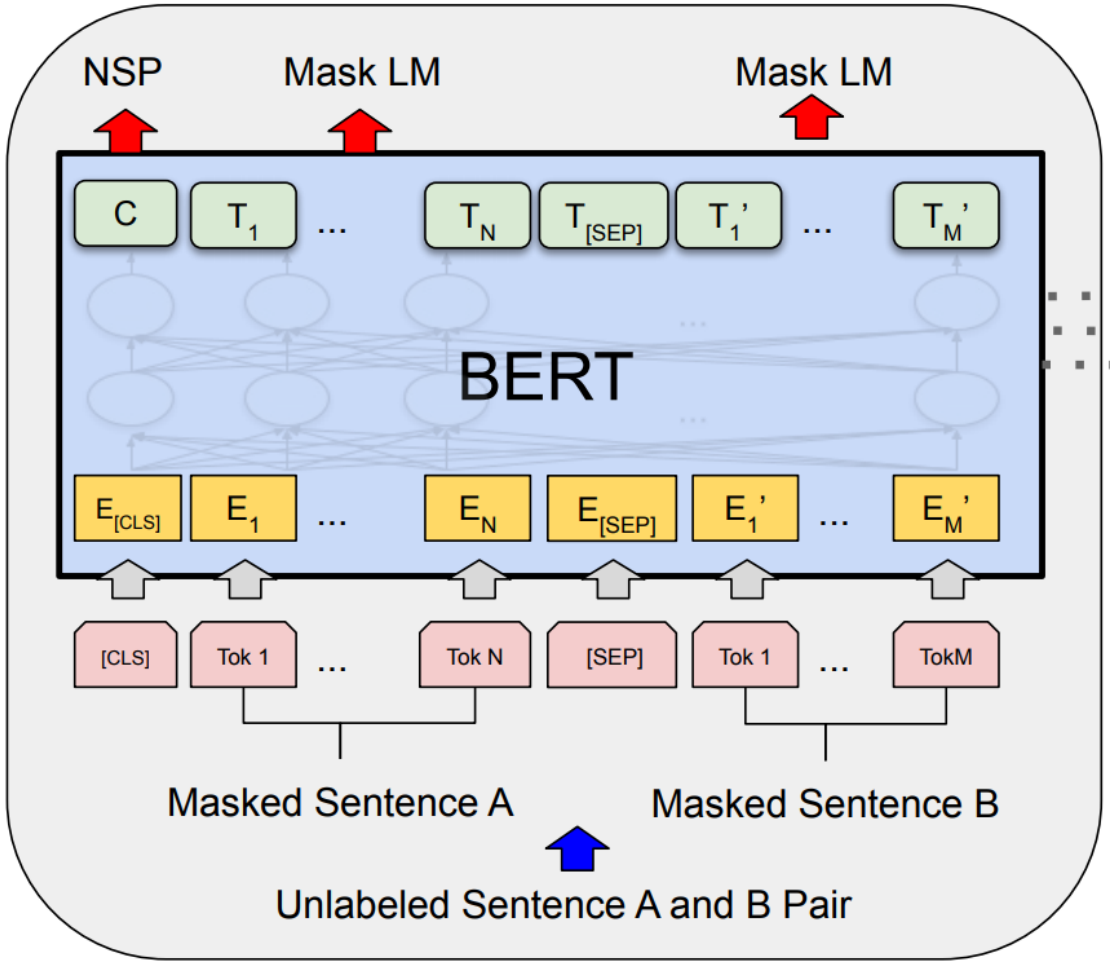
Content

- Self-supervised Vision Learning
  - 1. Self-supervised Contextual Modeling
  - 2. Contrastive learning-based self-supervised learning
  - 3. Returning of Self-supervised Contextual Modeling

BERT for Language Understanding (Devlin et al., Arxiv2018)

- Citations: 36172 (2018.10 —)

- Randomly masks some of the tokens from the input, and the objective is to predict the original vocabulary id of the masked word based only on its context.



Pre-training

Deep Context Modeling

- Jigsaw Puzzles
- Image Rotation
- Image Colorization
- Image Inpainting

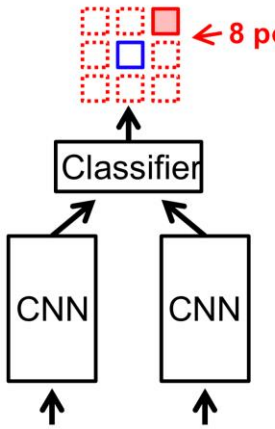
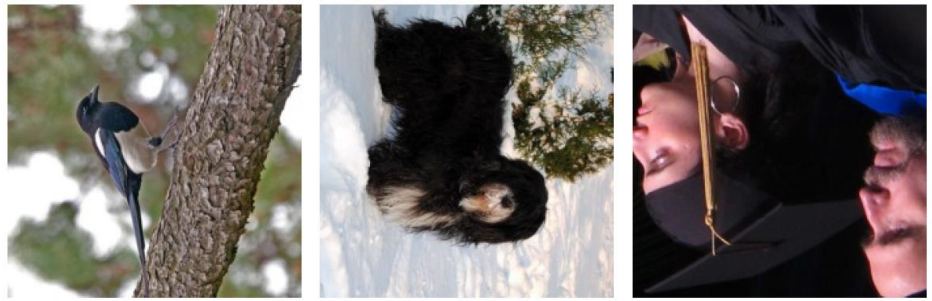
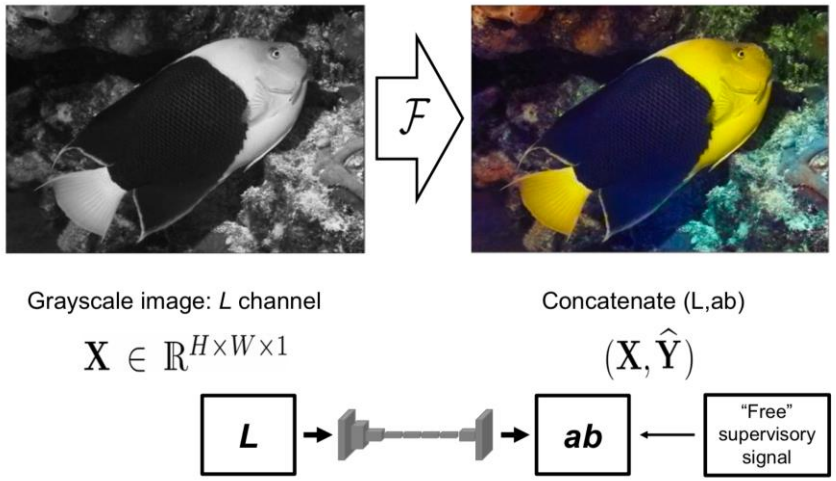
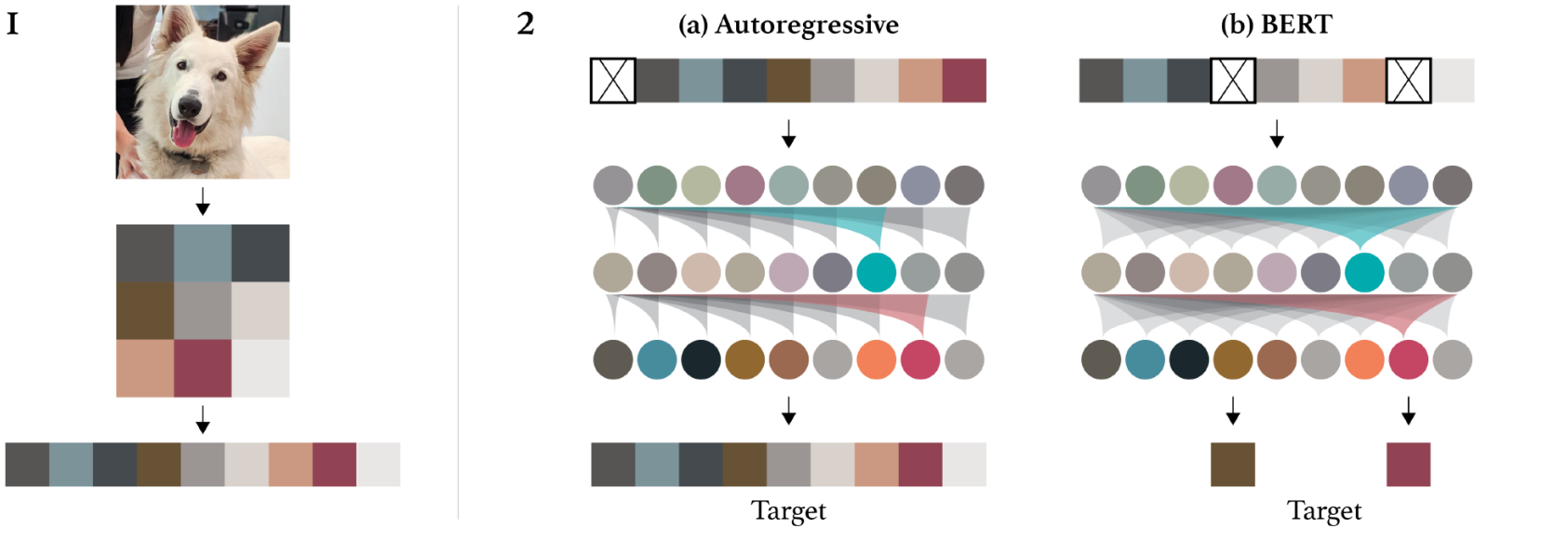


Image Rotation





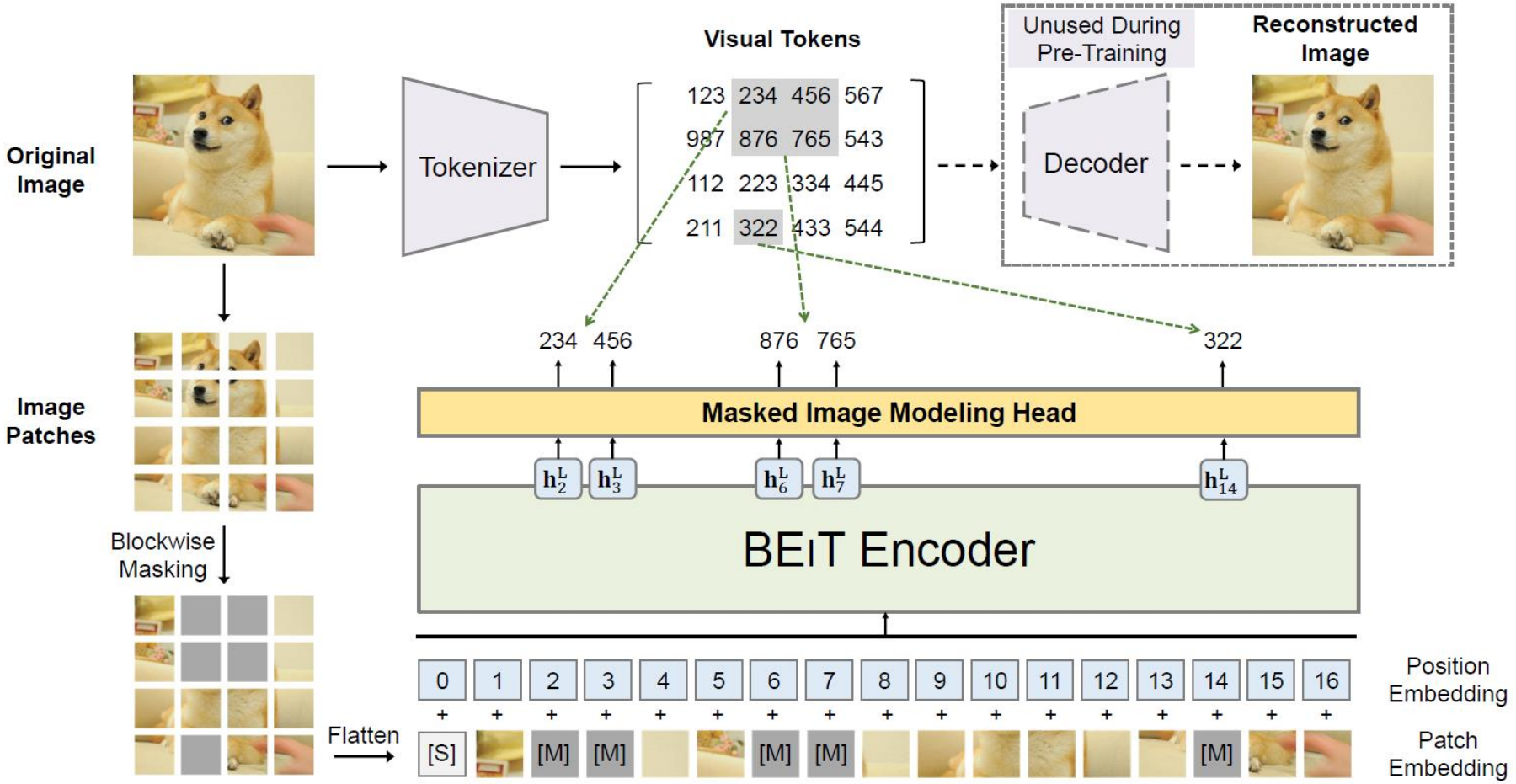
iGPT (ICML 2021)



$$L_{AR} = \mathop{\mathbb{E}}_{x \sim X}[-\log p(x)] \qquad L_{BERT} = \mathop{\mathbb{E}}_{x \sim X} \mathop{\mathbb{E}}_{M} \sum_{i \in M} \left[-\log p\left(x_i | x_{[1,n] \setminus M}\right)\right]$$

Generative Pretraining from Pixels, ICML 2021.

Masked Image Modeling: BEiT (ICLR 2022)



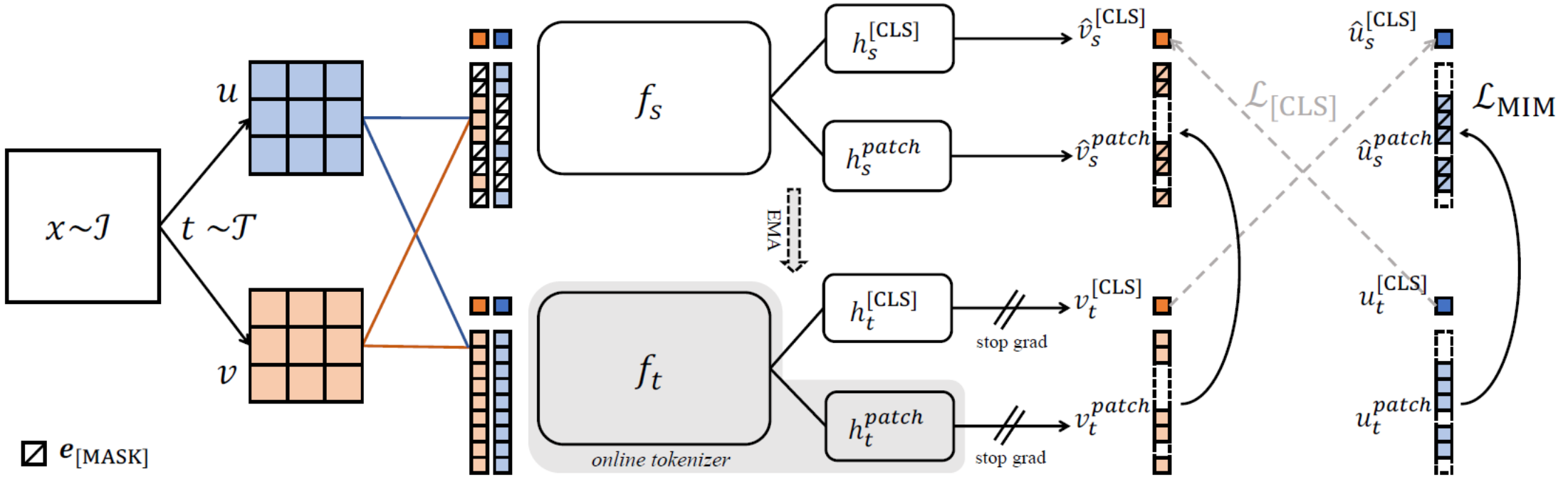
BEiT: BERT Pre-Training of Image Transformers, ICLR 2022

BEiT (ICLR 2022)

$$\sum_{(x_i, \tilde{x}_i) \in \mathcal{D}} \Big(\underbrace{\mathbb{E}_{z_i \sim q_\phi(z|x_i)}[\log p_\psi(x_i|z_i)]}_{\text{Stage 1: Visual Token Reconstruction}} + \underbrace{\log p_\theta(\hat{z}_i|\tilde{x}_i)}_{\text{Stage 2: Masked Image Modeling}} \Big)$$

Self-Supervised Pre-Training on ImageNet-1K (without labeled data)			
iGPT-1.36B† (Chen et al., 2020a)	1.36B	224^2	66.5
ViT$_{384}$-B-JFT300M‡ (Dosovitskiy et al., 2020)	86M	384^2	79.9
MoCo v3-B (Chen et al., 2021)	86M	224^2	83.2
MoCo v3-L (Chen et al., 2021)	307M	224^2	84.1
DINO-B (Caron et al., 2021)	86M	224^2	82.8
BEIT-B (ours)	86M	224^2	83.2
BEIT$_{384}$-B (ours)	86M	384^2	84.6
BEIT-L (ours)	307M	224^2	85.2
BEIT$_{384}$-L (ours)	307M	384^2	**86.3**

iBOT (ICLR 2022)



$$\mathcal{L}_{[\mathrm{CLS}]} = -P_{\theta'}^{[\mathrm{CLS}]}(\boldsymbol{v})^{\mathrm{T}} \log P_{\theta}^{[\mathrm{CLS}]}(\boldsymbol{u})$$

$$\mathcal{L}_{\mathrm{MIM}} = -\sum_{i=1}^{N} m_i \cdot P_{\theta'}^{\mathrm{patch}}(\boldsymbol{u}_i)^{\mathrm{T}} \log P_{\theta}^{\mathrm{patch}}(\hat{\boldsymbol{u}}_i)$$

iBOT : Image BERT Pre-Training with Online Tokenizer, ICLR 2022.

iBOT (ICLR 2022)

Table 1: k-**NN and linear probing on ImageNet-1K**. † denotes using selective kernel. ‡ denotes pre-training on ImageNet-22K.

Method	Arch.	Par.	im/s	Epo.[1]	k-NN	Lin.
SSL big ResNets						
MoCov3	RN50	23	1237	1600	-	74.6
SwAV	RN50	23	1237	2400	65.7	75.3
DINO	RN50	23	1237	3200	67.5	75.3
BYOL	RN200w2	250	123	2000	73.9	79.6
SCLRv2	RN152w3†	794	46	2000	73.1	79.8
SSL Transformers						
MoCov3	ViT-S/16	21	1007	1200	-	73.4
MoCov3	ViT-B/16	85	312	1200	-	76.7
SwAV	ViT-S/16	21	1007	2400	66.3	73.5
DINO	ViT-S/16	21	1007	3200	74.5	77.0
DINO	ViT-B/16	85	312	1600	76.1	78.2
EsViT	Swin-T/7	28	726	1200	75.7	78.1
EsViT	Swin-T/14	28	593	1200	77.0	78.7
iBOT	ViT-S/16	21	1007	3200	75.2	77.9
iBOT	Swin-T/7	28	726	1200	75.3	78.6
iBOT	Swin-T/14	28	593	1200	76.2	79.3
iBOT	ViT-B/16	85	312	1600	77.1	79.5
iBOT	ViT-L/16	307	102	1200	**78.0**	81.0
iBOT ‡	ViT-L/16	307	102	200	72.9	**82.3**

Table 2: **Fine-tuning on ImageNet-1K.**

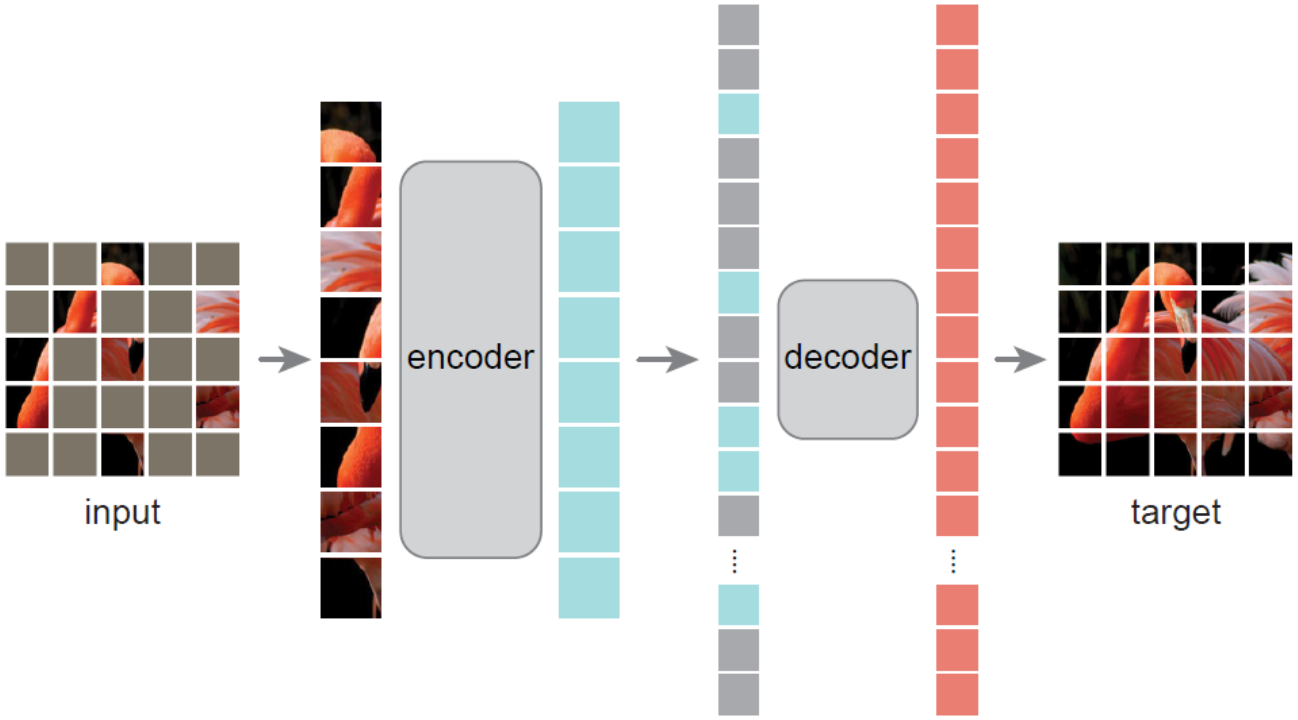
Method	Arch.	Epo.[1]	Acc.
Rand.	ViT-S/16	-	79.9
MoCov3	ViT-S/16	600	81.4
DINO	ViT-S/16	3200	82.0
iBOT	ViT-S/16	3200	**82.3**
Rand.	ViT-B/16	-	81.8
MoCov3	ViT-B/16	600	83.2
BEiT	ViT-B/16	800	83.4
DINO	ViT-B/16	1600	83.6
iBOT	ViT-B/16	1600	**84.0**
MoCov3	ViT-L/16	600	84.1
iBOT	ViT-L/16	1000	84.8
BEiT	ViT-L/16	800	**85.2**

Table 3: **Fine-tuning on ImageNet-1K.** Pre-training on ImageNet-22K.

Method	Arch.	Epo.[1]	Acc.
BEiT	ViT-B/16	150	83.7
iBOT	ViT-B/16	320	**84.4**
BEiT	ViT-L/16	150	86.0
iBOT	ViT-L/16	200	86.6
iBOT	ViT_{512}-L/16	200	**87.8**

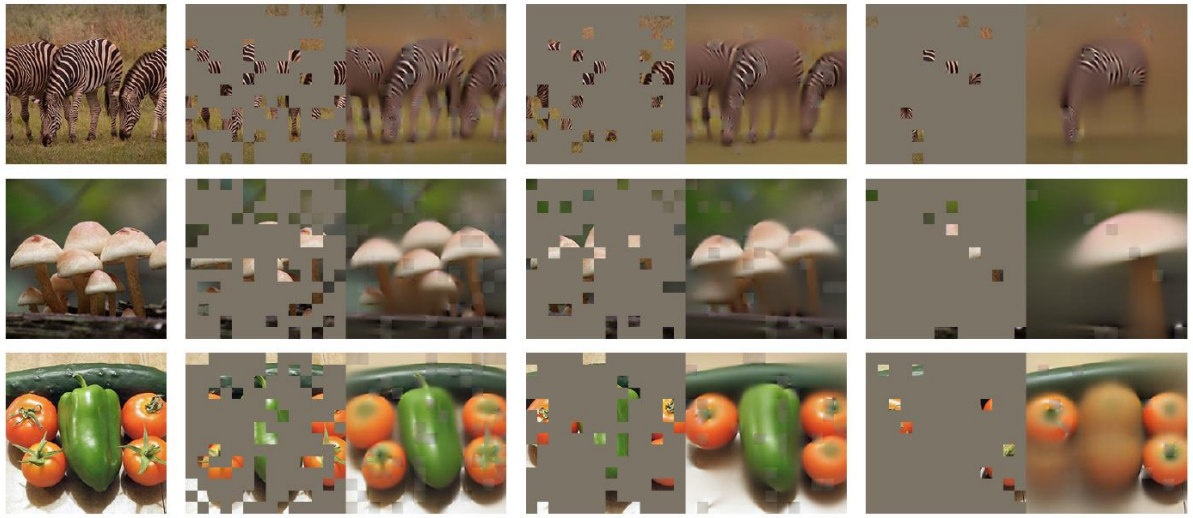
Masked Autoencoders (MAE)(CVPR'22)

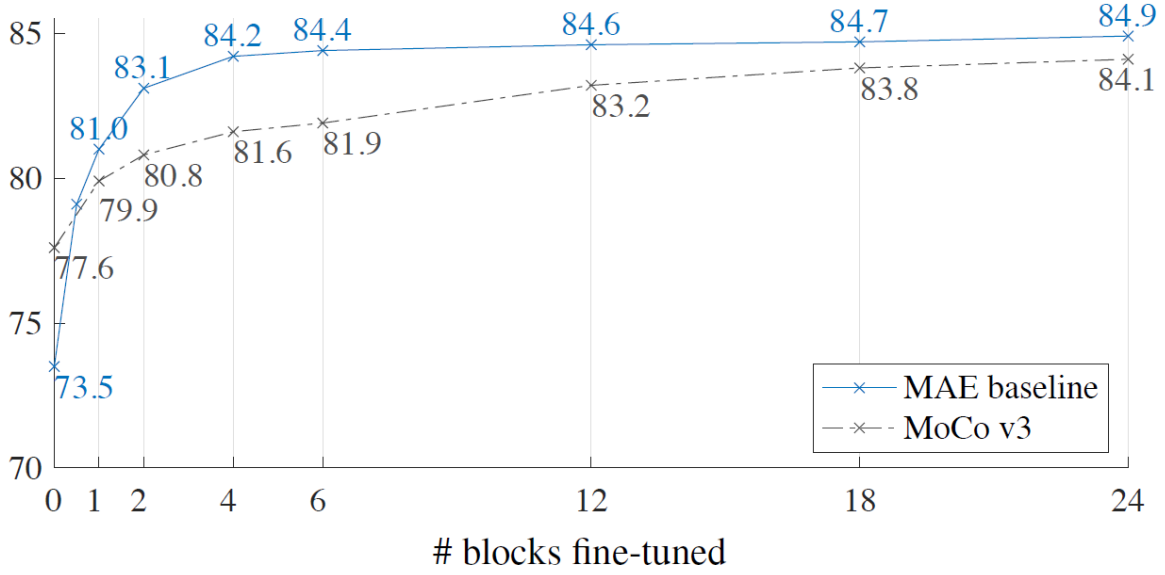
- Asymmetric encoder-decoder
- Masking a high proportion of the input image (e.g., 75%)



Masked Autoencoders Are Scalable Vision Learners, Arxiv 2021

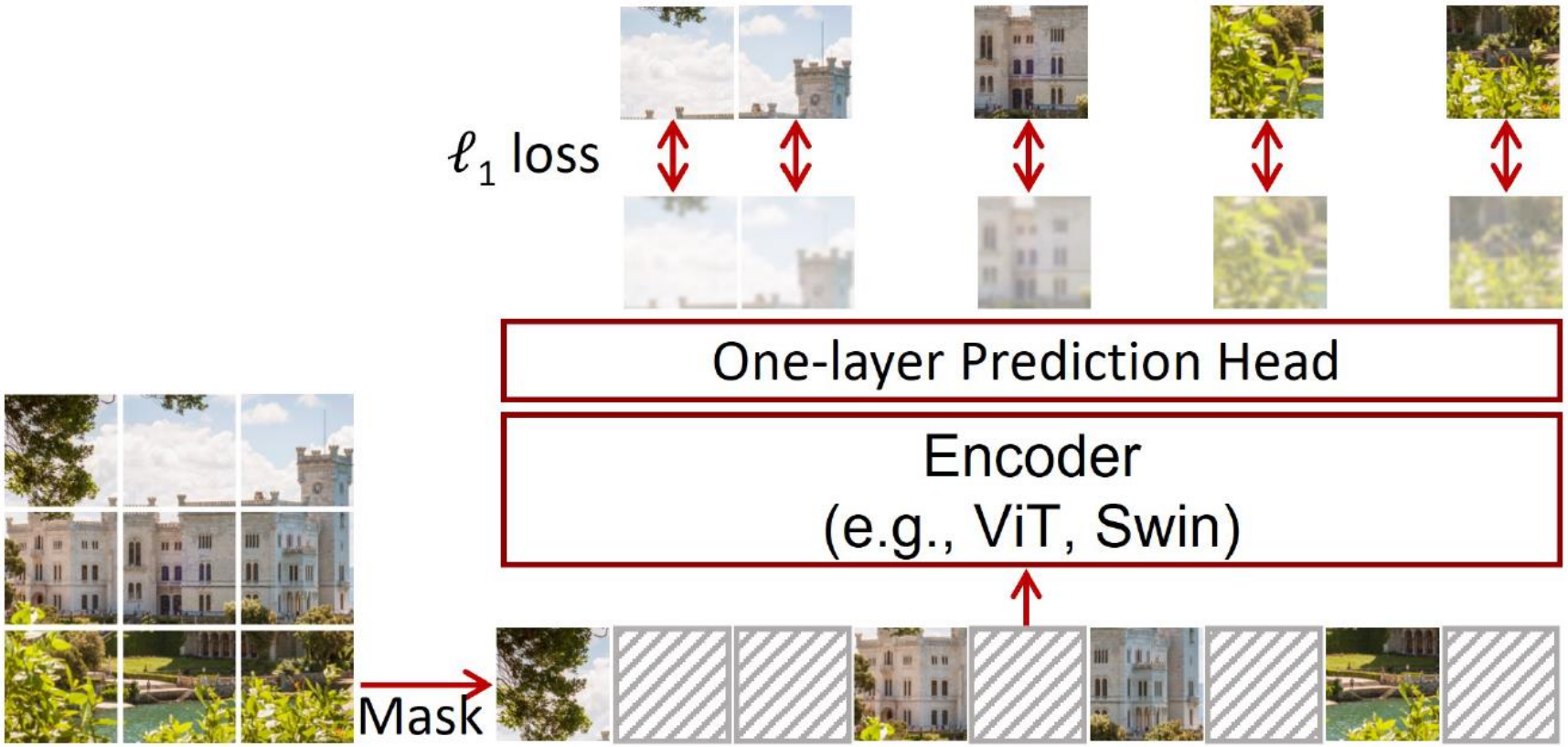
MAE





method	pre-train data	ViT-B	ViT-L	ViT-H	ViT-H_{448}
scratch, our impl.	-	82.3	82.6	83.1	-
DINO [5]	IN1K	82.8	-	-	-
MoCo v3 [9]	IN1K	83.2	84.1	-	-
BEiT [2]	IN1K+DALLE	83.2	85.2	-	-
MAE	IN1K	83.6	85.9	86.9	**87.8**

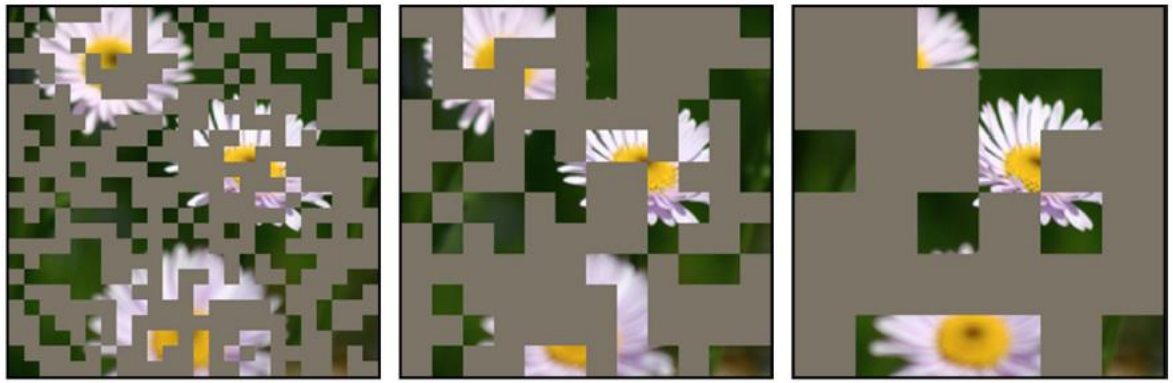
SimMIM (CVPR'22)



SimMIM: A Simple Framework for Masked Image Modeling, Arxiv 2022.

SimMIM

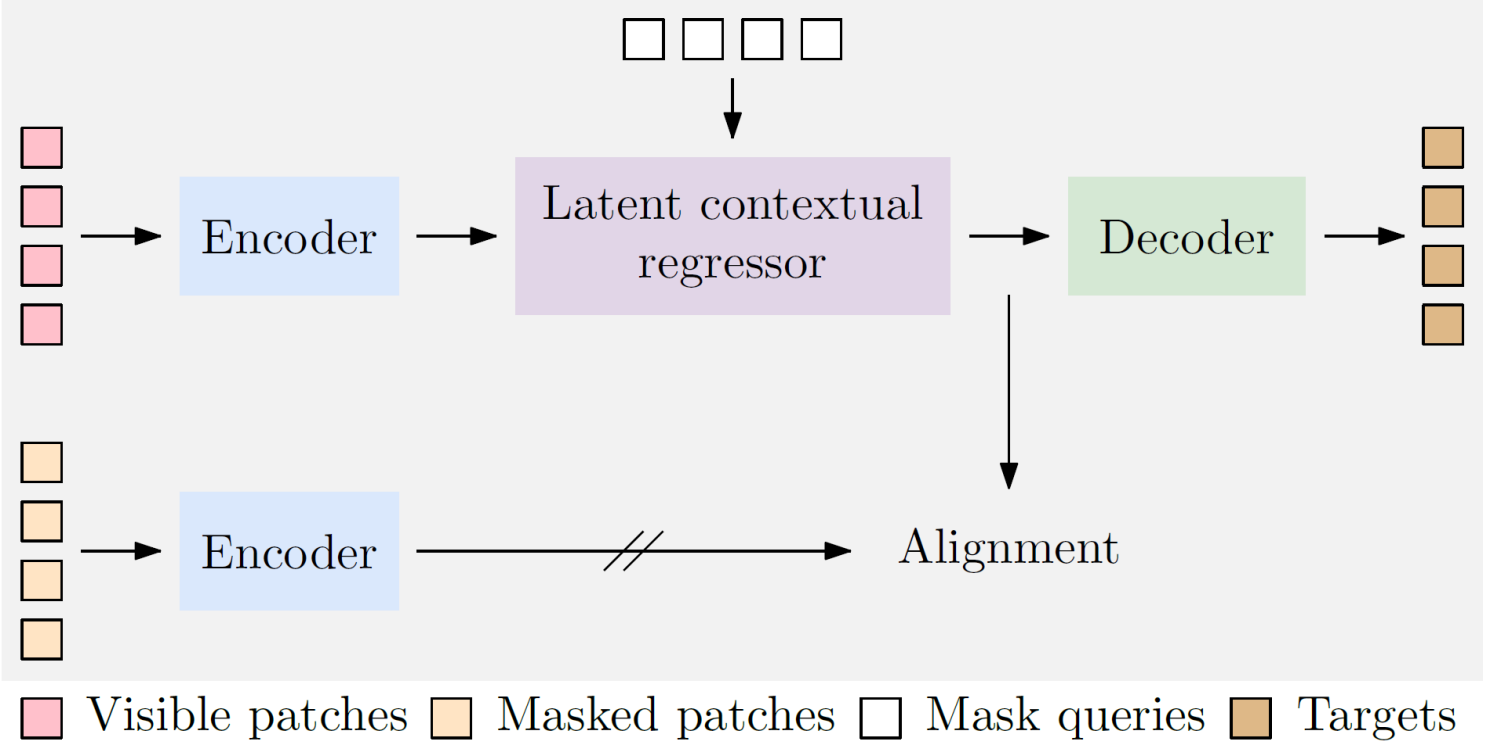
- Moderately large masked patch size (e.g. 32)
- Predicting raw pixels of RGB values
- Light prediction head (e.g., a linear layer)



Random (8) Random (16) Random (32)

Methods	Input Size	Fine-tuning Top-1 acc (%)	Linear eval Top-1 acc (%)	Pre-training costs
Sup. baseline [44]	224^2	81.8	-	-
DINO [5]	224^2	82.8	78.2	2.0×
MoCo v3 [9]	224^2	83.2	76.7	1.8×
ViT [15]	384^2	79.9	-	∼4.0×
BEiT [1]	224^2	83.2	56.7	1.5×†
Ours	224^2	**83.8**	56.7	1.0×

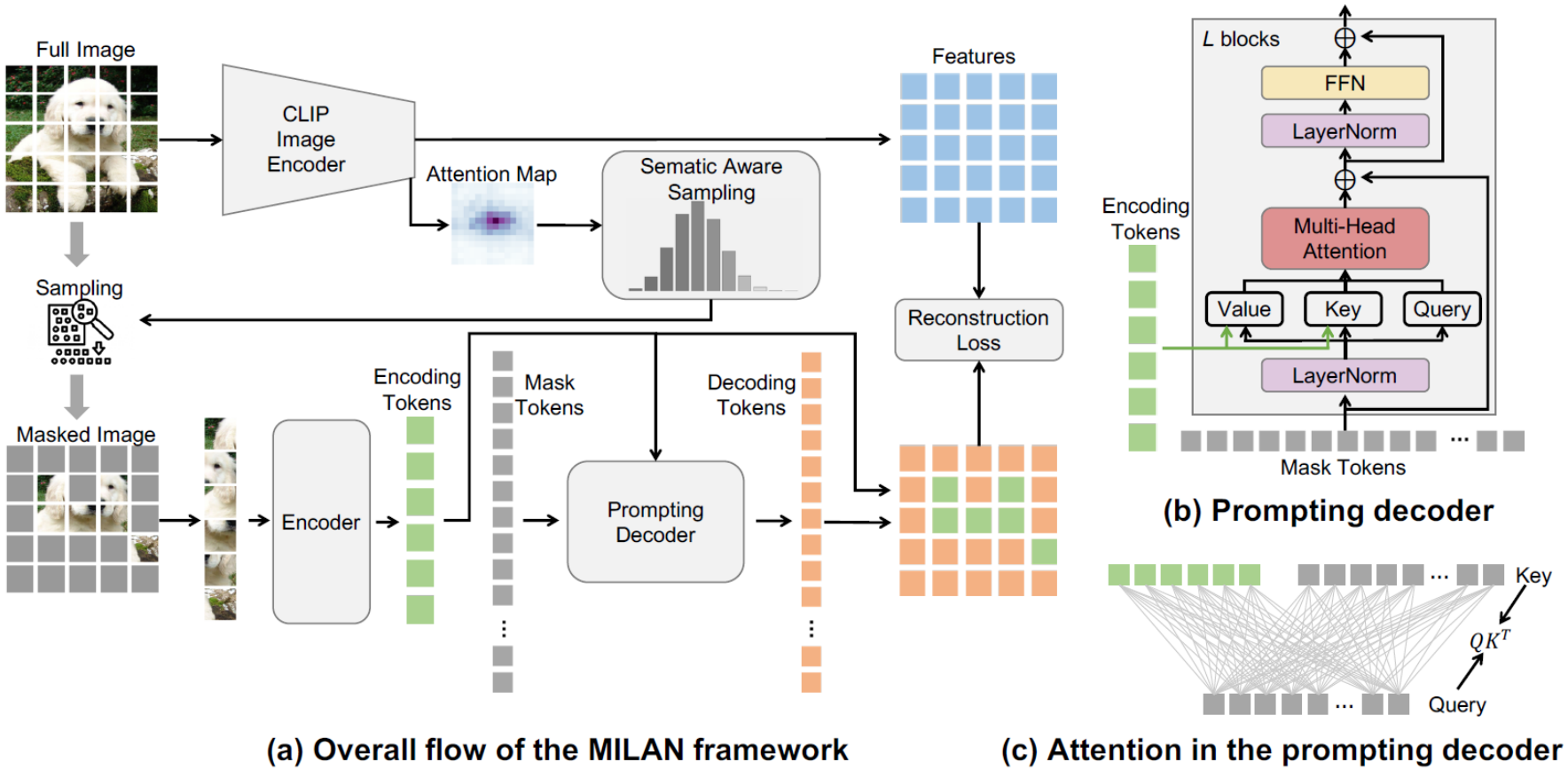
Context Autoencoder (CAE)



$$\ell_y(\mathbf{Y}_m, \bar{\mathbf{Y}}_m) + \lambda\, \ell_z(\mathbf{Z}_m, \bar{\mathbf{Z}}_m)$$

Context Autoencoder for Self-Supervised Representation Learning, Arxiv 2022.

MILAN (NeurIPS'22)



MILAN (NeurIPS'22)

Method	Training data	Resolution	ViT-B/16		ViT-L/16	
			Epochs	Top-1 (%)	Epochs	Top-1 (%)
Supervised [50]	IN1K	224	-	83.8 (+1.6)	-	84.9 (+1.8)
contrastive or clustering based						
MoCov3 [11]	IN1K	224	300	83.2 (+2.2)	300	84.1 (+2.6)
DINO [6]	IN1K	224	400	82.8 (+2.6)	-	-
iBOT [69]	IN22K+IN1K	224	320	84.4 (+1.0)	200	86.3 (+0.4)
reconstruction based						
BEiT [3]	DALLE250M+IN22K+IN1K	224	150	83.7 (+1.7)	150	86.0 (+0.7)
mc-BEiT [33]	OpenImages9M+IN1K	224	800	84.1 (+1.3)	800	85.6 (+1.1)
PeCo [18]	IN1K	224	800	84.5 (+0.9)	800	86.5 (+0.2)
SimMIM [61]	IN1K	224	800	83.8 (+1.6)	-	-
MaskFeat [56]	IN1K	224	1600	84.0 (+1.4)	1600	85.7 (+1.0)
data2vec [2]	IN1K	224	800	84.2 (+1.2)	1600	86.6 (+0.1)
CAE [9]	IN1K	224	800	83.6 (+1.8)	-	-
MAE [25]	IN1K	224	1600	83.6 (+1.8)	1600	85.9 (+0.8)
language-image pretraining based						
CLIP [43]	OpenAI400M+IN1K	224	-	82.1 (+3.3)	-	85.3 (+1.4)
MVP [57]	OpenAI400M+IN1K	224	300	84.4 (+1.0)	300	86.3 (+0.4)
MILAN	OpenAI400M+IN1K	224	400	**85.4**	400	**86.7**
Supervised [19]	JFT300M+IN1K	384	90	84.2 (+2.2)	90	87.1 (+0.2)
BEiT [3]	DALLE250M+IN1K	384	800	84.6 (+1.8)	800	86.3 (+1.0)
SWAG [47]	IG3.6B+IN1K	384	2	85.3 (+1.1)	-	-
MILAN	OpenAI400M+IN1K	384	400	**86.4**	400	**87.3**

Summary

- Before 2020, Self-supervised Contextual Modeling (百家争鸣)
- Contrastive learning-based self-supervised learning
  - Before 2021, CNN-based
  - 2021, transformer-based
- Masked Image Modeling
  - 2021, transformer-based