Self-supervised Learning for Visual Recognition

Hamed Pirsiavash
University of Maryland, Baltimore County

UMBC
AN HONORS UNIVERSITY IN MARYLAND

Significant progress in recognition due to large **annotated** datasets

14 million images

450 hours of video

10 million images

1.7 million question/answers

Self supervised learning

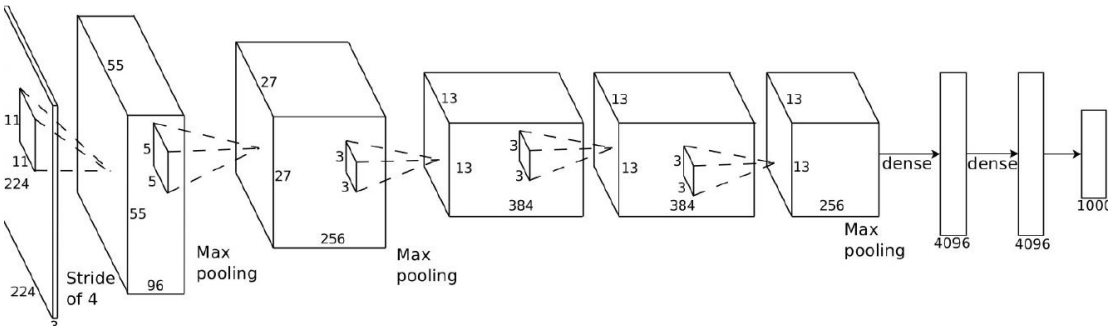
Input

Output

Zhang *et al*. ECCV'16

Supervised Learning (classification)

Input image

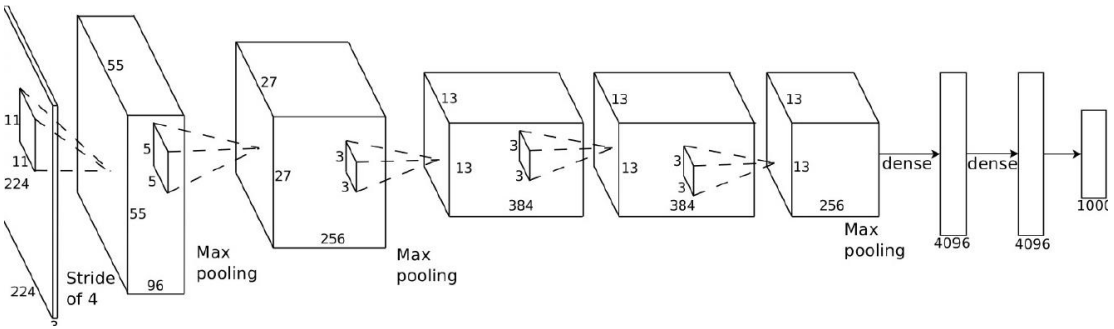


Chair: 0
Dog: **1**
Car: 0
.
.
.

Label

Supervised Learning (classification)

Input image

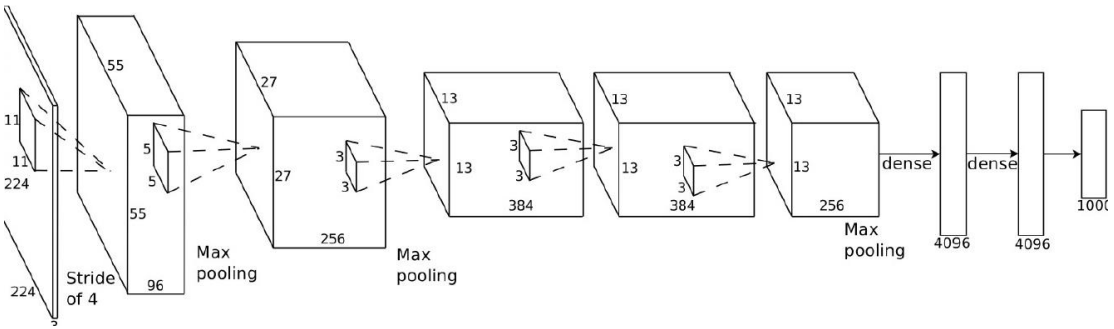


Chair: 0
Dog: **1**
Car: 0
.
.
.

Label

Supervised Learning (classification)

Input image

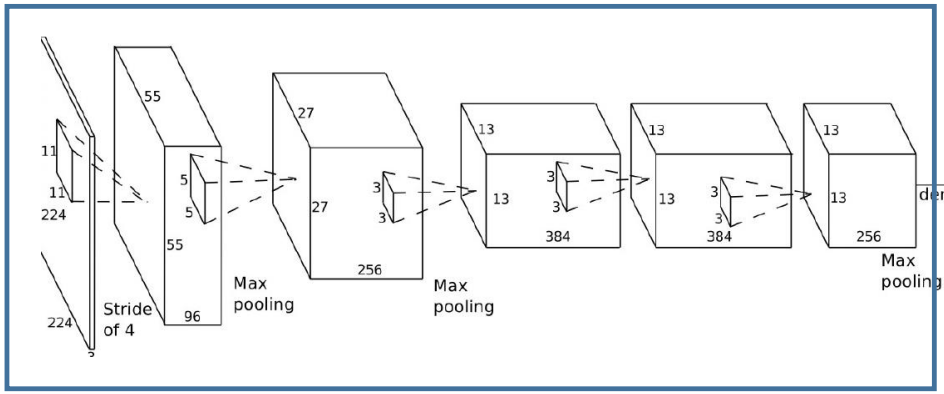


Chair: **1**
Dog: 0
Car: 0
.
.
.

Label

Supervised Learning (classification)

Input image



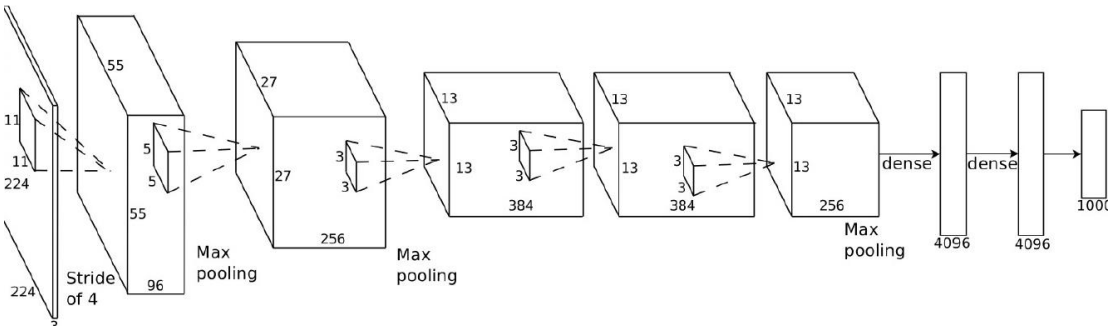
Chair: **1**
Dog: 0
Car: 0
.
.
.

Label

Transfer to another task

Supervised Learning (counting)

Input image



Chair: 0
Dog: **2**
Car: 0
.
.
.

Label

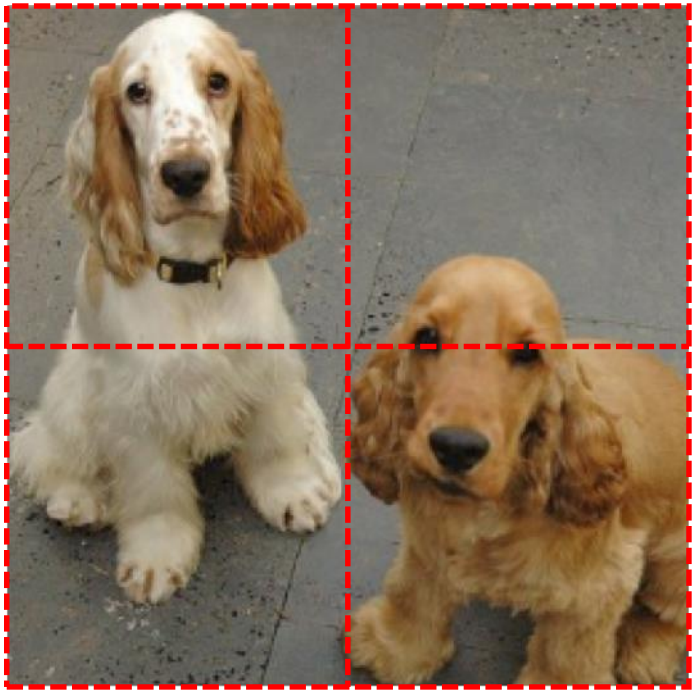
Inference on counting network

nose	2
eyes	4
paws	6
head	2

Constraint in the output

nose	2
eyes	4
paws	6
head	2

nose	1
eyes	2
paws	0
head	1

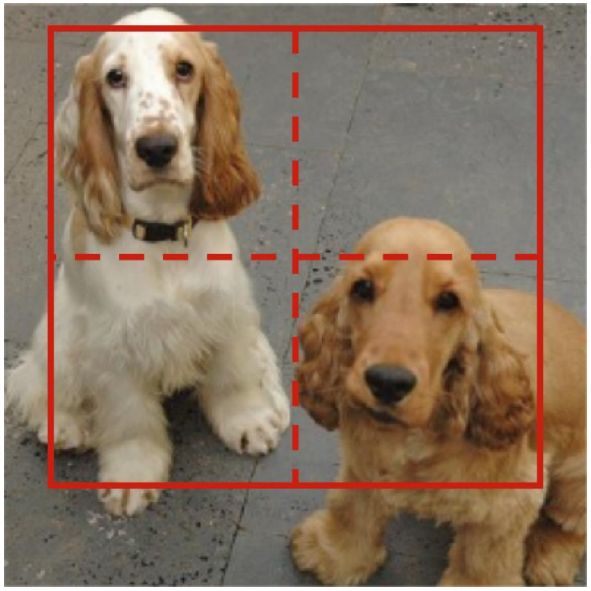


nose	0
eyes	1
paws	0
head	0.5

nose	0
eyes	0
paws	3
head	0

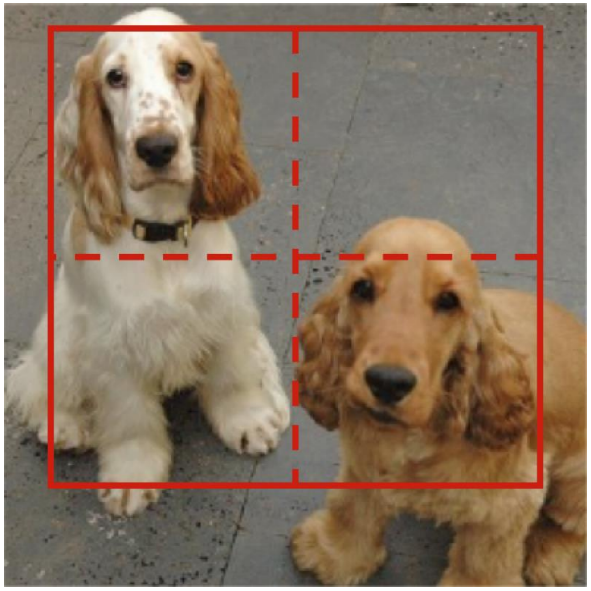
nose	1
eyes	1
paws	3
head	0.5

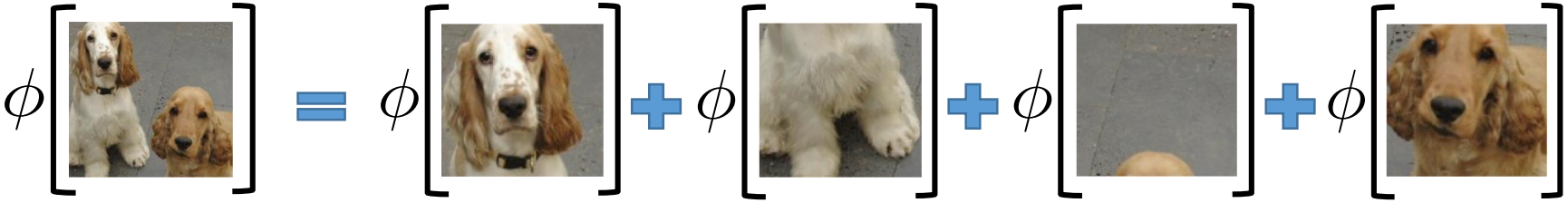
Constraint in the output



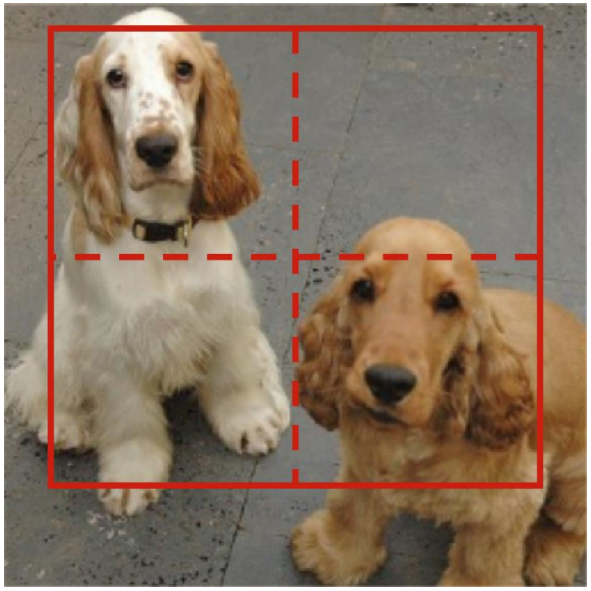
ϕ  =

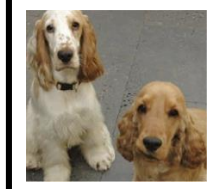
Constraint in the output

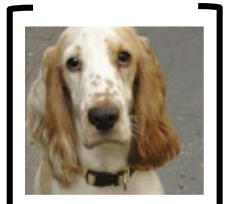




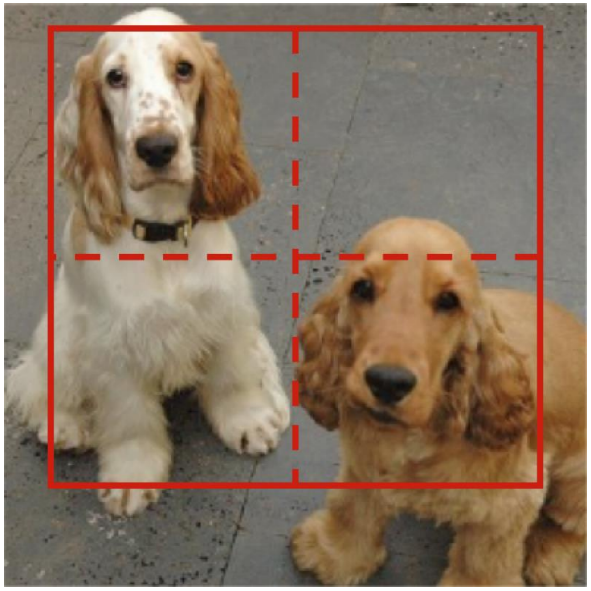
Two constraints in learning

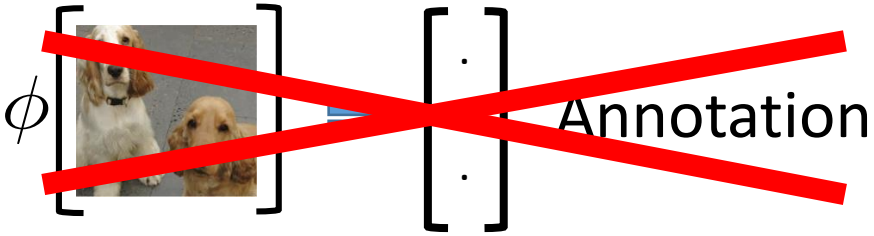


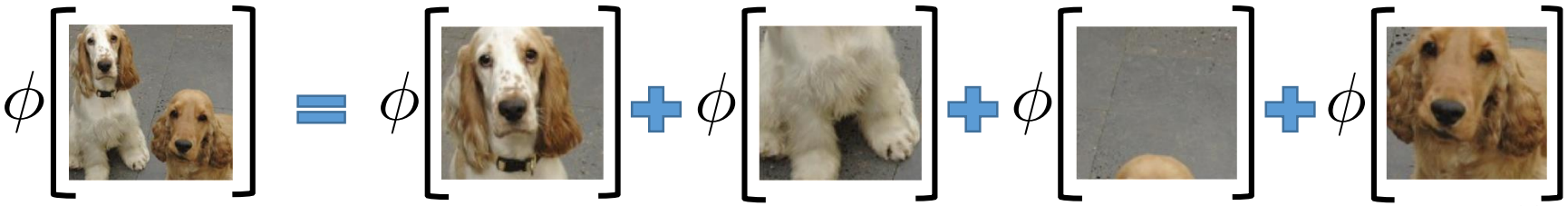
ϕ  = $\begin{bmatrix} \cdot \\ \cdot \\ \cdot \end{bmatrix}$ Annotation

ϕ  = ϕ  + ϕ  + ϕ  + ϕ 

Two constraints in learning

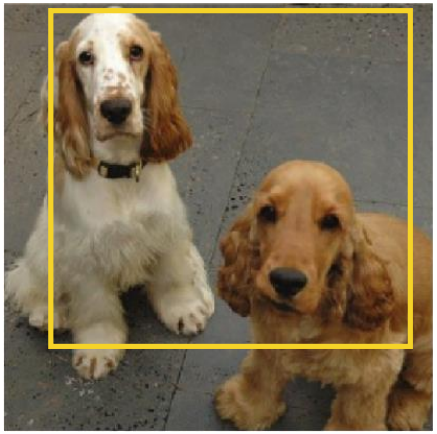




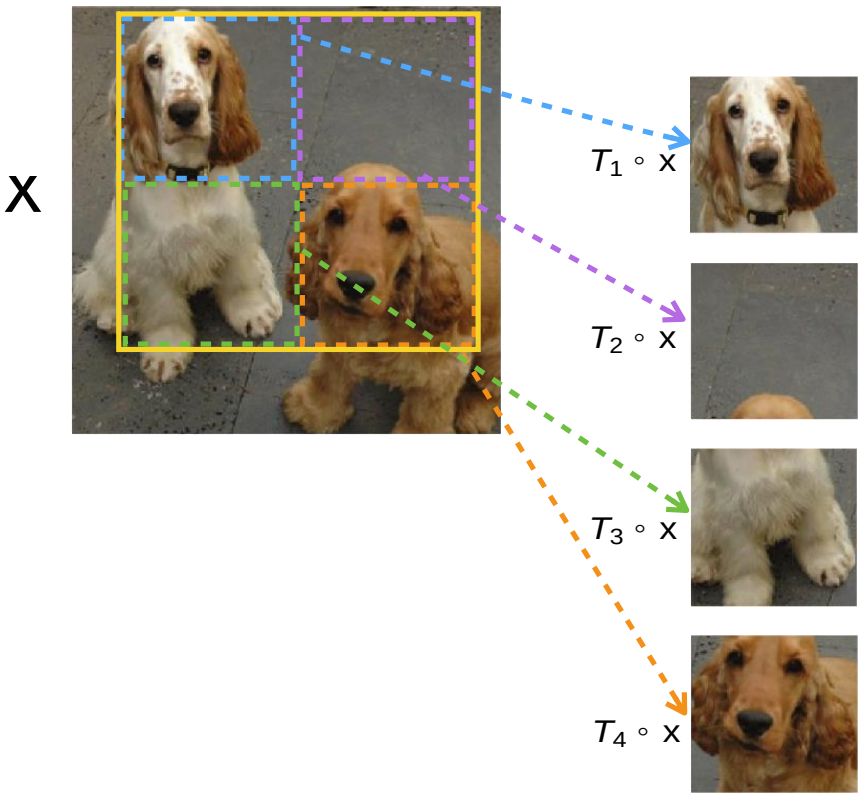


Self supervised learning

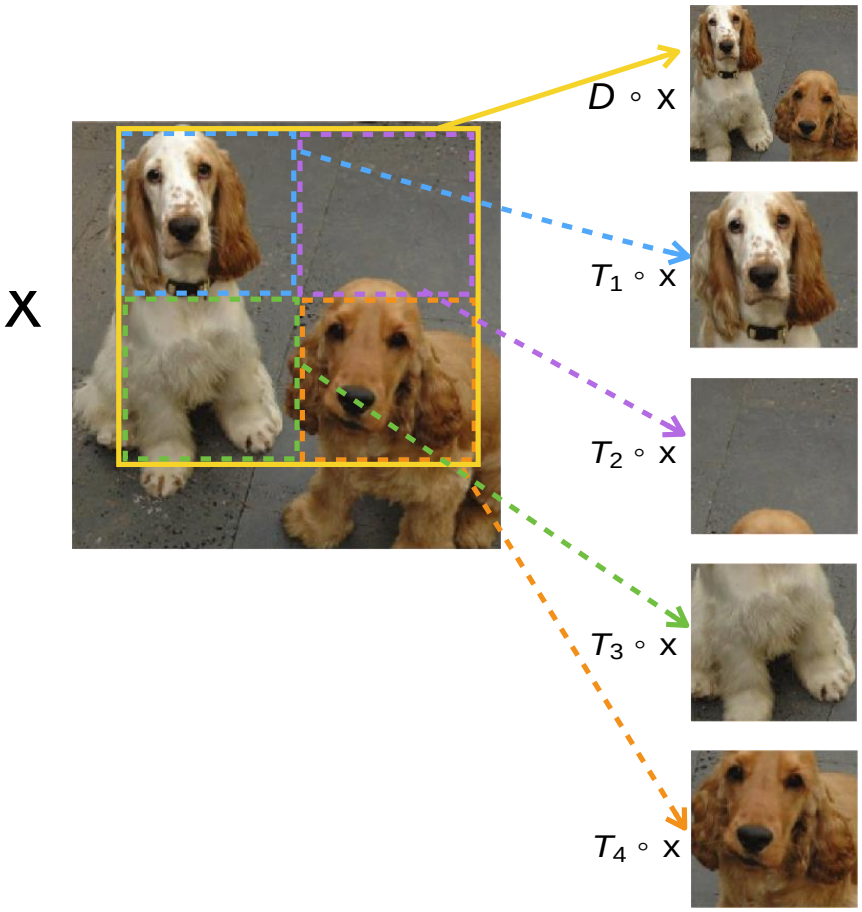
x



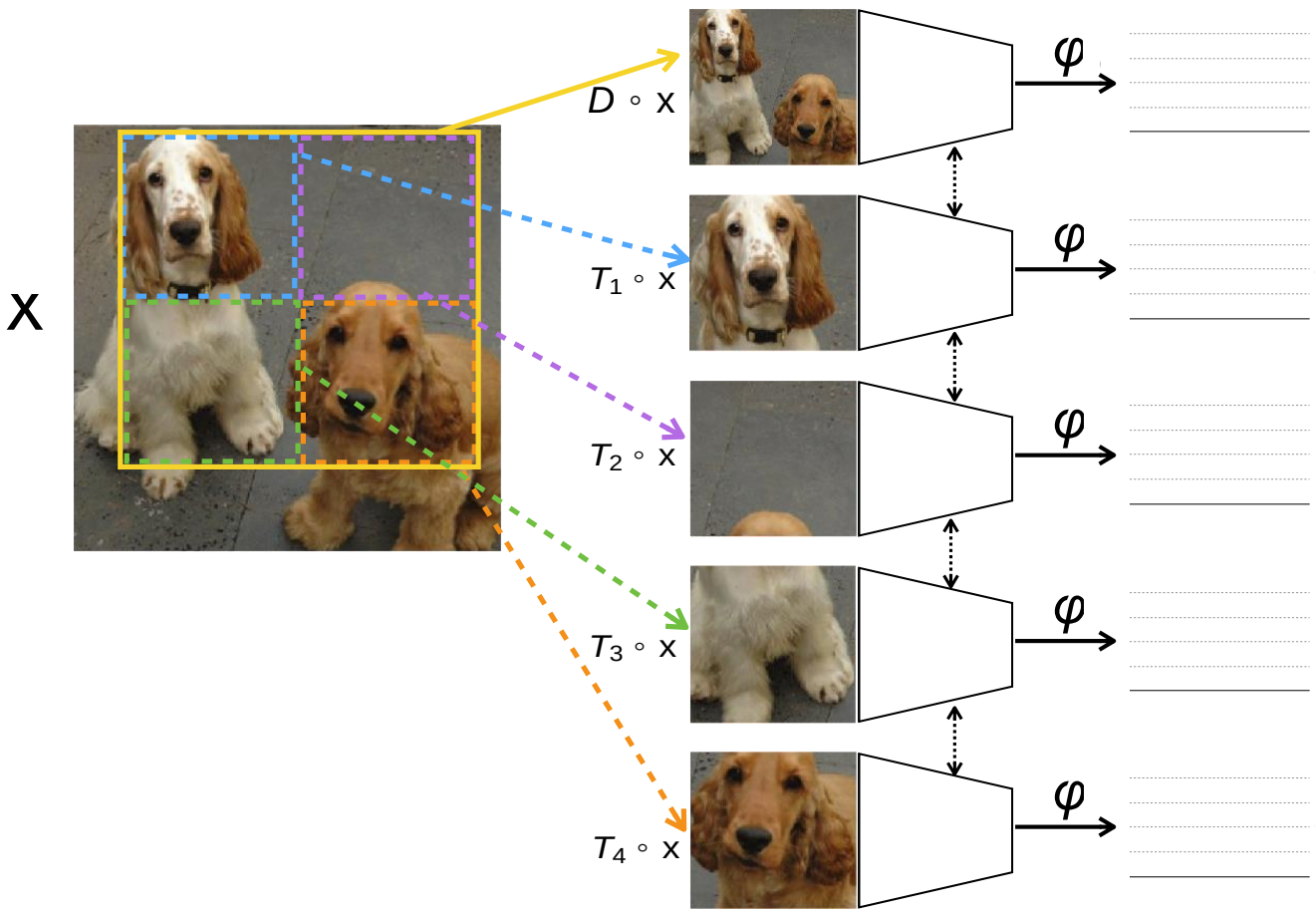
Self supervised learning



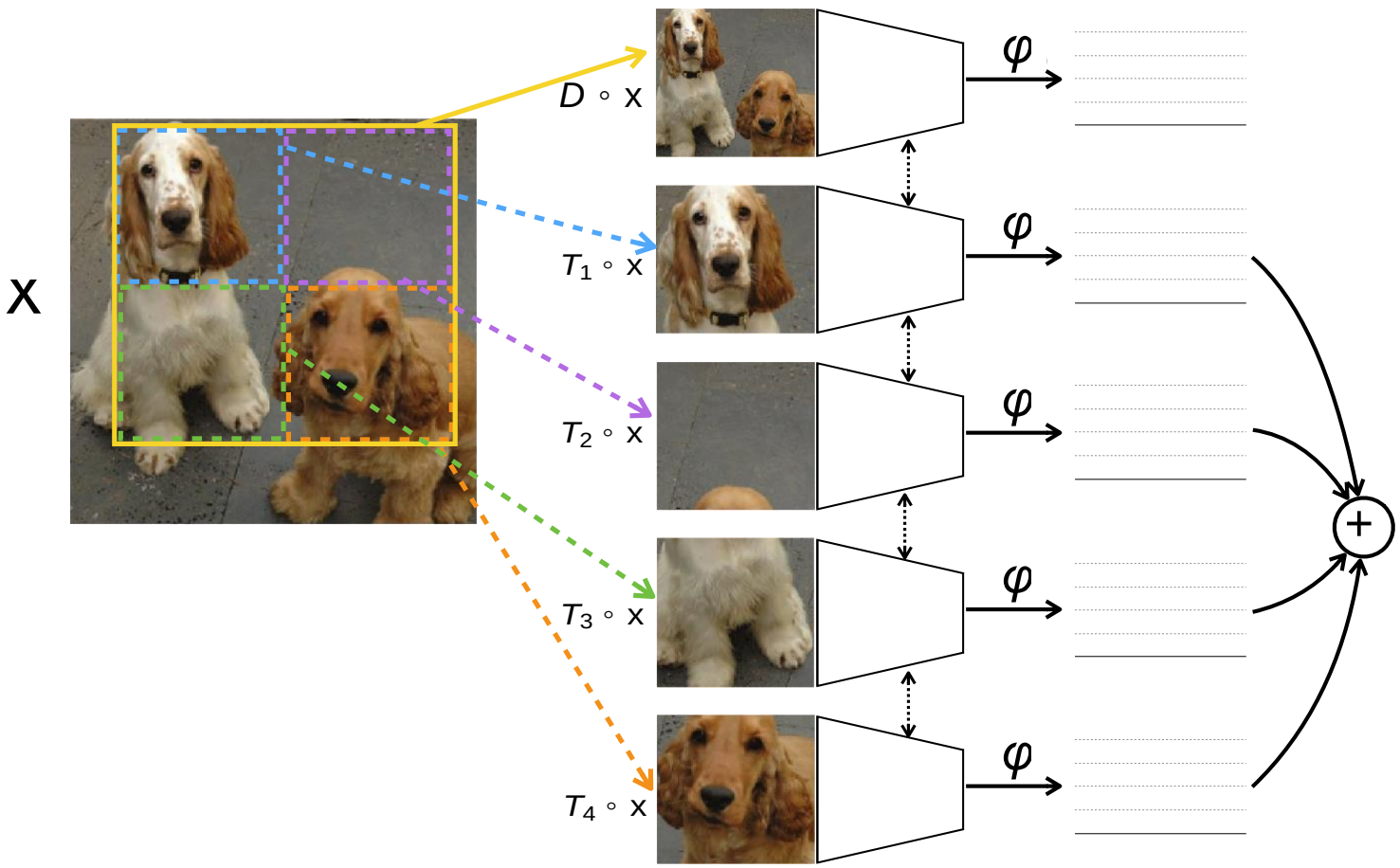
Self supervised learning



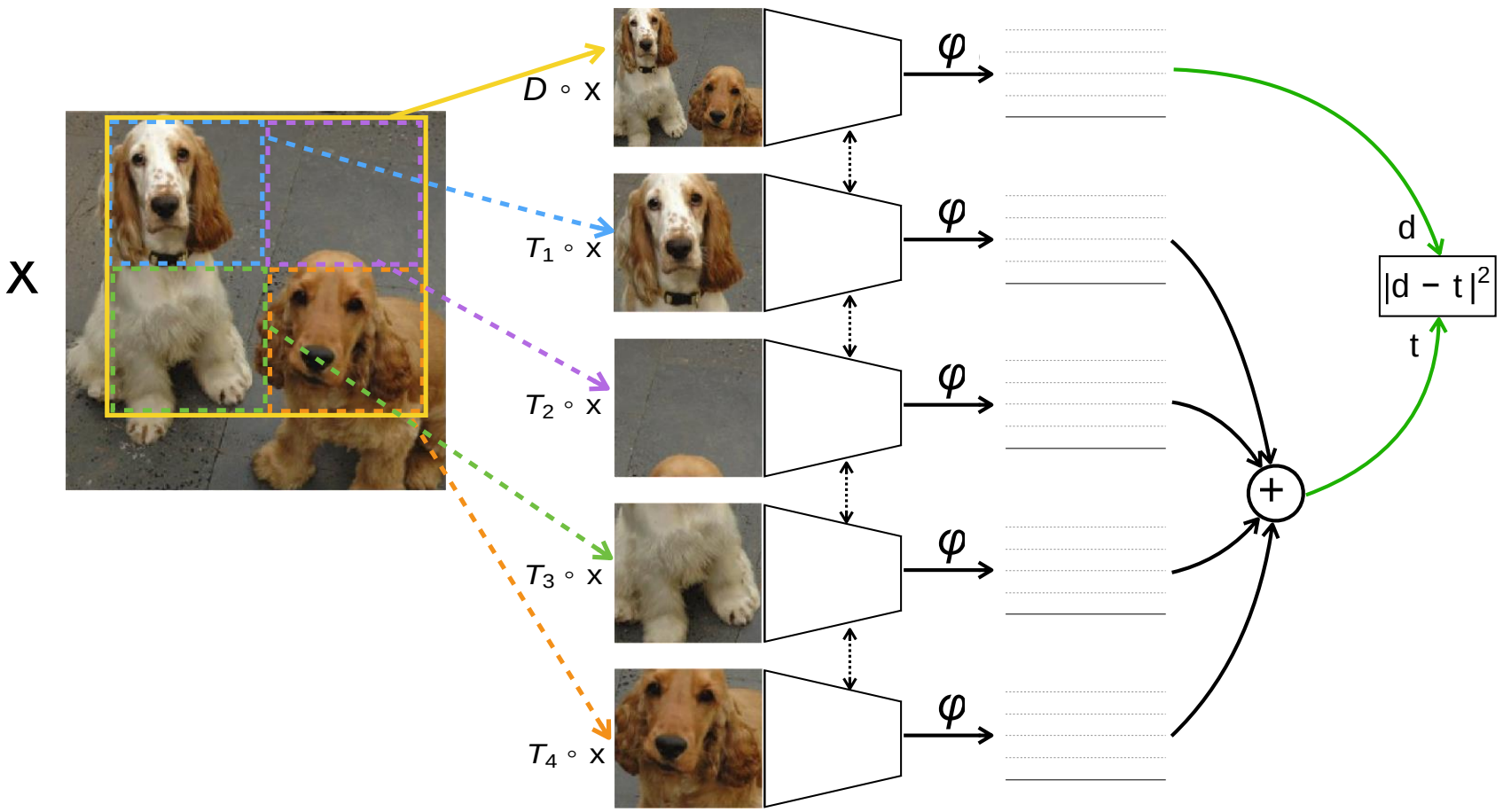
Self supervised learning



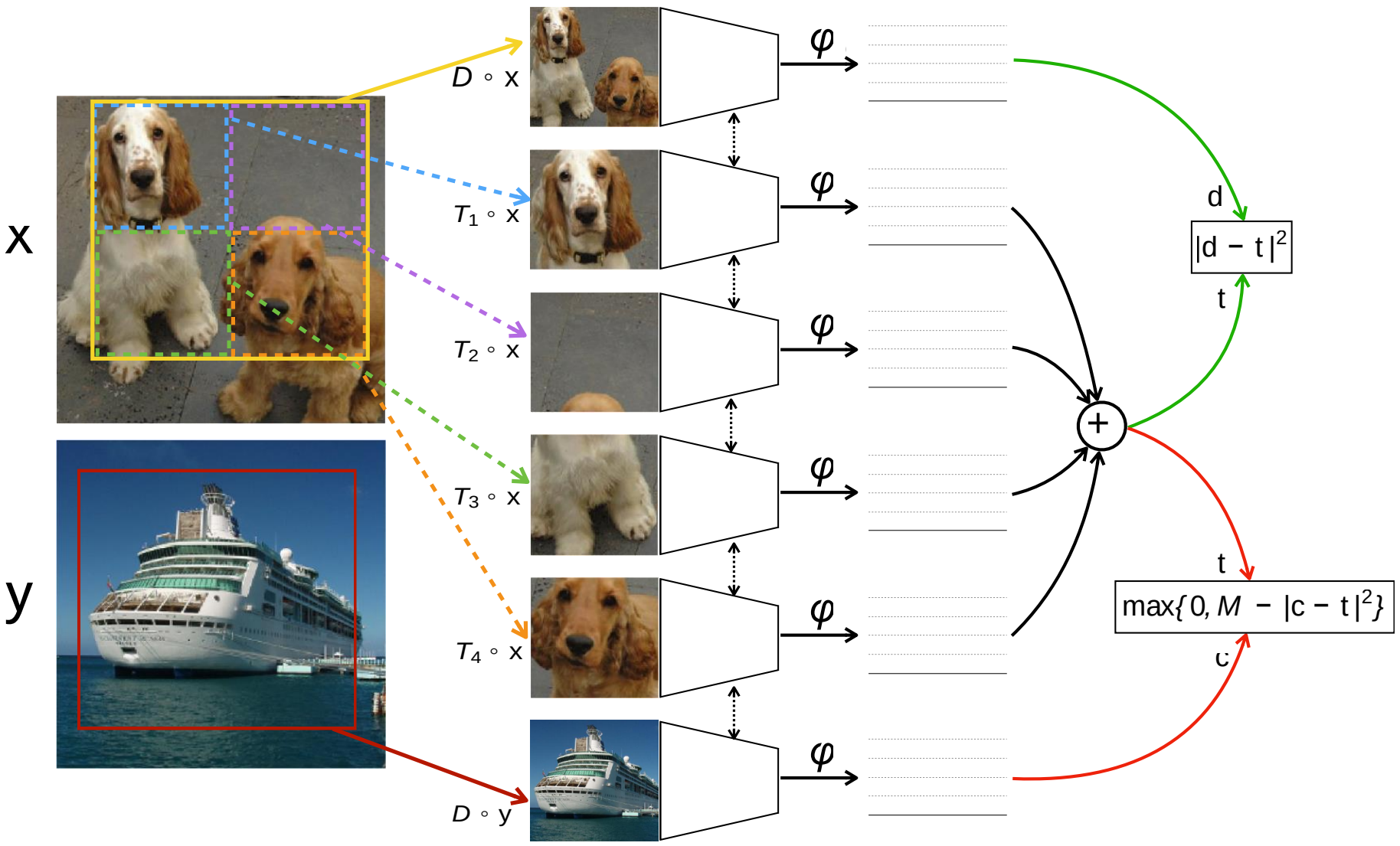
Self supervised learning



Self supervised learning



Self supervised learning



Images with largest activation

Trained on **ImageNet** without annotation

Unit **1**

Unit **2**

Unit **3**

Images with largest activation

Trained on **COCO** without annotation

Unit 1

Unit 2

Unit 3

Nearest neighbor search

Trained on **ImageNet** without annotation

query

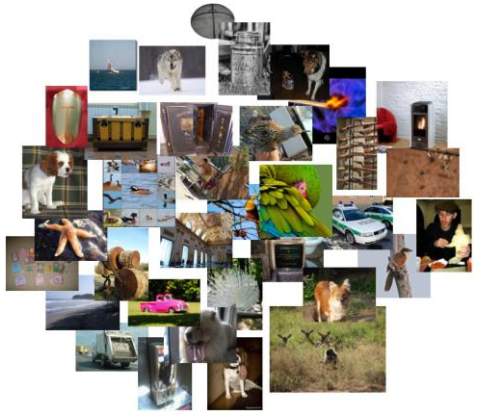
retrieved

Nearest neighbor search

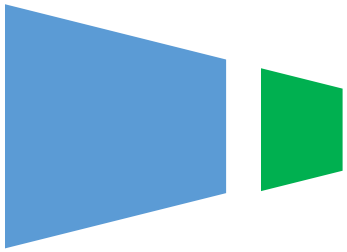
Trained on **COCO** without annotation

query retrieved

Dataset (no labels)



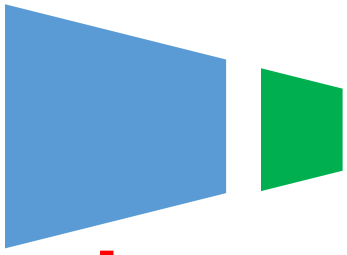
Feature network
(e.g., AlexNet)



Pretext task
(e.g., counting)

Dataset (no labels)

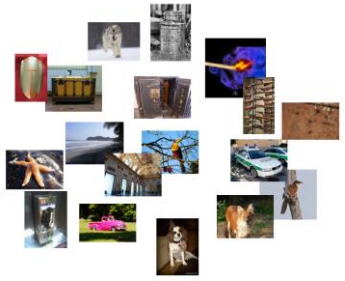
Feature network
(e.g., AlexNet)



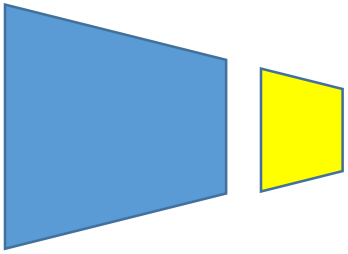
Pretext task
(e.g., counting)

Fine-tuning

Dataset (with labels)



Feature network
(e.g., AlexNet)

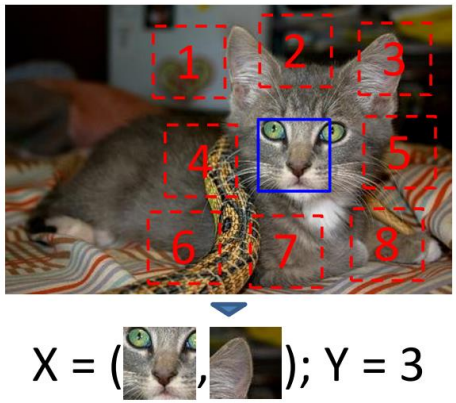


Target task
(e.g., object detection)

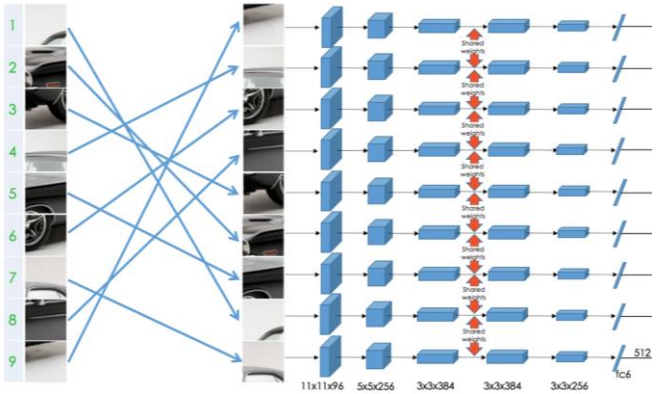
Results on transfer learning

Fine-tuning on PASCAL VOC07

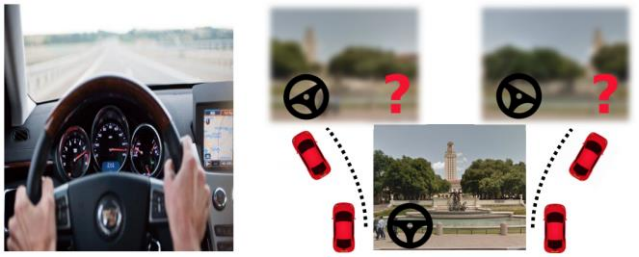
Method	**Class.**	**Det.**	**Segm.**
Supervised	79.9	57.1	48.0
Random	53.3	43.4	19.8
Sound	54.4	44.0	-
Video	63.1	47.2	-
Split-Brain	67.1	46.7	36.0
Watching-Objects	61.0	52.2	-
Jigsaw(new version)	**67.6**	**53.2**	**37.6**
Counting (Ours)	**67.7**	**52.4**	**36.6**



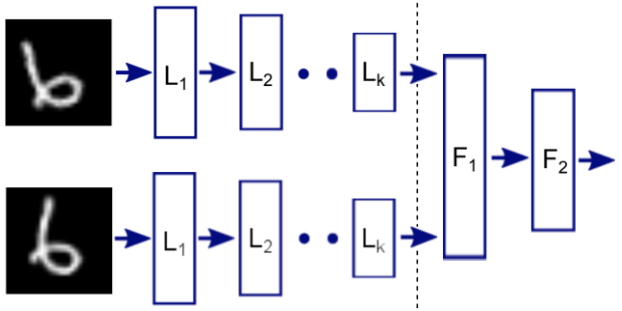
Doerch et al. ICCV'15



Noroozi and Favaro ECCV'16

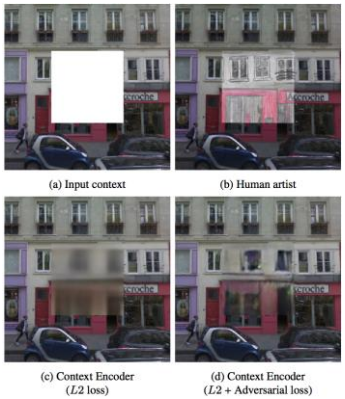


Jayaraman and Grauman ICCV'15

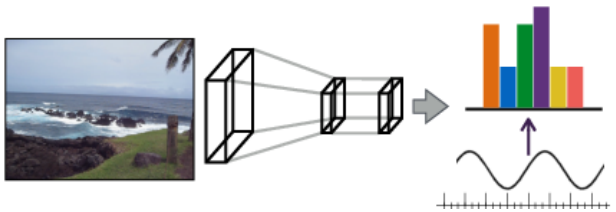


Agrawal et al. ICCV'15

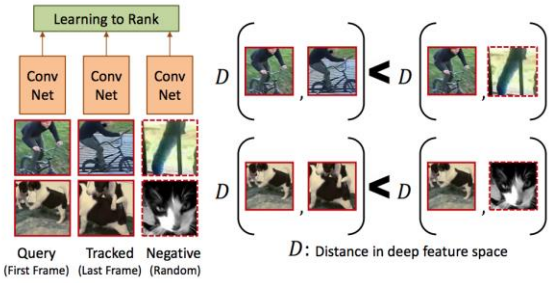
Zhang *et al*. ECCV'16



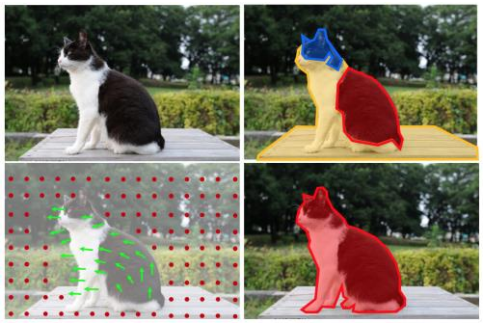
Pathak *et al*. CVPR'16



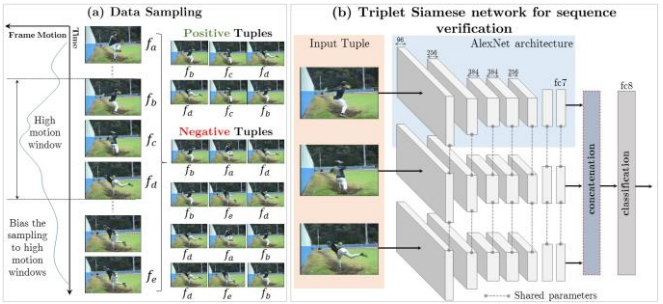
Owens et al. ECCV'16



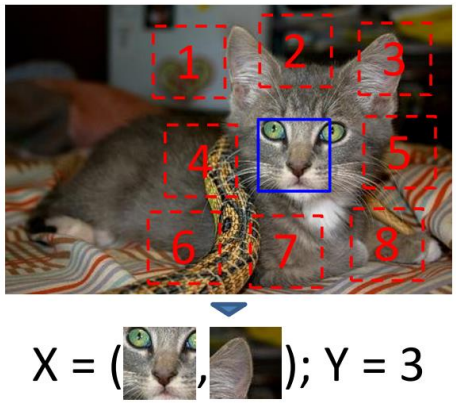
Wang and Gupta ICCV'15



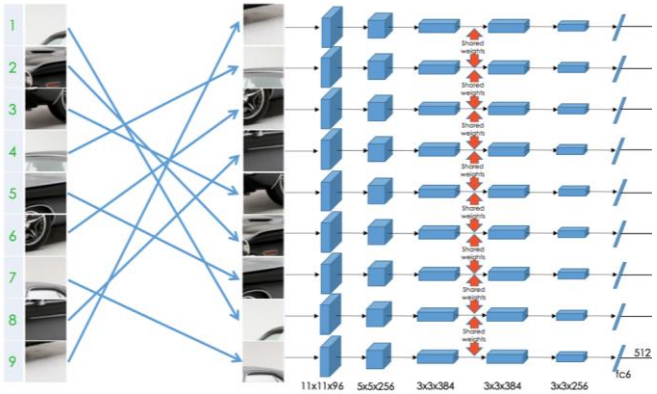
Pathak et al. CVPR'17



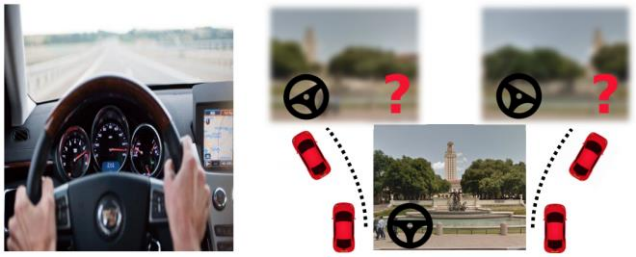
Mirsa et al.ECCV'16



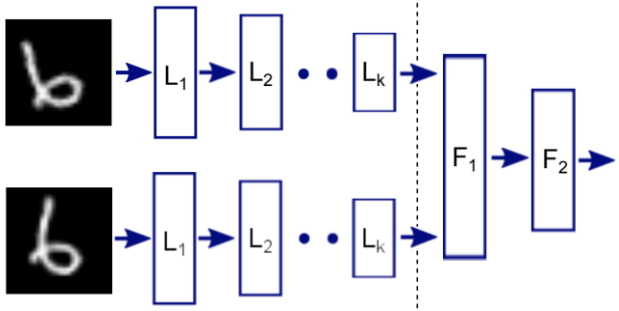
Doerch et al. ICCV'15



Noroozi and Favaro ECCV'16

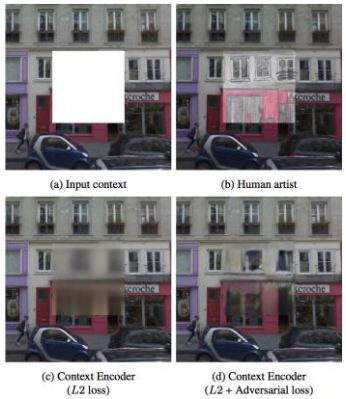


Jayaraman and Grauman ICCV'15

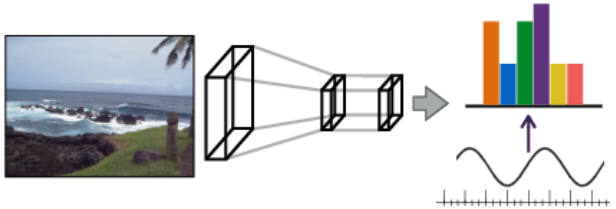


Agrawal et al. ICCV'15

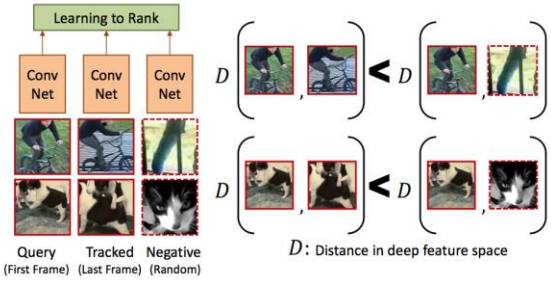
Zhang *et al*. ECCV'16



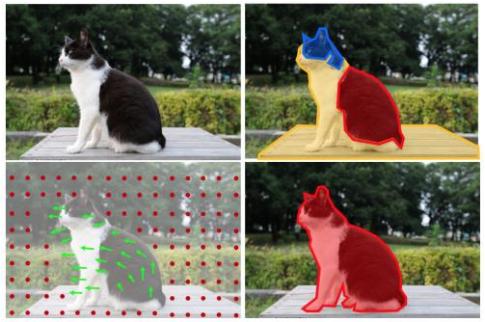
Pathak *et al*. CVPR'16



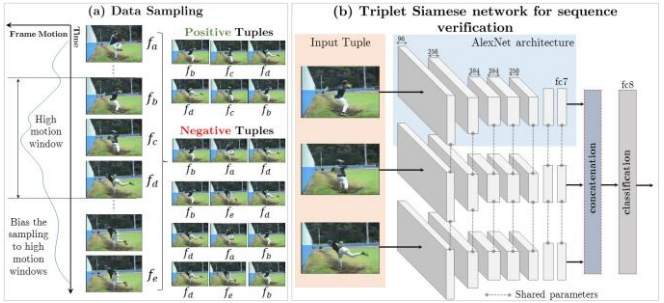
Owens et al. ECCV'16



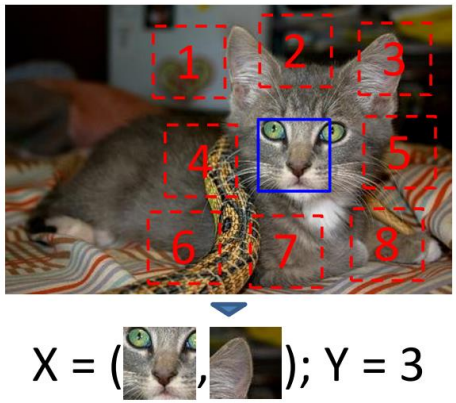
Wang and Gupta ICCV'15



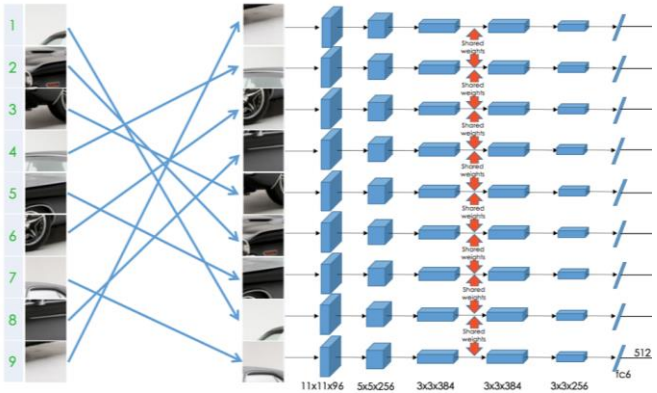
Pathak et al. CVPR'17



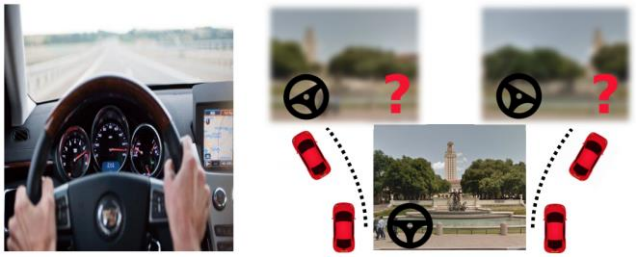
Mirsa et al.ECCV'16



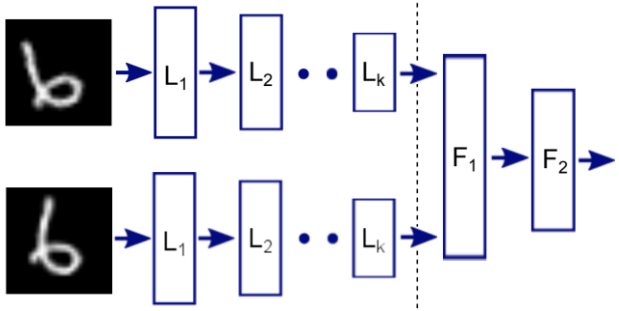
Doerch et al. ICCV'15



Noroozi and Favaro ECCV'16

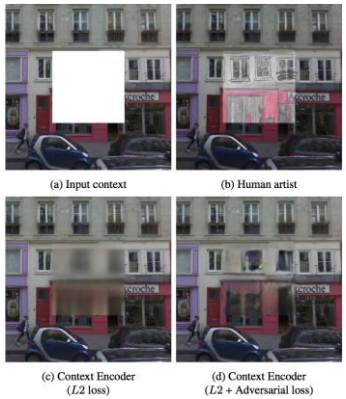


Jayaraman and Grauman ICCV'15

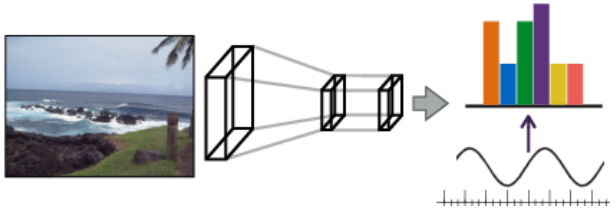


Agrawal et al. ICCV'15

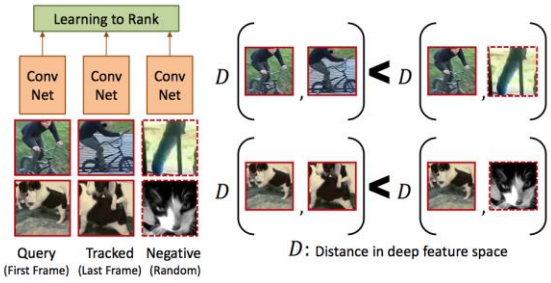
Zhang *et al*. ECCV'16



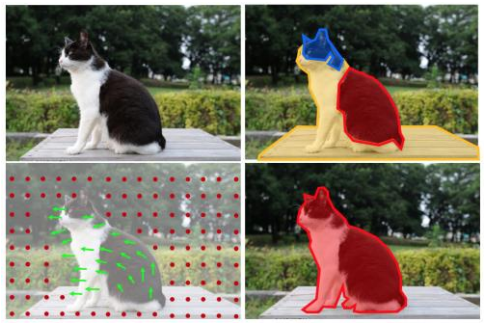
Pathak *et al*. CVPR'16



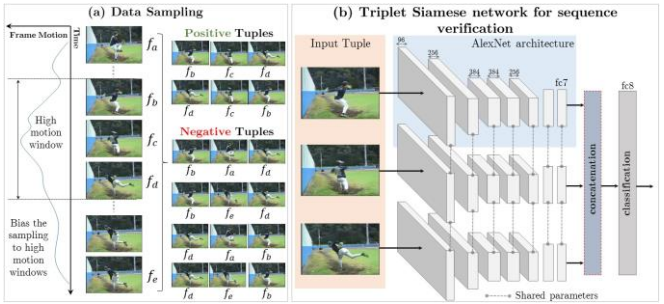
Owens et al. ECCV'16



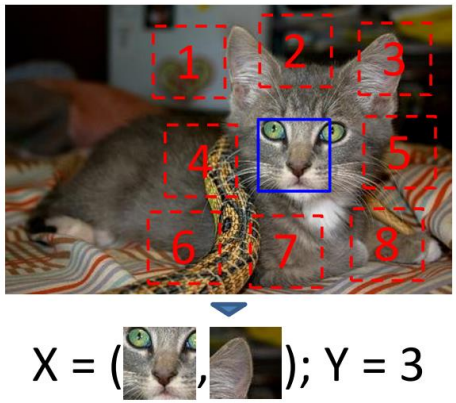
Wang and Gupta ICCV'15



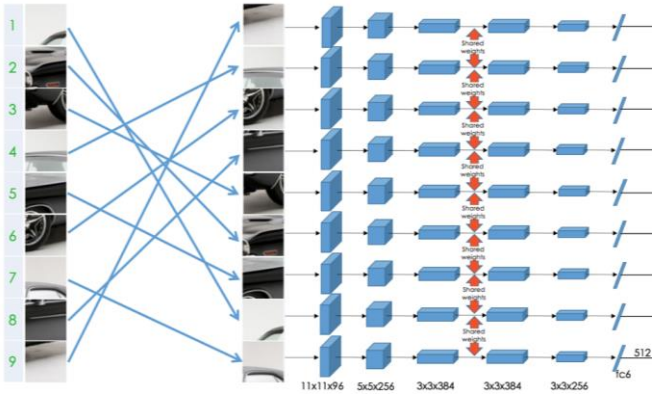
Pathak et al. CVPR'17



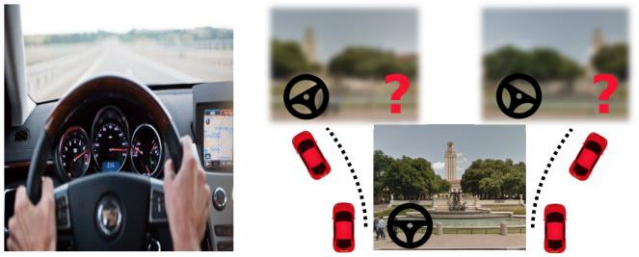
Mirsa et al.ECCV'16



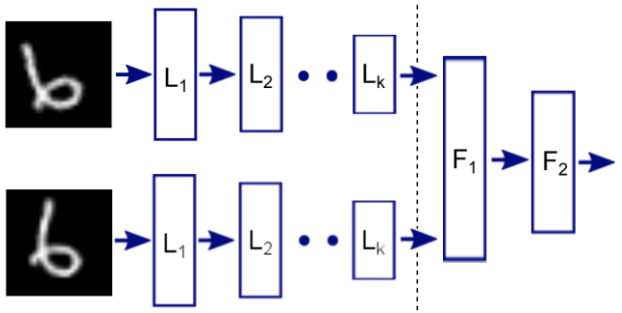
Doerch et al. ICCV'15



Noroozi and Favaro ECCV'16

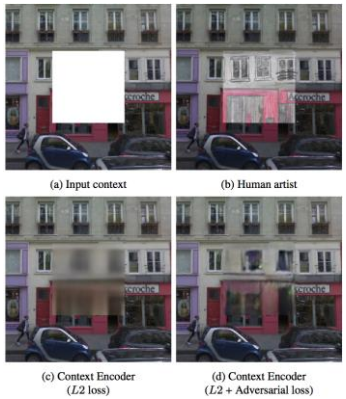


Jayaraman and Grauman ICCV'15

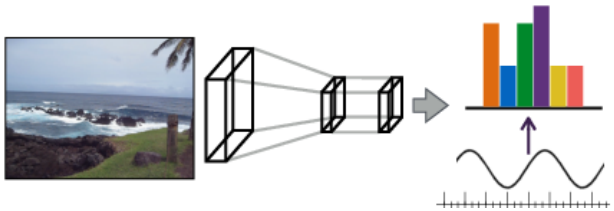


Agrawal et al. ICCV'15

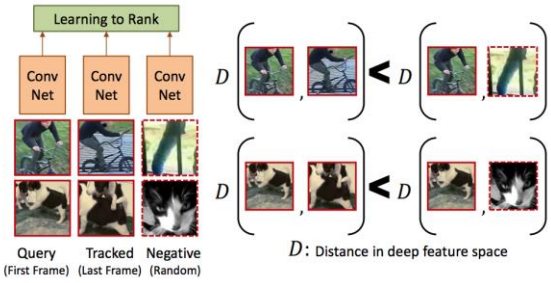
Zhang *et al*. ECCV'16



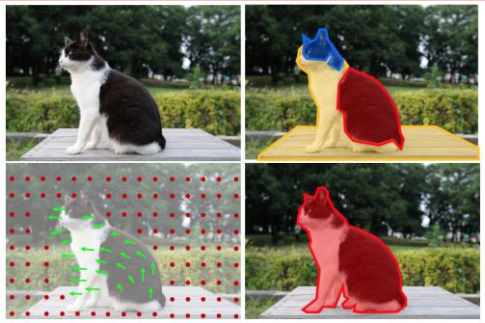
Pathak *et al*. CVPR'16



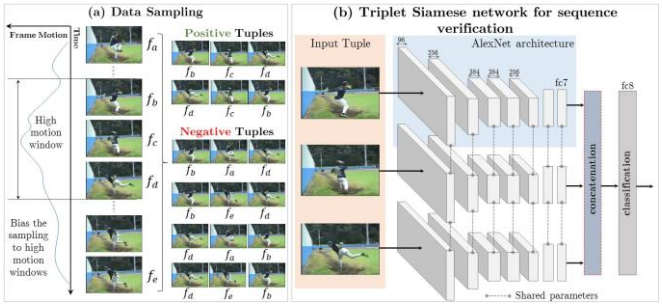
Owens et al. ECCV'16



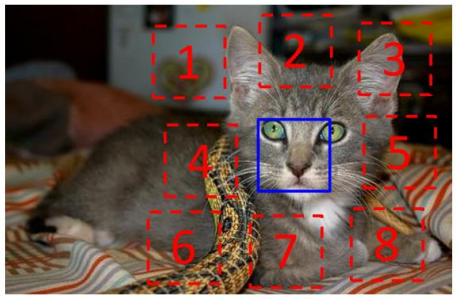
Wang and Gupta ICCV'15



Pathak et al. CVPR'17

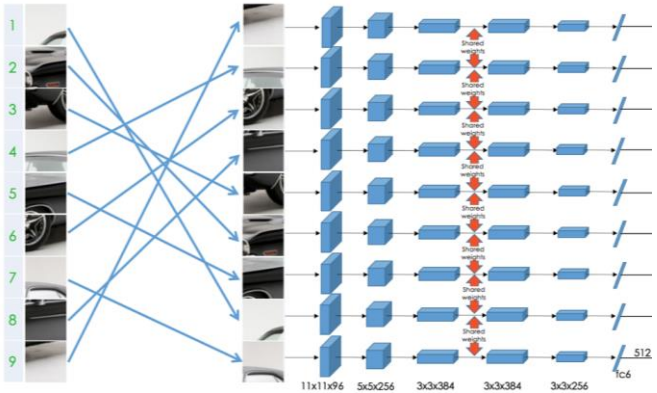


Mirsa et al.ECCV'16

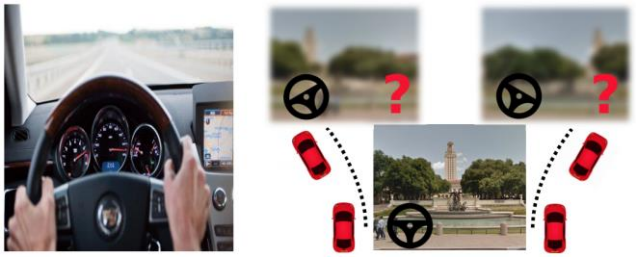


X = (,); Y = 3

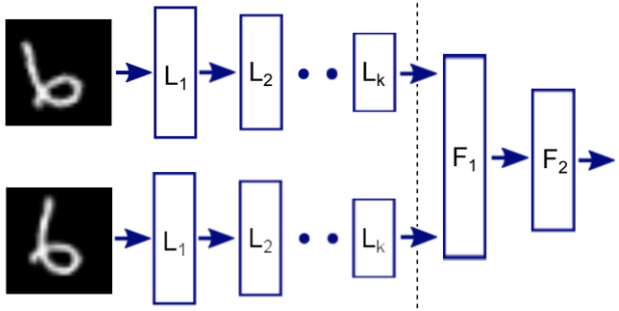
Doerch et al. ICCV'15



Noroozi and Favaro ECCV'16

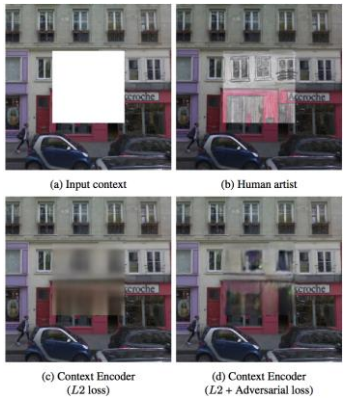


Jayaraman and Grauman ICCV'15

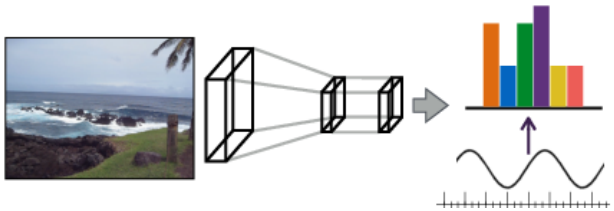


Agrawal et al. ICCV'15

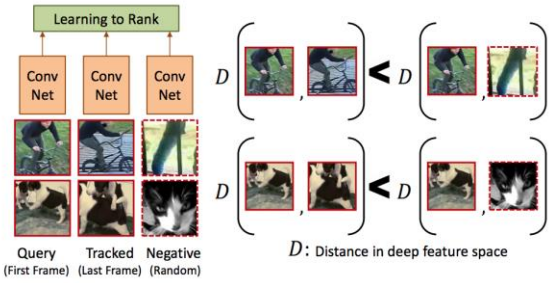
Zhang *et al*. ECCV'16



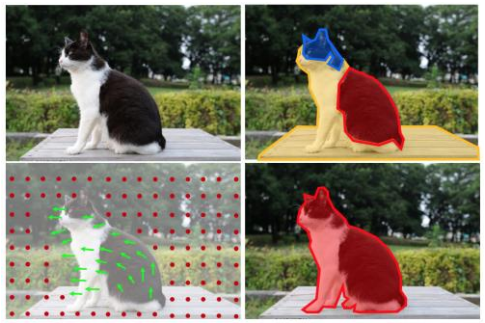
Pathak *et al*. CVPR'16



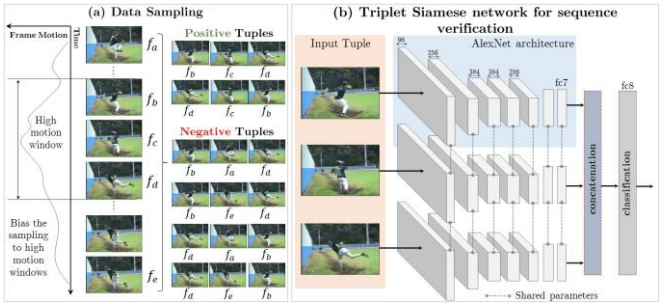
Owens et al. ECCV'16



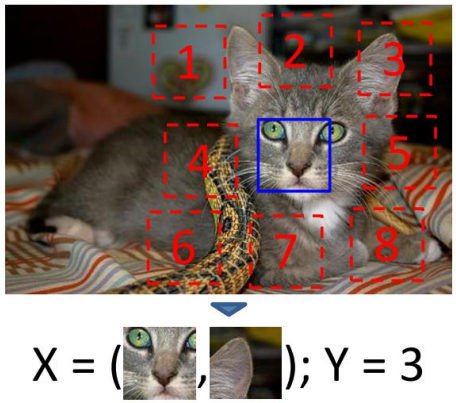
Wang and Gupta ICCV'15



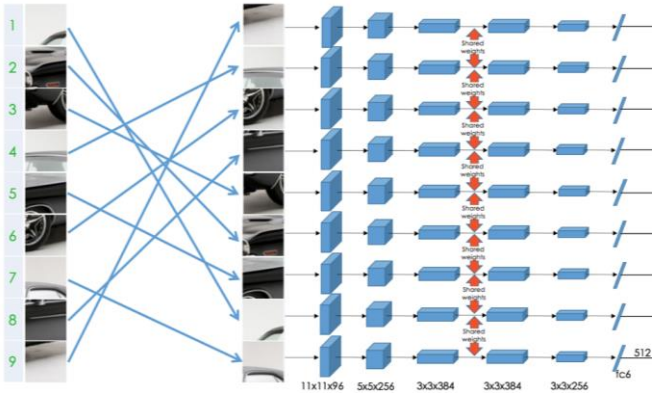
Pathak et al. CVPR'17



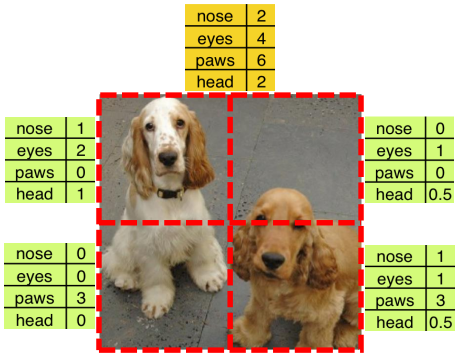
Mirsa et al.ECCV'16



Doerch et al. ICCV'15

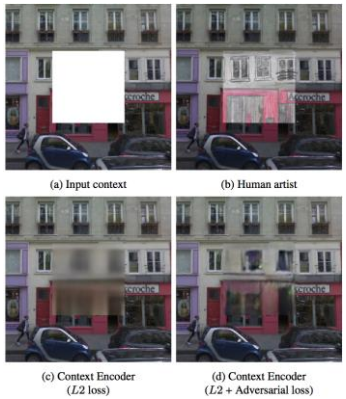


Noroozi and Favaro ECCV'16

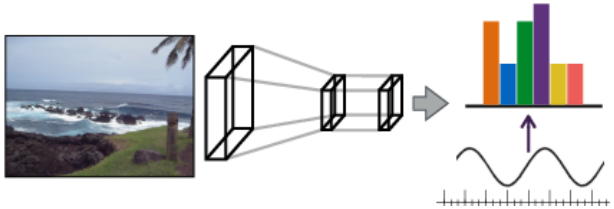


Jayaraman and Grauman ICCV'15

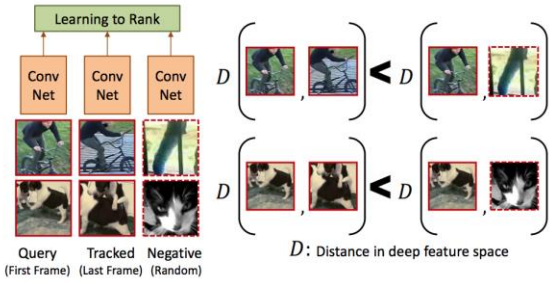
Zhang *et al*. ECCV'16



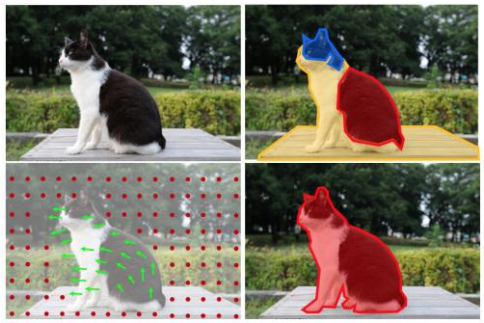
Pathak *et al*. CVPR'16



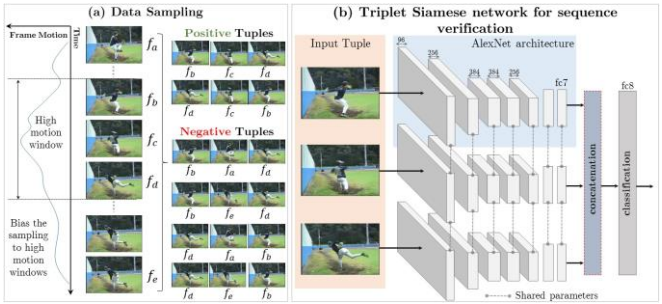
Owens et al. ECCV'16



Wang and Gupta ICCV'15



Pathak et al. CVPR'17



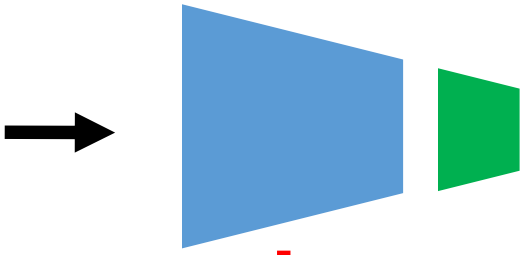
Mirsa et al.ECCV'16

Agenda

- Self supervised learning by counting
- **Boosting self-supervised learning by knowledge transfer**

Dataset (no labels)

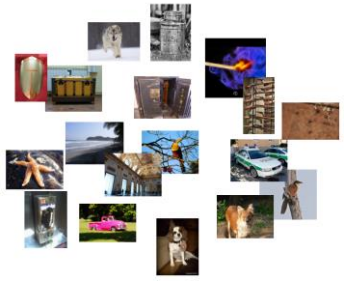
Feature network
(e.g., AlexNet)



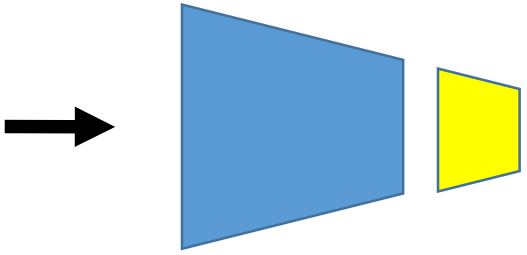
Pretext task
(e.g., counting)

Fine-tuning

Dataset (with labels)

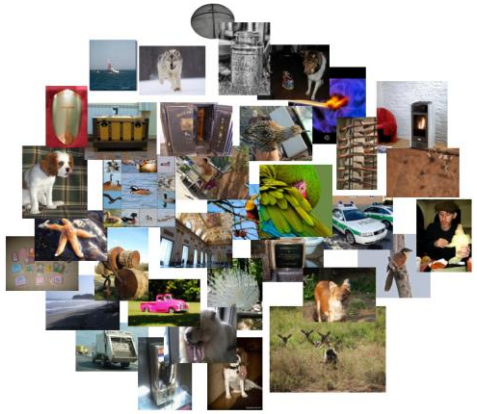


Feature network
(e.g., AlexNet)

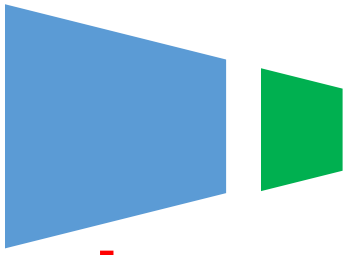


Target task
(e.g., object detection)

Larger Dataset (no labels)



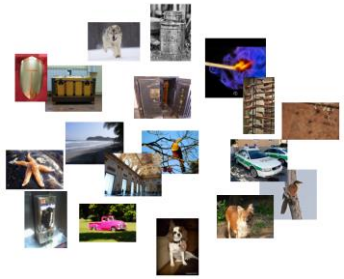
Feature network
(e.g., AlexNet)



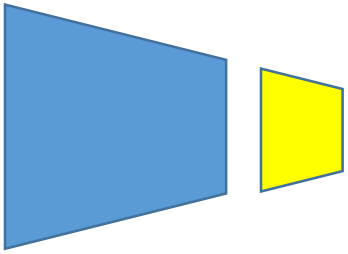
More complicated
Pretext task

Fine-tuning

Dataset (with labels)



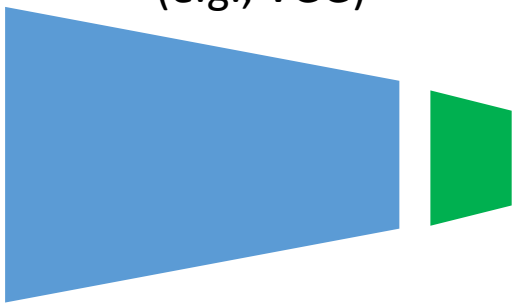
Feature network
(e.g., AlexNet)



Target task
(e.g., object detection)

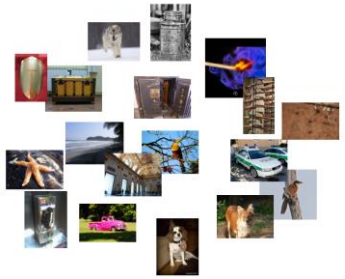
Larger Dataset (no labels)

More complicated
Feature network
(e.g., VGG)

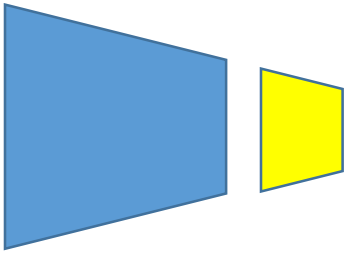


More complicated
Pretext task

Dataset (with labels)

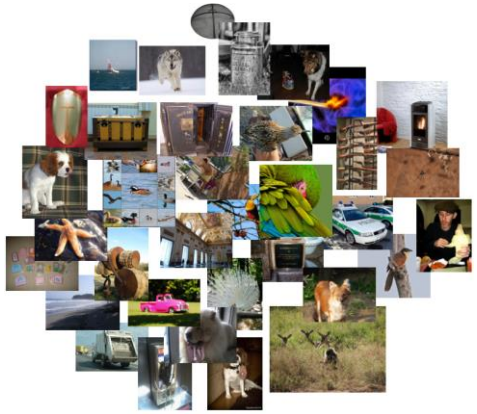


Feature network
(e.g., AlexNet)

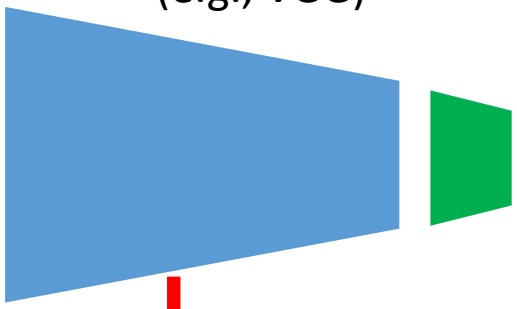


Target task
(e.g., object detection)

Larger Dataset (no labels)



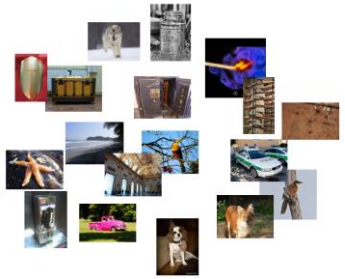
More complicated
Feature network
(e.g., VGG)



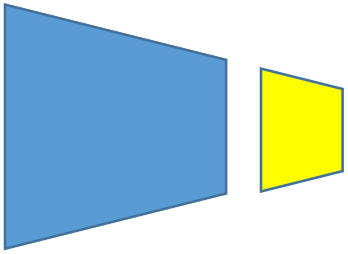
More complicated
Pretext task

Transferring

Dataset (with labels)



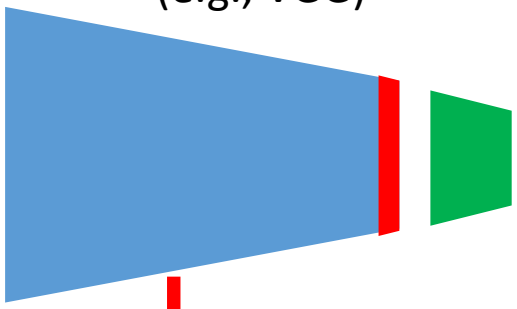
Feature network
(e.g., AlexNet)



Target task
(e.g., object detection)

Larger Dataset (no labels)

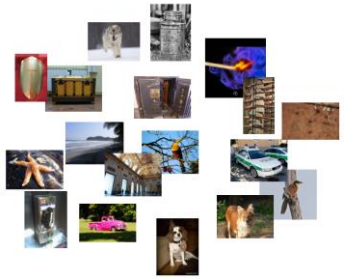
More complicated
Feature network
(e.g., VGG)



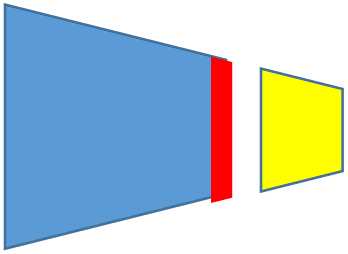
More complicated
Pretext task

Transferring

Dataset (with labels)

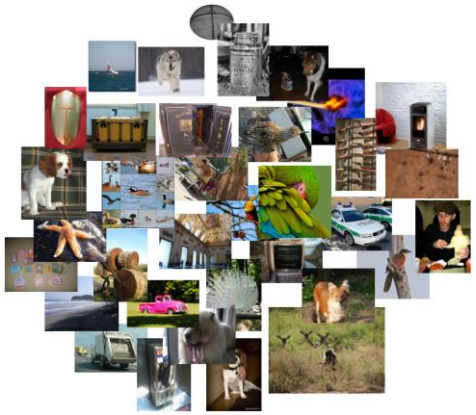


Feature network
(e.g., AlexNet)

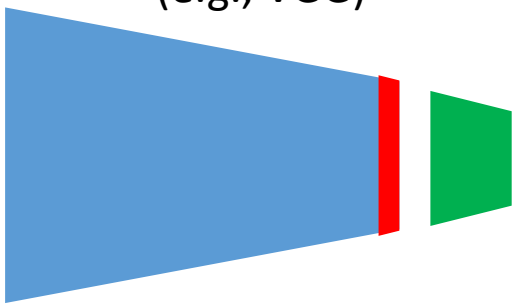


Target task
(e.g., object detection)

Larger Dataset (no labels)

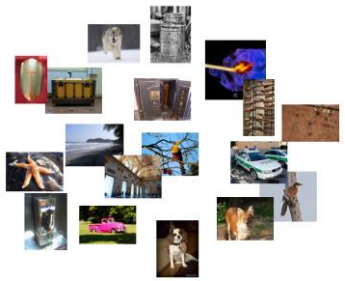


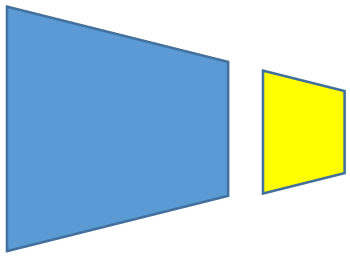
More complicated
Feature network
(e.g., VGG)



More complicated
Pretext task

Dataset (with labels)

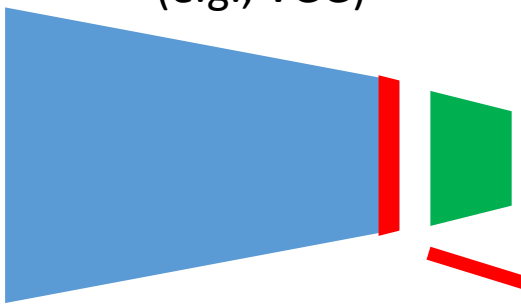




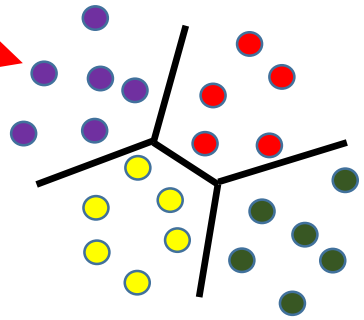
Target task
(e.g., object detection)

Dataset (no labels)

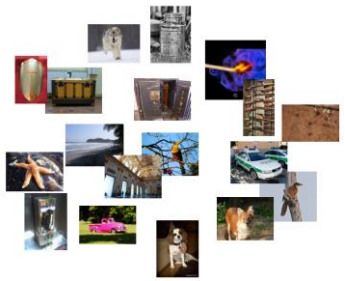
More complicated
Feature network
(e.g., VGG)

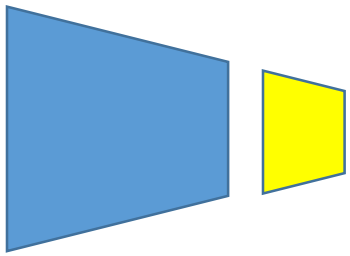


More complicated
Pretext task



Dataset (with labels)

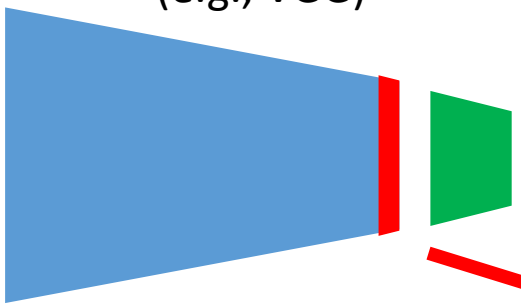




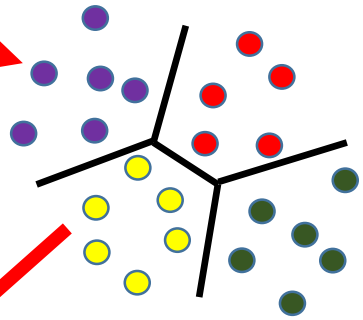
Target task
(e.g., object detection)

Dataset (no labels)

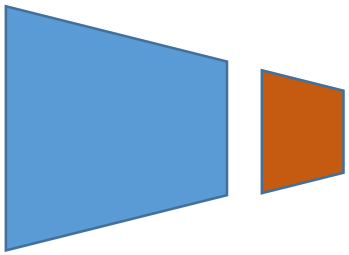
More complicated
Feature network
(e.g., VGG)



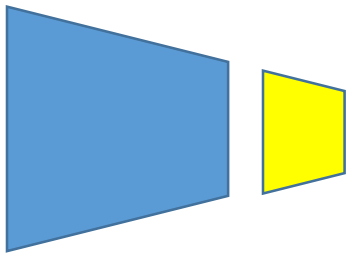
More complicated
Pretext task



Pseudo labels



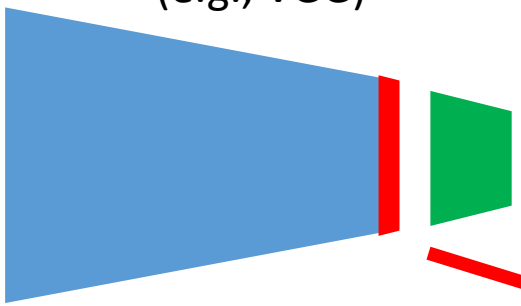
Dataset (with labels)



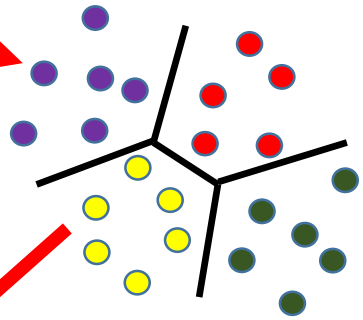
Target task
(e.g., object detection)

Dataset (no labels)

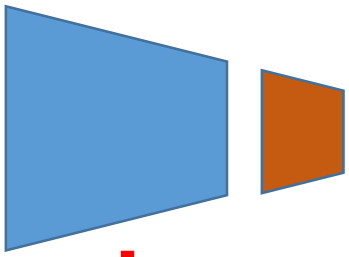
More complicated
Feature network
(e.g., VGG)



More complicated
Pretext task

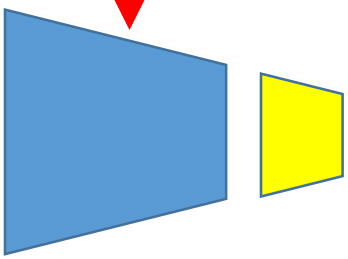


Pseudo labels



Fine-tuning

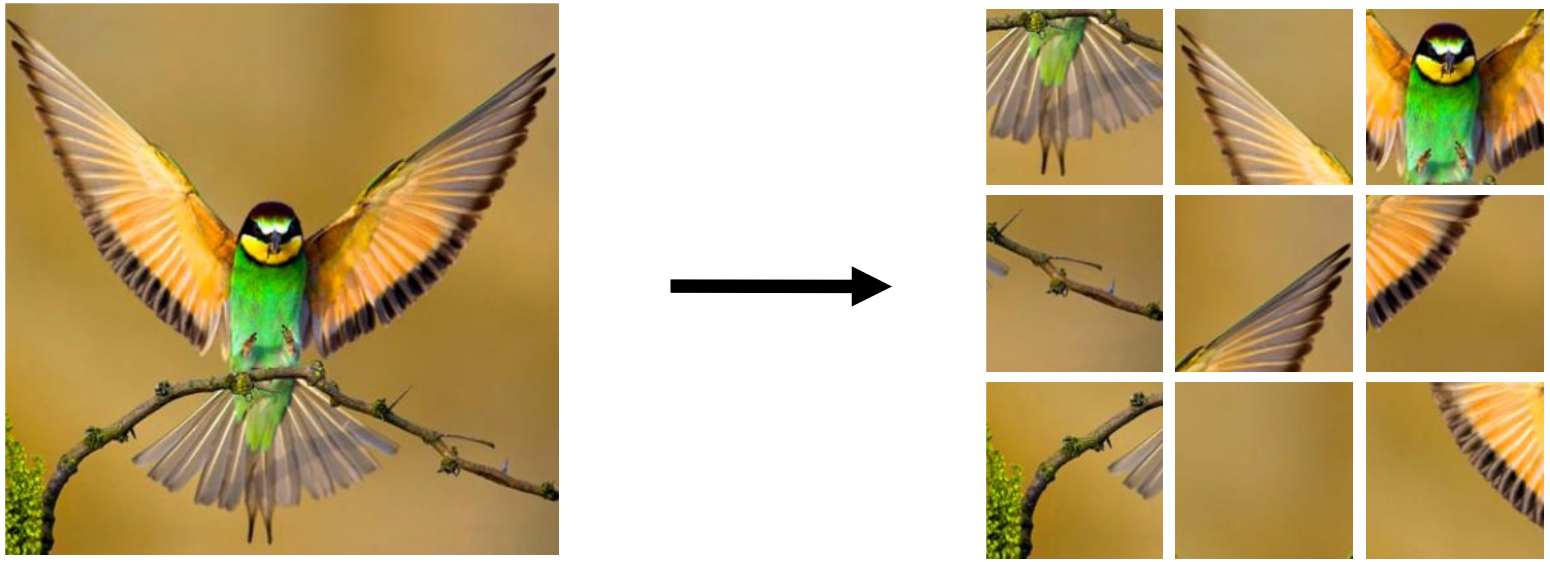
Dataset (with labels)



Target task
(e.g., object detection)

Jigsaw

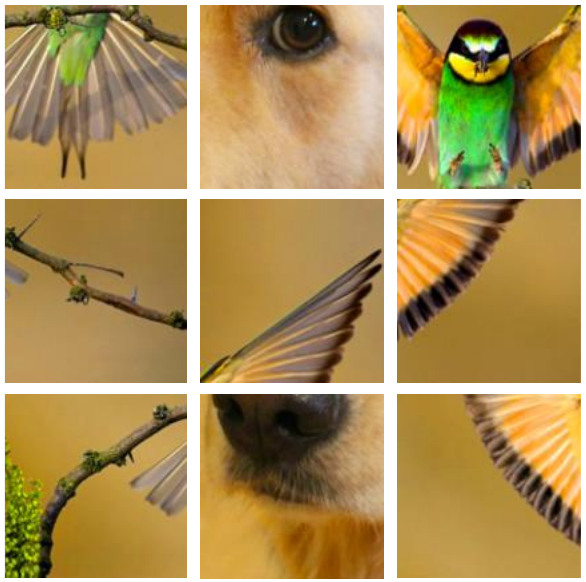
Permute and then predict the permutation



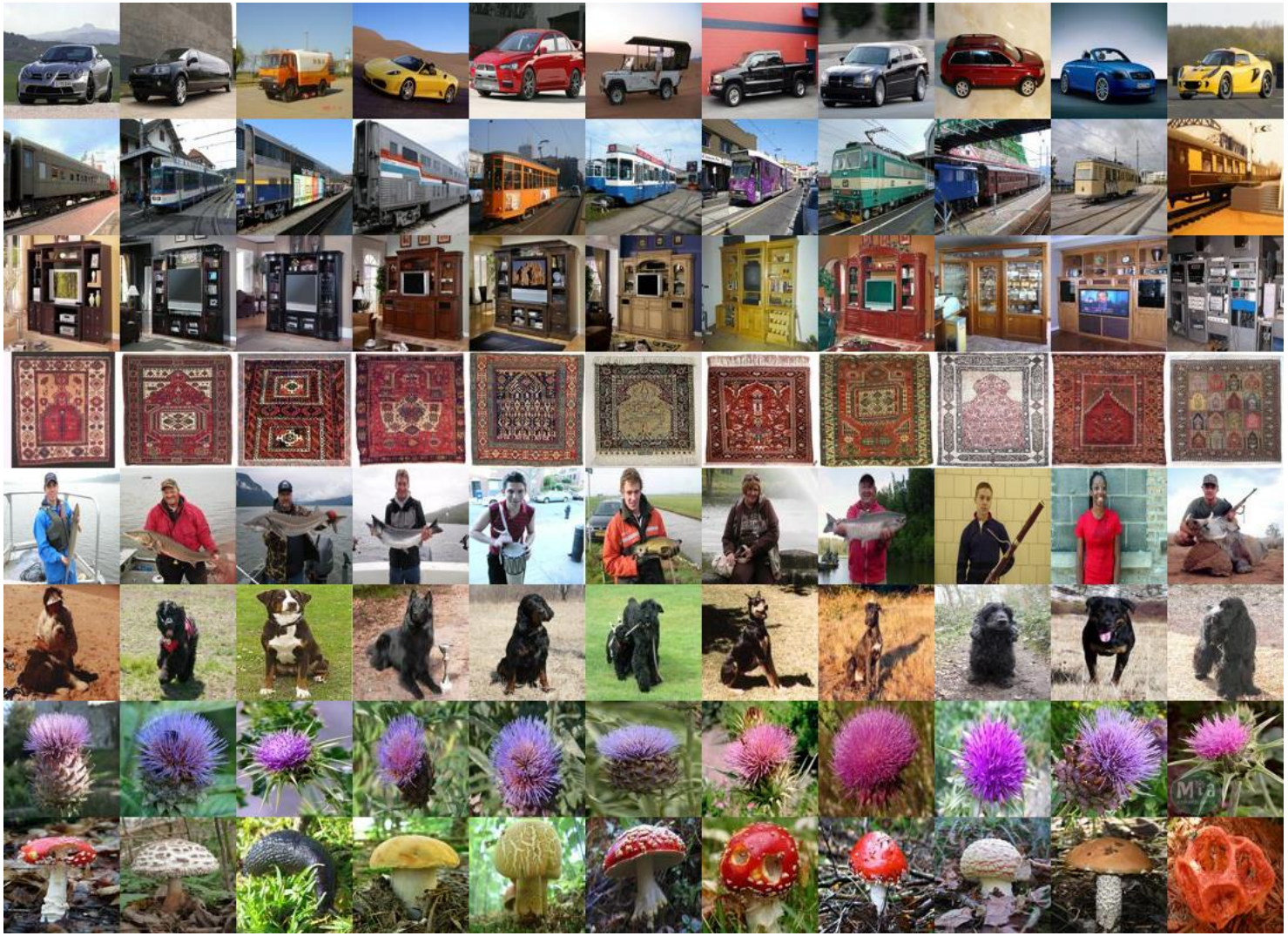
Noroozi, Mehdi, and Paolo Favaro. "Unsupervised learning of visual representations by solving jigsaw puzzles." ECCV 2016.

Jigsaw++

- Add distracting patches
- Increase number of permutations



Clusters on Jigsaw++



Results on transfer learning

Fine-tuning on PASCAL VOC07

Method	**Class.**	**Det.**	**Segm.**
Supervised	79.9	57.1	48.0
Random	53.3	43.4	19.8
Sound	54.4	44.0	-
Video	63.1	47.2	-
Split-Brain	67.1	46.7	36.0
Watching-Objects	61.0	52.2	-
Jigsaw(new version)	**67.6**	**53.2**	**37.6**
Counting (Ours)	**67.7**	**52.4**	**36.6**

Results on transfer learning

Fine-tuning on PASCAL VOC07

Method	**Class.**	**Det.**	**Segm.**
Supervised	79.9	57.1	48.0
Random	53.3	43.4	19.8
Sound	54.4	44.0	-
Video	63.1	47.2	-
Split-Brain	67.1	46.7	36.0
Watching-Objects	61.0	52.2	-
Jigsaw(new version)	**67.6**	**53.2**	**37.6**
Counting (Ours)	**67.7**	**52.4**	**36.6**
Jigsaw++ (Ours)	72.5	**56.5**	**42.6**

Results on transfer learning

Fine-tuning on PASCAL VOC07

Method	Class.	Det.	Segm.
Supervised	79.9	57.1	48.0
Random	53.3	43.4	19.8
Sound	54.4	44.0	-
Video	63.1	47.2	-
Split-Brain	67.1	46.7	36.0
Watching-Objects	61.0	52.2	-
Jigsaw(new version)	**67.6**	**53.2**	**37.6**
Counting (Ours)	**67.7**	**52.4**	**36.6**
Jigsaw++ (Ours)	72.5	**56.5**	**42.6**
RotNet (ICLR'18)	**72.9**	54.4	39.1
Deep clustering (ECCV'18)	73.7	55.4	45.1

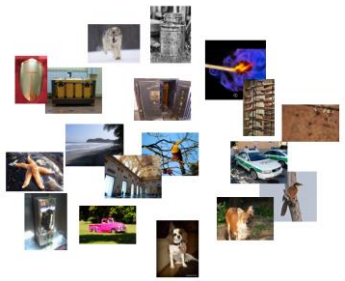
Dataset (no labels)

More complicated
Feature network
(e.g., VGG)

More complicated
Pretext task

Pseudo labels

Dataset (with labels)

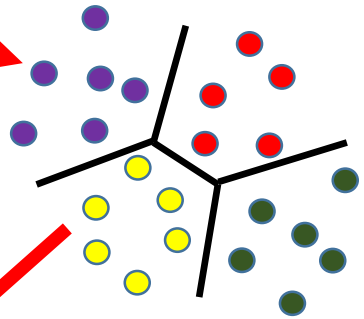


Fine-tuning

Target task
(e.g., object detection)

Dataset (no labels)

HOG



Pseudo labels

Dataset (with labels)

Fine-tuning

Target task
(e.g., object detection)

Results on transfer learning

Fine-tuning on PASCAL VOC07

Method	Class.	Det.	Segm.
Supervised	79.9	57.1	48.0
Random	53.3	43.4	19.8
Sound	54.4	44.0	-
Video	63.1	47.2	-
Split-Brain	67.1	46.7	36.0
Watching-Objects	61.0	52.2	-
Jigsaw(new version)	**67.6**	**53.2**	**37.6**
Counting (ours)	**67.7**	**52.4**	**36.6**
Jigsaw++ (ours)	72.5	**56.5**	**42.6**
HOG (ours)	70.2	53.2	39.2

Kaiming He Ross Girshick Piotr Dollar, "Rethinking ImageNet Pre-training", arXiv, Nov 2018.

Visualization of conv1 filters

From scratch

CC on
VGG-Jigsaw++

CC on
HOG

Thanks to

Ananth Kavalkazhani

Mehdi Noroozi

Paolo Favaro

Thanks!