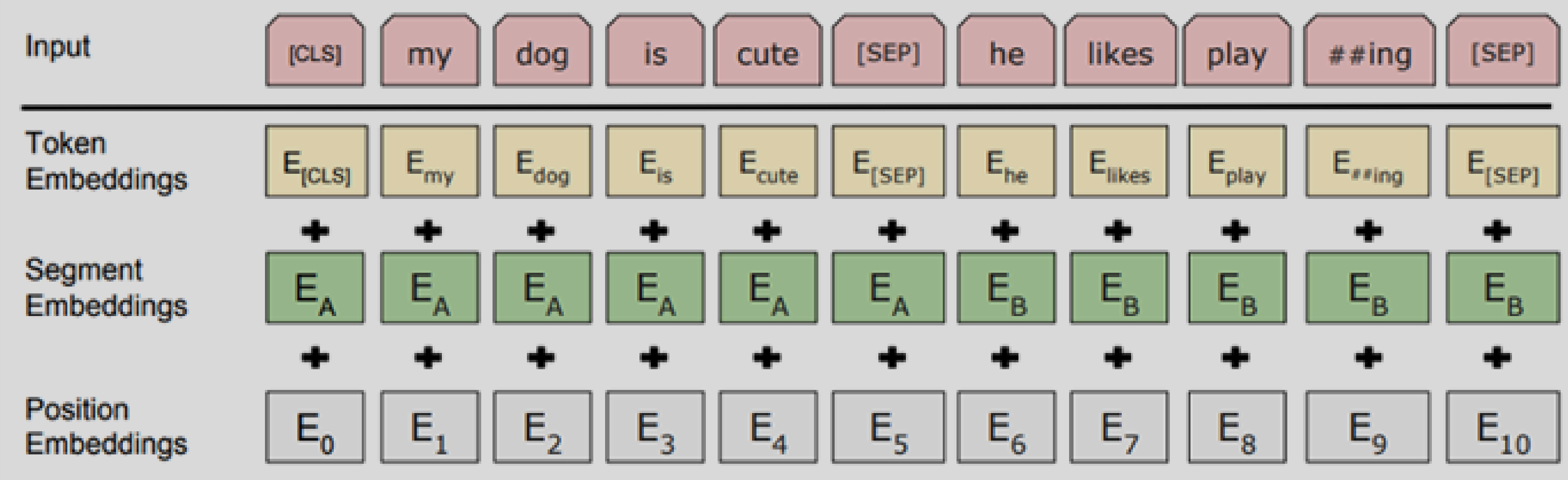
Computer Vision

Lecture 09: Weakly-/Self-supervised Learning

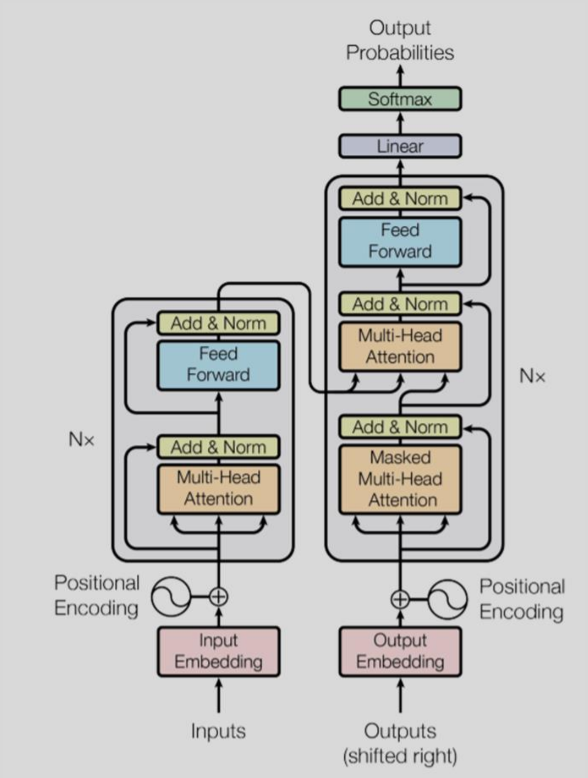
Self-supervised learning

Self-supervised learning

- Self-supervision: Learning without tagged data.
- The method could be applied to any inputs.
  - Speech, image, video, text and etc.



Self-supervised learning



Input = [CLS] the man went to [MASK] store [SEP]

he bought a gallon [MASK] milk [SEP]

Label = IsNext

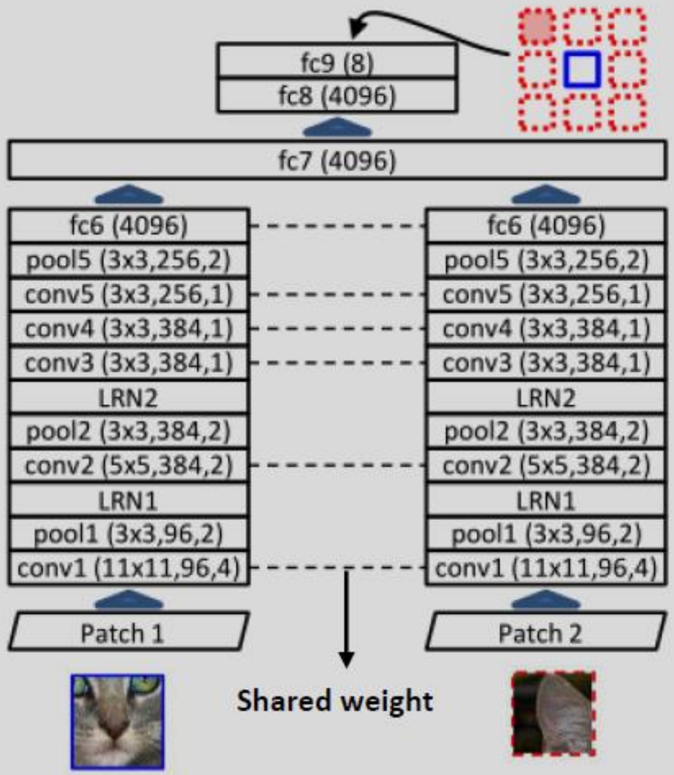
Input = [CLS] the man [MASK] to the store [SEP]

penguin [MASK] are flight ##less birds [SEP]

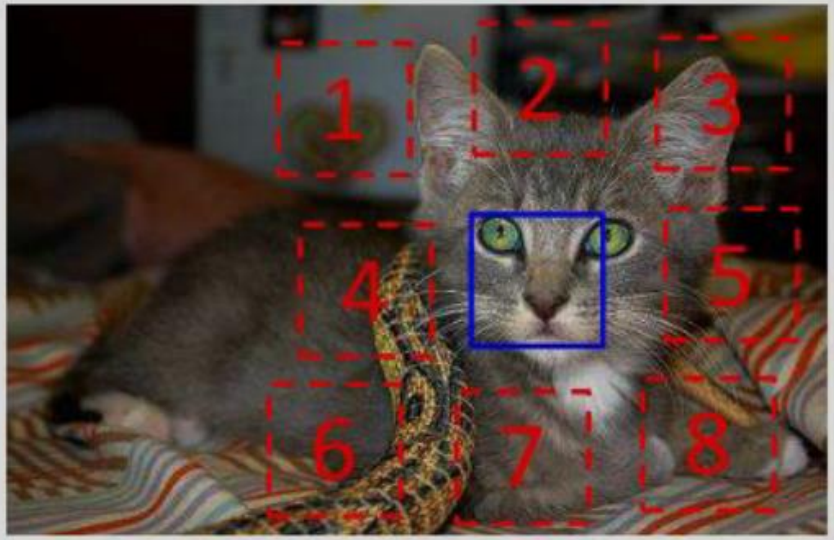
Label = NotNext

Transformer architecture is trained by 1) Masked language model, 2) Next sentence prediction

Self-supervised learning



Include a gap between patches



Randomly jitter each patch location

Unsupervised Visual Representation Learning by Context Prediction. *ICCV 2015*

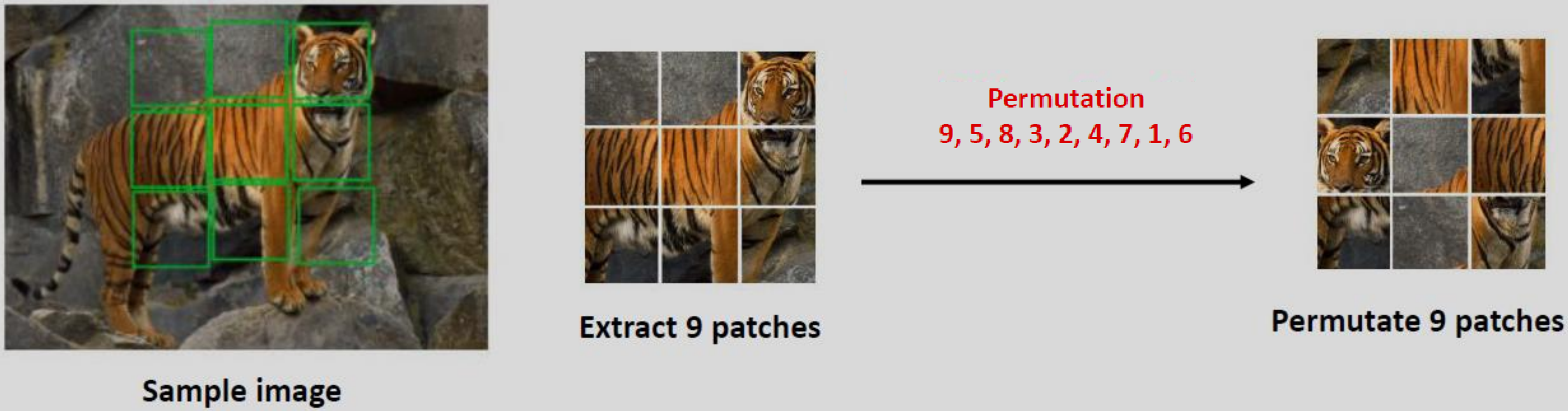
Self-supervised learning

Context Prediction: Predict relative positions of patches

- You have to understand the object to solve this problem!
- Be aware of trivial solution! CNN is especially good at it

Unsupervised Visual Representation Learning by Context Prediction. *ICCV 2015*

Self-supervised learning

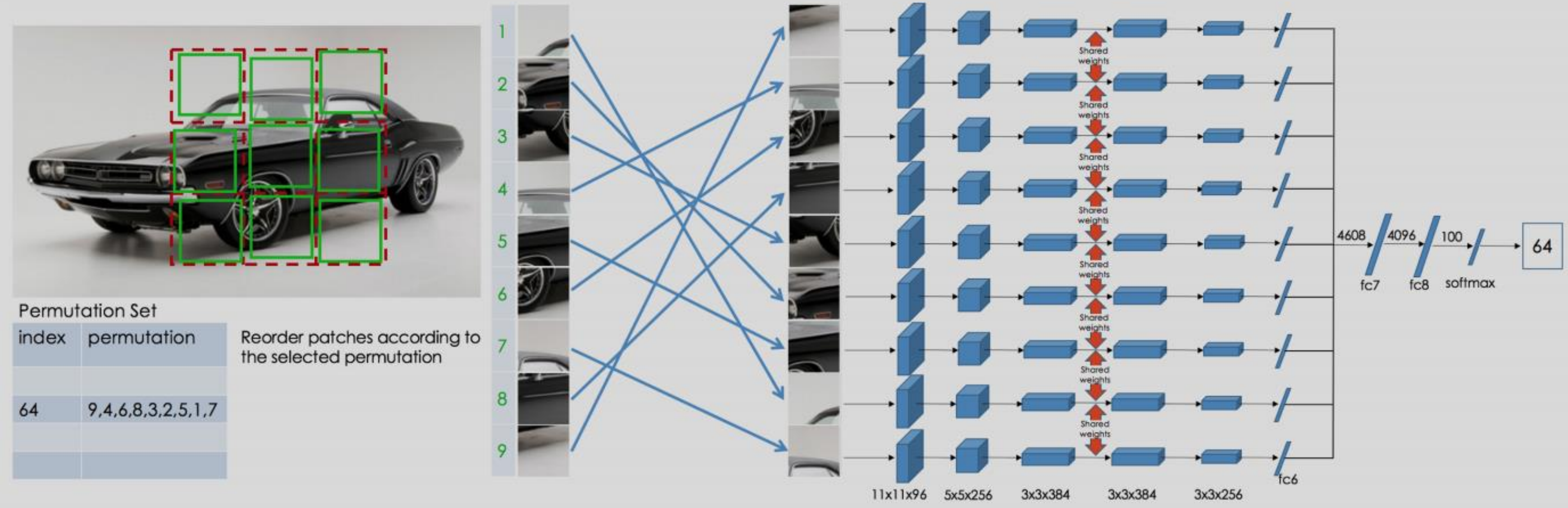


Unsupervised learning of visual representations by solving jigsaw puzzles. In *ECCV 2016*.

Self-supervised learning

Solving the Jigsaw

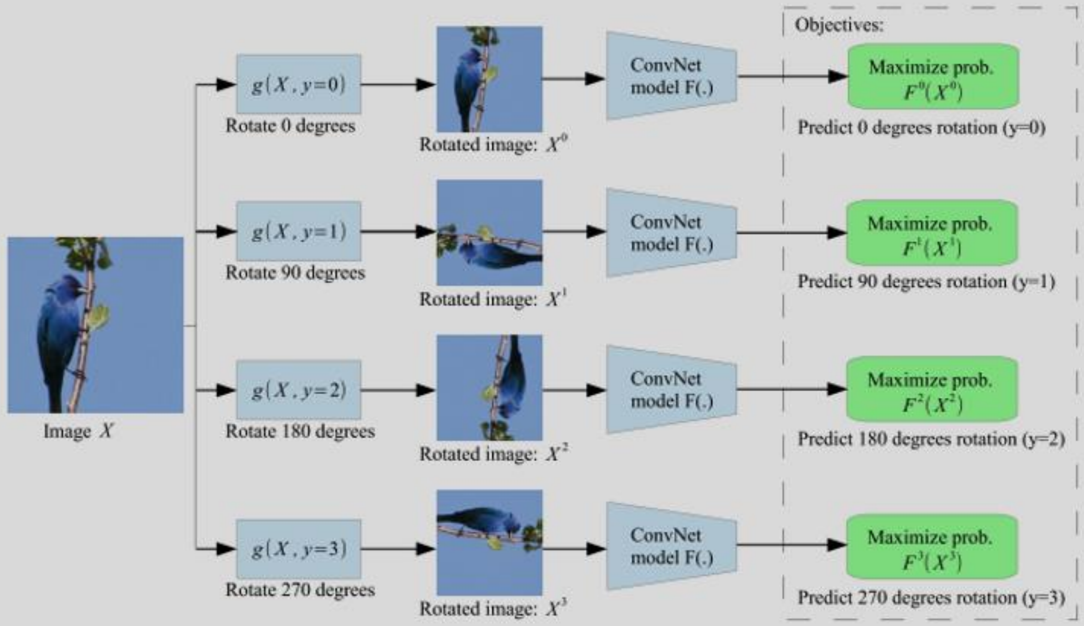
- Use stronger supervision, solve the real jigsaw problem
- Harder problem, better ability for networks



Self-supervised learning

Predicting the rotations

- Predict the 4 types of rotation angles.



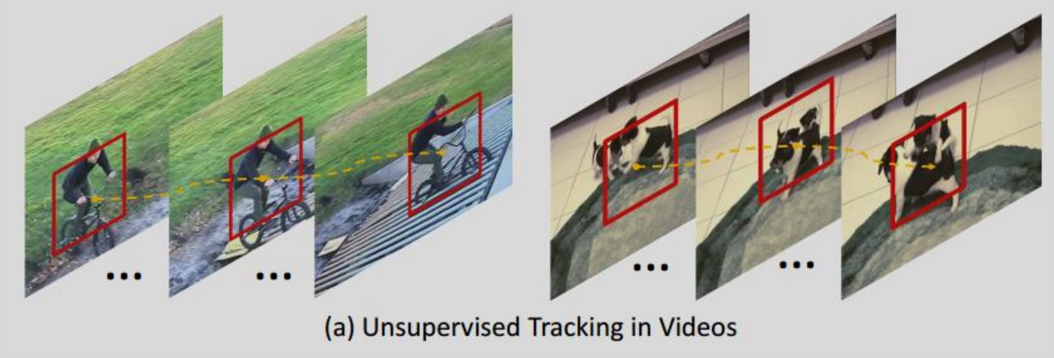
Method	Conv1	Conv2	Conv3	Conv4	Conv5
ImageNet labels	19.3	36.3	44.2	48.3	50.5
Random	11.6	17.1	16.9	16.3	14.1
Random rescaled Krähenbühl et al. (2015)	17.5	23.0	24.5	23.2	20.6
Context (Doersch et al., 2015)	16.2	23.3	30.2	31.7	29.6
Context Encoders (Pathak et al., 2016b)	14.1	20.7	21.0	19.8	15.5
Colorization (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	**18.8**	**31.7**	**38.7**	**38.2**	**36.5**

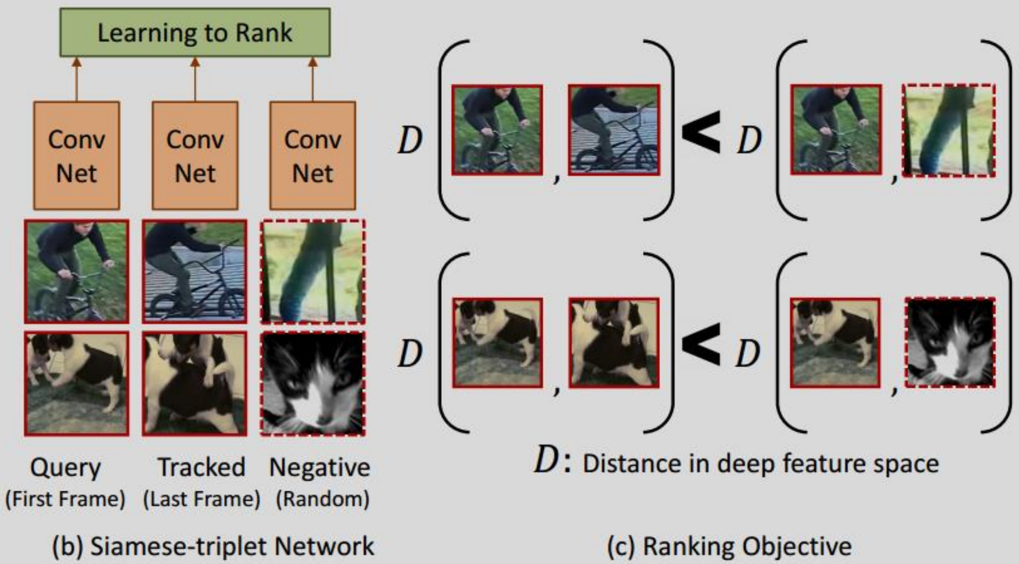
ImageNet classification top-1 accuracy

Unsupervised representation learning by predicting image rotations. In *ICLR 2018*.

Self-supervision for video

Find corresponding pairs using visual tracking



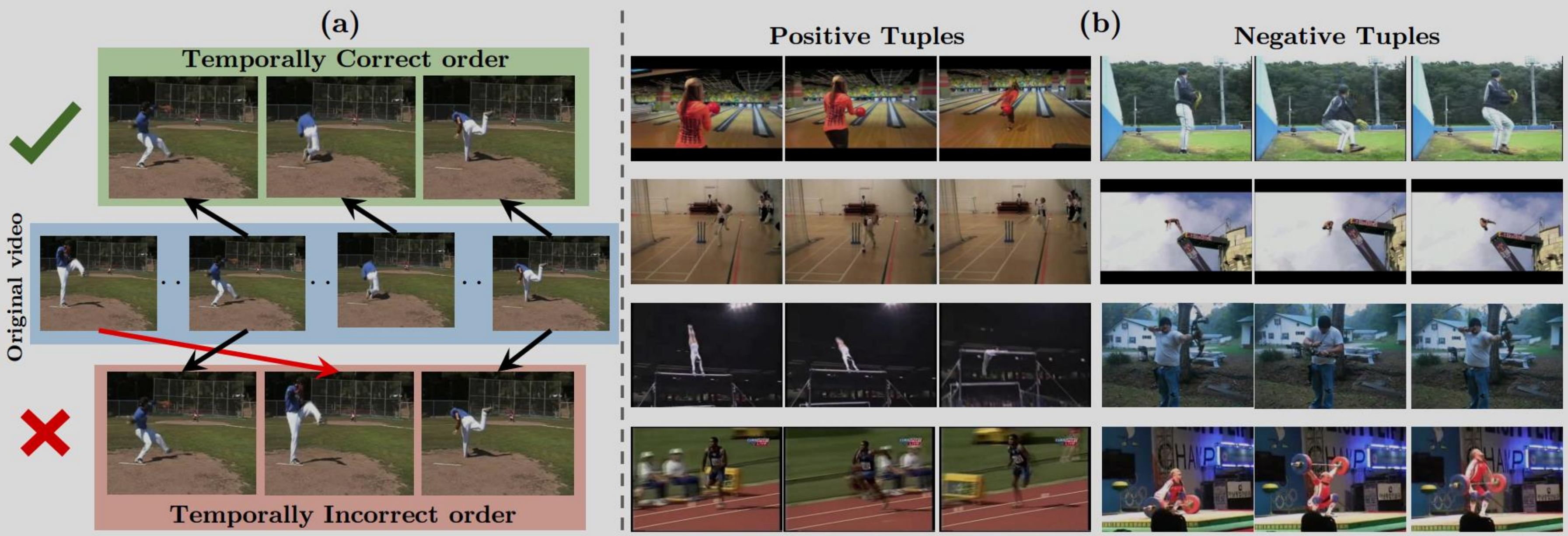


Wang, X., & Gupta, A. (2015). Unsupervised learning of visual representations using videos. In *ICCV2015*

Self-supervision for video

Is the temporal order of a video correct?

- Encode the cause and effect of action

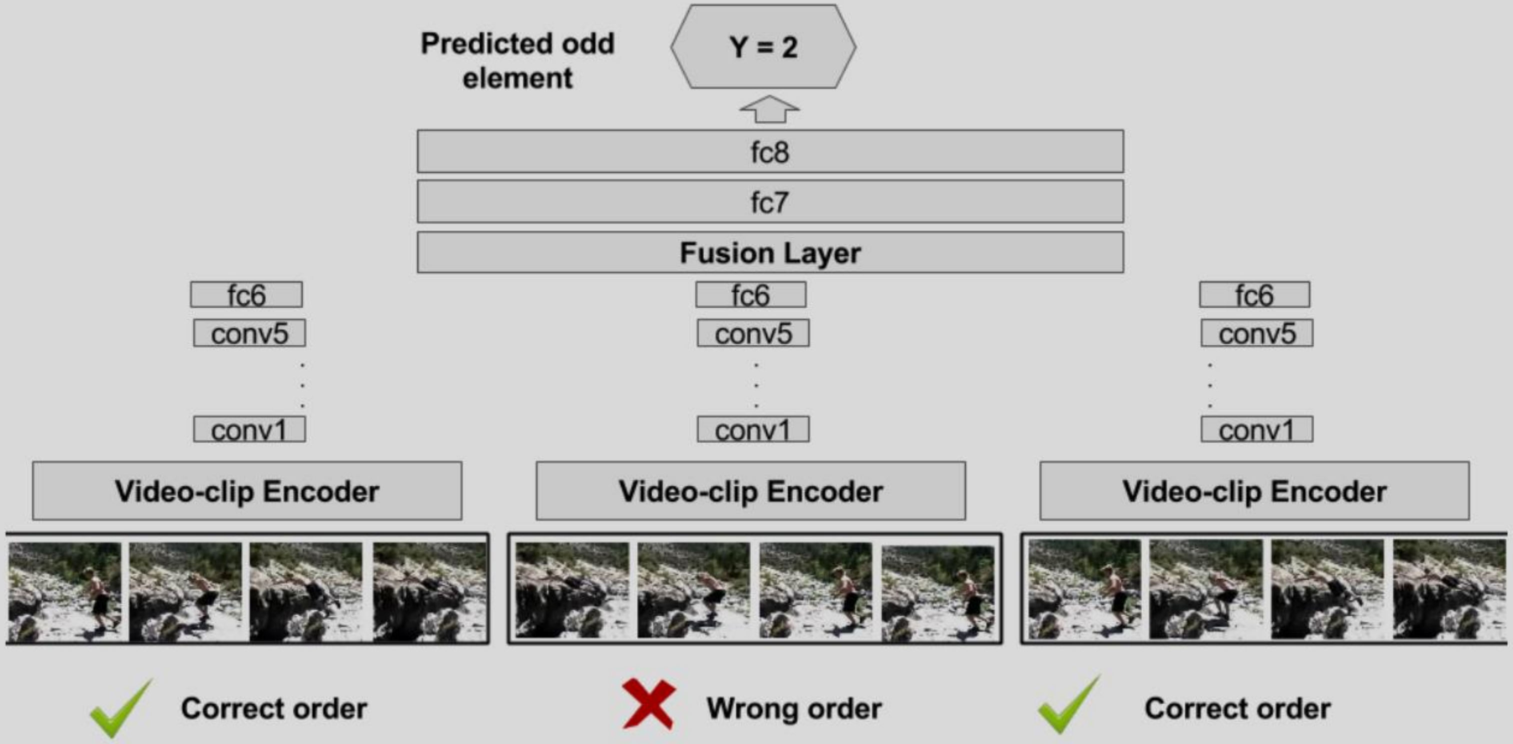


Misra, I., Zitnick, C. L., & Hebert, M. Shuffle and learn: unsupervised learning using temporal order verification. In *ECCV 2016*.

Self-supervision for video

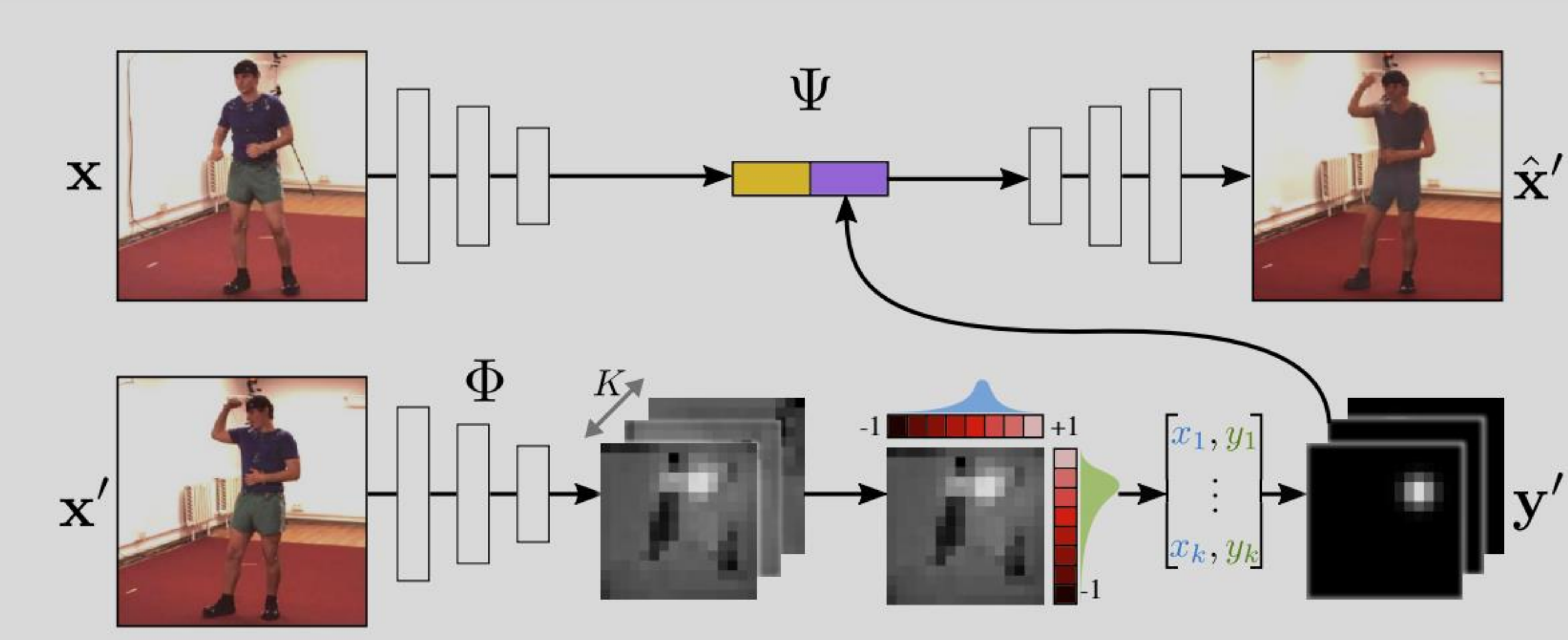
Is the temporal order of a video correct?

- Find the odd sequence



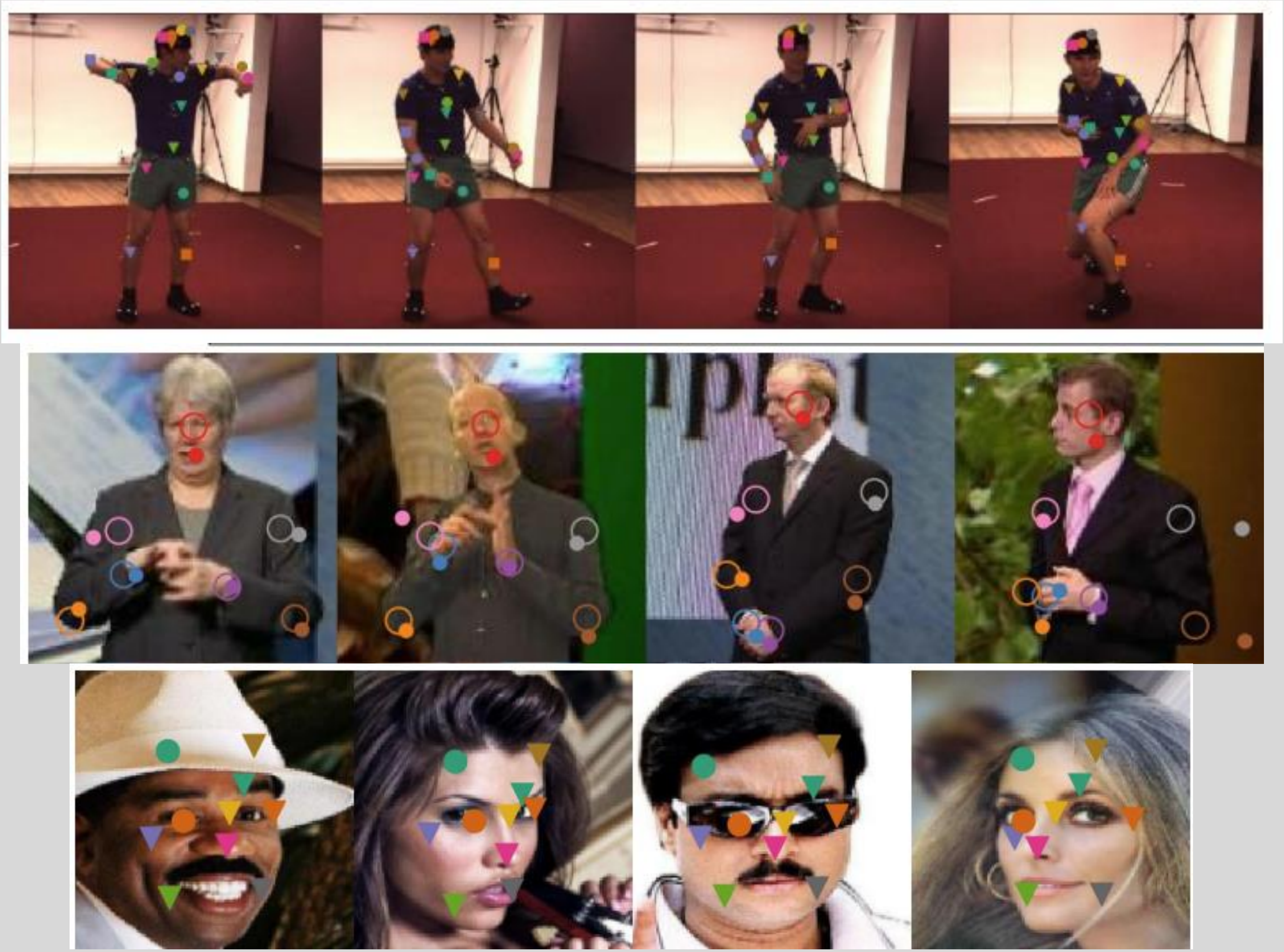
Fernando, B., Bilen, H., Gavves, E., & Gould, S. Self-Supervised Video Representation Learning With Odd-One-Out Networks. *In CVPR2017*.

Self-supervision for pose

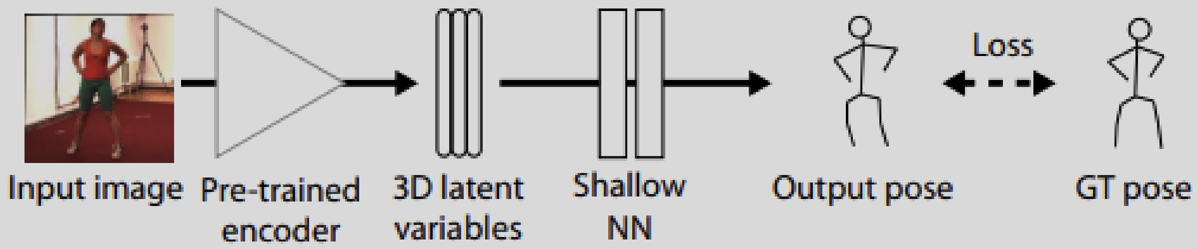


Unsupervised Learning of Object Landmarks through Conditional Image Generation, NeurIPS'18.

Self-supervision for pose

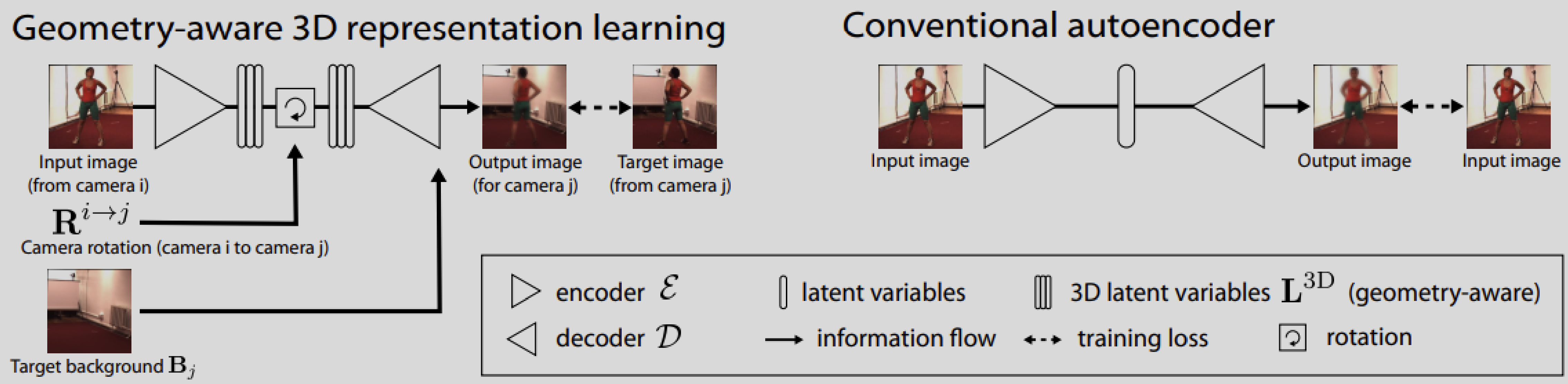


Self-supervision for pose



Unsupervised Geometry-Aware Representation for 3D Human Pose Estimation, ECCV'18.

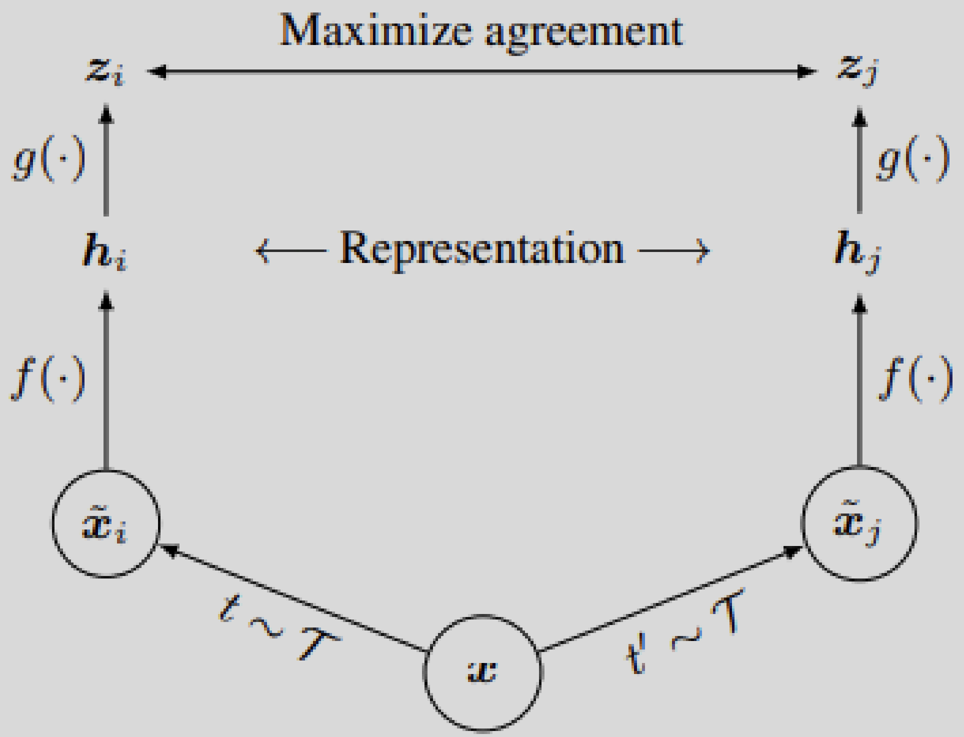
Self-supervision for pose



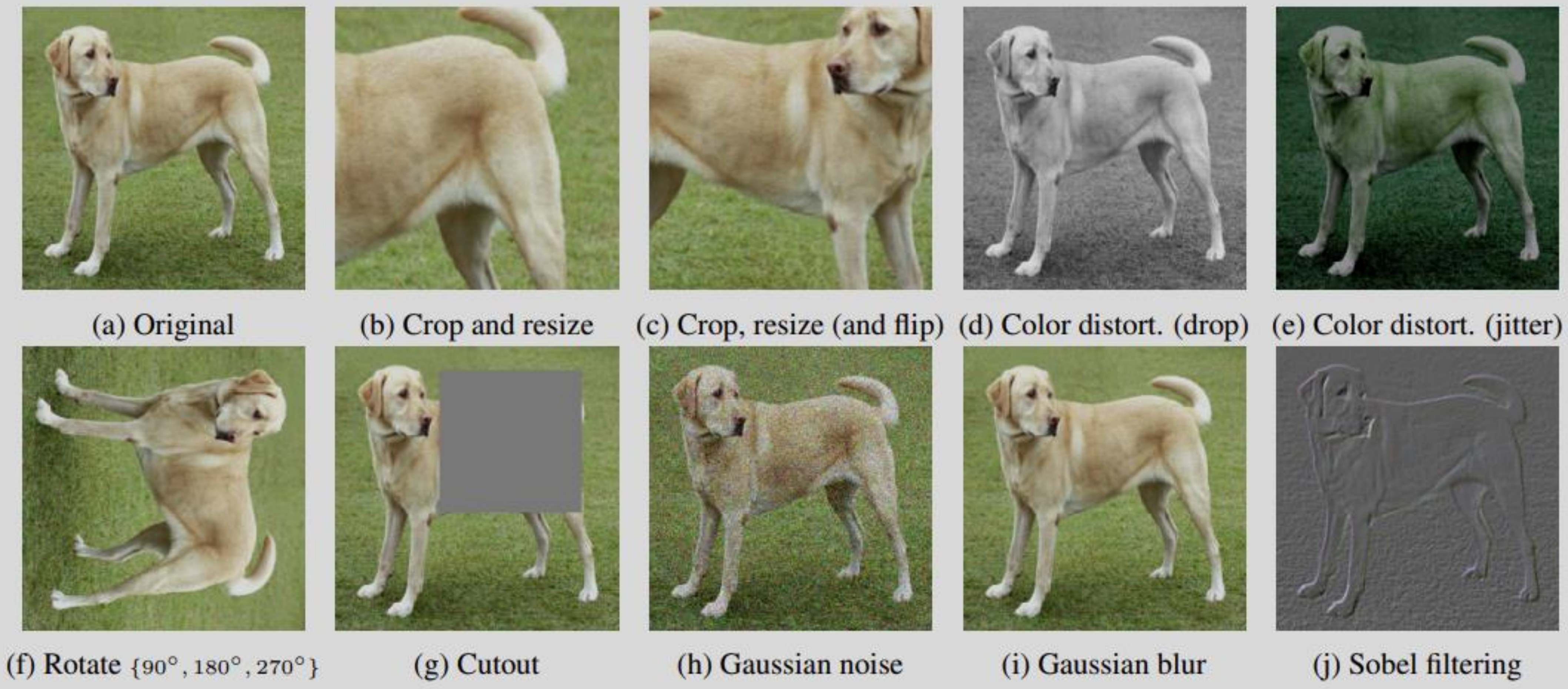
Unsupervised Geometry-Aware Representation for 3D Human Pose Estimation, ECCV'18.

Contrastive learning

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



Contrastive learning



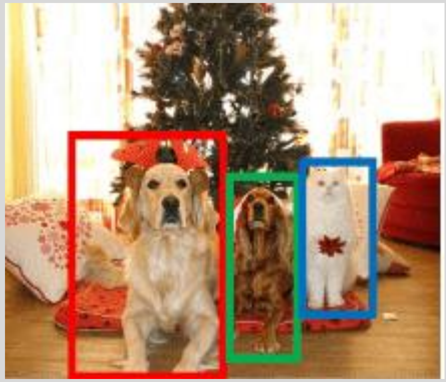
Contrastive learning

	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

Weakly-supervised learning

Weakly-supervised learning

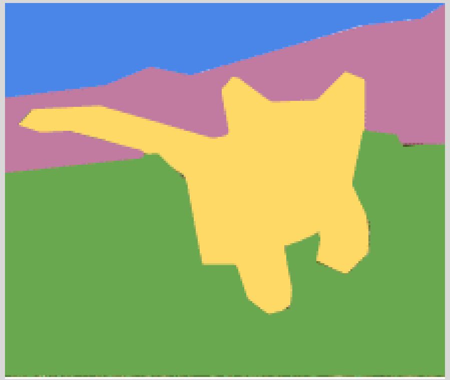
- Weak supervision: Incomplete supervision
- Training data with only coarse-grained labels.



Bounding box

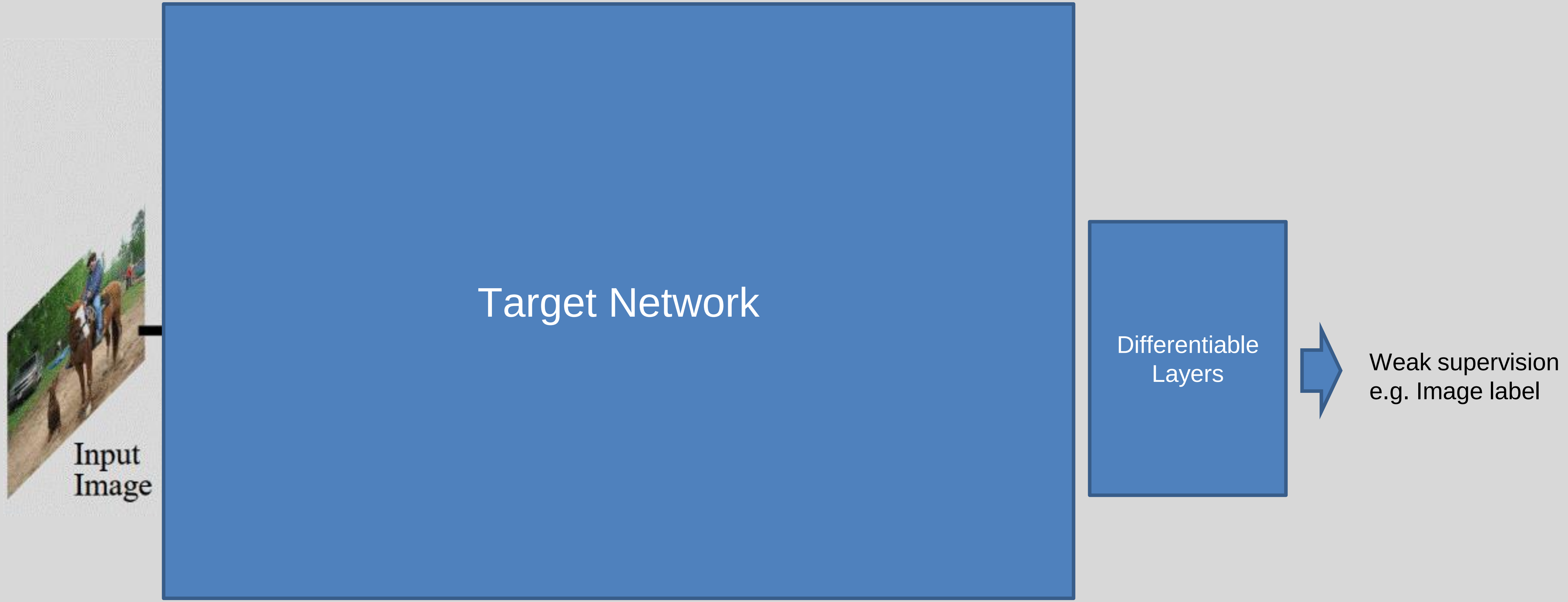
DOG, DOG, CAT

Image-level label

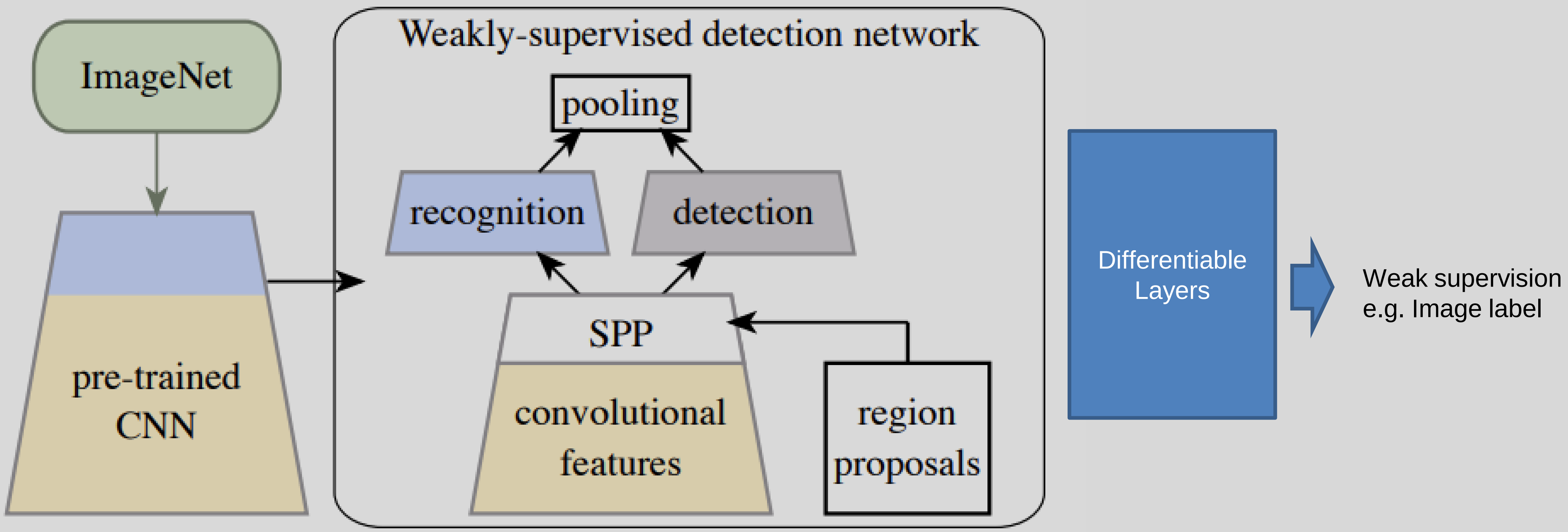


Semantic segmentation

Weakly-supervised learning



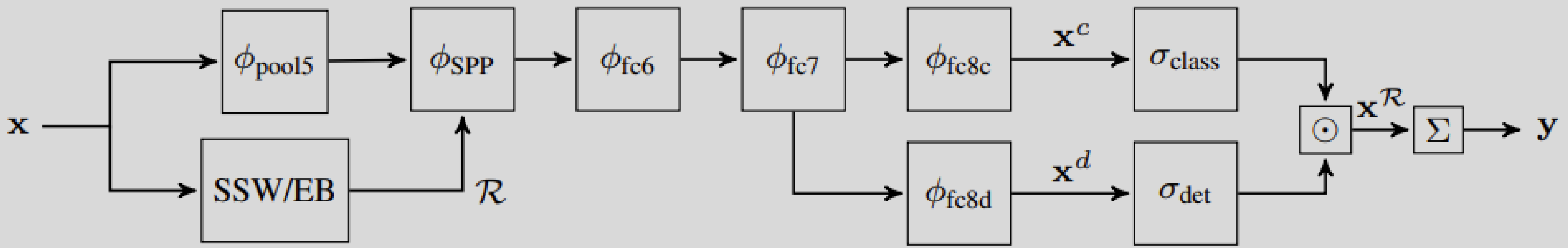
Weakly-supervised object detection



Weakly Supervised Deep Detection Networks, CVPR’16

Weakly-supervised object detection

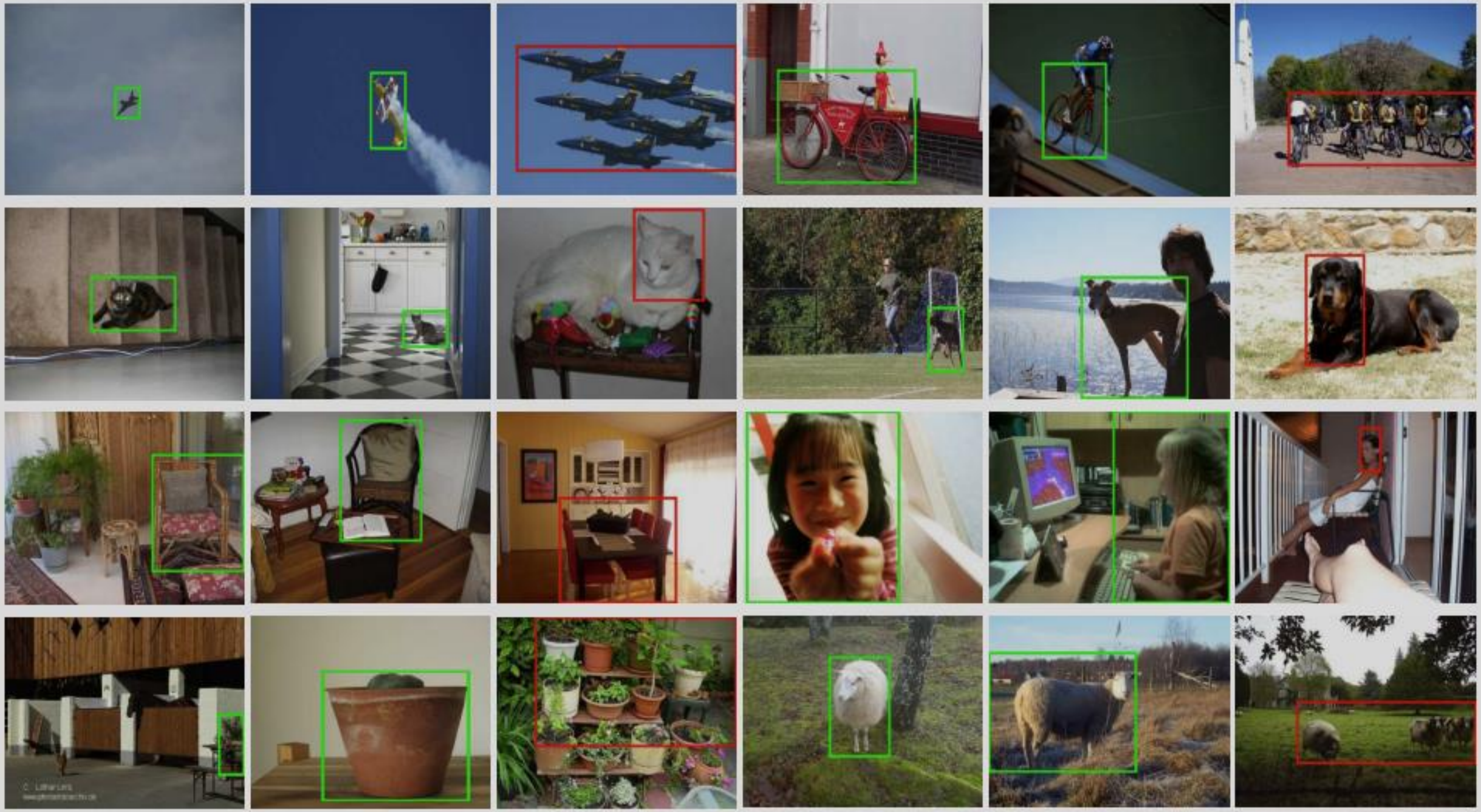
$$[\sigma_{\text{class}}(\mathbf{x}^c)]_{ij} = \frac{e^{x^c_{ij}}}{\sum_{k=1}^{C} e^{x^c_{kj}}}$$



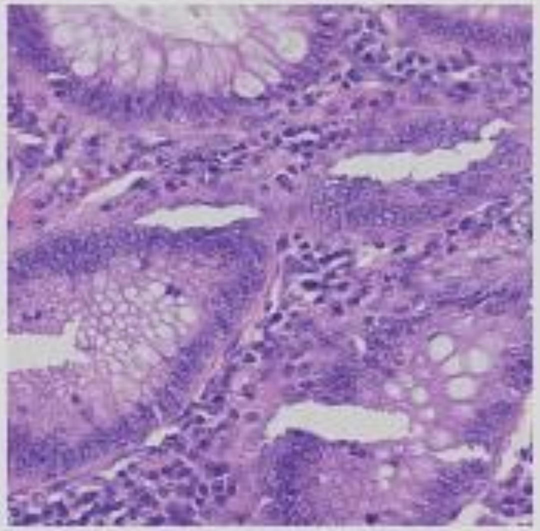
$$y_c = \sum_{r=1}^{|\mathcal{R}|} x^{\mathcal{R}}_{cr}$$

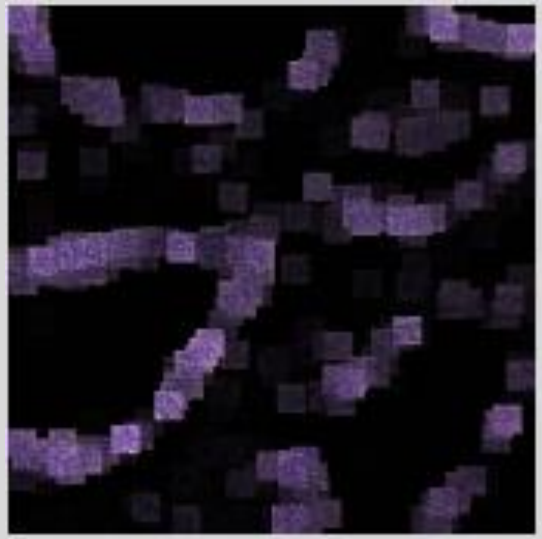
$$[\sigma_{\text{det}}(\mathbf{x}^d)]_{ij} = \frac{e^{x^d_{ij}}}{\sum_{k=1}^{|\mathcal{R}|} e^{x^d_{ik}}}$$

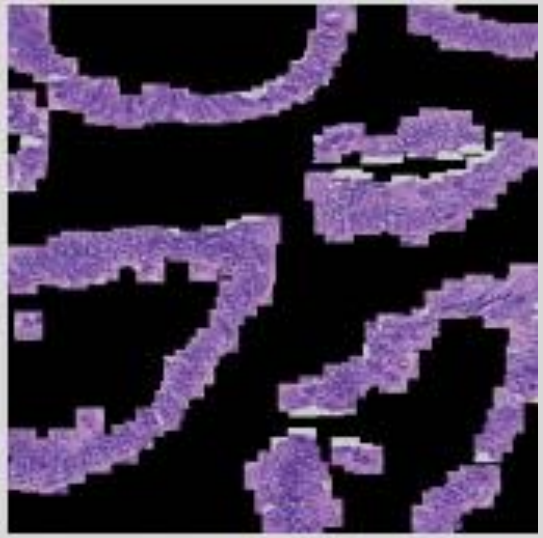
Weakly-supervised object detection



Weakly-supervised segmentation


Original image


Predicted patch weights

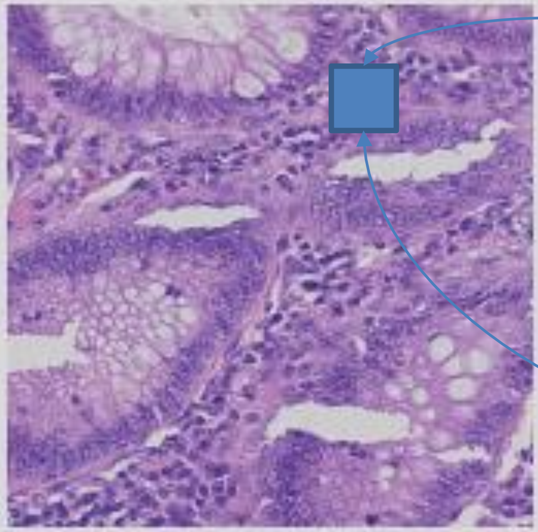

Ground-truth patches

Attention-based Deep Multiple Instance Learning, ICML'18

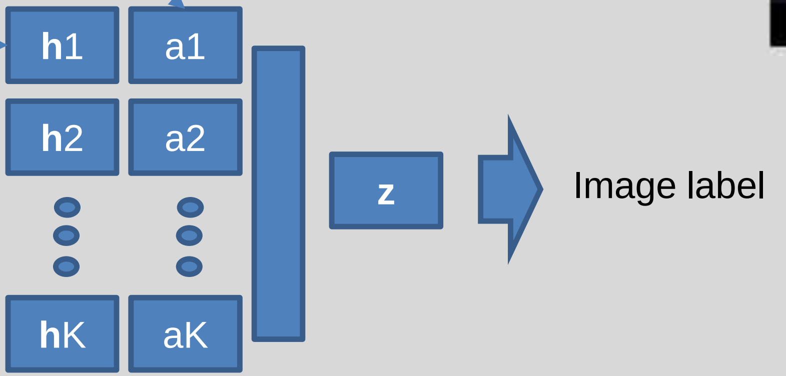
Weakly-supervised segmentation

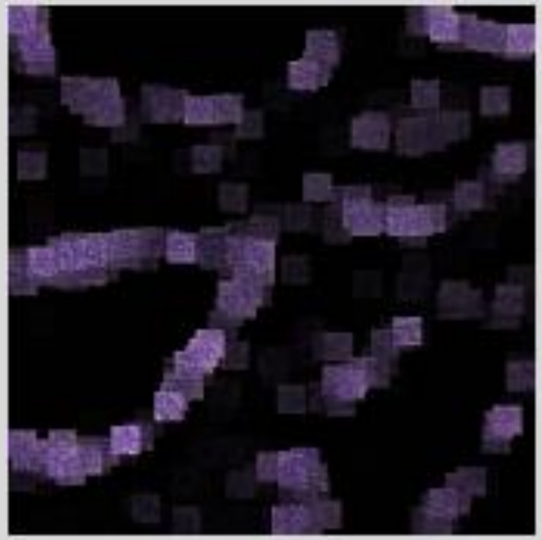
$$\mathbf{z} = \sum_{k=1}^{K} a_k \mathbf{h}_k$$

$$a_k = \frac{\exp\{\mathbf{w}^\top \tanh\left(\mathbf{V}\mathbf{h}_k^\top\right)\}}{\sum_{j=1}^{K} \exp\{\mathbf{w}^\top \tanh\left(\mathbf{V}\mathbf{h}_j^\top\right)\}}$$



Original image



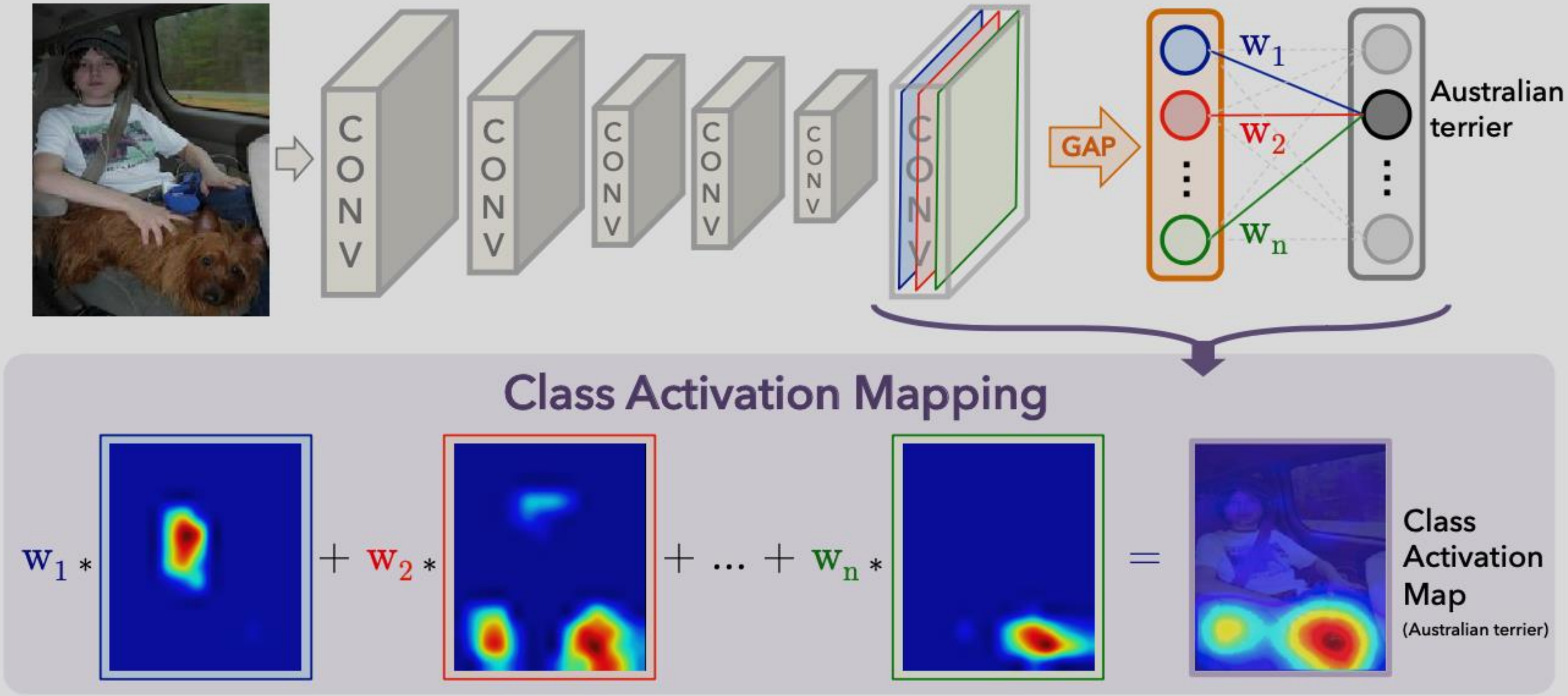


Predicted patch weights

Class Activation Map (CAM)

Learning Deep Features for Discriminative Localization, CVPR’16

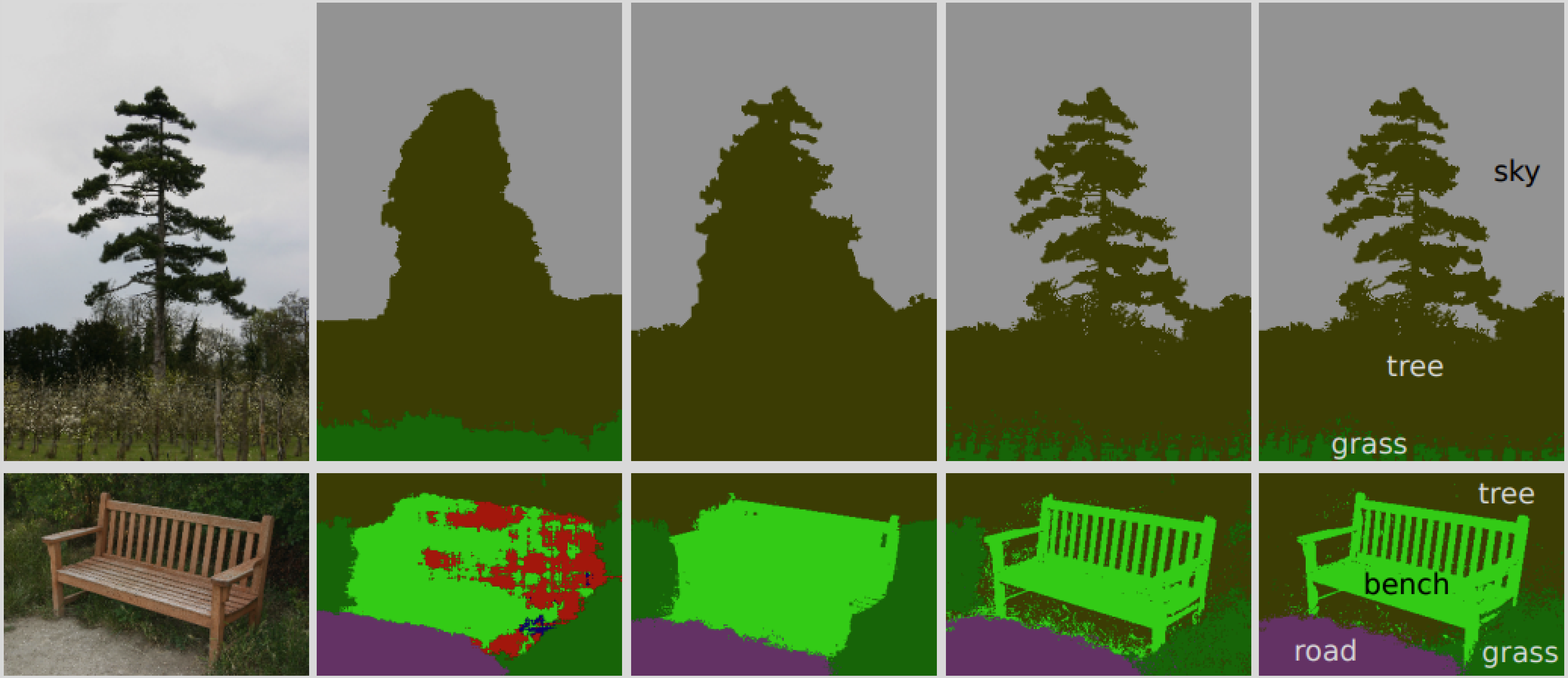
Class Activation Map (CAM)



$$S_c = \sum_k w_k^c \sum_{x,y} f_k(x,y)$$

$$M_c(x,y) = \sum_k w_k^c f_k(x,y)$$

CAM to Mask



(a) Image (b) Unary classifiers (c) Robust P^n CRF (d) Fully connected CRF, MCMC inference, 36 hrs (e) Fully connected CRF, our approach, 0.2 seconds

Efficient Inference in Fully Connected CRFs with Gaussian Edge Potentials, NIPS'12

CAM to Mask

$$E(\mathbf{x}) = \sum_{i} \psi_u(x_i) + \sum_{i<j} \psi_p(x_i, x_j)$$

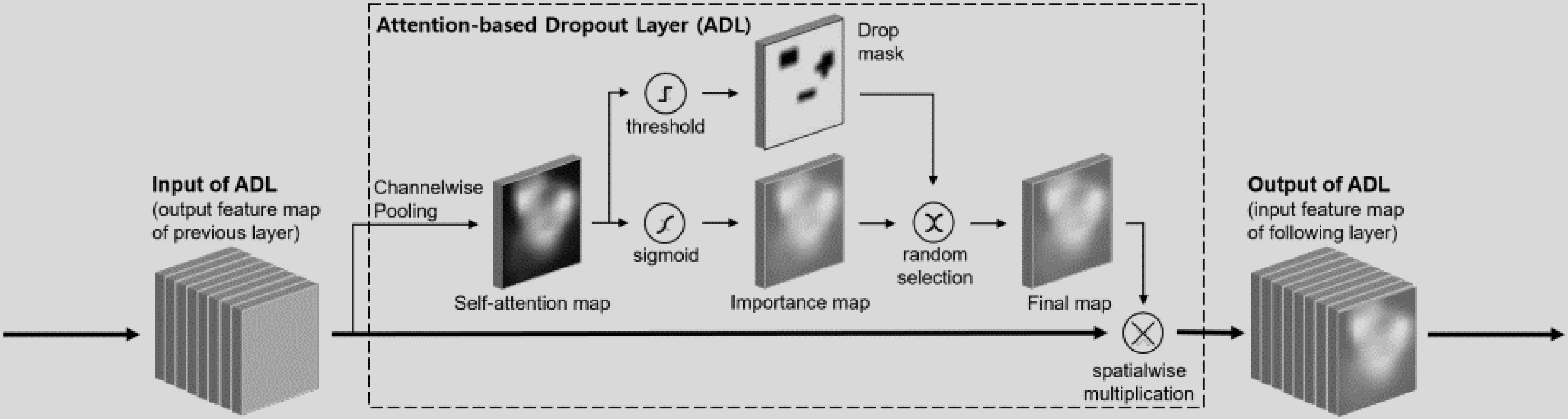
$$\psi_p(x_i, x_j) = \mu(x_i, x_j) \underbrace{\sum_{m=1}^{K} w^{(m)} k^{(m)}(\mathbf{f}_i, \mathbf{f}_j)}_{k(\mathbf{f}_i, \mathbf{f}_j)}$$

$$\mu(x_i, x_j) = [x_i \neq x_j]$$

$$k(\mathbf{f}_i, \mathbf{f}_j) = w^{(1)} \underbrace{\exp\left(-\frac{|p_i - p_j|^2}{2\theta_\alpha^2} - \frac{|I_i - I_j|^2}{2\theta_\beta^2}\right)}_{\text{appearance kernel}} + w^{(2)} \underbrace{\exp\left(-\frac{|p_i - p_j|^2}{2\theta_\gamma^2}\right)}_{\text{smoothness kernel}}$$

Efficient Inference in Fully Connected CRFs with Gaussian Edge Potentials, NIPS’12

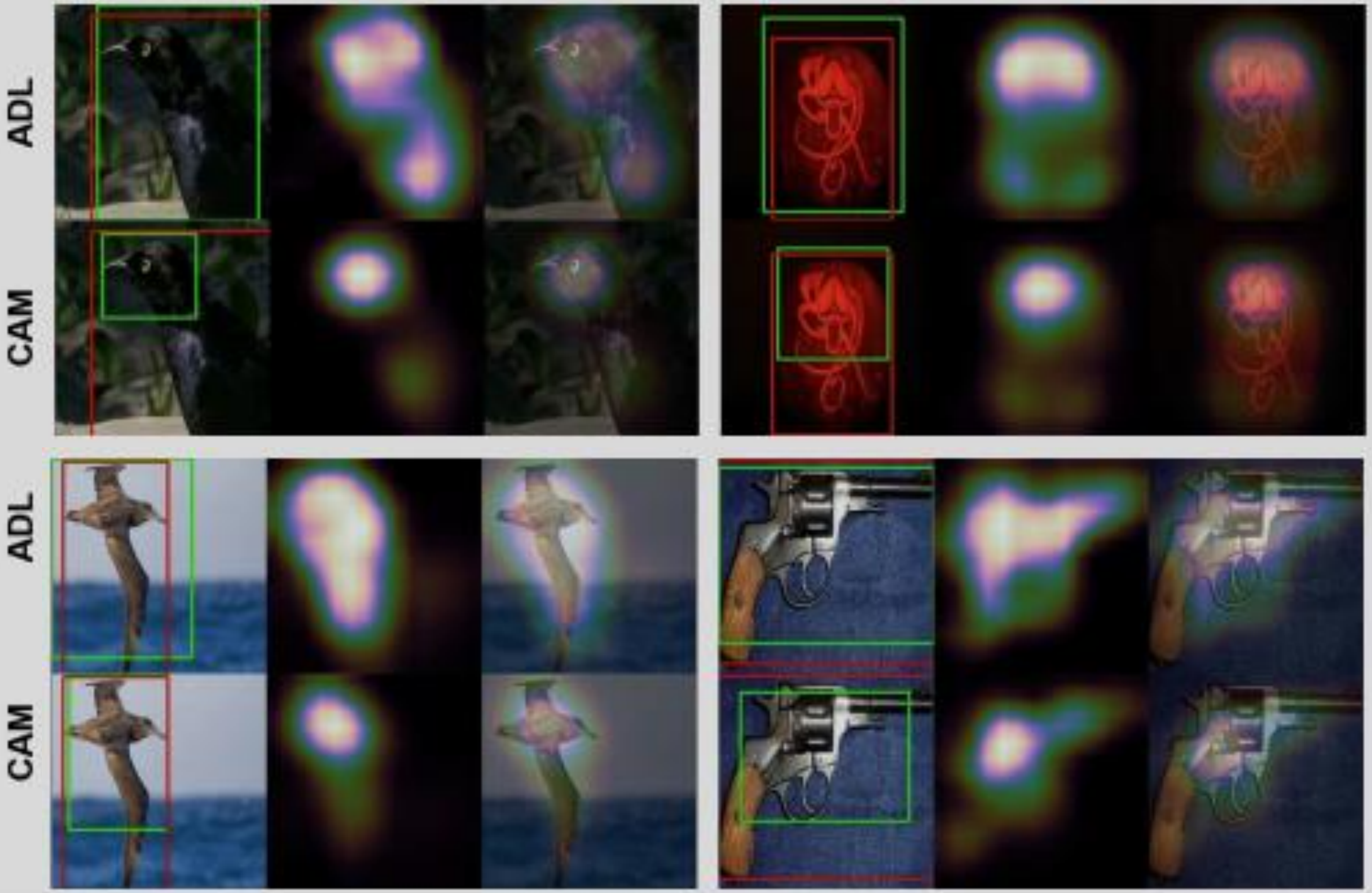
Limitation of CAM



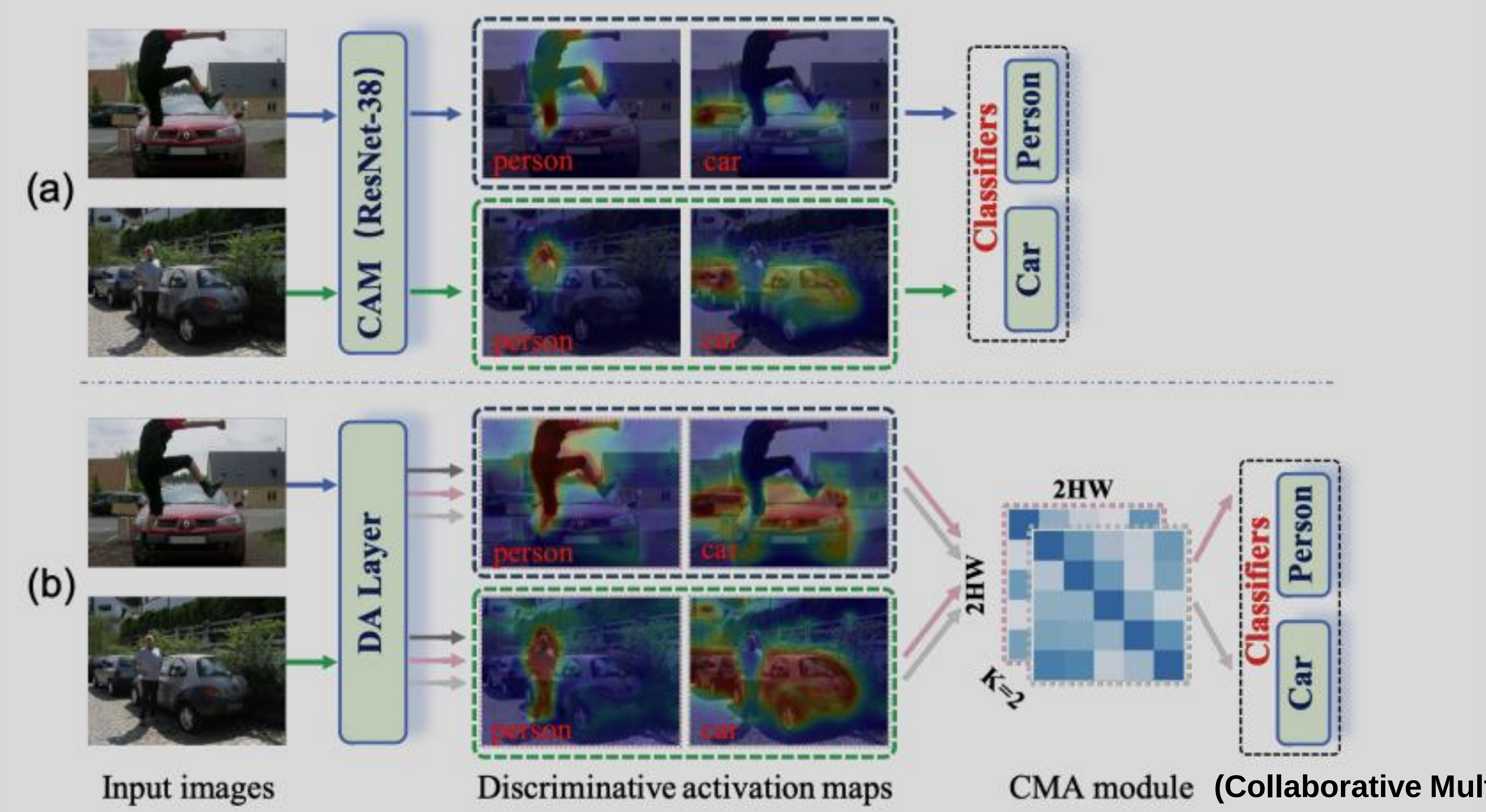
CAM covers only the most discriminative part of the object, not the entire object.

Attention-based Dropout Layer for Weakly Supervised Object Localization, CVPR'19

Limitation of CAM



Weakly-supervised segmentation

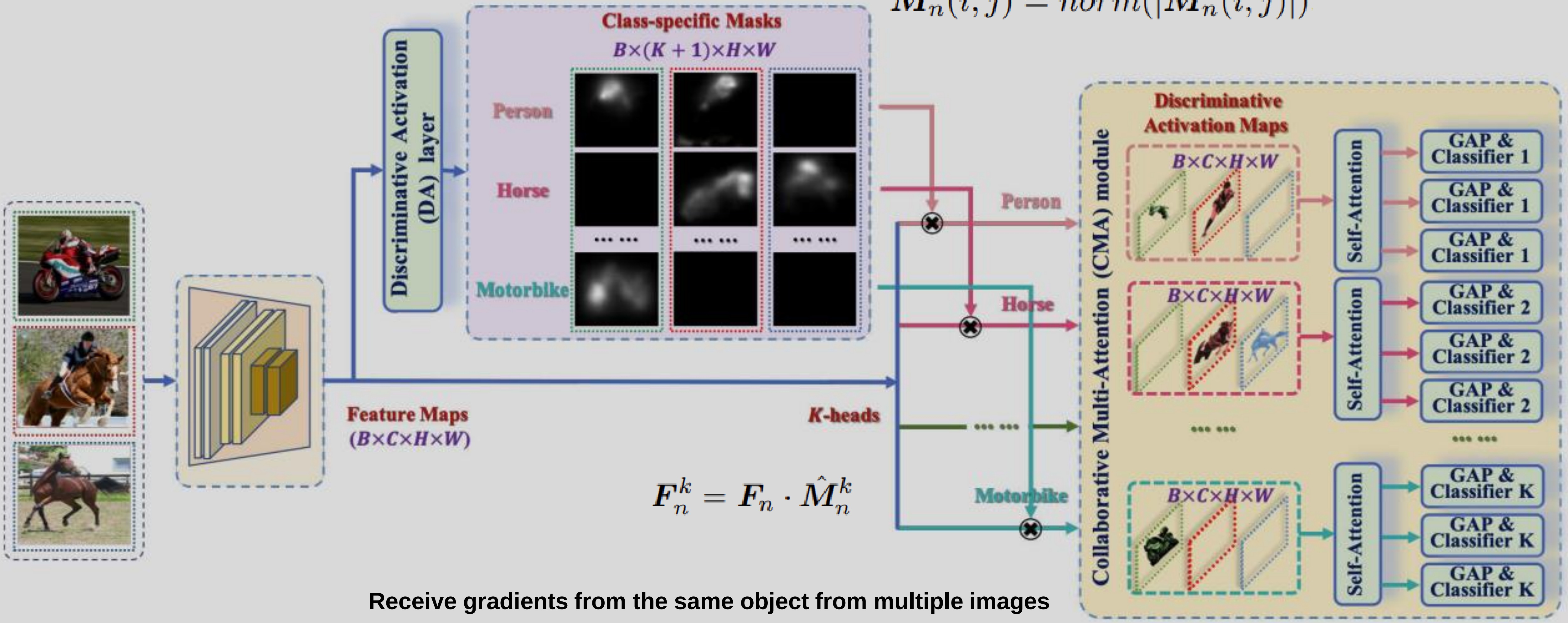


(Collaborative Multi-Attention)

Embedded Discriminative Attention Mechanism for Weakly Supervised Semantic Segmentation, CVPR'21

Weakly-supervised segmentation

$$\hat{\boldsymbol{M}}_n(i,j) = norm(|\boldsymbol{M}_n(i,j)|)$$



$$\boldsymbol{F}_n^k = \boldsymbol{F}_n \cdot \hat{\boldsymbol{M}}_n^k$$

Receive gradients from the same object from multiple images

Weakly-supervised segmentation

$$\mathcal{F}^k = [\boldsymbol{F}_1^k, \boldsymbol{F}_2^k, ..., \boldsymbol{F}_B^k] \in \mathbb{R}^{B \times C \times H \times W}$$

$$\hat{\mathcal{F}}^k \in \mathbb{R}^{1 \times (B \times H \times W) \times d}$$

$$[\boldsymbol{A}_1^k, \boldsymbol{A}_2^k, ..., \boldsymbol{A}_B^k] = SelfAttention(\hat{\mathcal{F}}^k)$$

$$\mathcal{L}_{cls} = \frac{1}{B \times K} \sum_{n=1}^{B} \sum_{k=1}^{K} \mathcal{L}_{BCE}(Linear(GAP(\boldsymbol{A}_n^k)), \boldsymbol{l}_n^k)$$

Summary

- Supervised learning requires us to collect large-scale data for our own applications.
- Weakly-supervised, Self-supervised learning methods provide a way to prevent collecting data.

Thank you!

UNIST
ULSAN NATIONAL INSTITUTE OF SCIENCE AND TECHNOLOGY
2007