Weakly-supervised Learning for Detection and Segmentation

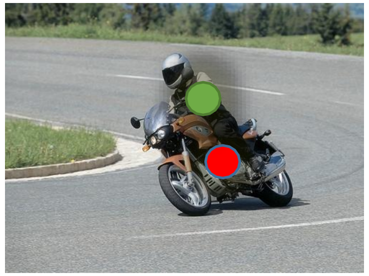
Meng Tang

May 30, 2019

Manual supervision for object recognition

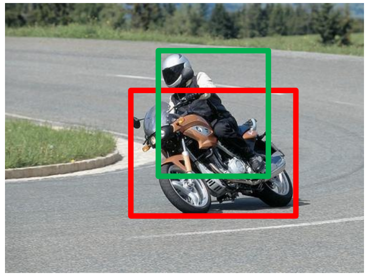
{motorbike,person}

1 sec
per class



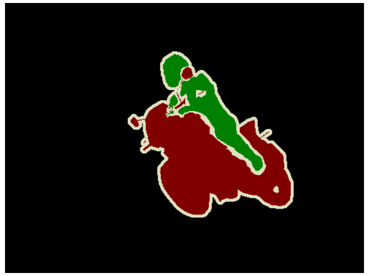
{motorbike (point),
person (point)}

2.4 sec
per instance



{motorbike (b-box),
person (b-box)}

10 sec
per instance



{motorbike (pixel labels),
person (pixel labels)}

78 sec
per instance

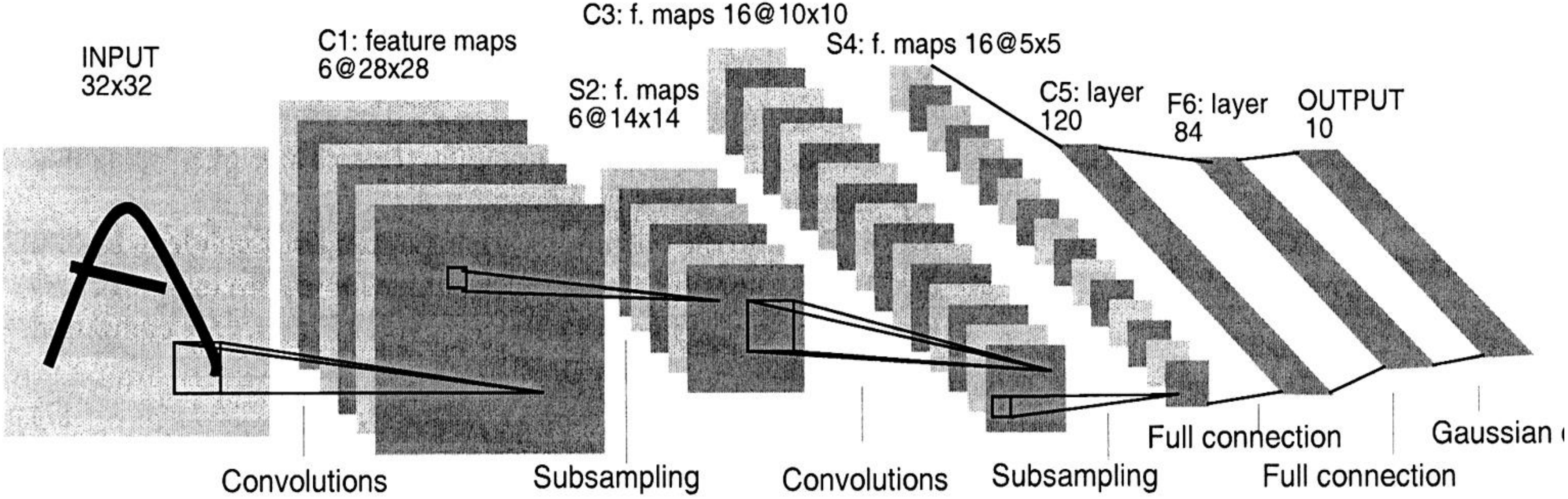
Weak supervision

Lower degree (or cheaper) annotation at train time than the required output at test time

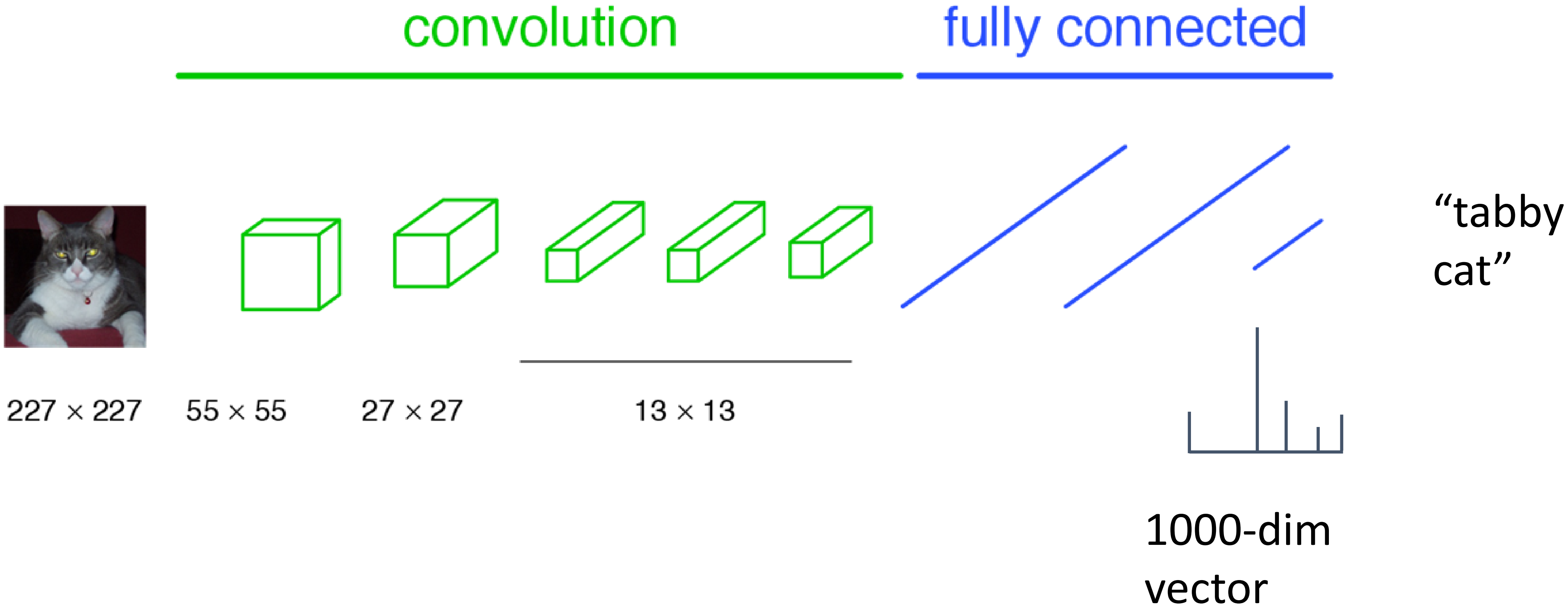
Slide credit: Hakan Bilen

Part I: Weakly-supervised Semantic Segmentation

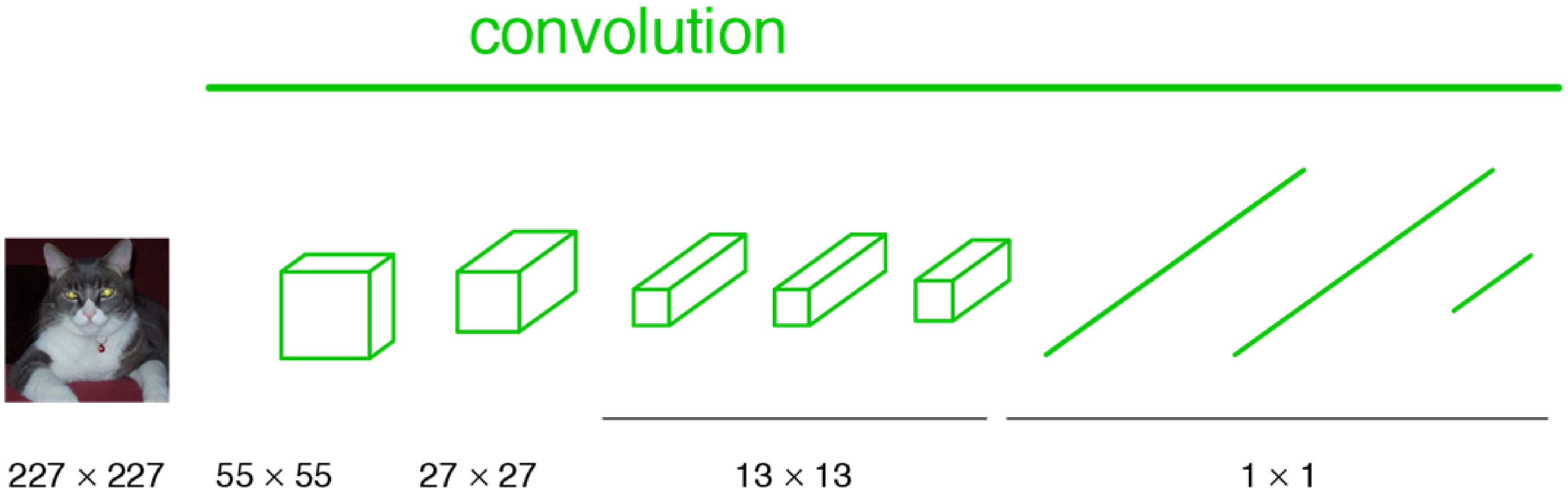
The architecture of LeNet5



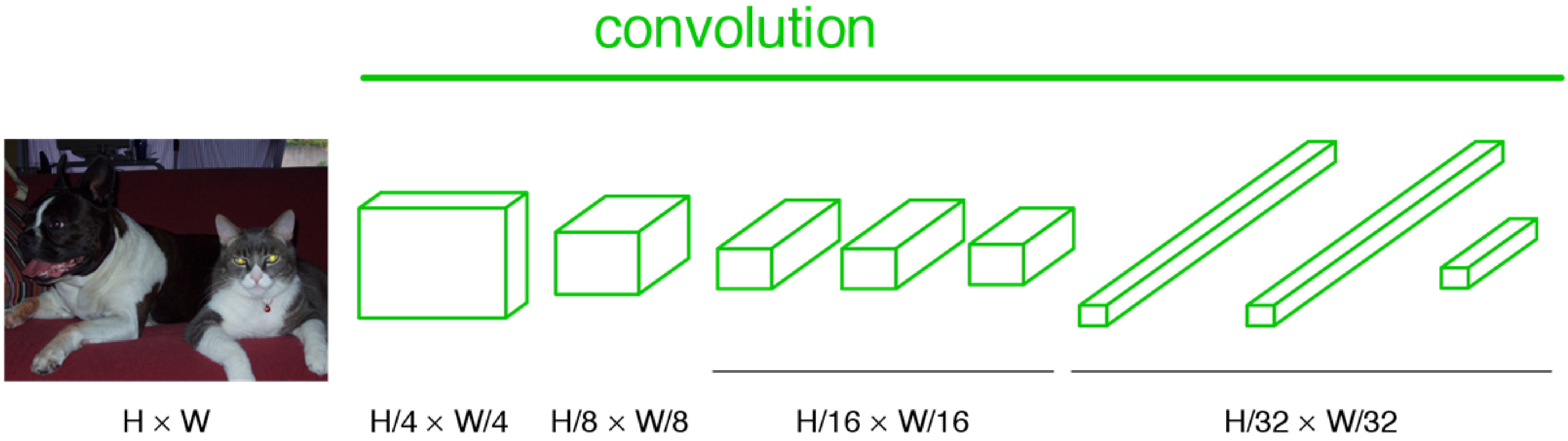
a classification network



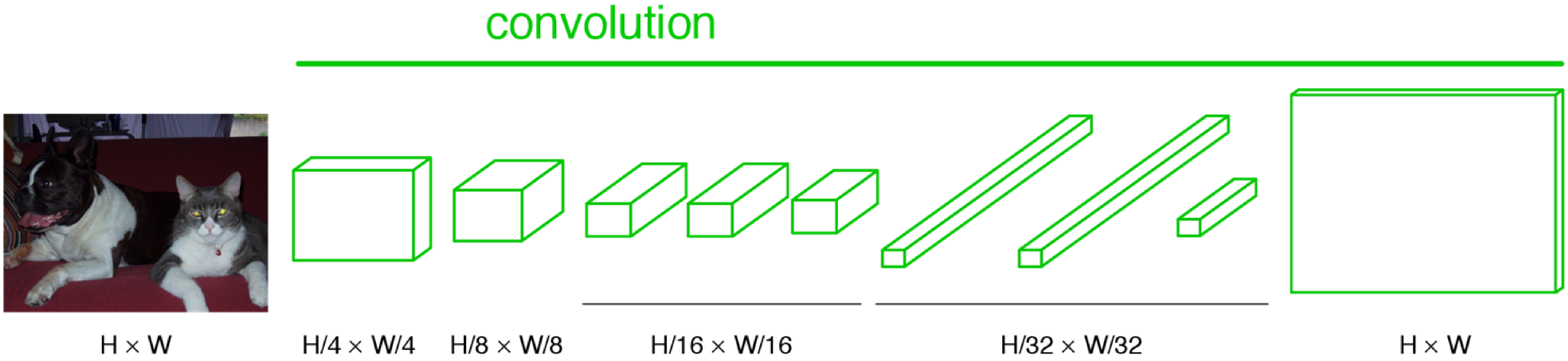
becoming fully convolutional



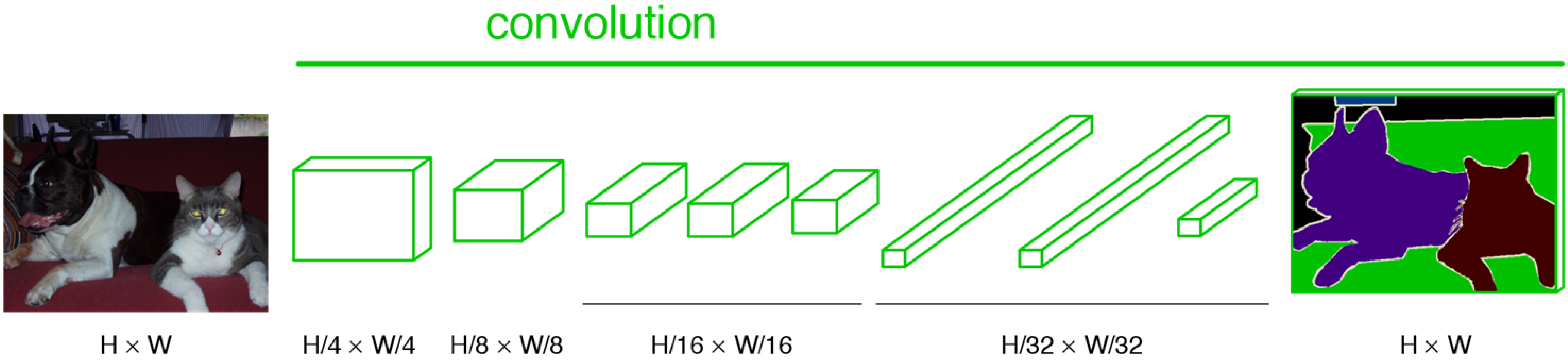
becoming fully convolutional



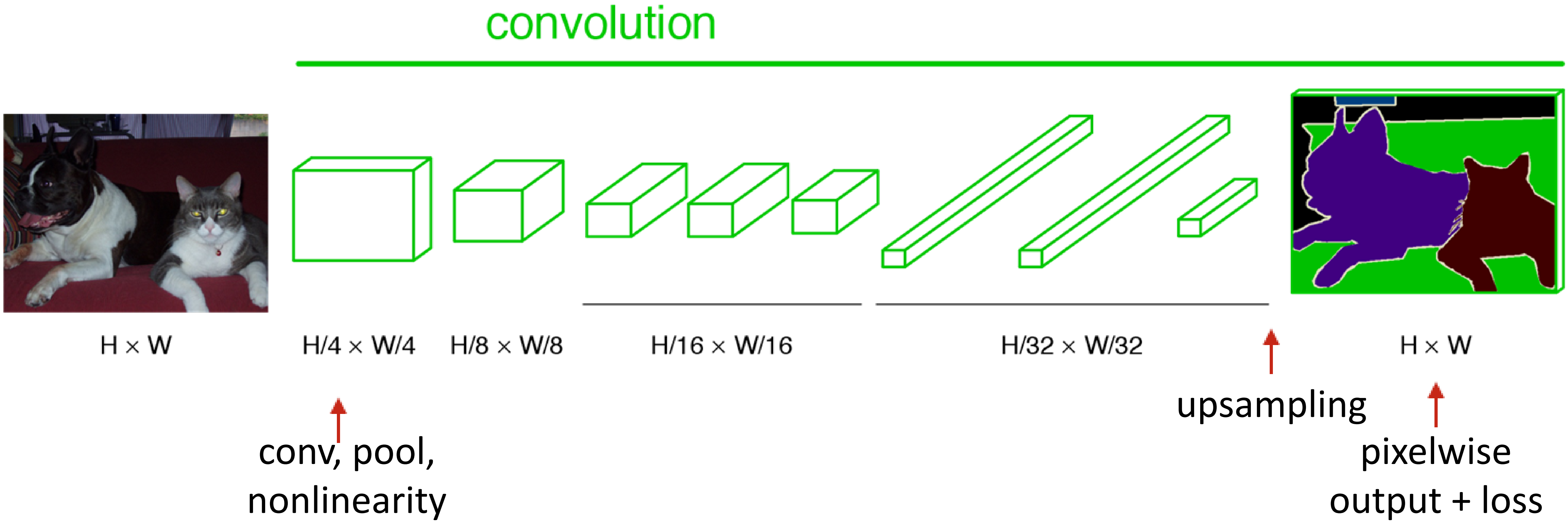
upsampling output



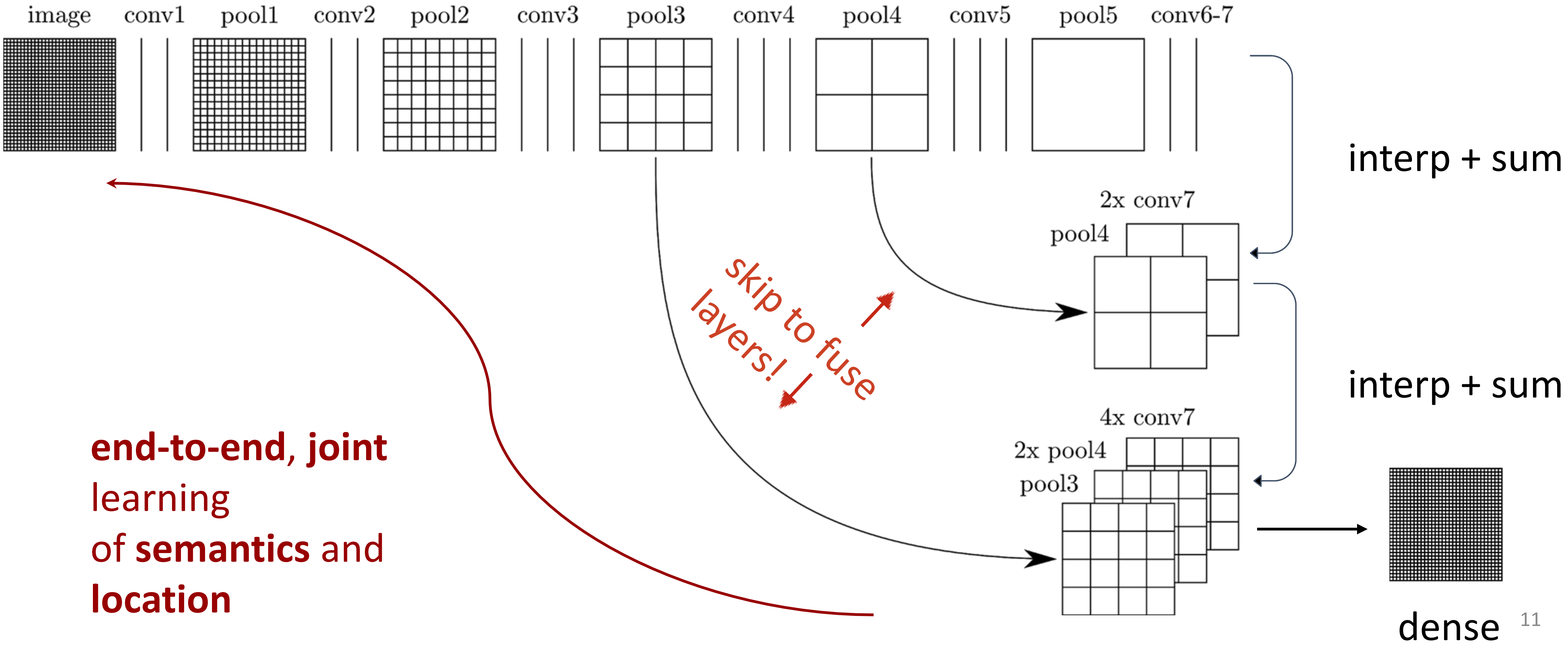
end-to-end, pixels-to-pixels network



end-to-end, pixels-to-pixels network



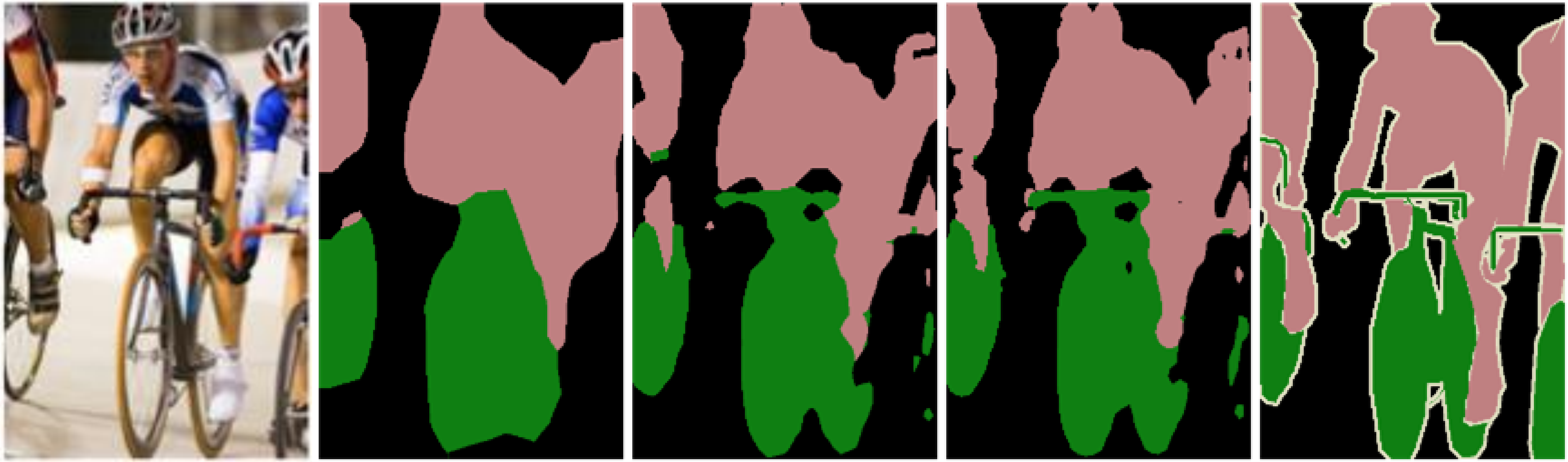
skip layers



end-to-end, joint learning
of **semantics** and **location**

skip layer refinement

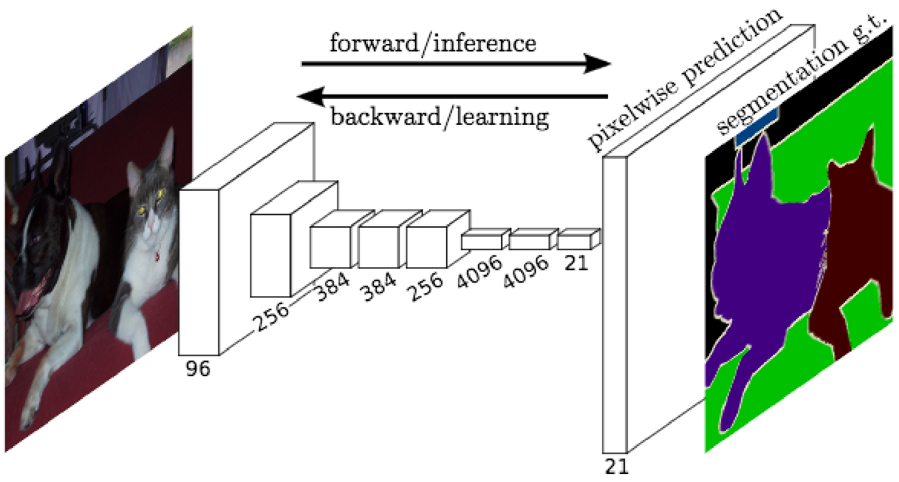
input image | stride 32 | stride 16 | stride 8 | ground truth



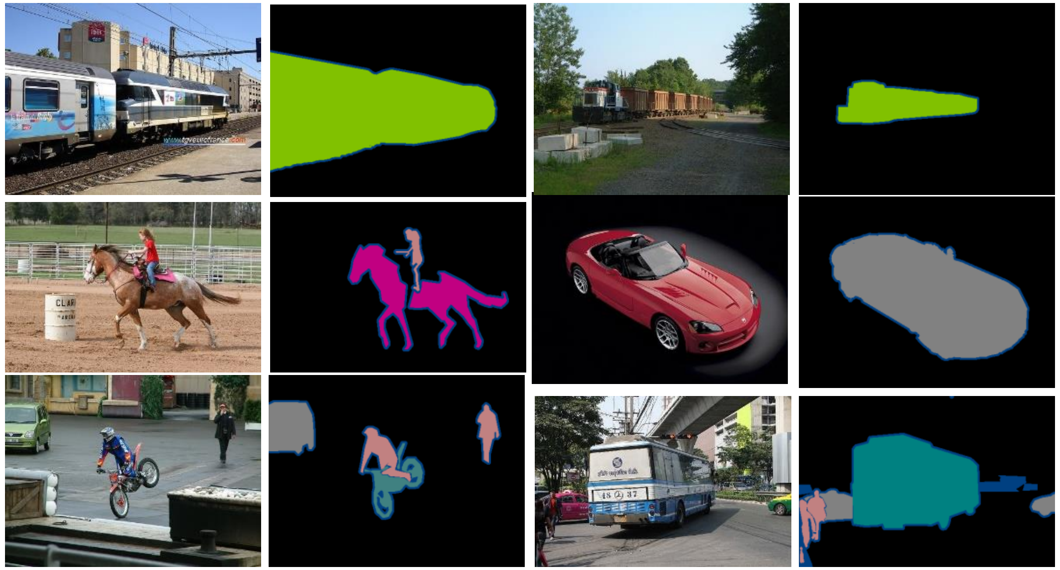
no skips | 1 skip | 2 skips

Fully-supervised CNN Segmentation

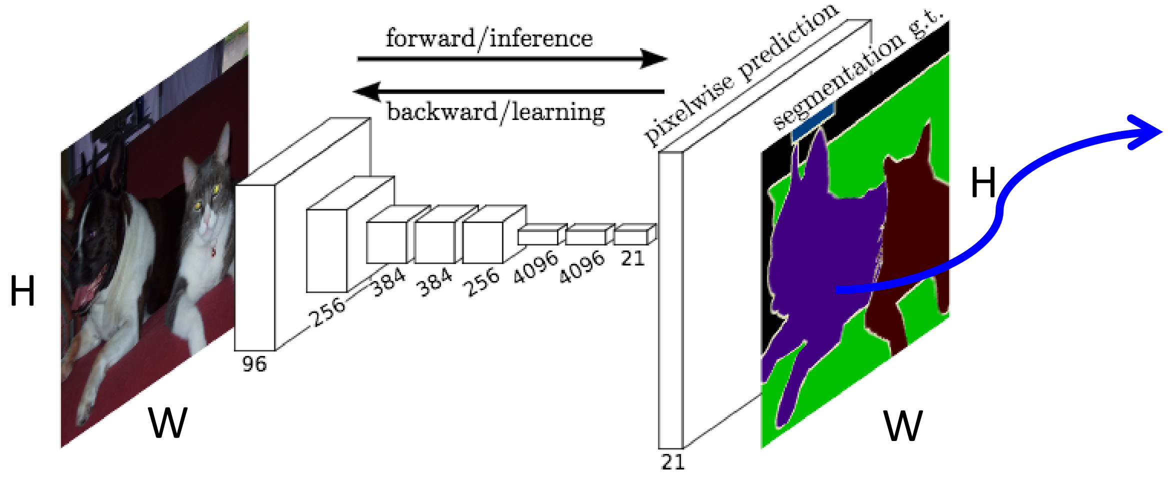
Network



Training Data



Losses for CNN Segmentation



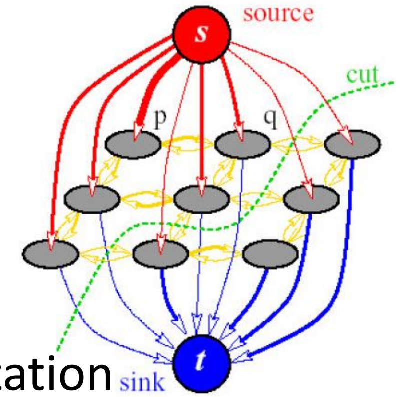
	Ground truth	Output distribution
cat	0	0.1
dog	1	0.8
sofa	0	0.05
…	…	…

pixel-wise Cross Entropy (CE) loss:

$- 0 \times \log 0.1 - 1 \times \log 0.8 - 0 \times \log 0.05 \ldots$

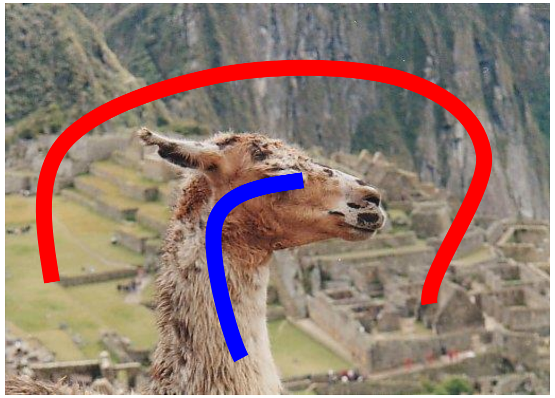
Scribbles Supervised Semantic Segmentation

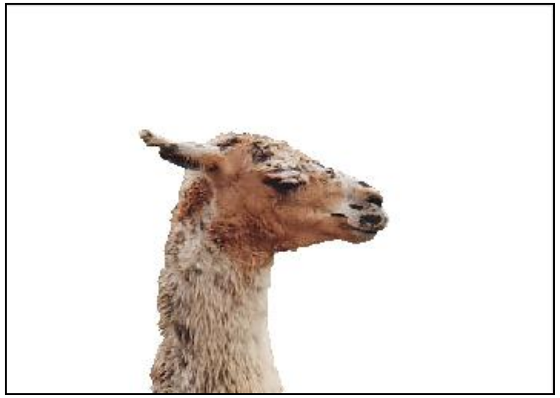
Markov Random Field for Segmentation

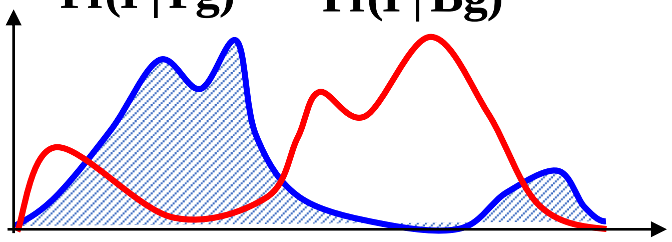


Without Regularization

With Regularization







$$E(S,\theta_0,\theta_1) = \sum_{k=0,1} \sum_{p \in S^k} -\ln \mathrm{P}(I_p|\theta_k) + \boxed{\lambda \cdot \sum_{pq \in \mathcal{N}} w_{pq} \cdot [s_p \neq s_q]}$$

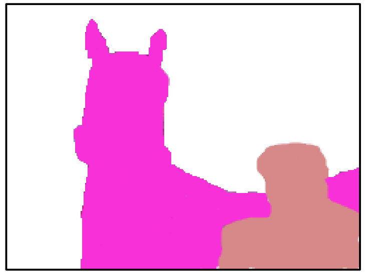
MRF regularization

[Boykov, Jolly, *ICCV* 2001]

Pipeline of previous work

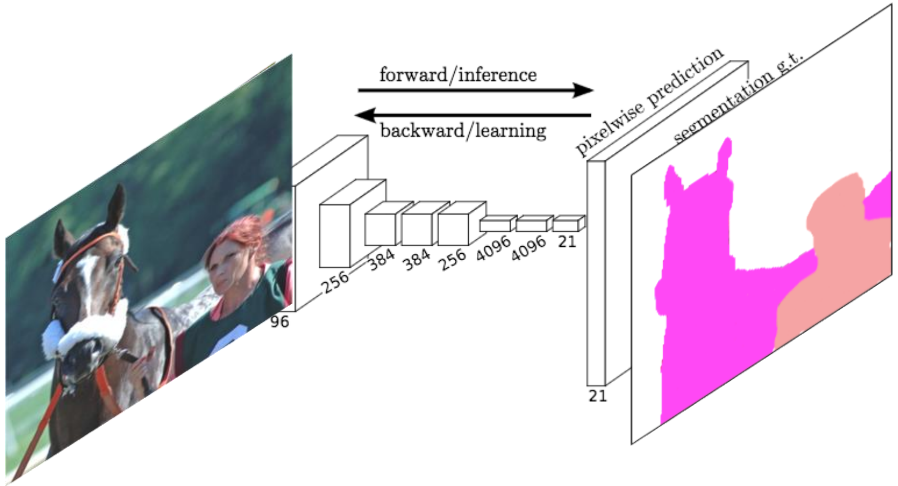
[Dai *et al. ICCV* 2015]
[Khoreva *et al. CVPR* 2017]
[Kolesnikov *et al. ECCV* 2016]
[Lin *et al. CVPR* 2016]

Interactive Segmentation



proposals

Network Training



Proposal Generation

All weak supervision method generates “fake” proposals

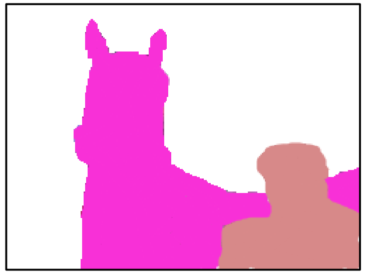
Input scribbles GraphCut Ground Truth

What’s wrong with proposal generation?

- Training is sensitive to the quality of proposals
- How to obtain good proposals?
- Mistakes mislead training to fit errors

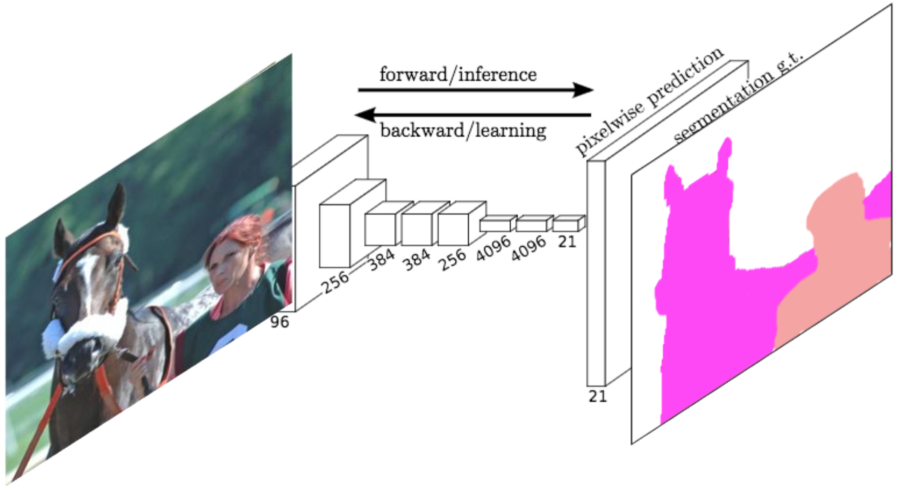
Train without (Full but Fake) Proposals?

Interactive Segmentation

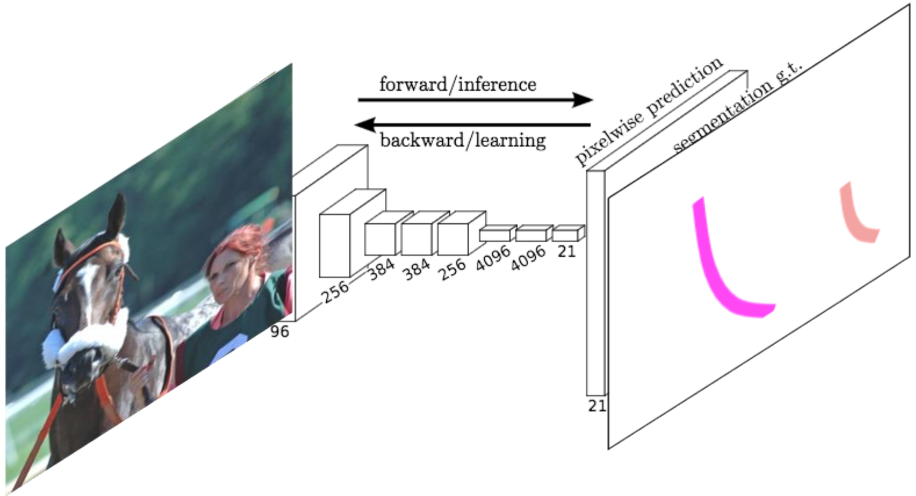


proposals

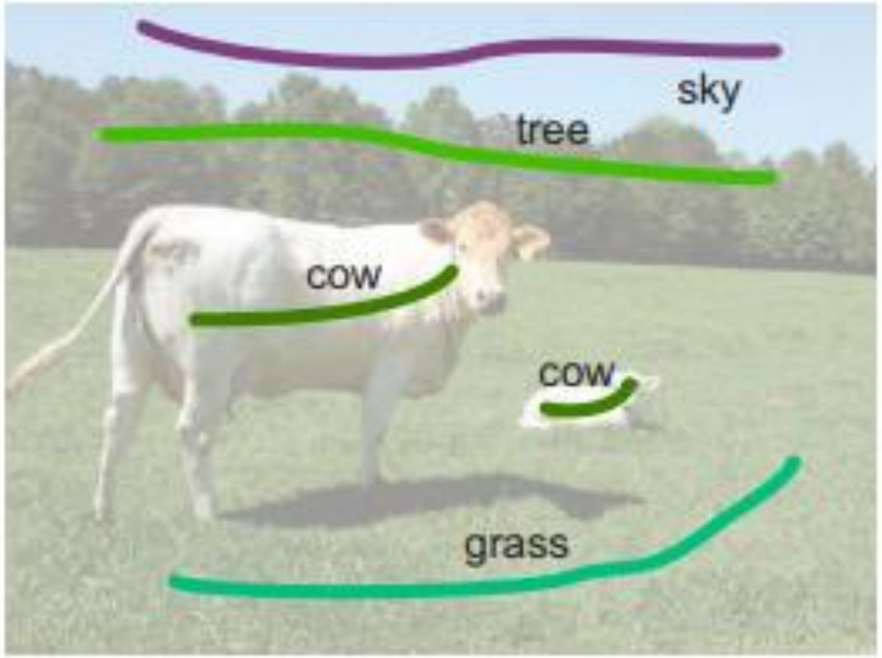
Network Training



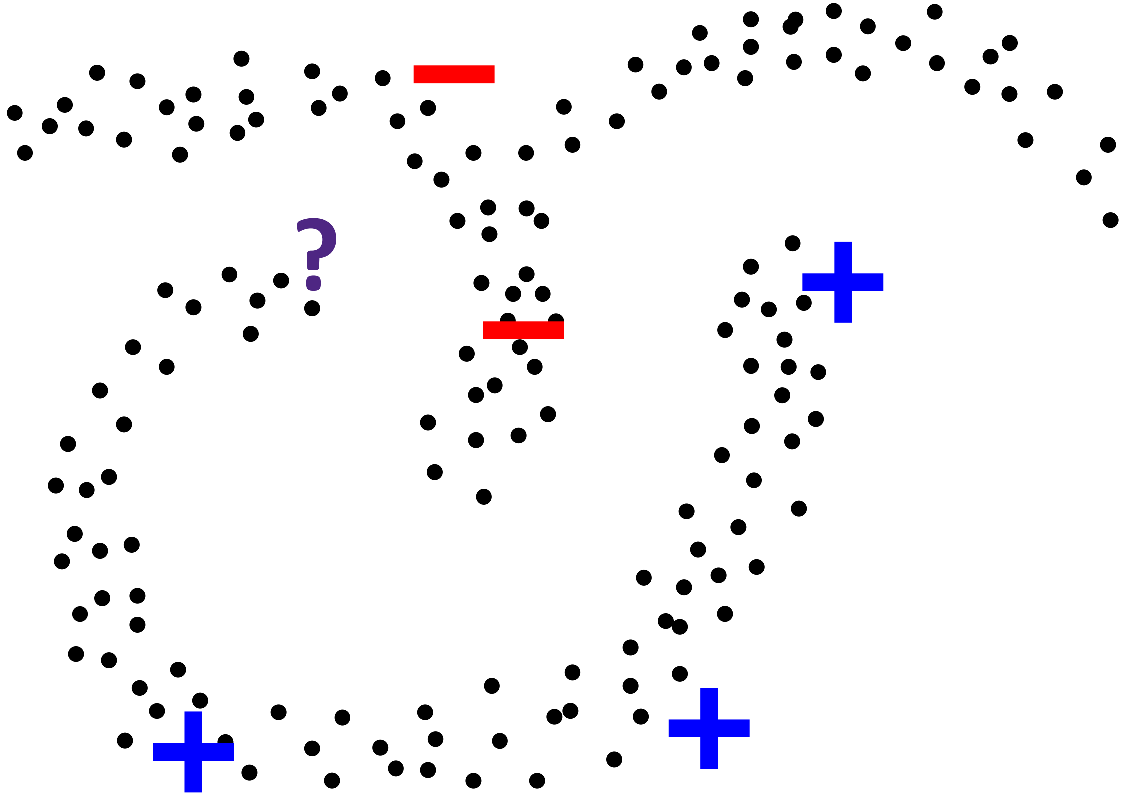
Can we train directly?



Weakly-supervised segmentation

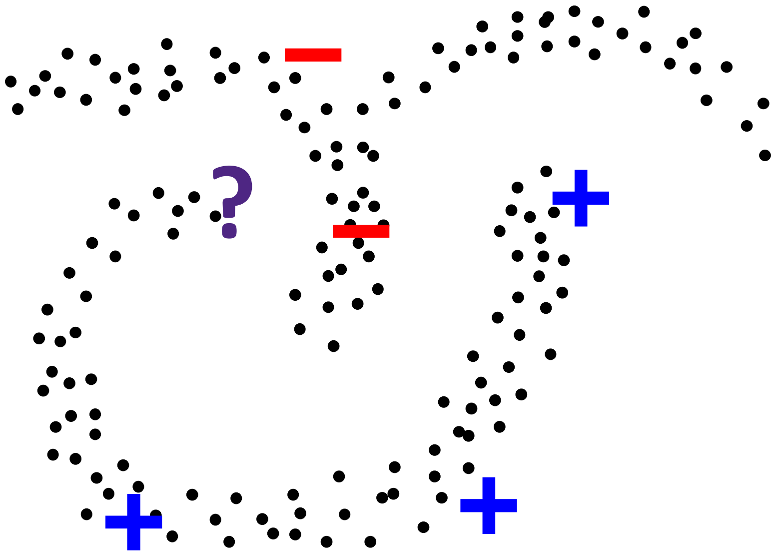


Semi-supervised learning



Semi-supervised learning

Definition Given M labeled data $(x_i, y_i) \in (\mathcal{X}, \mathcal{Y})$,$i = 1, ..., M$ and U unlabeled data $x_i, i = M+1, ..., M+U$, learn $f(x) : \mathcal{X} \to \mathcal{Y}$.

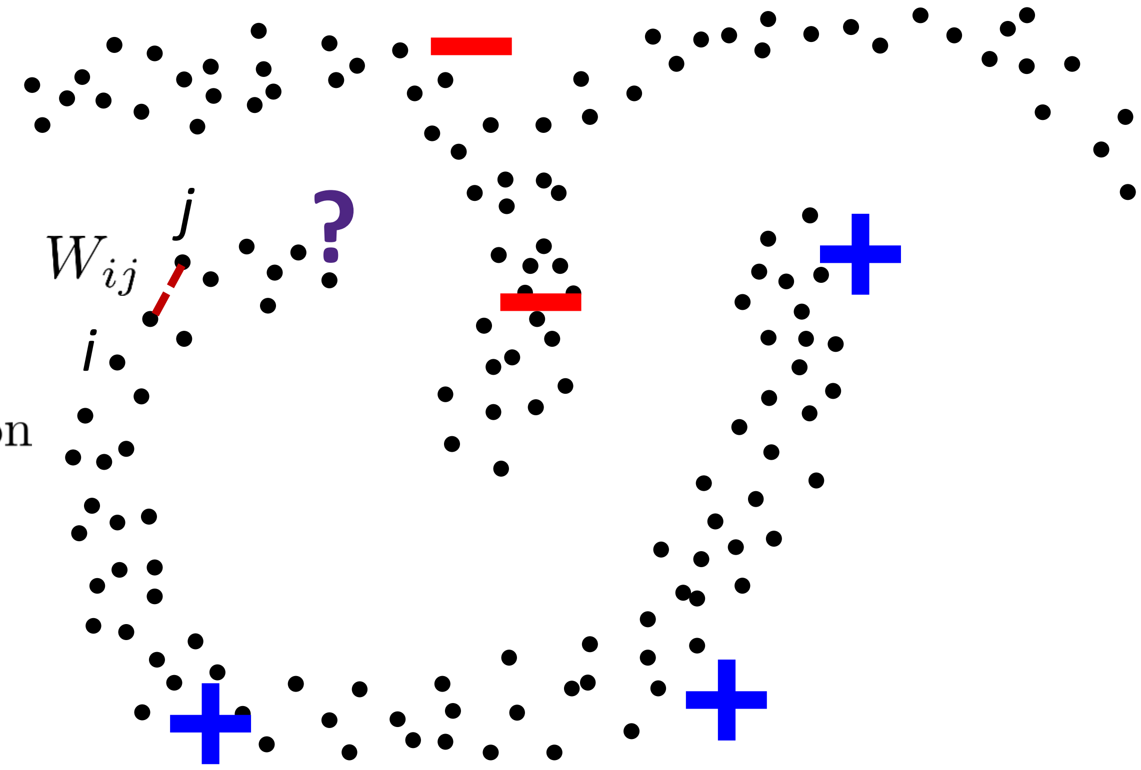


[Zhu & Goldberg, "Introduction to semi-supervised learning", 2009]
[Chapelle, Scholkopf & Zien, "Semi-supervised learning", 2009]

Graph-Based Semi-supervised Learning

labelled points $\rightarrow$ target labeling
e.g. $\sum_{i=1}^{M} \delta(f(x_i) \neq y_i)$

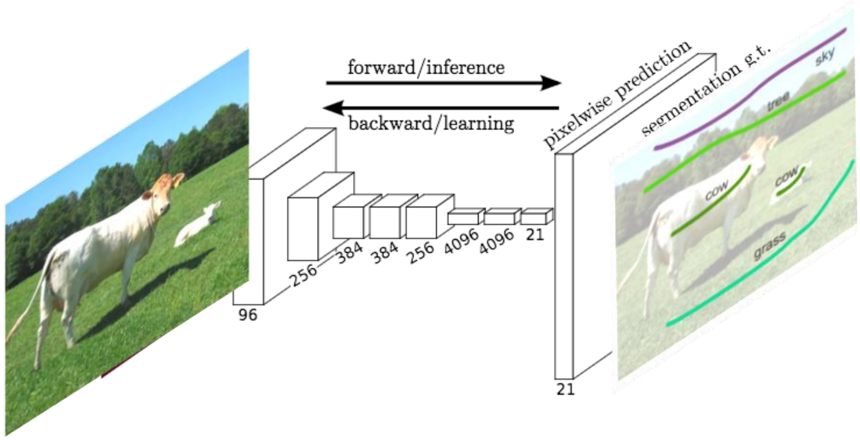
unlabelled points $\rightarrow$ pairwise regularization
e.g. $\sum_{ij} W_{ij} \cdot ||f(x_i) - f(x_j)||^2$

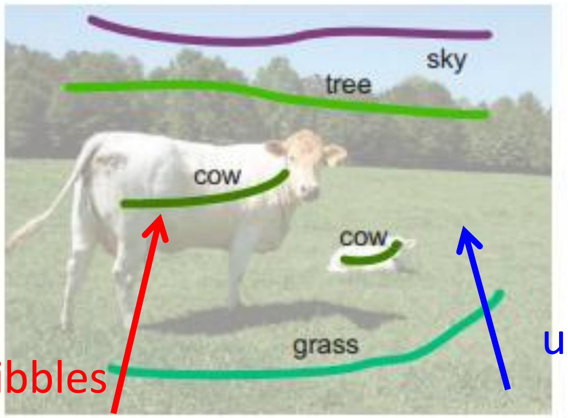


[Tang, Djelouah, Perazzi, Boykov, Schroers, *CVPR* 2018]

[Tang, Perazzi, Djelouah, Ben Ayed, Schroers, Boykov, *ECCV* 2018]

Regularized loss for weakly-supervised CNN segmentation





scribbles

unknown pixels

empirical risk Loss
for *labeled* data

regularization Loss
for *unlabeled* data

$$\boxed{\sum_{i=1}^{M} \ell(f_\theta(x_i), y_i)} \quad + \quad \lambda \cdot \boxed{R(f)}$$

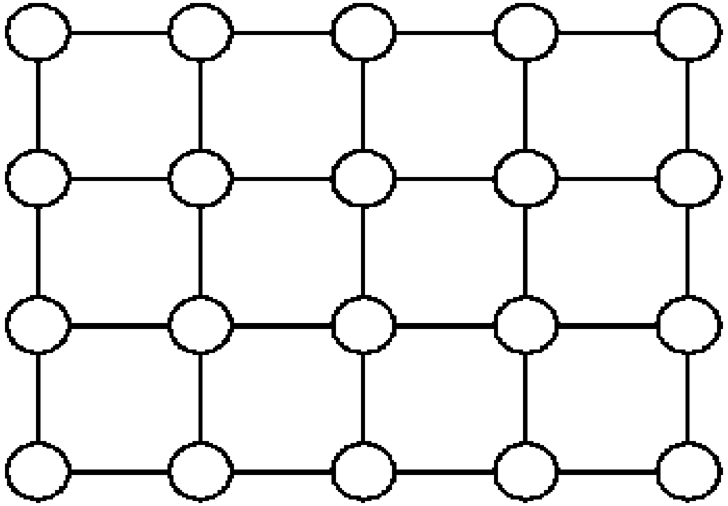
partial Cross Entropy (PCE)

e.g. MRF, NC or both

$$\sum_{ij} W_{ij} \cdot ||f(x_i) - f(x_j)||^2$$

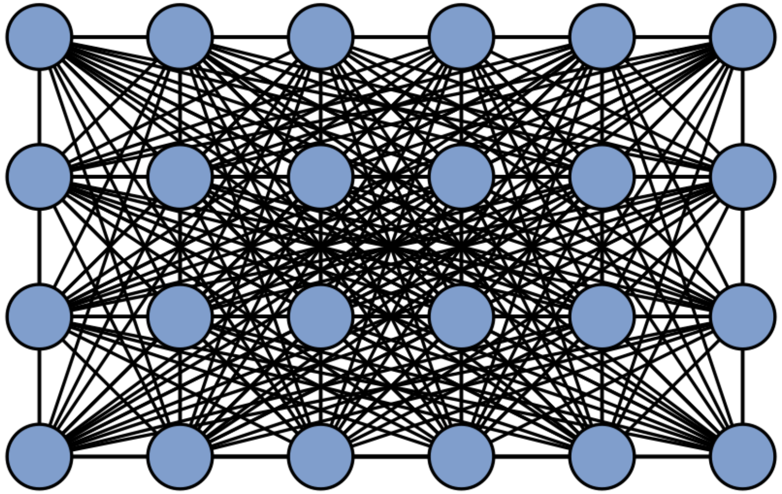
Pairwise MRF regularization as loss

$$\sum_{ij} W_{ij} \cdot ||f(x_i) - f(x_j)||^2$$



Sparse Connected Potts

[Boykov and Jolly, ICCV 2001]



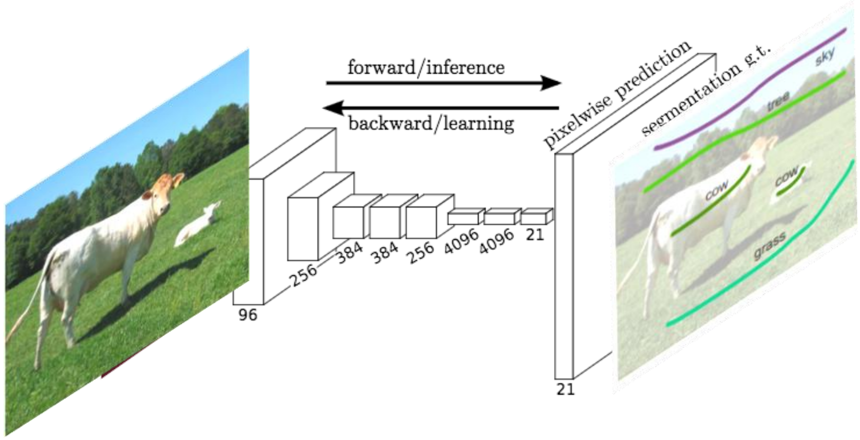
Fully Connected DenseCRF

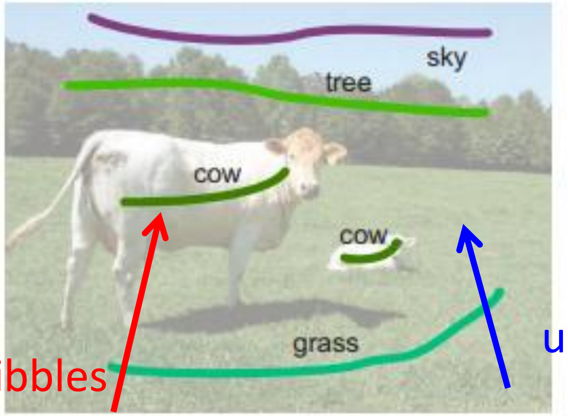
[Krähenbühl and Vladlen Koltun, NIPS 2011]

[Tang, Djelouah, Perazzi, Boykov, Schroers, *CVPR* 2018]

[Tang, Perazzi, Djelouah, Ben Ayed, Schroers, Boykov, *ECCV* 2018]

Regularized loss for weakly-supervised CNN segmentation





scribbles

unknown pixels

empirical risk Loss for *labeled* data

regularization Loss for *unlabeled* data

$$\boxed{\sum_{i=1}^{M} \ell(f_\theta(x_i), y_i)} \quad + \quad \lambda \boxed{\cdot R(f)}$$

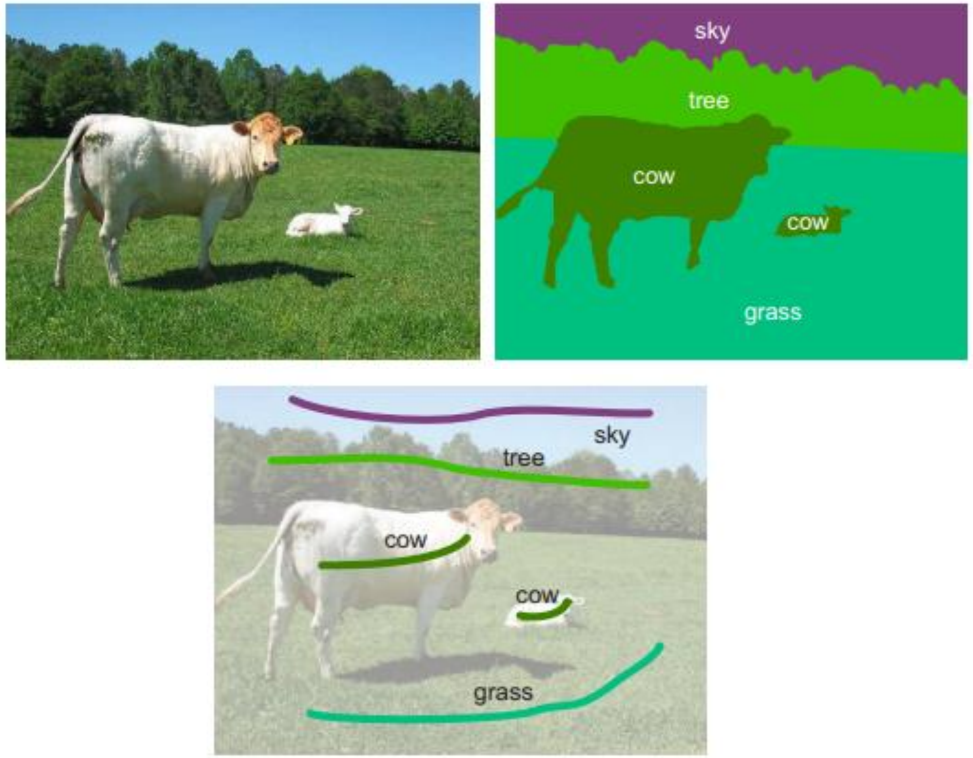
partial Cross Entropy (PCE)

e.g. MRF, NC or both

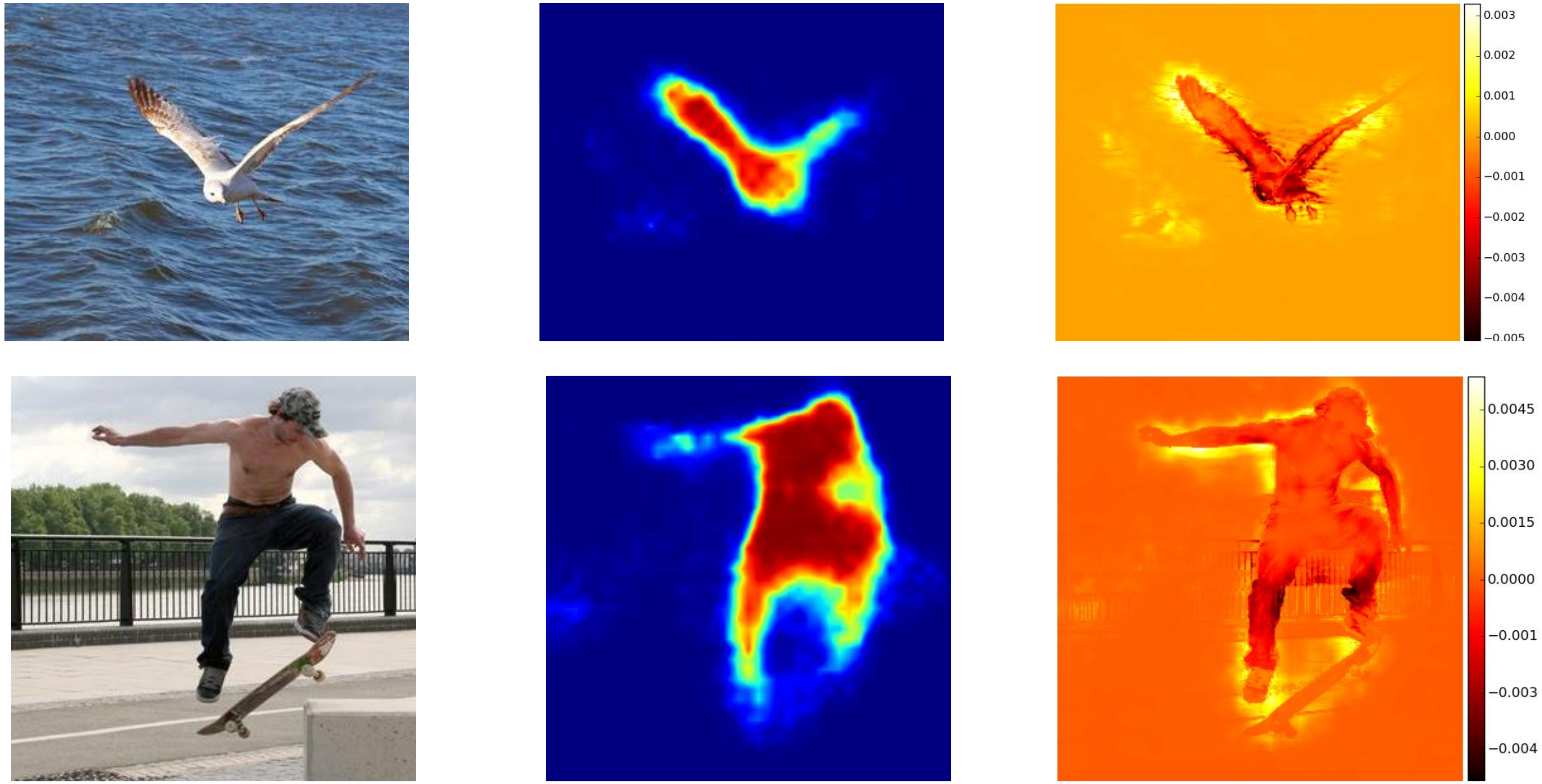
$$\sum_{ij} W_{ij} \cdot ||f(x_i) - f(x_j)||^2$$

Experiments

- PASCAL VOC 2012 Segmentation Dataset
  - 10K training images (full masks)
  - 1.5K validation images
  - 1.5K test images
- ScribbleSup Dataset [Dai *et al.* ICCV 2015]
  - scribbles for each object
  - ~3% of pixels labelled



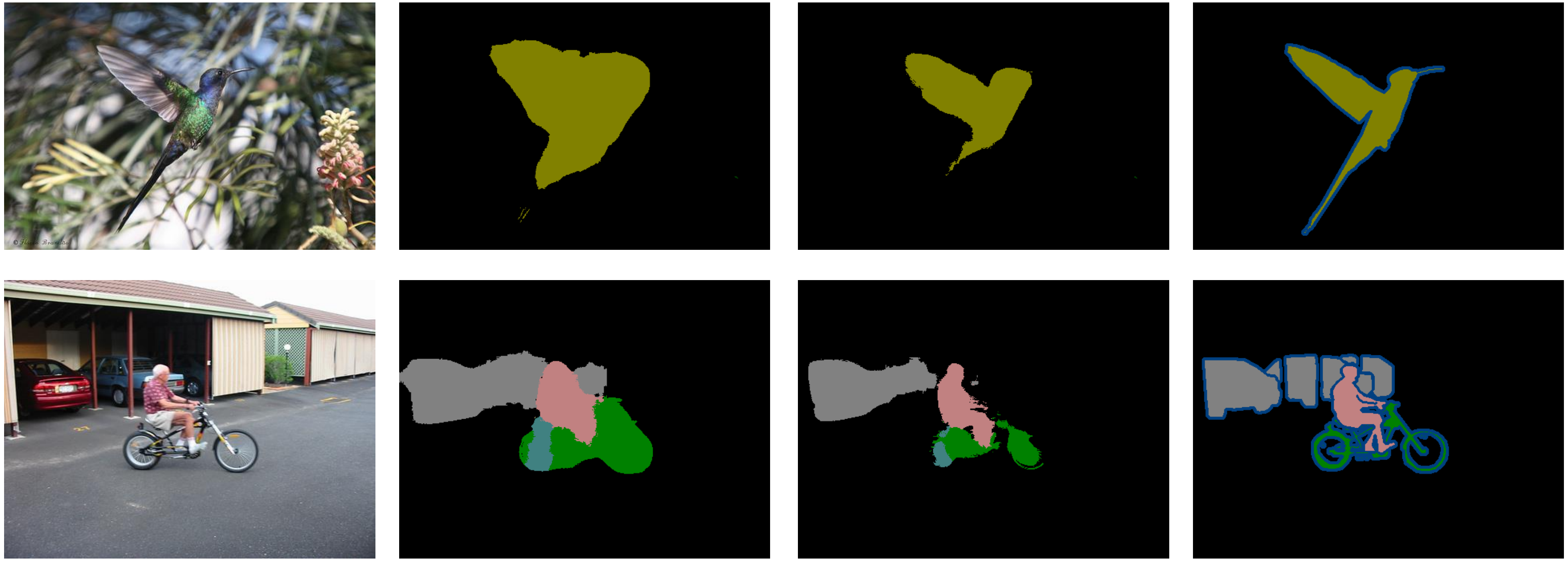
Visualization of Gradients



input

network output f_θ gradient of regularization loss $\frac{\partial R(f)}{\partial f}$

Training with regularized losses



Test image | pCE loss (unregularized) | w/ regularized loss | Ground truth

better color clustering

better edge alignment

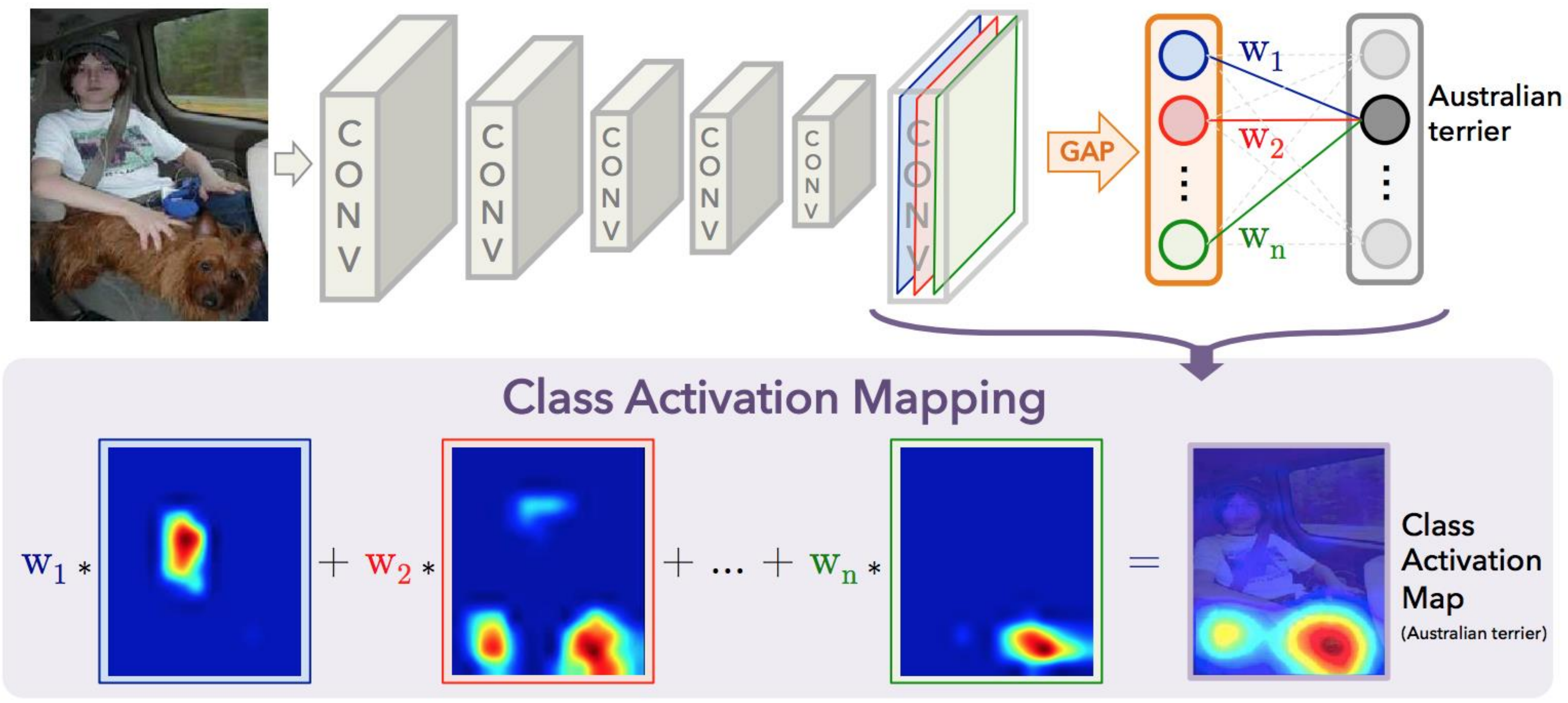
Compare weak and full supervision

	Weak				Full
	CE only	w/ NC [68]	w/ CRF	w/ KernelCut	
DeepLab-MSc-largeFOV	56.0 (8.1)	60.5 (3.6)	63.1 (1.0)	**63.5 (0.6)**	64.1
DeepLab-VGG16	60.4 (8.4)	62.4 (6.4)	64.4 (4.4)	**64.8 (4.0)**	68.8
DeepLab-ResNet101	69.5 (6.1)	72.8 (2.8)	72.9 (2.7)	**73.0 (2.6)**	75.6

Table 2: mIOU on PASCAL VOC2012 *val* set. Our flexible framework allows various types of regularization losses for weakly supervised segmentation, e.g. normalized cut, CRF or their combinations (KernelCut [69]) as joint loss. We achieved the state-of-the-art with scribbles. In () shows the offset to the result with full masks.

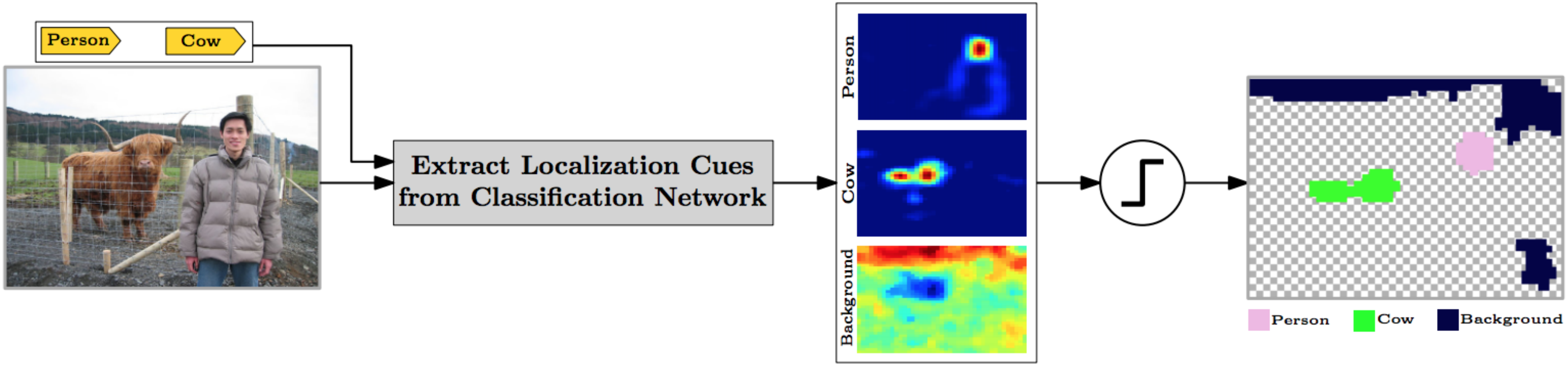
Class Activation Map

[Zhou et al. CVPR16]



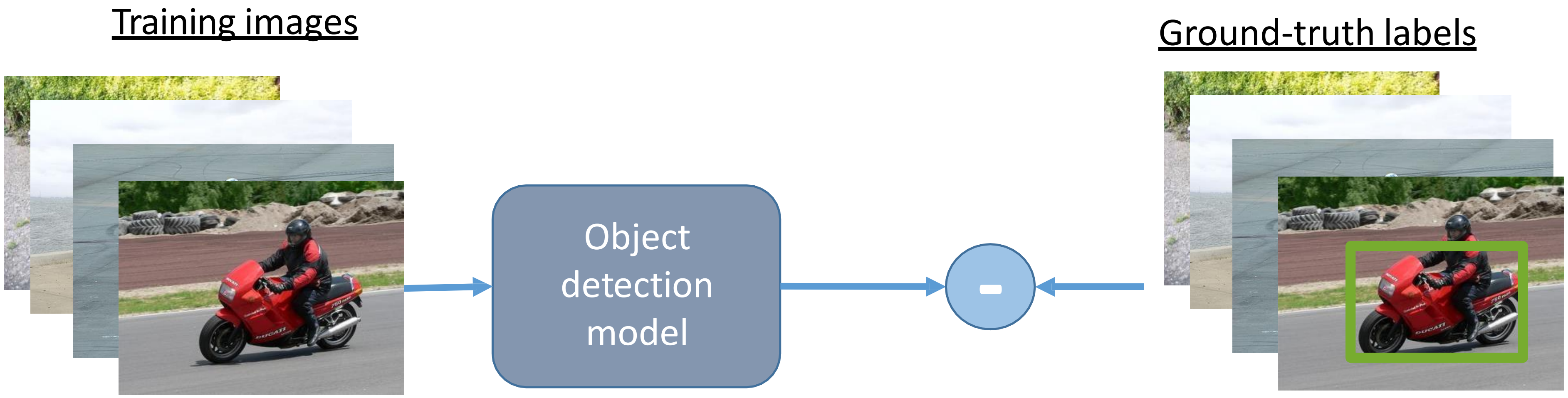
Generating seeds using CAM

[Kolesnikov et al., ECCV16]

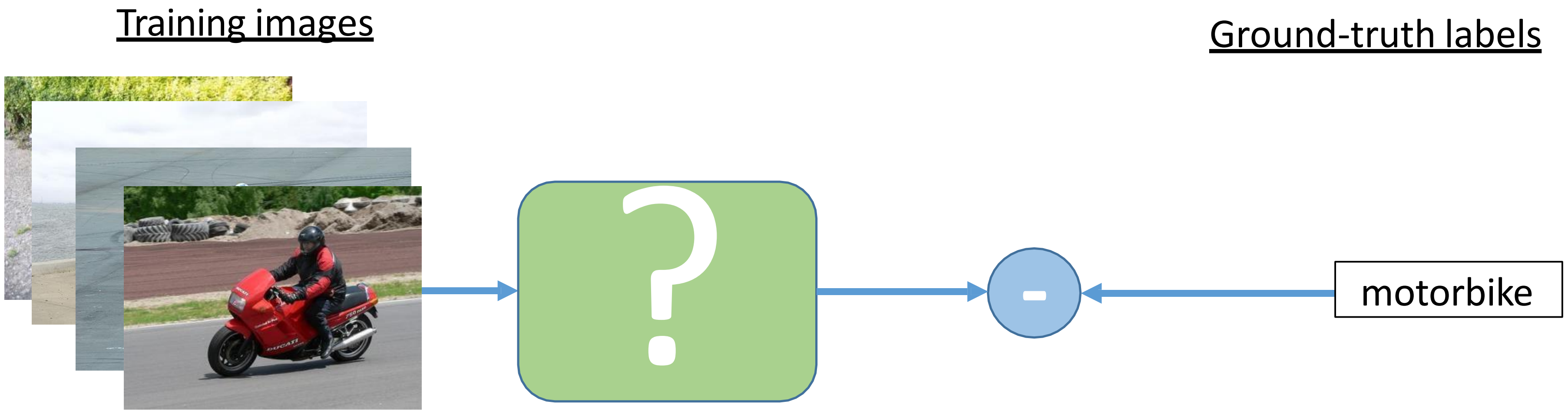


Part II: Weakly-supervised Object Detection

Standard supervised object detection



Slide credit: Hakan Bilen



What can we say at minimum?

1- When image is positive, at least one object instance from target category is present

2- When image is negative, no object instance from target category is present

Assumptions

1- There exists a set of features present in positive images and absent in negative images

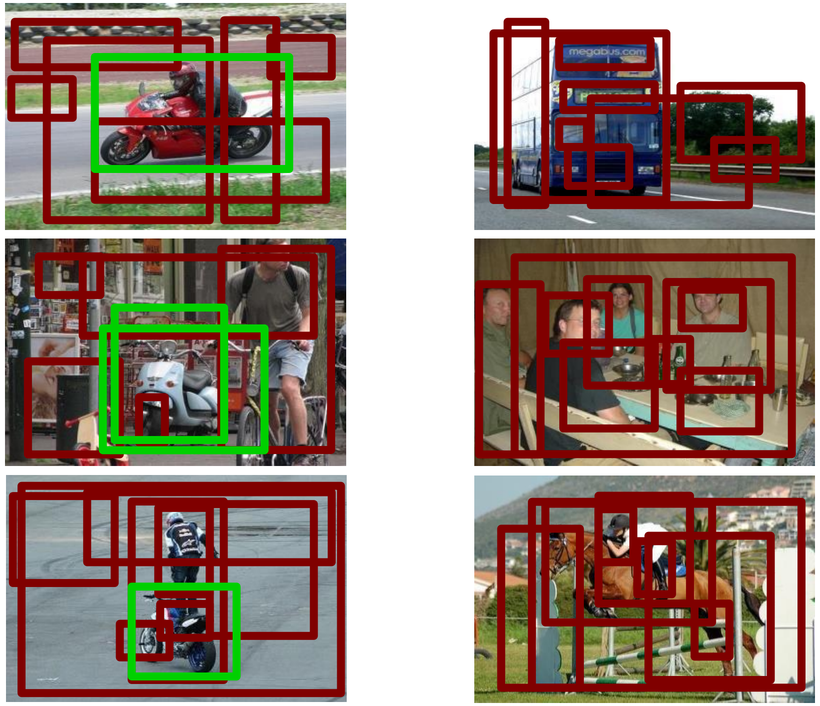
2- The same features are only present on the target object instances

Slide credit: Hakan Bilen

Multiple-instance learning (MIL)

Dietterich et al. Solving the multiple instance problem with axis-parallel rectangles. Artificial Intelligence

Positive bags **Negative bags**



bags = images
instances = windows

Goals:

- find true positive instances
- train window classifier

[Blaschko NIPS 10, Cinbis CVPR 14, Deselaers ECCV 10, Nguyen ICCV 09, Bilen BMVC 11, Russakovsky ECCV 12, Siva ICCV 11, Siva ECCV 12, Song NIPS 14, Song ICML 14, Bilen BMVC 14]

Slide credit: Vitto Ferrari

Multiple Instance Learning

Serge's key-chain

Serge **cannot** enter
the *Secret Room*

Sanjoy's key-chain

Sanjoy **can** enter
the *Secret Room*

Lawrence's key-chain

Lawrence **can** enter
the *Secret Room*

Supervised learning:

Definition Given n labeled data $(x_i, y_i) \in (\mathcal{X}, \mathcal{Y})$,$i = 1, ..., n$, $\mathcal{X} = R^d$, $\mathcal{Y} = \{0, 1\}$ learn $f(x) : \mathcal{X} \rightarrow \mathcal{Y}$.

Multiple Instance learning:

Definition Given n bags $\{X_1, ..., X_n\}$ and bag labels $\{y_1, ..., y_n\}$ where $X_i = \{x_{i1}, ..., x_{im}\}$, $x_{ij} \in \mathcal{X}$ and $y_i \in \{0, 1\}$, learn classifier for a bag $f(X) \rightarrow \{0, 1\}$.

Multiple Instance Learning

Definition Given n bags $\{X_1, ..., X_n\}$ and bag labels $\{y_1, ..., y_n\}$ where $X_i = \{x_{i1}, ..., x_{im}\}$, $x_{ij} \in \mathcal{X}$ and $y_i \in \{0, 1\}$, learn classifier for a bag $f(X) \to \{0, 1\}$.

$$\mathcal{L} = \sum_{i|y_i=1} \log(p_i) + \sum_{i|y_i=0} \log(1 - p_i)$$

$$p_i = \max_j \{p_{ij}\}$$

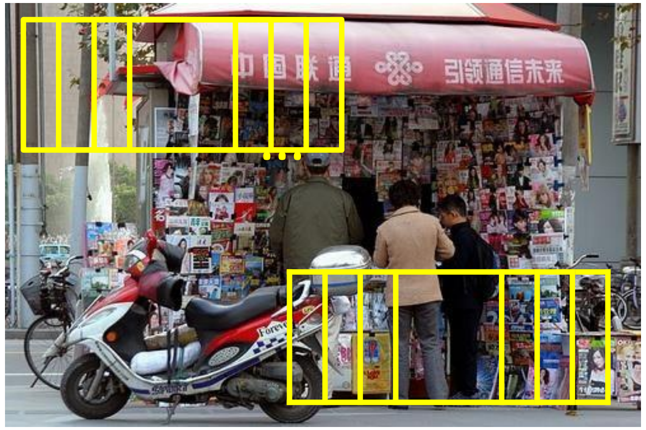
e.g. logistic regression: $\sigma(x) = \dfrac{1}{1 + \exp\{-x\}}$ $\quad p_{ij} = \sigma\Big(w \cdot x_{ij}\Big)$

How to generate bags?

Sliding windows

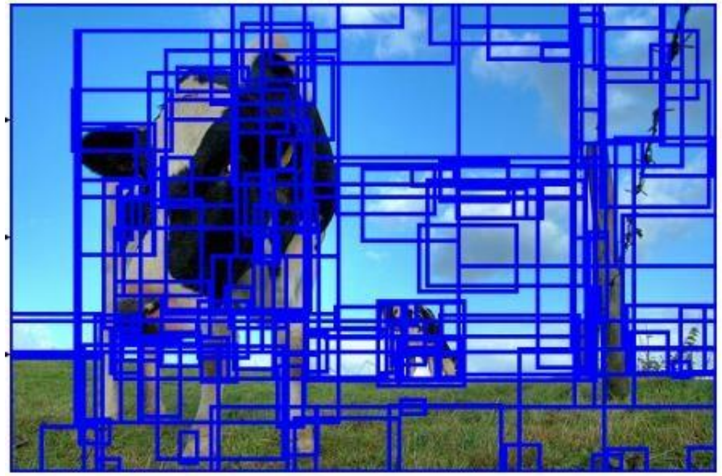
- >100k per image
- dense
- translations, scales and aspect-ratios (4D space)

[Chum CVPR 07, Nguyen ICCV 09, Pandey ICCV 11]



Object proposals

- ~2k per image
- sparse
- [Alexe CVPR 10, van de Sande ICCV 11, Dollar ECCV 14]
- Commonly used in WSOD [Deselaers ECCV 10, Siva ICCV 11, Russakovsky ECCV 12, Cinbis CVPR 14, Wang ECCV 14, Bilen CVPR 16]



Slide credit: Vitto Ferrari

Cascaded Object Detection [Diba CVPR17]

- Stage 1: Better class activation maps, provides a subset of windows
- Stage 2: Selects highest scoring proposal window
- Additional final step: Trains a Fast-RCNN
- Back to 64% of supervised counterpart (Fast-RCNN)

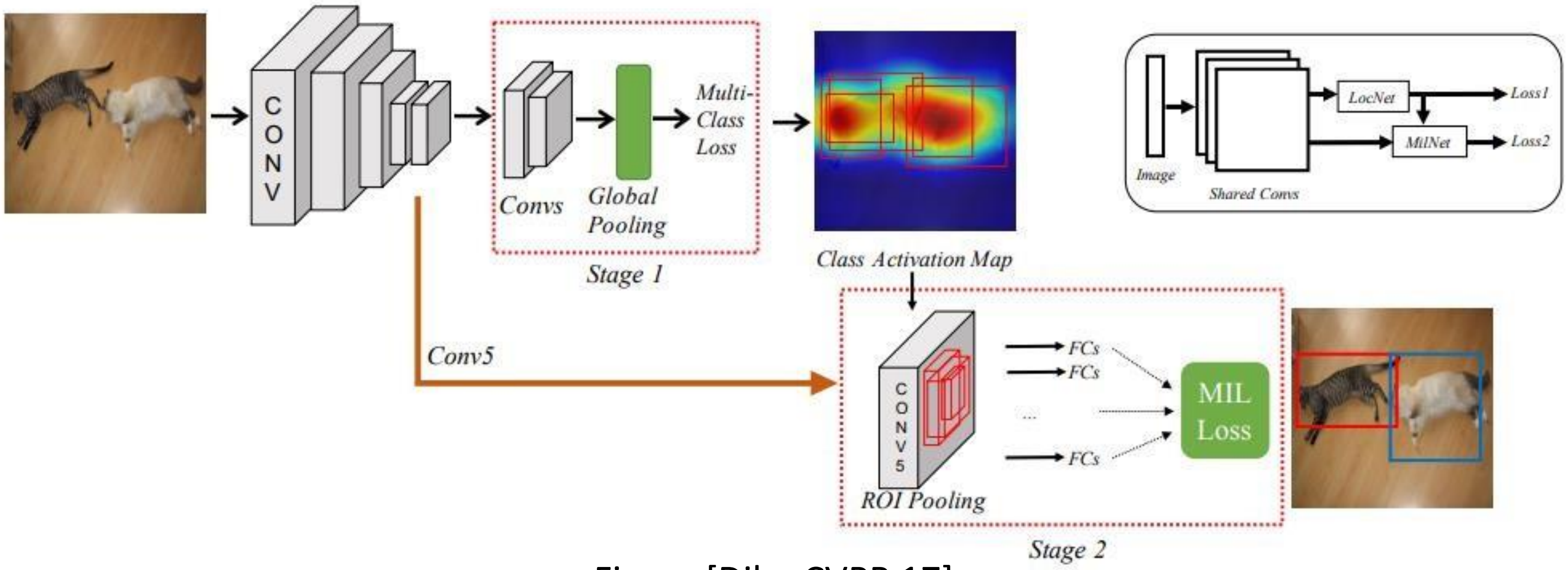
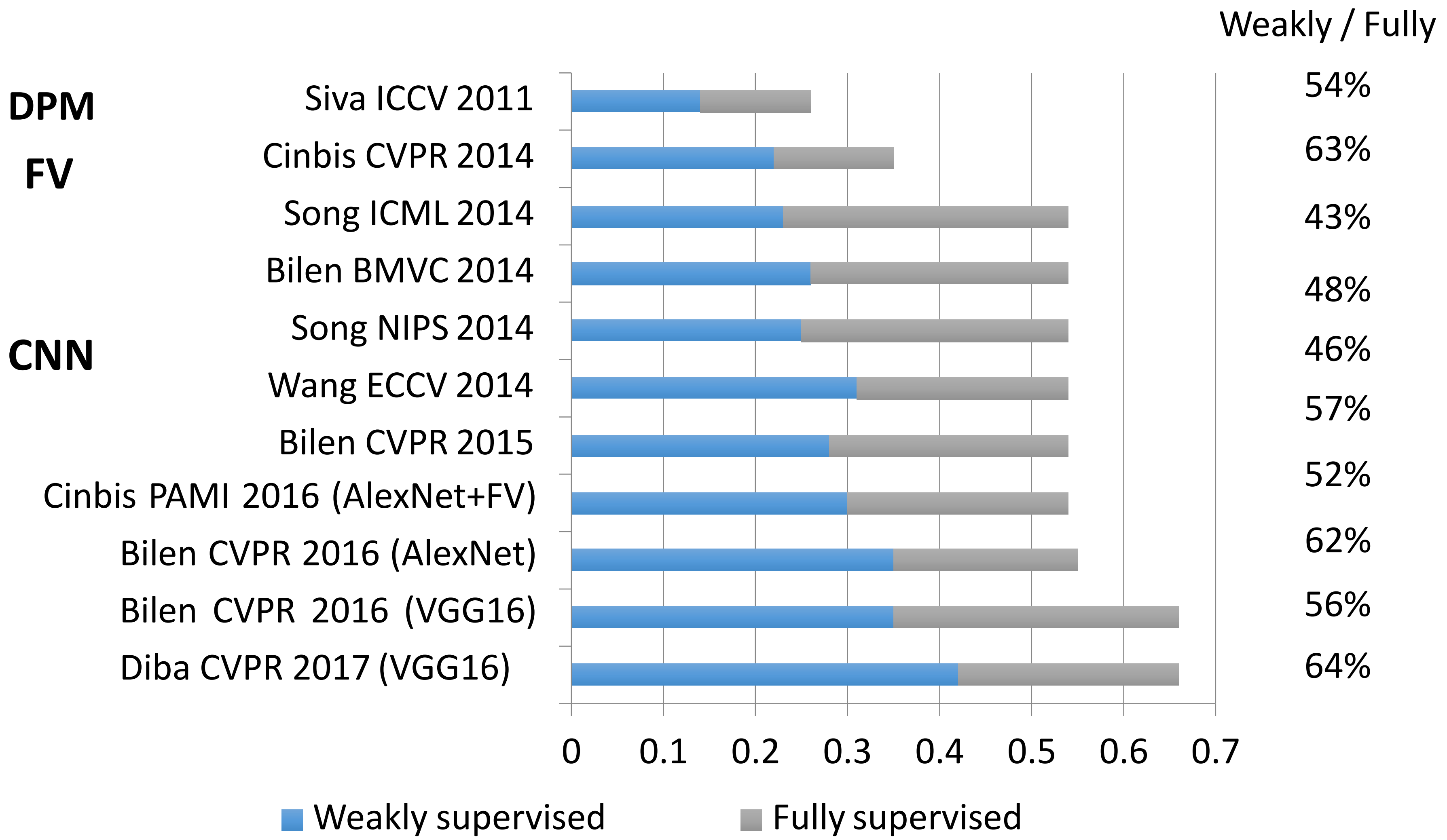


Figure [Diba CVPR 17]

Performance at test time

WSL on PASCAL 07 trainval all views, test on test (mAP)



Performance still far from fully supervised detector

Slide credit: Hakan Bilen

Conclusion

- Supervised learning of CNN is a great success but data is expensive
- A classification network implicitly encodes about localization, CAM
- Regularized losses for weakly-supervised segmentation
- Multiple Instance learning for weakly-supervised detection
- Other tasks such as single view 3D reconstruction and optical flow