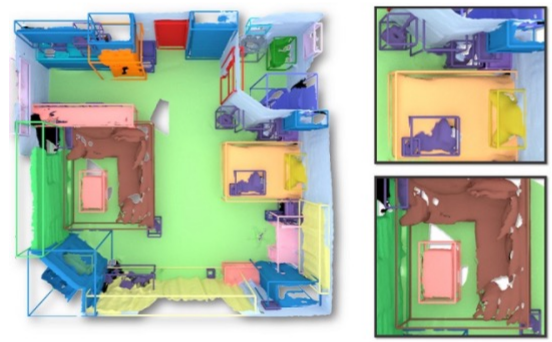
Weakly-supervised Segmentation

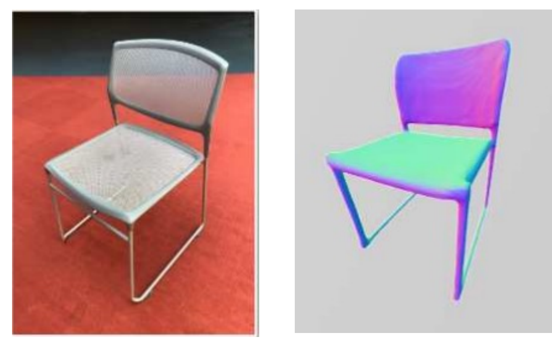
Visual perception has been a success

Object Detection & Segmentation



3D Scene Understanding

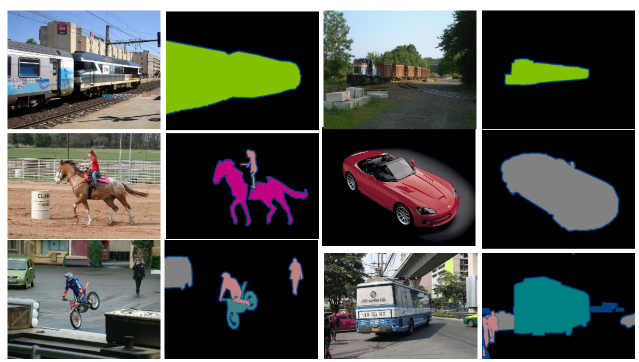
Object Tracking



Object Reconstruction

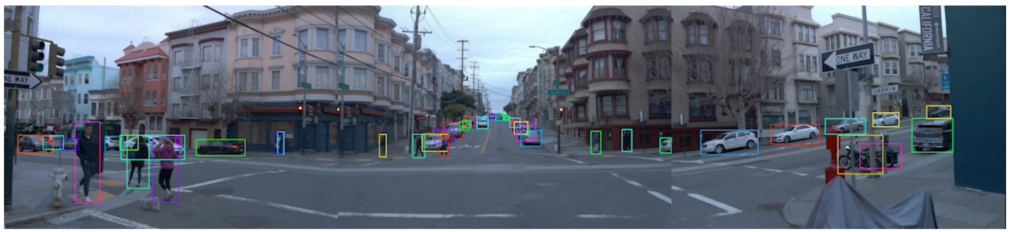
Big Data and Full Supervision

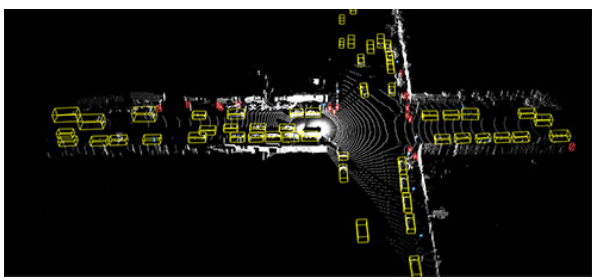
ImageNet dataset (14 million images w/ classification)



PASCAL dataset (10,582 images w/ segmentation)

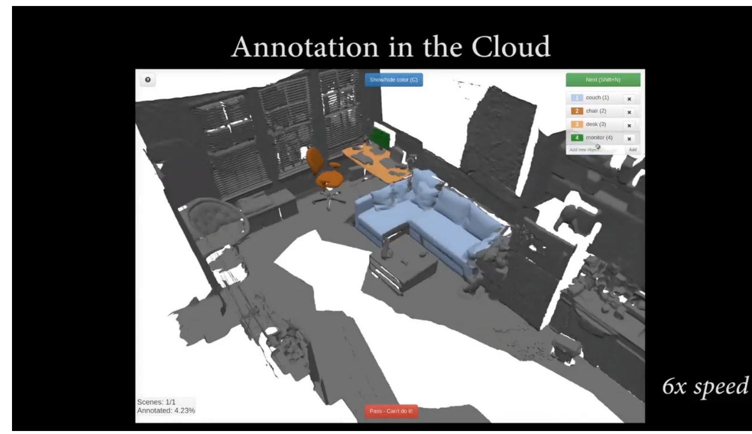
3D Annotation is much more costly than 2D annotation





	Vehicle	Pedestrian	Cyclist	Sign
3D Object	6.1M	2.8M	67k	3.2M
3D TrackID	60k	23k	620	23k
2D Object	9.0M	2.7M	81k	–
2D TrackID	194k	58k	1.7k	–

Waymo open dataset



3D annotation for ScanNet dataset

"The Next AI Revolution Will Not Be Supervised"

- Yann LeCun

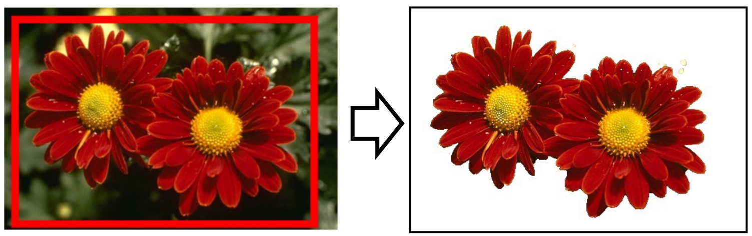
https://engineering.nyu.edu/news/revolution-will-not-be-supervised-promises-facebooks-yann-lecun-kickoff-ai-seminar

Outline

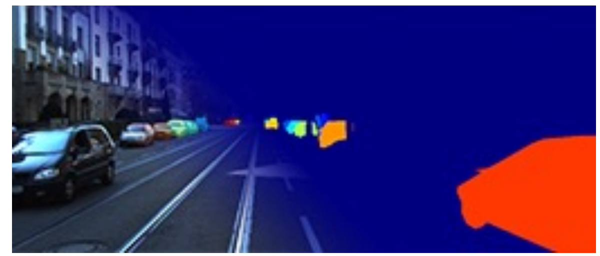
Part I: Scribble-supervised Segmentation

Part II: Segmentation from Image-level labels

Segmentation



interactive segmentation



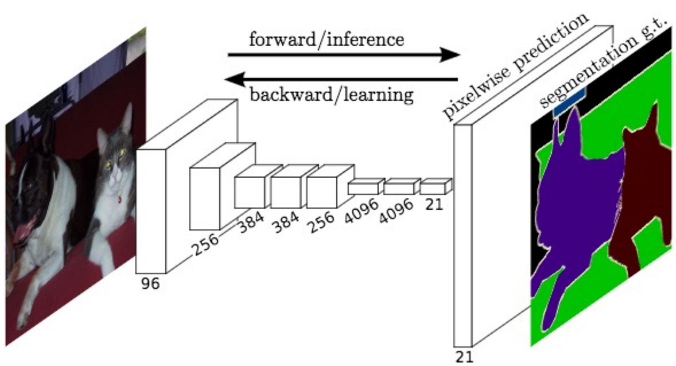
instance segmentation

semantic segmentation

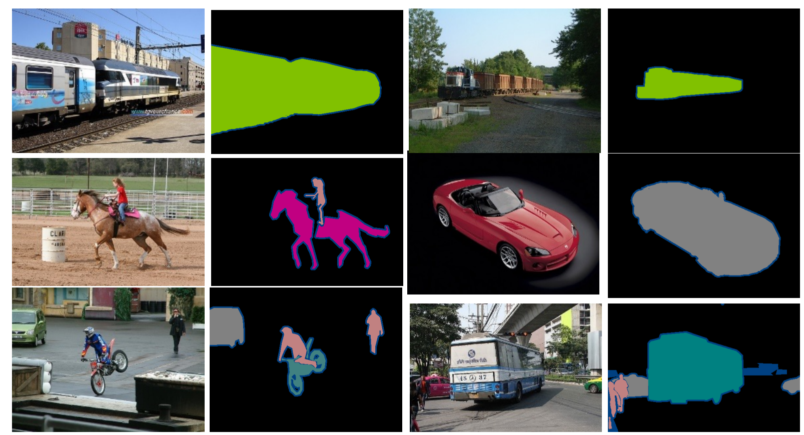
video object segmentation

Fully-supervised CNN Segmentation

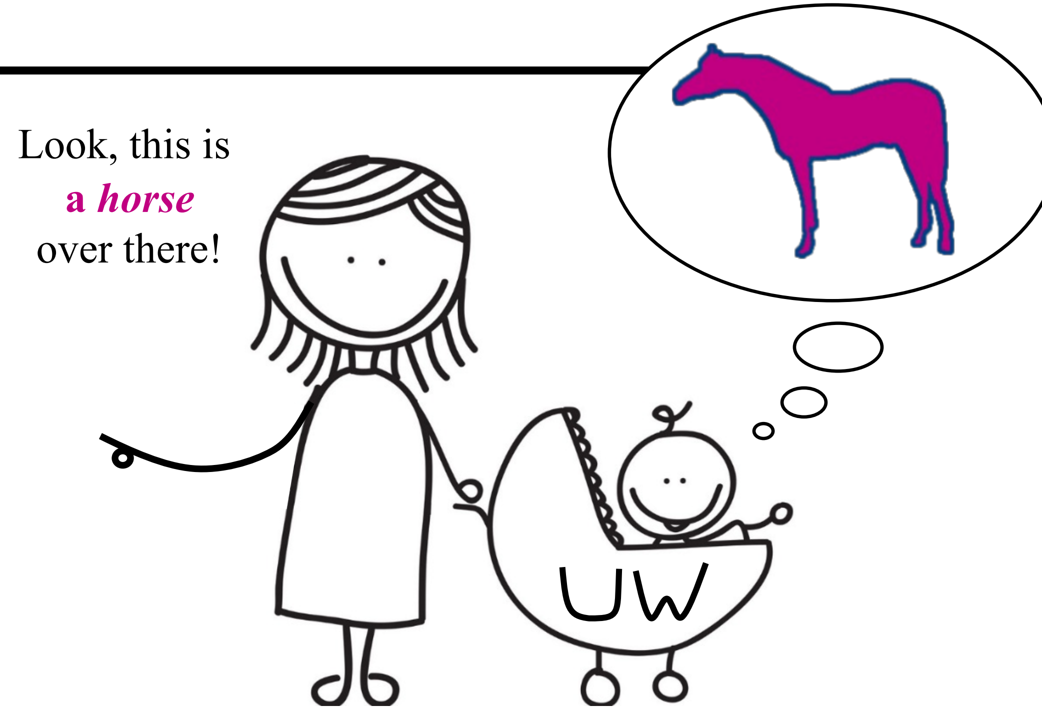
Network



Training Data



[Long et al. 2015]



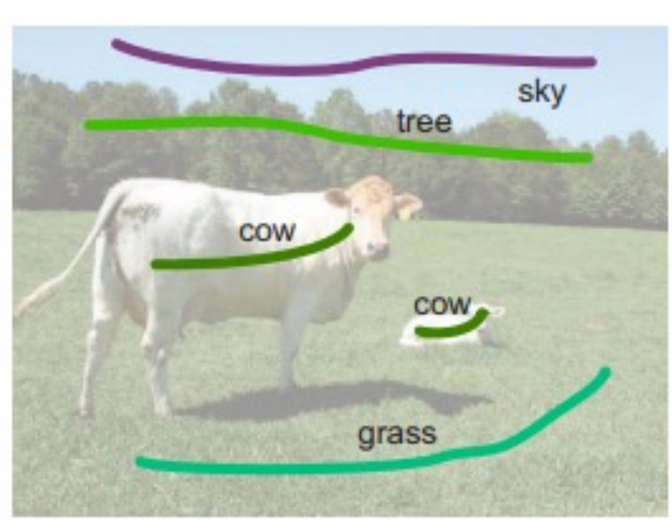
from fully-supervised to

Weakly-supervised Semantic Segmentation

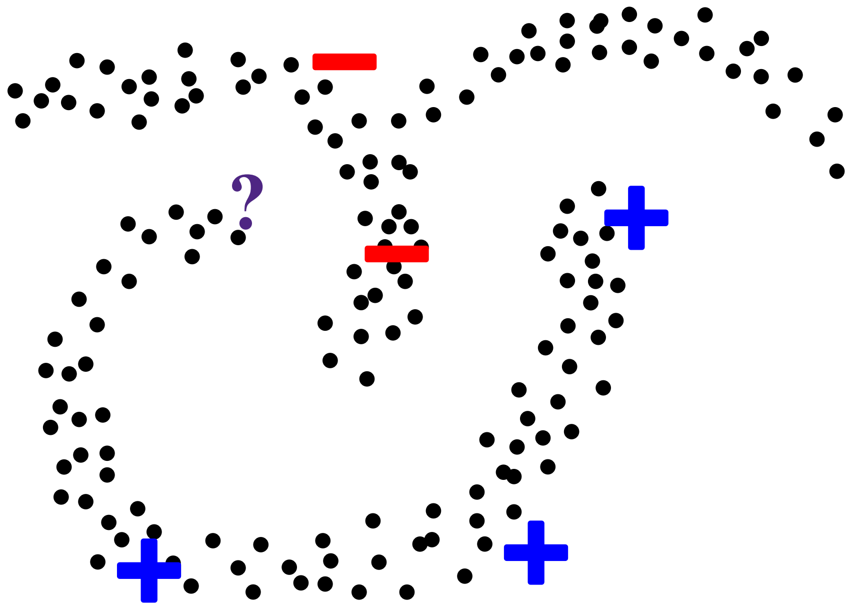
Weakly Supervised Semantic Segmentation

Key Idea

Weakly-supervised segmentation

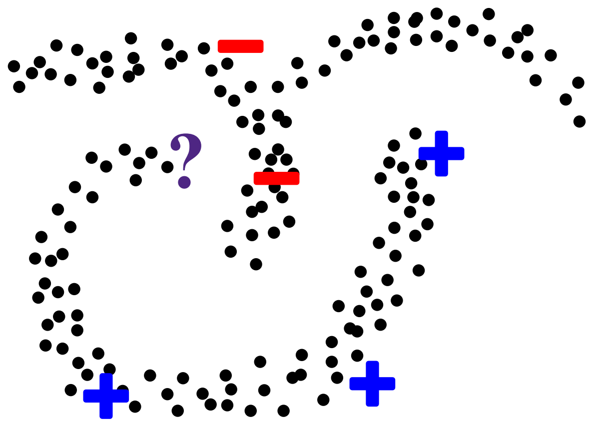


Semi-supervised learning



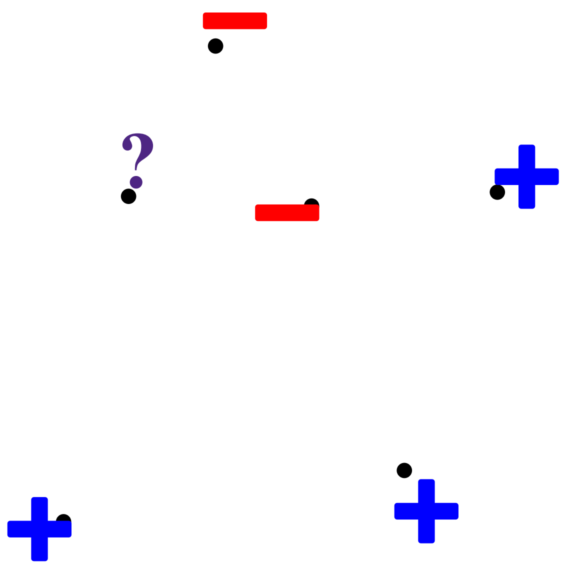
Semi-supervised learning

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 $f(x) : \mathcal{X} \to \mathcal{Y}$.

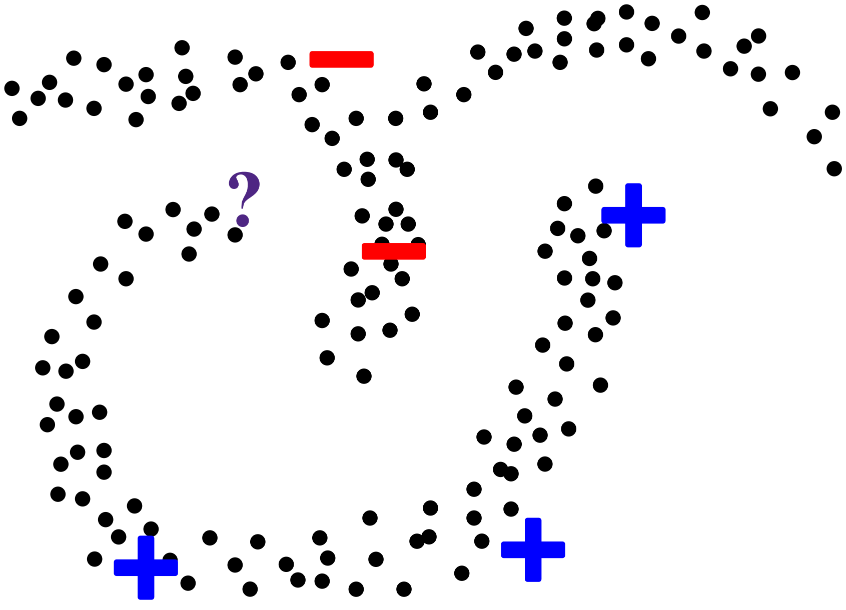


[Zhu & Goldberg, "Introduction to semi-supervised learning", 2009]
[Chapelle, Scholkopf & Zien, "Semi-supervised learning", 2009]

Does unlabeled data matter?



w/o unlabeled data



w/ unlabeled data

Semi-supervised Learning Methods

Self-training

Graph-based Semi-supervised learning

Entropy minimization

Many others…

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

Graph-Based Semi-supervised Learning

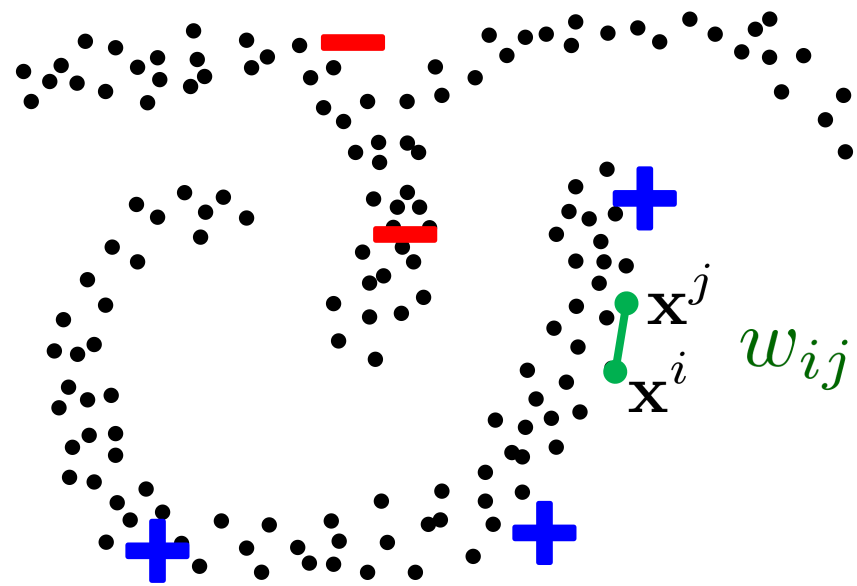
Loss function ?

- labelled points should have **consistency with the target** e.g.

$$\sum_{i=1}^{M} \delta(f(\mathbf{x}^i) \neq \mathbf{y}^i)$$

- unlabeled points should be labeled so that there is some agreement between neighbors i.e. **pairwise regularization**:

$$\sum_{ij \in \mathcal{N}} w_{ij} \, ||f(\mathbf{x}^i) - f(\mathbf{x}^j)||^2$$

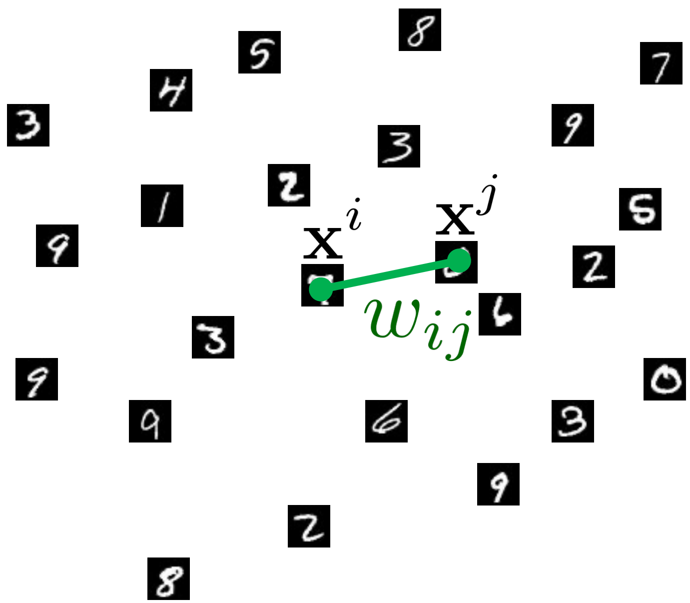


w_{ij} - pre-computed penalty, e.g. based on distance between feature vectors $\mathbf{x}^i$ and $\mathbf{x}^j$

Deep Semi-supervised Learning

Classification

(Weston et al. 2012)



e.g. for **classification CNN** output

$$f(\mathbf{x}^i) = \bar{\sigma}^i \equiv (\bar{\sigma}_1^i, \ldots, \bar{\sigma}_K^i)$$

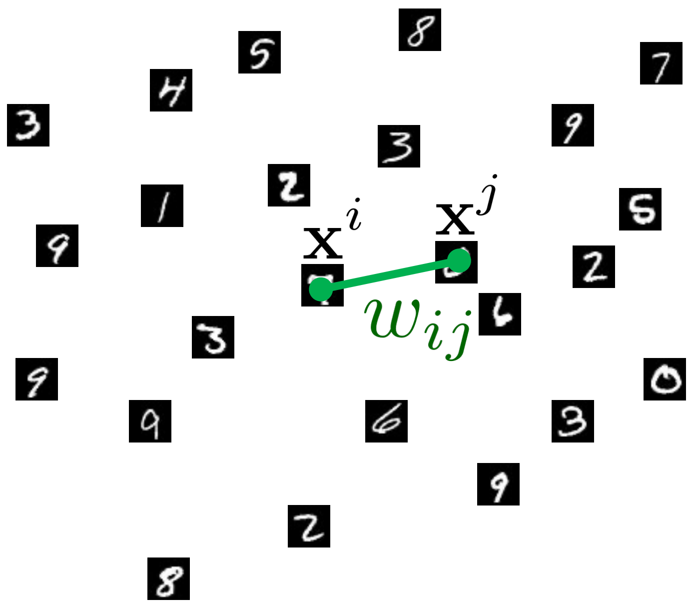
class probabilities at point i

$$\boxed{\sum_{ij \in \mathcal{N}} w_{ij} \quad ||\bar{\sigma}^i - \bar{\sigma}^j||^2}$$

Deep Semi-supervised Learning

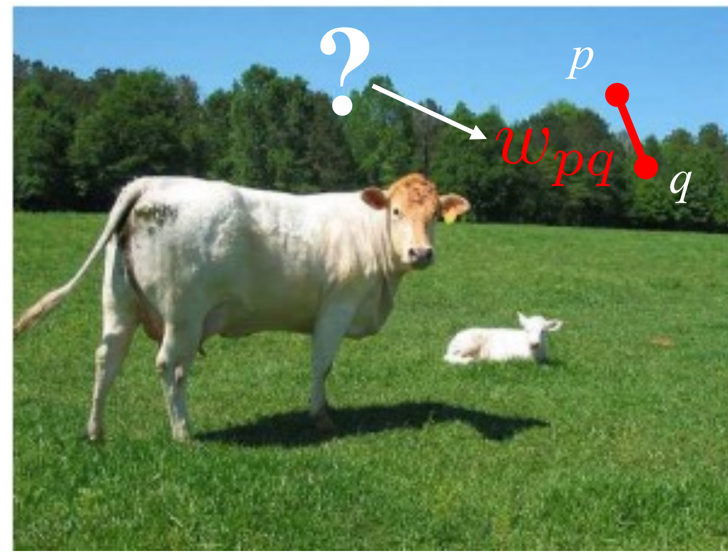
Classification

(Weston et al. 2012)



Segmentation

(Tang et al. CVPR18, ECCV18)



e.g. for **classification CNN** output

$$f(\mathbf{x}^i) = \bar{\sigma}^i \equiv (\bar{\sigma}_1^i, \ldots, \bar{\sigma}_K^i)$$

class probabilities at point i

$$\boxed{\sum_{ij \in \mathcal{N}} w_{ij} \quad ||\bar{\sigma}^i - \bar{\sigma}^j||^2}$$

e.g. for **segmentation CNN** output

$$\bar{\sigma}^p \equiv (\bar{\sigma}_1^p, \ldots, \bar{\sigma}_K^p)$$

class probabilities at pixel p

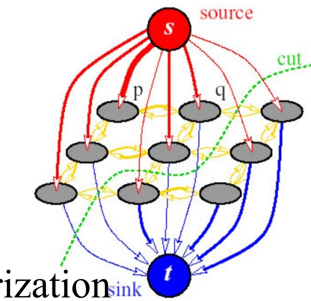
$$\boxed{\sum_{pq \in \mathcal{N}} w_{pq} \quad ||\bar{\sigma}^p - \bar{\sigma}^q||^2}$$

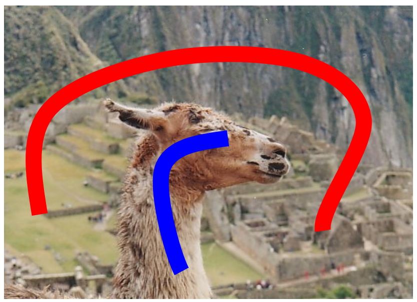
Regularized Loss Functions

We can use regularization ideas from
unsupervised and interactive segmentation
to exploit low-level segmentation cues
(contrast alignment, boundary regularity, regional color consistency, etc.)
for unlabeled parts of an image

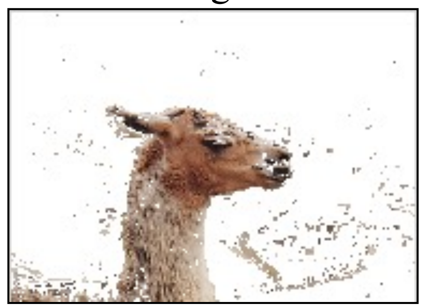
low-level segmentation

Markov Random Field for Segmentation

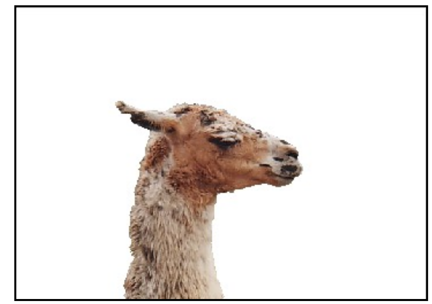


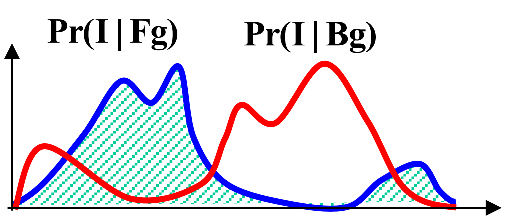


Without Regularization



With Regularization





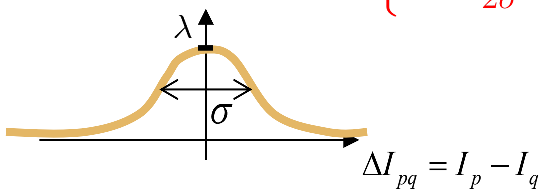
$$E(S,\theta_0,\theta_1) = \sum_{k=0,1}\sum_{p\in S^k} -\ln \mathrm{P}(I_p|\theta_k) + \boxed{\lambda \cdot \sum_{pq\in\mathcal{N}} w_{pq}\cdot[s_p \neq s_q]}$$

MRF regularization

[Boykov, Jolly, *ICCV* 2001]

Regularization energies

$w_{pq} = \lambda \exp\left\{-\frac{\|I_p - I_q\|^2}{2\sigma^2}\right\}$ - contrast weights w_{pq} from topic 9

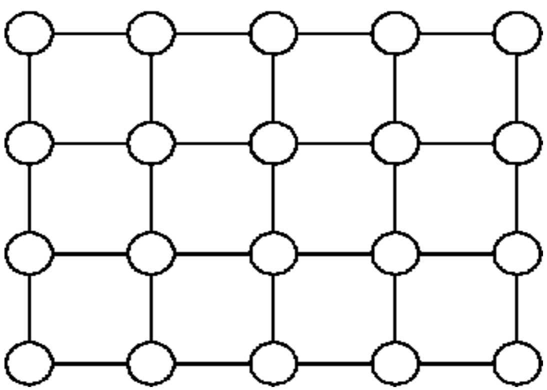


coherence between
discrete labels
at pixels p and q

$$\sum_{pq \in \mathcal{N}} w_{pq} \quad [S^p \neq S^q]$$

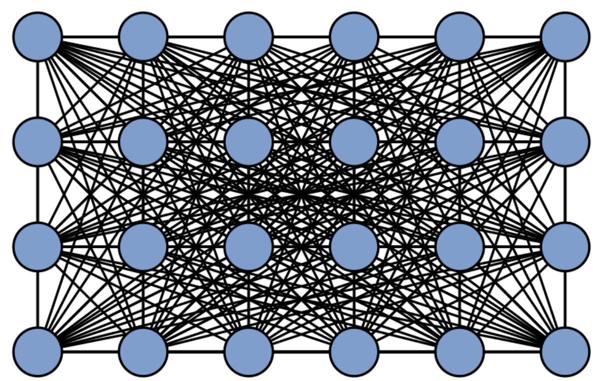
Iverson brackets

Examples of neighborhood systems $\mathcal{N}$ on pixel grid



sparsely connected

[Geman&Giman'81, BVZ PAMI'01, B&J ICCV'01]



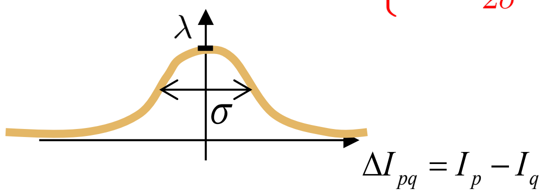
densely connected

[Dense CRF, Krähenbühl & Koltun, NIPS 2011]

weakly-supervised CNN segmentation:

Regularization Loss

$w_{pq} = \lambda \exp\left\{-\frac{\|I_p - I_q\|^2}{2\sigma^2}\right\}$ - contrast weights w_{pq} from topic 9

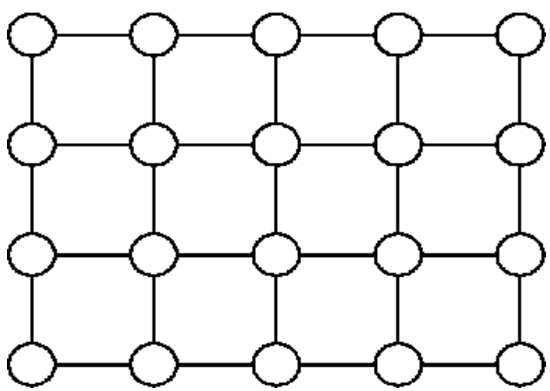


coherence between
probabilistic predictions
at pixels p and q

$$\sum_{pq \in \mathcal{N}} w_{pq} \quad ||\bar{\sigma}^p - \bar{\sigma}^q||^2$$

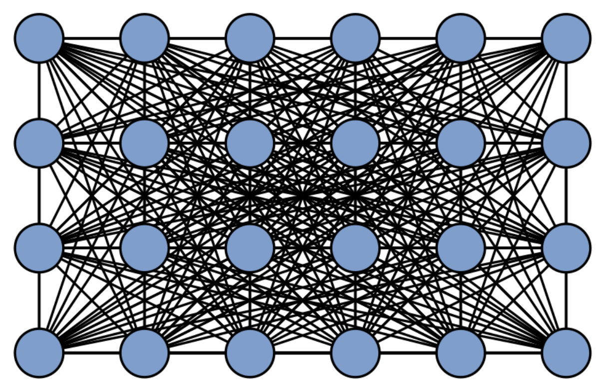
relaxation of Iverson brackets
for probabilistic predictions

Examples of neighborhood systems $\mathcal{N}$ on pixel grid



sparsely connected

[Geman&Giman'81, BVZ PAMI'01, B&J ICCV'01]



densely connected

[Dense CRF, Krähenbühl & Koltun, NIPS 2011]

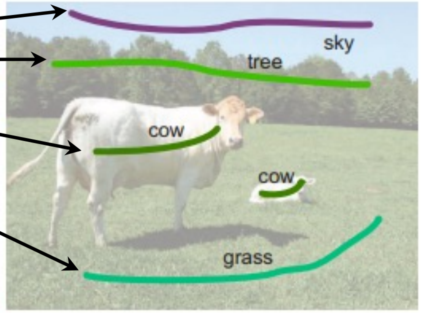
weakly-supervised CNN segmentation:

Partial Cross Entropy Loss

cross entropy

over seeds only

$$-\sum_{p \in seeds} \ln \bar{\sigma}^p_{\mathbf{y}^p}$$



$$\bar{\sigma}^p \equiv (\bar{\sigma}^p_1, \ldots, \bar{\sigma}^p_K)$$

predicted "probabilities" for p
to be in each class, e.g. (0,0,..,1,...) in **one-hot** case

NOTE: if prediction is one-hot
then cross entropy at seed p
is equivalent to $0/\infty$ hard constraint
(as in interactive graph cut, Topic 9)

$$\sum_{p \in seeds} \delta(\bar{\sigma}^p \neq \bar{\mathbf{y}}^p)$$

hard constraint
on seed p

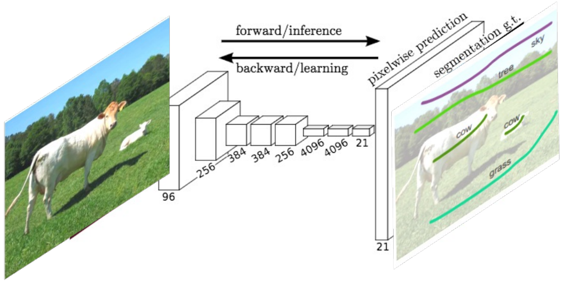
Implications:

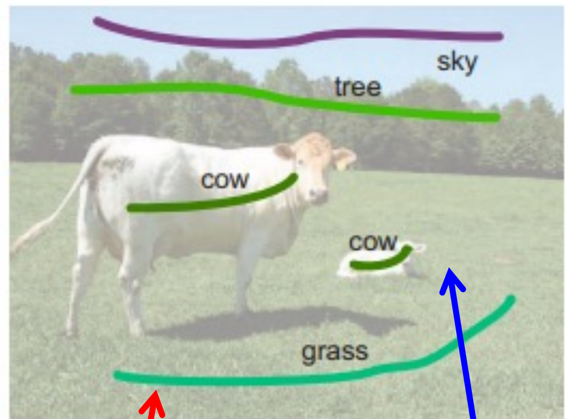
- Cross entropy is a relaxation of hard constraints for probabilistic predictions.

- Cross entropy is a bad idea for pixels where targets y^p could be wrong.
Remember "fake" ground truths - network tries hard to learn all their mistakes.

weakly-supervised CNN segmentation:

Total **Regularized Loss**





scribbles / seeds

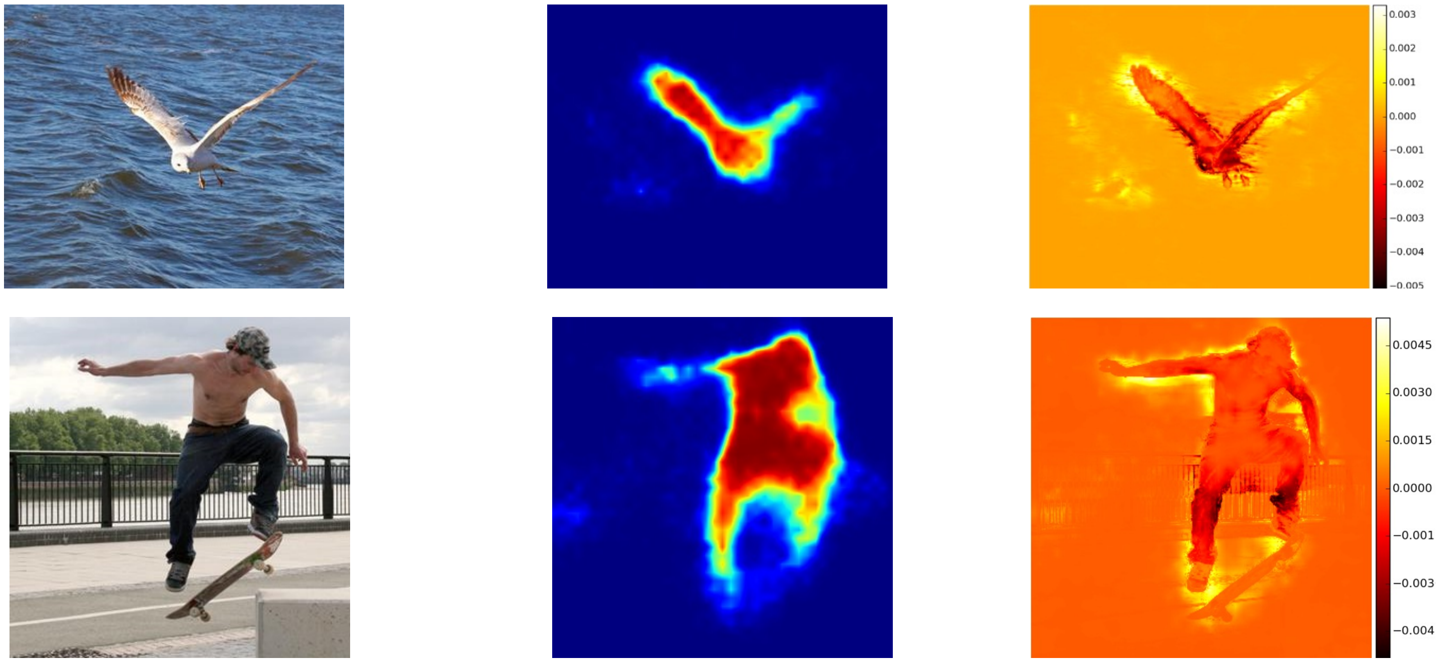
unlabeled pixels

$$L(\bar\sigma) = \boxed{-\sum_{p \in seeds} \ln \bar\sigma^p_{\mathbf{y}^p}} + \boxed{\sum_{\substack{pq \in \mathcal{N} \\ \text{n-links}}} w_{pq} \quad ||\bar\sigma^p - \bar\sigma^q||^2}$$

Partial Cross Entropy (PCE)

Regularization Loss

Regularization Loss Gradients



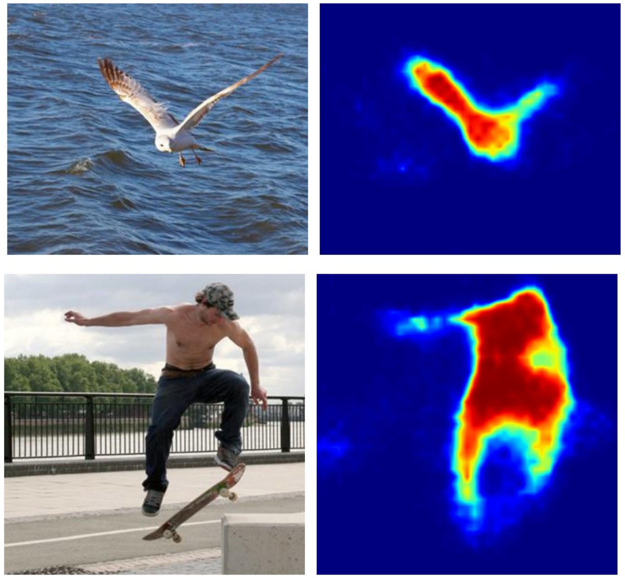
input

network prediction for class k during training $\bar{\sigma}_k^p$

regularization loss gradient $\frac{\partial R(\sigma)}{\partial \sigma_k}$

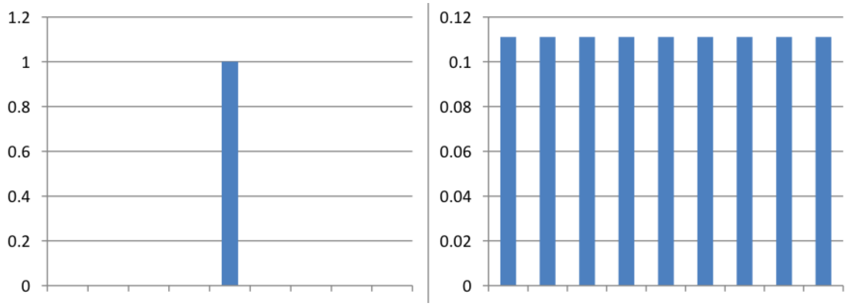
$$R(\sigma) = \sum_{pq \in \mathcal{N}} w_{pq} \cdot ||\bar{\sigma}^p - \bar{\sigma}^q||^2$$

CNN Segmentation may be blurred



Pointwise Entropy Regularization

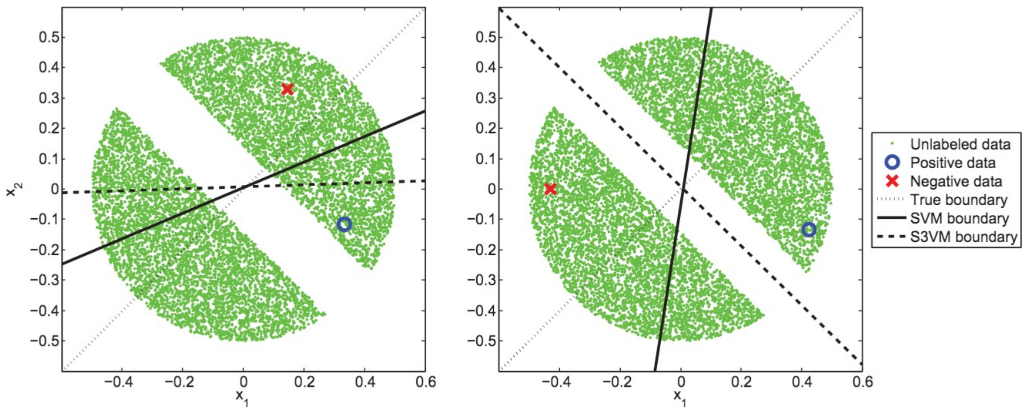
$$\sum_i H(f(x_i))$$



Low entropy High entropy

$$\mathrm{H}(P) = \sum_{k=0}^{K} -P_k \cdot logP_k$$

Entropy Minimization for Semi-supervised Learning



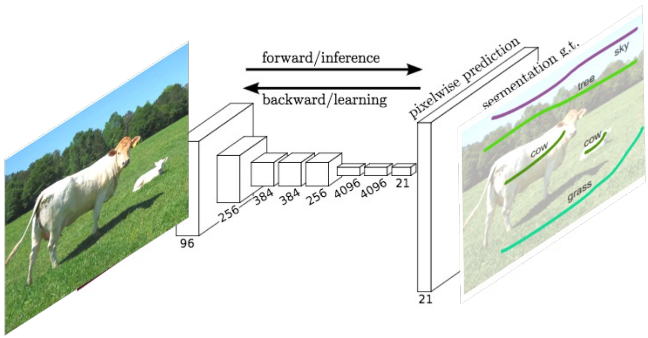
(a) S3VM in local minimum (b) S3VM in "wrong" low density region

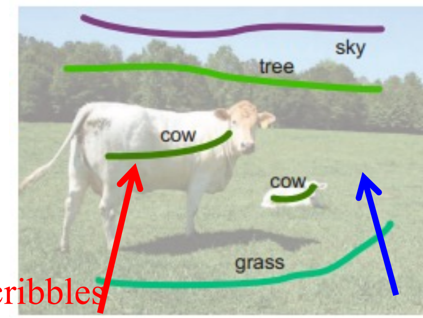
Remark 6.1. The assumption of both S3VMs and entropy regularization is that the classes are well-separated, such that the decision boundary falls int o a low density region in the feature space, and does not cut through dense unlabeled data

Introduction to Semi-Supervised Learning Xiaojin Zhu and Andrew B. Goldberg

Grandvalet, Yves, and Yoshua Bengio. "Semi-supervised learning by entropy minimization." *Advances in neural information processing systems*. 2005.

Regularized loss for weakly-supervised CNN segmentation





scribbles

unknown pixels

empirical risk Loss
for *labeled* data

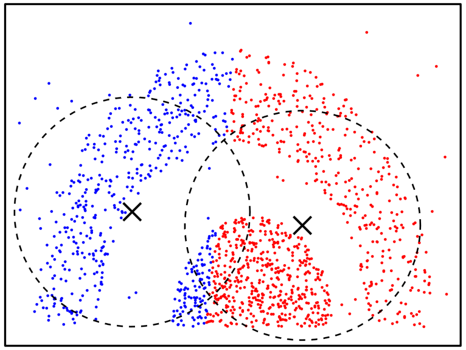
regularization Loss
for *unlabeled* data

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

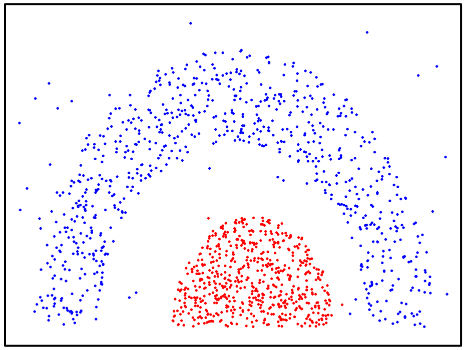
partial Cross Entropy (PCE)

Clustering and Segmentation are Largely Synonym

Linear Clustering



Nonlinear Clustering



Normalized Cut
Segmentation

Kernel K-means

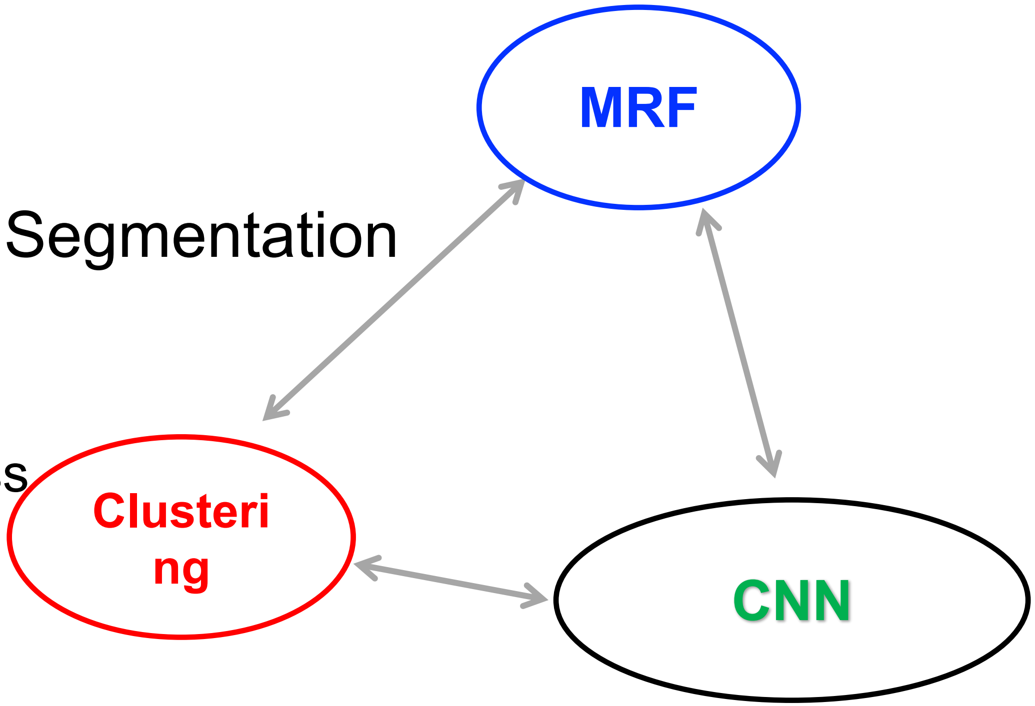
$$\sum_{p \in \mathbf{S}} \|\phi(I_p) - \mu_{\mathbf{S}}\|^2 + \sum_{p \in \overline{\mathbf{S}}} \|\phi(I_p) - \mu_{\overline{\mathbf{S}}}\|^2$$

$$\stackrel{c}{=} \quad -\frac{\sum_{p,q \in \mathbf{S}} k(I_p, I_q)}{|\mathbf{S}|} - \frac{\sum_{p,q \in \bar{\mathbf{S}}} k(I_p, I_q)}{|\bar{\mathbf{S}}|}$$

Regularized Losses

Regularized Loss for CNN Segmentation

- Pointwise entropy loss
- Pairwise MRF loss
- High-order Clustering loss



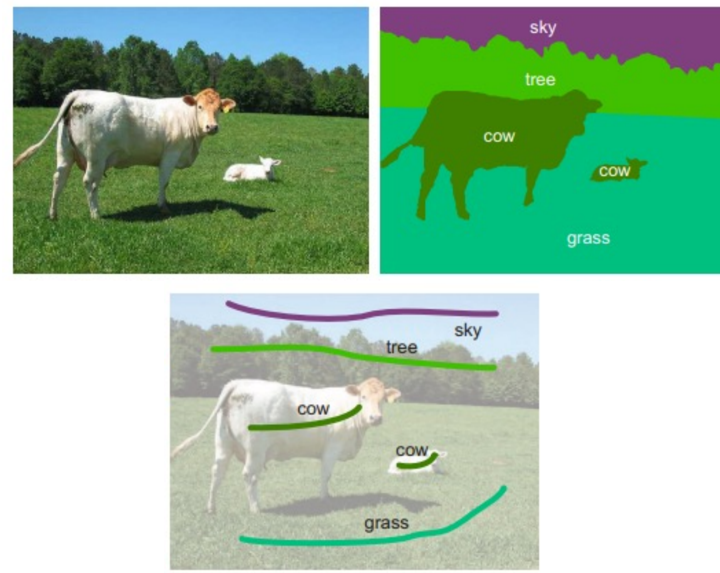
Experiments

PASCAL VOC 2012 Segmentation Dataset

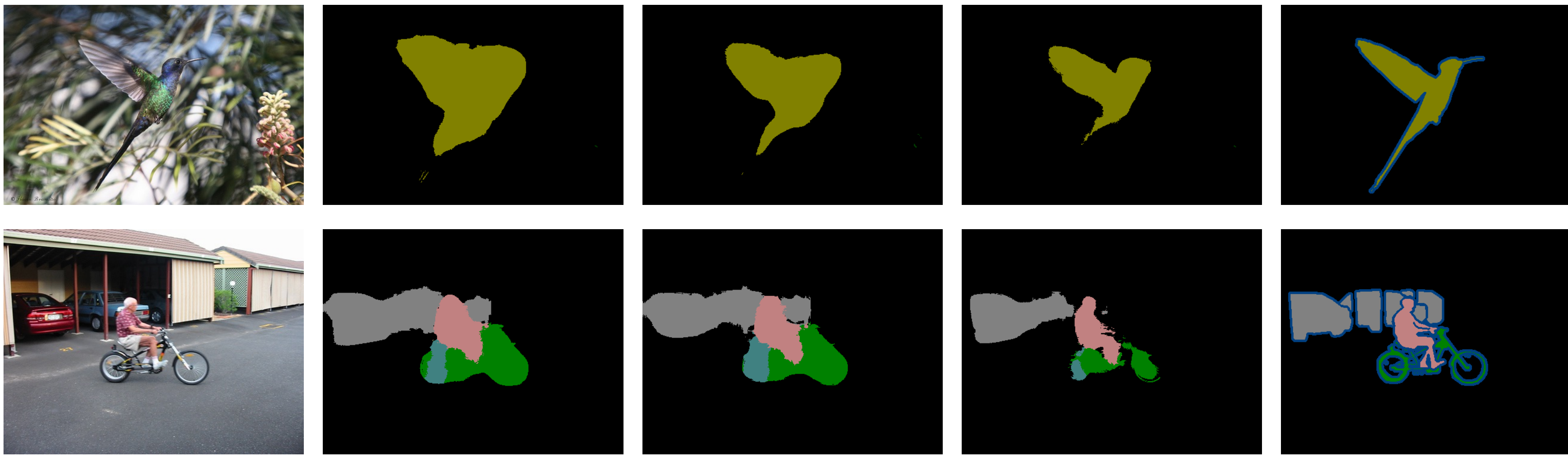
- 10K training images (full masks)
- 1.5K validation images
- 1.5K test images

ScribbleSup Dataset [Dai *et al*. ICCV 2015]

- scribbles for each object
- ~3% of pixels labelled



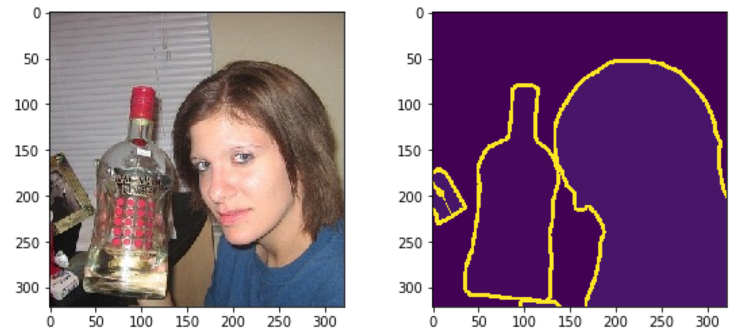
Training with combination of losses

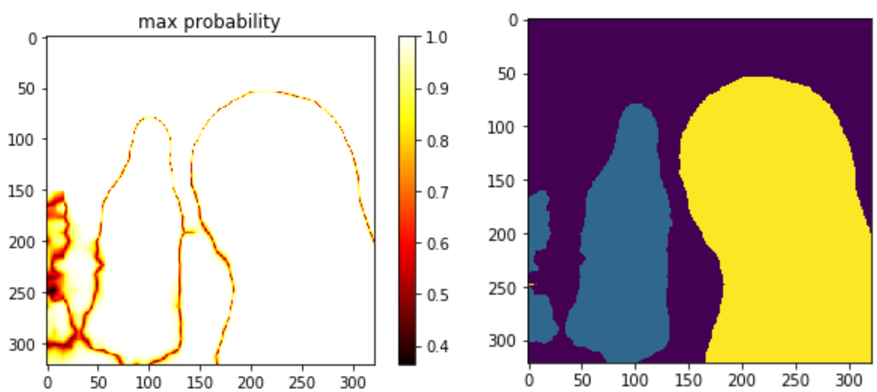


Test image | pCE loss | + clustering loss | + clustering loss + MRF loss | Ground truth

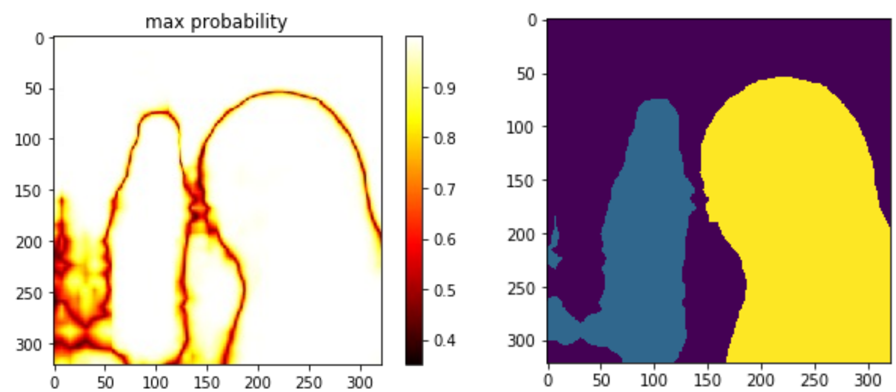
better color clustering | better edge alignment

Peakedness of distribution





w/ entropy regularization



w/o entropy regularization

Compare weak and full supervision

Almost as good as full supervision!

network	Full supervision	Weak supervision			
		PCE	PCE+CRF [1]	PCE+ENTROPY	PCE+CRF+ENTROPY
Deeplab2-largeFOV	63.0	55.8	62.2	59.9	**63.0**
Deeplab2-Msc-largeFOV	64.1	56.0	63.1	n/a	**63.5**
Deeplab2-VGG16	68.8	60.4	64.4	63.3	**65.5**
Deeplab2-Resnet101	75.6	69.5	72.9	73.1	**74.4**
Deeplab3$^+$-Resnet101	78.6	71.9	74.6	74.0	**75.6**

PCE: partial cross entropy. CRF: pairwise conditional random field

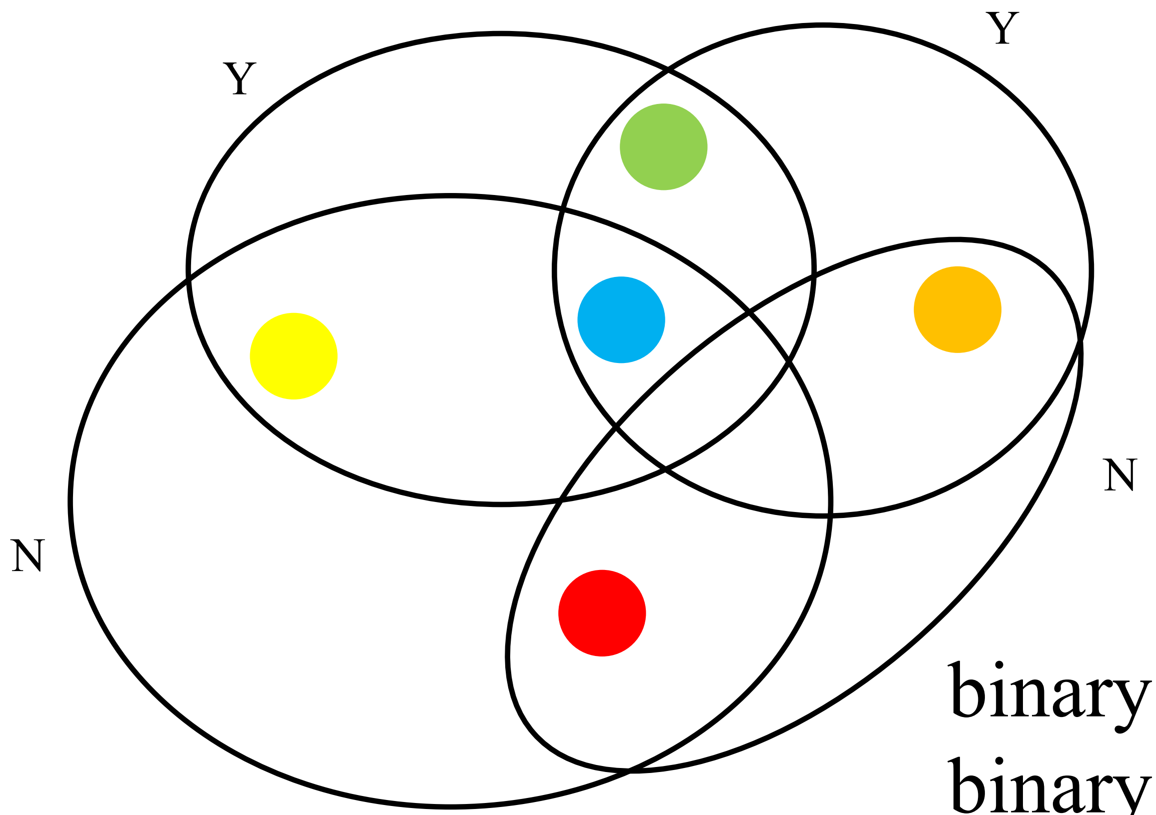
[1] Tang et al., "On Regularized Losses for Weakly-supervised CNN Segmentation", in *ECCV* 2018.

What if **image-level labels only** ?

First, consider a simple related example:

find working molecule (drug discovery)

instead of individual examples, training labels are available only for sets (bags) of examples



binary tags

binary classification

Multiple Instance Learning (MIL)

What if **image-level labels only** ?

For simplicity, assume pixel colors are discriminative enough features.

To segment, we have to learn **what color is sky, grass, and sand** ?

From these three images, we can segment pixels by matching **green to grass**, **blue to sky**, and **beige to sand**.

{ sky, grass, sand }

{ sky, sand }

{ grass, sand }

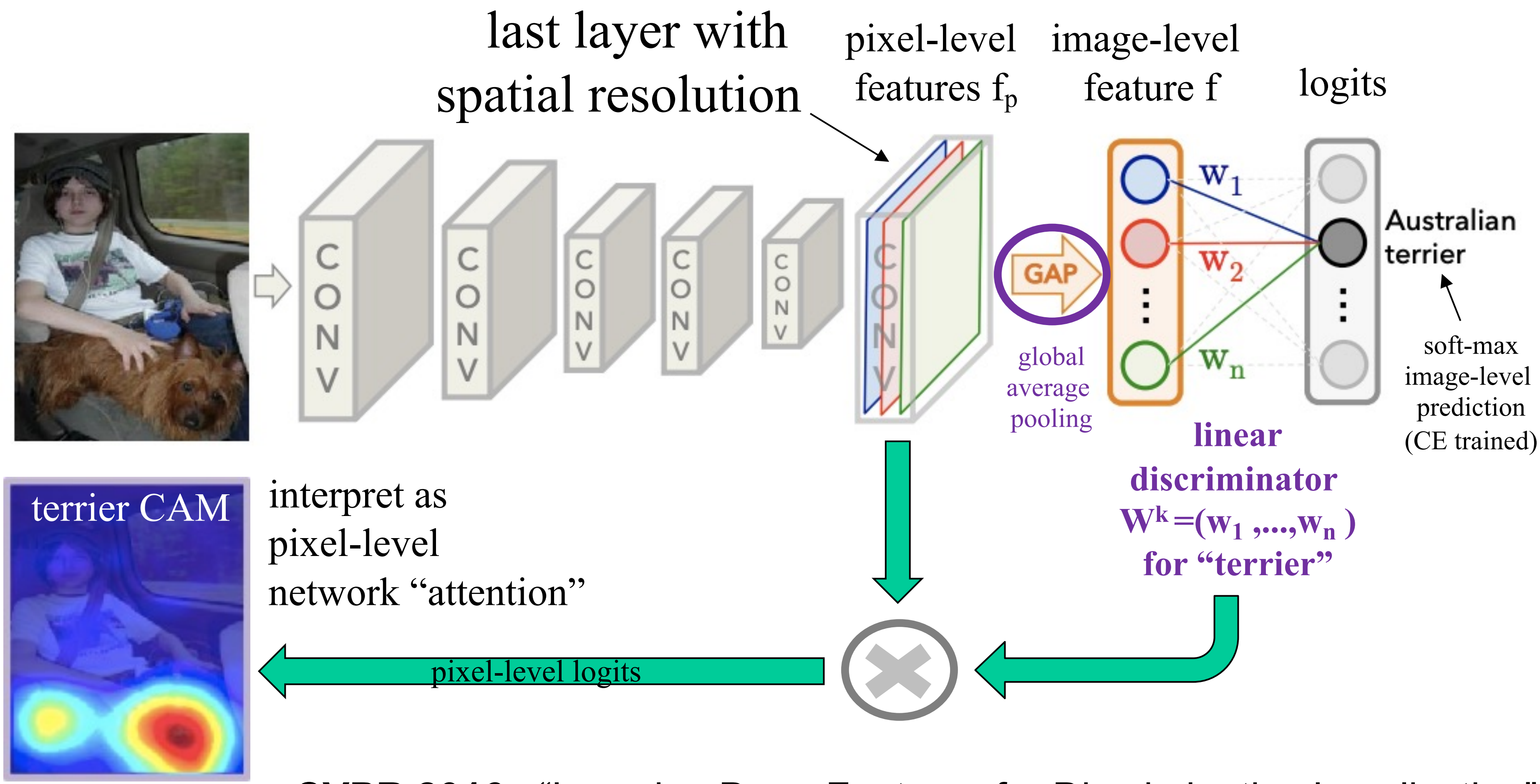
image-level tags

multi-class tags

multi-class classification

In general, segmentation network must learn BOTH (deep) **discriminative pixel-level features** AND **their match with class tags**

Class-activation Map (CAM)



CVPR 2016: "Learning Deep Features for Discriminative Localization"
B.Zhou, A.Khosla, A. Lapedriza, A.Oliva, A.Torralba

NOTE: motivates ideas for **object localization**, as well as **image-level supervision for semantic segmentation**

What if **image-level labels only** ?

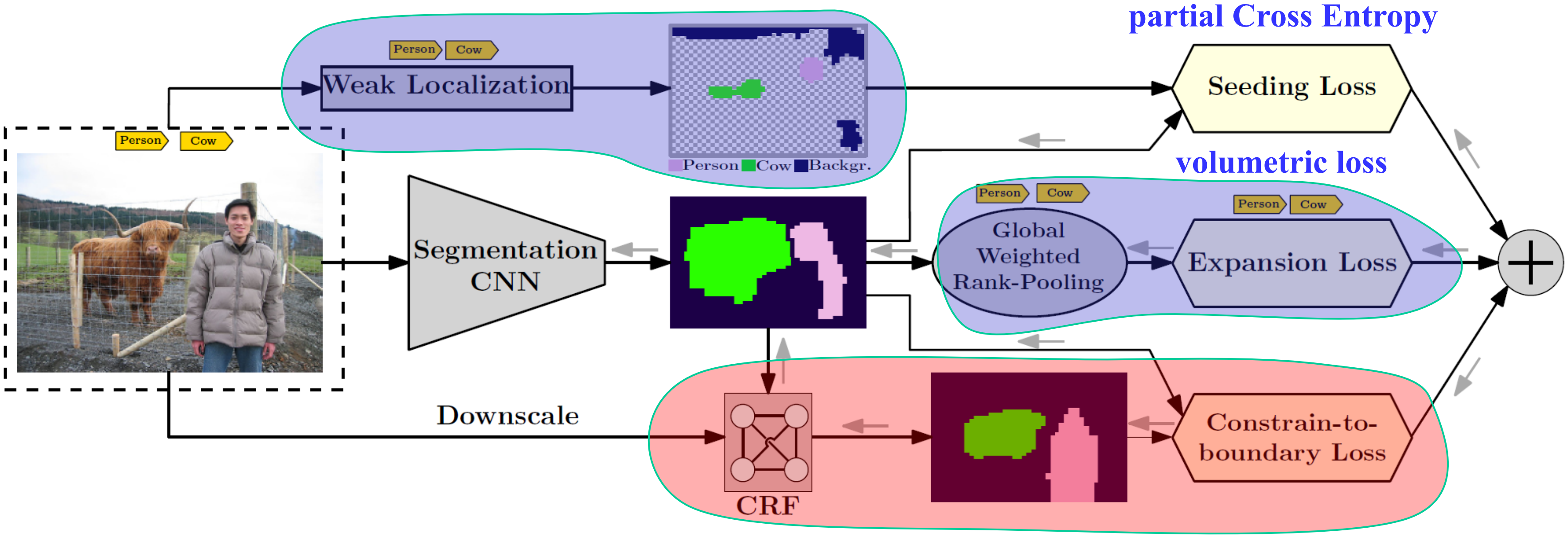
Some ideas: [Kolesnikov & Lampert ECCV 2016]

seeds from "network attention"

see CAM at the end of Topic 10

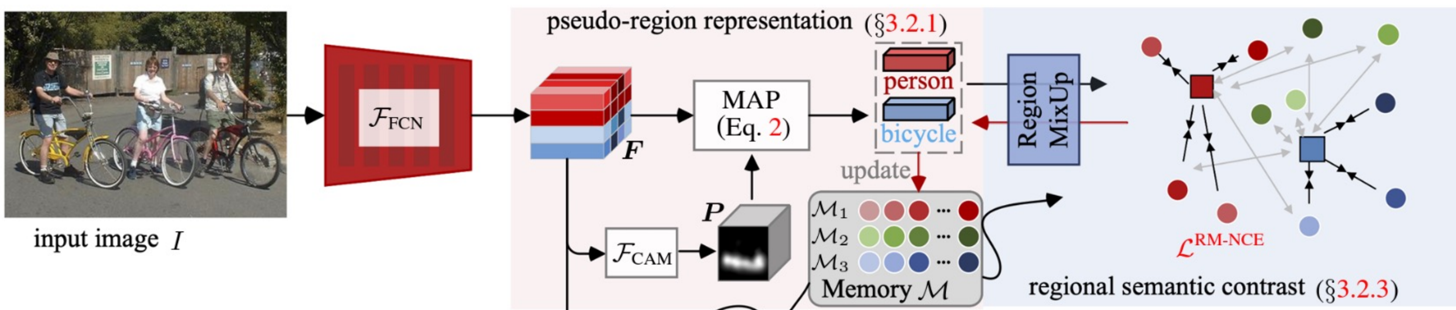
partial Cross Entropy

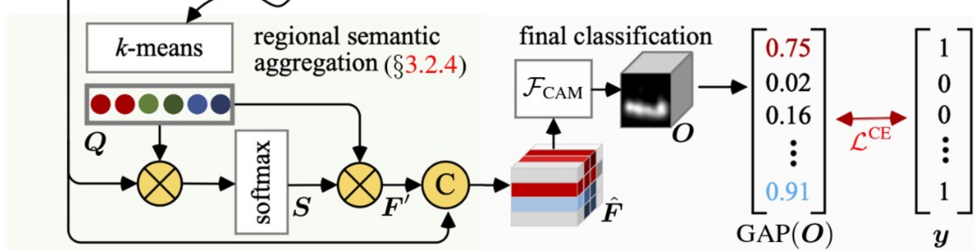
volumetric loss



Can be simplified using
regularization loss
in the previous slides

What if **image-level labels only** ?





$$\mathcal{L}_l^{\mathrm{NCE}}(\boldsymbol{f}_l, y_l) = \frac{1}{|\mathcal{M}_l|}\sum_{\boldsymbol{m}_l^+\in\mathcal{M}_l} -\log \frac{e^{\mathrm{sim}(\boldsymbol{f}_l,\boldsymbol{m}_l^+)/\tau}}{e^{\mathrm{sim}(\boldsymbol{f}_l,\boldsymbol{m}_l^+)/\tau}+\sum_{\boldsymbol{m}_l^-\in\mathcal{M}\setminus\mathcal{M}_l} e^{\mathrm{sim}(\boldsymbol{f}_l,\boldsymbol{m}_l^-)/\tau}}$$

Contrastive Learning for Features

Zhou, Tianfei, et al. "Regional semantic contrast and aggregation for weakly supervised semantic segmentation." CVPR 2022.

More recently, the state of the art for segmentation from image-level supervision is approaching full pixel-level supervision.