Constrained Convolutional Neural Networks for Weakly Supervised Segmentation

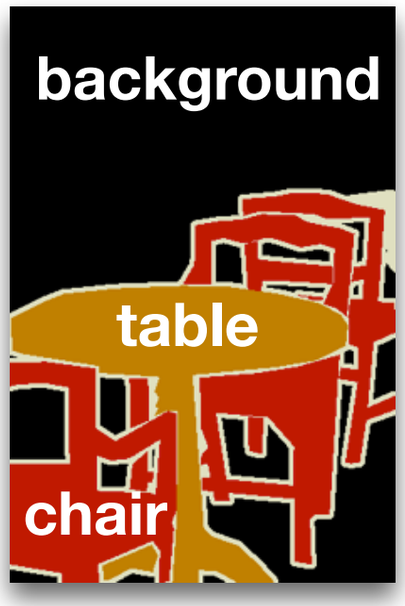
Deepak Pathak, Philipp Krähenbühl and Trevor Darrell

Multi-class Image Segmentation

- Assign a class label to each pixel in the image

Multi-class Image Segmentation

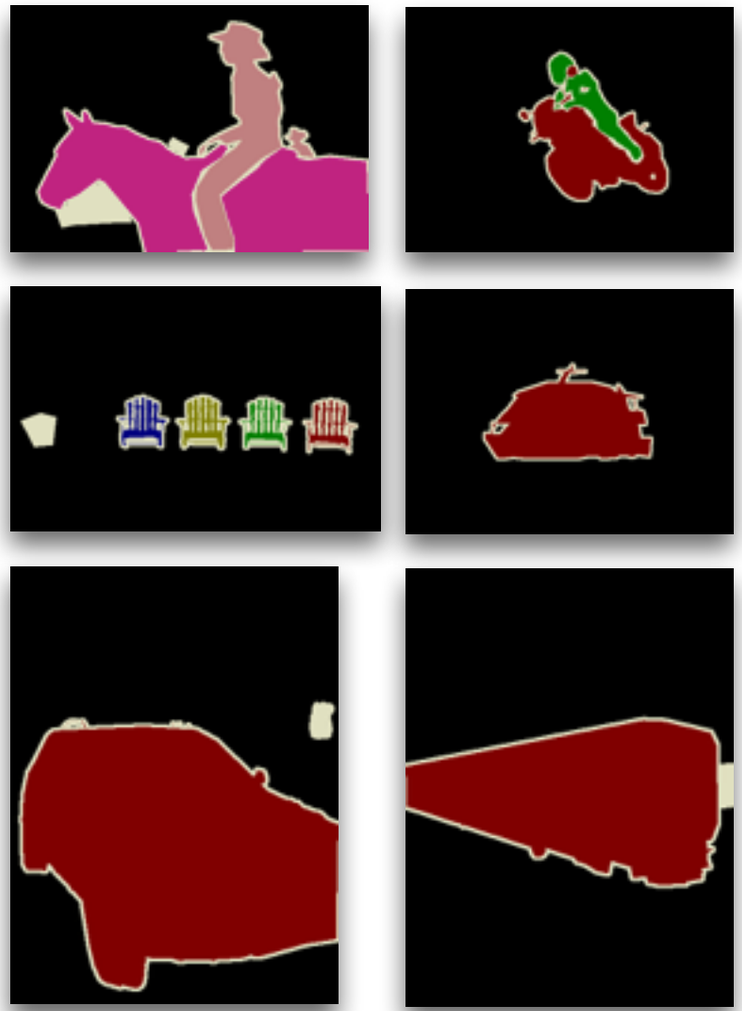
- Pixel-level classification
- Train a classifier $\mathbf{Q_i(I)}$
  - for each pixel **i**
    - and label **I**
- Convolutional neural network (CNN)
  - trained end-to-end



Fully Convolutional Networks for Semantic Segmentation [Long et al. CVPR 2015]

How does prior work train

CNN



- back propagation
- stochastic gradient descent (SGD)
- large labeled dataset

Limitation : Training Supervision

- Need full supervision
  - Time consuming to obtain
    - “79s per label per image”
      [Russakovsky et al. Arxiv 2015]
  - Expensive to obtain
  - Bottleneck for learning models at large scale

Weak Training Supervision

- Weak supervision
  - Class labels or tags
  - Cheap to obtain
    - “1s per label per image” [Russakovsky et al. Arxiv 2015]
- Scalable to large number of categories

person
horse
background

Training a CNN using weak supervision - Prior work

- Multiple instance learning
  - Tag present
    - at least one pixel takes label
  - Tag absent
    - No pixel takes that label
- Shown promise for weak detection

person
car
background

Multiple instance learning - Issues

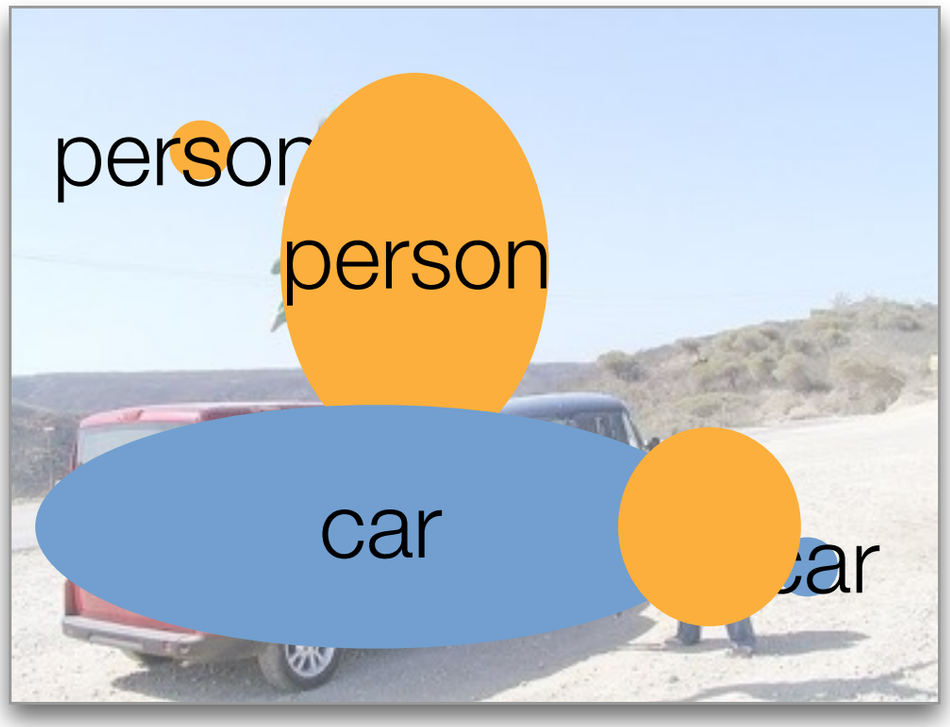
- Very weak signal
  - one pixel per class per image
- Converges to bad local minima
- Requires good initialization !
- Heuristics to get out of local optima

person
car
background

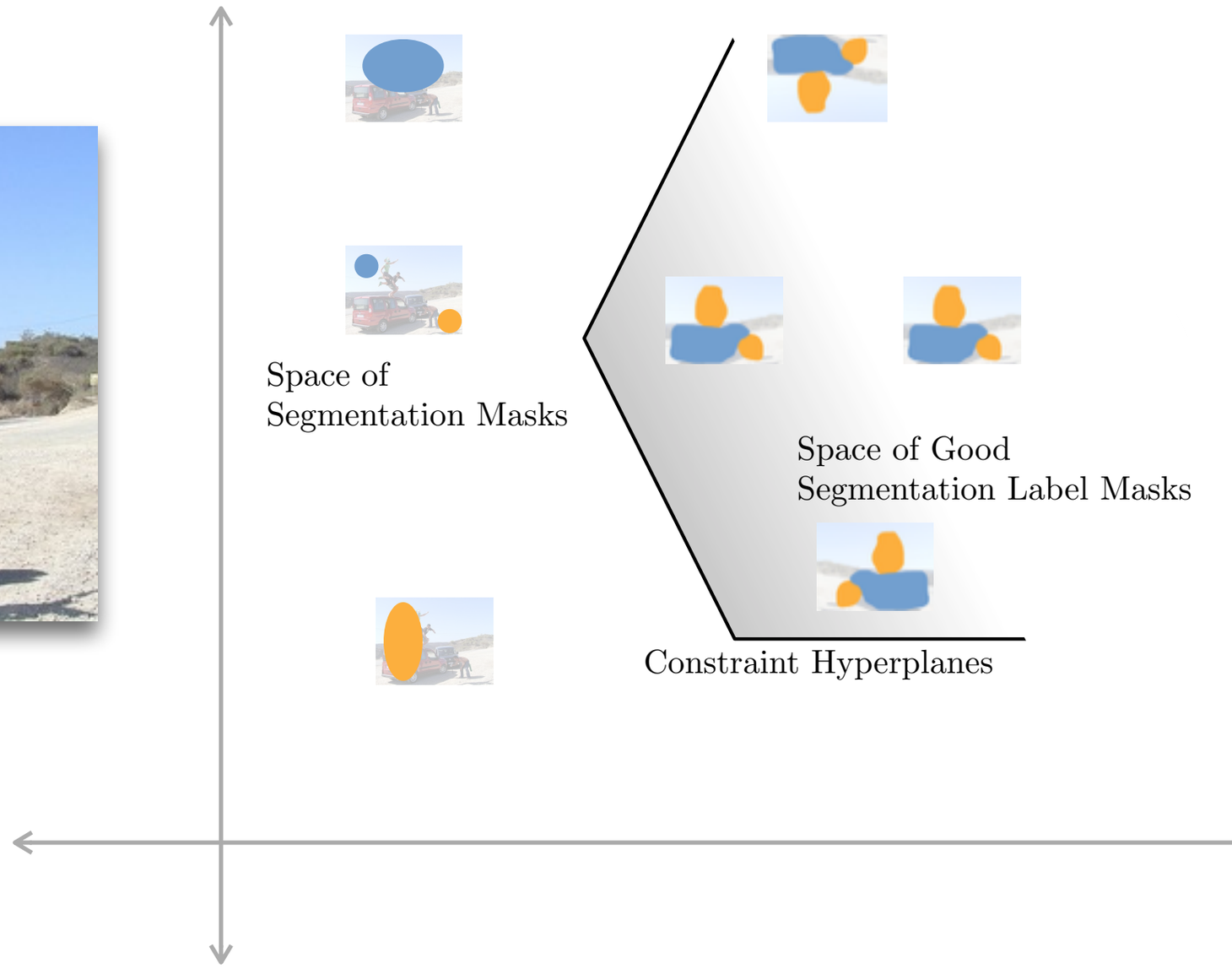
Weakly Supervised Training

- Is there a better description of the desired solution?

person
car
background



Idea : Weakly Supervised Training with constraints



Description Constraints

Suppression Constraint:

- suppress labels that do not appear in the image.

$$\sum_{i=1}^{n} p_i(l) \leq 0 \qquad \forall_{l \notin \mathcal{L}_I}$$

Description Constraints

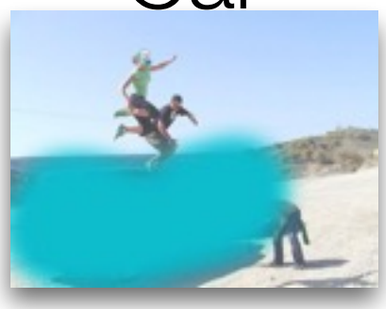
Foreground Constraint:

- label at least some pixels for each object present

$$a_l \leq \sum_{i=1}^{n} p_i(l) \qquad \forall_{l \in \mathcal{L}_I}$$

Person

Car



Description Constraints

Background Constraint:

- The number of background pixels in an image should be bounded say between 10% to 75%

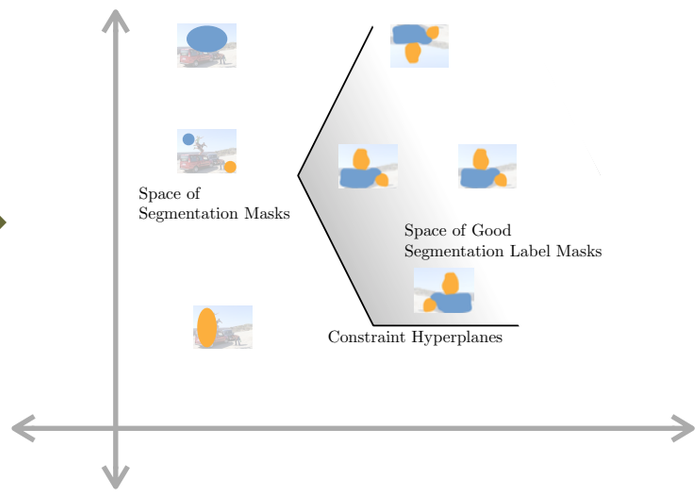
$$a_0 \leq \sum_{i=1}^{n} p_i(0) \leq b_0$$

Background

Constrained Convolutional Neural Network [CCNN]

Convolutional Neural Network + Constraints

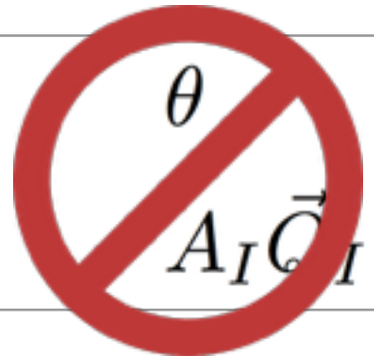
CNN



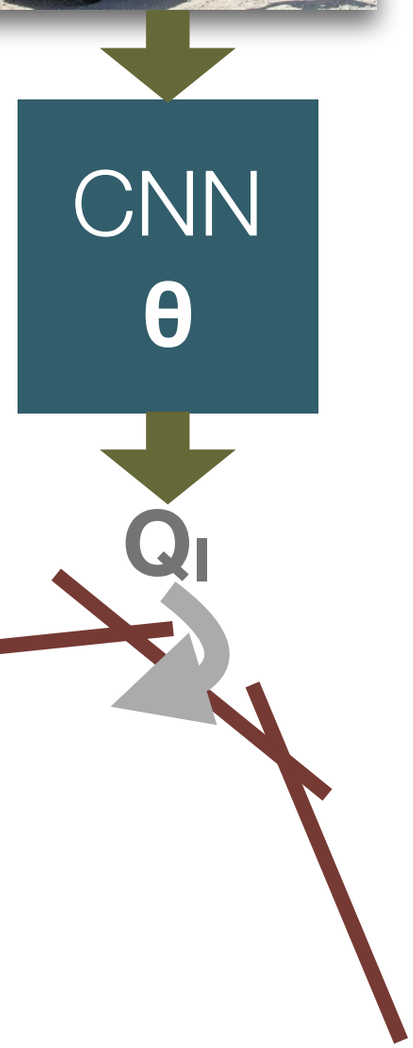
How to constrain CNN output?

- Constraints on CNN distribution $\mathbf{Q_I}$

$$\begin{aligned} \text{find} \quad & \theta \\ \text{subject to} \quad & A_I \vec{Q}_I \geq \vec{b}_I \qquad \forall I \end{aligned}$$

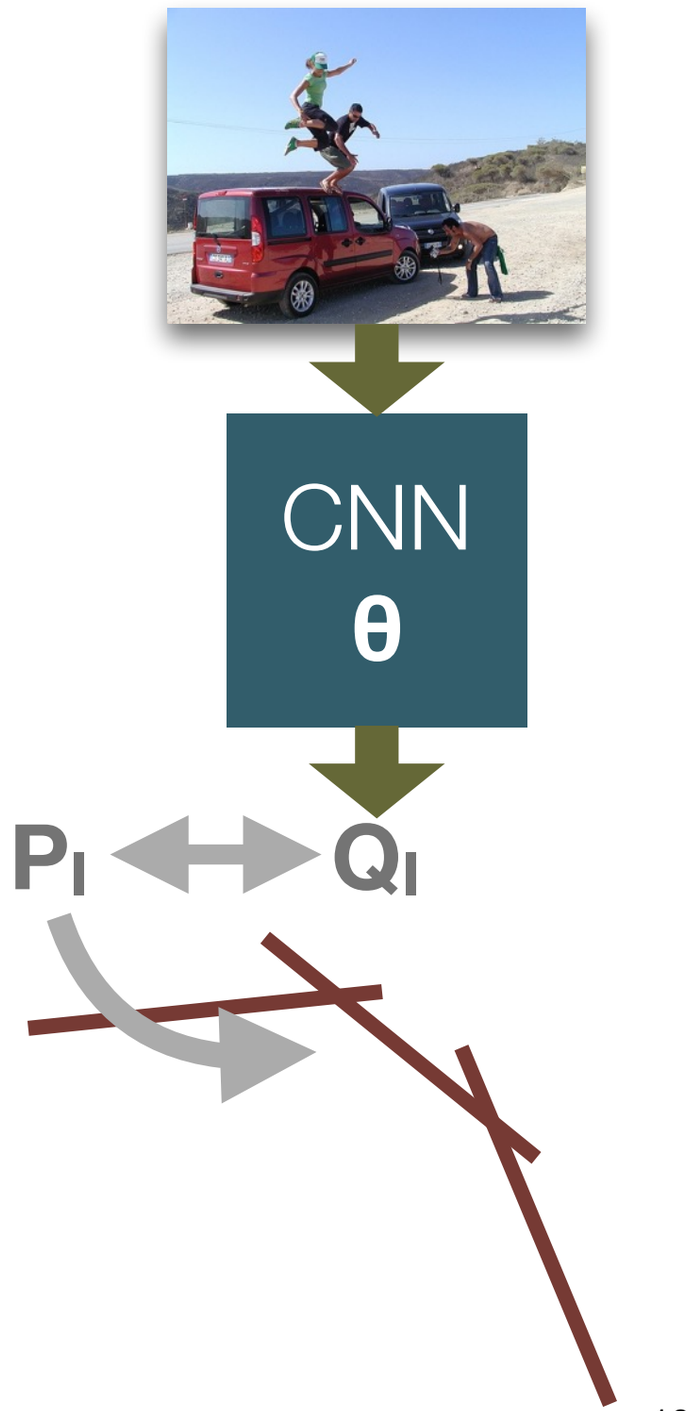


Expensive and Non-Convex



CCNN : Output as latent distribution

- Introduce latent variable $\mathbf{P_I}$ for distribution of network output
- Apply constraints on the latent distribution
- Minimize the distance between $\mathbf{P_I}$ and $\mathbf{Q_I}$

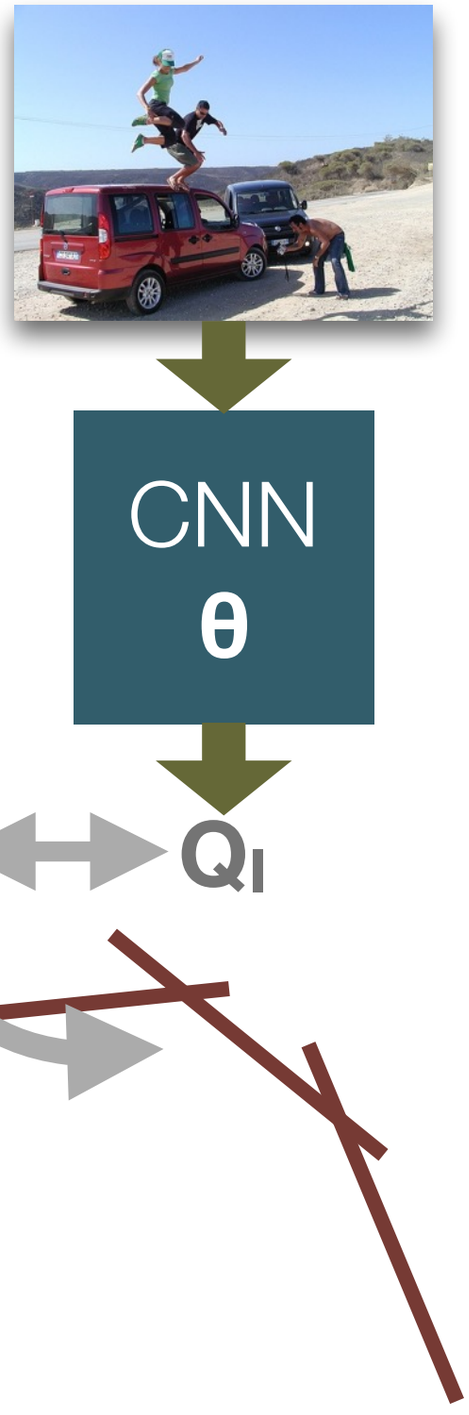


CCNN Optimization

$$\begin{aligned} \text{find} \quad & \theta \\ \text{subject to} \quad & A_I \vec{Q}_I \geq \vec{b}_I \qquad \forall I \end{aligned}$$

- KL-Divergence minimization between latent distribution and network output distribution

$$\begin{aligned} \underset{\theta, P}{\text{minimize}} \quad & D(P_I \| Q_I) \\ \text{subject to} \quad & A_I \vec{P}_I \geq \vec{b}, \\ & \vec{1}^\top \vec{P}_I = 1 \end{aligned}$$



CCNN Optimization

- Solves same optimization problem
- Convex in **P**
- Standard convnet loss for **Q**

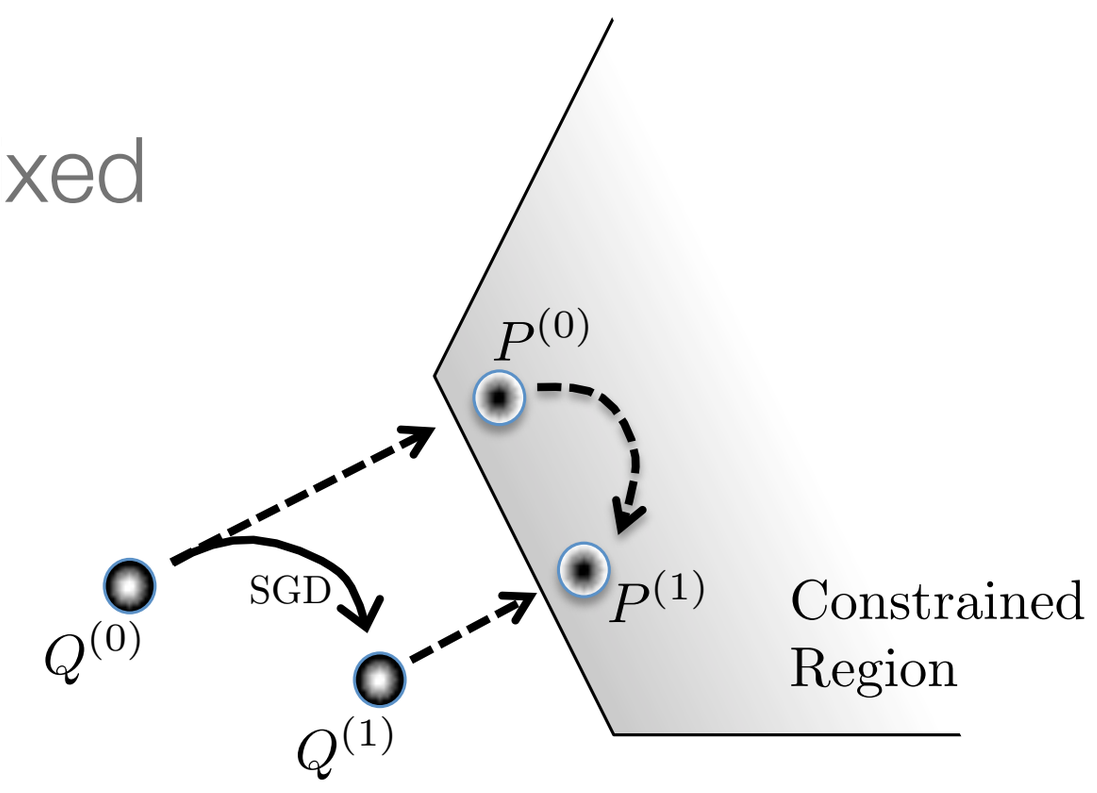
$$\begin{aligned} \underset{\theta, P}{\text{minimize}} \quad & D(P_I \| Q_I) \\ \text{subject to} \quad & A_I \vec{P}_I \geq \vec{b}, \\ & \vec{1}^\top \vec{P}_I = 1 \end{aligned}$$

- log-likelihood / cross entropy
- Convex for log-linear model
  - logistic regression

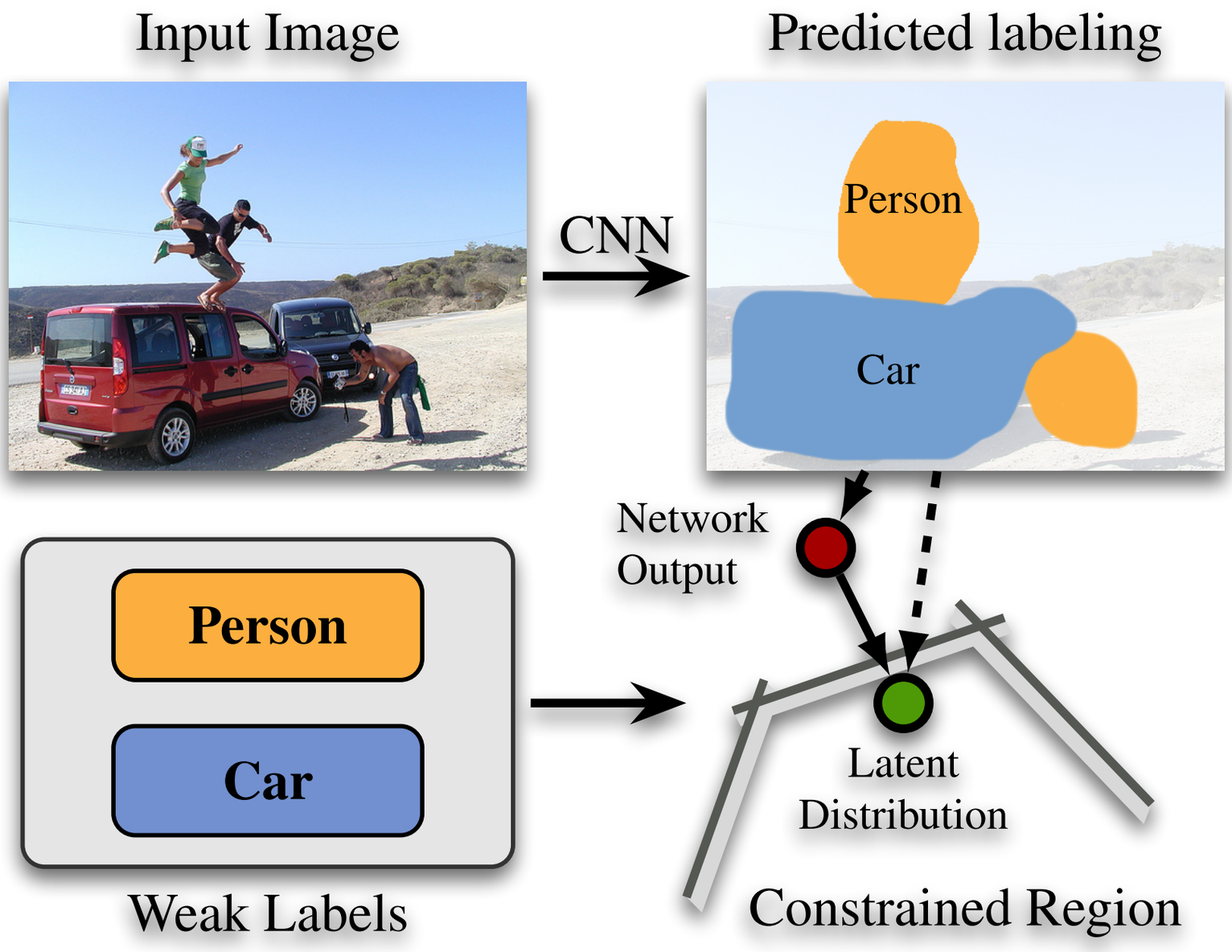
CCNN Optimization : Alternative Minimization

- Optimization using block coordinate descent :
  - Solve for **P** while convnet parameters **θ** fixed
  - Gradient step in **θ** while **P** fixed
- Each step guaranteed to decrease the overall objective

$$\begin{aligned} \underset{\theta, P}{\text{minimize}} \quad & D(P_I \| Q_I) \\ \text{subject to} \quad & A_I \vec{P}_I \geq \vec{b}, \\ & \vec{1}^\top \vec{P}_I = 1 \end{aligned}$$



Summary : Constrained CNN



Evaluation

- VOC 2012 dataset
  - Trained using 10,582 tagged images
  - Training time: 8hrs
    - Constraint satisfaction: **30ms** per image on CPU
    - Forward - Backward: **400ms** per image on GPU
- Evaluated on Intersection over Union score

Results : State of the Art

- State of the art weakly supervised semantic segmentation

Method	bgnd	aero	bike	bird	boat	bottle	bus	car	cat	chair	cow	table	dog	horse	mbike	person	plant	sheep	sofa	train	tv	mIoU
MIL-FCN [25]	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	24.9
MIL-Base [26]	37.0	10.4	12.4	10.8	05.3	05.7	25.2	21.1	25.2	04.8	21.5	08.6	29.1	25.1	23.6	25.5	12.0	28.4	08.9	22.0	11.6	17.8
MIL-Base w/ ILP [26]	**73.2**	25.4	**18.2**	22.7	21.5	28.6	39.5	44.7	46.6	11.9	**40.4**	11.8	**45.6**	**40.1**	35.5	35.2	20.8	**41.7**	17.0	34.7	30.4	32.6
EM-Adapt w/o CRF [24]	65.3	28.2	16.9	27.4	21.1	28.1	45.4	40.5	42.3	13.2	32.1	23.3	38.7	32.0	39.9	31.3	22.7	34.2	22.8	37.0	30.0	32.0
EM-Adapt [24]	67.2	**29.2**	17.6	**28.6**	**22.2**	29.6	**47.0**	44.0	44.2	14.6	35.1	**24.9**	41.0	34.8	41.6	32.1	24.8	37.4	**24.0**	38.1	31.6	33.8
CCNN w/o CRF	66.3	24.6	17.2	24.3	19.5	34.4	45.6	44.3	44.7	14.4	33.8	21.4	40.8	31.6	42.8	39.1	28.8	33.2	21.5	37.4	34.4	33.3
CCNN	68.5	25.5	18.0	25.4	20.2	**36.3**	46.8	**47.1**	**48.0**	**15.8**	37.9	21.0	44.5	34.5	**46.2**	**40.7**	**30.4**	36.3	22.2	**38.8**	**36.9**	**35.3**

Additional 1-bit Supervision

1-bit additional information:

- object size is ‘small’ (<10%) or ‘large’ (>10%)

- Size Constraints
  - Boost large objects $a_l \leq \sum_{i=1}^{n} p_i(l)$
  - Limit small objects $\sum_{i=1}^{n} p_i(l) \leq b_l$

Car

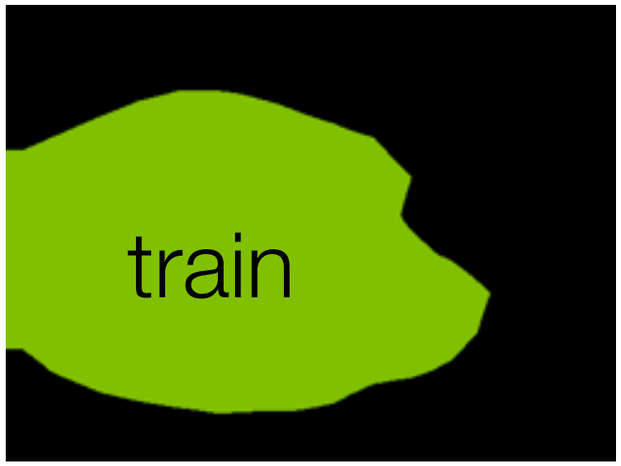
Person

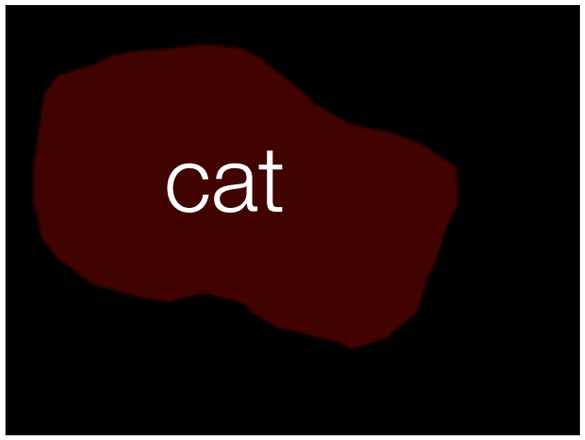
Results : Comparison with Fully Supervised

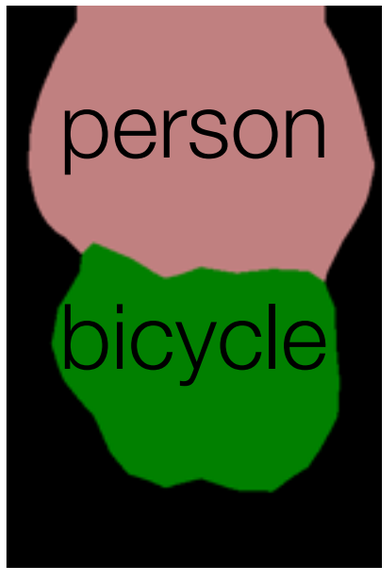
- 10% improvement using 1-bit additional supervision at training time.

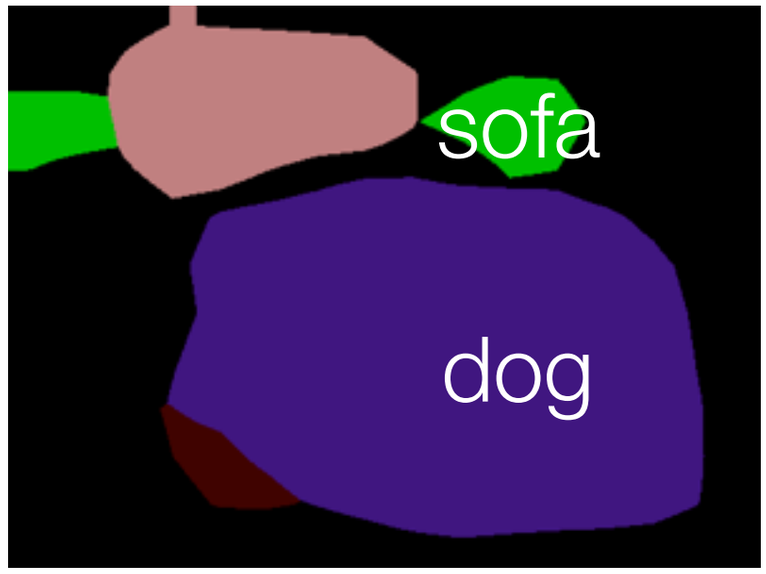
Method	aero	bike	bird	boat	bottle	bus	car	cat	chair	cow	table	dog	horse	mbike	person	plant	sheep	sofa	train	tv	mIoU
Fully Supervised:																					
SDS [14]	63.3	25.7	63.0	39.8	59.2	70.9	61.4	54.9	16.8	45.0	48.2	50.5	51.0	57.7	63.3	31.8	58.7	31.2	55.7	48.5	51.6
FCN-8s [21]	76.8	34.2	68.9	49.4	60.3	75.3	74.7	77.6	21.4	62.5	46.8	71.8	63.9	76.5	73.9	45.2	72.4	37.4	70.9	55.1	62.2
TTIC Zoomout [23]	81.9	35.1	78.2	57.4	56.5	80.5	74.0	79.8	22.4	69.6	53.7	74.0	76.0	76.6	68.8	44.3	70.2	40.2	68.9	55.3	64.4
DeepLab-CRF [4]	78.4	33.1	78.2	55.6	65.3	81.3	75.5	78.6	25.3	69.2	52.7	75.2	69.0	79.1	77.6	54.7	78.3	45.1	73.3	56.2	**66.4**
Weakly Supervised:																					
CCNN w/ tags	24.2	19.9	26.3	18.6	38.1	51.7	42.9	48.2	15.6	37.2	18.3	43.0	38.2	52.2	40.0	33.8	36.0	21.6	33.4	38.3	35.6
CCNN w/ size	36.7	23.6	47.1	30.2	40.6	59.5	54.3	51.9	15.9	43.3	34.8	48.2	42.5	59.2	43.1	35.5	45.2	31.4	46.2	42.2	43.3
CCNN w/ size (CRF tuned)	42.3	24.5	56.0	30.6	39.0	58.8	52.7	54.8	14.6	48.4	34.2	52.7	46.9	61.1	44.8	37.4	48.8	30.6	47.7	41.7	**45.1**

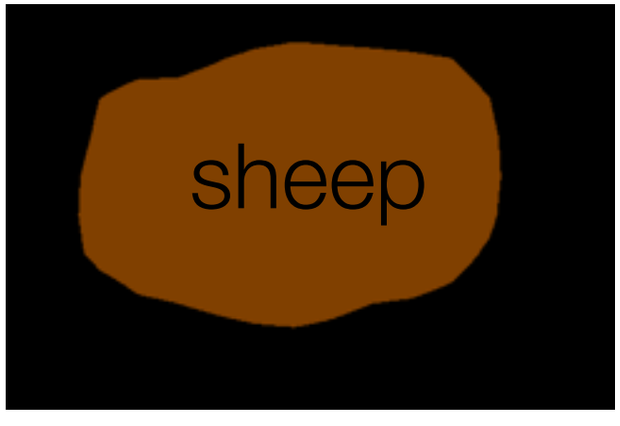

124
124


train

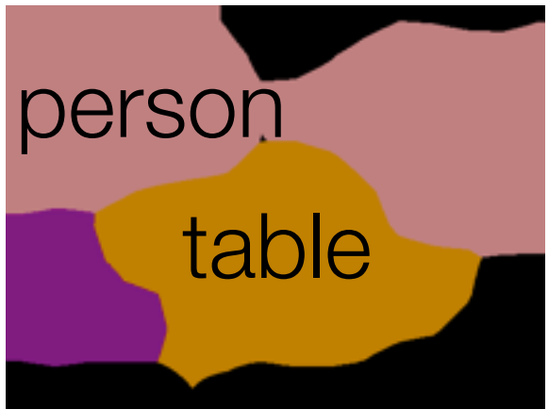

cat

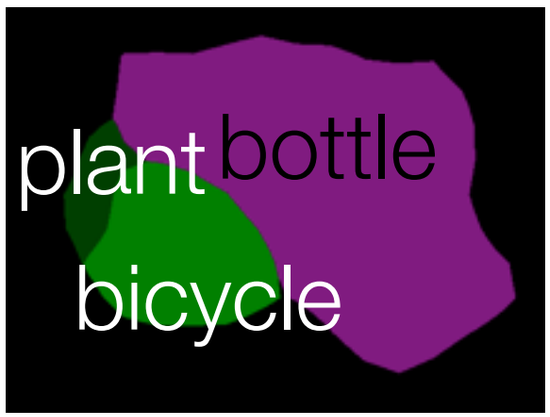

person
bicycle

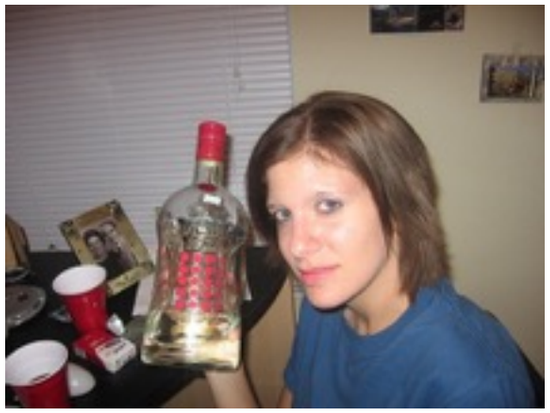

sofa
dog

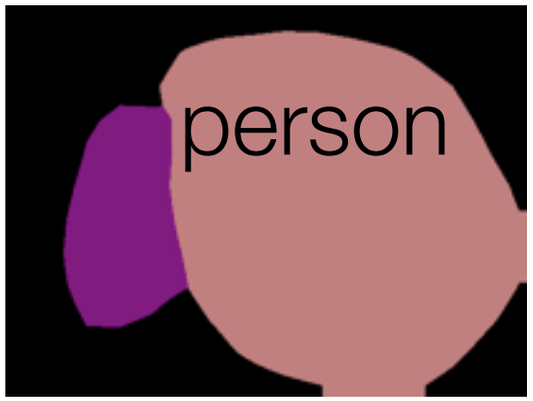

sheep

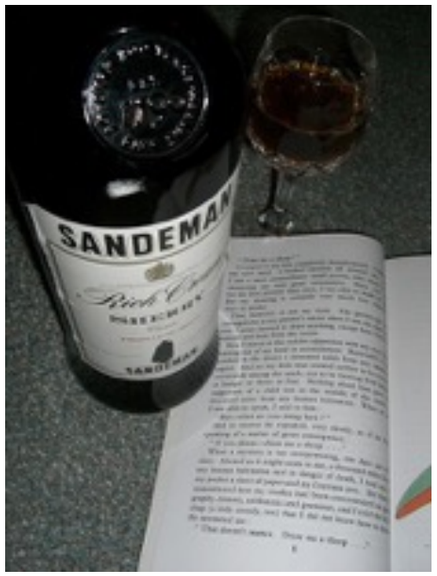

horse


person
table


plant
bottle
bicycle




person


SANDEMAN

Questions?

- Paper (and code) is available :

  Constrained Convolutional Neural Networks for Weakly Supervised Segmentation, ICCV 2015

  http://arxiv.org/abs/1506.03648