Category-level recognition

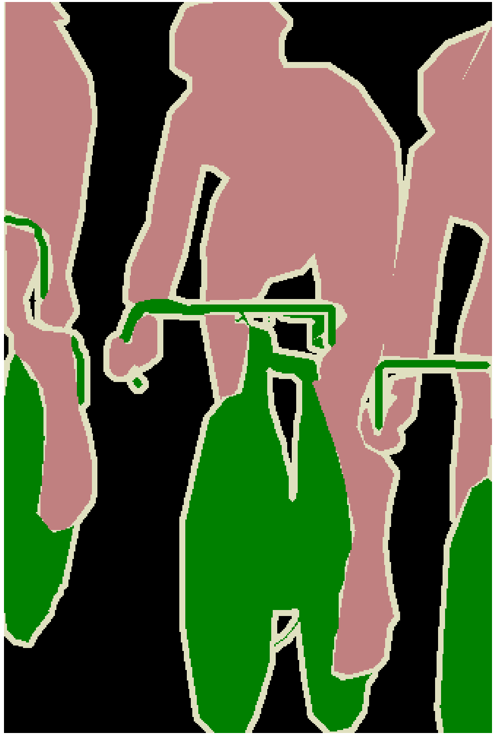
1. Classification

- Contains a car? [yes/no]
- List categories present
- Which city is this from?

2. Detection

- Localize horses (if present)
- Segment people (if present)
- Parse objects into parts

Semantic Image Segmentation



Datasets

- PASCAL VOC segmentation
  - O(10K) images, 20 classes + bgnd
  - Also bounding box annotations

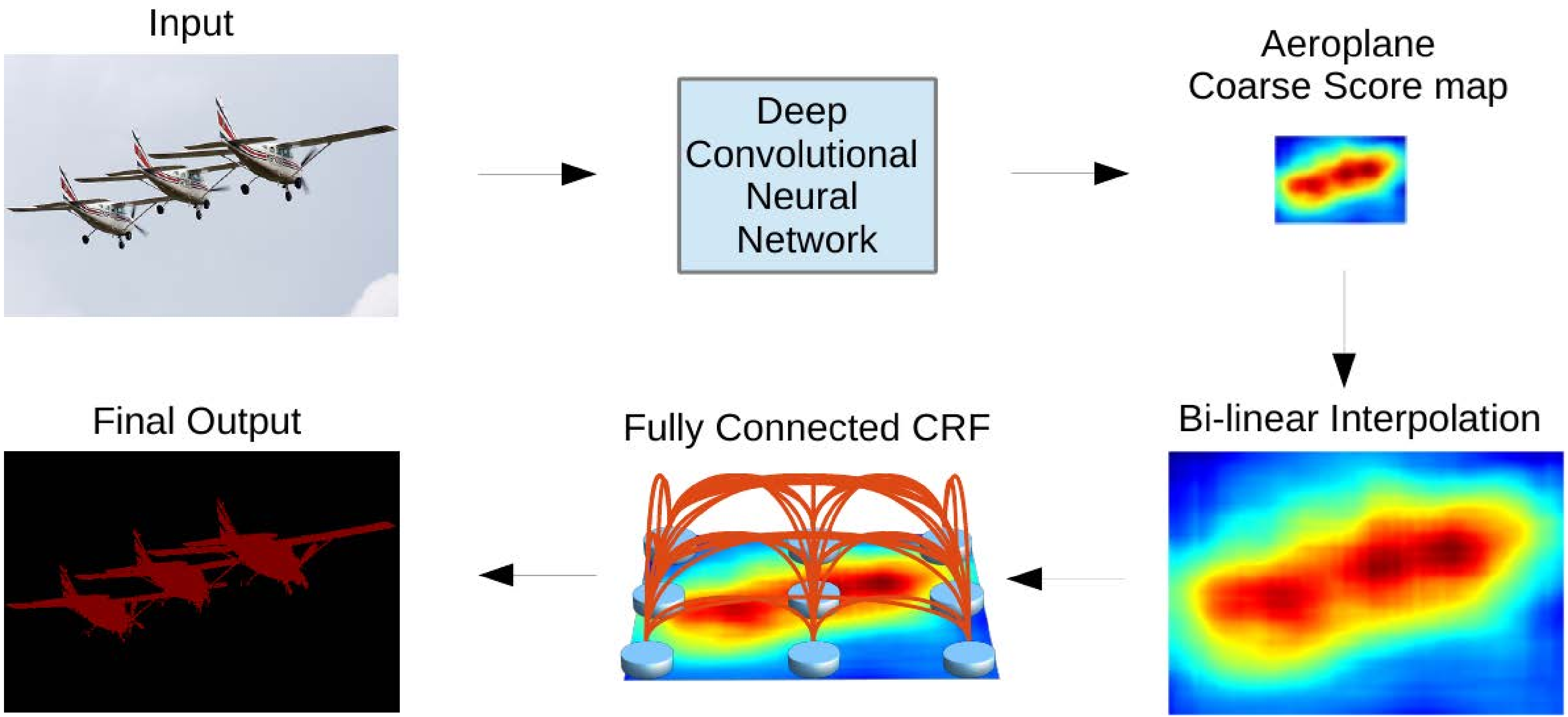
- MS COCO
  - O(100K) images, 80 classes + bgnd
  - Also 5 text captions / image

Applications

- Fine-grained image recognition
  - Explicit localization
  - More natural description of “stuff”

- Image manipulation and editing

System Overview



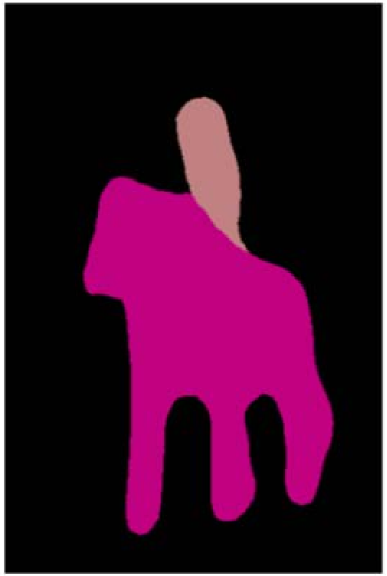
Basic Ingredients: (1) Conv Nets

- Train convnet to predict label of center pixel
- Apply in sliding window fashion

See also: J Long, E Shelhamer, T Darrell: Fully Convolutional Networks for Semantic Segmentation (arXiv)

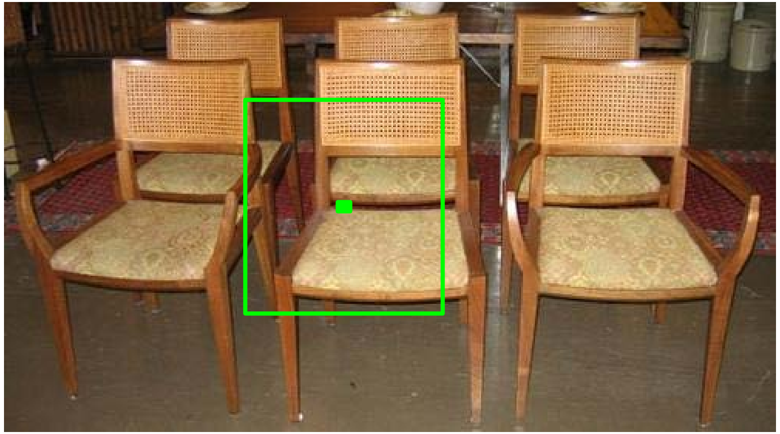
The accuracy/localization tradeoff

- Large CNN receptive field
  → poor performance near boundaries



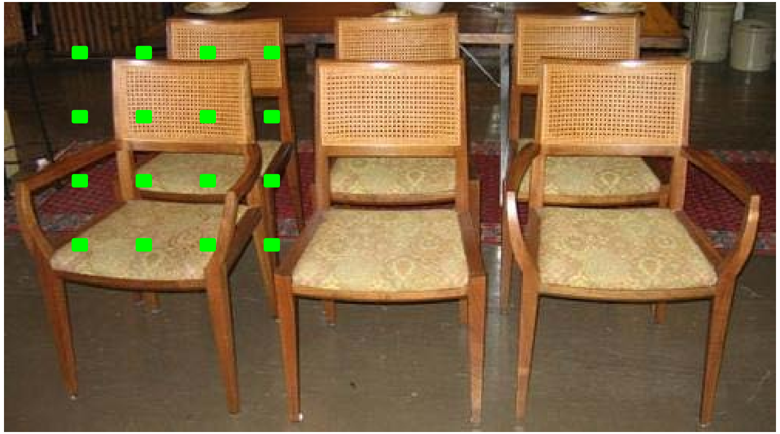
Explicit control of receptive field size

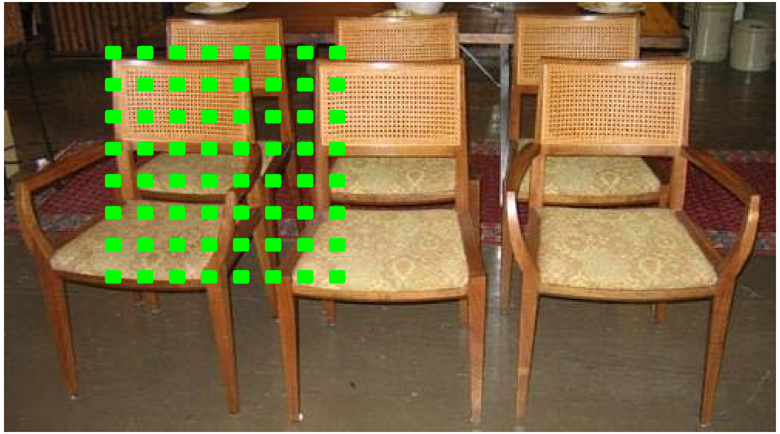
- Reduce RF size by conv layer manipulation
- In VGG: Subsample first FC layer 7x7 → 3x3



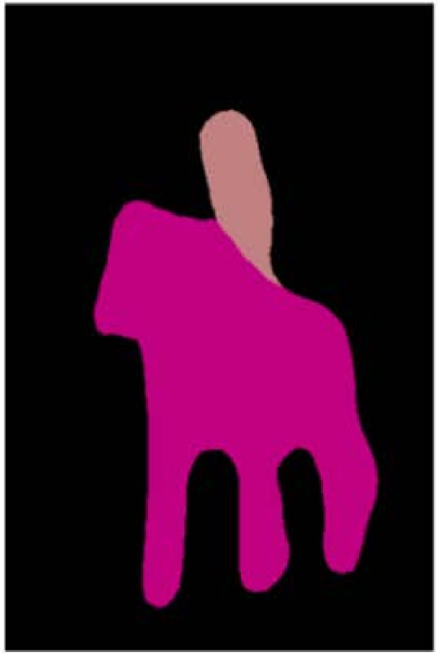
Explicit control of response density

- Decrease score map stride: 32 → 8
- Efficient implementation with “atrous” algorithm





Accurate Boundary Recovery w. CRF



Raw score maps

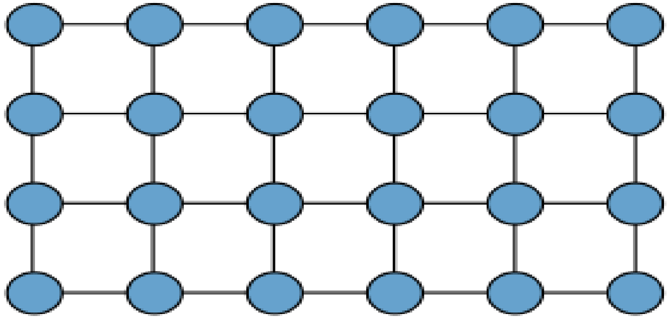
After dense CRF

CRFs in a nutshell

CRF slides credit: Iasonas Kokkinos

- a set of i.i.d. samples $\mathcal{D} = \{(x^n, y^n)\}_{n=1,\ldots,N}, \quad (x^n, y^n) \sim d(x, y)$
- feature functions $(\, \phi_1(x, y), \ldots, \phi_D(x, y) \,) \equiv: \phi(x, y)$
- parametrized family $p(y|x, w) = \frac{1}{Z(x,w)} \exp(\, \langle w, \phi(x, y) \rangle \,)$

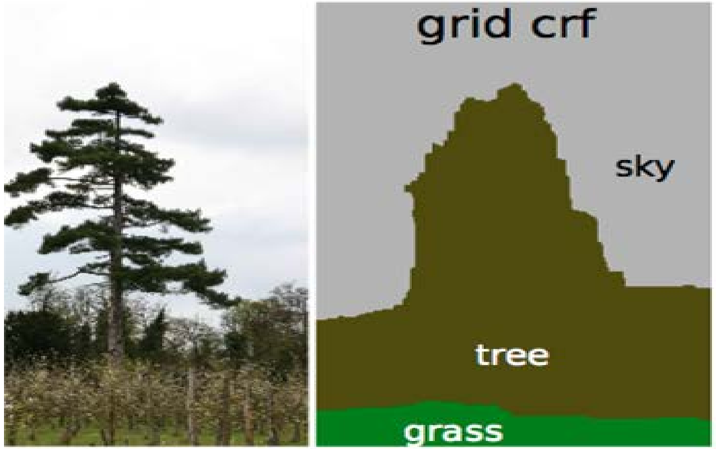
$$E(\mathbf{x}) = \sum_i \underbrace{\psi_u(x_i)}_{\text{unary term}} + \sum_i \sum_{j \in \mathcal{N}_i} \underbrace{\psi_p(x_i, x_j)}_{\text{pairwise term}}$$



- Unary term
  - From classifier
  - TextonBoost [Shotton et al. 09]
- Pairwise term
  - Consistent labeling

Grid CRF

$$E(\mathbf{x}) = \sum_i \underbrace{\psi_u(x_i)}_{\text{unary term}} + \sum_i \sum_{j \in \mathcal{N}_i} \underbrace{\psi_p(x_i, x_j)}_{\text{pairwise term}}$$



- Efficient inference
  - 1 second for 50′000 variables
- Limited expressive power
- Only local interactions
- Excessive smoothing of object boundaries
  - Shrinking bias

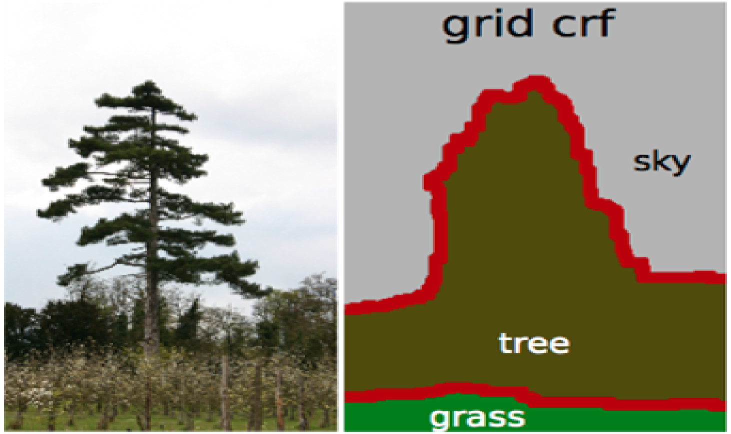
Grid CRF limitations

$$E(\mathbf{x}) = \sum_i \underbrace{\psi_u(x_i)}_{\text{unary term}} + \sum_i \sum_{j \in \mathcal{N}_i} \underbrace{\psi_p(x_i, x_j)}_{\text{pairwise term}}$$

- Efficient inference
  - 1 second for 50′000 variables
- Limited expressive power
- Only local interactions
- Excessive smoothing of object boundaries
  - Shrinking bias

Grid CRF limitations

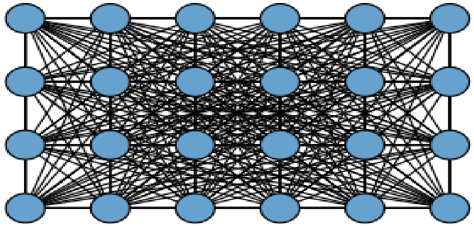
$$E(\mathbf{x}) = \sum_i \underbrace{\psi_u(x_i)}_{\text{unary term}} + \sum_i \sum_{j \in \mathcal{N}_i} \underbrace{\psi_p(x_i, x_j)}_{\text{pairwise term}}$$



- Efficient inference
  - 1 second for 50'000 variables
- Limited expressive power
- Only local interactions
- Excessive smoothing of object boundaries
  - Shrinking bias

2011: Fully-connected CRF (Krahnebuhl & Koltun)

$$E(\mathbf{x}) = \sum_i \underbrace{\psi_u(x_i)}_{\text{unary term}} + \sum_i \sum_{j>i} \underbrace{\psi_p(x_i, x_j)}_{\text{pairwise term}}$$



- Every node is connected to every other node
  - Connections weighted differently

P Krähenbühl and V Koltun, Efficient Inference in Fully Connected CRFs with Gaussian Edge Potentials, NIPS 2011

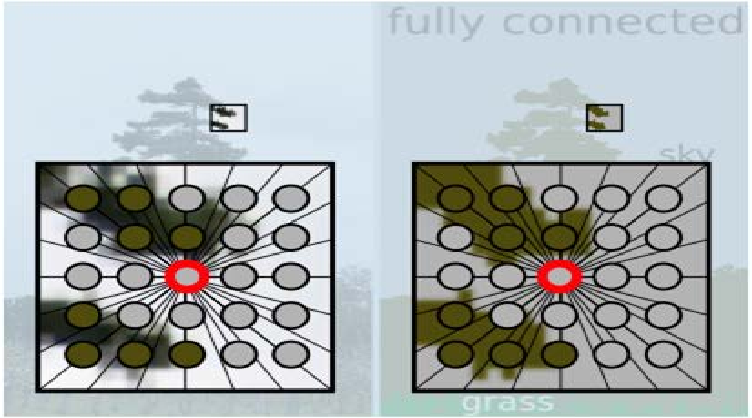
Fully-connected CRF

$$E(\mathbf{x}) = \sum_i \underbrace{\psi_u(x_i)}_{\text{unary term}} + \sum_i \sum_{j>i} \underbrace{\psi_p(x_i, x_j)}_{\text{pairwise term}}$$

- Long-range interactions
- No more shrinking bias

Fully-connected CRF

$$E(\mathbf{x}) = \sum_i \underbrace{\psi_u(x_i)}_{\text{unary term}} + \sum_i \sum_{j>i} \underbrace{\psi_p(x_i, x_j)}_{\text{pairwise term}}$$



- Long-range interactions
- No more shrinking bias

Fully-connected CRF

$$E(\mathbf{x}) = \sum_i \underbrace{\psi_u(x_i)}_{\text{unary term}} + \sum_i \sum_{j>i} \underbrace{\psi_p(x_i, x_j)}_{\text{pairwise term}}$$

- Long-range interactions
- No more shrinking bias

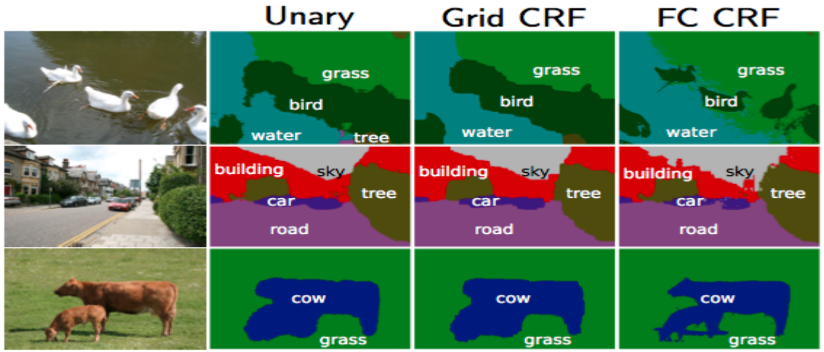
Fully-connected CRF: FAST

- Inference in 0.2 seconds
  - 50'000 variables
  - MCMC inference: 36 hrs
- Pairwise potentials: linear combinations of Gaussians

MSRC dataset

- 591 images
- 21 classes

	Time	Global	Avg
Unary	–	84.0	76.6
Grid CRF	1s	84.6	77.2
FC CRF	**0.2s**	**86.0**	**78.3**



How? Mean Field + some tricks

Trick: Pairwise Term

$$\theta_{ij}(x_i, x_j) = \mu(x_i, x_j) \sum_{m=1}^{K} w_m \cdot k^m(\boldsymbol{f}_i, \boldsymbol{f}_j)$$

Potts model

$$\mu(x_i, x_j) = 1 \text{ if } x_i \neq x_j$$

Gaussian kernels

$$w_1 \exp\left(-\frac{||p_i - p_j||^2}{2\sigma_\alpha^2} - \frac{||I_i - I_j||^2}{2\sigma_\beta^2}\right) + w_2 \exp\left(-\frac{||p_i - p_j||^2}{2\sigma_\gamma^2}\right)$$

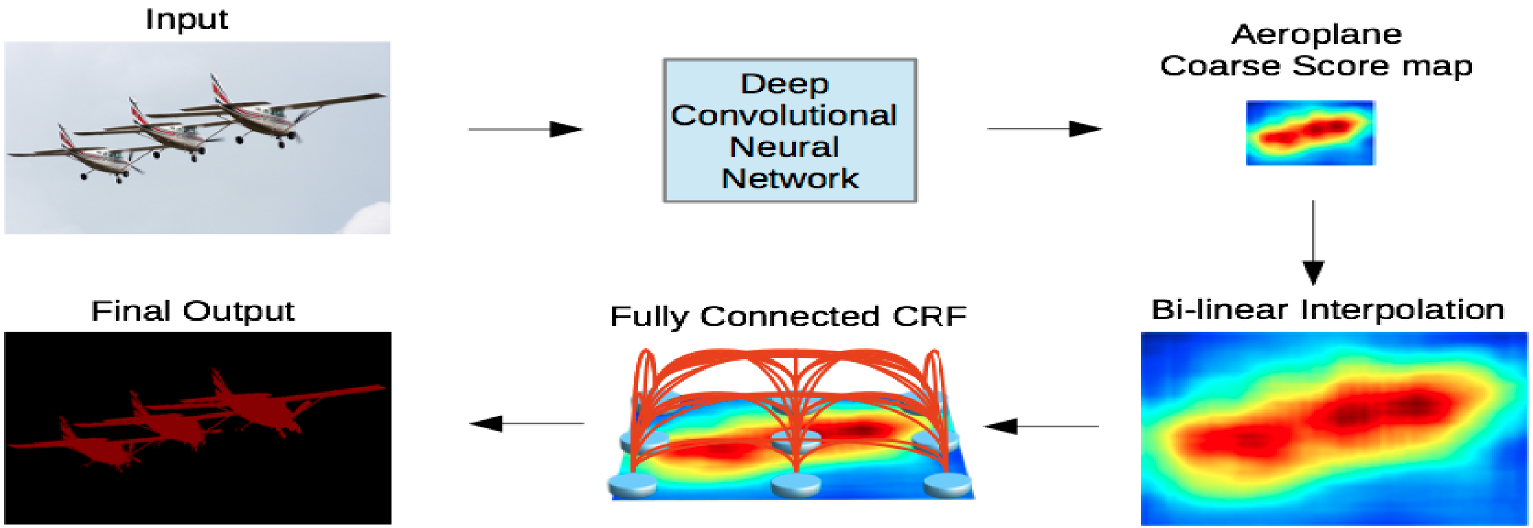
$$Q_i(x_i = l) = \frac{1}{Z_i} \exp\left\{-\psi_u(x_i) - \sum_{l' \in \mathcal{L}} \mu(l, l') \sum_{m=1}^{K} w^{(m)} \sum_{j \neq i} k^{(m)}(\mathbf{f}_i, \mathbf{f}_j) Q_j(l')\right\}$$

Fast summation through separable convolution

- Initialize $Q_i(x_i) \leftarrow \frac{1}{Z_i} \exp\{-\phi_u(x_i)\}$
- **while** not converged
  - Message passing: $\tilde{Q}_i^{(m)}(l) \leftarrow \sum_{j \neq i} k^{(m)}(\mathbf{f}_i, \mathbf{f}_j) Q_j(l)$
  - Compatibility transform: $\hat{Q}_i(x_i) \leftarrow \sum_{l \in \mathcal{L}} \mu^{(m)}(x_i, l) \sum_m w^{(m)} \tilde{Q}_i^{(m)}(l)$
  - Local update: $Q_i(x_i) \leftarrow \exp\{-\psi_u(x_i) - \hat{Q}_i(x_i)\}$
  - Normalize: $Q_i(x_i)$

P Krähenbühl, V Koltun, NIPS 2011

2014: Fully connected CRFs + Deep Classifiers



$$E(\boldsymbol{x}) = \sum_i \theta_i(x_i) + \sum_{ij} \theta_{ij}(x_i, x_j) \qquad \theta_i(x_i) = -\log P(x_i)$$

L.-C. Chen, G. Papandreou, I. Kokkinos, K. Murphy and A. Yuille Semantic Image Segmentation with Deep Convolutional Nets and Fully Connected CRFs, http://arxiv.org/abs/1412.7062

Evolution from mean field updates

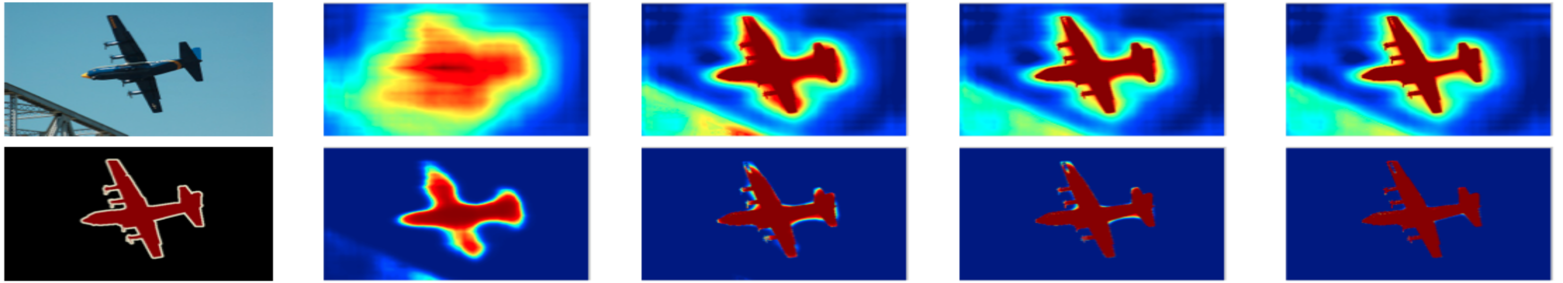
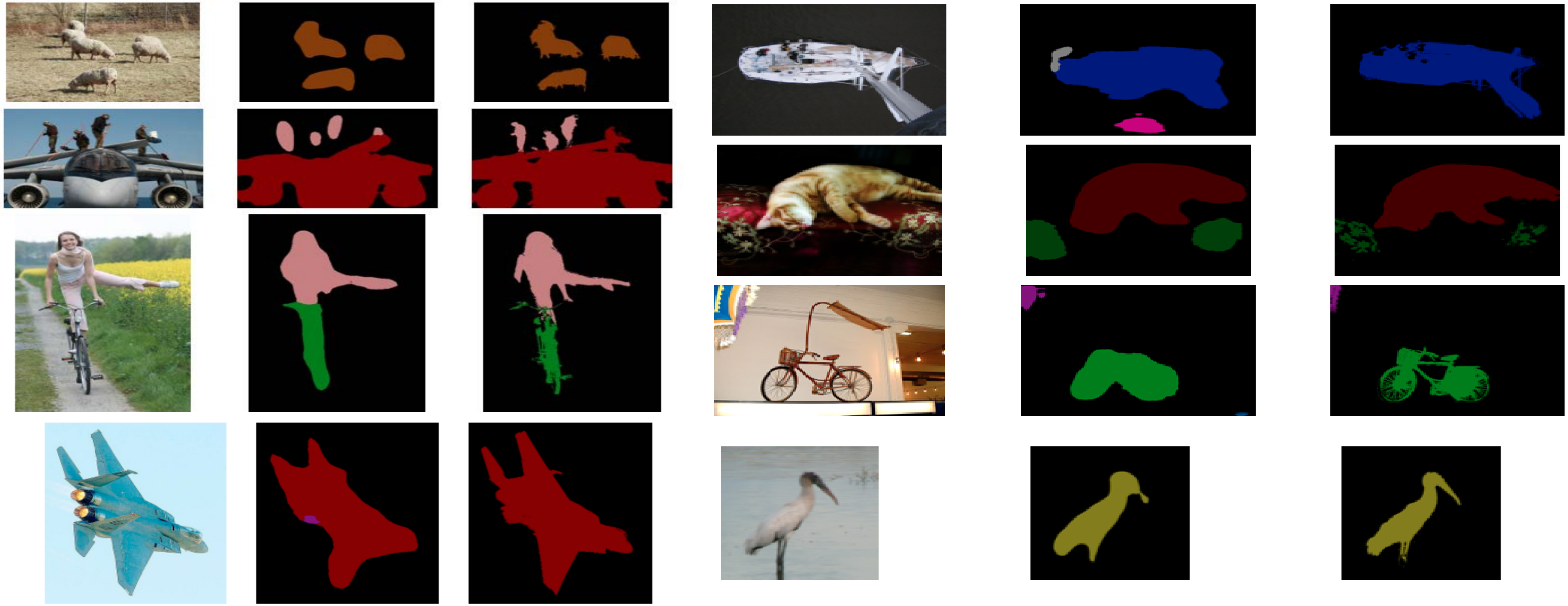


Figure 1: Score map (input before softmax function) and belief map (output of softmax function) for Aeroplane. We show the score (1st row) and belief (2nd row) maps after each mean field iteration. The output of the last DCNN layer is used as input to the mean field inference method.

Our Results (input, DCNN, CRF-DCNN)

Our Results (input, DCNN, CRF-DCNN)



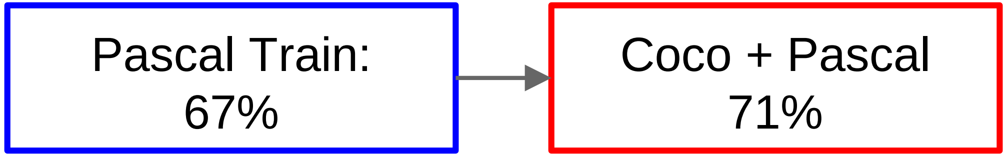
Comparisons to other techniques on VOC test

	mean	aero plane	bicycle	bird	boat	bottle	bus	car	cat	chair	cow	dining table	dog	horse	motor bike	person	potted plant	sheep	sofa	train	tv/ monitor	submission date
DeepLab-CRF-MSc [?]	67.1	80.4	36.8	77.4	55.2	66.4	81.5	77.5	78.9	27.1	68.2	52.7	74.3	69.6	79.4	79.0	56.9	78.8	45.2	72.7	59.3	30-Dec-2014
DeepLab-CRF [?]	66.4	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	23-Dec-2014
TTI_zoomout_16 [?]	64.4	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	24-Nov-2014
FCN-8s [?]	62.2	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	12-Nov-2014
MSRA_CFM [?]	61.8	75.7	26.7	69.5	48.8	65.6	81.0	69.2	73.3	30.0	68.7	51.5	69.1	68.1	71.7	67.5	50.4	66.5	44.4	58.9	53.5	17-Dec-2014
TTI_zoomout [?]	58.4	70.3	31.9	68.3	46.4	52.1	75.3	68.4	75.3	19.2	58.4	49.9	69.6	63.0	70.1	67.6	41.5	64.0	34.9	64.2	47.3	17-Nov-2014
SDS [?]	51.6	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	21-Jul-2014
NUS_UDS [?]	50.0	67.0	24.5	47.2	45.0	47.9	65.3	60.6	58.5	15.5	50.8	37.4	45.8	59.9	62.0	52.7	40.8	48.2	36.8	53.1	45.6	29-Oct-2014
TTIC-divmbest-rerank [?]	48.1	62.7	25.6	46.9	43.0	54.8	58.4	58.6	55.6	14.6	47.5	31.2	44.7	51.0	60.9	53.5	36.6	50.9	30.1	50.2	46.8	15-Nov-2012
BONN_O2PCPMC_FGT_SEGM [?]	47.8	64.0	27.3	54.1	39.2	48.7	56.6	57.7	52.5	14.2	54.8	29.6	42.2	58.0	54.8	50.2	36.6	58.6	31.6	48.4	38.6	08-Aug-2013
BONN_O2PCPMC_FGT_SEGM [?]	47.5	63.4	27.3	56.1	37.7	47.2	57.9	59.3	55.0	11.5	50.8	30.5	45.0	58.4	57.4	48.6	34.6	53.3	32.4	47.6	39.2	23-Sep-2012
BONNGC_O2P_CPMC_CSI [?]	46.8	63.6	26.8	45.6	41.7	47.1	54.3	58.6	55.1	14.5	49.0	30.9	46.1	52.6	58.2	53.4	32.0	44.5	34.6	45.3	43.1	23-Sep-2012
BONN_CMBR_O2P_CPMC_LIN [?]	46.7	63.9	23.8	44.6	40.3	45.5	59.6	58.7	57.1	11.7	45.9	34.9	43.0	54.9	58.0	51.5	34.6	44.1	29.9	50.5	44.5	23-Sep-2012

Pre-CNN: Up to 50% → CNN: 60-64% → CNN + CRF: >67%

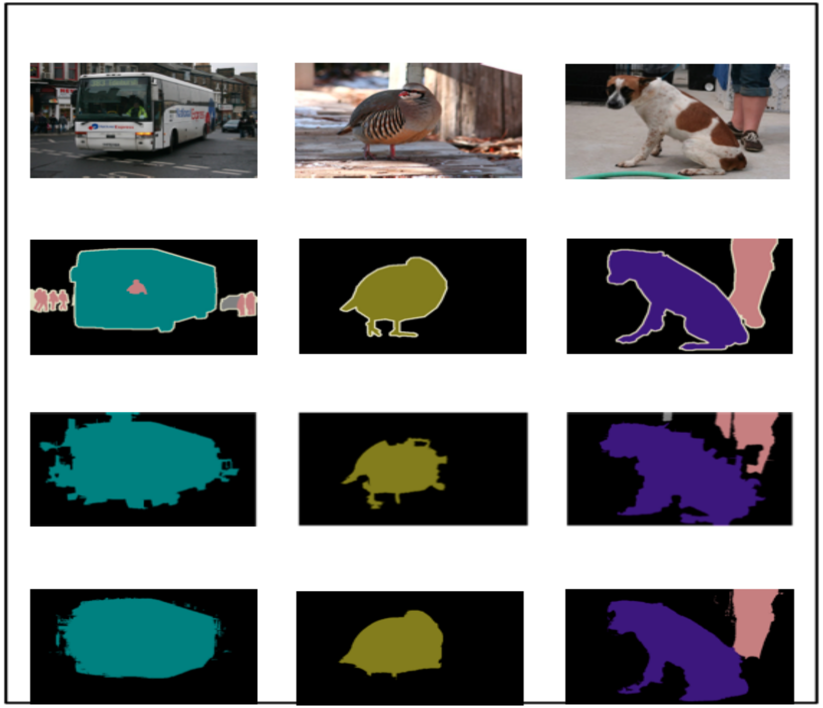
More data helps

- Pre-train on MS-COCO, refine in PASCAL:

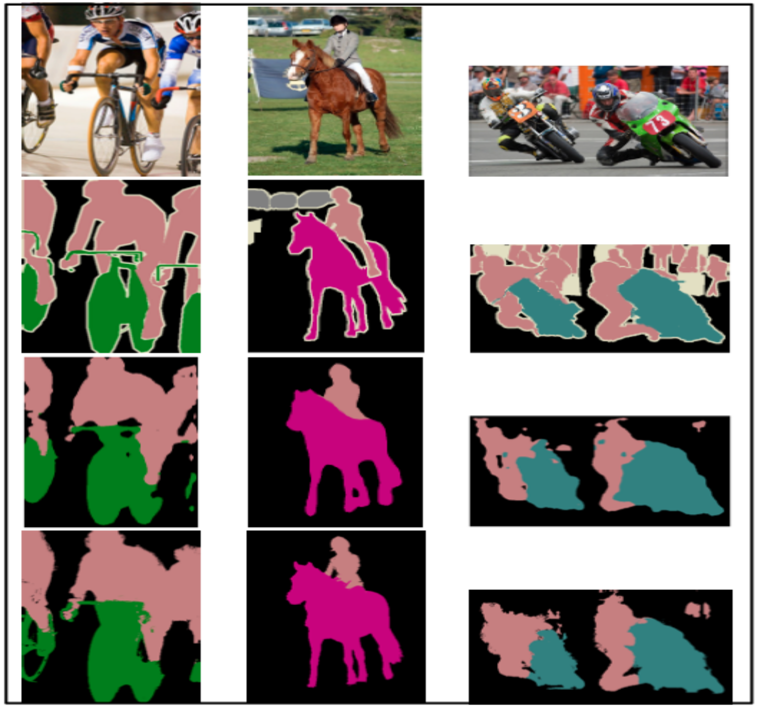


- Preliminary eval on COCO: ~40% mean IoU

Comparisons to previous state-of-the-art

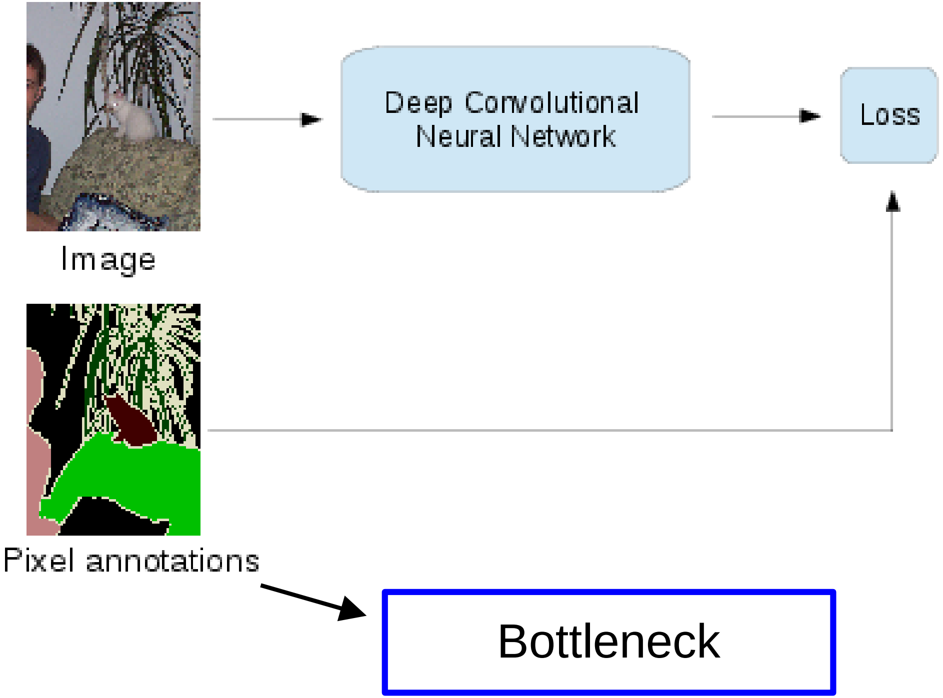


(b) TTI-Zoomout-16 vs. DeepLab-CRF



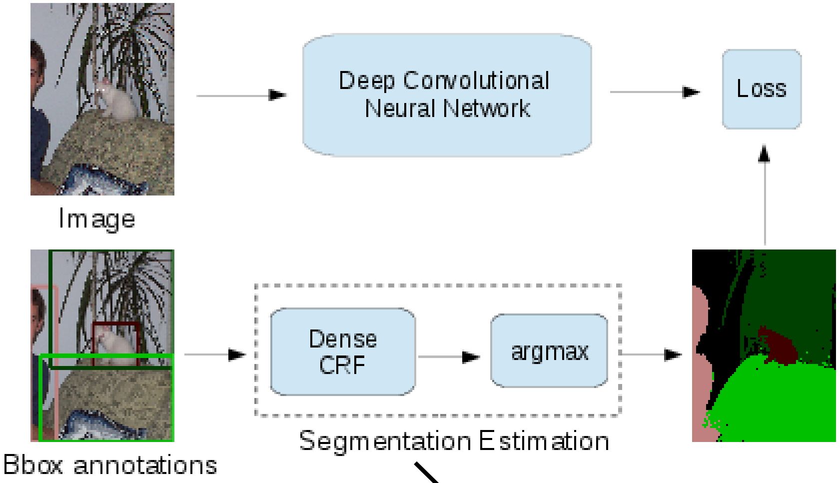
(a) FCN-8s vs. DeepLab-CRF

Towards Weaker Annotations



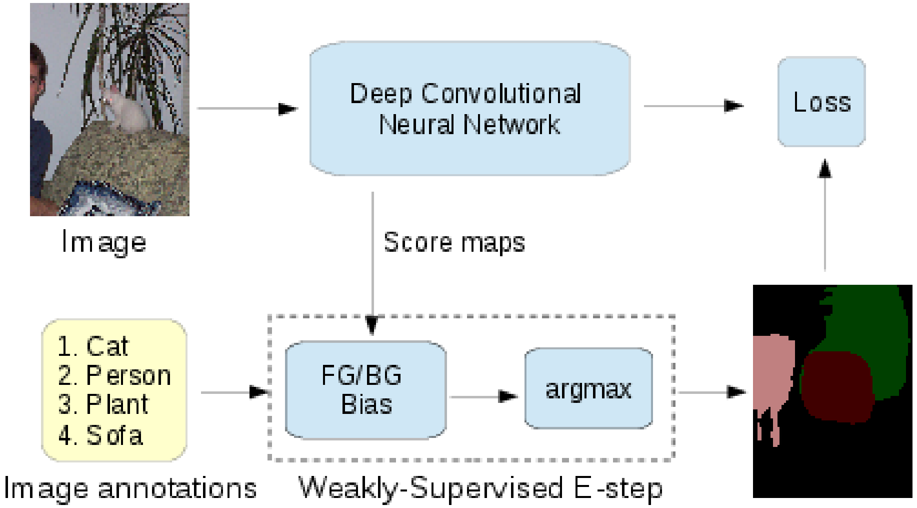
G. Papandreou, L.-C. Chen, K. Murphy and A. Yuille
Weakly- and Semi-Supervised Learning of a DCNN for Semantic Image Segmentation, http://arxiv.org/abs/1502.02734

Weaker Annotations: Bounding Boxes



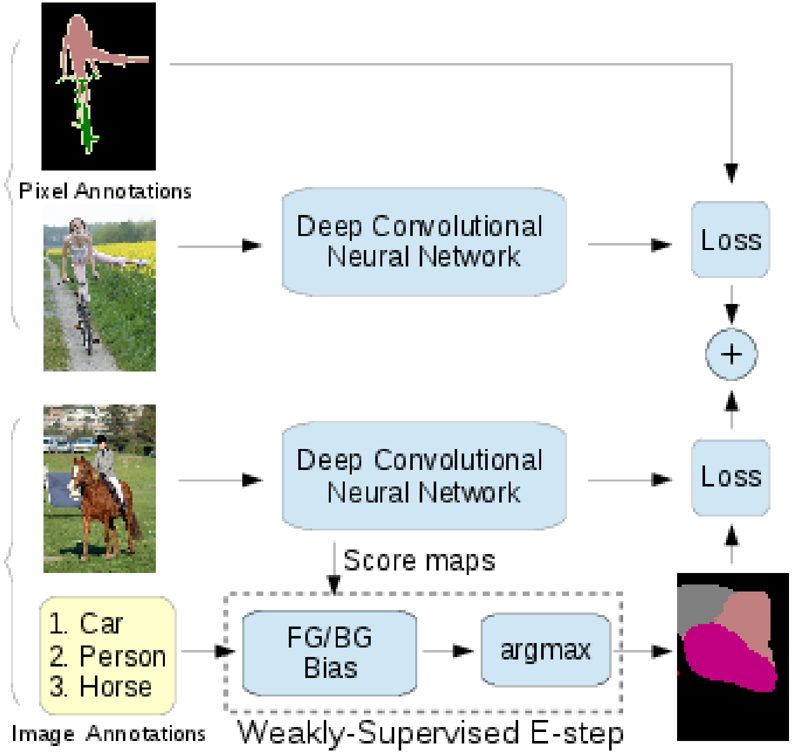
Grab-Cut as preprocessing

Weaker Annotations: Image Level



Expectation-Maximization during net training

Weaker Annotations: Hybrid Approach



Weak Annotation Pascal Results

Image
39%

BBox (Base/Grab)
54.2% / 60.4%

Hybrid (1.5K / 3K)
63.5% / 66.4%

Strong (10K)
66.4%