Learning to Segment Under Various Forms of Weak Supervision

THE UNIVERSITY of WISCONSIN MADISON

Jia Xu[1] Alexander G. Schwing[2] Raquel Urtasun[2]

[1] University of Wisconsin-Madison [2]University of Toronto

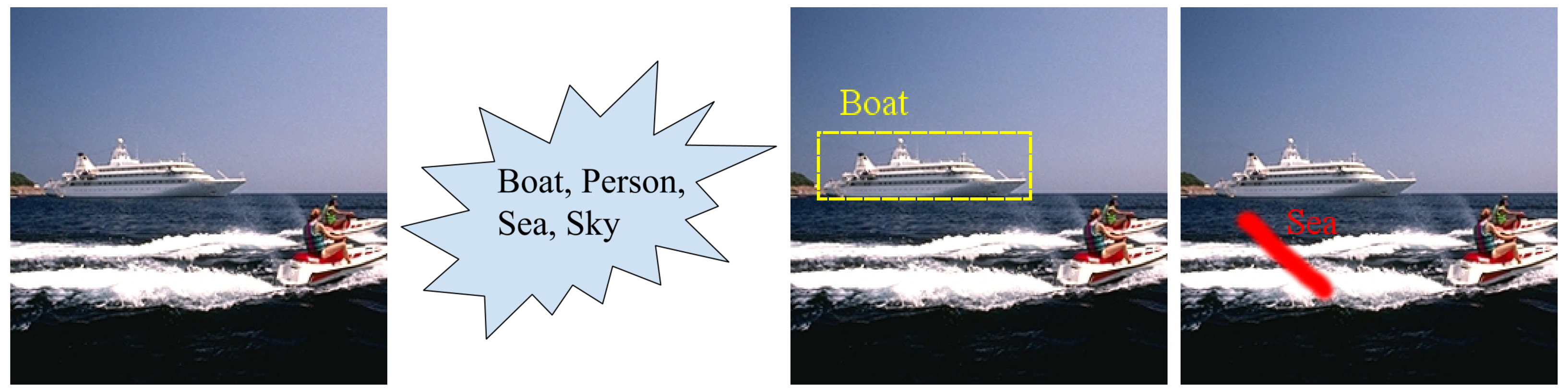
UNIVERSITY OF TORONTO

Motivation

Observations:

- Full annotations are expensive to collect
- Weak labelings are easy to obtain and available at larger scale
- Different algorithms have been developed for different forms of weak supervision

Unified pixel-wise semantic segmentation algorithm to learn from various forms of weak supervision like image level tags, bounding boxes and partial labels



Problem Formulation

Task: Segment m images into n super-pixels with C categories

Max-Margin Objective:

$$\min_{W,H} \frac{1}{2}\mathrm{tr}(W^TW) + \lambda \sum_{p=1}^{n}\sum_{c=1}^{C} \xi(\mathbf{w}_c; \mathbf{x}_p, h_p^c) \;\; \text{s.t.} \;\; H\mathbf{1}_C = \mathbf{1}_n, \;\; H \in \mathcal{S}$$

- Feature matrix $X = [\mathbf{x}_1^T, \mathbf{x}_p^T, \cdots, \mathbf{x}_n^T] \in \mathbb{R}^{n\times d}$
- Latent assignment matrix $H = [\mathbf{h}_1^T, \mathbf{h}_p^T, \cdots, \mathbf{h}_n^T] \in \{0,1\}^{n\times c}$
- Appearance model matrix $W \in \mathbb{R}^{d\times c}$
- Surrogate loss

$$\xi(\mathbf{w}_c; \mathbf{x}_p, h_p^c) = \begin{cases} \max(0, 1 + (\mathbf{w}_c^T\mathbf{x}_p)) & h_p^c = 0 \\ \mu^c \max(0, 1 - (\mathbf{w}_c^T\mathbf{x}_p)) & h_p^c = 1 \end{cases}$$

Inference: $h_p^{\hat{c}} = 1$ iif $\hat{c} \in \arg\max_c \mathbf{w}_c^T\mathbf{x}_p$

- Asymmetric loss

$$\mu^c = \frac{\sum_{p=1}^n 1(h_p^c == 0)}{\sum_{p=1}^n 1(h_p^c == 1)}$$

Penalizes according to ratio of negative vs. positive examples

Supervision as Constraints:

- Unlabeled $\mathcal{S} = \emptyset$
- Image level tags $\mathcal{S} = \{H \le BZ, B^TH \ge Z\}$
- Bounding boxes $\mathcal{S} = \{H \le \hat{B}\hat{Z}, \hat{B}^TH \ge \hat{Z}\}$
- Semi-supervised $\mathcal{S} = \{H_\Omega = \hat{H}_\Omega\}$

An Example: 2 images of 2 and 3 super-pixels, tagged with classes $\{1,2\}$ and $\{2,3\}$

$$B = \begin{bmatrix} 1 & 0 \\ 1 & 0 \\ 0 & 1 \\ 0 & 1 \\ 0 & 1 \end{bmatrix}, \quad H = \begin{bmatrix} 0 & 1 & 0 \\ 1 & 0 & 0 \\ 0 & 0 & 1 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}, \quad BZ = \begin{bmatrix} 1 & 1 & 0 \\ 1 & 1 & 0 \\ 0 & 1 & 1 \\ 0 & 1 & 1 \\ 0 & 1 & 1 \end{bmatrix}$$

$$Z = \begin{bmatrix} 1 & 1 & 0 \\ 0 & 1 & 1 \end{bmatrix}, \quad B^TH = \begin{bmatrix} 1 & 1 & 0 \\ 0 & 1 & 2 \end{bmatrix}$$

Unified Segmentation Algorithm

Model:

$$\min_{W,H} \frac{1}{2}\mathrm{tr}(W^TW) + \lambda \sum_{p=1}^{n} \xi(W; \mathbf{x}_p, \mathbf{h}_p) \;\; \text{s.t.} \;\; \begin{matrix} H\mathbf{1}_C = \mathbf{1}_n, H \in \{0,1\}^{n\times C} \\ H \in \mathcal{S} \end{matrix}$$

Challenges: non-convex mixed integer programming

Observations:

- Optimization problem is bi-convex, i.e., it is convex w.r.t. W if H is fixed, and convex w.r.t. H if W is fixed
- Constraints are linear and they only involve the super-pixel assignment matrix H

Learning to segment by alternating optimization:

1. Fix H solve for W independently for all classes (1-vs-all linear SVM)
2. Fix W infer super-pixel labels H in parallel for all images (small LP instances, inference task)

Inference:

$$\max_{H} \;\; \mathrm{tr}((XW)^TH) \;\; \text{s.t.} \;\; H\mathbf{1}_C = \mathbf{1}_n, \;\; H \in \{0,1\}^{n\times C}, \;\; H \in \mathcal{S}$$

Proposition: The inference task (integer linear program) can be solved to global optimality in polynomial time

Reason: Constraint matrix is totally unimodular

Model Nature:

- Decomposable and easily parallelizable
- Theoretical guarantee of relaxation quality

Computation Efficiency:

- Orders of magnitude faster than the state-of-the-art for training 20 min vs. 24 hours
- 10 ms for inference on one image

Results on SIFT-Flow

Method	Supervision	Per-class	Per-pixel
Liu et al. 2011 (PAMI)	full	24	76.7
Farabet et al. 2012 (ICML)	full	29.5	78.5
Farabet et al. 2012 (ICML) balanced	full	46.0	74.2
Eigen et al. 2012 (CVPR)	full	32.5	77.1
Tighe et al. 2014 (CVPR)	full	39.3	78.6
Yang et al. 2014 (CVPR)	full	48.7	79.8
Vezhnevets et al. 2011 (ICCV)	weak (tags)	14	N/A
Vezhnevets et al. 2012 (CVPR)	weak (tags)	22	51
Xu et al. 2014 (CVPR)	weak (tags)	27.9	N/A
Ours (1-vs-all)	weak (tags)	**32.0**	**64.4**
Ours (ILT)	weak (tags)	**35.0**	**65.0**
Ours (1-vs-all + transductive)	weak (tags)	**40.0**	**59.0**
Ours (ILT + transductive)	weak (tags)	**41.4**	**62.7**

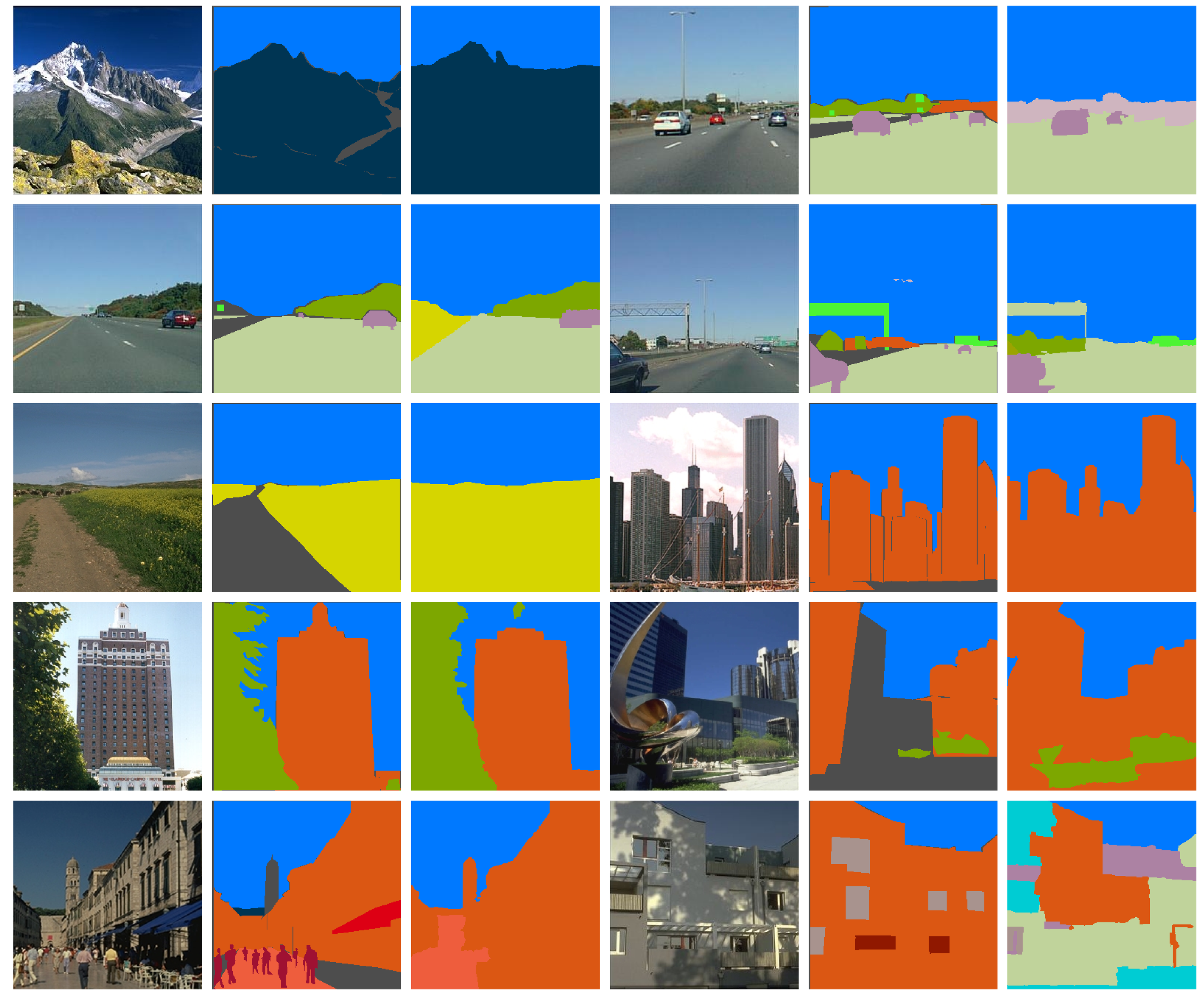
Results

Comparison to state-of-the-art on MSRC:

Method	Supervision	per-class	per-pixel
Shotton et al. 2008 (ECCV)	full	67	72
Yao et al. 2012 (CVPR)	full	79	86
Vezhnevets et al. 2011 (ICCV)	weak (tags)	67	67
Liu et al. 2012 (TMM)	weak (tags)	N/A	**71**
Ours	weak (tags)	**73**	70

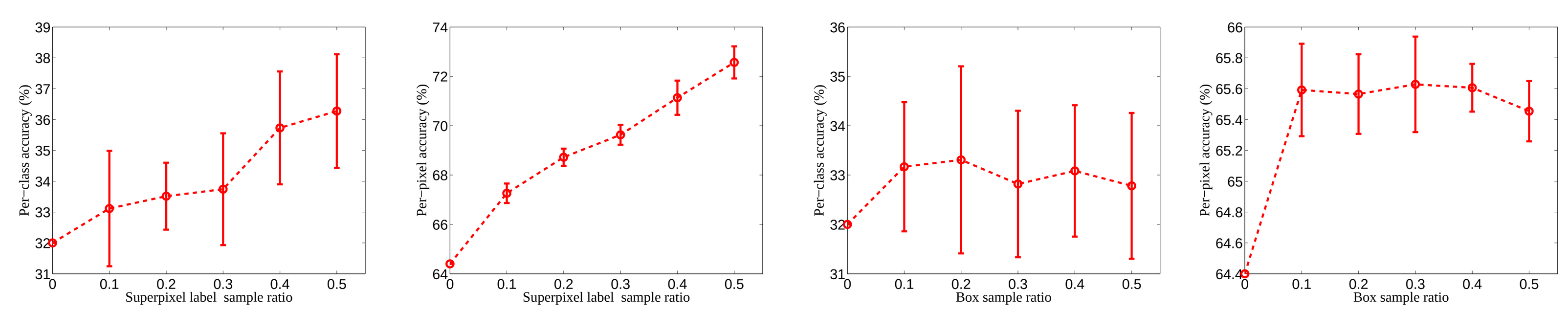
Visual results:

Input Truth Ours Input Truth Ours

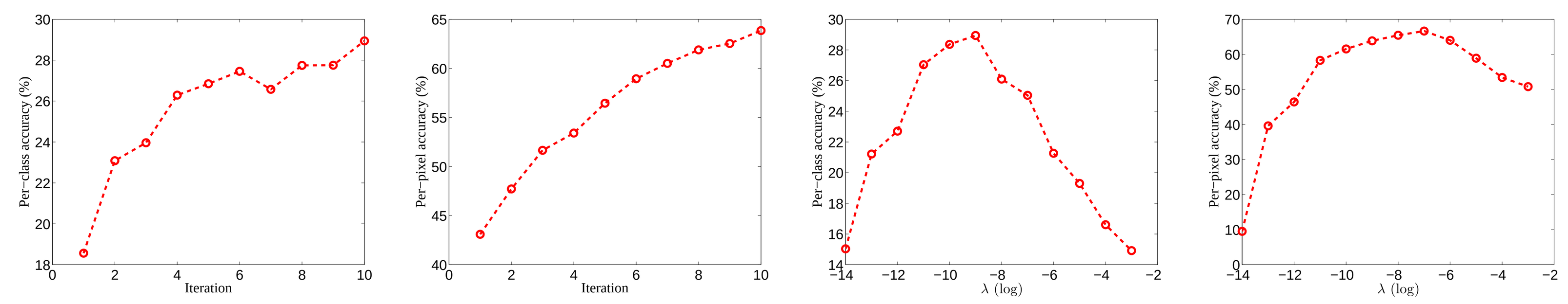


unlabeled sky mountain road tree car sign person field building

Other forms of weak supervision:



Model behavior:



Summary

- A unified model to learn semantic segmentation under various forms of weak supervision
- An efficient algorithm achieving state-of-the-art results