Weakly Supervised CNN Segmentation: Models and Optimization

Ismail Ben Ayed
Christian Desrosiers
Jose Dolz
Hoel Kervadec

ÉTS
Le génie pour l'industrie

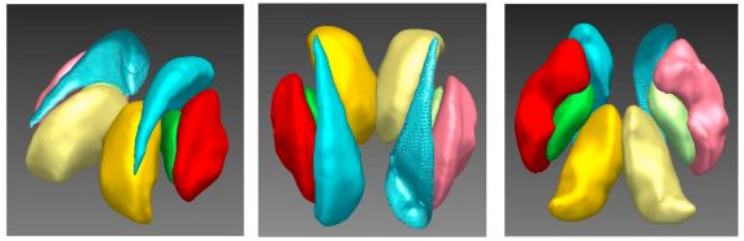
Why are we doing this tutorial at MICCAI?

Deep CNNs are dominating computer vision

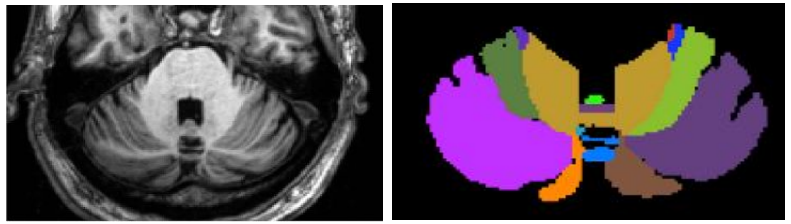
e.g., semantic segmentation

Cordts et al., “The Cityscapes Dataset for Semantic Urban Scene Understanding,” CVPR 2016

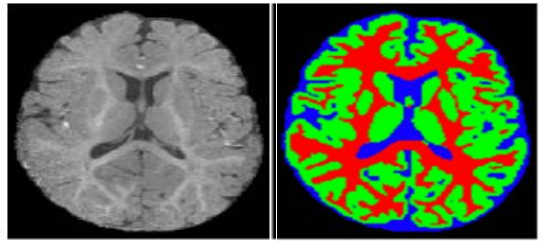
… and medical image analysis



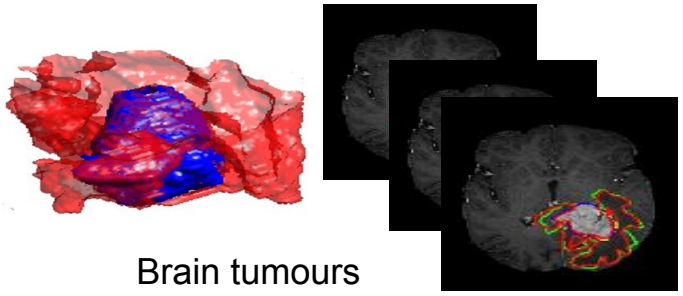
Subcortical structures
(Dolz et al., Neuroimage 2018)



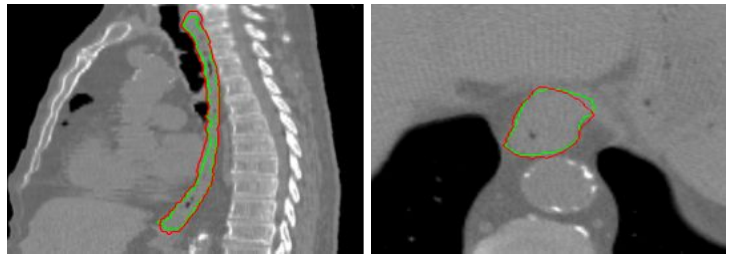
Cerebellum parcellation
(Carass et al., Neuroimage 2018)



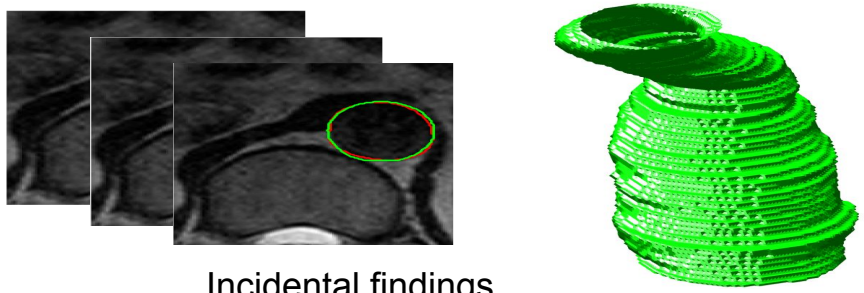
Brain tissues (6-month infant)
(Li et al., TMI 2019)



Brain tumours
(Njeh et al., CMIG 2015)



Organs at risk
(Dolz et al., Med. Phys. 2017)



Incidental findings
(Ben Ayed et al., MICCAI 2014)

But, massive and dense annotations are not always available

Full supervision

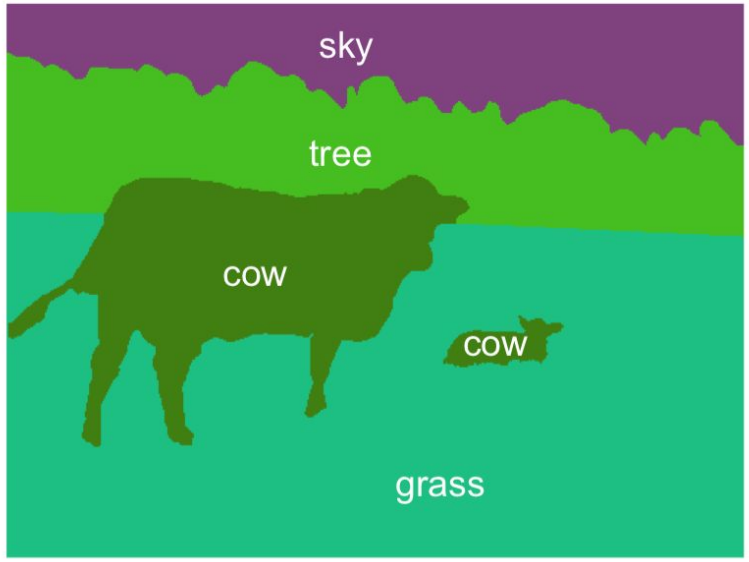
- more than 1h per image (even several hours for a medical image)
- Bottleneck for learning at large scale

Weak supervision (e.g., image-level tags)

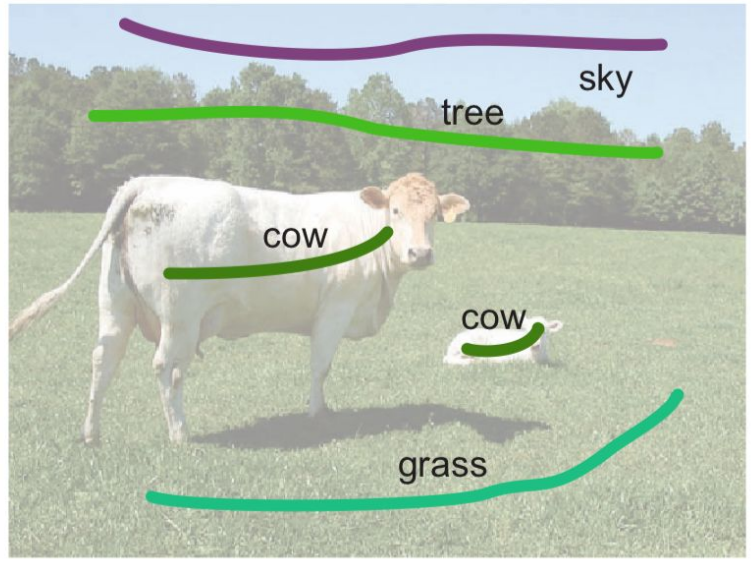
- 1s per label per image
- Scalable for large numbers of labels

person
horse
background

Semi-supervision with a lot of **non-annotated** data, and a **fraction** of points annotated



Full annotations



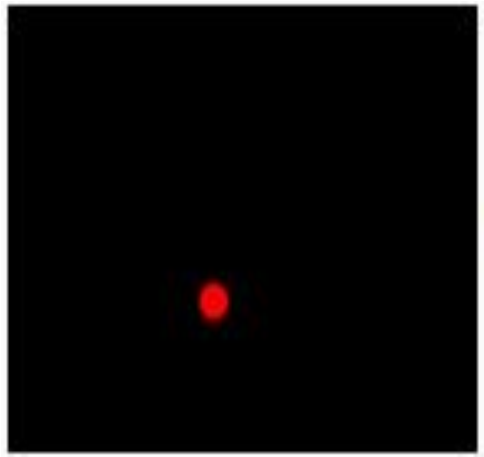
Semi-supervised

Figures from Lin et al. Scribblesup: Scribble-supervised convolutional networks for semantic segmentation, CVPR 2016

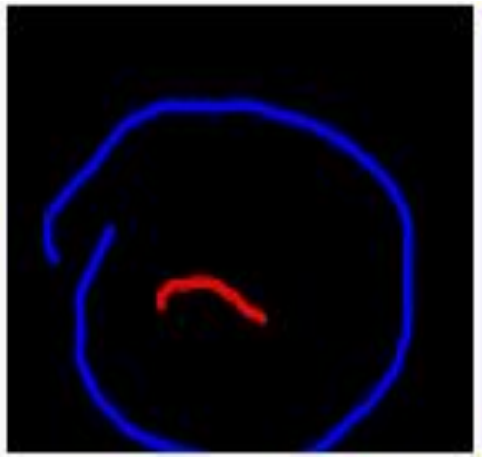
Forms of semi/weak supervision: Examples in segmentation

Car
Parking
Sky
No person

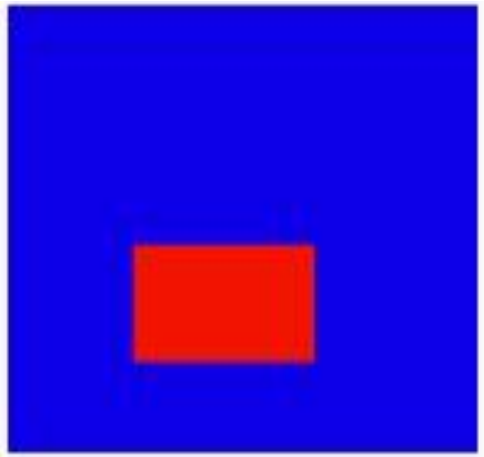
Image tags



points



scribbles

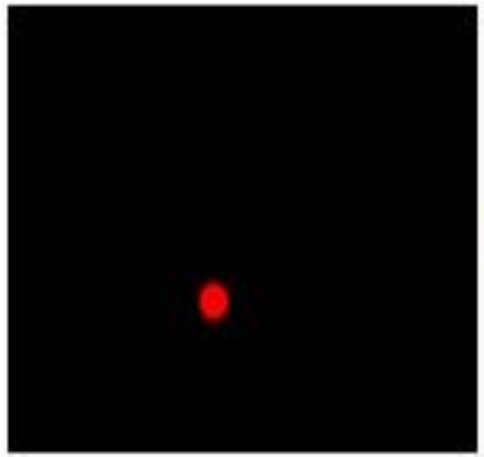


boxes

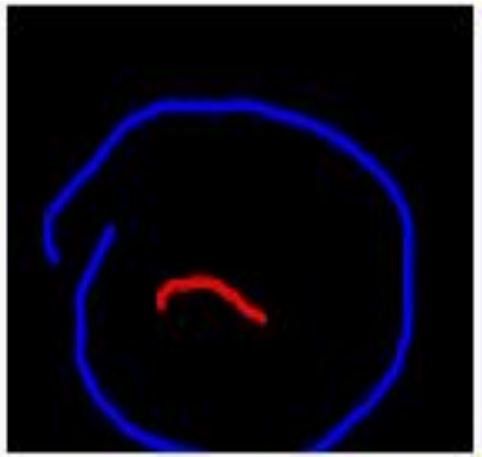
Forms of semi/weak supervision: Examples in segmentation

Car
Parking
Sky
No person

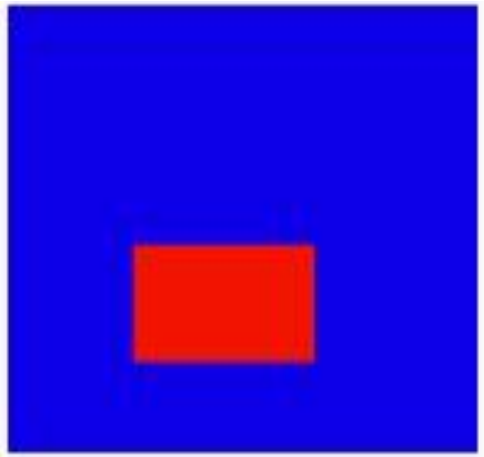
Image tags



points



scribbles



boxes

[Marin et al., CVPR 2019], [Tang et al., ECCV 2018],
[Lin et al., CVPR 2016], [Khoreva et al. CVPR 2017],
[Vernaza et al., CVPR 2017], [Kolesnikov and Lampert, ECCV 2016]
[Dai et al., CVPR 2015], [Bearman et al., ECCV 2016]
[Pathak et al., ICCV 2015], [Papandreou et al., ICCV 2015]

[Rajchl et al., TMI 2017]
[Bai et al., MICCAI 2017]
[Kervadec et al., MedIA]

Full annotations are much more problematic in medical imaging

Not anywhere close to the 10k images of Pascal VOC and the 5k of Cityscapes

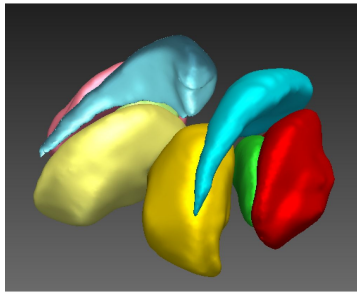
Crowdsourcing?

Full annotations are much more problematic in medical imaging

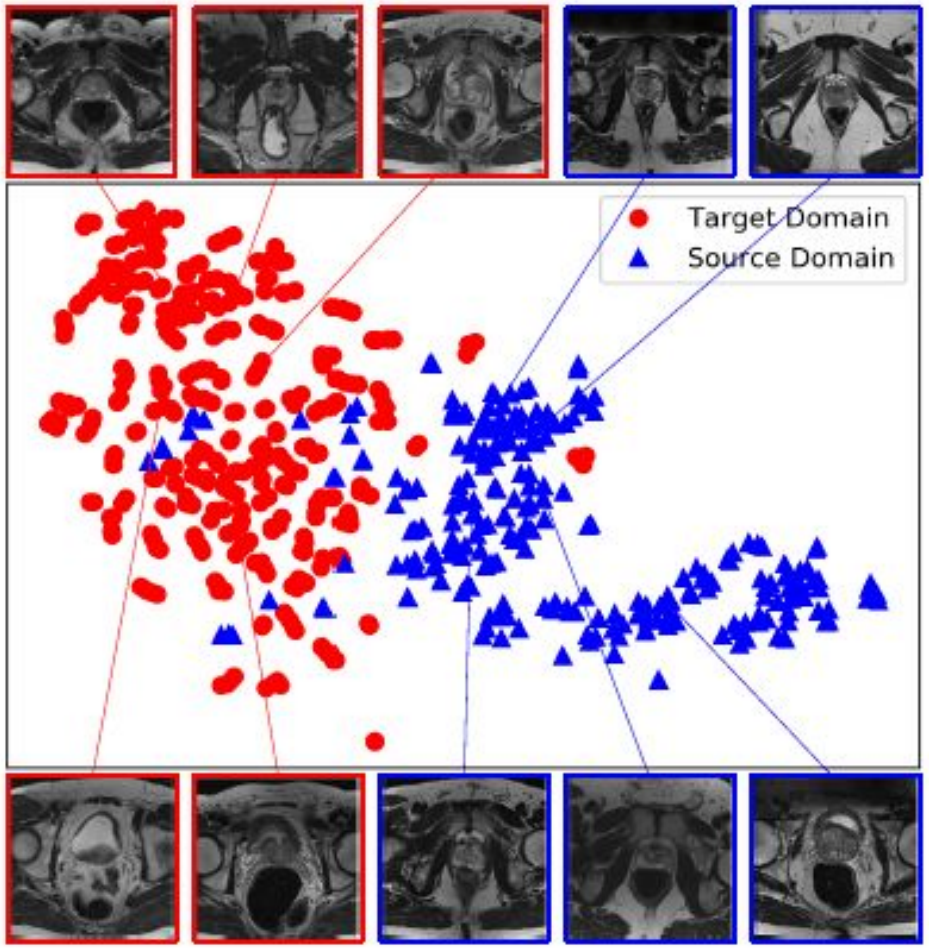
Not anywhere close to the 10k images of Pascal VOC and the 5k of Cityskapes

Crowdsourcing?

Dense 3D annotations: several hours
(of radiologist time)

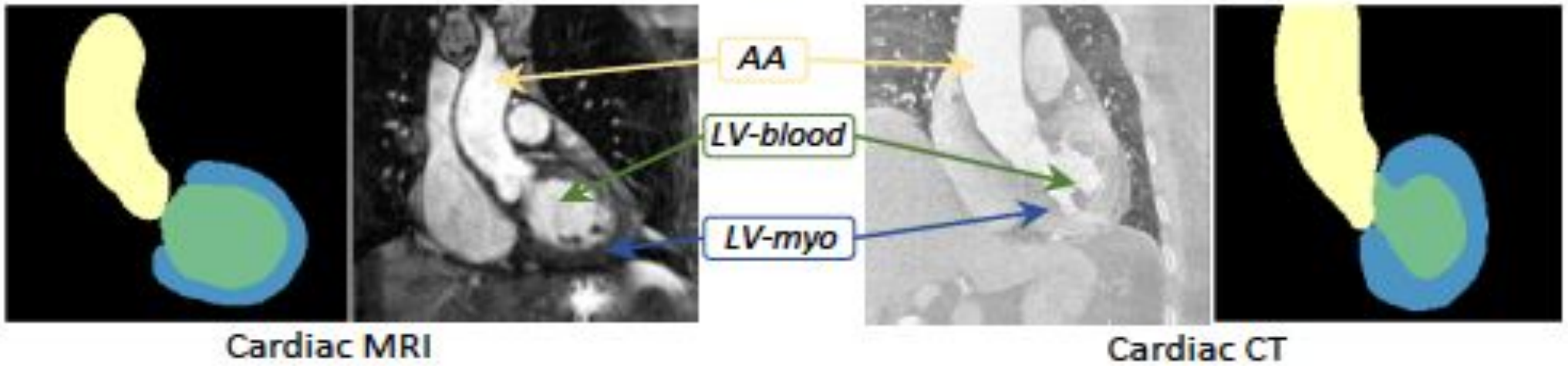


Domain shifts make things worse (even with full annotations in one domain)



[MRI Prostate segmentation: Figure from Zhu et al., Boundary-weighted Domain Adaptive Neural Network for Prostate MR Image Segmentation ArXiv 2019]

Domain shifts: within and across modalities

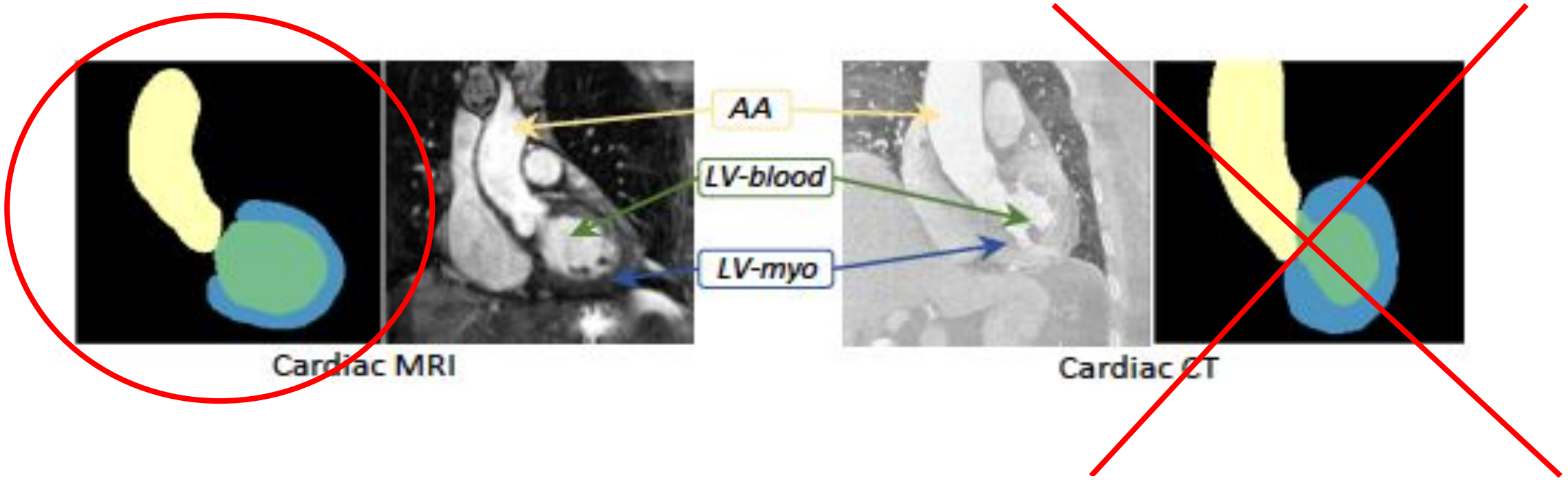


[Images from Dou et al., PnP-AdaNet: Plug-and-play adversarial domain adaptation network with a benchmark at cross-modality cardiac segmentation ArXiv 2018]

Unsupervised domain adaptation

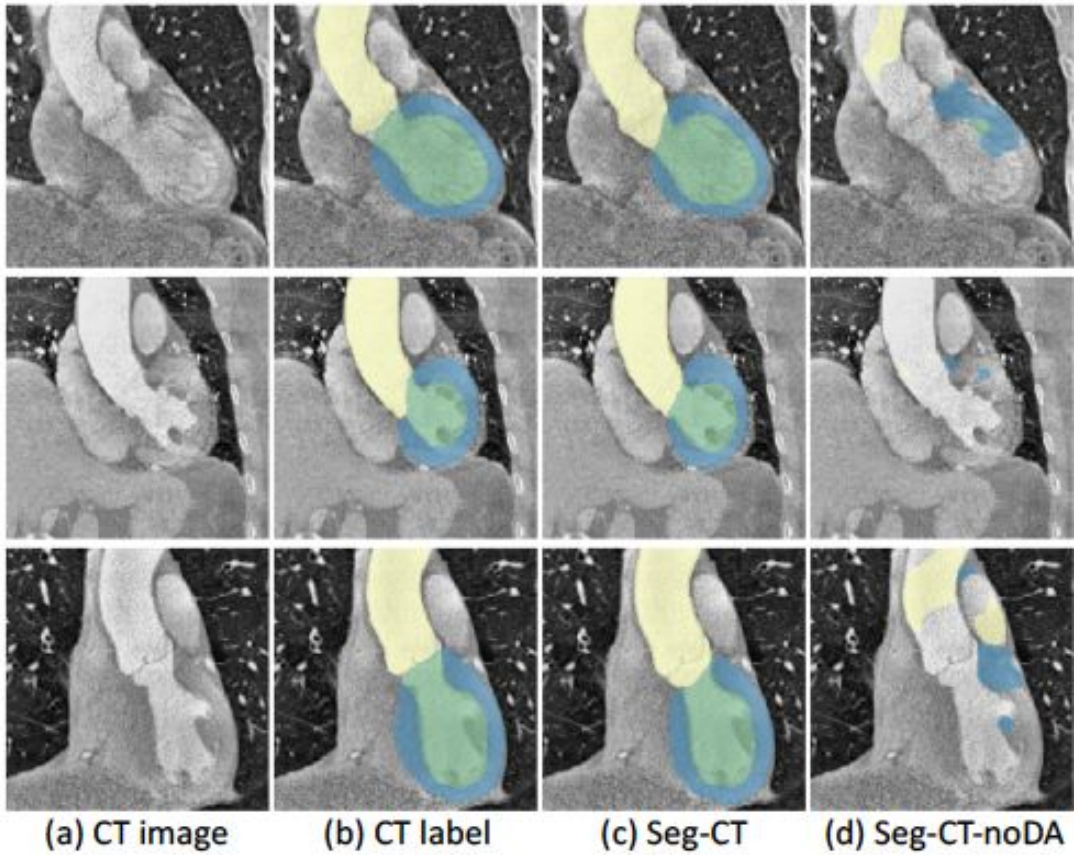
We have labels for the source domain

No labels for the target



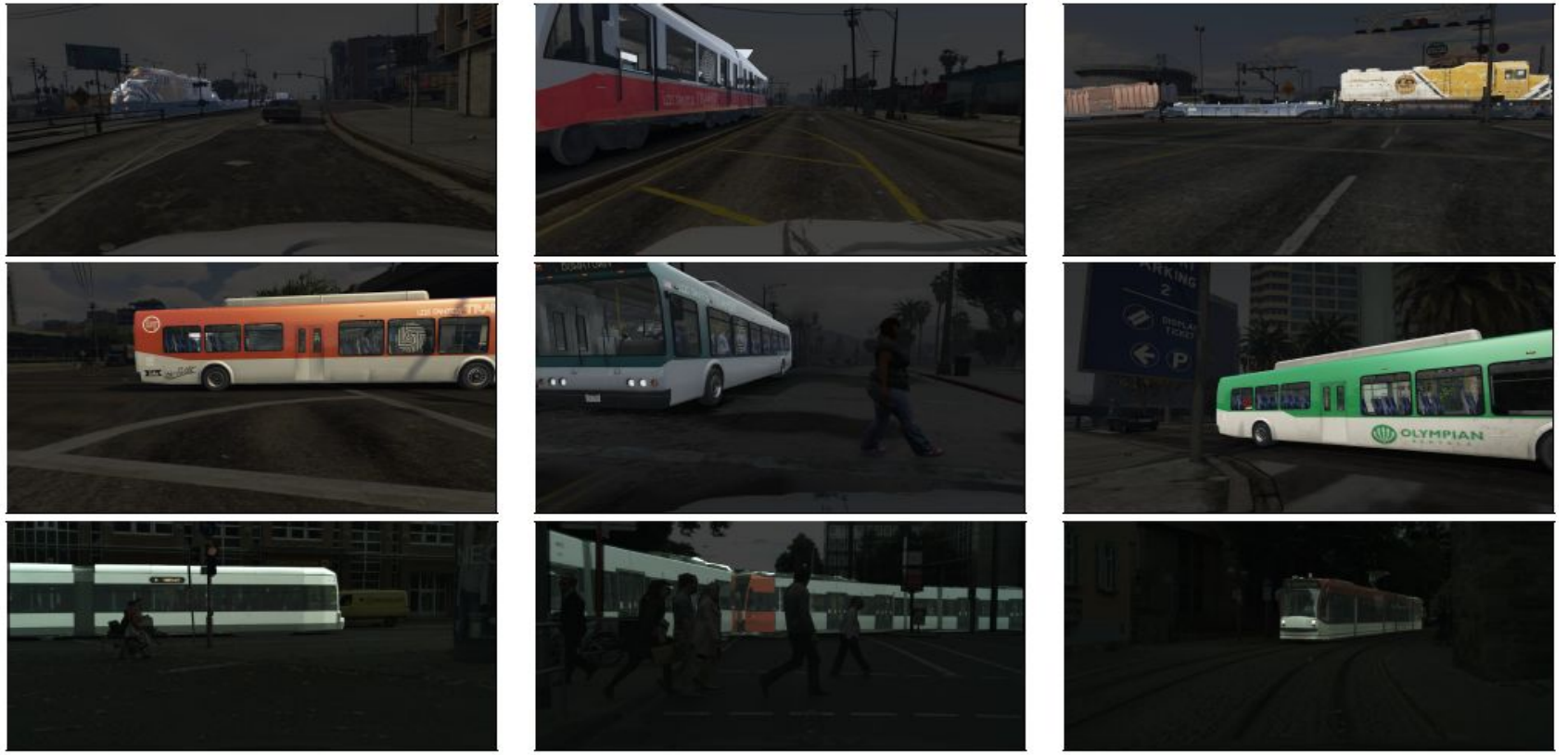
[Images from Dou et al., PnP-AdaNet: Plug-and-play adversarial domain adaptation network with a benchmark at cross-modality cardiac segmentation ArXiv 2018]

Bad generalization to the target



[Images from Dou et al., PnP-AdaNet: Plug-and-play adversarial domain adaptation network with a benchmark at cross-modality cardiac segmentation ArXiv 2018]

A lot of interest in vision as well:
Domain shifts are *everywhere* **BUT** we cannot label *everywhere*



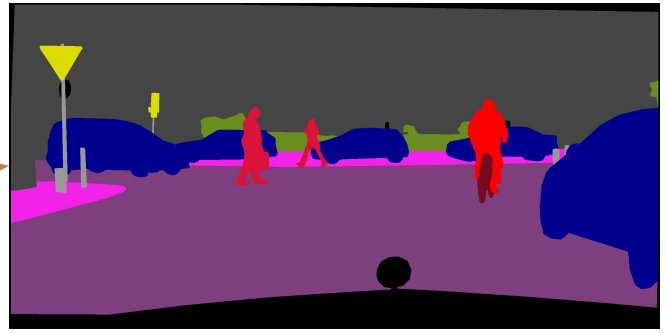
"train"
GTA

"bus"
GTA

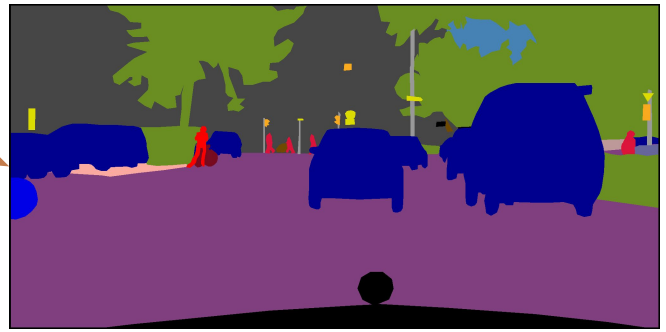
"train"
Cityscapes

Figures from [Zhang et al., A Curriculum Domain Adaptation Approach to the Semantic Segmentation of Urban Scenes TPAMI 2019]

A lot of interest in vision as well:
Domain shifts are *everywhere* **BUT** we cannot label *everywhere*



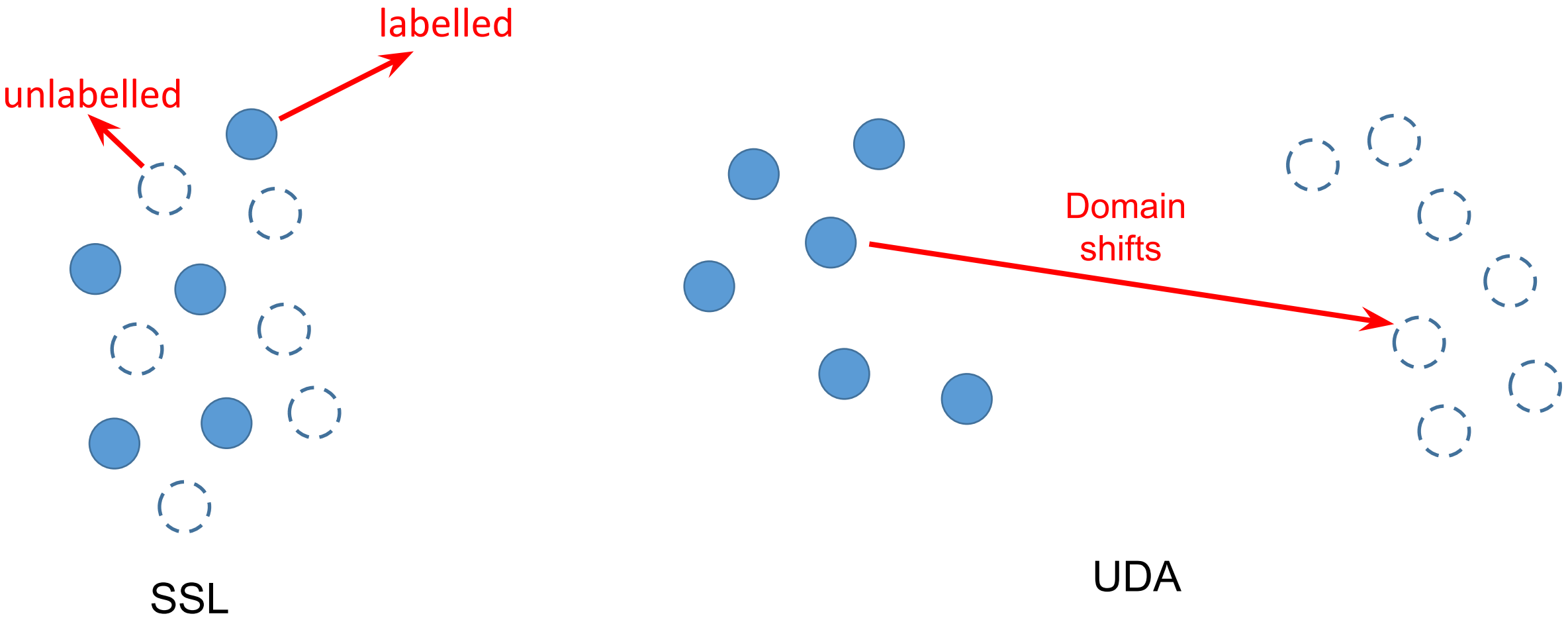
Frankfurt



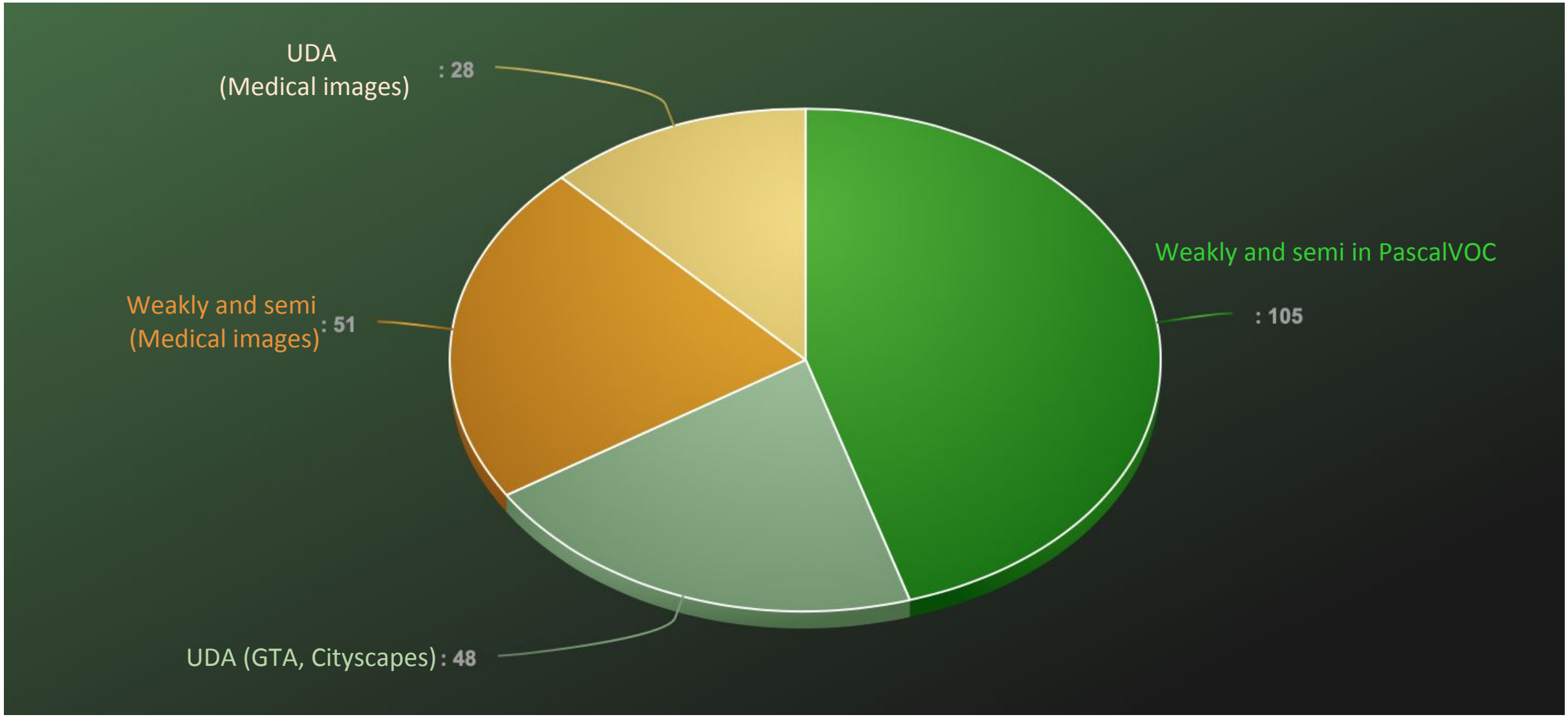
Zurich

Cityscapes (5000 images): labeling of 1 image takes 90 min at average [Cordt et al., CVPR 2016]

UDA = SSL + domain shift



Surprisingly in medical image analysis, we are behind



Semi/weak supervision in a nutshell:
We are leveraging **unlabelled** data with **priors**

- Structure-driven priors: *Regularization (Part 1)*

✔ Models and optimization

- Knowledge-driven priors (e.g., anatomy): *Constraints (Part 2)*

✔ Models and optimization

- Data-driven priors: *Adversarial learning (Part 3)*

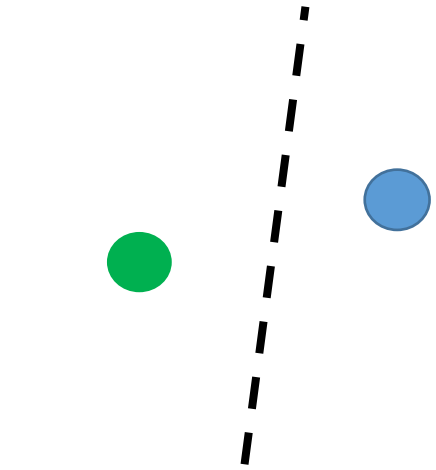
✔ Models and optimization

Part 1

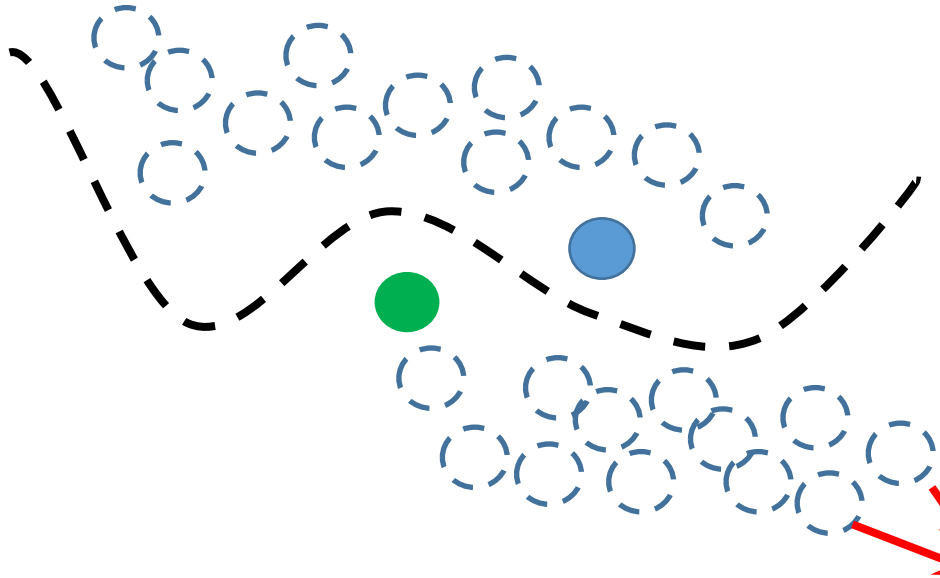
Regularization

Laplacian (and CRFs)

Semi-supervised learning (general form)

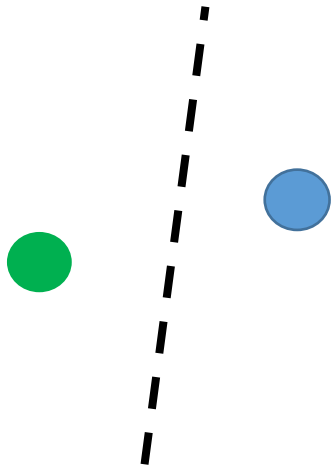


Only labeled points

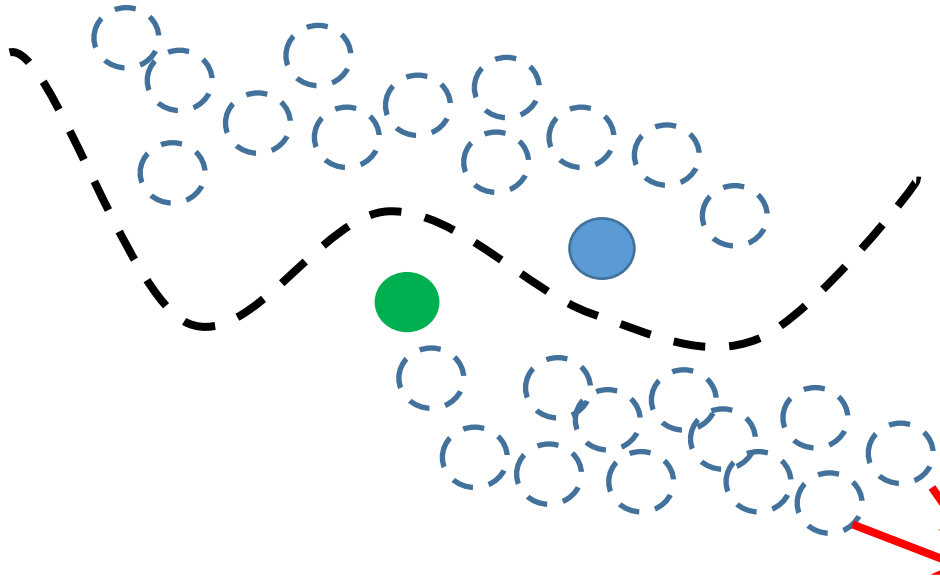


Unlabeled points

Semi-supervised learning (general form)



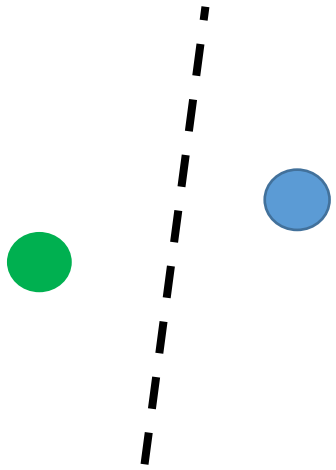
Only labeled points



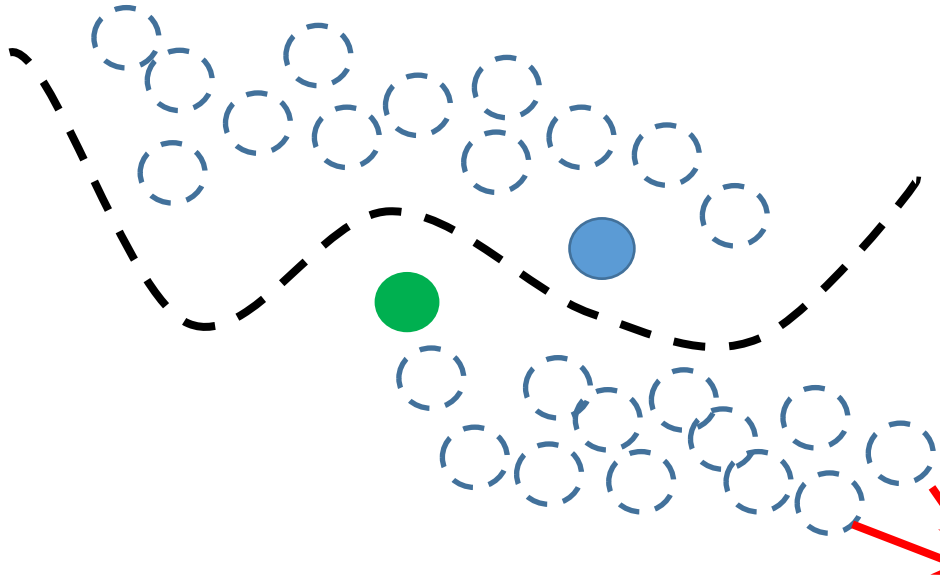
Unlabeled points

$$\sum_{p \in \mathcal{L}} l(\mathbf{f}_\theta^p, \mathbf{y}^p) + \sum_{p,q \in \mathcal{L} \cup \mathcal{U}} g(\mathbf{f}_\theta^p, \mathbf{f}_\theta^q, w_{p,q})$$

Semi-supervised learning (general form)



Only labeled points



Unlabeled points

$$\sum_{p \in \mathcal{L}} l(\mathbf{f}_\theta^p, \mathbf{y}^p) + \sum_{p,q \in \mathcal{L} \cup \mathcal{U}} g(\mathbf{f}_\theta^p, \mathbf{f}_\theta^q, w_{p,q})$$

Labeled points

Unlabeled points

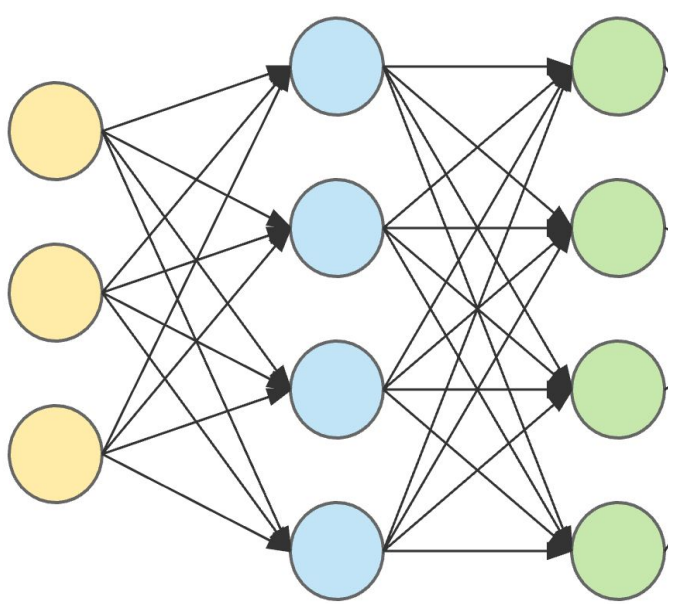
Semi-supervised learning (general form)

$$\sum_{p \in \mathcal{L}} l(\mathbf{f}_\theta^p, \mathbf{y}^p) + \sum_{p,q \in \mathcal{L} \cup \mathcal{U}} g(\mathbf{f}_\theta^p, \mathbf{f}_\theta^q, w_{p,q})$$

Labels
(binary simplex vectors)

e.g.: cross-entropy

e.g.: simplex probability vectors
(*softmax outputs of the network*)



$$f_\theta^{p,1}$$
$$f_\theta^{p,2}$$
$$\vdots$$
$$f_\theta^{p,C}$$

$$\mathbf{f}_\theta^p = \mathbf{s}_\theta^p \in [0,1]^K$$

Semi-supervised learning (general form)

$$\sum_{p \in \mathcal{L}} l(\mathbf{f}_\theta^p, \mathbf{y}^p) + \sum_{p,q \in \mathcal{L} \cup \mathcal{U}} g(\mathbf{f}_\theta^p, \mathbf{f}_\theta^q, w_{p,q})$$

e.g.: cross-entropy

Labels
(binary simplex vectors)

e.g.: **Laplacian**

$$w_{p,q} \| \mathbf{f}_\theta^p - \mathbf{f}_\theta^q \|^2$$

e.g.: simplex probability vectors
(*softmax outputs of the network)*

Semi-supervised learning (general form)

$$\sum_{p \in \mathcal{L}} l(\mathbf{f}_\theta^p, \mathbf{y}^p) + \sum_{p,q \in \mathcal{L} \cup \mathcal{U}} g(\mathbf{f}_\theta^p, \mathbf{f}_\theta^q, w_{p,q})$$

e.g.: cross-entropy

Labels
(binary simplex vectors)

e.g.: **Laplacian**

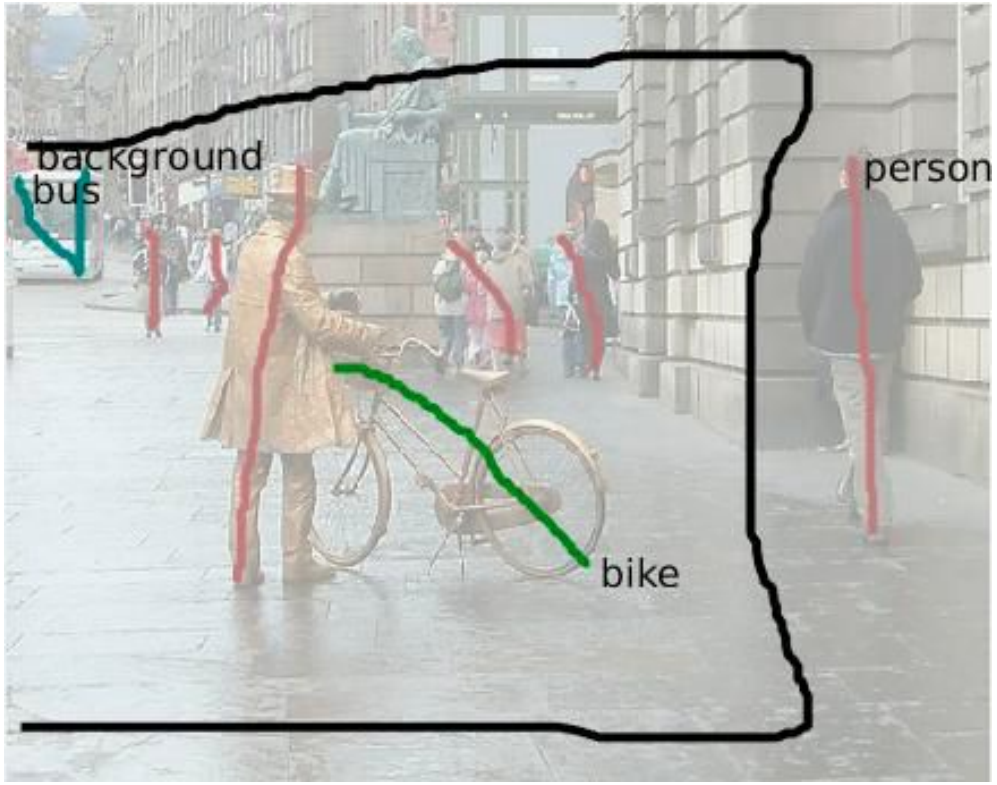
$$w_{p,q} \| \mathbf{f}_\theta^p - \mathbf{f}_\theta^q \|^2$$

e.g.: simplex probability vectors
(*softmax outputs of the network*)

- [Weston et al., Deep Learning via semi-supervised embedding, ICML 2008]
- [Belkin et al., Manifold regularization: a geometric framework for learning from Labeled and Unlabeled Examples, JMLR 2006]
- [Zhu et al., Semi-Supervised Learning Using Gaussian Fields and Harmonic Functions, ICML 2003]

Semi-supervision loss for segmentation

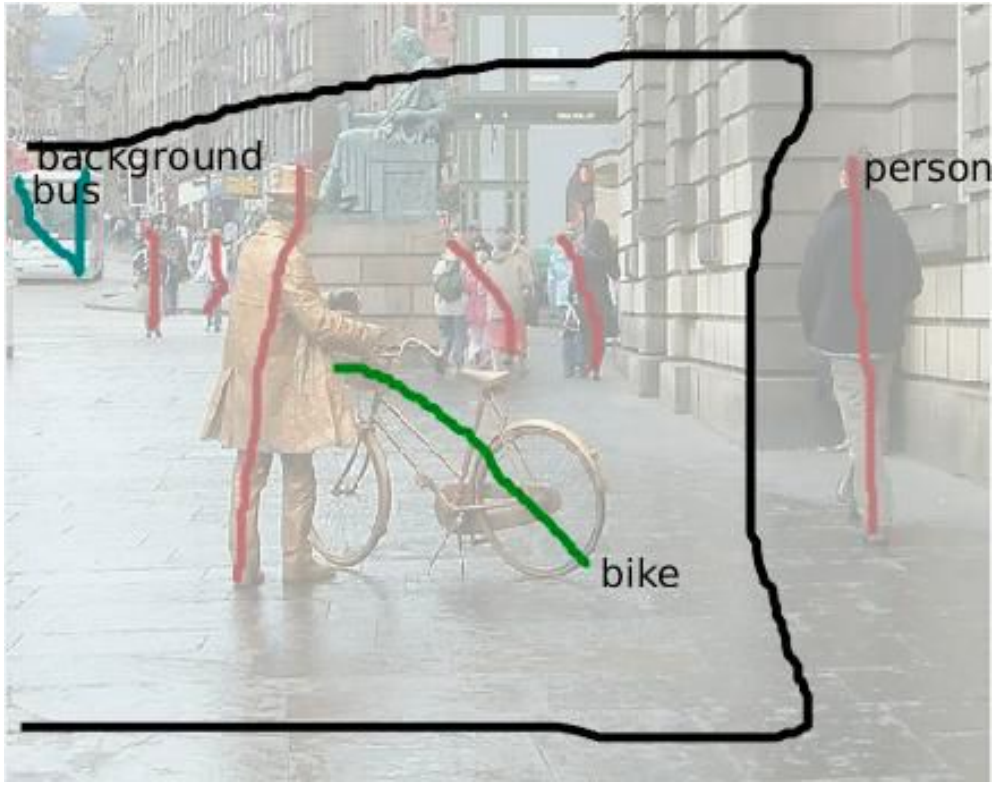
$$\min_{\theta} \sum_{p \in \mathcal{L}} l(\mathbf{y}^p, \mathbf{s}_\theta^p) + \sum_{p,q \in \mathcal{L} \cup \mathcal{U}} w_{p,q} \|\mathbf{s}_\theta^p - \mathbf{s}_\theta^q\|^2$$



[Tang et al., On regularized losses for weakly supervised segmentation, ECCV 2018]

Semi-supervision loss for segmentation

$$\min_{\theta} \sum_{p \in \mathcal{L}} l(\mathbf{y}^p, \mathbf{s}_\theta^p) + \sum_{p,q \in \mathcal{L} \cup \mathcal{U}} w_{p,q} \|\mathbf{s}_\theta^p - \mathbf{s}_\theta^q\|^2$$



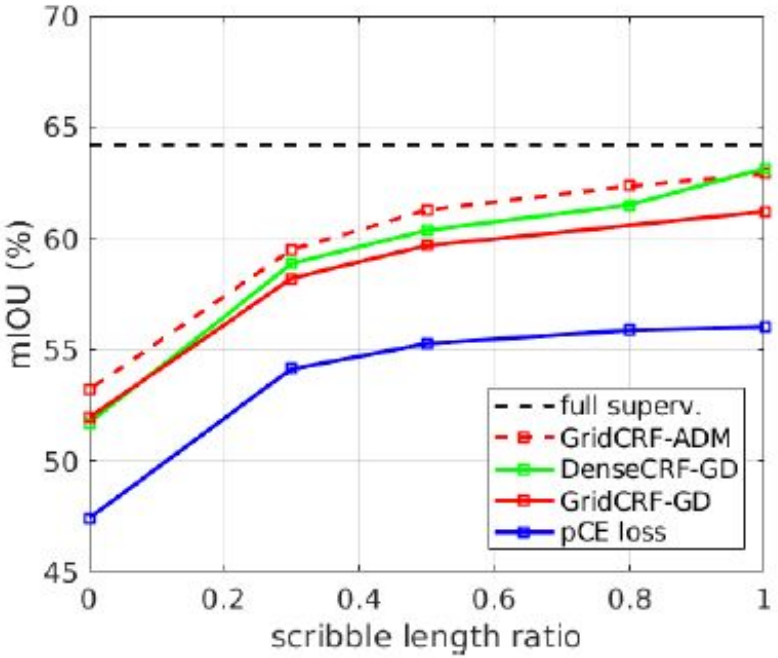
On the vertices of the simplex (binary variables), this is exactly the Potts model in Conditional Random Fields (e.g., Dense CRFs)!

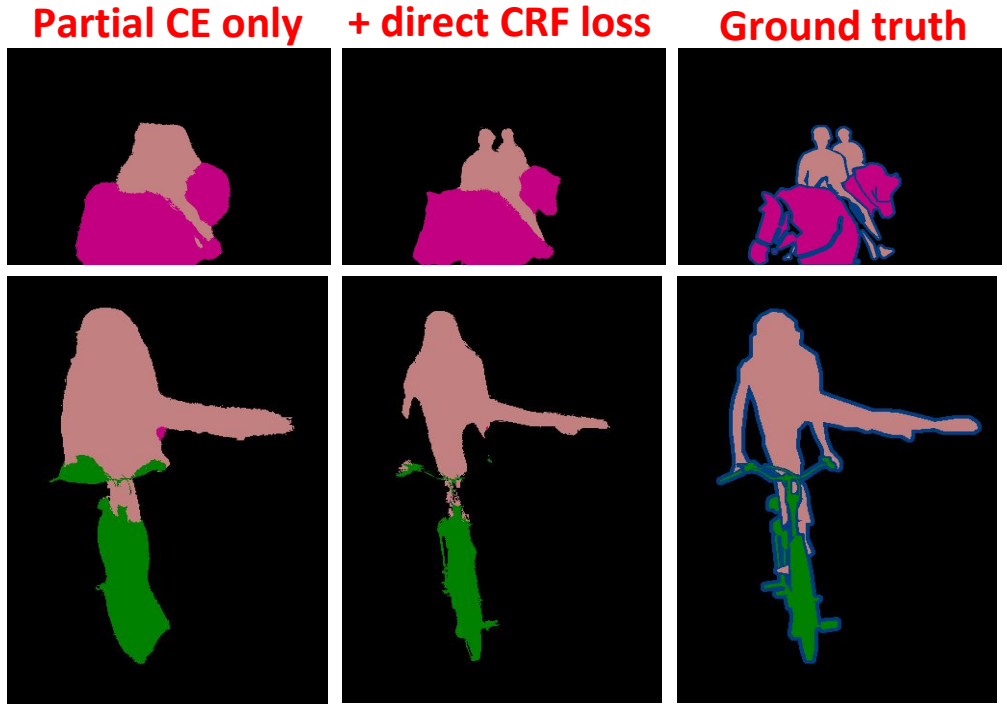
[Tang et al., On regularized losses for weakly supervised segmentation, ECCV 2018]

Semi-supervision loss for segmentation

$$\min_{\theta} \sum_{p \in \mathcal{L}} l(\mathbf{y}^p, \mathbf{s}_\theta^p) + \sum_{p,q \in \mathcal{L} \cup \mathcal{U}} w_{p,q} \|\mathbf{s}_\theta^p - \mathbf{s}_\theta^q\|^2$$

SGD

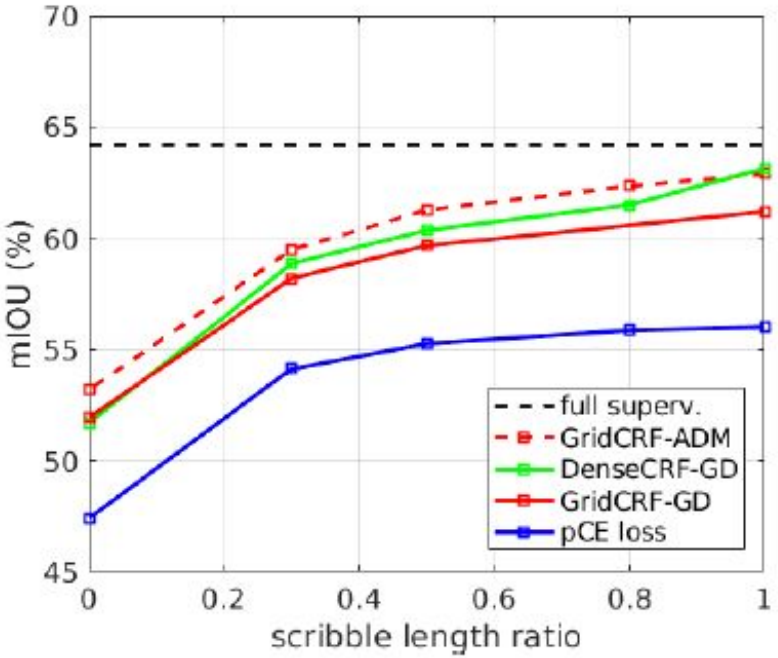




[Tang et al., On regularized losses for weakly supervised segmentation, ECCV 2018]
[Marin et al., Beyond gradient descent for regularized segmentation losses, CVPR 2019]

Semi-supervision loss for segmentation

$$\min_{\theta} \sum_{p \in \mathcal{L}} l(\mathbf{y}^p, \mathbf{s}_\theta^p) + \sum_{p,q \in \mathcal{L} \cup \mathcal{U}} w_{p,q} \|\mathbf{s}_\theta^p - \mathbf{s}_\theta^q\|^2$$



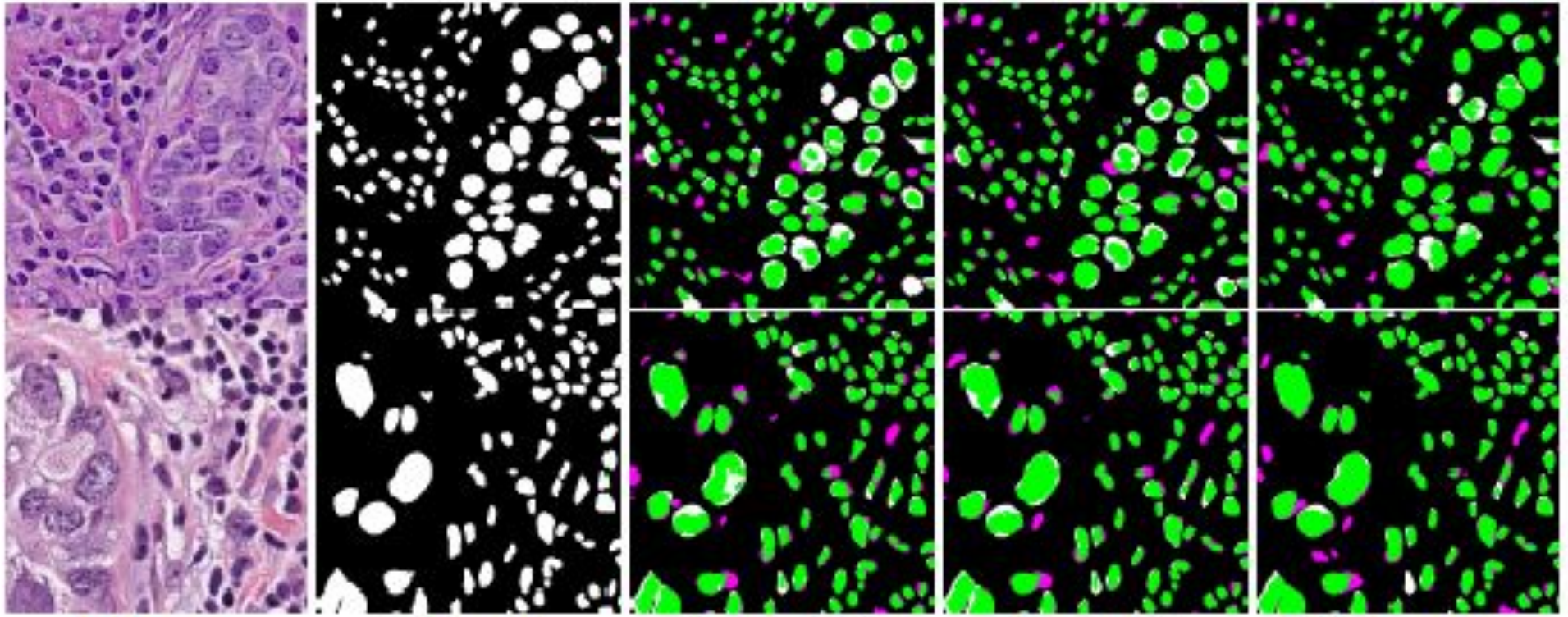
The exciting part in this plot:

Dense CRF with SGD gets you **97.6%** of full supervision performance with **3%** of the labels!

[Tang et al., On regularized losses for weakly supervised segmentation, ECCV 2018]
[Marin et al., Beyond gradient descent for regularized segmentation losses, CVPR 2019]

Some applications of CRF loss in MICCAI

White (FN); Magenta (FP); Green (TP)



- Figures from Qu et al., Weakly Supervised Deep Nuclei Segmentation using Points Annotation in Histopathology Images, MIDL 2019 [Histology, point annotation]

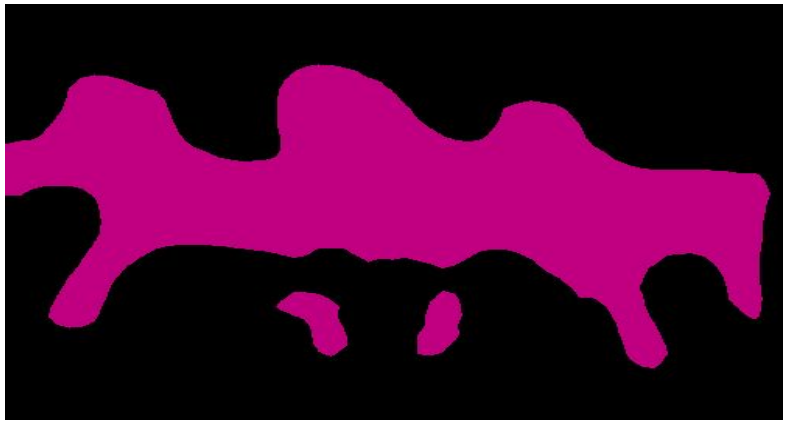
- Ji et al., Scribble-Based Hierarchical Weakly Supervised Learning for Brain Tumor Segmentation, MICCAI 2019 [Brain tumor images, scribble annotations]

Regularization

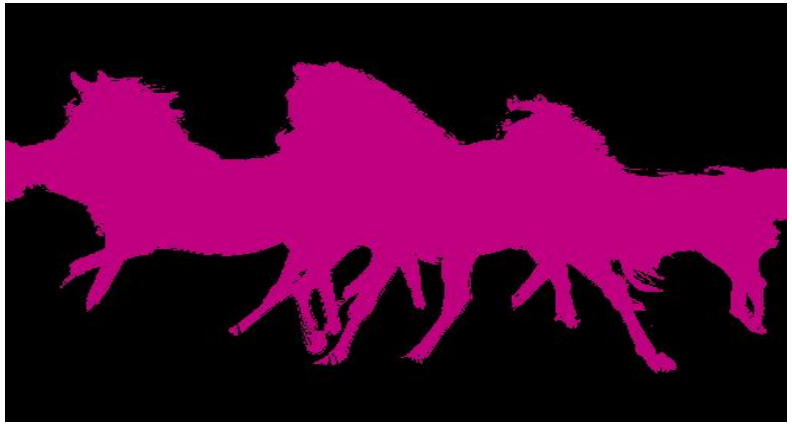
Optimization matters and the choice of an optimizer depends on the form of your loss (SGD is not not your only choice)

Conditional Random Fields (CRFs) is a form of Laplacian Regularization!

You probably know DeepLab:
DeepLab = supervised CNN + Dense CRF



CNN
(fully supervised)

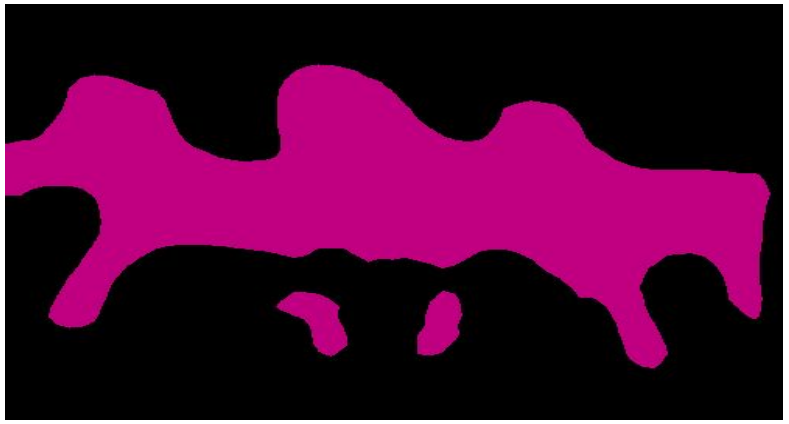


CNN
+
CRF (post-processing)

[Chen et al., Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected CRFs. TPAMI 2018]

CRFs meet fully supervised CNNS

$\mathbf{x}^p \in \mathbb{R}^N$ (Image colors)



$\mathbf{s}_\theta^p \in [0,1]^L$ (Network outputs)

$\mathbf{y}^p \in \{0,1\}^L$ (Binary labels)

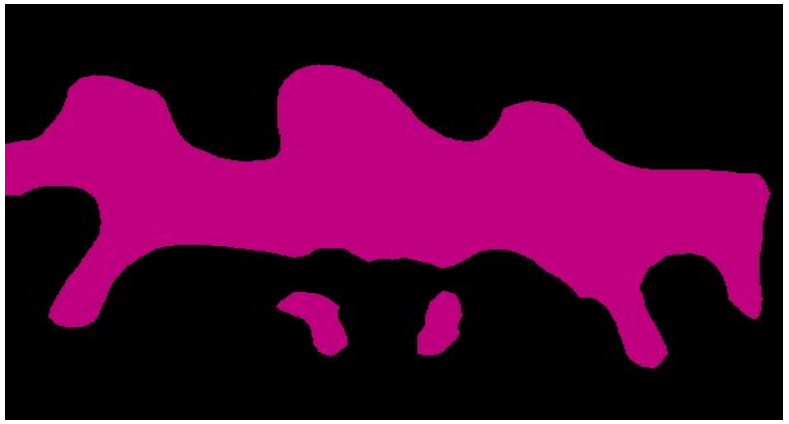
[Chen et al., Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected CRFs. TPAMI 2018]

CRFs meet fully supervised CNNS

Called **unary potentials** in discrete optimization
(the problem is trivial)

$$\min_{Y=[\mathbf{y}^1,\ldots\mathbf{y}^{|\Omega|}]} \sum_{p\in\Omega} l(\mathbf{y}^p, \mathbf{s}_\theta^p)$$

$\mathbf{x}^p \in \mathbb{R}^N$ (Image colors)



$\mathbf{s}_\theta^p \in [0,1]^L$ (Network outputs)

$\mathbf{y}^p \in \{0,1\}^L$ (Binary labels)

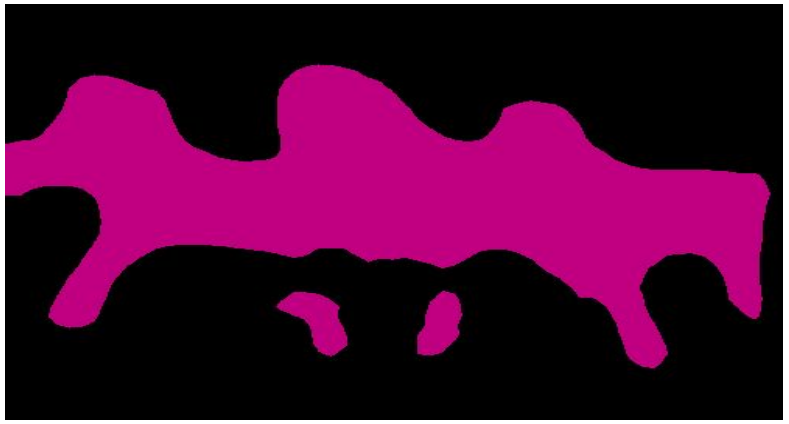
[Chen et al., Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected CRFs. TPAMI 2018]

CRFs meet fully supervised CNNS

Pairwise potentials (the very popular *Potts*)

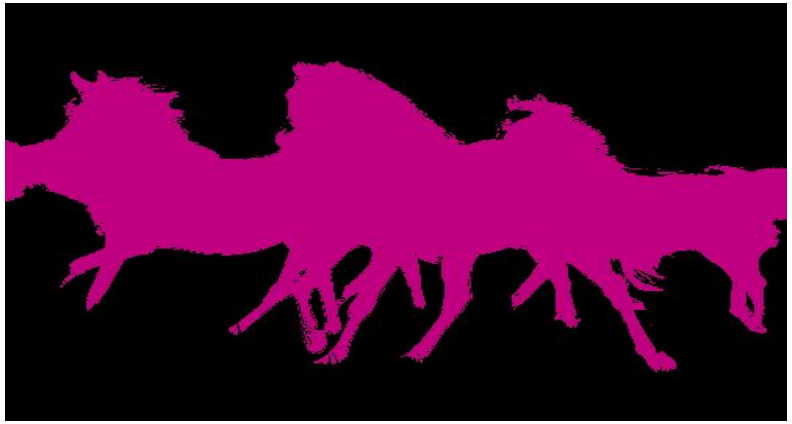
$$\min_{Y=[\mathbf{y}^1,\dots\mathbf{y}^{|\Omega|}]} \sum_{p\in\Omega} l(\mathbf{y}^p, \mathbf{s}_\theta^p) + \sum_{p,q\in\Omega^2} w_{pq}[\mathbf{y}^p \neq \mathbf{y}^q]$$

$\mathbf{x}^p \in \mathbb{R}^N$ (Image colors)



$\mathbf{s}_\theta^p \in [0,1]^L$ (Network outputs)

$\mathbf{y}^p \in \{0,1\}^L$ (Binary labels)



w_{pq}

Decreasing function of

$\|\mathbf{x}_p - \mathbf{x}_q\|$

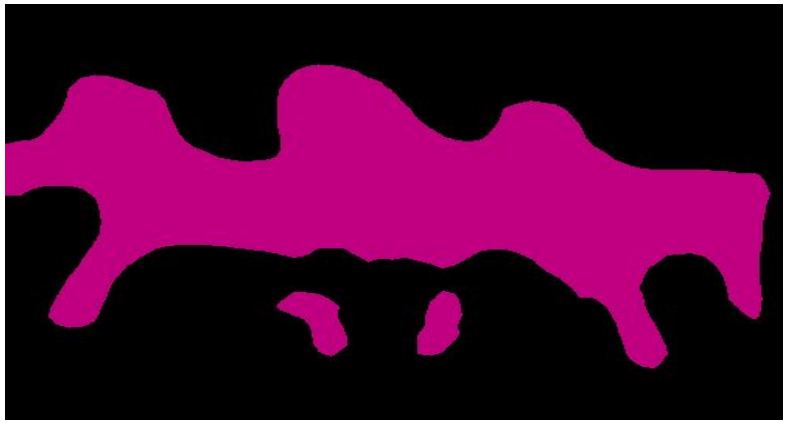
[Chen et al., Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected CRFs. TPAMI 2018]

CRFs meet fully supervised CNNS

Pairwise potentials (also the very popular *DenseCRF*)

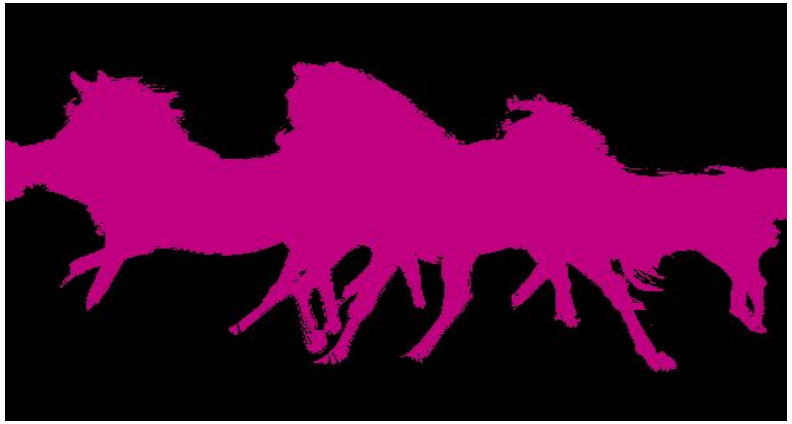
$$\min_{Y=[\mathbf{y}^1,\dots\mathbf{y}^{|\Omega|}]} \sum_{p\in\Omega} l(\mathbf{y}^p, \mathbf{s}_\theta^p) + \sum_{p,q\in\Omega^2} w_{pq}[\mathbf{y}^p \neq \mathbf{y}^q]$$

$\mathbf{x}^p \in \mathbb{R}^N$ (Image colors)



$\mathbf{s}_\theta^p \in [0,1]^L$ (Network outputs)

$\mathbf{y}^p \in \{0,1\}^L$ (Binary labels)



w_{pq}

Decreasing function of

$\|\mathbf{x}_p - \mathbf{x}_q\|$

[Chen et al., Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected CRFs. TPAMI 2018]

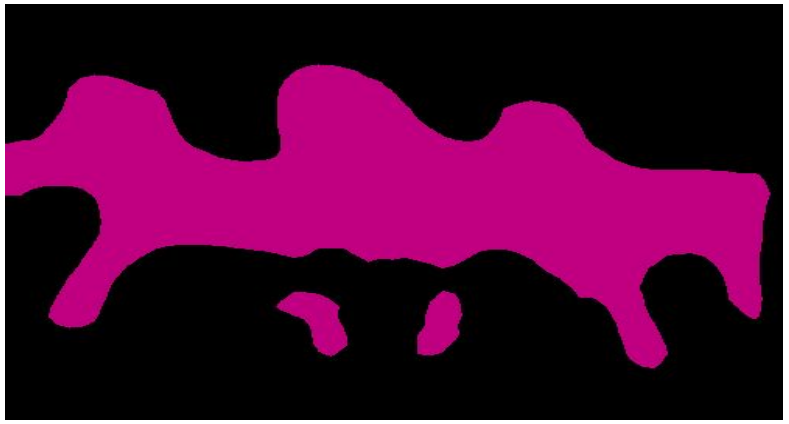
CRFs meet fully supervised CNNS

Pairwise potentials (also the very popular *Laplacian*) $\|\mathbf{y}^p - \mathbf{y}^q\|^2$

For one-hot encoding vectors

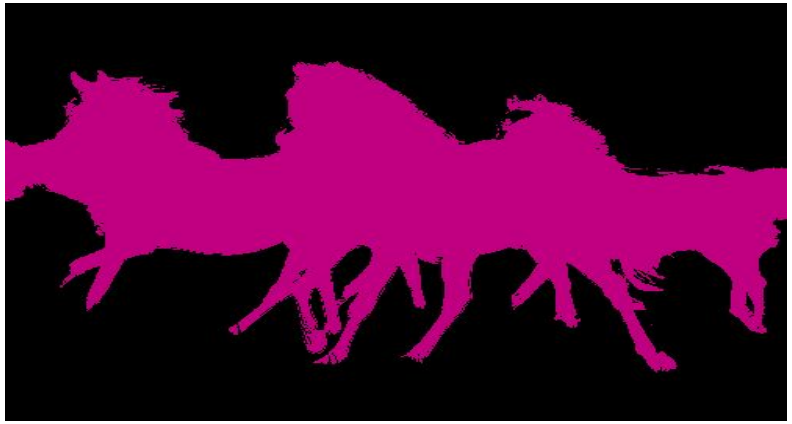
$$\min_{Y=[\mathbf{y}^1,\dots\mathbf{y}^{|\Omega|}]} \sum_{p\in\Omega} l(\mathbf{y}^p, \mathbf{s}_\theta^p) + \sum_{p,q\in\Omega^2} w_{pq}[\mathbf{y}^p \neq \mathbf{y}^q]$$

$\mathbf{x}^p \in \mathbb{R}^N$ (Image colors)



$\mathbf{s}_\theta^p \in [0,1]^L$ (Network outputs)

$\mathbf{y}^p \in \{0,1\}^L$ (Binary labels)



w_{pq}
Decreasing function of
$\|\mathbf{x}_p - \mathbf{x}_q\|$

[Chen et al., Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected CRFs. TPAMI 2018]

A long history in computer vision for optimizing pairwise potentials (Potts, DenseCRF, Laplacian)

- The most influential works:

✔ **Graph cuts**:
Boykov et al., TPAMI'01 (over 9000 citations, test-of-time award)

✔ **Mean-field approximation**:
Krahenbuhl and Koltun, NIPS'11 (~2300 citations)

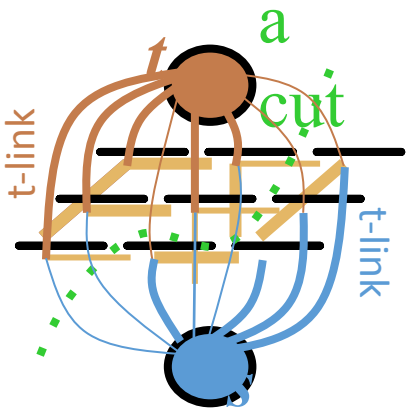
Graph cuts *vs.* first-order methods

$$\sum_{p,q \in \mathcal{N}} w_{pq} \|\mathbf{y}^p - \mathbf{y}^q\|^2$$

$g_{pq}(y_p, y_q)$

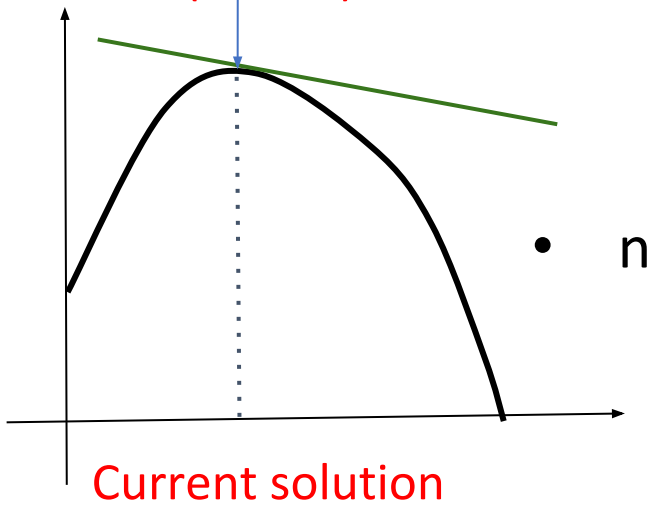
submodular

$$g_{pq}(0,0) + g_{pq}(1,1) < g_{pq}(0,1) + g_{pq}(1,0)$$



- Global optimality (binary)
- quality bounds (multi-label)

Linear approximation (bound)

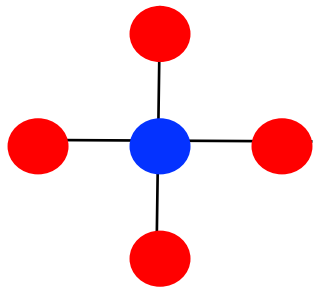


- no optimality guarantee

Graph cuts *vs.* first-order methods

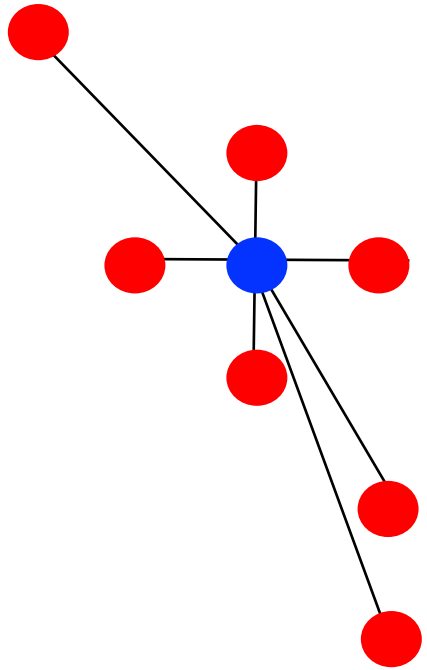
$$\sum_{p,q \in \mathcal{N}} w_{pq} \|\mathbf{y}^p - \mathbf{y}^q\|^2$$

Grid CRF



- Very powerful for grid (sparse) CRF
- Not scalable for Dense CRF
- Cannot be easily parallelized

Dense CRF



- Scalable for Dense CRF
- Parallel implementation

Graph cuts *vs.* first-order methods

Classical ‘shallow’ segmentation

- Better alignment with edges
- More regular boundaries (geometric length interpretation)

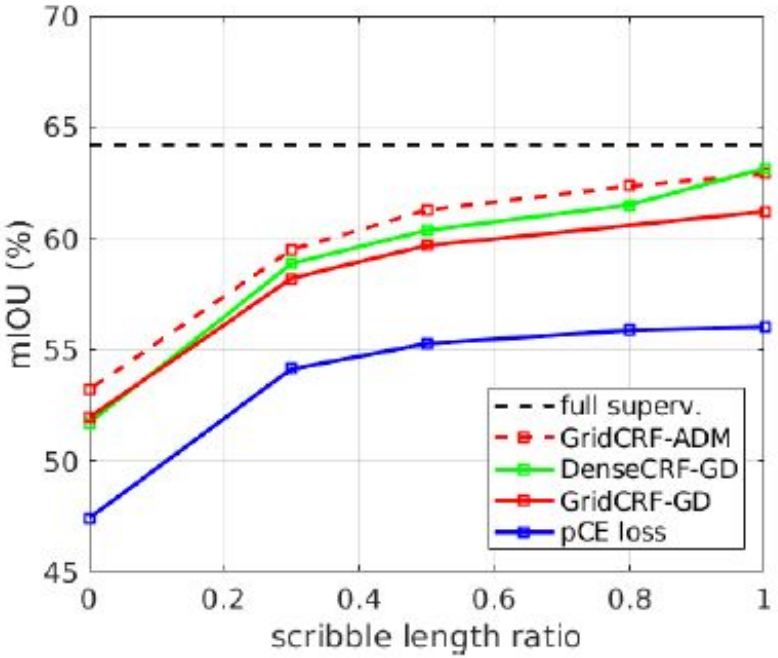
Grid CRF + graph cut

Dense CRF + first-order

- Irregular boundaries and poor edge alignment
- The popularity is due to computational efficiency and…

Semi-supervision loss for segmentation

$$\min_{\theta} \sum_{p \in \mathcal{L}} l(\mathbf{y}^p, \mathbf{s}_\theta^p) + \sum_{p,q \in \mathcal{L} \cup \mathcal{U}} w_{p,q} \|\mathbf{s}_\theta^p - \mathbf{s}_\theta^q\|^2$$



The disturbing part (for those who know classical CRFs):

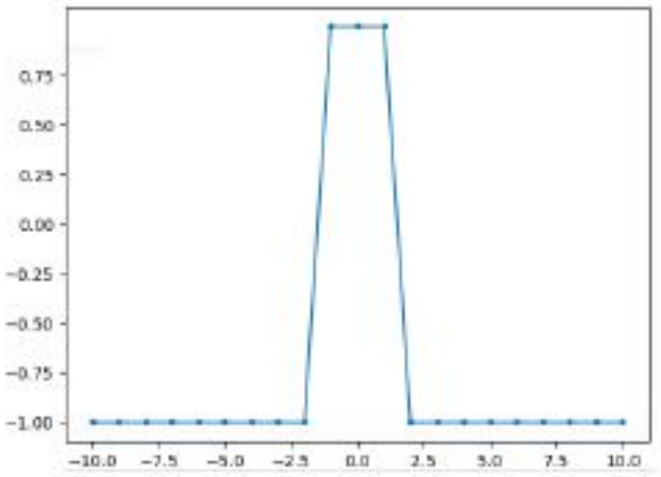
Dense CRF is not supposed to be better than grid CRF

[Tang et al., On regularized losses for weakly supervised segmentation, ECCV 2018]
[Marin et al., Beyond gradient descent for regularized segmentation losses, CVPR 2019]

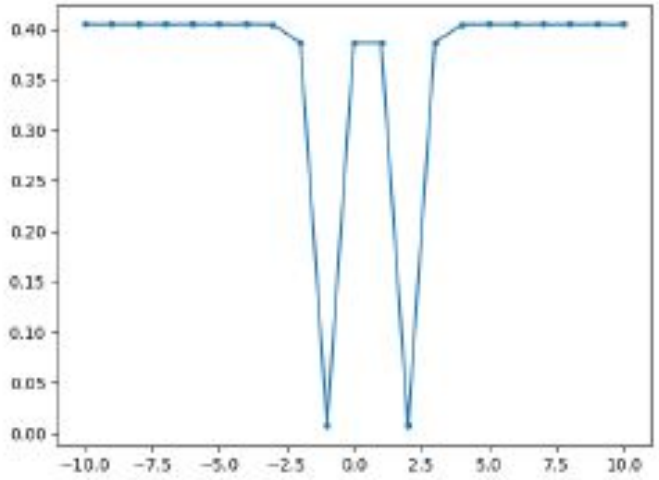
Dense CRF is **NOT** supposed to be better, but…

- Consider a 1-D image: $I(x)$
- Plot the CRF term as function of several segmentations: $S^t = \{x | x < t\}$

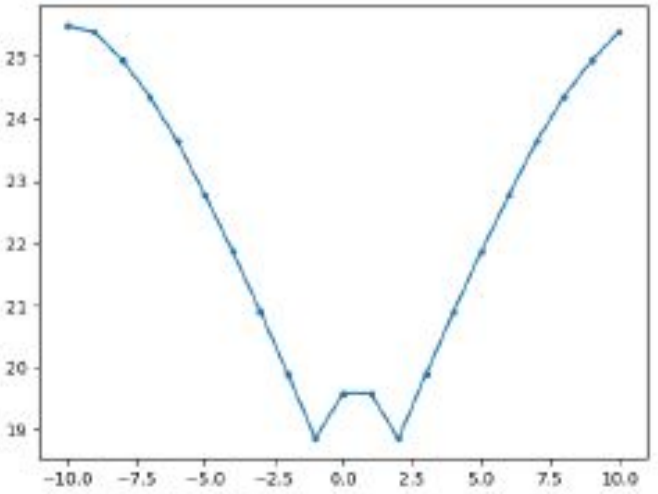
$I(x)$



Grid CRF term



Dense CRF term

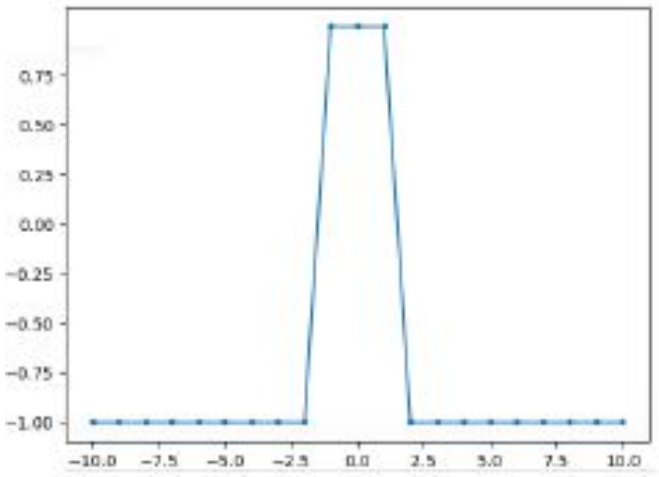


[Marin et al., Beyond gradient descent for regularized segmentation losses, CVPR 2019]

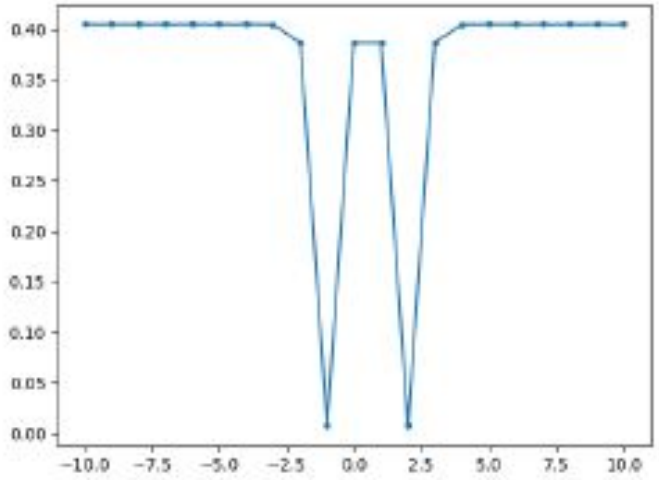
Dense CRF is **NOT** supposed to be better, but…

- Dense CRF yields a **smoother** cost function (facilitates optimization)
- The flatter minimum may complicate discontinuity localization

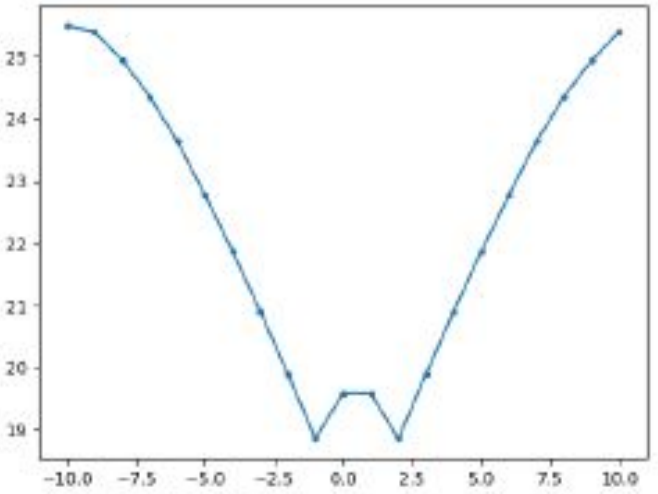
$I(x)$



Grid CRF term



Dense CRF term



[Marin et al., Beyond gradient descent for regularized segmentation losses, CVPR 2019]

Beyond gradient descent for regularized losses

- Let us first simplify the notation:

$$\min_{\theta} \sum_{p \in \mathcal{L}} l(\mathbf{y}^p, \mathbf{s}_\theta^p) + \sum_{p,q \in \mathcal{L} \cup \mathcal{U}} w_{p,q} \|\mathbf{s}_\theta^p - \mathbf{s}_\theta^q\|^2$$

$L(S_\theta, Y)$

A few labeled points

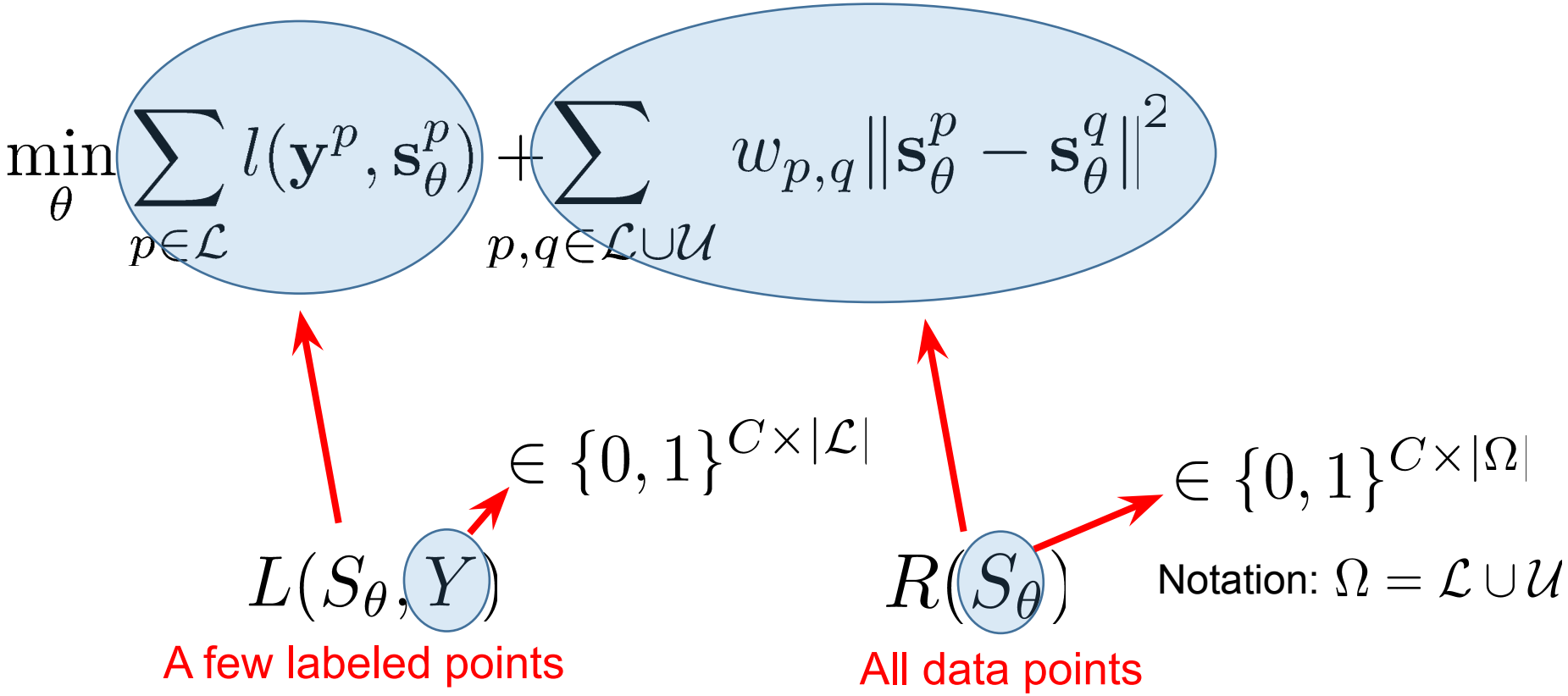
$R(S_\theta)$

All data points

[Marin et al., Beyond gradient descent for regularized segmentation losses, CVPR 2019]

Beyond gradient descent for regularized losses

- Let us first simplify the notation:



[Marin et al., Beyond gradient descent for regularized segmentation losses, CVPR 2019]

Beyond gradient descent for regularized losses

$$\min_{\theta} L(S_\theta, Y) + R(S_\theta)$$

A splitting of the problem (Alternating Direction Method)

$$\min_{\theta, \hat{Y}} L(S_\theta, Y) + R(\hat{Y})$$

$$\text{s.t.} \quad \hat{\mathbf{y}}^p = \mathbf{s}_\theta^p \quad \forall p \in \mathcal{U}$$

$$\hat{\mathbf{y}}^p = \mathbf{y}^p \quad \forall p \in \mathcal{L}$$

[Marin et al., Beyond gradient descent for regularized segmentation losses, CVPR 2019]

Beyond gradient descent for regularized losses

$$\min_{\theta} L(S_\theta, Y) + R(S_\theta)$$

A splitting of the problem (Alternating Direction Method)

$$\min_{\theta, \hat{Y}} L(S_\theta, Y) + R(\hat{Y})$$

$$\text{s.t.} \quad \hat{\mathbf{y}}^p = \mathbf{s}_\theta^p \quad \forall p \in \mathcal{U}$$

$$\hat{\mathbf{y}}^p = \mathbf{y}^p \quad \forall p \in \mathcal{L}$$

Fake labels (or proposals)

[Marin et al., Beyond gradient descent for regularized segmentation losses, CVPR 2019]

Beyond gradient descent for regularized losses

$$\min_{\theta} L(S_\theta, Y) + R(S_\theta)$$

A splitting of the problem (Alternating Direction Method)

$$\min_{\theta, \hat{Y}} L(S_\theta, Y) + R(\hat{Y})$$

$$\text{s.t.} \quad \hat{\mathbf{y}}^p = \mathbf{s}_\theta^p \quad \forall p \in \mathcal{U}$$

$$\hat{\mathbf{y}}^p = \mathbf{y}^p \quad \forall p \in \mathcal{L}$$

$\hat{Y} \in \{0, 1\}^{K \times |\Omega|}$

Discrete binary variables
(amenable to graph cut optimization)

[Marin et al., Beyond gradient descent for regularized segmentation losses, CVPR 2019]

Beyond gradient descent for regularized losses: Alternate **two steps**, each decreasing the loss

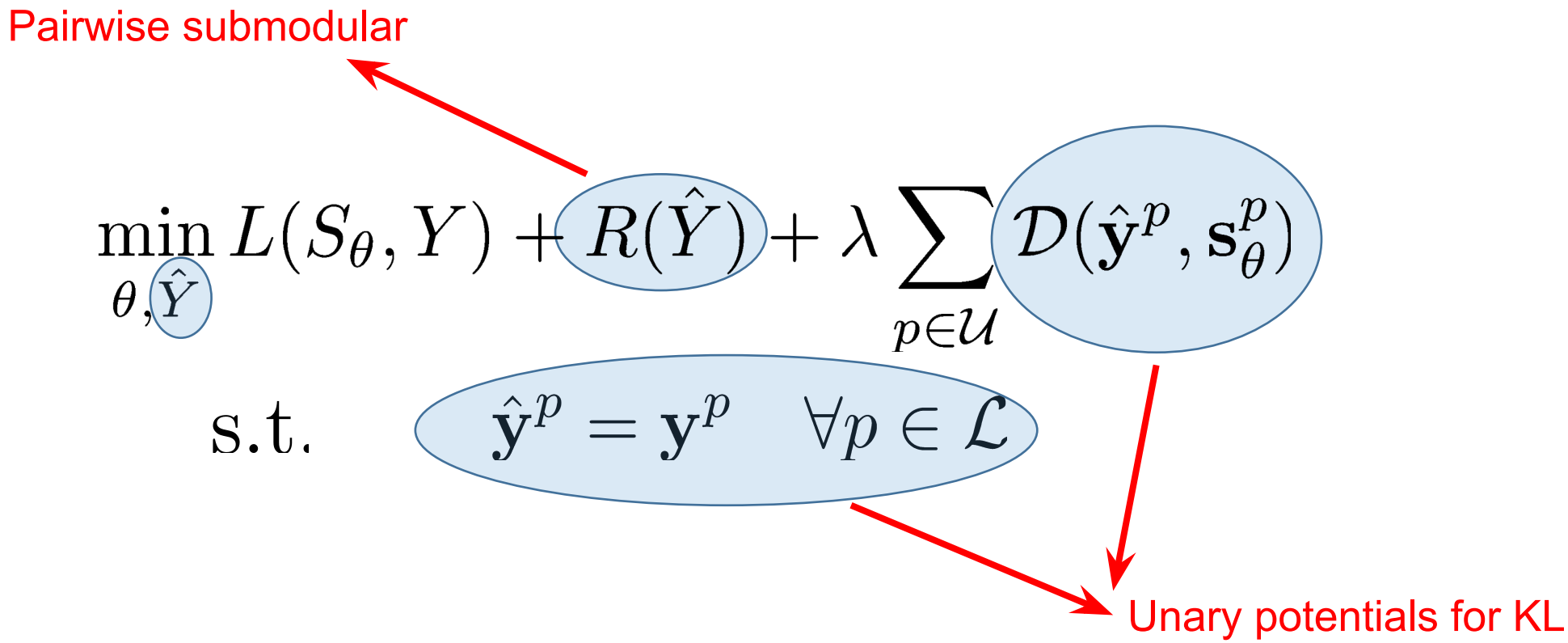
$$\min_{\theta, \hat{Y}} L(S_\theta, Y) + R(\hat{Y}) + \lambda \sum_{p \in \mathcal{U}} \mathcal{D}(\hat{\mathbf{y}}^p, \mathbf{s}_\theta^p)$$

$$\text{s.t.} \qquad \hat{\mathbf{y}}^p = \mathbf{y}^p \quad \forall p \in \mathcal{L}$$

[Marin et al., Beyond gradient descent for regularized segmentation losses, CVPR 2019]

Beyond gradient descent for regularized losses

Step 1: **Graph cuts with network parameters fixed**



[Marin et al., Beyond gradient descent for regularized segmentation losses, CVPR 2019]

Beyond gradient descent for regularized losses
Step 2: **Standard SGD for cross-entropy learning**

(Note: equivalent to a cross-entropy with 'fake' ground-truth labels)

Kullback-Leibler (KL) divergence

$$\min_{\theta, \hat{Y}} L(S_\theta, Y) + R(\hat{Y}) + \lambda \sum_{p \in \mathcal{U}} \mathcal{D}(\hat{\mathbf{y}}^p, \mathbf{s}_\theta^p)$$

[Marin et al., Beyond gradient descent for regularized segmentation losses, CVPR 2019]

Link to standard ADMM?

Alternating **D**irection **M**ethod of Multipliers (ADMM):

$$\min_s g(s) + r(s)$$

[Boyd et al., Distributed optim. and stat. learning via ADMM, Foundation and trends in ML 2010]

Link to standard ADMM?

Alternating **D**irection **M**ethod of Multipliers (ADMM):

$$\min_{s} g(s) + r(s)$$

$$\min_{s,y} g(s) + r(y)$$

$$\text{s.t} \quad s = y$$

[Boyd et al., Distributed optim. and stat. learning via ADMM, Foundation and trends in ML 2010]

Link to standard ADMM?

Alternating **D**irection **M**ethod of Multipliers (ADMM):

$$\min_s g(s) + r(s)$$

$$g(s) + r(y) + \mu(s - y) + \lambda \mathcal{D}(s, y)$$

(Augmented Lagrangian)

[Boyd et al., Distributed optim. and stat. learning via ADMM, Foundation and trends in ML 2010]

Link to standard ADMM?

Alternating **D**irection **M**ethod of Multipliers (**ADM**M):

$$\min_s g(s) + r(s)$$

$$g(s) + r(y) + \lambda(s - y) + \mu\mathcal{D}(s, y)$$

(Penalty method)

[Boyd et al., Distributed optim. and stat. learning via ADMM, Foundation and trends in ML 2010]

Link to standard ADMM?

[Marin et al., CVPR 2019]

Alternating **D**irection **M**ethod of Multipliers (**ADM**M):

$$\min_s g(s) + r(s)$$

$$g(s) + r(y) + \lambda(s - y) + \mu\mathcal{D}(s, y)$$

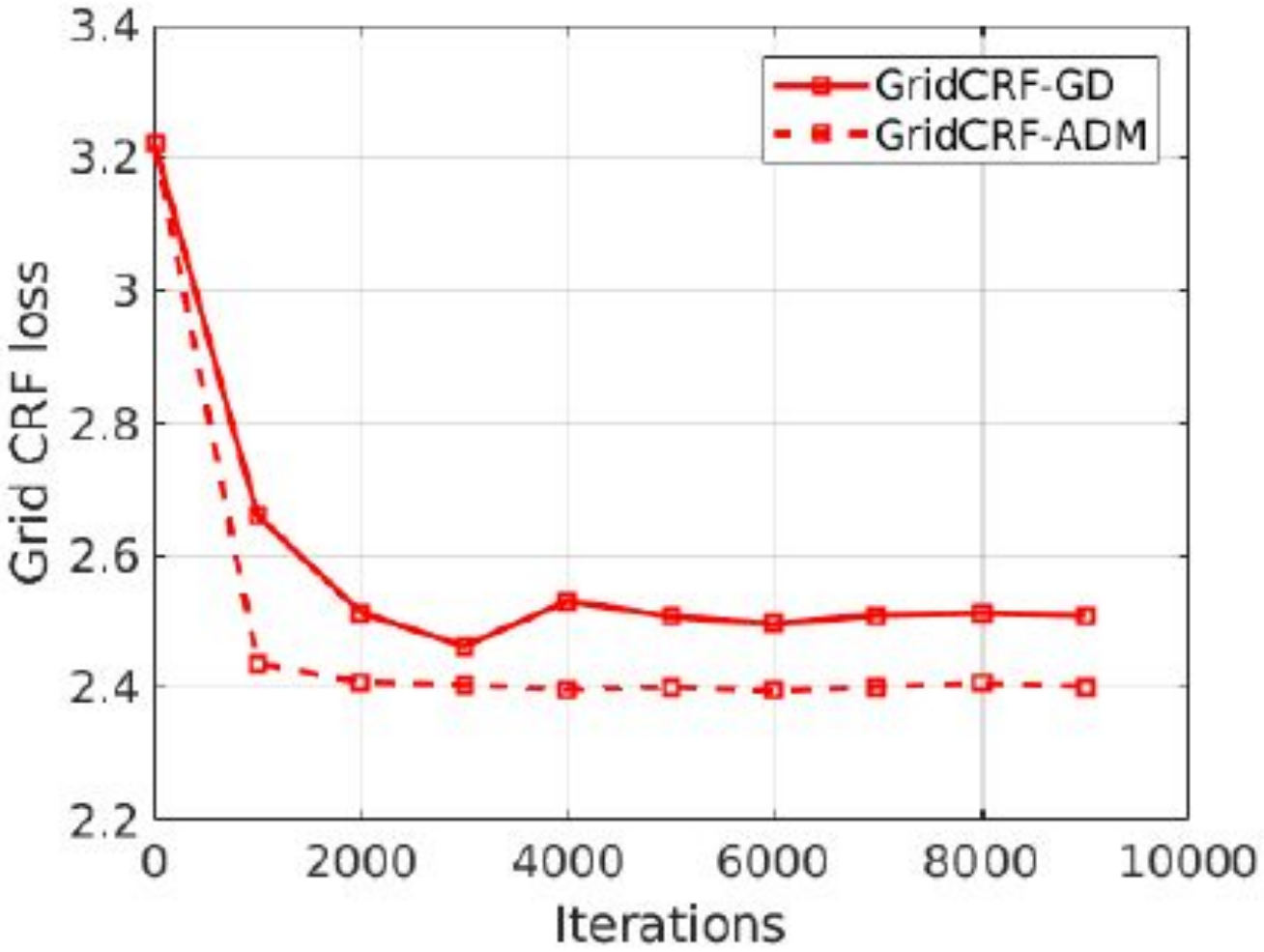
(Typically L_2 in ADMM)

Note: ½ KL is an upper bound on L_2 for simplex vectors (Pinsker's inequality)

(Penalty method)

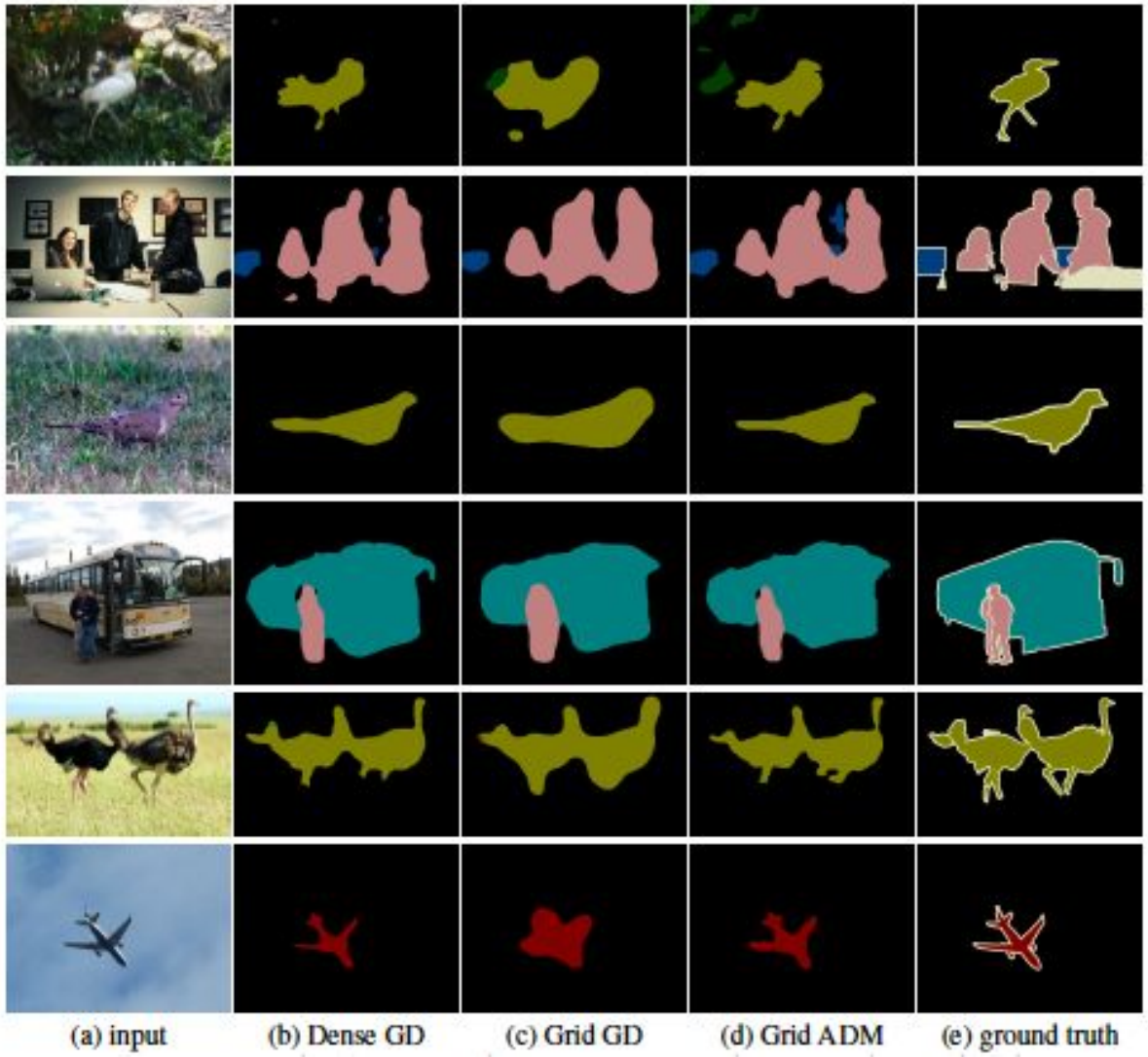
[Boyd et al., Distributed optim. and stat. learning via ADMM, Foundation and trends in ML 2010]

The optimizer matters: Beyond gradient descent



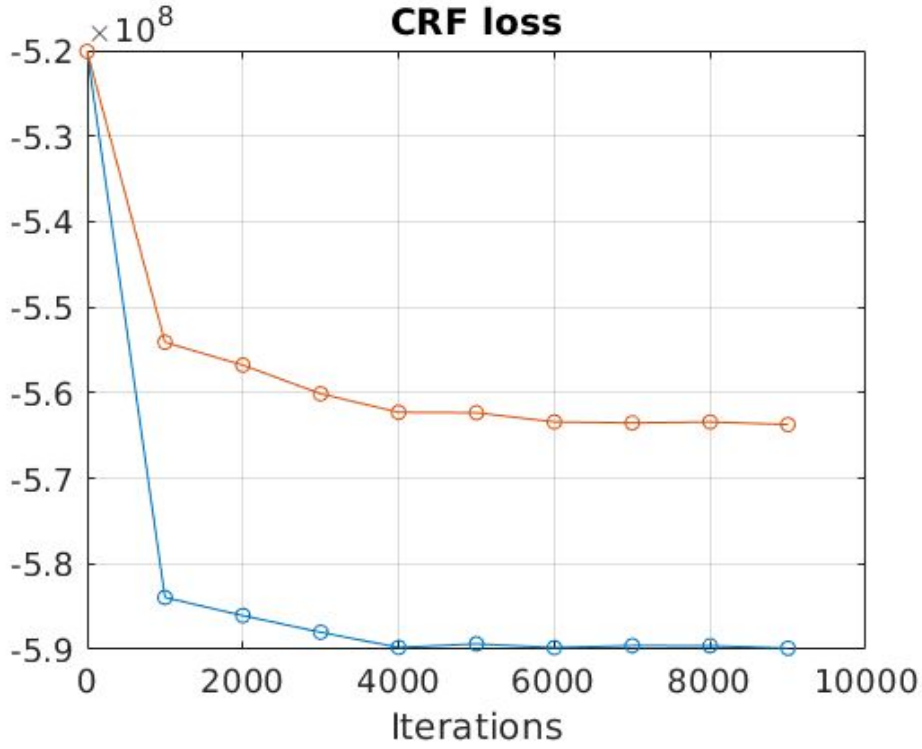
[Marin et al., Beyond gradient descent for regularized segmentation losses, CVPR 2019]

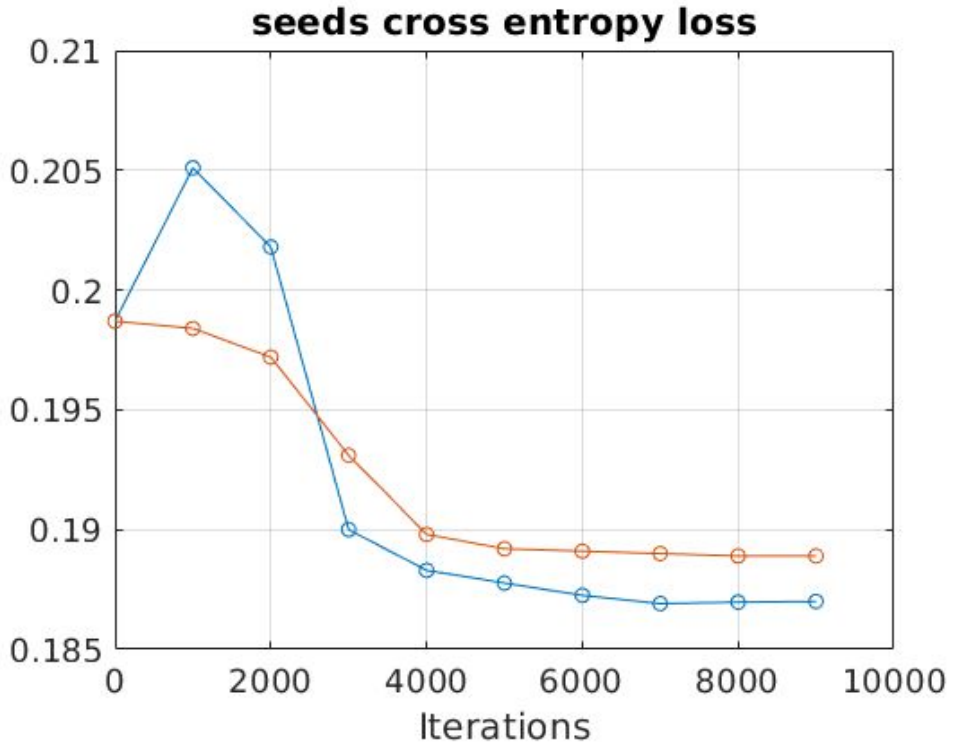
Some visual examples)



[Marin et al., Beyond gradient descent for regularized segmentation losses, CVPR 2019]

ADM does not help with a first-order solver (Dense CRF + Mean-field approximation)





SGD

ADM (SGD + mean-field inference)

[Tang et al., On regularized losses for weakly supervised segmentation, ECCV 2018]

Link to large body of works based on Proposals - ‘Fake’ ground-truth labels

[Lin et al., CVPR 2016], [Khoreva et al. CVPR 2017], [Vernaza et al., CVPR 2017], [Dai et al., CVPR 2015], [Kolesnikov and Lampert, ECCV 2016], [Papandreou et al., ICCV 2015]

[Rajchl et al., TMI 2017]
Bai et al., MICCAI 2017

iter 1 iter 2 iter 3

Training a CNN at each iteration from CRF regularized proposal is ADM

Figures from [Lin et al., ScribbleSup: Scribble-Supervised Convolutional Networks for Semantic Segmentation, CVPR 2016]
Detailed explanation in [Tang et al., On regularized losses for weakly supervised segmentation, ECCV 2018]

Regularization

entropy

Entropy minimization for SSL

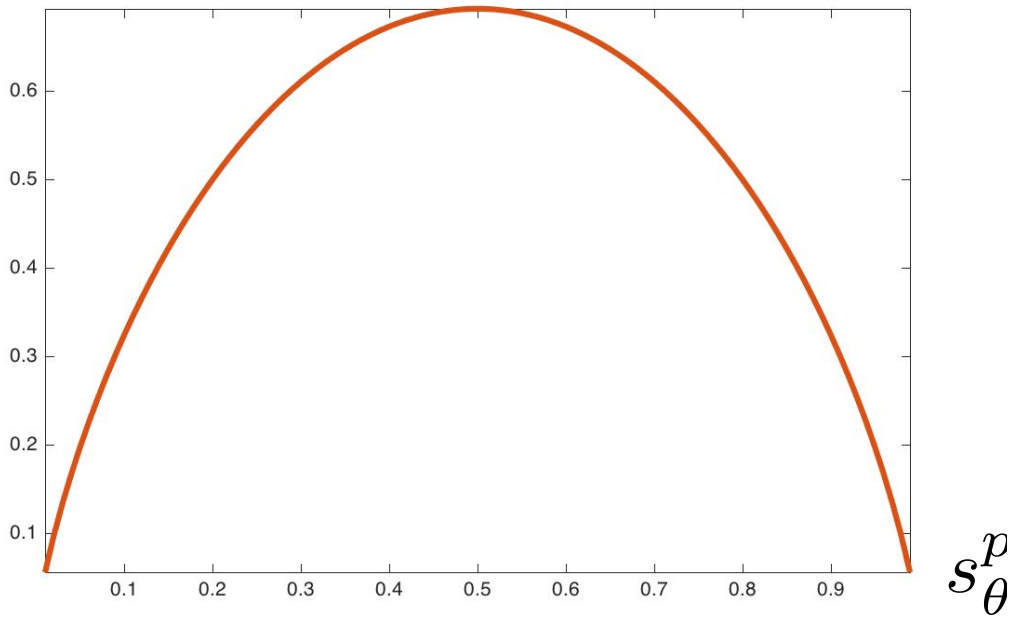
$$\min_{\theta} -\sum_{p\in\mathcal{L}}\sum_{c=1}^{C} y^{p,c}\log s_\theta^{p,c} \;\; -\sum_{p\in\mathcal{U}}\sum_{c=1}^{C} s_\theta^{p,c}\log s_\theta^{p,c}$$

Shannon Entropies: “unsupervised cross-entropies (with unknown labels)”

- Grandvalet & Bengio, Semi-supervised learning by entropy minimization, NIPS 2005
- Gomes et al., Discriminative clustering by regularized information maximization, NIPS 2010

Effect of the entropy (why is it good for SSL?):
It makes the predictions confident (like cross-entropy)

$$-s_\theta^p \log s_\theta^p - (1 - s_\theta^p) \log(1 - s_\theta^p)$$

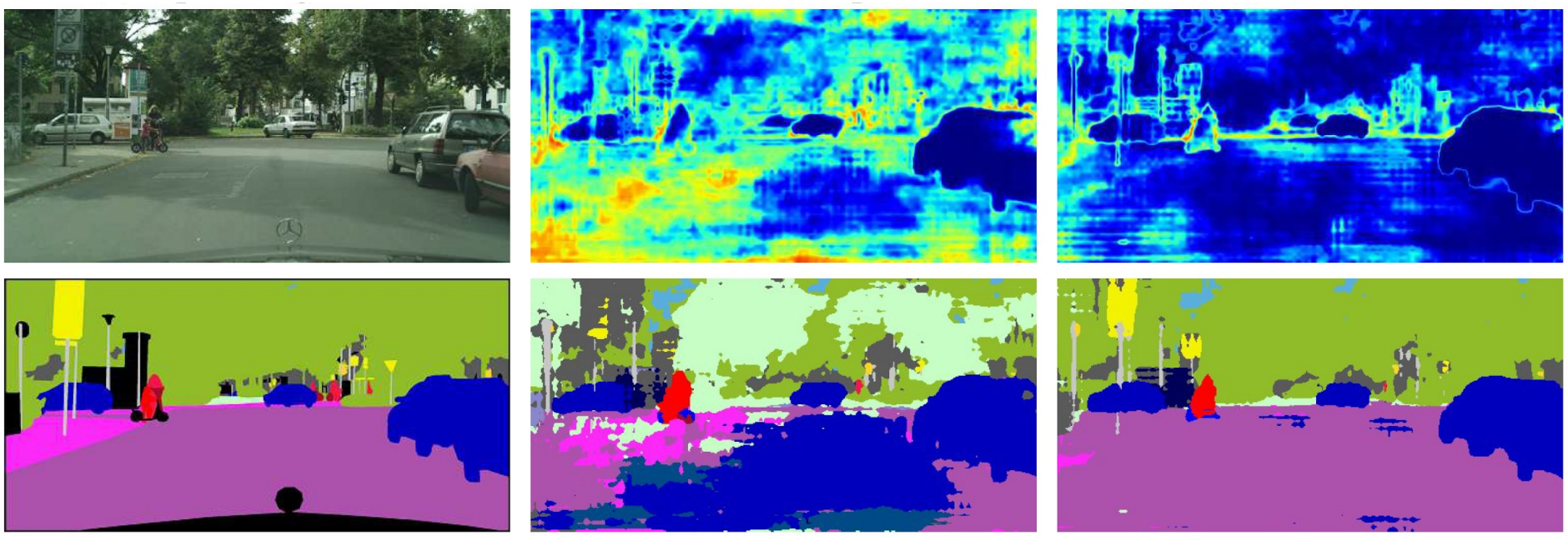


Entropy minimization for UDA

Input image + GT

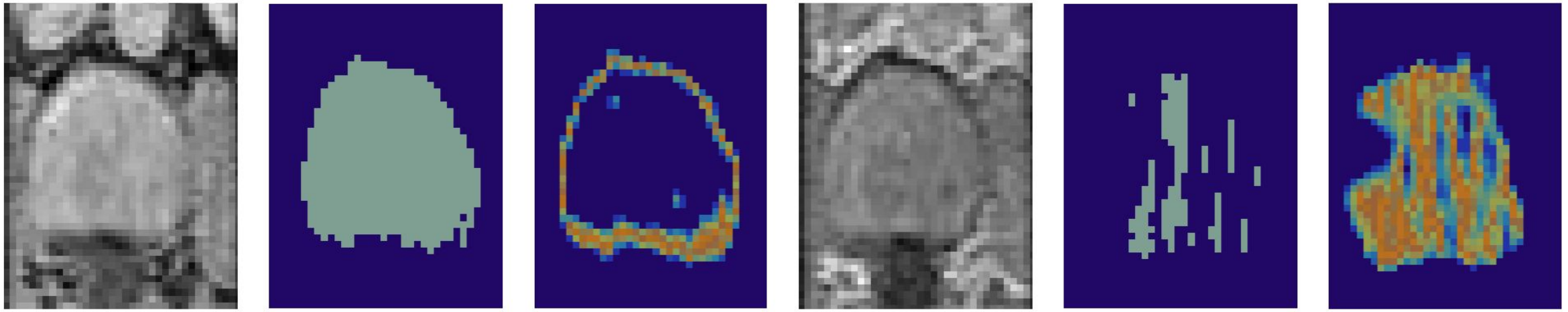
Without adaptation

Entropy minimization



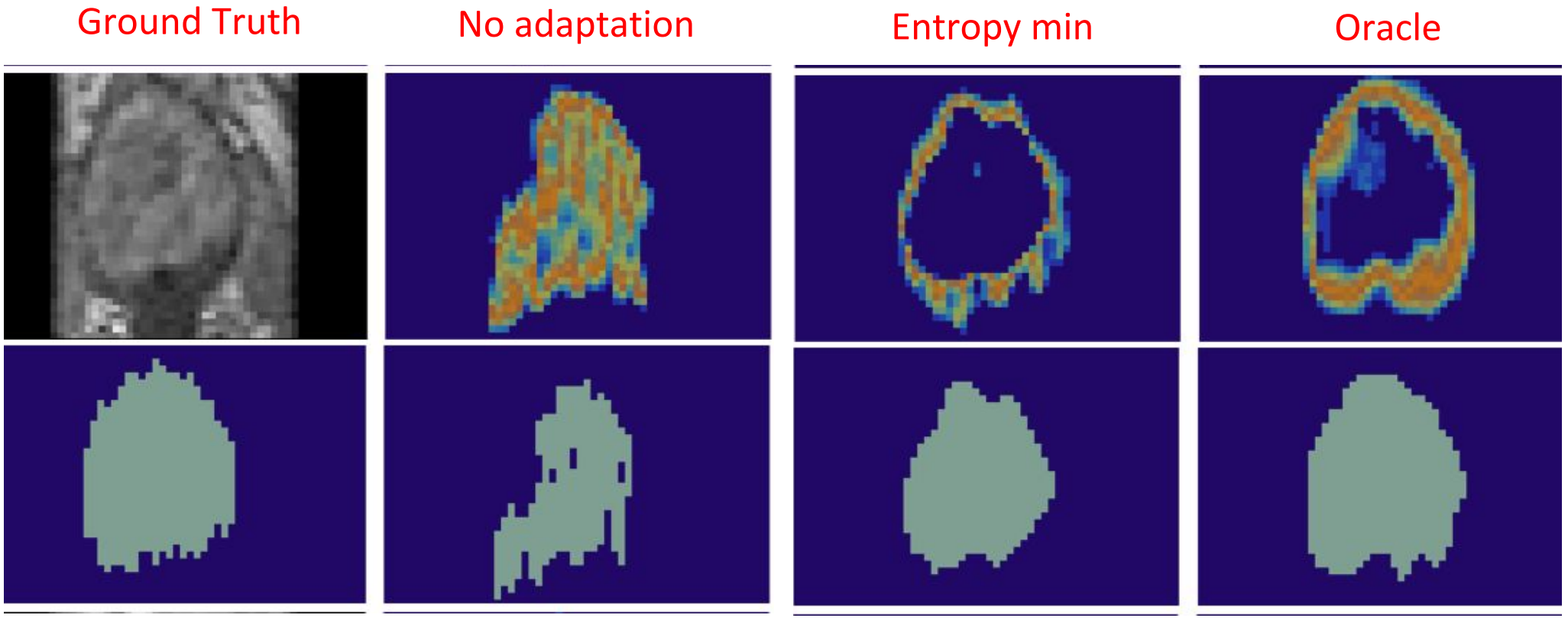
Images from Vu et al., ADVENT: Adversarial Entropy Minimization for Domain Adaptation in Semantic Segmentation, CVPR 2019

Entropy minimization for UDA



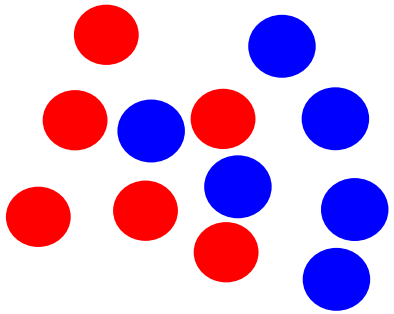
Images from Bateson et al., Source-relaxed domain adaptation for segmentation, MICCAI 2020

Entropy minimization for UDA



Images from Bateson et al., Source-relaxed domain adaptation for segmentation, MICCAI 2020

Why entropy minimization is good
(It increases the margin between the classes)



High entropy
(low confidence)

Low entropy
(high confidence)

Effect of the entropy (why is it good for SSL?): It increases the margin between the classes

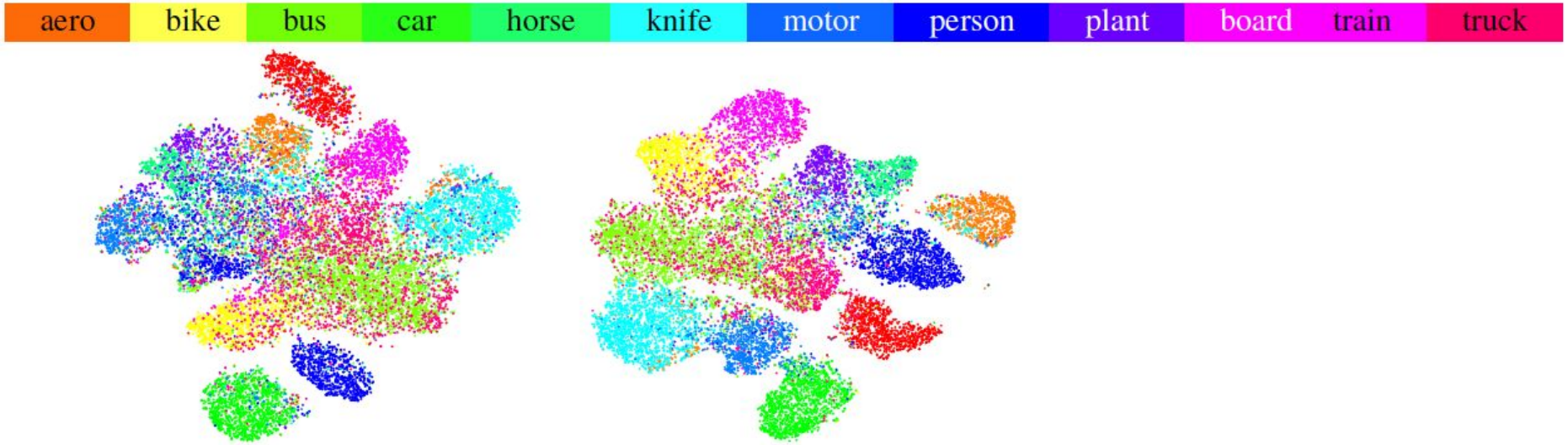
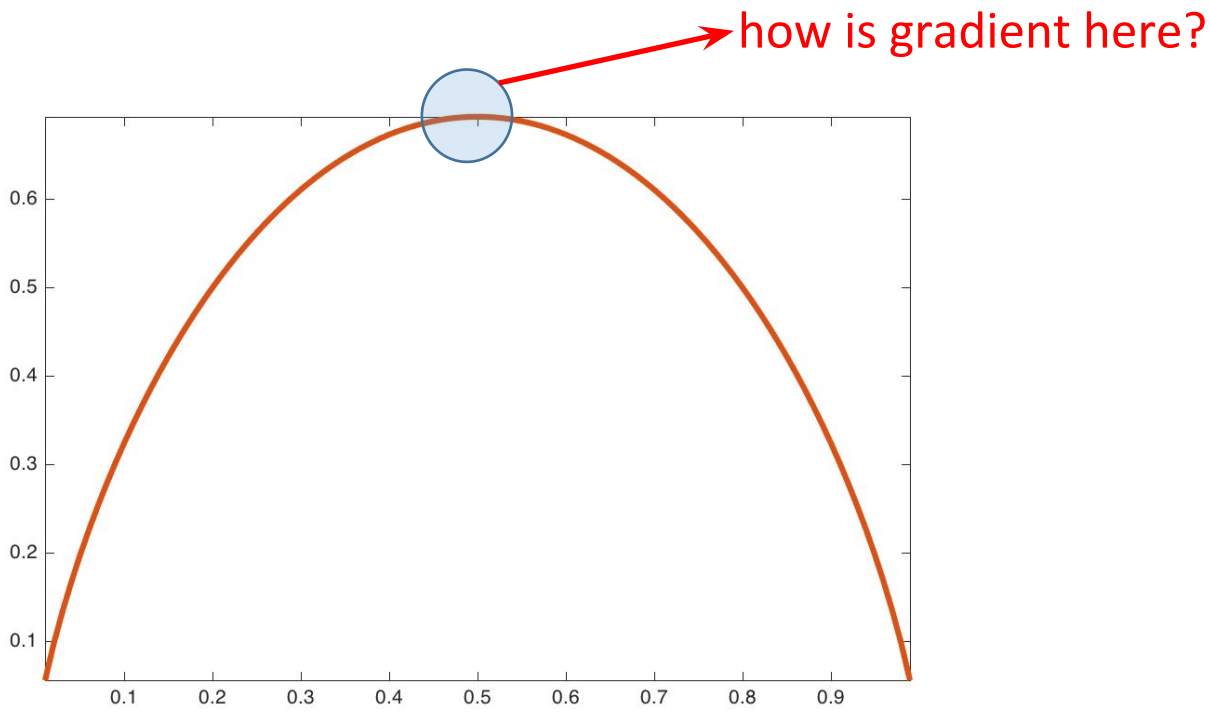


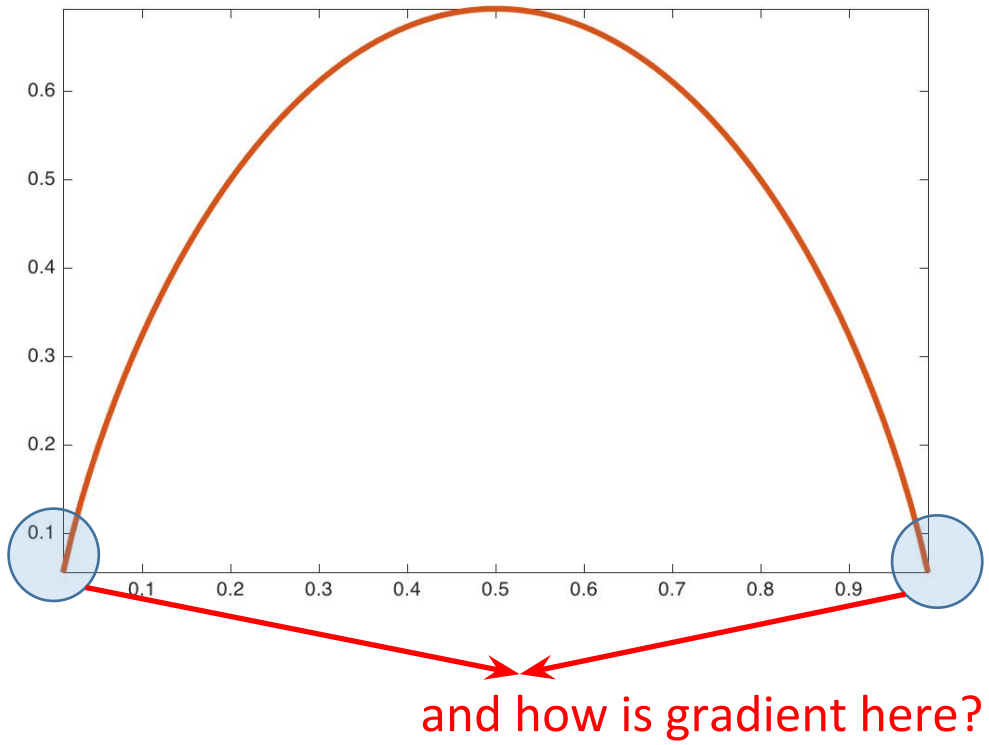
Image classification UDA on VisDA17 data set: Feature visualization for source model (left) and *min-entropy (lower bound on Shannon)* minimization (right) - equivalent to self training (clarified in the next slide)

Figures from Zou et al., Confidence regularized self training, ICCV 2019

Difficulty of optimizing entropy

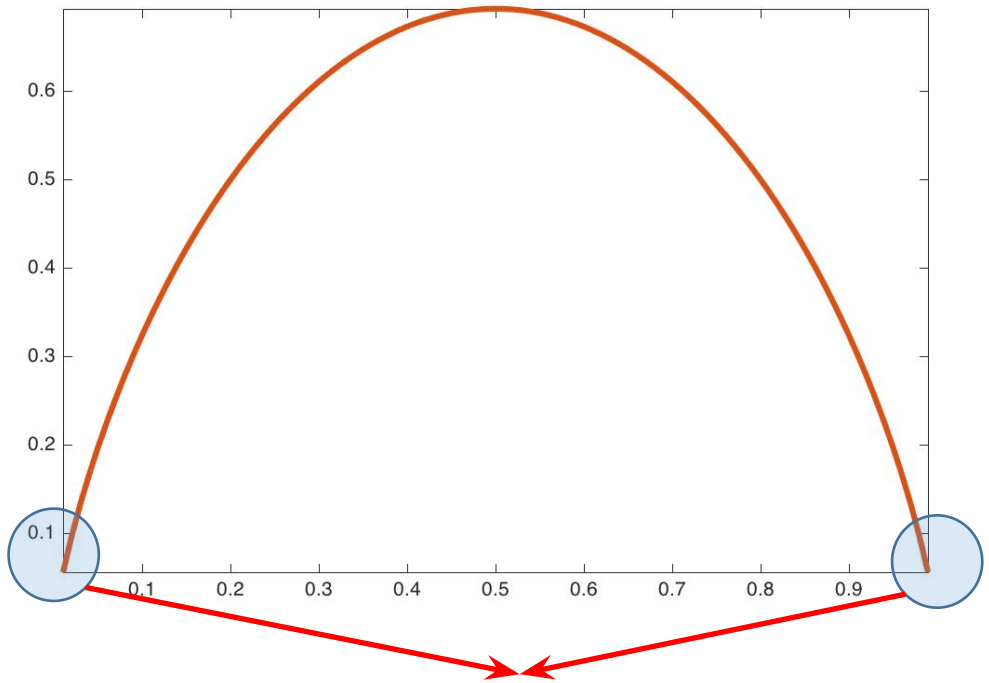


Difficulty of optimizing entropy



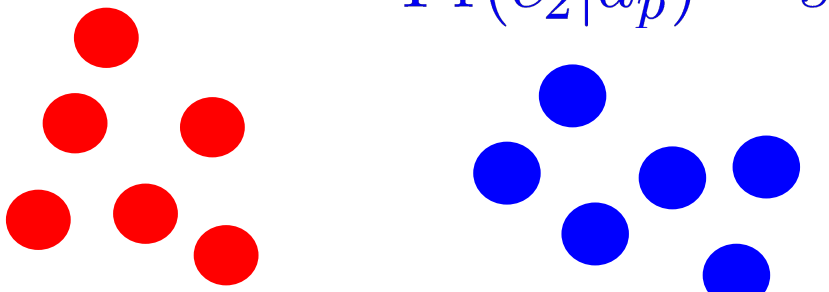
Difficulty of optimizing entropy

Typically we add other cues to facilitate optimization and avoid trivial solutions (more on this later)



Avoiding the trivial solutions of entropy minimization

$$\Pr(c_2|d_p) = s^{p,2} = 1 - s^{p,1} = 1$$



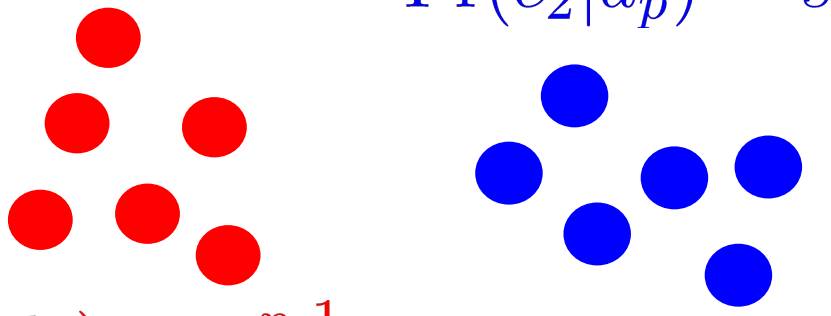
$$\Pr(c_1|d_p) = s^{p,1} = 1$$

Softmax output

Note: Ignoring the network parameters to simplify notation

Avoiding the trivial solutions of entropy minimization

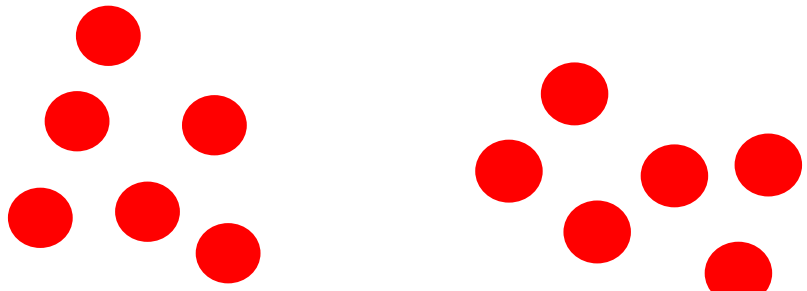
$$\Pr(c_2|d_p) = s^{p,2} = 1 - s^{p,1} = 1$$



$$\Pr(c_1|d_p) = s^{p,1} = 1$$

Min entropy
(max confidence)

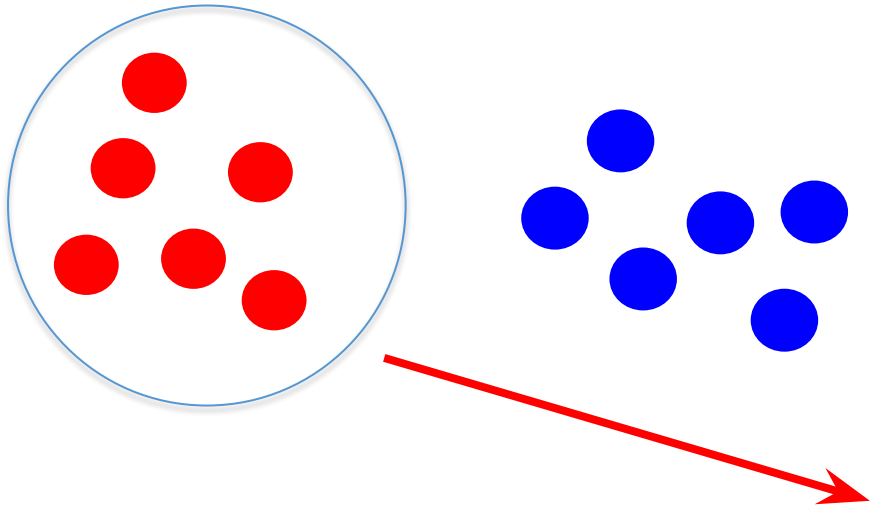
Avoiding the trivial solutions of entropy minimization



$$\Pr(c_1|d_p) = s^{p,1} = 1$$

This bad solution also has a minimum entropy!!!

Avoiding the trivial solutions of entropy minimization



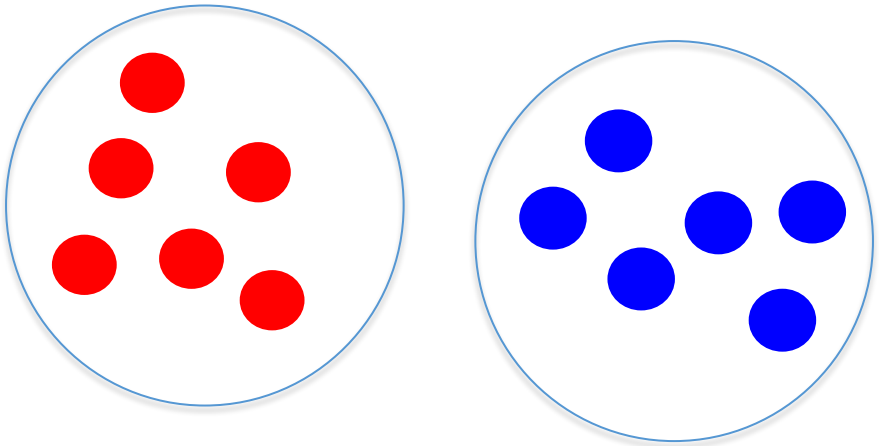
Marginal probabilities of the labels

-- Class proportion

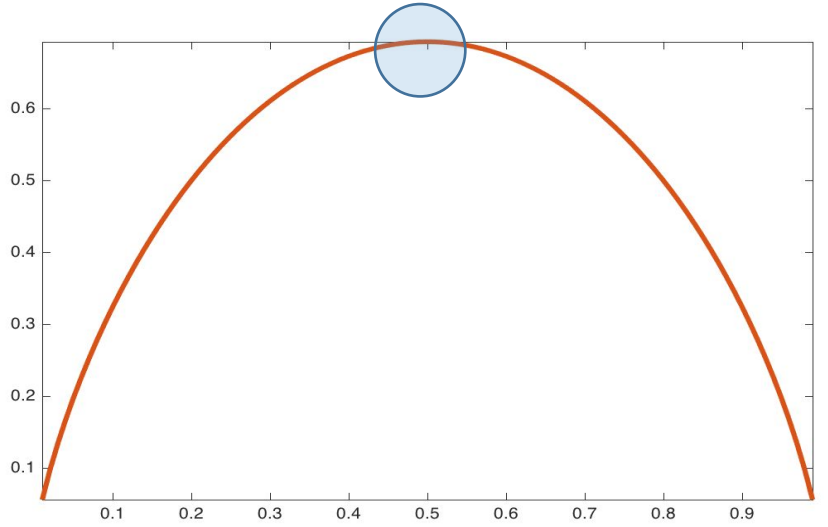
-- Region size (normalized) in segmentation

$$\Pr(c_1) \propto \sum_p s^{p,1}$$

Avoiding the trivial solutions of entropy minimization



Balanced solution **maximizes** the entropy of **label marginal**



$\Pr(c_1)$

Maximizing the mutual info (MI) (between data points and their latent labels)

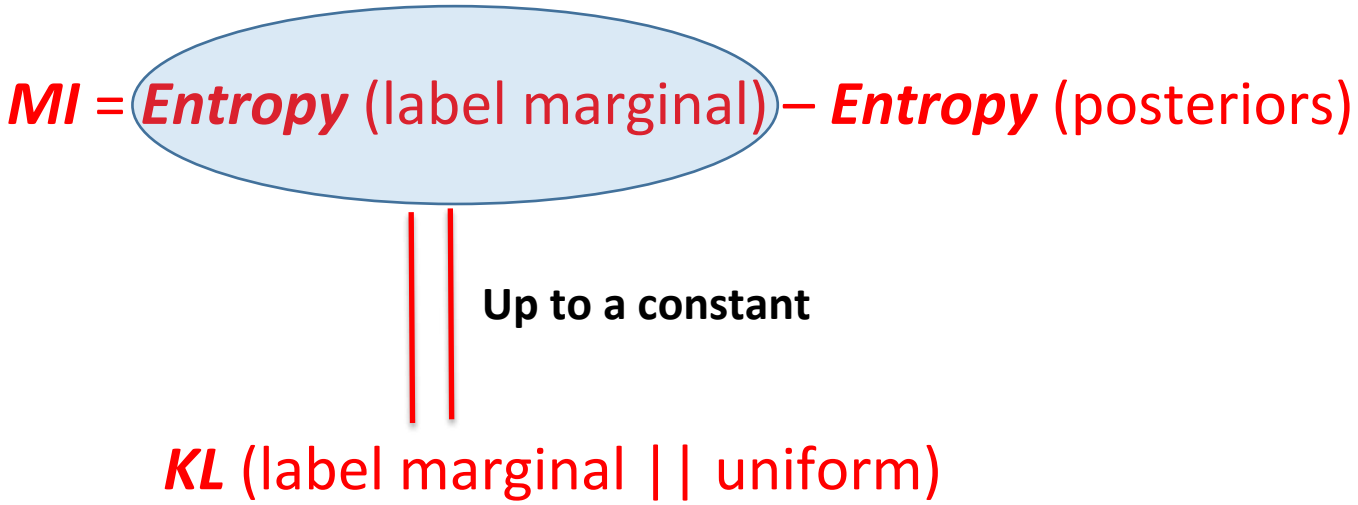
$$I(X,Y) = H(Y) - H(Y/X)$$

MI = ***Entropy*** (label marginal) – ***Entropy*** (posterior)

Standard and old in clustering, e.g.,:

Gomes et al., Discriminative clustering by regularized information maximization, NIPS 2010

Maximizing the mutual info (MI) (between data points and their latent labels)

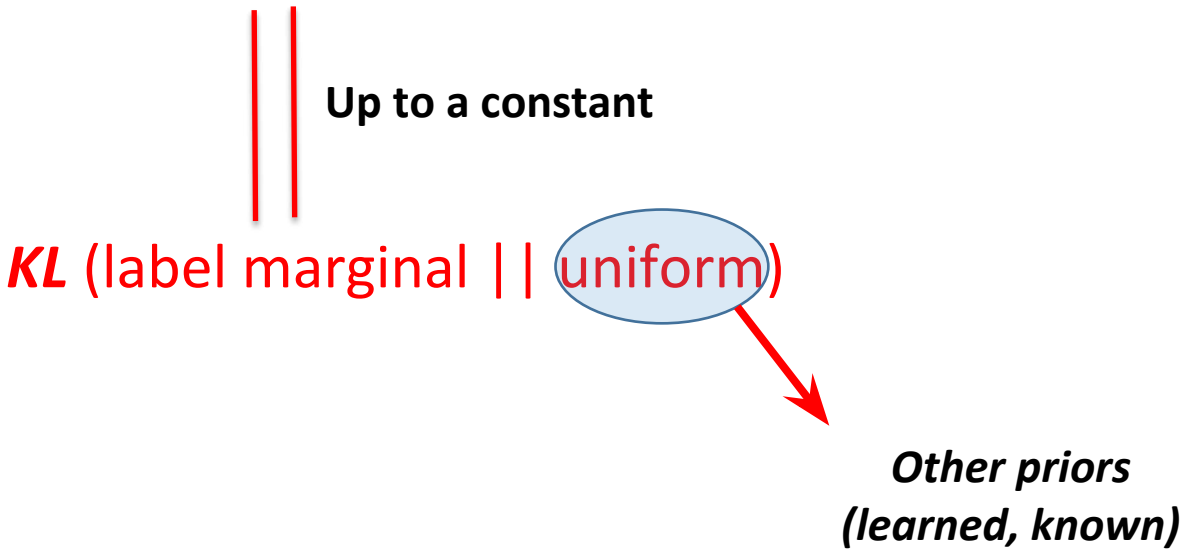


Standard and old in clustering:

Gomes et al., Discriminative clustering by regularized information maximization, NIPS 2010

Maximizing the mutual info (MI) (between data points and their latent labels)

MI = ***Entropy*** (label marginal) – ***Entropy*** (posteriors)

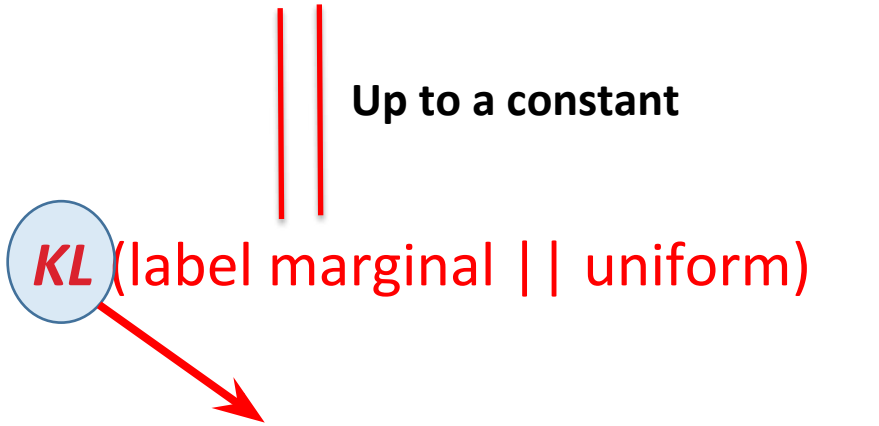


Standard and old in clustering:

Gomes et al., Discriminative clustering by regularized information maximization, NIPS 2010

Maximizing the mutual info (MI) (between data points and their latent labels)

MI = ***Entropy*** (label marginal) – ***Entropy*** (posteriors)



Standard and old in clustering:

Gomes et al., Discriminative clustering by regularized information maximization, NIPS 2010

Maximizing the mutual info (MI) (between data points and their latent labels)

Semi-supervised learning, e.g.
[Berthelot et al., NeurIPS'19]
[Kervadec et al., MedIA'19]

Few-shot learning, e.g.,
[Boudiaf et al., NeurIPS'20]
[Dhillion et al., ICLR'20]

Maximizing MI or its parts/proxies/generalizations
is **SOTA almost everywhere!**

Unsupervised domain adaptation, e.g.,
Liang et al., ICML'20
Bateson et al., MICCAI'20

Deep clustering
&
Unsupervised Representation Learning, e.g.,
Asano et al., ICLR'20
Jabi et al., TPAMI'20

Link to Self-Training

$$-\sum_{p\in\mathcal{U}}\sum_{c=1}^{C}\hat{y}^{p,c}\log s_\theta^{p,c}$$

Pseudo (Fake) labels for unlabeled data points

$$\hat{y}^{p,c*}=1 \quad \text{if} \quad c*=\arg\max_c s_\theta^{p,c} \quad \text{and} \quad 0 \quad \text{otherwise}$$

- Lee, Pseudo-Label : The Simple and Efficient Semi-Supervised Learning Method for Deep Neural Networks, ICML-W 2013
- Zou et al., Unsupervised Domain Adaptation for Semantic Segmentation via Class-Balanced Self-Training, ECCV 2018
- Zou et al., Confidence regularized self training, ICCV 2019

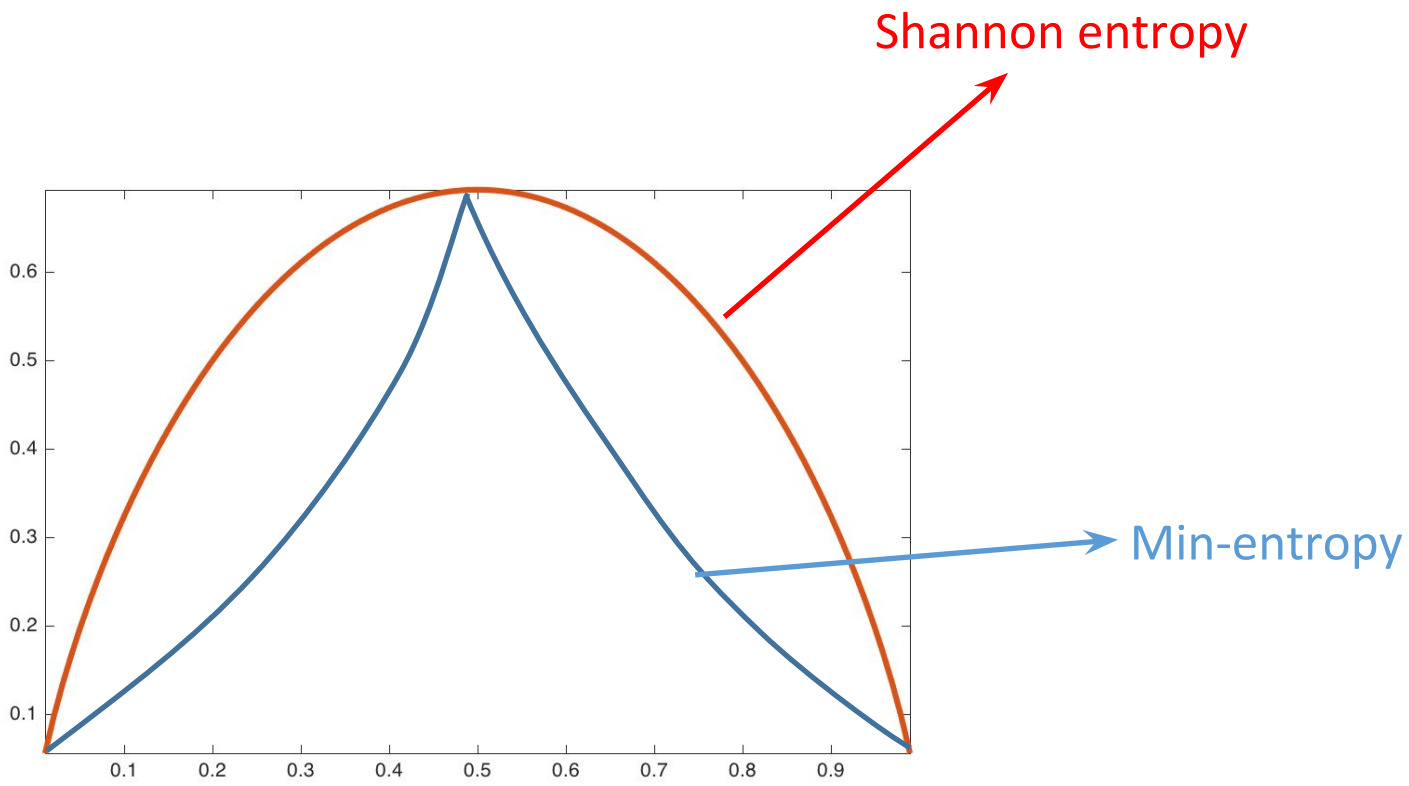
Link to Self-Training

$$-\sum_{p \in \mathcal{U}} \log(\max_{c} s_{\theta}^{p,c})$$

Or equivalently (re-writing without pseudo-labels):
Min-entropy (a lower bound on Shannon entropy)

- Lee, Pseudo-Label : The Simple and Efficient Semi-Supervised Learning Method for Deep Neural Networks, ICML-W 2013
- Zou et al., Unsupervised Domain Adaptation for Semantic Segmentation via Class-Balanced Self-Training, ECCV 2018
- Zou et al., Confidence regularized self training, ICCV 2019

Link to Self-Training



Self-Training + keeping the most confident predictions

$$\hat{y}^{p,c*} = 1 \quad \text{if} \quad c* = \arg\max_c s_\theta^{p,c} \quad \text{and} \quad s_\theta^{p,c*} \geq \exp(-w_c)$$

We keep only the first t% most confident for each class

- Zou et al., Unsupervised Domain Adaptation for Semantic Segmentation via Class-Balanced Self-Training, ECCV 2018
- Zou et al., Confidence regularized self training, ICCV 2019

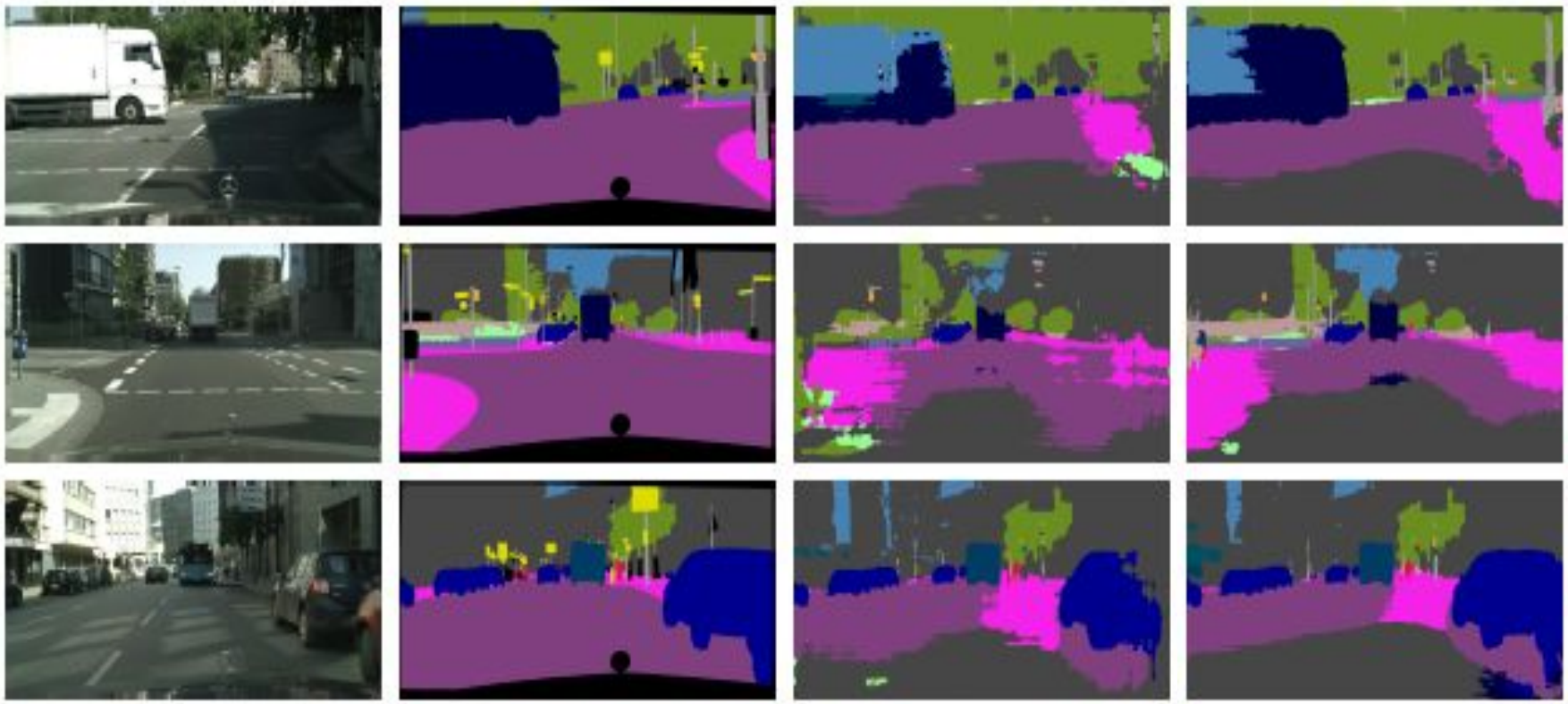
Self-Training + keeping the most confident predictions (corresponds to optimizing this simple loss)

$$\min_{\hat{Y},\theta} -\sum_{p\in\mathcal{L}} y^{p,c}\log(s_\theta^{p,c}) - \sum_{p\in\mathcal{U}} [\hat{y}^{p,c}\log(s_\theta^{p,c}) + w_c\hat{y}^{p,c}]$$

Avoid trivial solution setting all pseudo-labels to 0

- Zou et al., Unsupervised Domain Adaptation for Semantic Segmentation via Class-Balanced Self-Training, ECCV 2018
- Zou et al., Confidence regularized self training, ICCV 2019

Examples of results
(These self-training models are competitive for UDA)



From Zou et al., Unsupervised Domain Adaptation for Semantic Segmentation via Class-Balanced Self-Training, ECCV 2018