Weakly Supervised CNN Segmentation: Models and Optimization

ÉTS
Le génie pour l'industrie

Ismail Ben Ayed
Christian Desrosiers
Jose Dolz

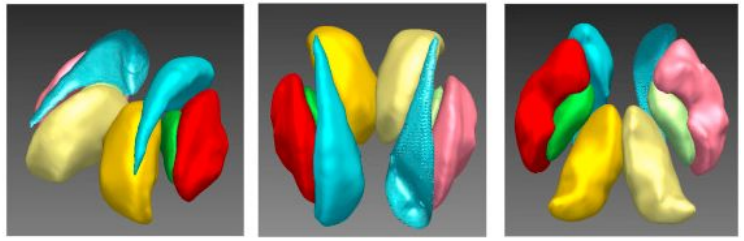
Why are we doing this tutorial at MICCAI?

Deep CNNs are dominating computer vision

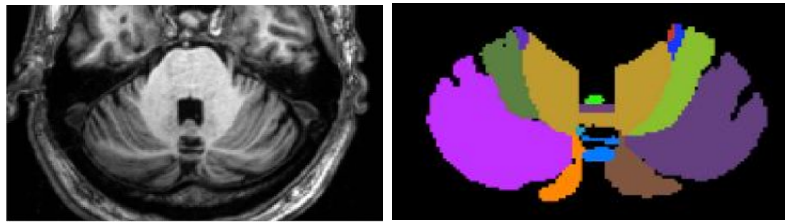
e.g., semantic segmentation

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

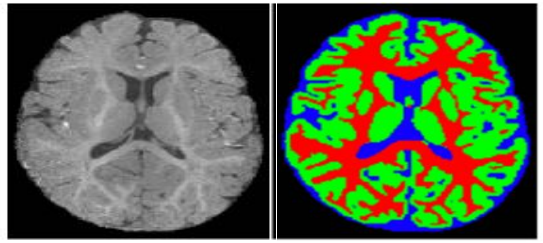
… and medical image analysis



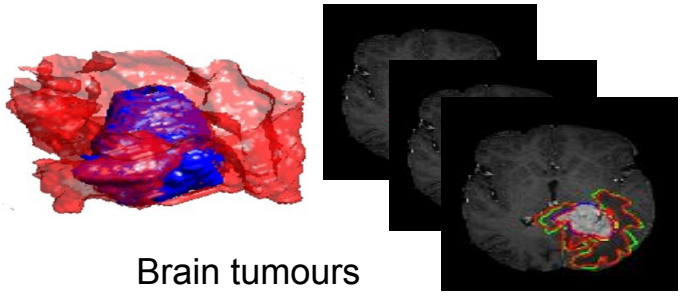
Subcortical structures
(Dolz et al., Neuroimage 2018)



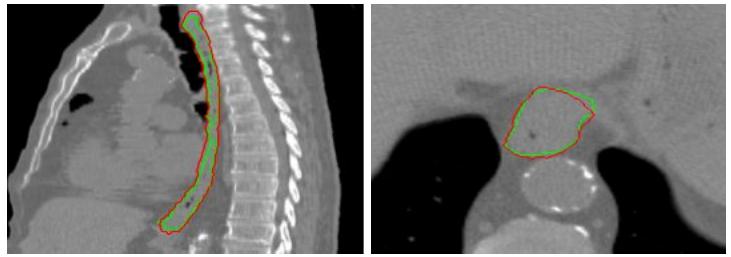
Cerebellum parcellation
(Carass et al., Neuroimage 2018)



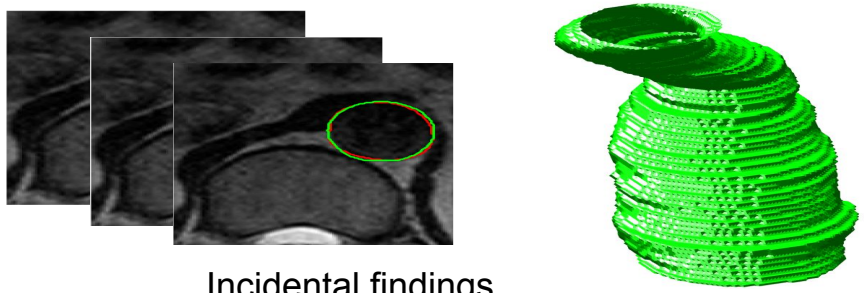
Brain tissues (6-month infant)
(Li et al., TMI 2019)



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)

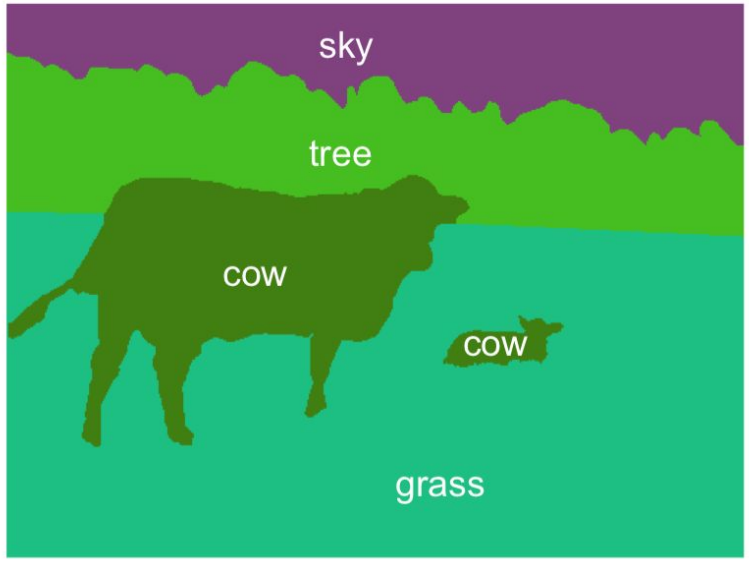
- Bottleneck for learning at large scale

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

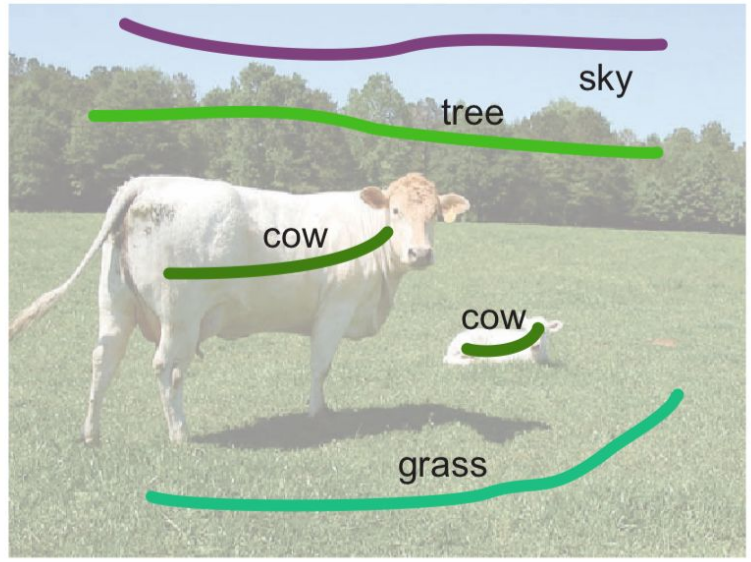
- 1s per label per image
- Scalable for large numbers of labels

person
horse
background

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



Full annotations



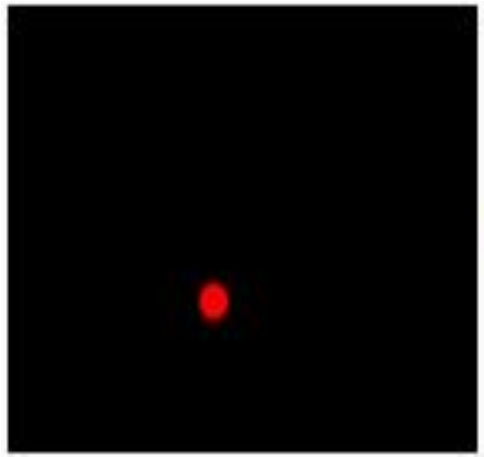
Semi-supervised

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

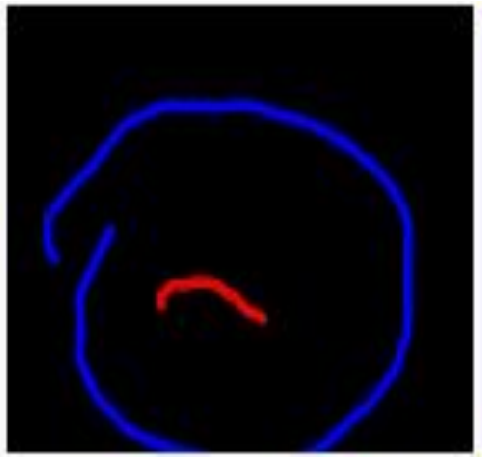
Forms of semi/weak supervision: Examples in segmentation

Car
Parking
Sky
No person

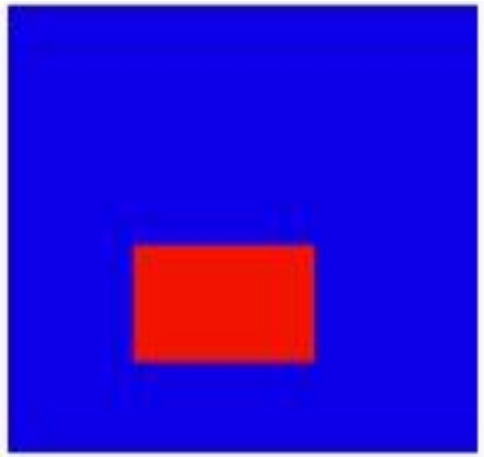
Image tags



points



scribbles

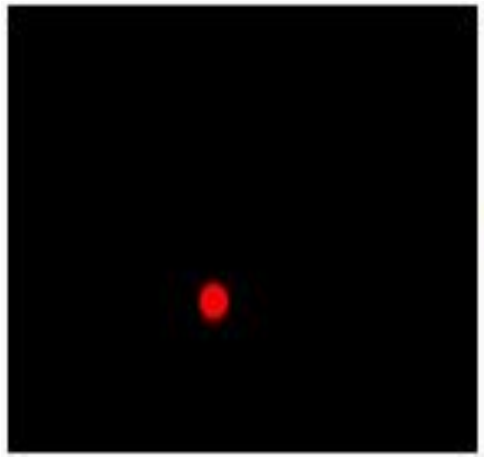


boxes

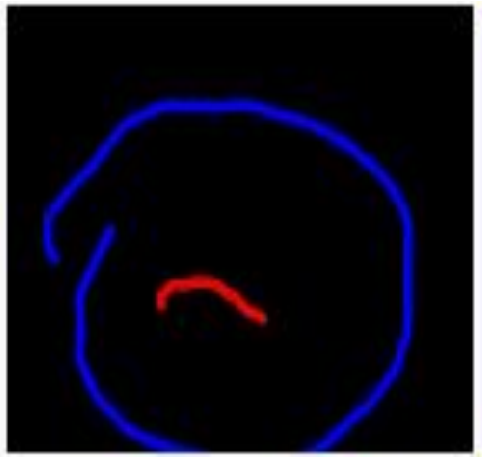
Forms of semi/weak supervision: Examples in segmentation

Car
Parking
Sky
No person

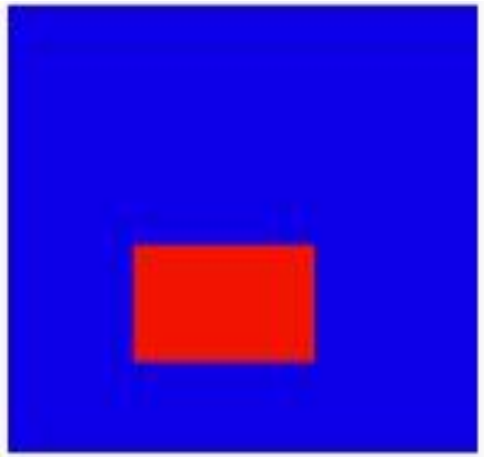
Image tags



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

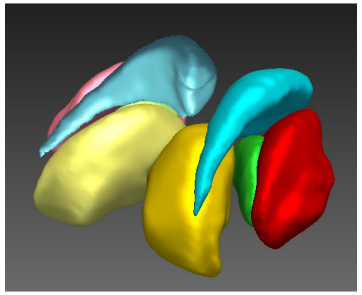
Crowdsourcing?

Full annotations are much more problematic in medical imaging

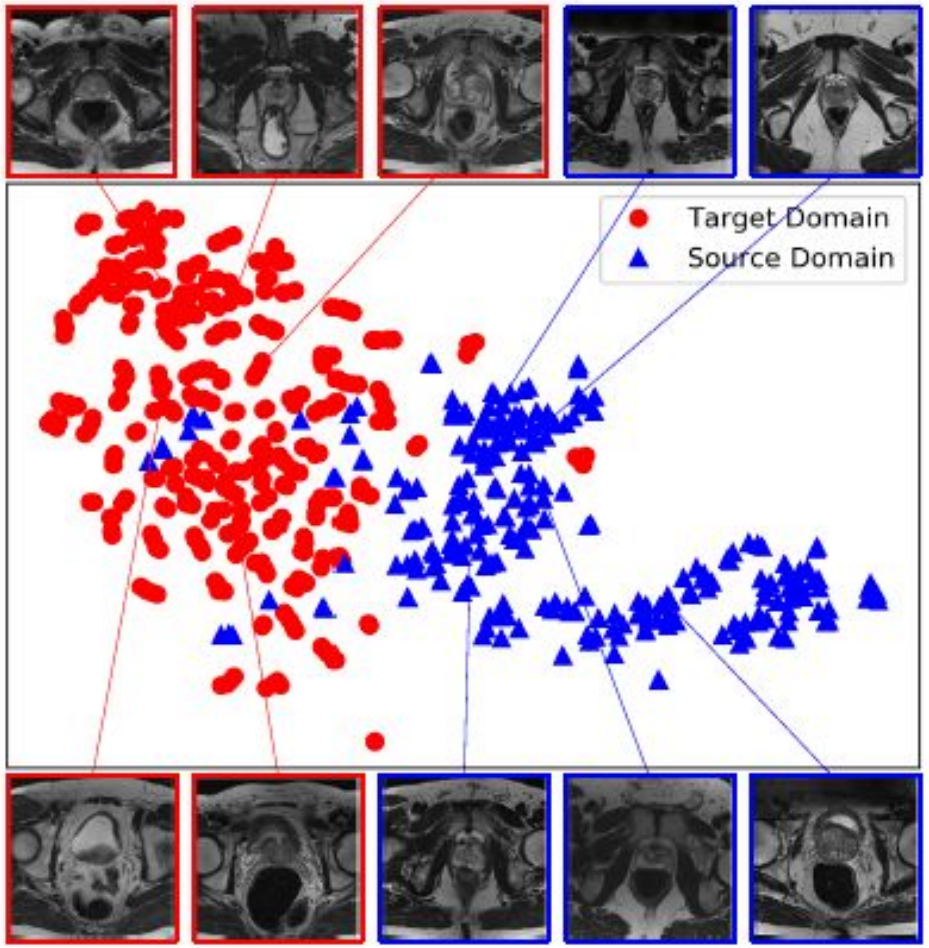
Not anywhere close to the 10k images of Pascal VOC and the 5k of Cityskapes

Crowdsourcing?

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

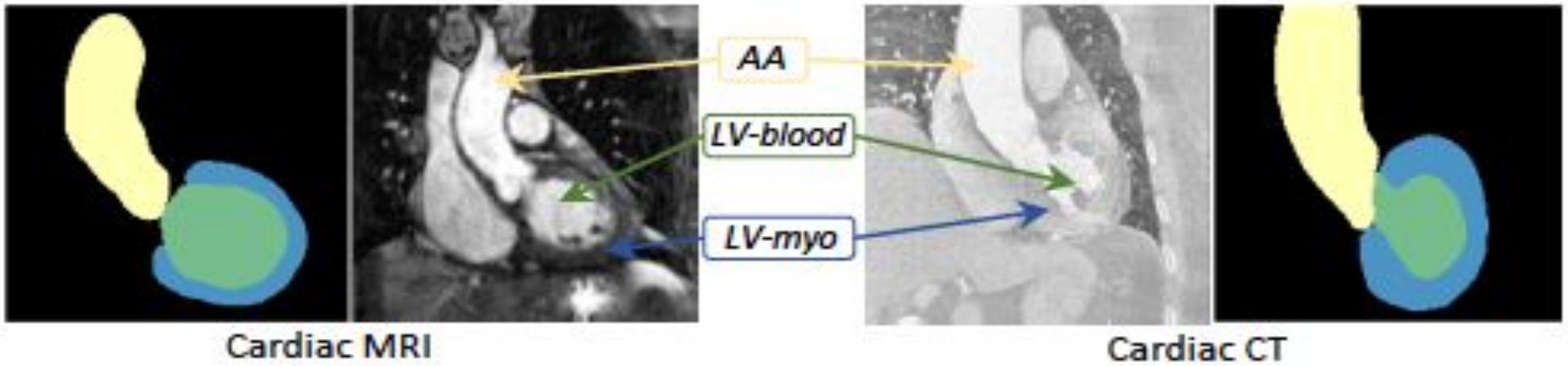


Domain shifts make things worse (even with full annotations in one domain)



[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

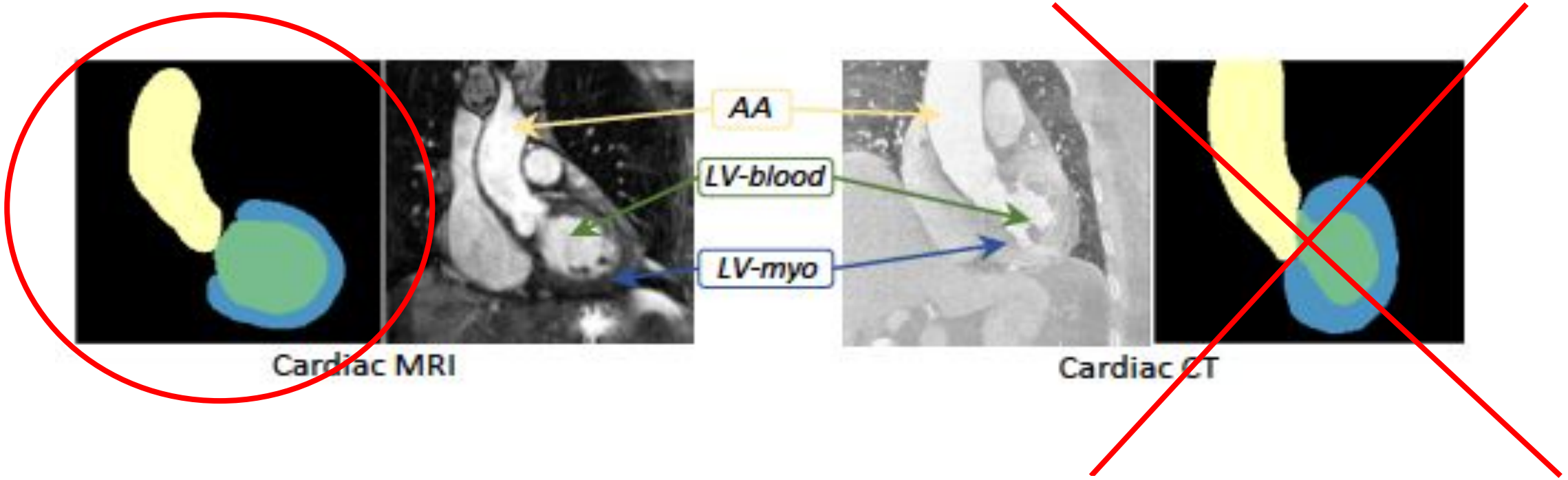


[Images from Dou et al., PnP-AdaNet: Plug-and-play adversarial domain adaptation network with a benchmark at cross-modality cardiac segmentation ArXiv 2018]

Unsupervised domain adaptation

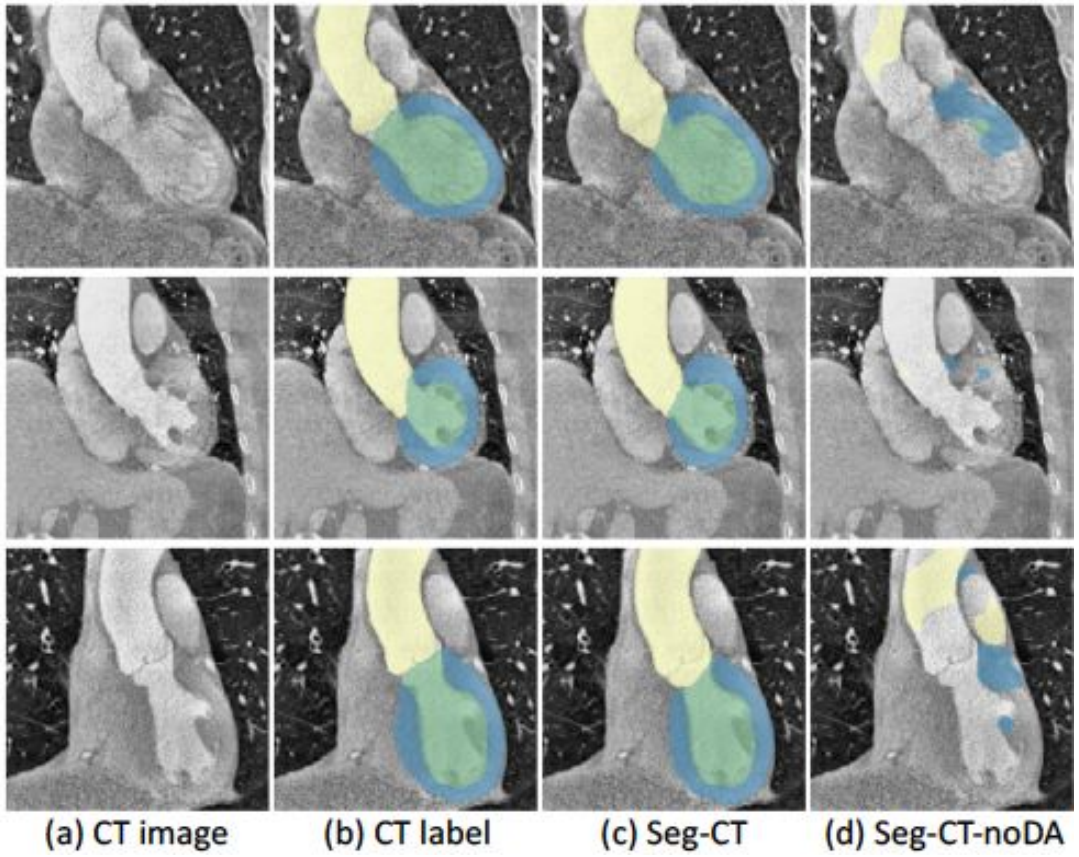
We have labels for the source domain

No labels for the target



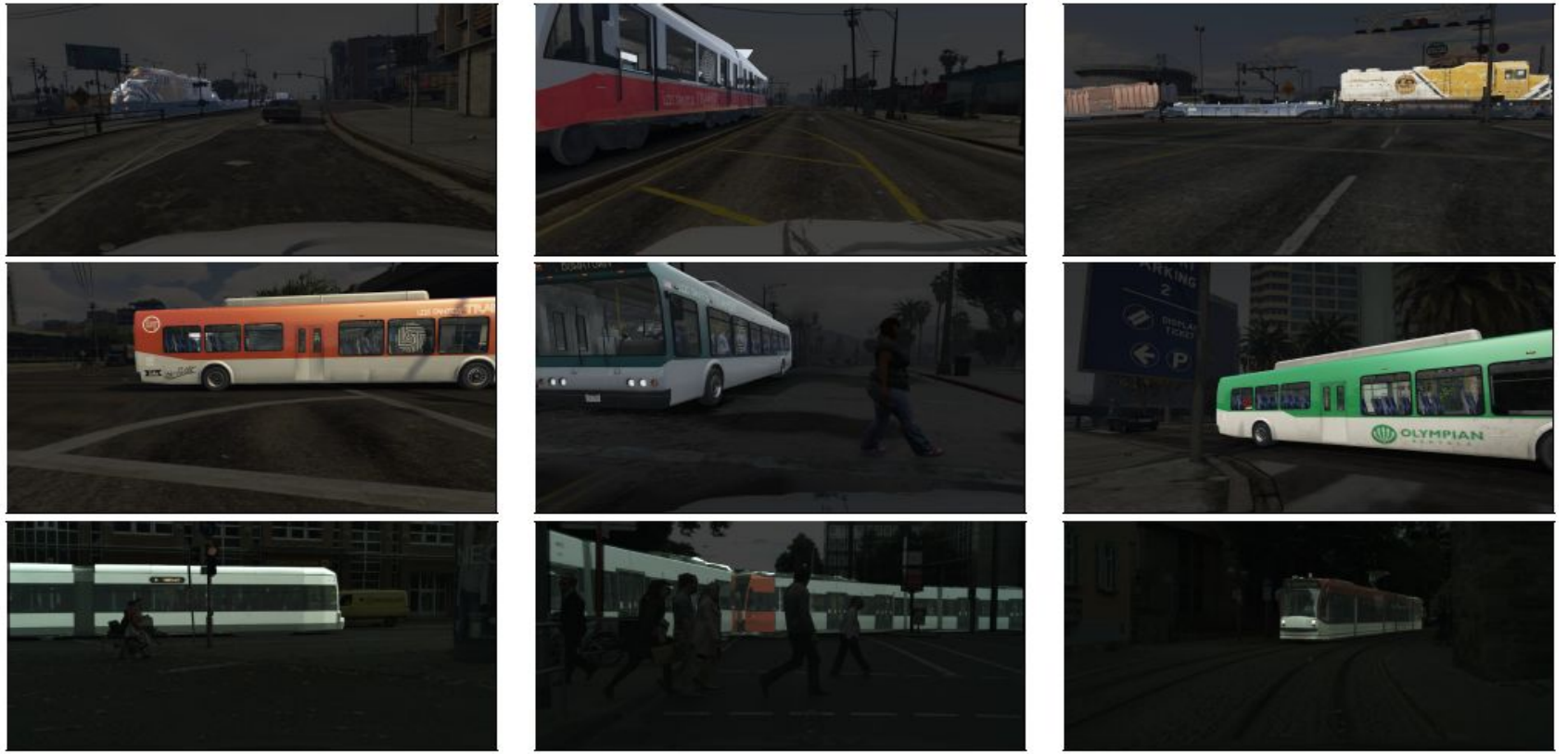
[Images from Dou et al., PnP-AdaNet: Plug-and-play adversarial domain adaptation network with a benchmark at cross-modality cardiac segmentation ArXiv 2018]

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*



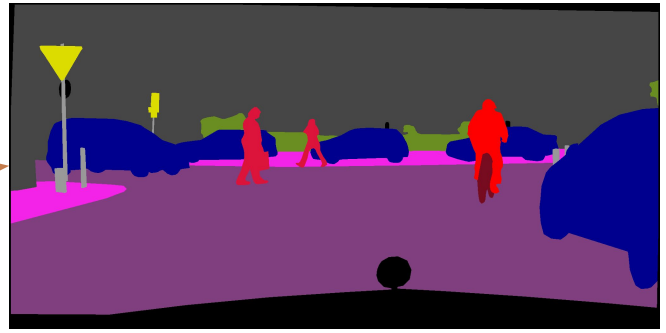
"train"
GTA

"bus"
GTA

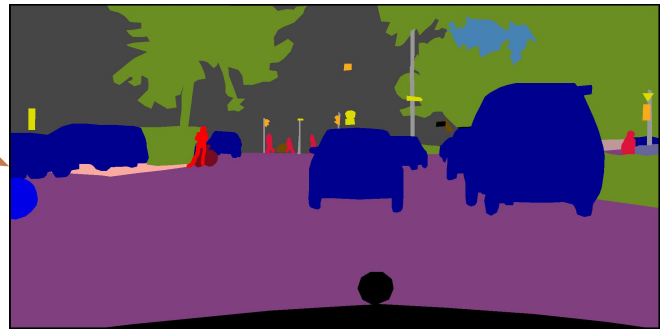
"train"
Cityscapes

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*



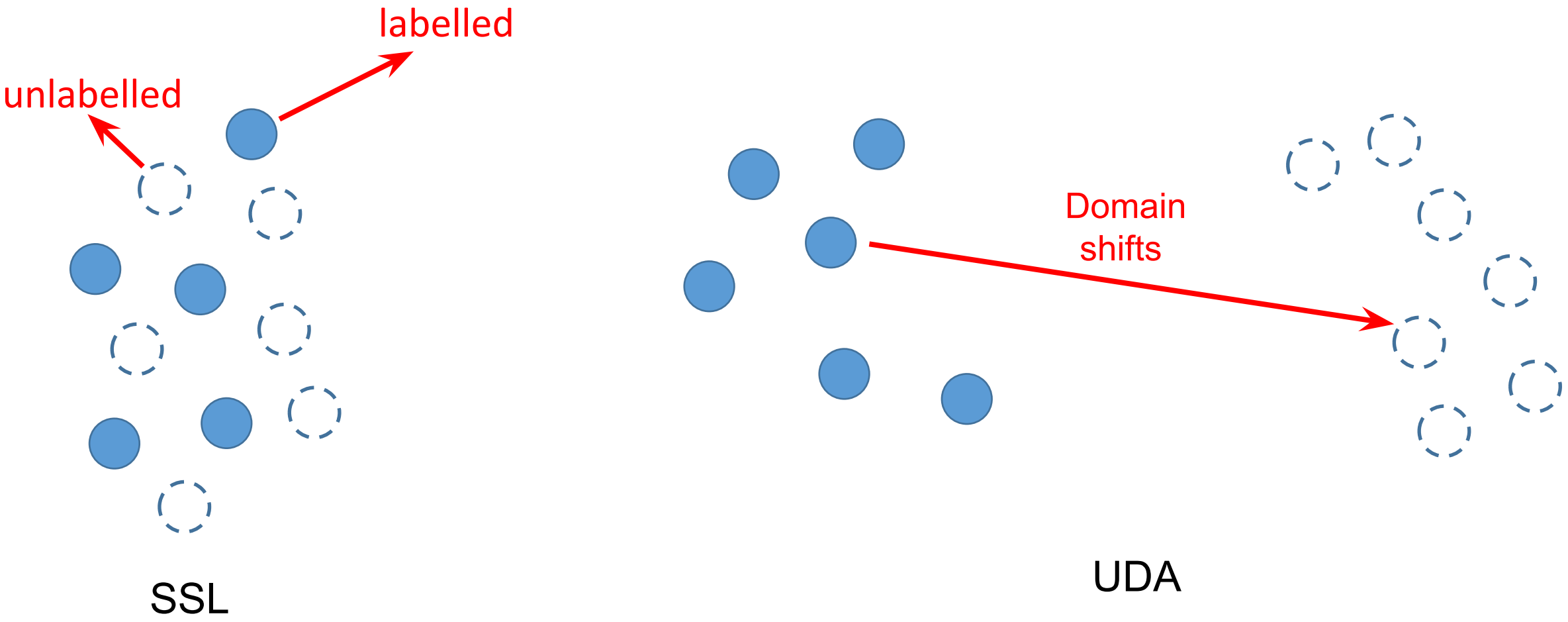
Frankfurt



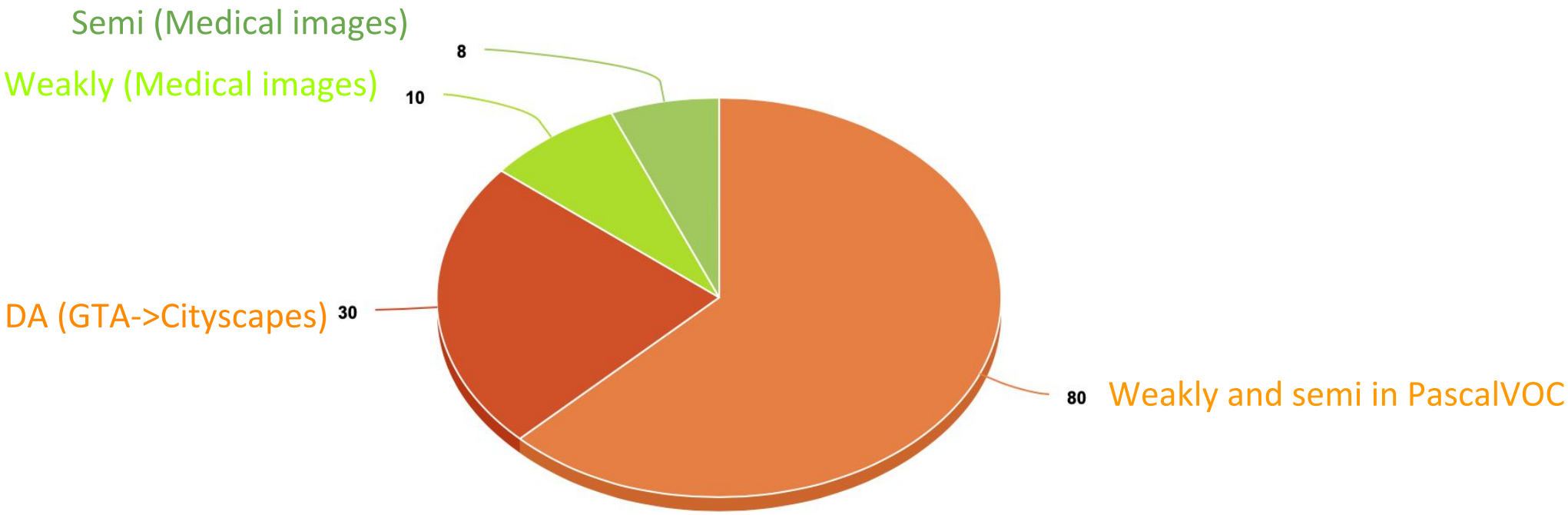
Zurich

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)*

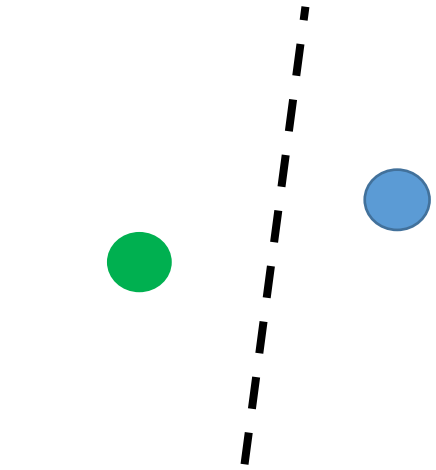
✔ Models and optimization

Part 1

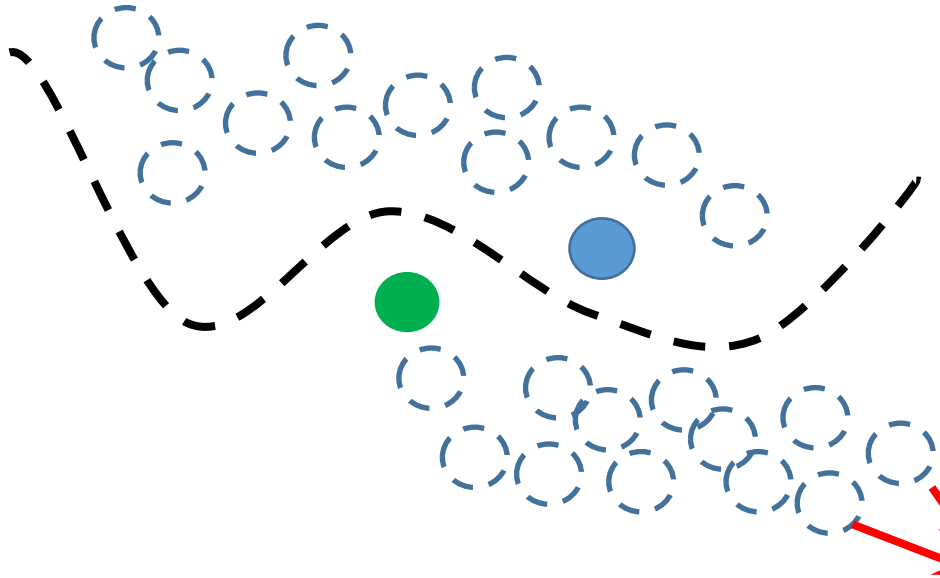
Regularization

Laplacian (and CRFs)

Semi-supervised learning (general form)

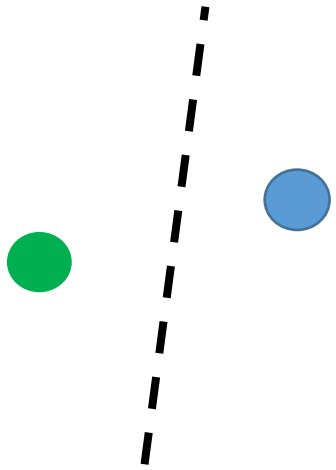


Only labeled points

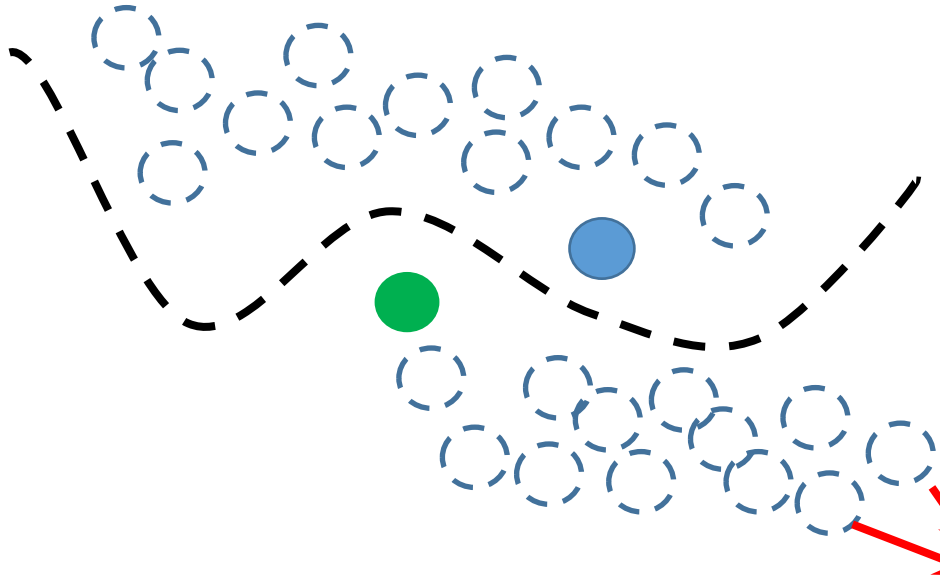


Unlabeled points

Semi-supervised learning (general form)



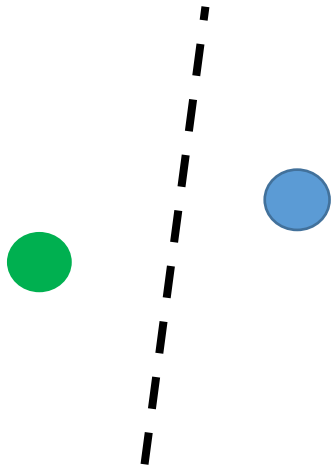
Only labeled points



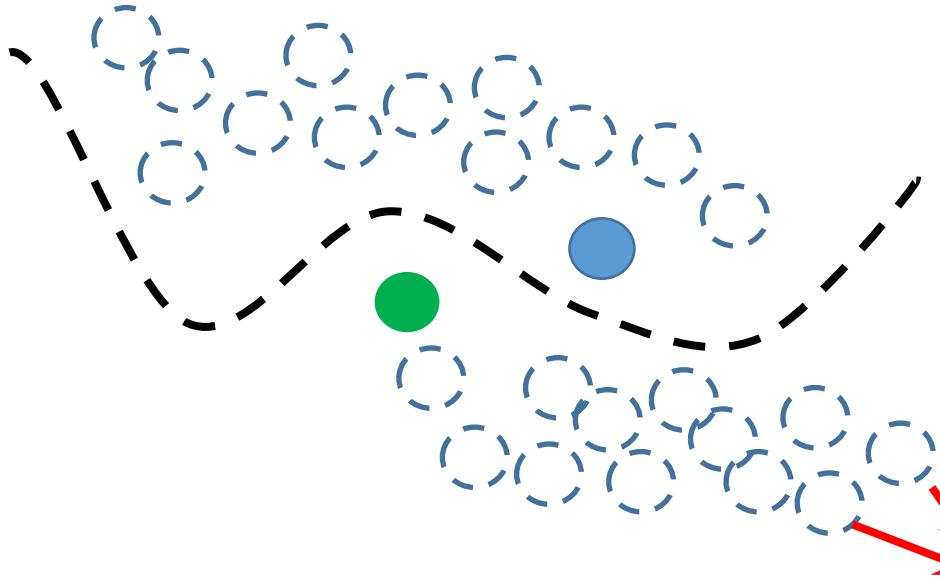
Unlabeled points

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

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

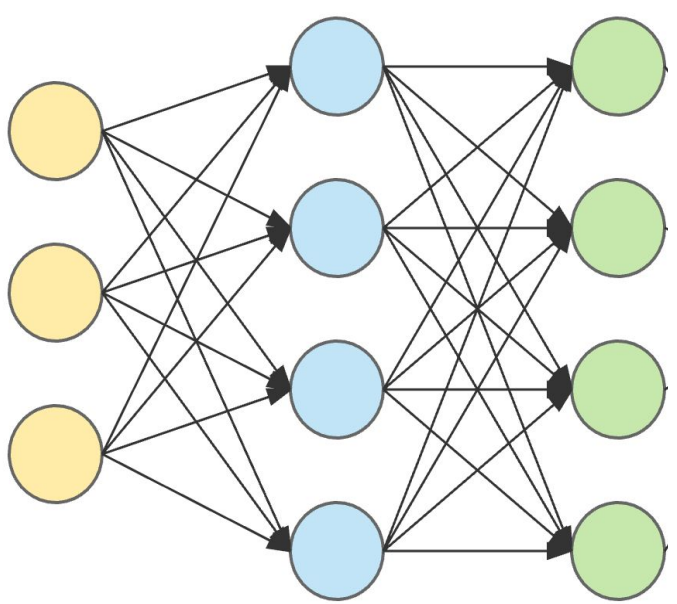
Semi-supervised learning (general form)

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

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**

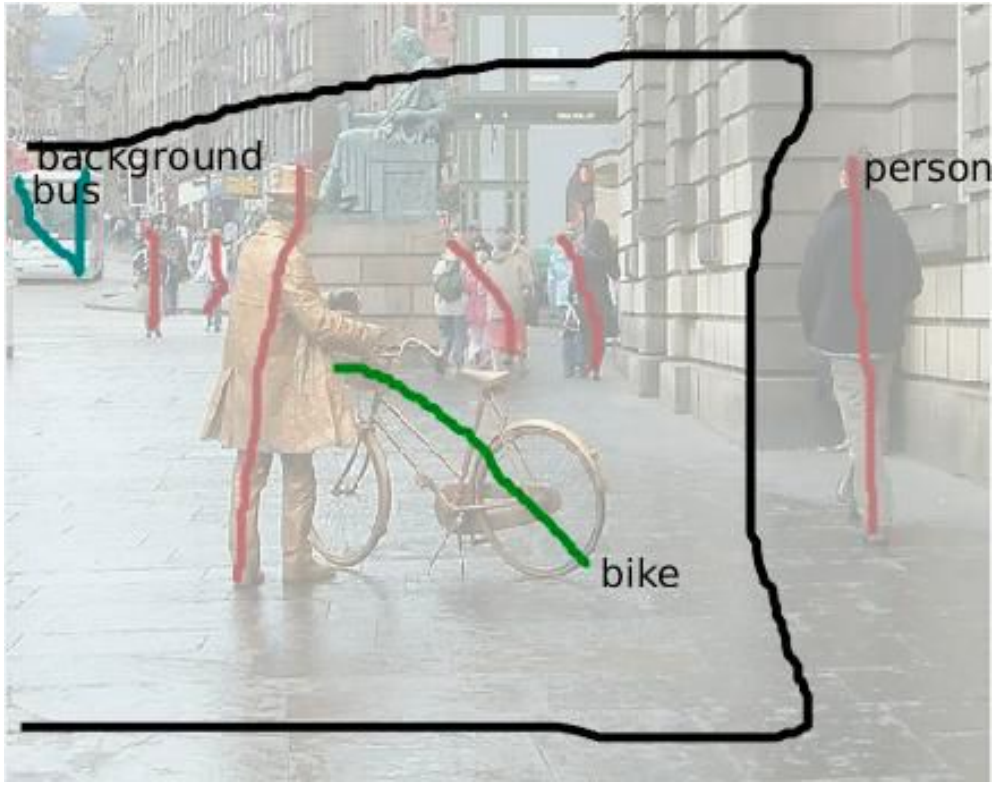
$$w_{p,q} \| \mathbf{f}_\theta^p - \mathbf{f}_\theta^q \|^2$$

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

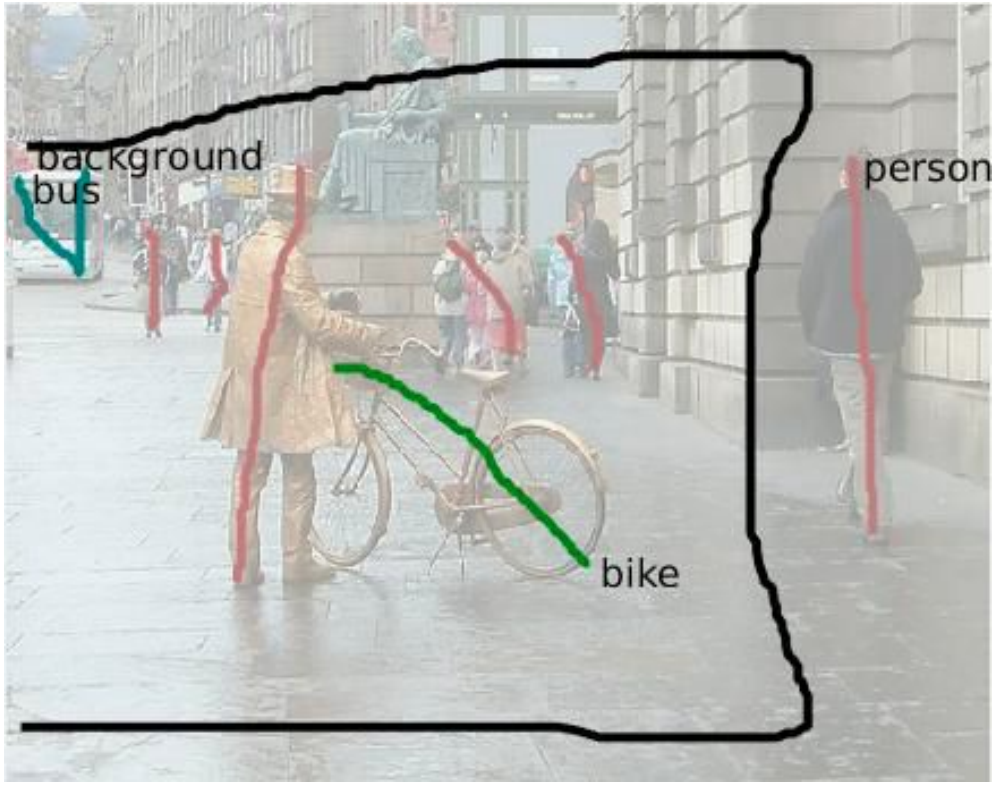
$$\min_{\theta} \sum_{p \in \mathcal{L}} l(\mathbf{y}^p, \mathbf{s}_\theta^p) + \sum_{p,q \in \mathcal{L} \cup \mathcal{U}} w_{p,q} \|\mathbf{s}_\theta^p - \mathbf{s}_\theta^q\|^2$$



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

Semi-supervision loss for segmentation

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



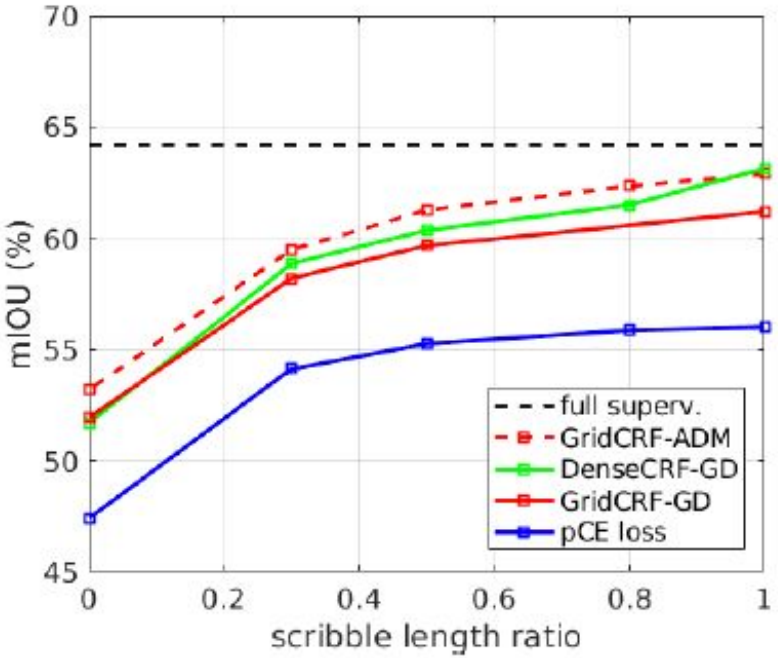
On the vertices of the simplex (binary variables), this is exactly the Potts model in Conditional Random Fields (e.g., Dense CRFs)!

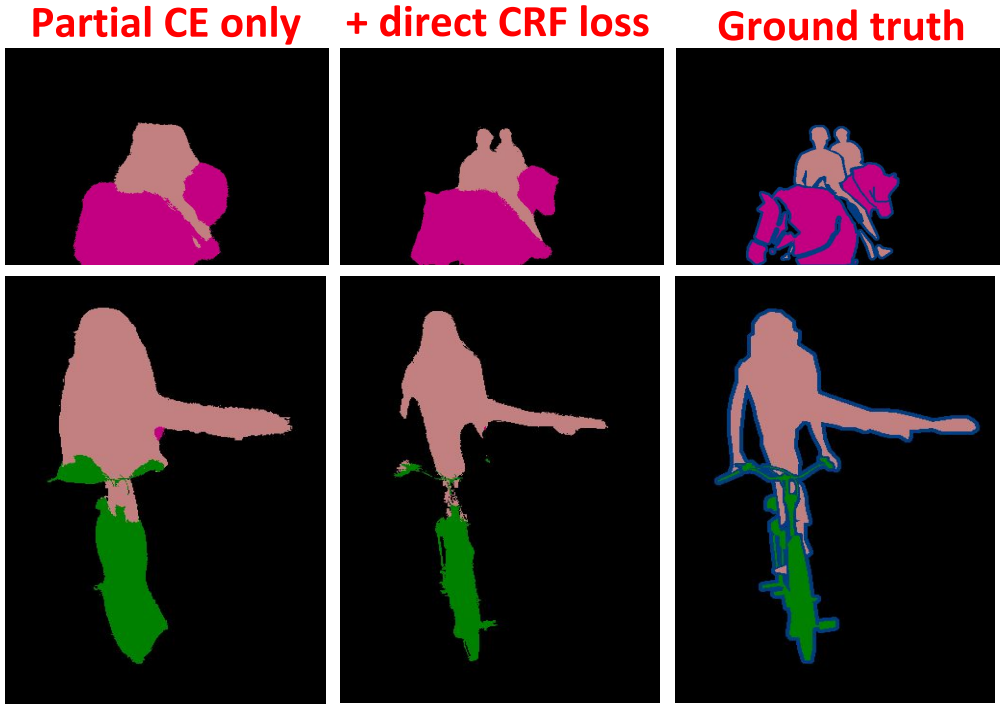
[Tang et al., On regularized losses for weakly supervised segmentation, ECCV 2018]

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

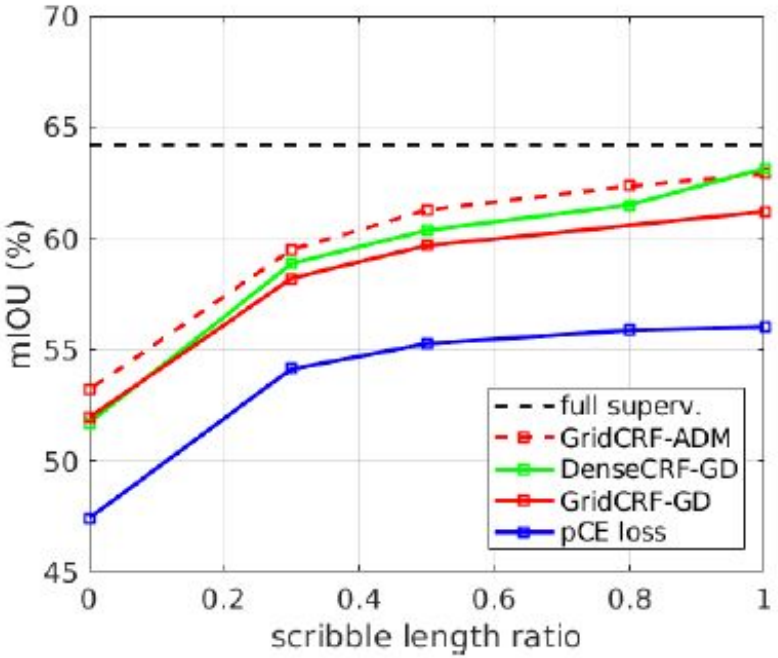




[Tang et al., On regularized losses for weakly supervised segmentation, ECCV 2018]
[Marin et al., Beyond gradient descent for regularized segmentation losses, CVPR 2019]

Semi-supervision loss for segmentation

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



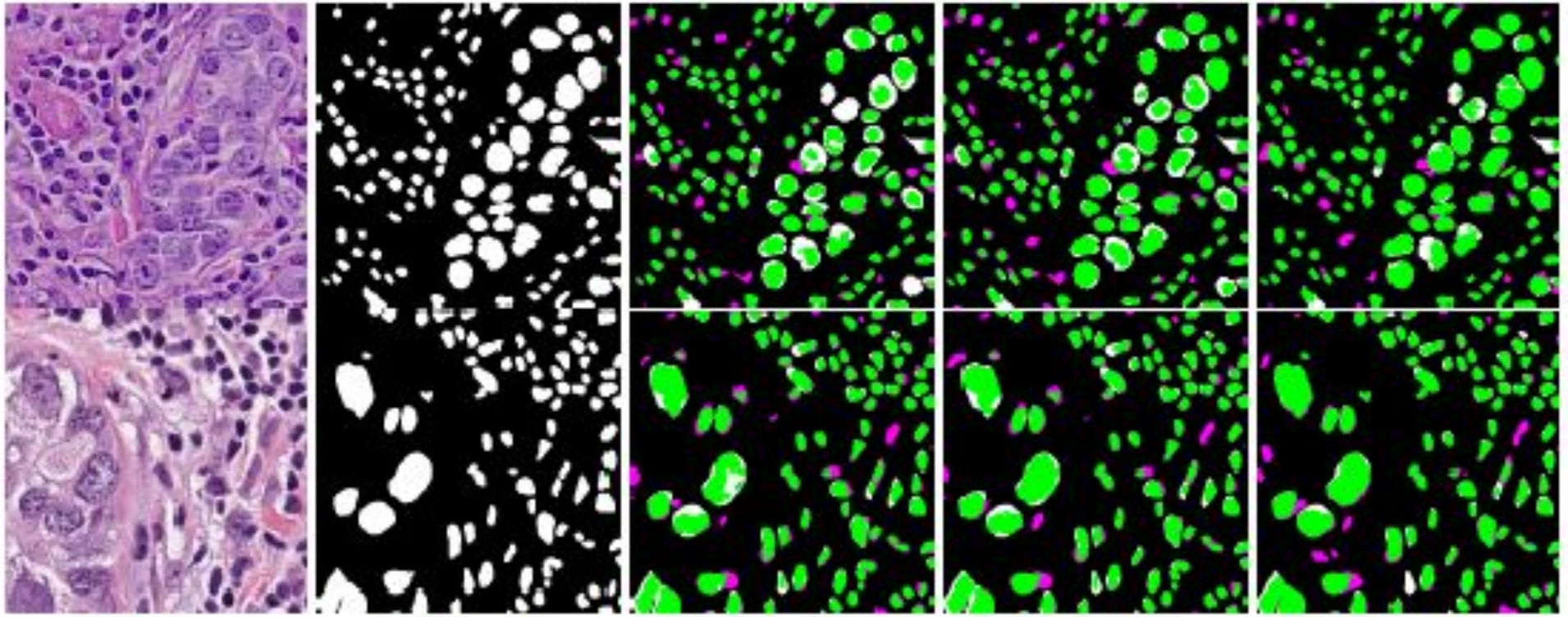
The exciting part in this plot:

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]

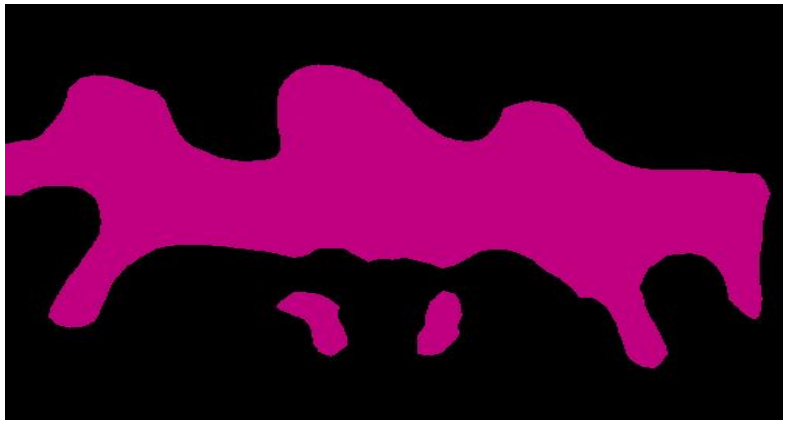
- Ji et al., Scribble-Based Hierarchical Weakly Supervised Learning for Brain Tumor Segmentation, MICCAI 2019 [Brain tumor images, scribble annotations]

Regularization

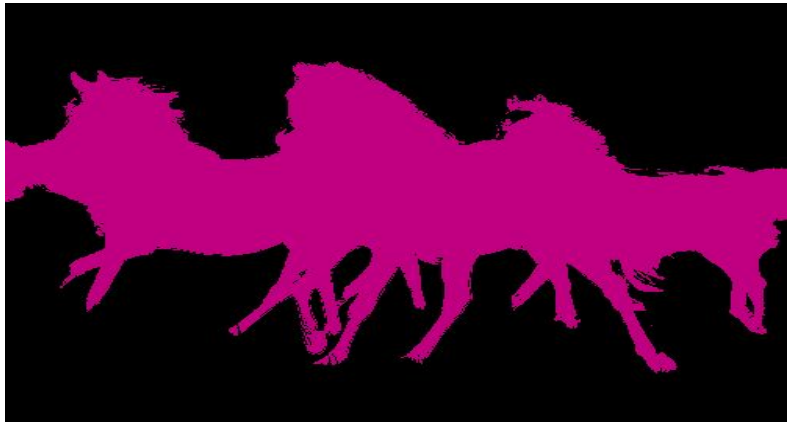
Optimization matters and the choice of an optimizer depends on the form of your loss (SGD is not not your only choice)

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

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



CNN
(fully supervised)

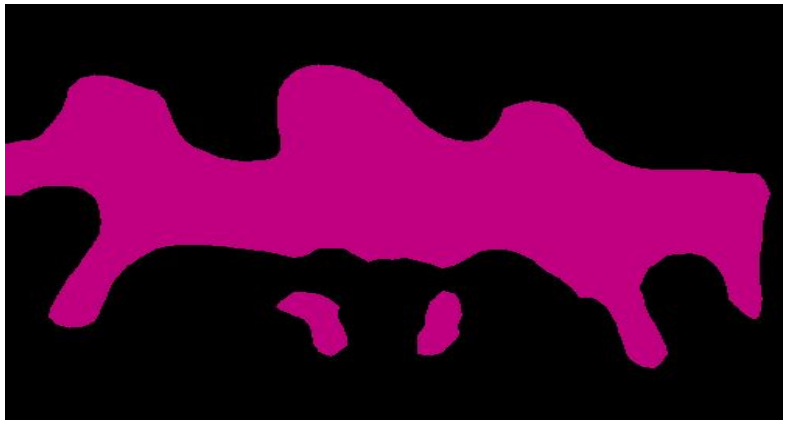


CNN
+
CRF (post-processing)

[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)

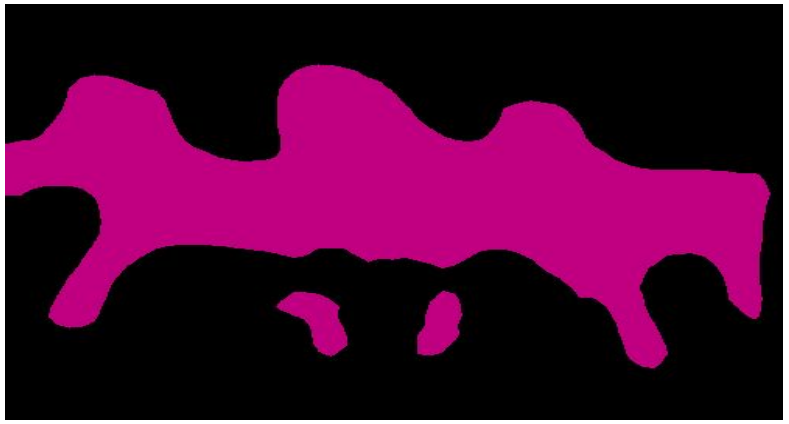
[Chen et al., Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected CRFs. TPAMI 2018]

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)

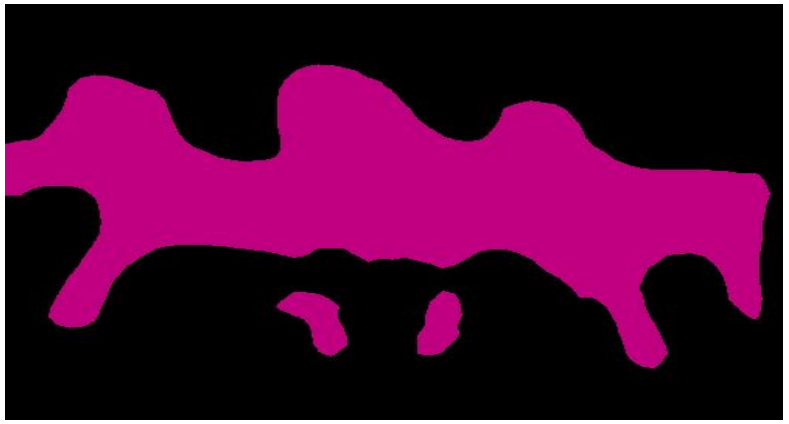
[Chen et al., Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected CRFs. TPAMI 2018]

CRFs meet fully supervised CNNS

Pairwise potentials (the very popular *Potts*)

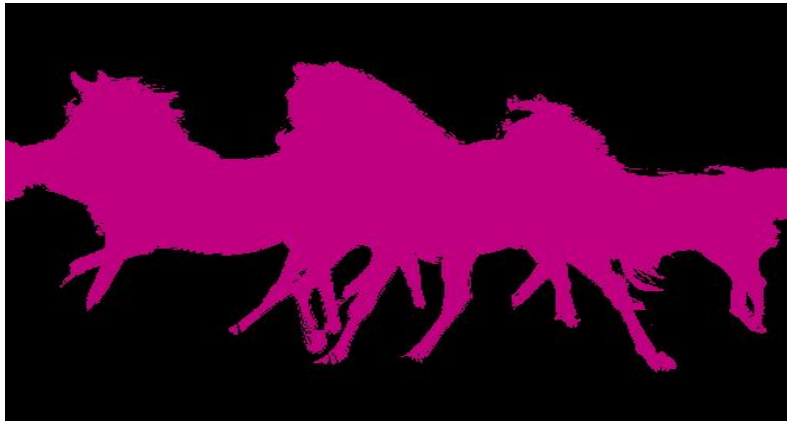
$$\min_{Y=[\mathbf{y}^1,\dots\mathbf{y}^{|\Omega|}]} \sum_{p\in\Omega} l(\mathbf{y}^p, \mathbf{s}_\theta^p) + \sum_{p,q\in\Omega^2} w_{pq}[\mathbf{y}^p \neq \mathbf{y}^q]$$

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

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

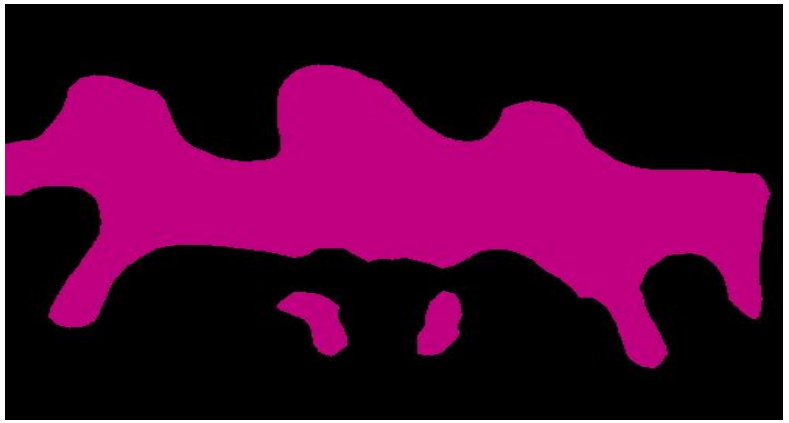
[Chen et al., Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected CRFs. TPAMI 2018]

CRFs meet fully supervised CNNS

Pairwise potentials (also the very popular *DenseCRF*)

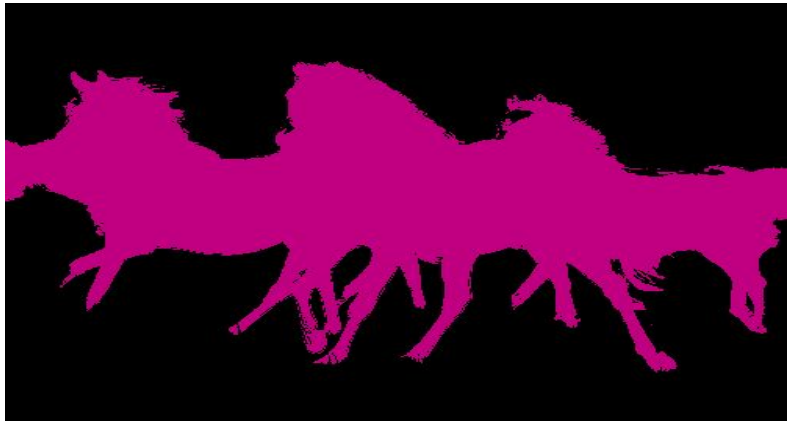
$$\min_{Y=[\mathbf{y}^1,\dots\mathbf{y}^{|\Omega|}]} \sum_{p\in\Omega} l(\mathbf{y}^p, \mathbf{s}_\theta^p) + \sum_{p,q\in\Omega^2} w_{pq}[\mathbf{y}^p \neq \mathbf{y}^q]$$

$\mathbf{x}^p \in \mathbb{R}^N$ (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]

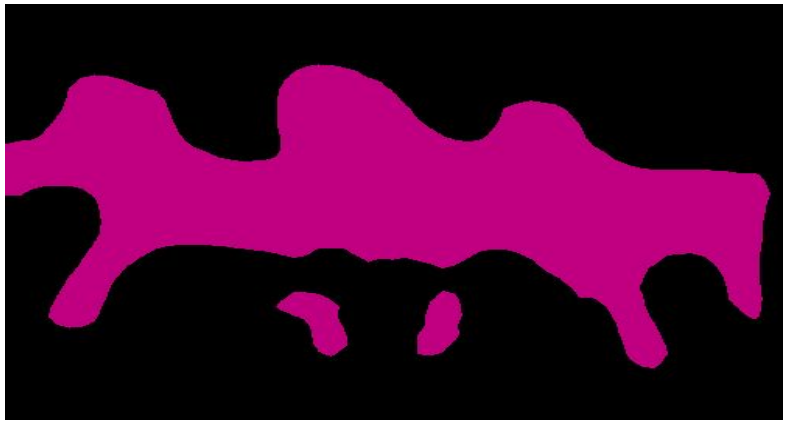
CRFs meet fully supervised CNNS

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

For one-hot encoding vectors

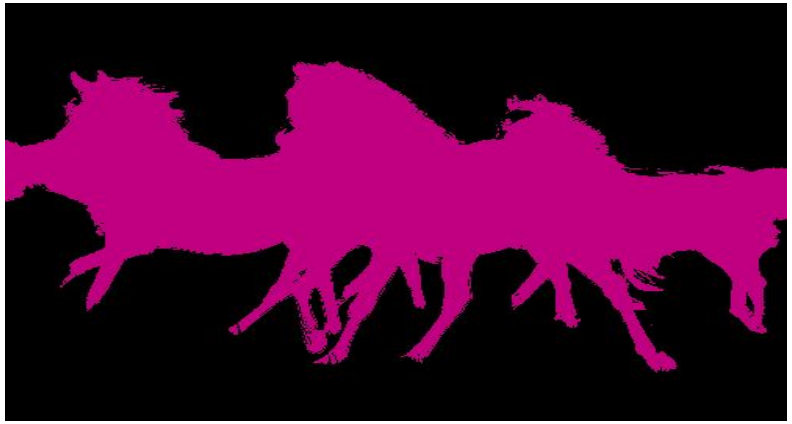
$$\min_{Y=[\mathbf{y}^1,\dots\mathbf{y}^{|\Omega|}]} \sum_{p\in\Omega} l(\mathbf{y}^p, \mathbf{s}_\theta^p) + \sum_{p,q\in\Omega^2} w_{pq}[\mathbf{y}^p \neq \mathbf{y}^q]$$

$\mathbf{x}^p \in \mathbb{R}^N$ (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 8000 citations, test-of-time award)

✔ **Mean-field approximation**:
Krahenbuhl and Koltun, NIPS'11 (~1500 citations)

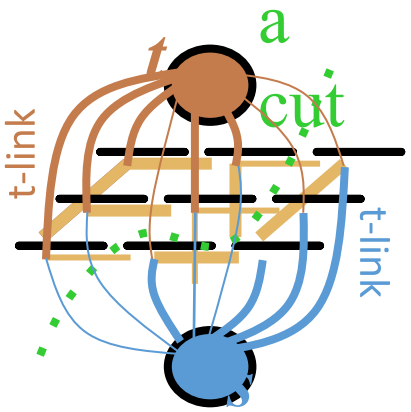
Graph cuts *vs.* first-order methods

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

$g_{pq}(y_p, y_q)$

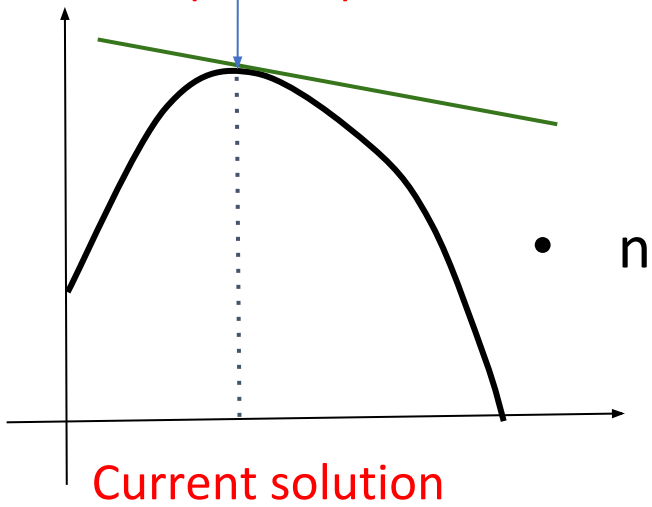
submodular

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



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

Linear approximation (bound)

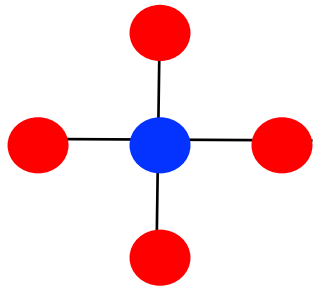


- no optimality guarantee

Graph cuts *vs.* first-order methods

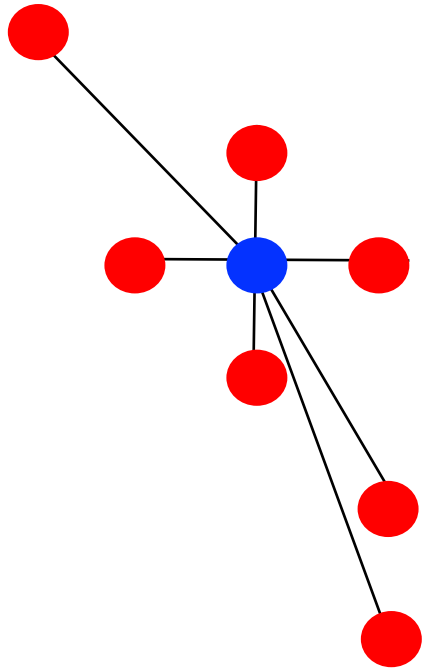
$$\sum_{p,q \in \mathcal{N}} w_{pq} \|\mathbf{y}^p - \mathbf{y}^q\|^2$$

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)

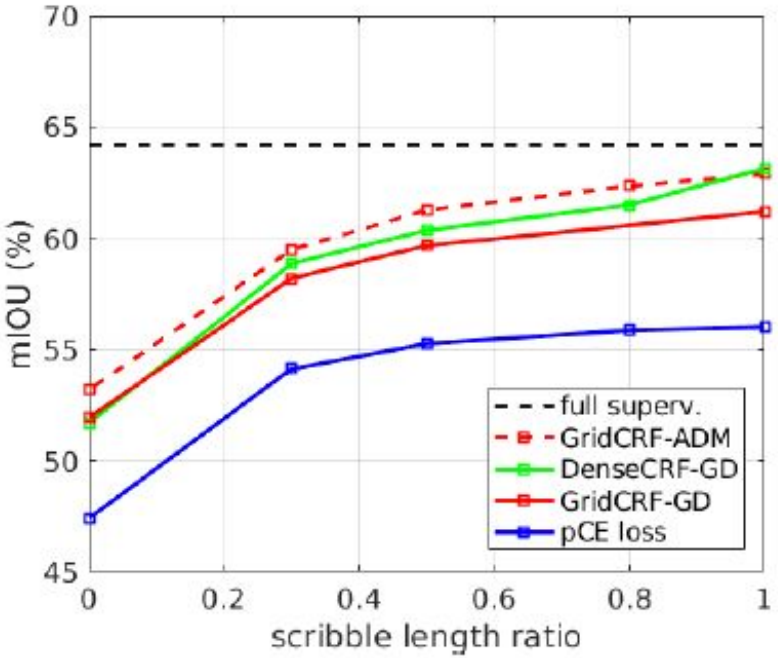
Grid CRF + graph cut

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

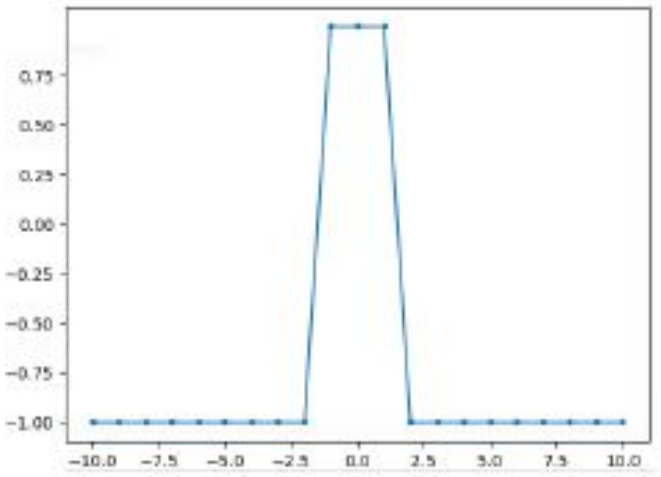
[Tang et al., On regularized losses for weakly supervised segmentation, ECCV 2018]

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

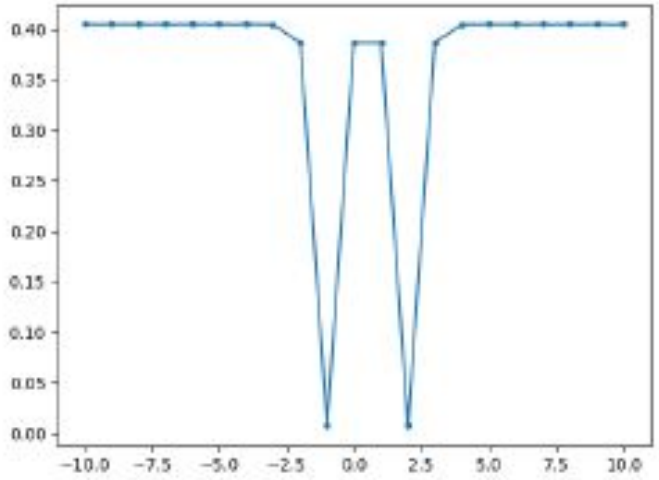
Dense CRF is **NOT** supposed to be better, but…

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

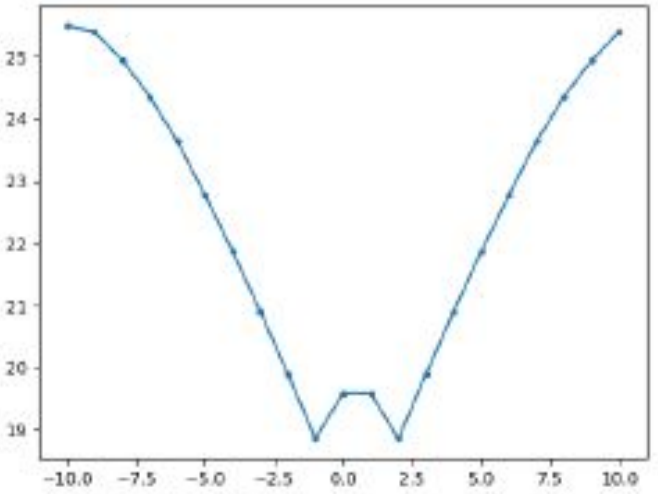
$I(x)$



Grid CRF term



Dense CRF term

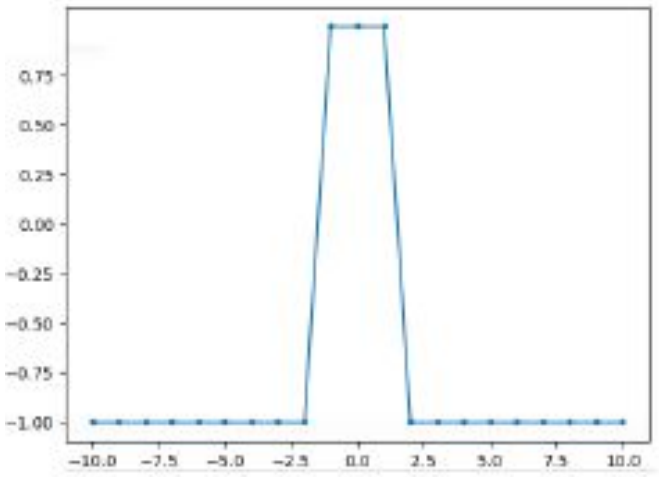


[Marin et al., Beyond gradient descent for regularized segmentation losses, CVPR 2019]

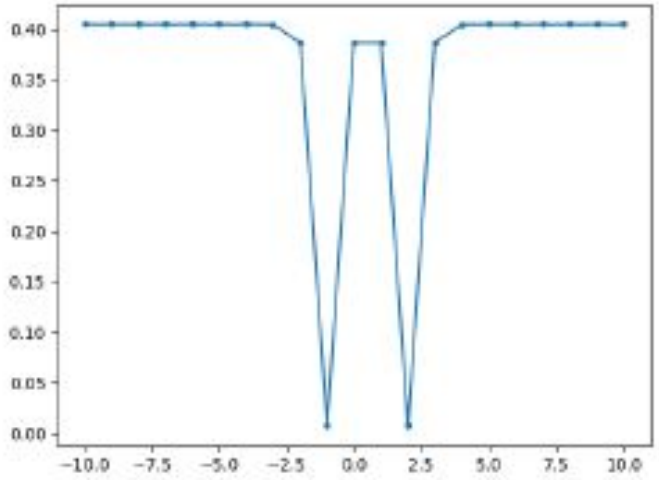
Dense CRF is **NOT** supposed to be better, but…

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

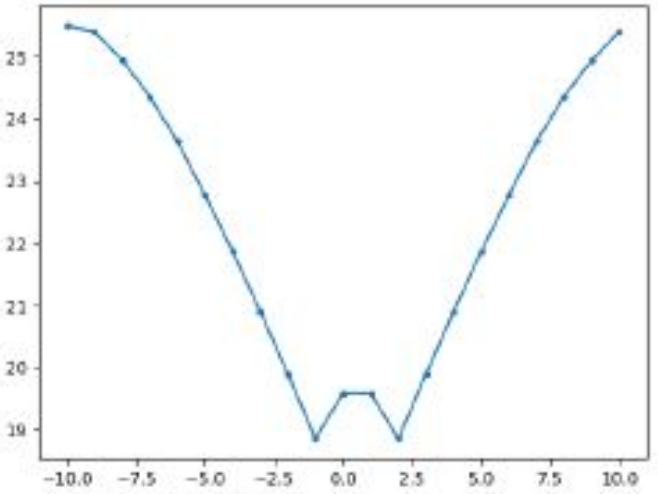
$I(x)$



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:

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

$$L(S_\theta, Y) \in \{0,1\}^{C \times |\mathcal{L}|}$$

A few labeled points

$$R(S_\theta) \in \{0,1\}^{C \times |\Omega|}$$

All data points

Notation: $\Omega = \mathcal{L} \cup \mathcal{U}$

[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

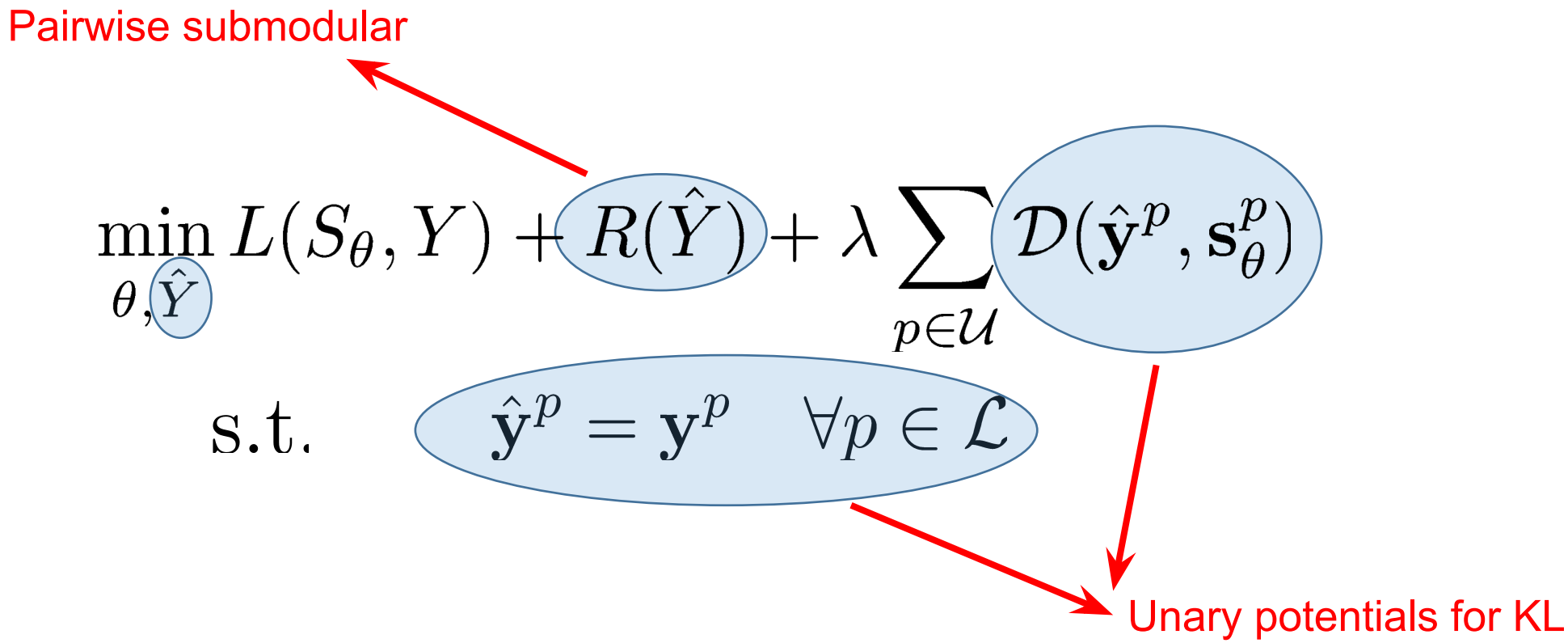
$$\min_{\theta, \hat{Y}} L(S_\theta, Y) + R(\hat{Y}) + \lambda \sum_{p \in \mathcal{U}} \mathcal{D}(\hat{\mathbf{y}}^p, \mathbf{s}_\theta^p)$$

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

[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):

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

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

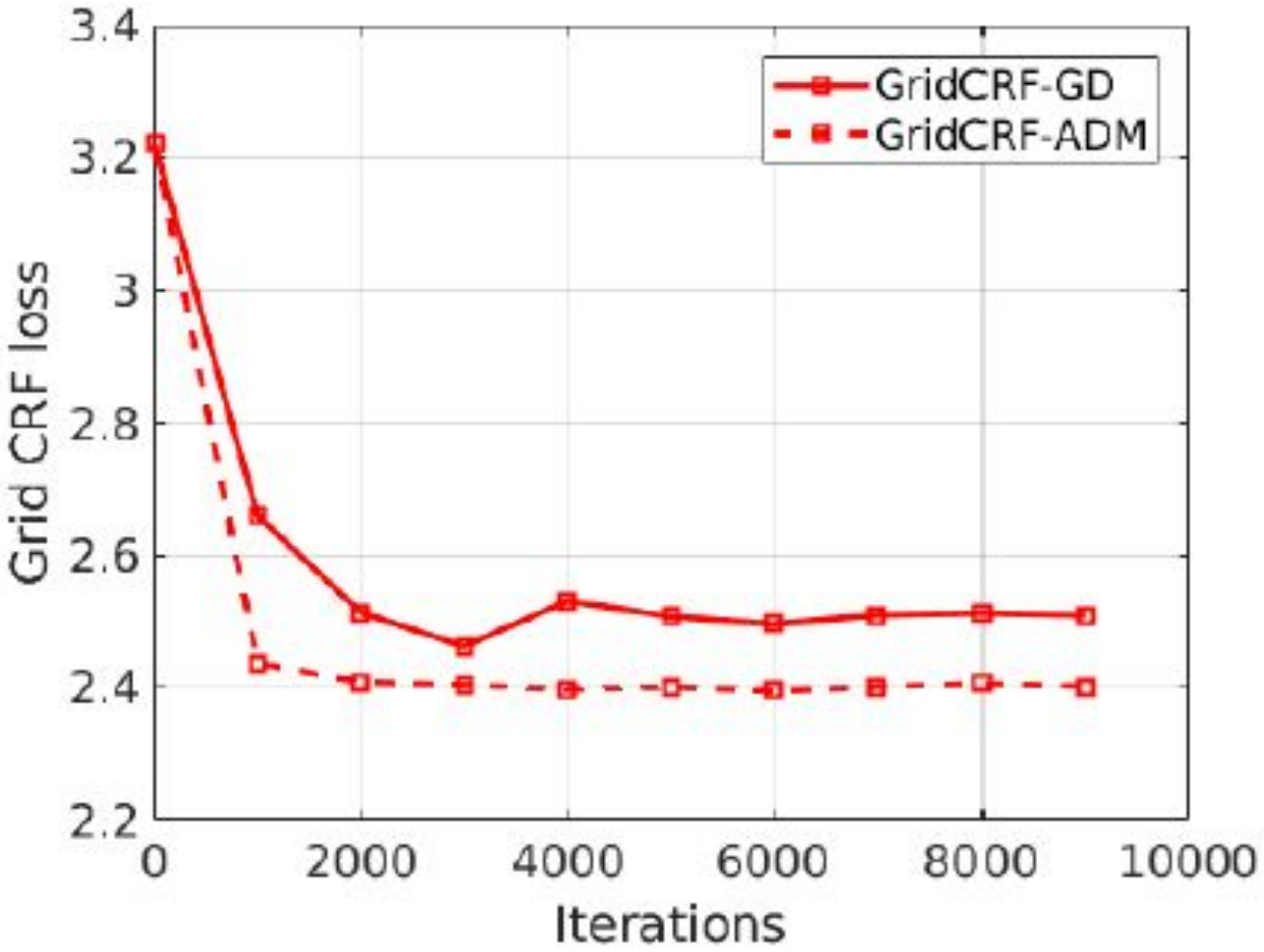
(Typically L_2 in ADMM)

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

(Penalty method)

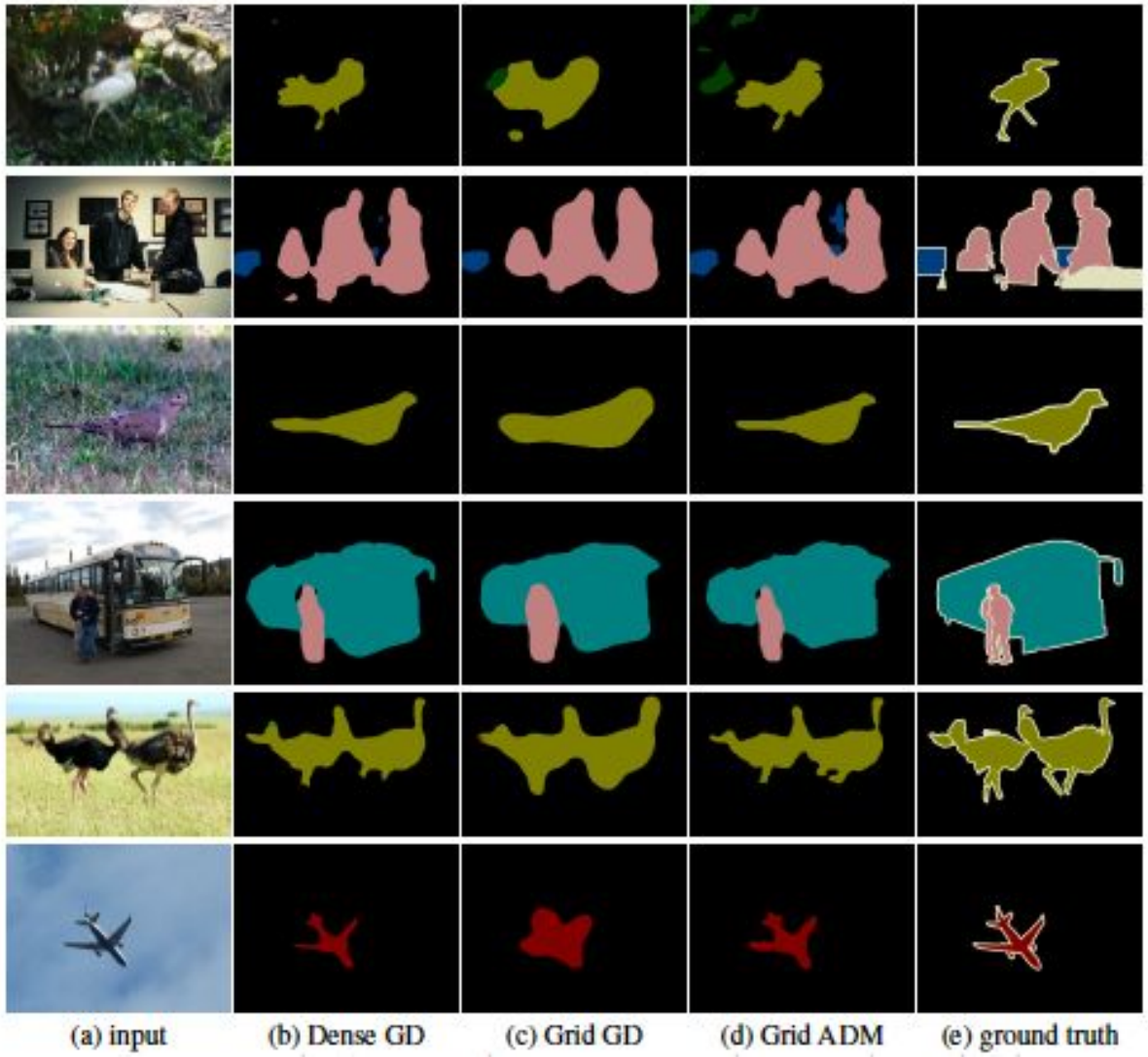
[Boyd et al., Distributed optim. and stat. learning via ADMM, Foundation and trends in ML 2010]

The optimizer matters: Beyond gradient descent



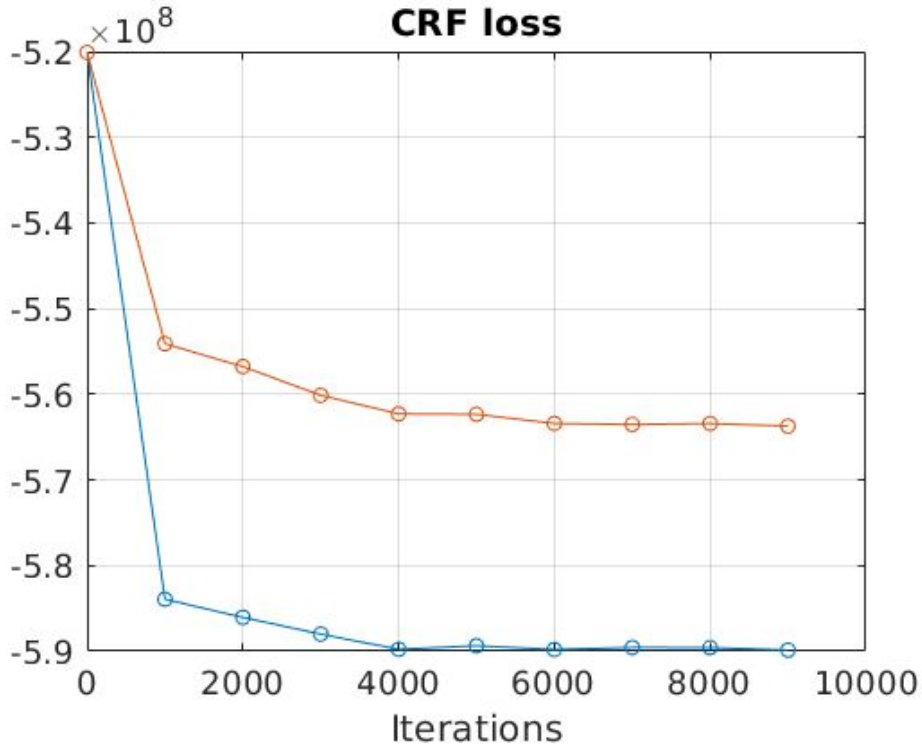
[Marin et al., Beyond gradient descent for regularized segmentation losses, CVPR 2019]

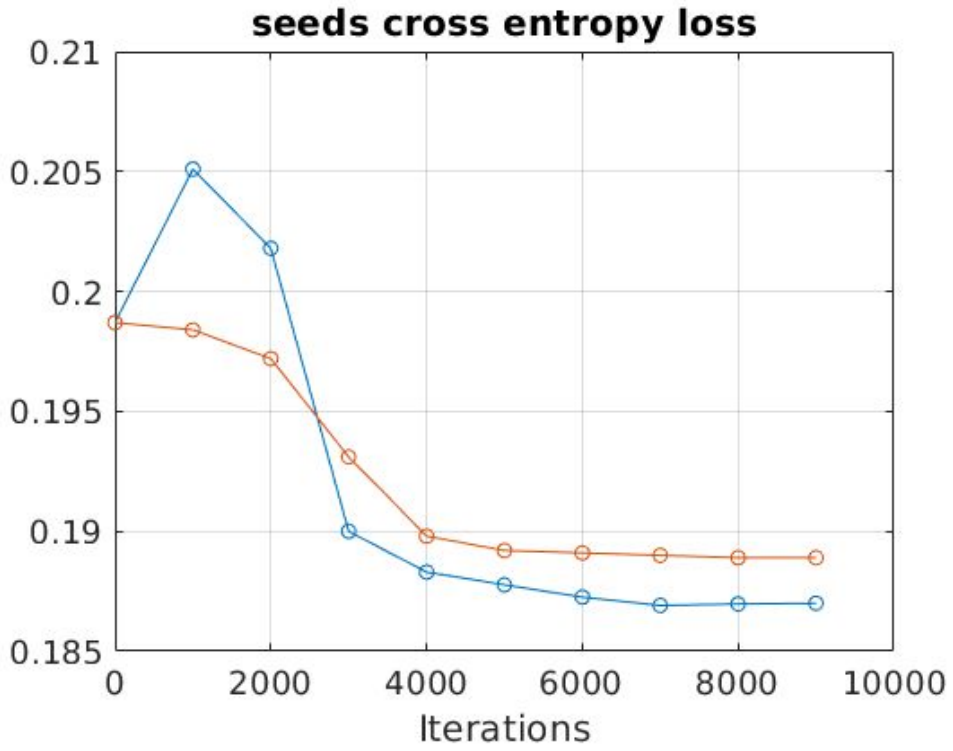
Some visual examples)



[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

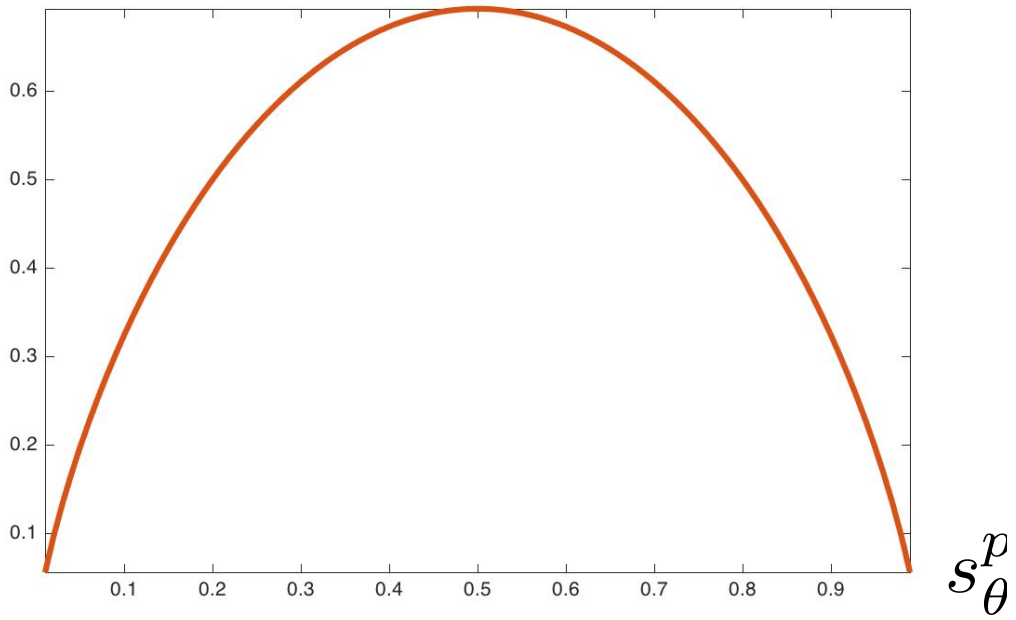
$$\min_{\theta} -\sum_{p\in\mathcal{L}}\sum_{c=1}^{C} y^{p,c}\log s_\theta^{p,c} \;\; -\sum_{p\in\mathcal{U}}\sum_{c=1}^{C} s_\theta^{p,c}\log s_\theta^{p,c}$$

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)

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

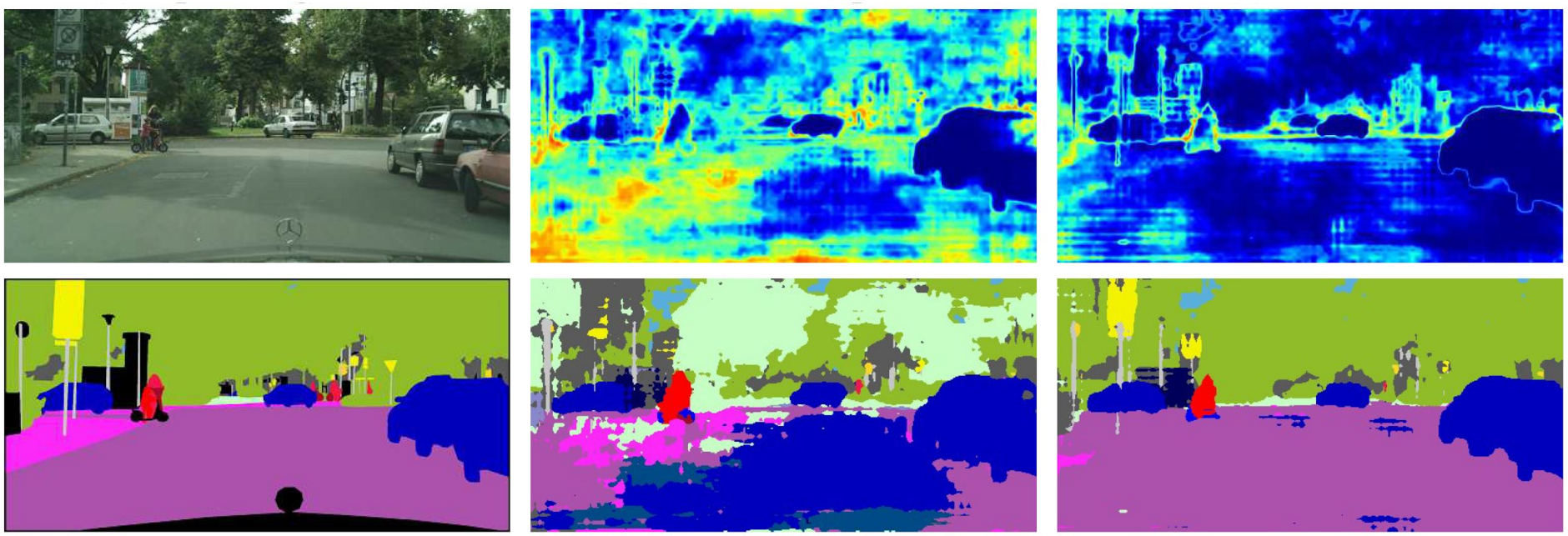


Entropy minimization for UDA

Input image + GT

Without adaptation

Entropy minimization



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

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

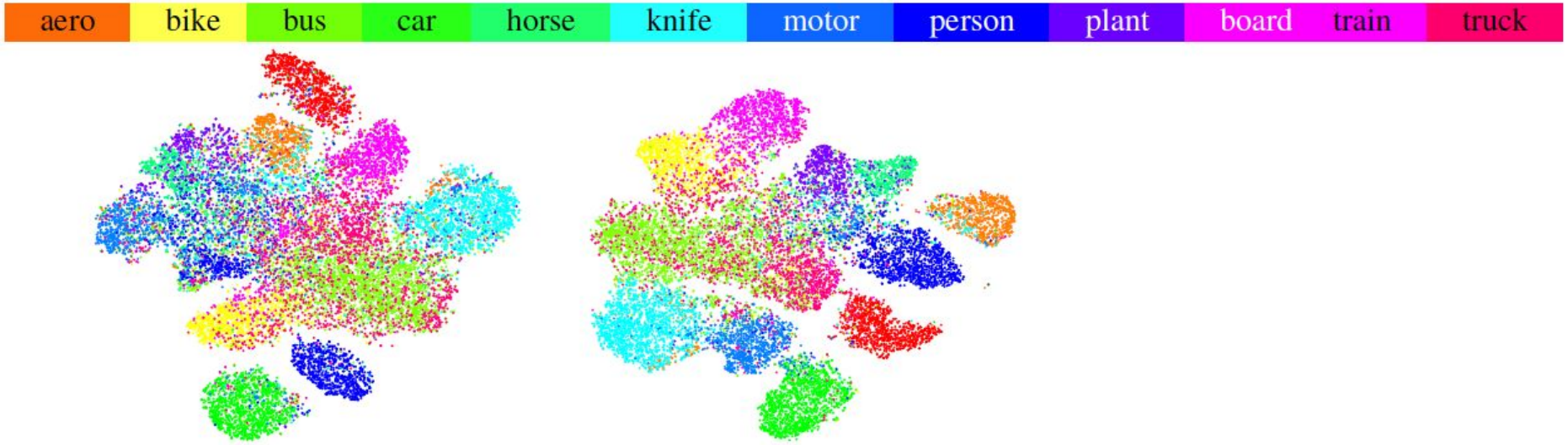
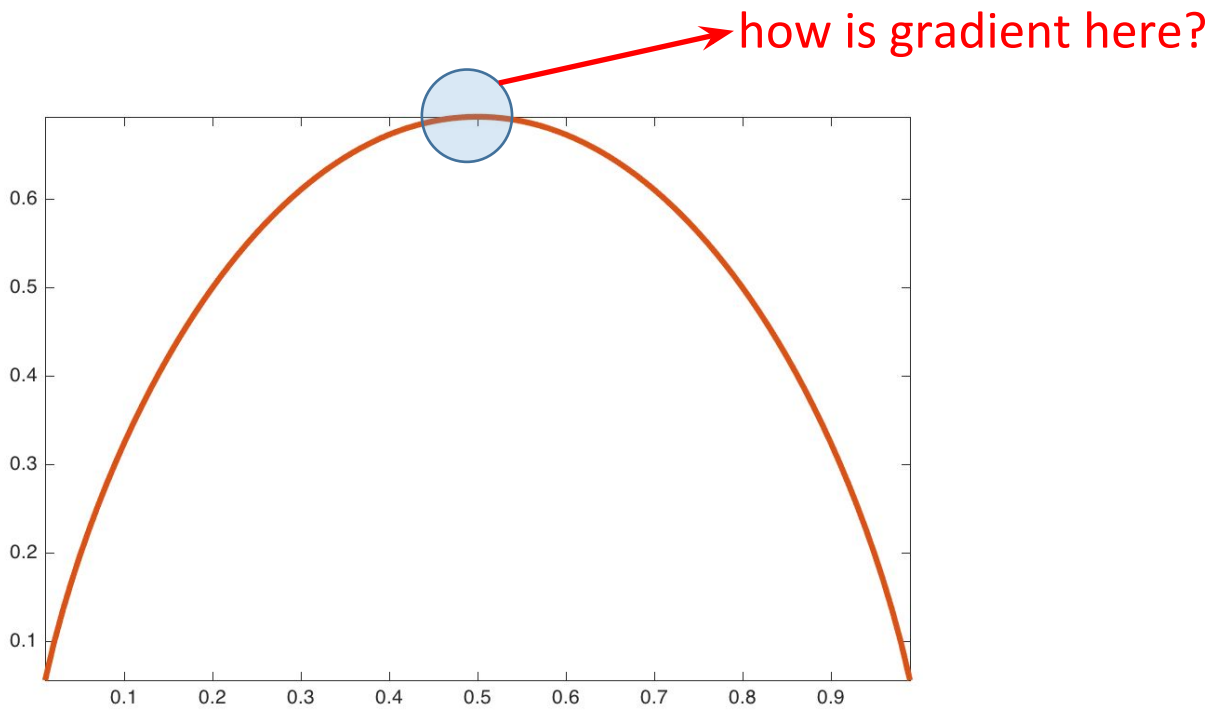


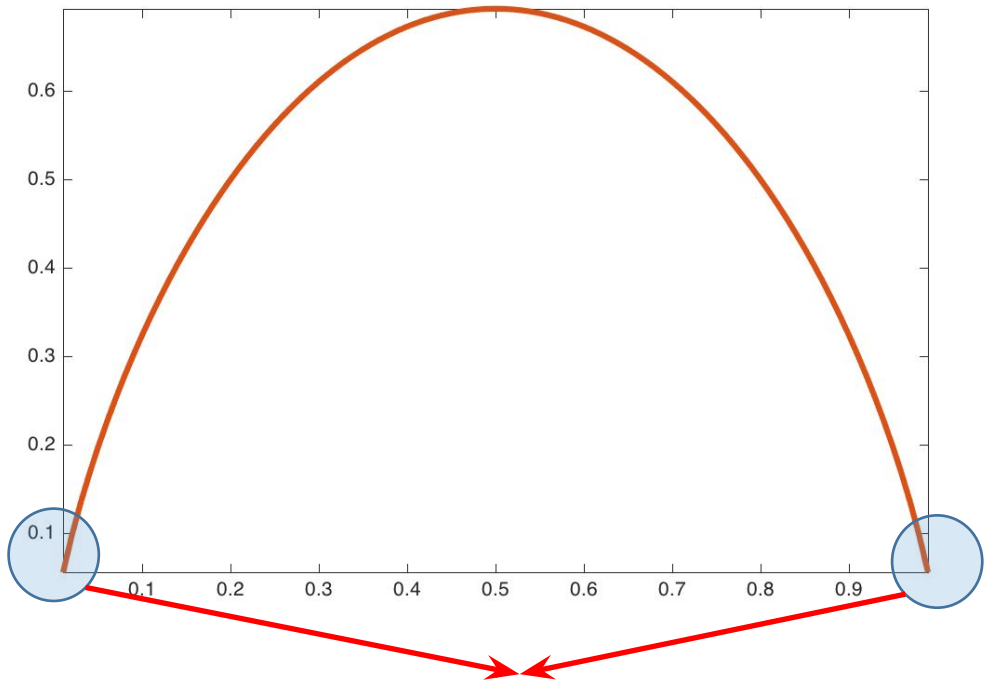
Image classification UDA on VisDA17 data set: Feature visualization for source model (left) and *min-entropy (lower bound on Shannon)* minimization (right) - equivalent to self training (clarified in the next slide)

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

Difficulty of optimizing entropy

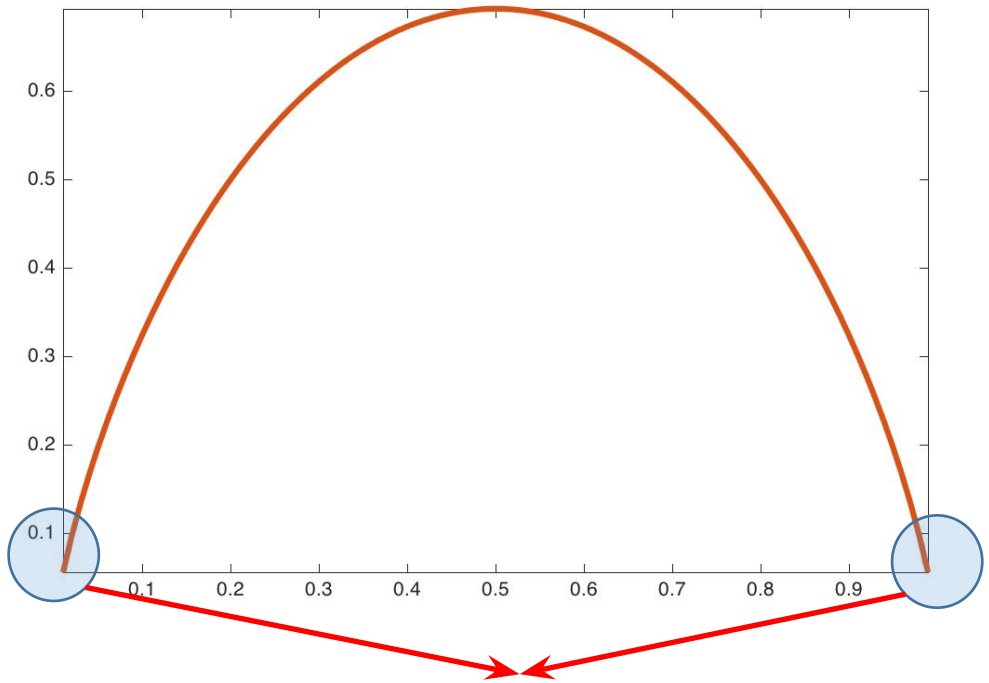


Difficulty of optimizing entropy



Difficulty of optimizing entropy

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



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

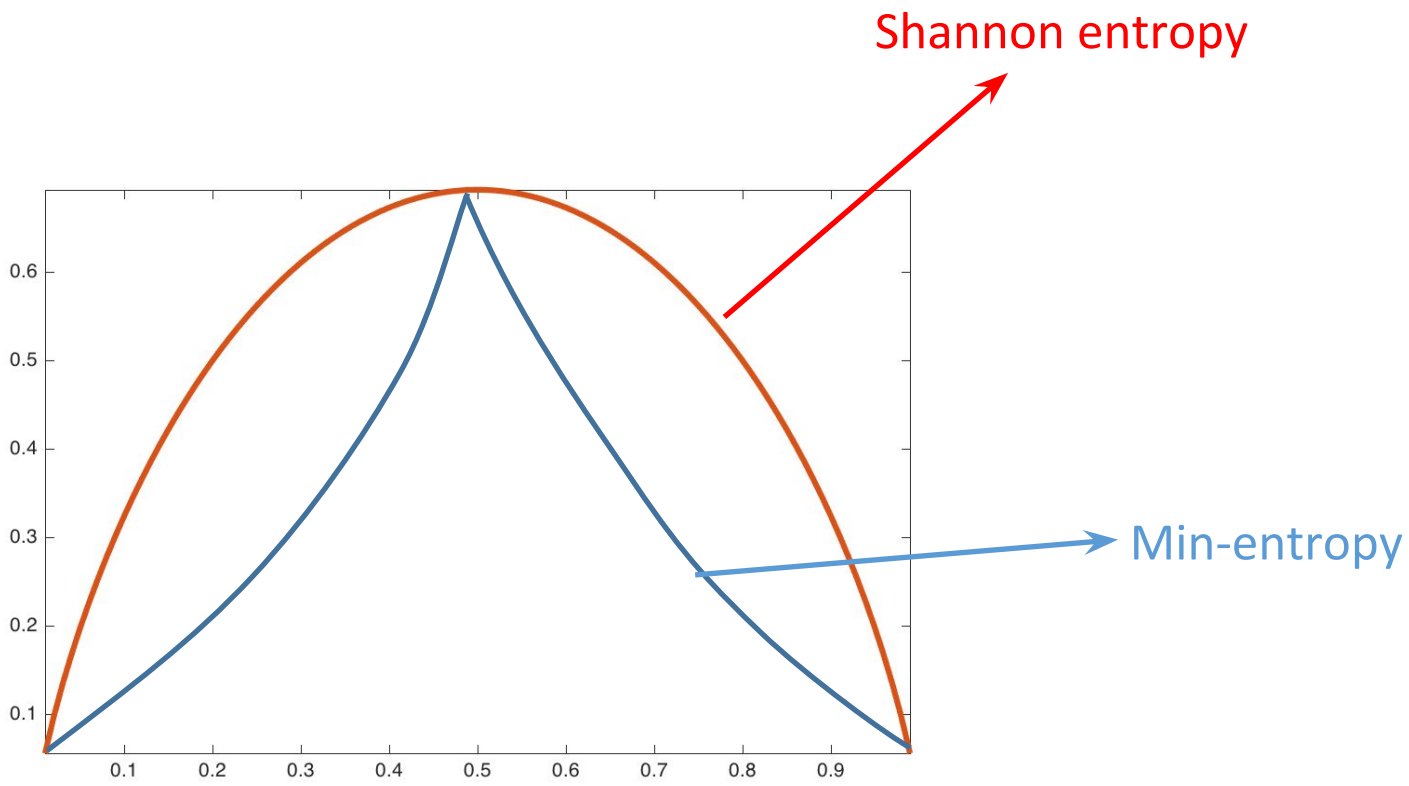
Link to Self-Training

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

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

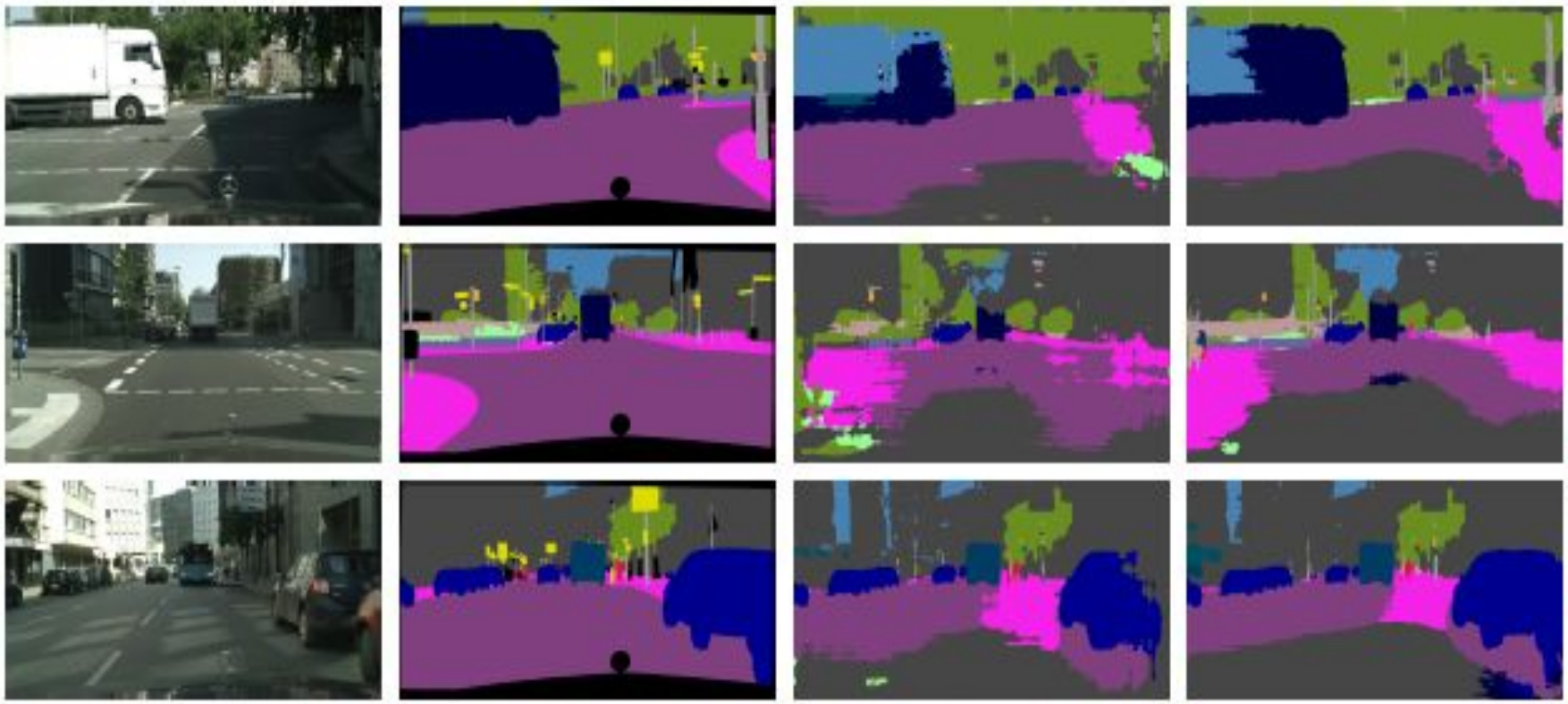
Self-Training + keeping the most confident predictions (corresponds to optimizing this simple loss)

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

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 the state-of-the-art for UDA)



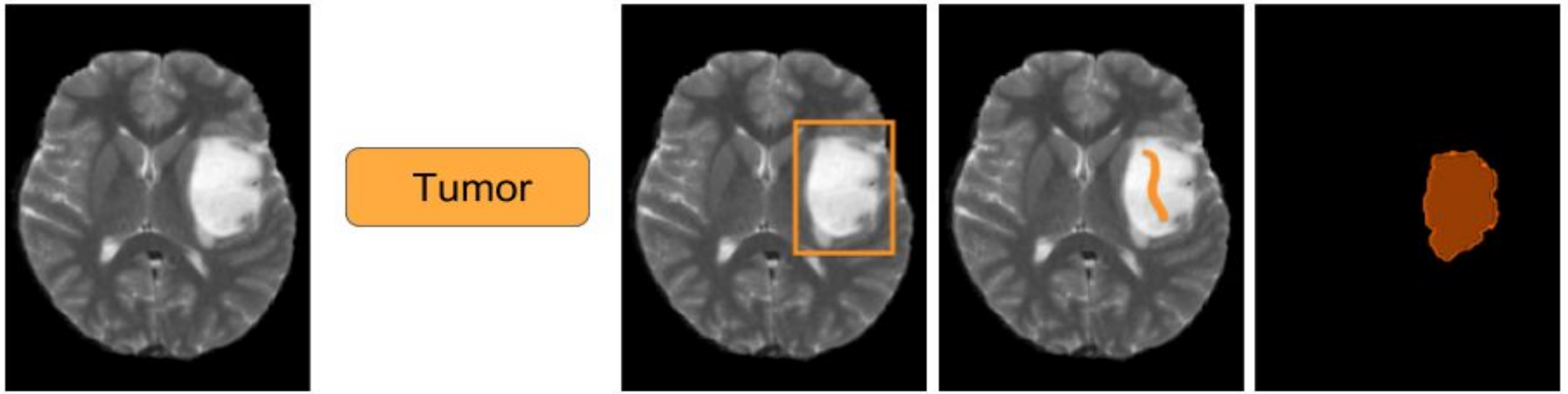
From Zou et al., Unsupervised Domain Adaptation for Semantic Segmentation via Class-Balanced Self-Training, ECCV 2018

Constrained CNNs

Constrained optimization (in CNNs)

Reminder

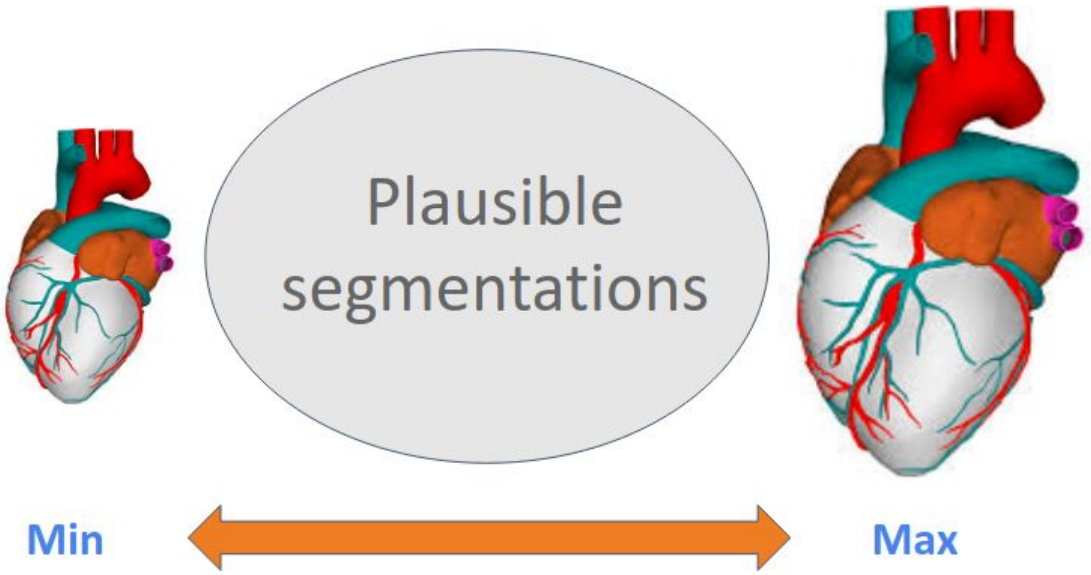
Standard levels of supervision



Constrained optimization (in CNNs)

Motivation

Prior information we can leverage on the medical domain

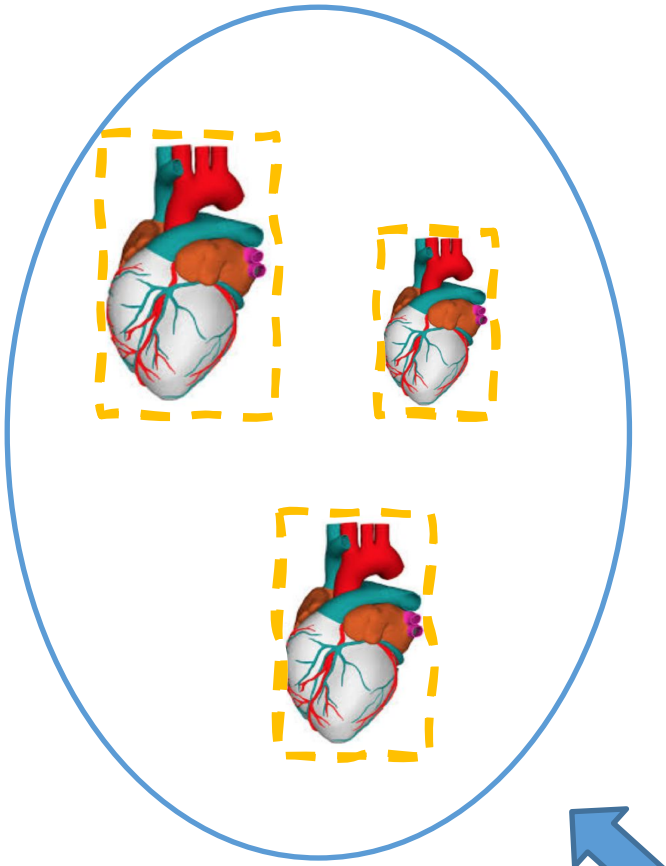


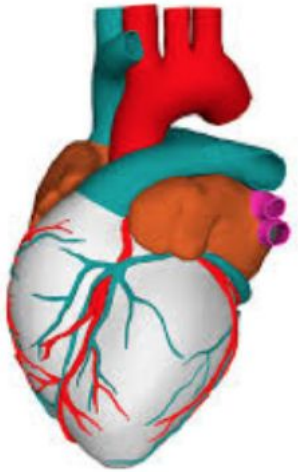
Anatomical priors

Partial labeled data
(exploit target relationships)

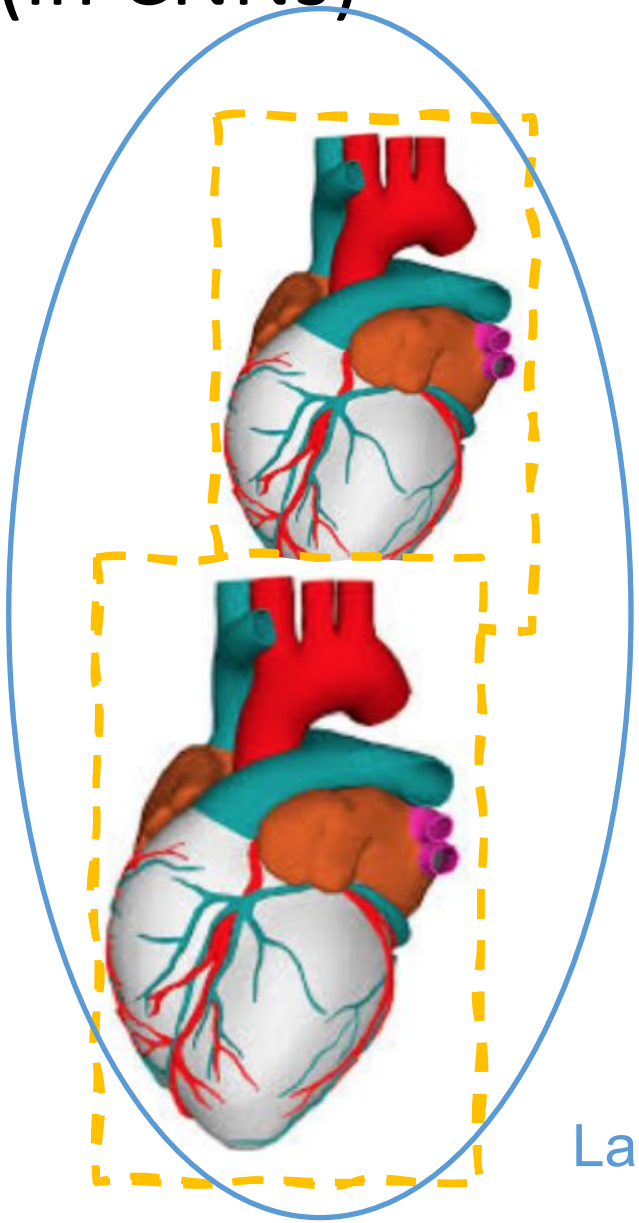
Constrained optimization (in CNNs)

Equality constraints





Known size



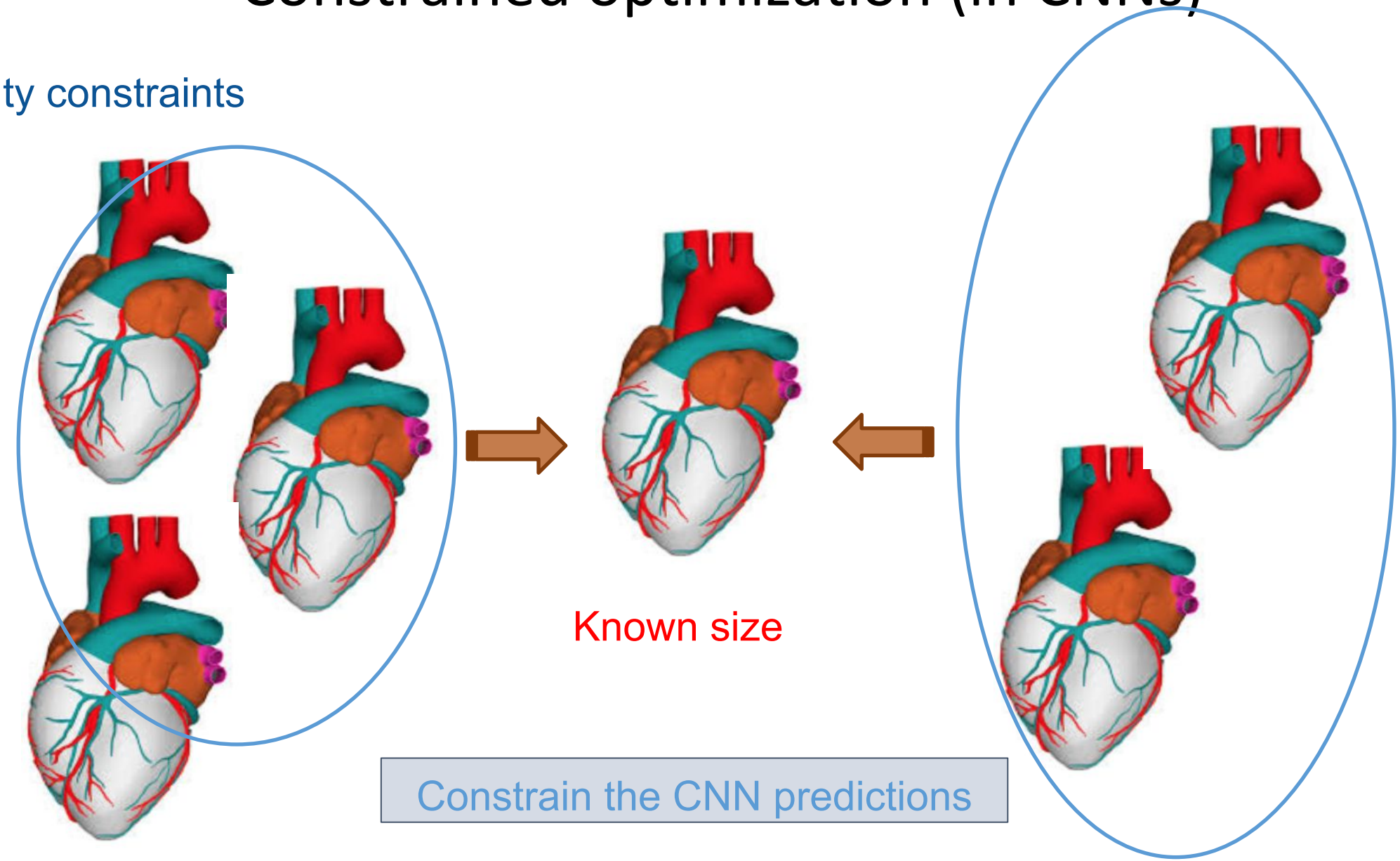
Smaller

CNN predictions

Larger

Constrained optimization (in CNNs)

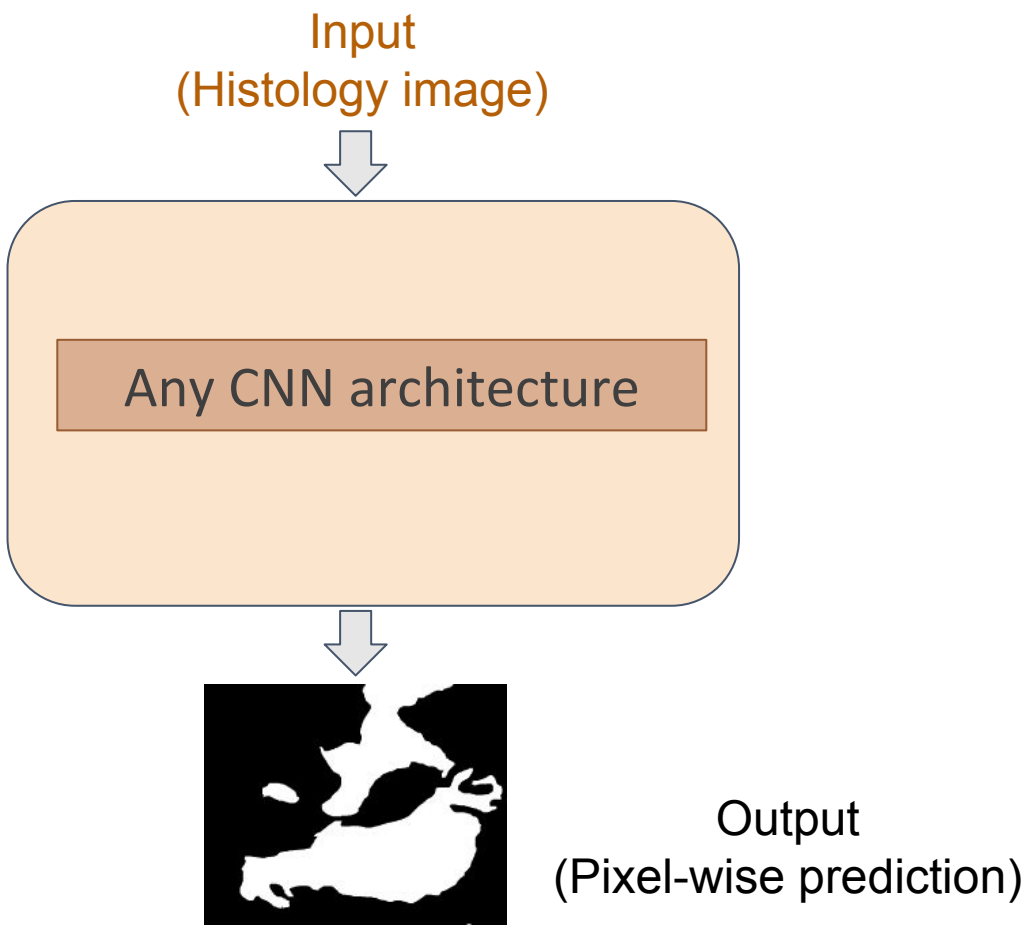
Equality constraints



Constrain the CNN predictions

Constrained optimization (in CNNs)

Equality constraints (e.g, L2 penalty)



Jia Z,et al. Constrained deep weak supervision for histopathology image segmentation. IEEE TMI 2017

Constrained optimization (in CNNs)

Equality constraints (e.g, L2 penalty)

Input
(Histology image)

Any CNN architecture

Output
(Pixel-wise prediction)

Additional term

$$l_{ac} = \boldsymbol{I}(Y_i = 1) \sum_i (v_i - a_i)^2$$

Predicted relative size (%)

$$v_i = \frac{1}{N} \sum_{p \in \Omega} s_\theta^{p,1}$$

Predicted size given by experts

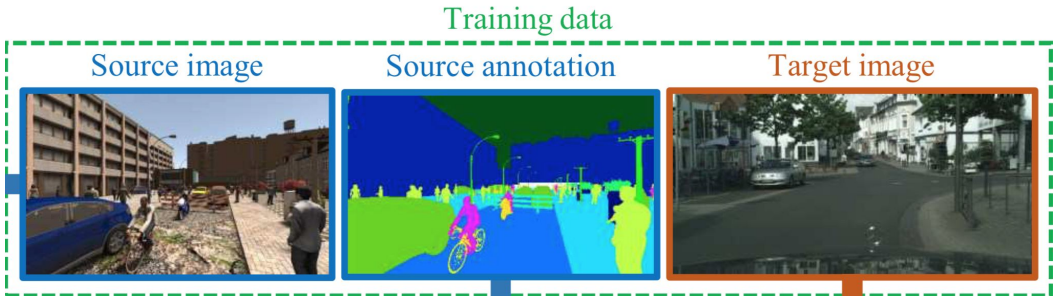
Reminder: ½ KL is an upper bound on L_2 for simplex vectors (Pinsker's inequality)

Jia Z,et al. Constrained deep weak supervision for histopathology image segmentation. IEEE TMI 2017

Constrained optimization (in CNNs)

Equality constraints (e.g, KL)

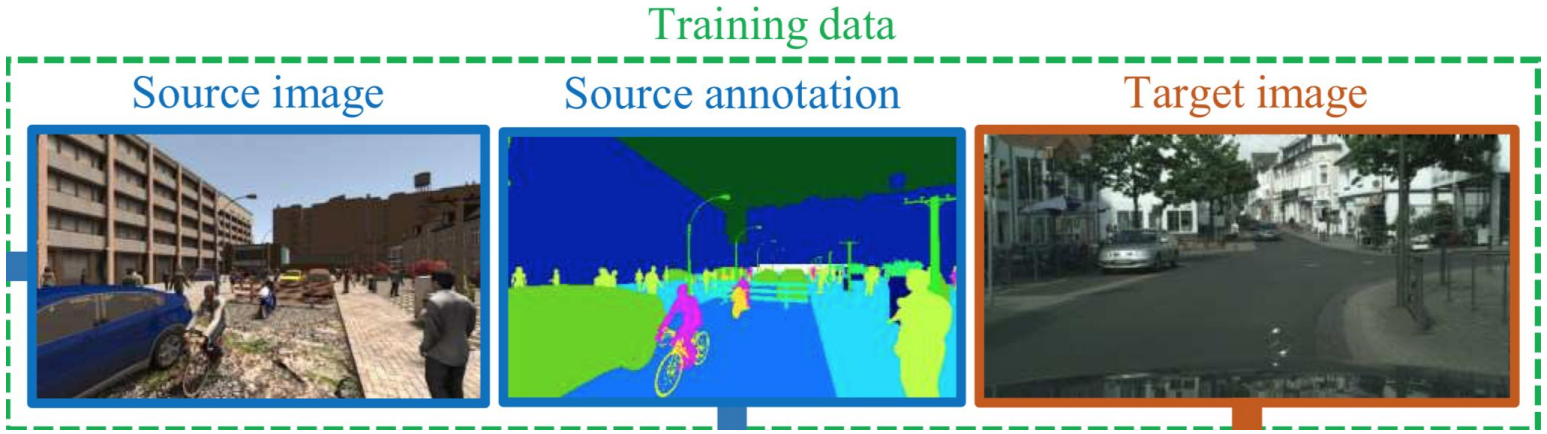
Unsupervised domain adaptation



Partially labeled data

Constrained optimization (in CNNs)

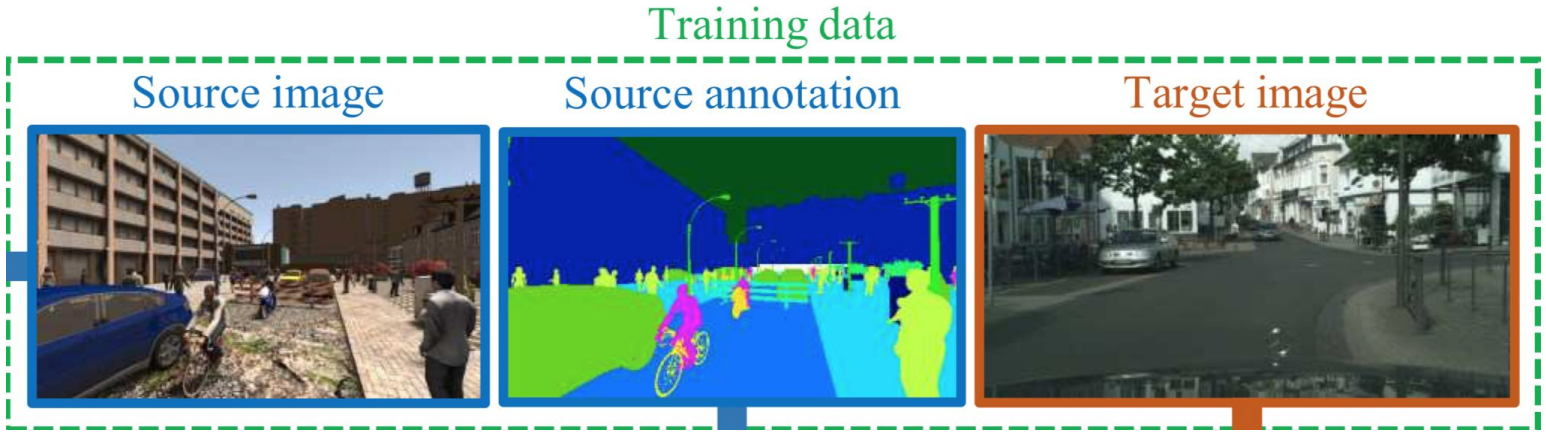
Equality constraints (e.g, KL): Curriculum DA



Zhang et al., Curriculum Domain Adaptation for Semantic Segmentation of Urban Scenes. ICCV'17

Constrained optimization (in CNNs)

Equality constraints (e.g, KL): Curriculum DA

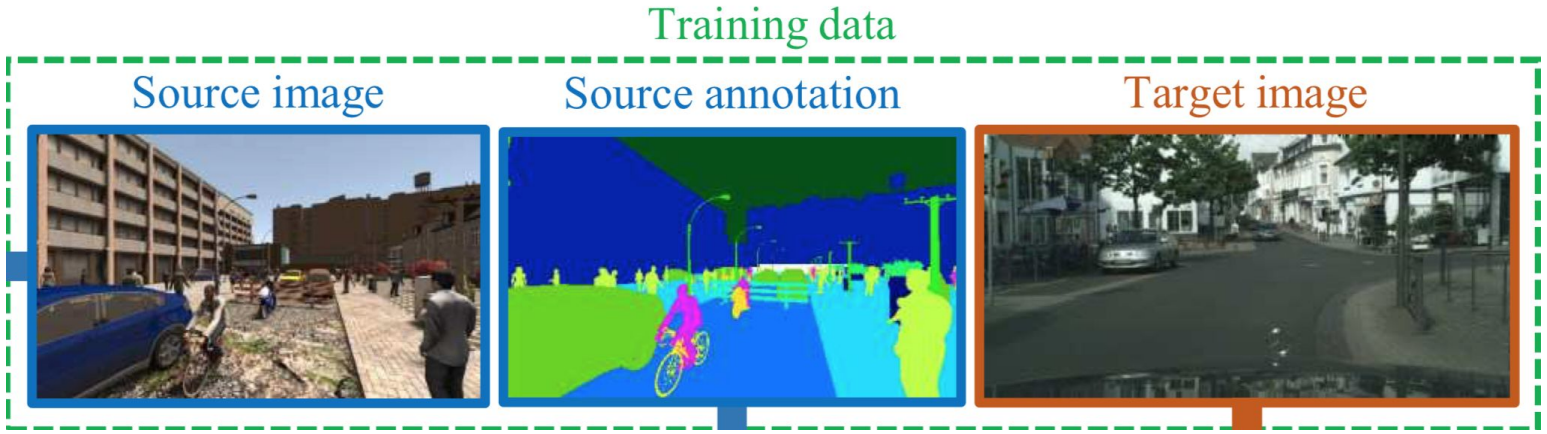


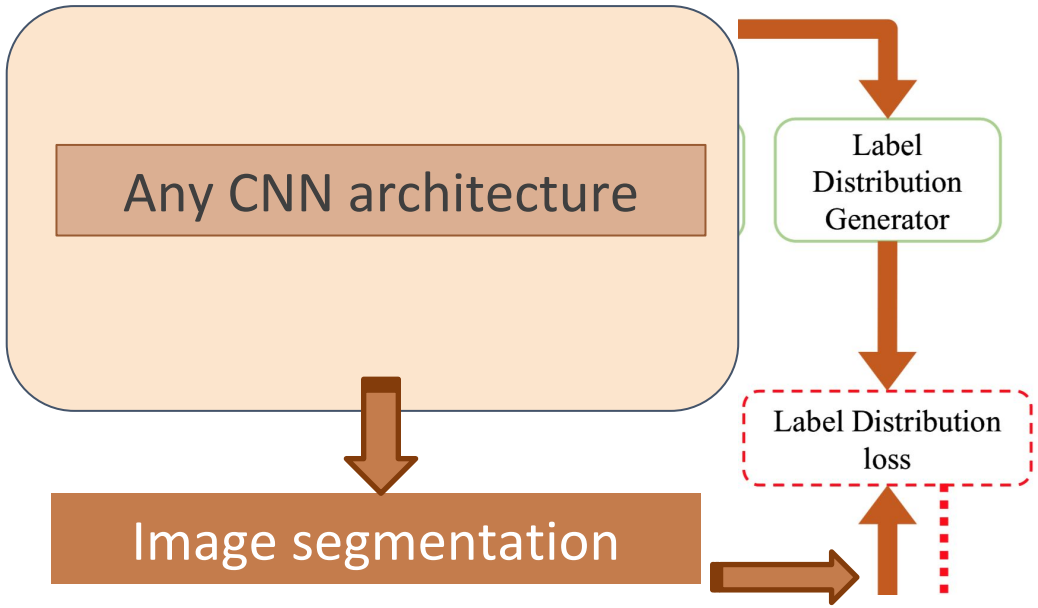
$$min \frac{\gamma}{|\mathcal{L}|} \sum_{p \in \mathcal{L}} l(\mathbf{y}^p, s_\theta^p) + \frac{1-\gamma}{|\mathcal{U}|} \sum_{q \in \mathcal{U}} \sum_k \mathbf{C}(a^{q,k}, \hat{a}^{q,k})$$

Zhang et al., Curriculum Domain Adaptation for Semantic Segmentation of Urban Scenes. ICCV'17

Constrained optimization (in CNNs)

Equality constraints (e.g, KL): Curriculum DA





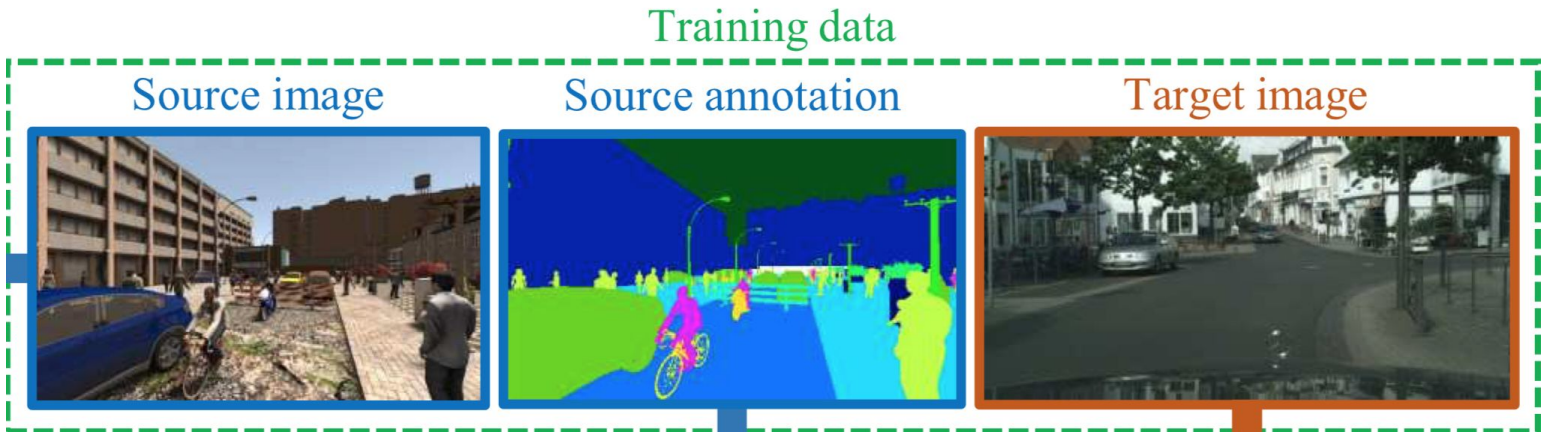
$$min \frac{\gamma}{|\mathcal{L}|} \sum_{p \in \mathcal{L}} l(\mathbf{y}^p, s_\theta^p) + \frac{1-\gamma}{|\mathcal{U}|} \sum_{q \in \mathcal{U}} \sum_k \mathbf{C}(a^{q,k}, \hat{a}^{q,k})$$

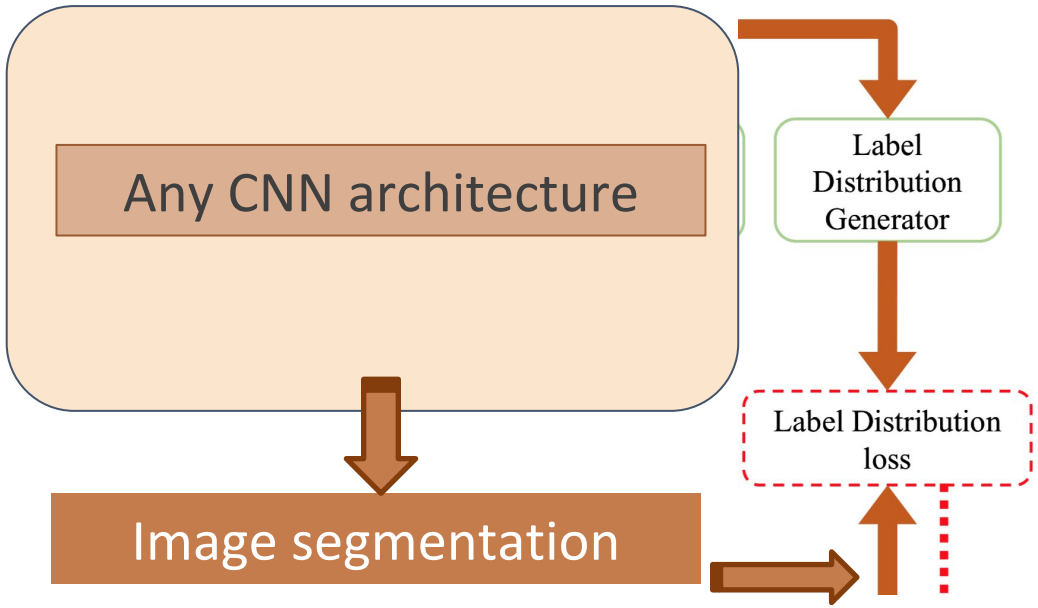
CE on supervised images (i.e., source)

Zhang et al., Curriculum Domain Adaptation for Semantic Segmentation of Urban Scenes. ICCV'17

Constrained optimization (in CNNs)

Equality constraints (e.g, KL): Curriculum DA





Additional term

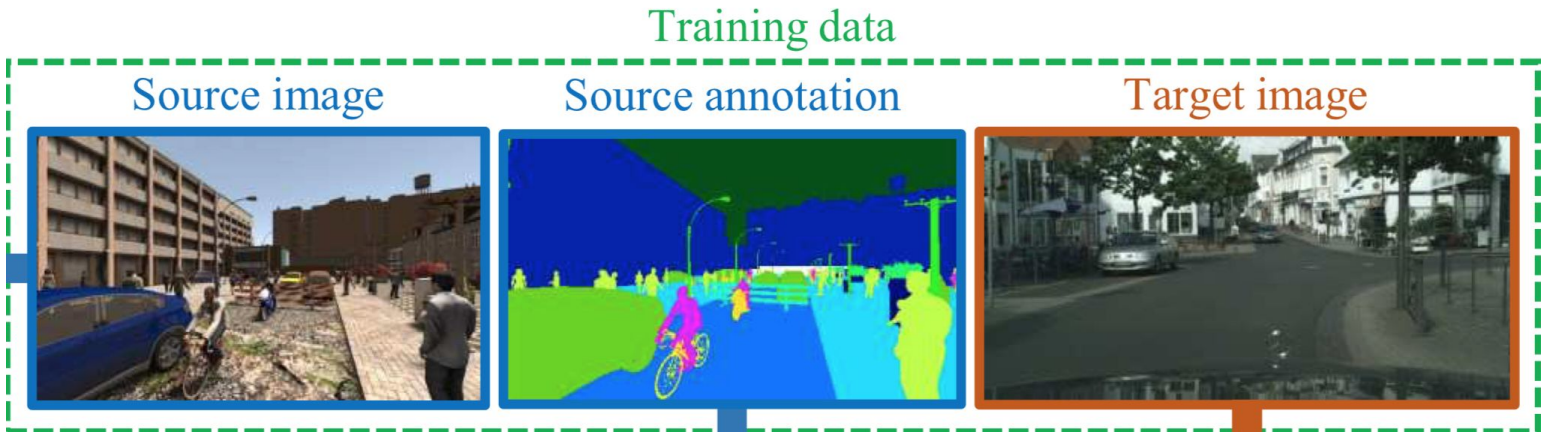
$$min \frac{\gamma}{|\mathcal{L}|} \sum_{p \in \mathcal{L}} l(\mathbf{y}^p, s_\theta^p) + \frac{1-\gamma}{|\mathcal{U}|} \sum_{q \in \mathcal{U}} \sum_k \mathbf{C}(a^{q,k}, \hat{a}^{q,k})$$

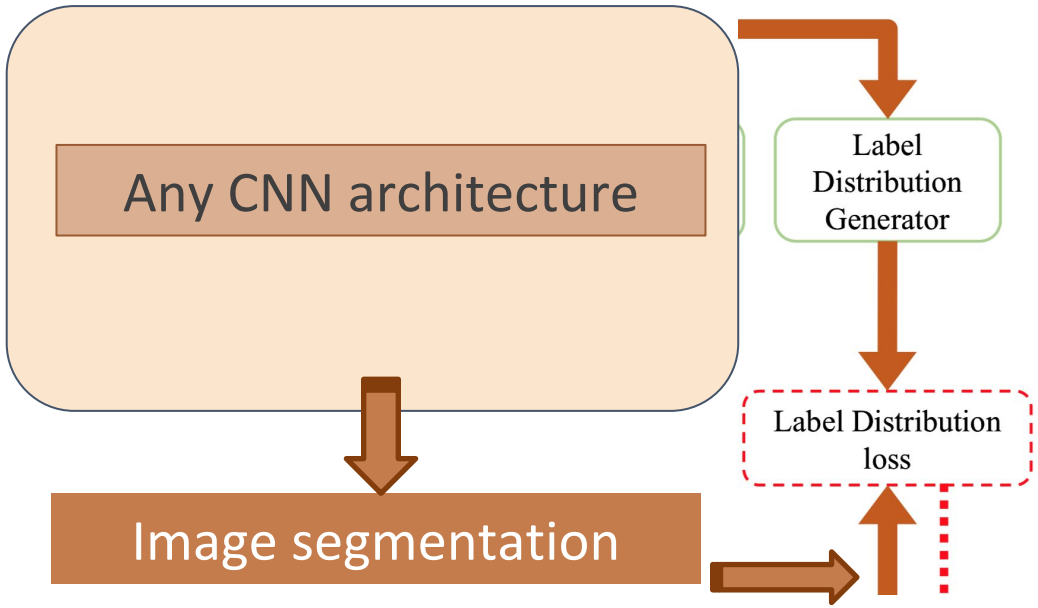
CE on supervised images (i.e., source)

Zhang et al., Curriculum Domain Adaptation for Semantic Segmentation of Urban Scenes. ICCV'17

Constrained optimization (in CNNs)

Equality constraints (e.g, KL): Curriculum DA





Additional term

$$min \frac{\gamma}{|\mathcal{L}|} \sum_{p \in \mathcal{L}} l(\mathbf{y}^p, s_\theta^p) + \frac{1-\gamma}{|\mathcal{U}|} \sum_{q \in \mathcal{U}} \sum_k \mathbf{C}(a^{q,k}, \hat{a}^{q,k})$$

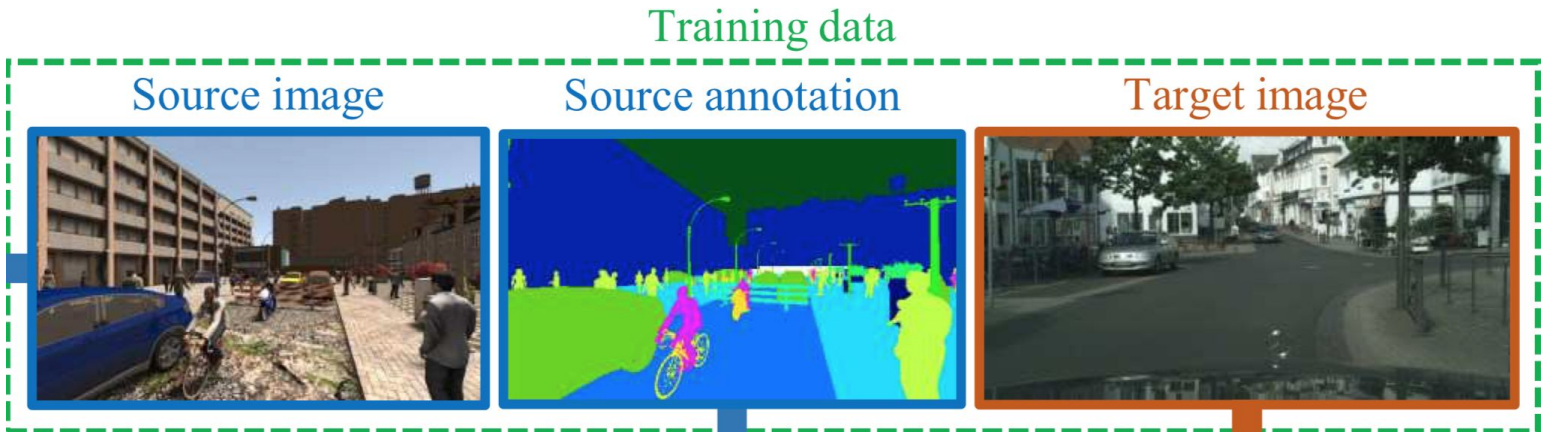
CE on supervised images (i.e., source)

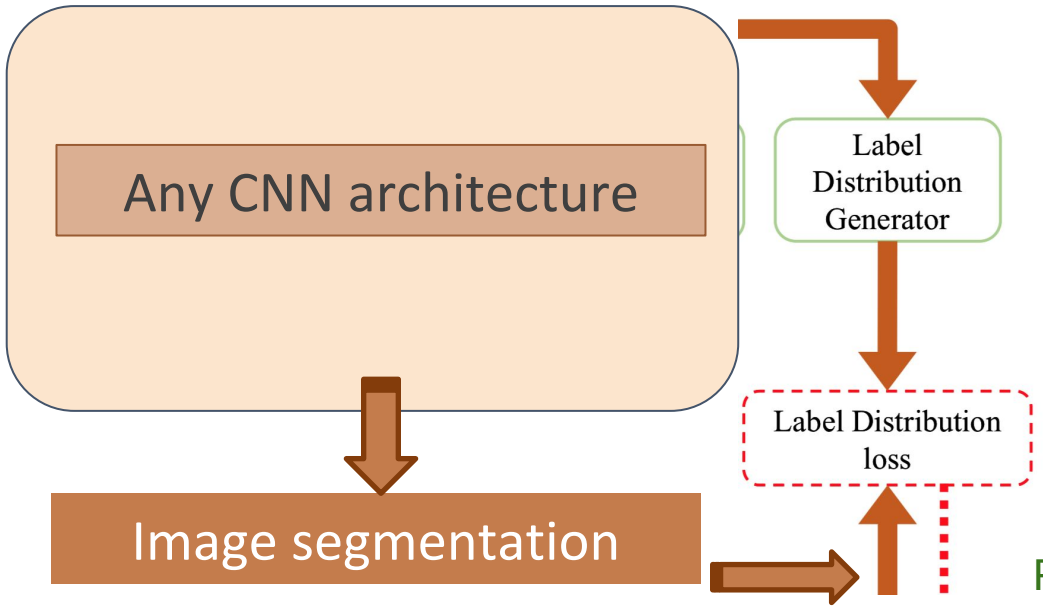
$$\mathbf{C}(\mathbf{a}^q, \hat{\mathbf{a}}^q) = H(\mathbf{a}^q) + KL(\mathbf{a}^q, \hat{\mathbf{a}}^q)$$

Zhang et al., Curriculum Domain Adaptation for Semantic Segmentation of Urban Scenes. ICCV'17

Constrained optimization (in CNNs)

Equality constraints (e.g, KL): Curriculum DA





$$min \frac{\gamma}{|\mathcal{L}|} \sum_{p \in \mathcal{L}} l(\mathbf{y}^p, s_\theta^p) + \frac{1-\gamma}{|\mathcal{U}|} \sum_{q \in \mathcal{U}} \sum_k \mathbf{C}(a^{q,k}, \hat{a}^{q,k})$$

Additional term

CE on supervised images (i.e., source)

$$\mathbf{C}(\mathbf{a}^q, \hat{\mathbf{a}}^q) = H(\mathbf{a}^q) + KL(\mathbf{a}^q, \hat{\mathbf{a}}^q)$$

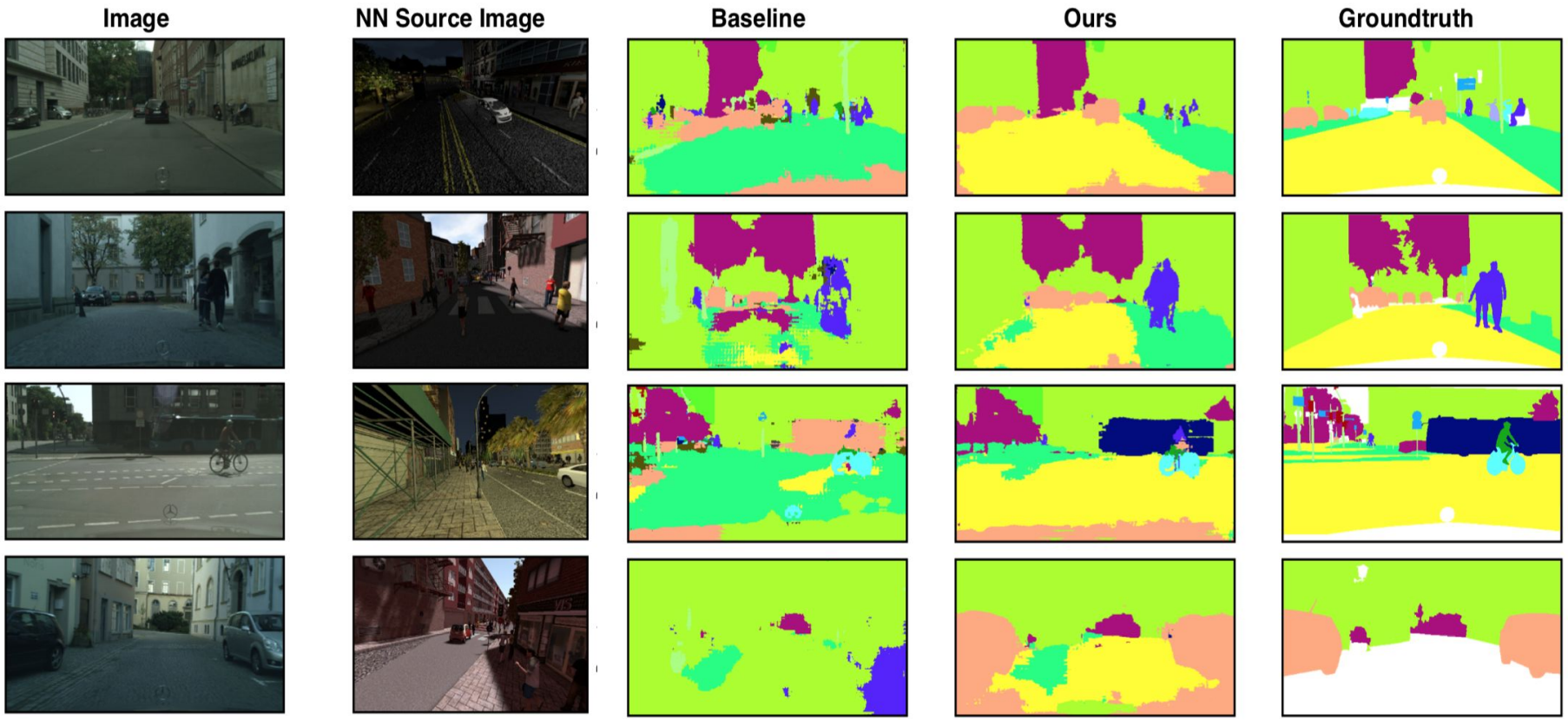
Predicted size

From predicted image

Zhang et al., Curriculum Domain Adaptation for Semantic Segmentation of Urban Scenes. ICCV'17

Constrained optimization (in CNNs)

Equality constraints (e.g, KL): Curriculum DA



Images from [Zhang et al., Curriculum Domain Adaptation for Semantic Segmentation of Urban Scenes. ICCV'17]

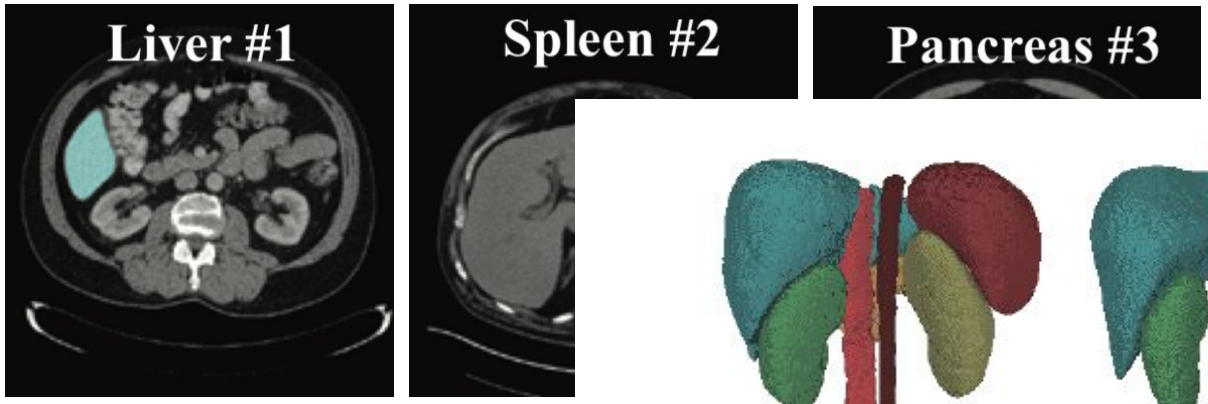
Constrained optimization (in CNNs)

Equality constraints (e.g, KL): Partial annotations

[Zhou et al., Prior-aware Neural Network for Partially-Supervised Multi-Organ Segmentation, ICCV'19]

Constrained optimization (in CNNs)

Equality constraints (e.g, KL): Partial annotations



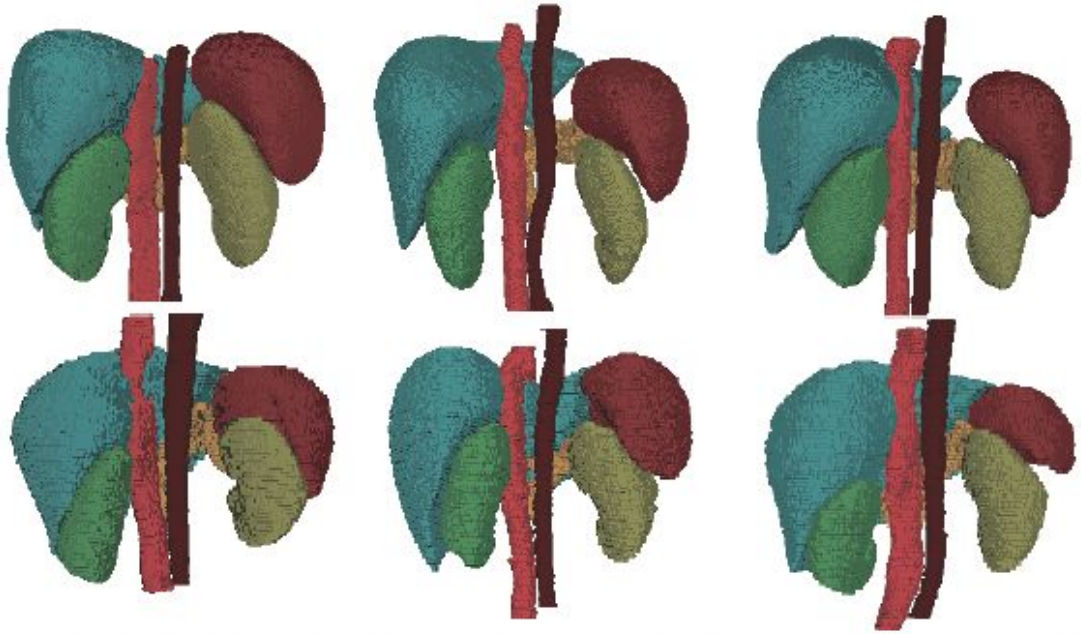


Figure 1. 3D Visualization of several abdominal organs (liver, spleen, left kidney, right kidney, aorta, inferior vena cava) to show the similarity of patient-wise abdominal organ size distributions.

Exploit relative size between organs in fully annotated volumes

Images from [Zhou et al., Prior-aware Neural Network for Partially-Supervised Multi-Organ Segmentation, ICCV'19]

Constrained optimization (in CNNs)

Equality constraints (e.g, KL): Partial annotations

Main objective:

$$min \frac{1}{|\mathcal{L}|} \sum_{p \in \mathcal{L}} l(\mathbf{y}^p, \mathbf{s}_\theta^p) + \lambda_1 \frac{1}{|\mathcal{P}|} \sum_{q \in \mathcal{P}} l(\mathbf{y}^q, \mathbf{s}_\theta^q) + \lambda_2 \mathcal{J}(\theta)$$

Fully labeled images

Partially labeled images

Prior-aware loss

[Zhou et al., Prior-aware Neural Network for Partially-Supervised Multi-Organ Segmentation, ICCV'19]

Constrained optimization (in CNNs)

Equality constraints (e.g, KL): Partial annotations

Prior-aware loss

Averaged predicted distribution

$$\hat{\mathbf{p}} = \frac{1}{N} \sum_{p \in \mathcal{P}} \mathbf{s}_\theta^p$$

$$[s_\theta^{p,0}, s_\theta^{p,1}, \ldots, s_\theta^{p,|K|}]$$

On partially labeled images

[Zhou et al., Prior-aware Neural Network for Partially-Supervised Multi-Organ Segmentation, ICCV'19]

Constrained optimization (in CNNs)

Equality constraints (e.g, KL): Partial annotations

Prior-aware loss

Embed prior knowledge

$$KL(\mathbf{q}|\hat{\mathbf{p}})$$

Real label distribution Average predicted distribution

[Zhou et al., Prior-aware Neural Network for Partially-Supervised Multi-Organ Segmentation, ICCV'19]

Constrained optimization (in CNNs)

Equality constraints (e.g, KL): Partial annotations

Prior-aware loss

KL can be expanded

$$-\sum_{c=0}^{|K|}\{q^c log\frac{1}{N}\sum_{p\in\mathcal{P}}\mathbf{s}_\theta^{p,c}+(1-q^c)log(1-\frac{1}{N}\sum_{p\in\mathcal{P}}\mathbf{s}_\theta^{p,c})\}+const$$

[Zhou et al., Prior-aware Neural Network for Partially-Supervised Multi-Organ Segmentation, ICCV'19]

Constrained optimization (in CNNs)

Equality constraints (e.g, KL): Partial annotations

Prior-aware loss

KL can be expanded

$$-\sum_{c=0}^{|K|}\{q^c log\frac{1}{N}\sum_{p\in\mathcal{P}}\mathbf{s}_\theta^{p,c}+(1-q^c)log(1-\frac{1}{N}\sum_{p\in\mathcal{P}}\mathbf{s}_\theta^{p,c})\}+const$$

This is problematic (average distribution of $\hat{p}$ organ sizes inside log!!)

[Zhou et al., Prior-aware Neural Network for Partially-Supervised Multi-Organ Segmentation, ICCV'19]

Constrained optimization (in CNNs)

Equality constraints (e.g, KL): Partial annotations

Stochastic primal-dual gradient

(split terms updated independently)

Prior-aware loss

KL can be expanded

$$-\sum_{c=0}^{|K|}\{q^c log\frac{1}{N}\sum_{p\in\mathcal{P}}\mathbf{s}_\theta^{p,c}+(1-q^c)log(1-\frac{1}{N}\sum_{p\in\mathcal{P}}\mathbf{s}_\theta^{p,c})\}+const$$

This is problematic (average distribution of $\hat{p}$ organ sizes inside log!!)

[Zhou et al., Prior-aware Neural Network for Partially-Supervised Multi-Organ Segmentation, ICCV'19]

Constrained optimization (in CNNs)

Equality constraints (e.g, KL): Partial annotations

primal dual

$$-\log\alpha = \max_{\beta}\left(\alpha\beta + 1 + \log(-\beta)\right)$$

Stochastic primal-dual gradient

(split terms updated independently)

Prior-aware loss

KL can be expanded

$$-\sum_{c=0}^{|K|}\{q^c log\frac{1}{N}\sum_{p\in\mathcal{P}}\mathbf{s}_\theta^{p,c} + (1-q^c) log\left(1-\frac{1}{N}\sum_{p\in\mathcal{P}}\mathbf{s}_\theta^{p,c}\right)\} + const$$

This is problematic (average distribution of $\hat{p}$ organ sizes inside log!!)

[Zhou et al., Prior-aware Neural Network for Partially-Supervised Multi-Organ Segmentation, ICCV'19]

Constrained optimization (in CNNs)

Equality constraints (e.g, KL): Partial annotations

$$-\log \vec{p}^{\,l} = \max_{\nu^l} \left(\vec{p}^{\,l} \nu^l + 1 + \log(-\nu^l) \right)$$

$$-\log(1-\vec{p}^{\,l}) = \max_{\mu^l} \left((1-\vec{p}^{\,l}) \mu^l + 1 + \log(-\mu^l) \right),$$

Stochastic primal-dual gradient

(split terms updated independently)

Prior-aware loss

KL can be expanded

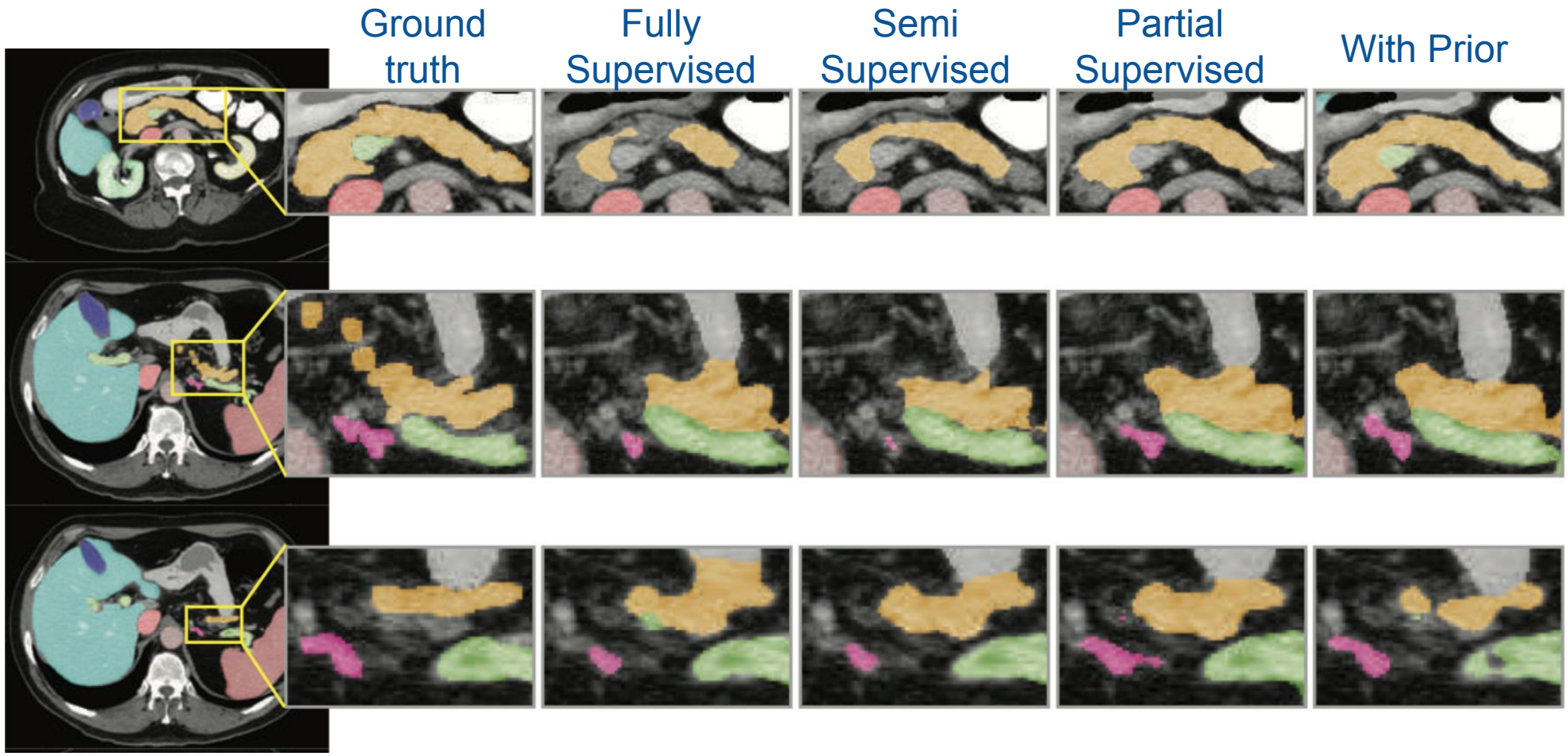
$$-\sum_{c=0}^{|K|} \{ q^c log \frac{1}{N} \sum_{p \in \mathcal{P}} \mathbf{s}_\theta^{p,c} + (1-q^c) log (1 - \frac{1}{N} \sum_{p \in \mathcal{P}} \mathbf{s}_\theta^{p,c}) \} + const$$

This is problematic (average distribution of p̂ organ sizes inside log!!)

[Zhou et al., Prior-aware Neural Network for Partially-Supervised Multi-Organ Segmentation, ICCV'19]

Constrained optimization (in CNNs)

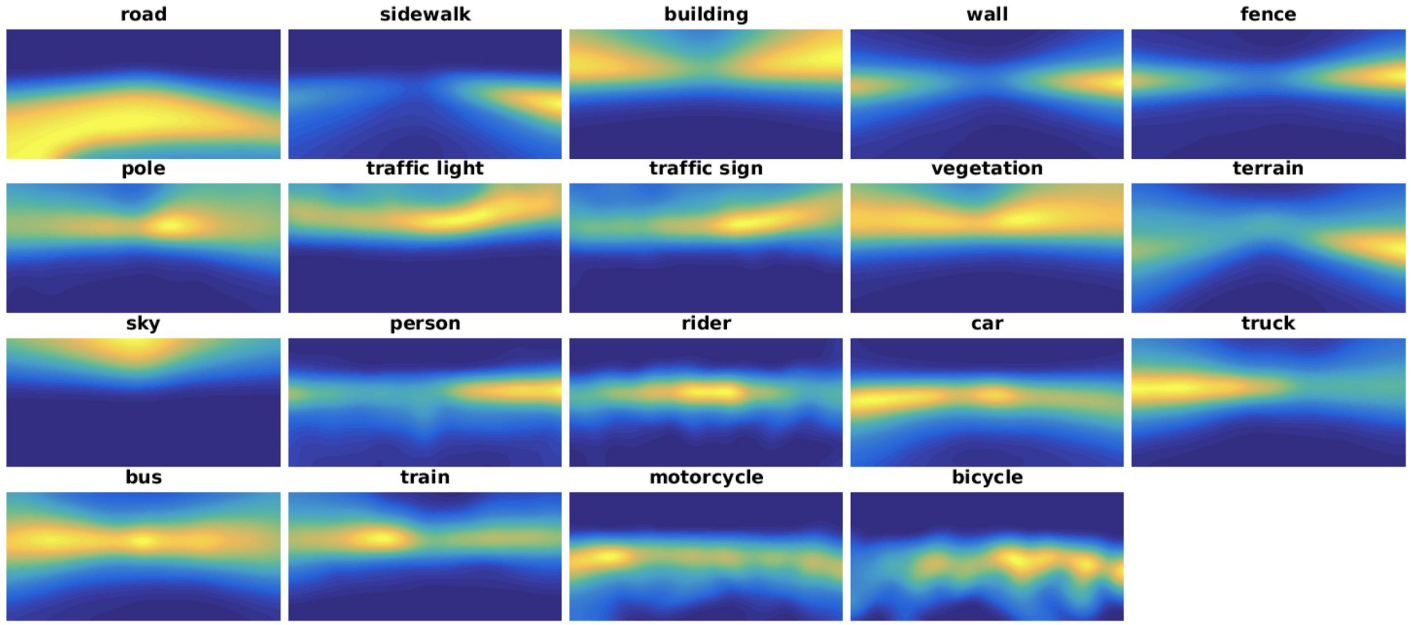
Equality constraints (e.g, KL): Partial annotations



Images from [Zhou et al., Prior-aware Neural Network for Partially-Supervised Multi-Organ Segmentation, ICCV'19]

Constrained optimization (in CNNs)

Equality constraints (at pixel-level)

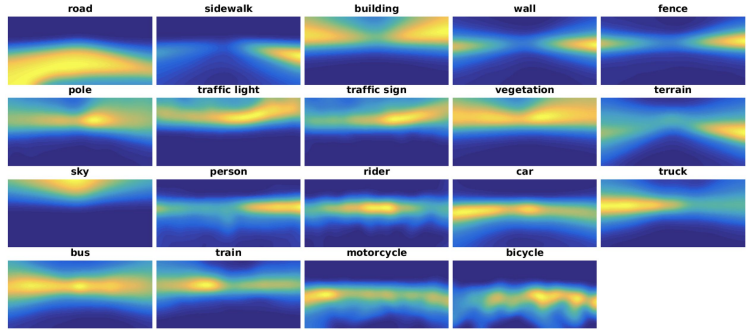


Spatial priors on GTA5

Images from [Zou et al., Unsupervised Domain Adaptation for Semantic Segmentation via Class-Balanced Self-Training, ECCV'18]

Constrained optimization (in CNNs)

Equality constraints (at pixel-level)



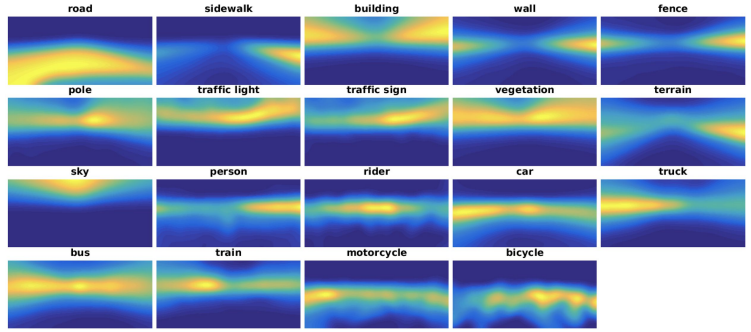
Objective:

$$-\sum_{p\in\mathcal{L}} l(\mathbf{y}^p, s_\theta^p) - \sum_{q\in\mathcal{U}}\sum_{c=1}^{K}(\hat{y}^{q,c} log(prior^{q,c} s_\theta^{q,c}) + w_c \hat{y}^{q,c})$$

Zou et al., Unsupervised Domain Adaptation for Semantic Segmentation via Class-Balanced Self-Training, ECCV'18

Constrained optimization (in CNNs)

Equality constraints (at pixel-level)



Proposals

Objective:

$$-\sum_{p\in\mathcal{L}} l(\mathbf{y}^p, s_\theta^p) - \sum_{q\in\mathcal{U}}\sum_{c=1}^{K}(\hat{y}^{q,c} log(prior^{q,c} s_\theta^{q,c}) + w_c \hat{y}^{q,c})$$

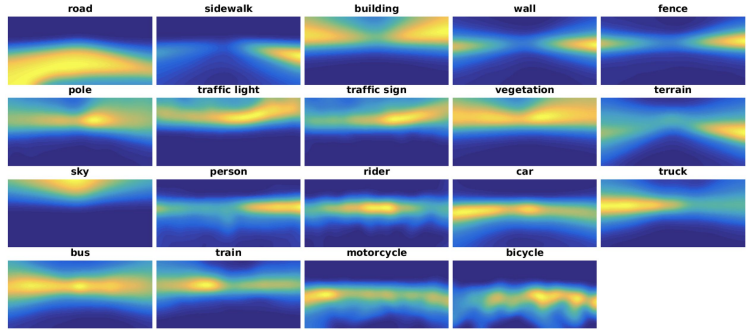
Target images

Source images

Zou et al., Unsupervised Domain Adaptation for Semantic Segmentation via Class-Balanced Self-Training, ECCV'18

Constrained optimization (in CNNs)

Equality constraints (at pixel-level)



Objective:

$$-\sum_{p\in\mathcal{L}} l(\mathbf{y}^p, s_\theta^p) - \sum_{q\in\mathcal{U}}\sum_{c=1}^{K}(\hat{y}^{q,c} log(prior^{q,c} s_\theta^{q,c}) + w_c \hat{y}^{q,c})$$

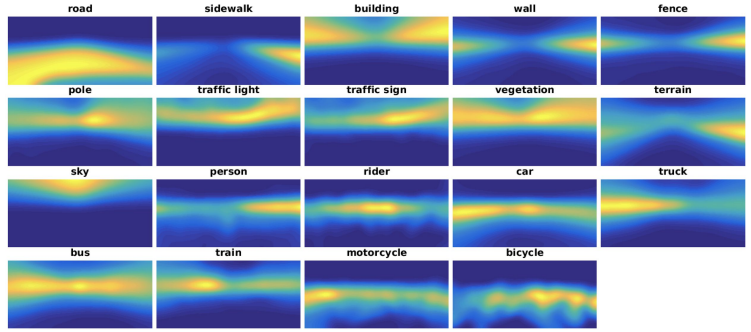
This becomes two KL

$$KL(\hat{y}^{q,c}|prior^{q,c}) \qquad KL(\hat{y}^{q,c}|s_\theta^{q,c})$$

Zou et al., Unsupervised Domain Adaptation for Semantic Segmentation via Class-Balanced Self-Training, ECCV'18

Constrained optimization (in CNNs)

Equality constraints (at pixel-level)



Objective:

$$-\sum_{p\in\mathcal{L}} l(\mathbf{y}^p, s_\theta^p) - \sum_{q\in\mathcal{U}}\sum_{c=1}^{K}(\hat{y}^{q,c} log(prior^{q,c} s_\theta^{q,c}) + w_c \hat{y}^{q,c})$$

Weights the proposals

Zou et al., Unsupervised Domain Adaptation for Semantic Segmentation via Class-Balanced Self-Training, ECCV'18

Constrained optimization (in CNNs)

Equality constraints (at pixel-level)

Spatial prior

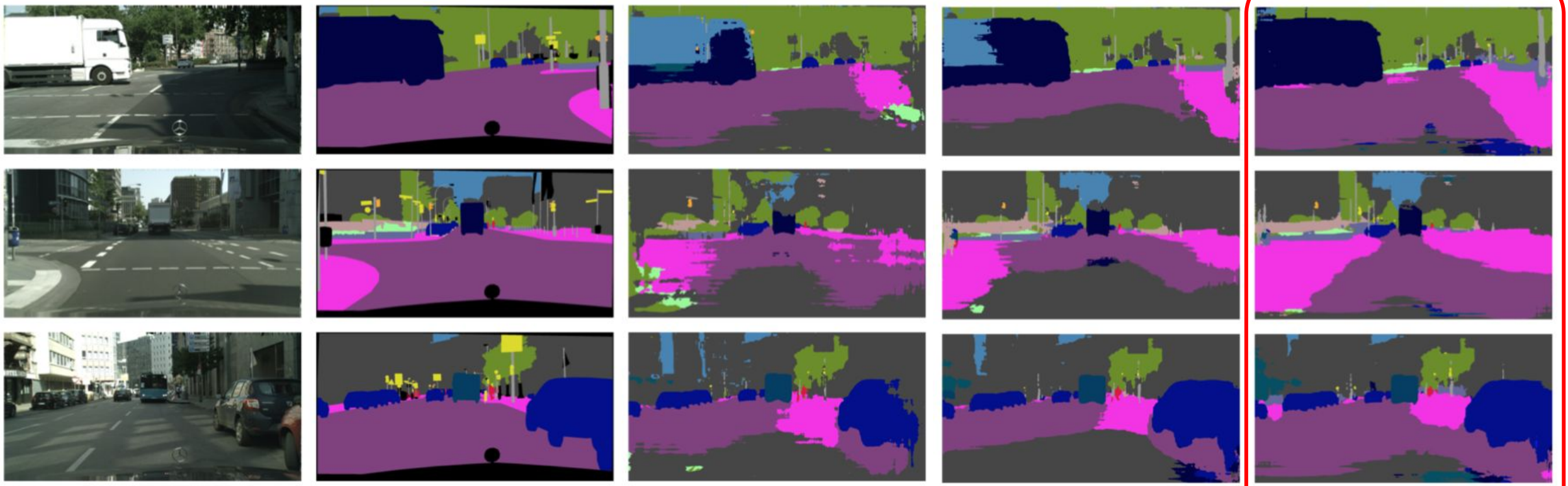


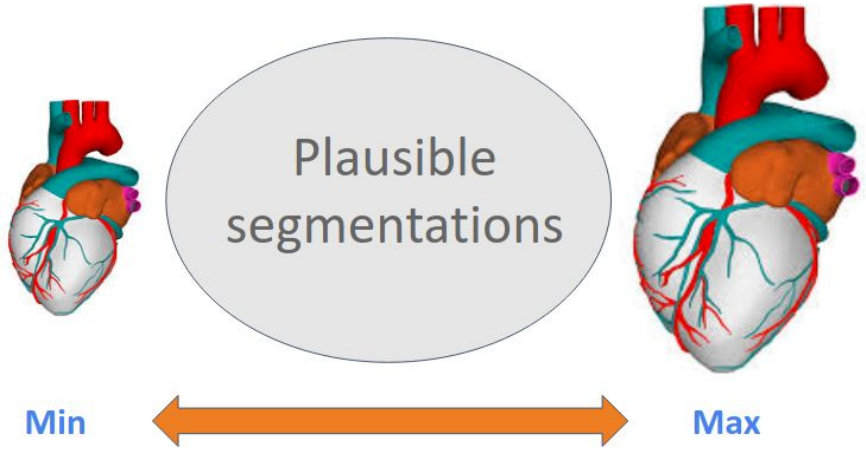
Image | Ground Truth | Only source | CBST | CBST-SP

Images from [Zou et al., Unsupervised Domain Adaptation for Semantic Segmentation via Class-Balanced Self-Training, ECCV'18]

Constrained optimization (in CNNs)

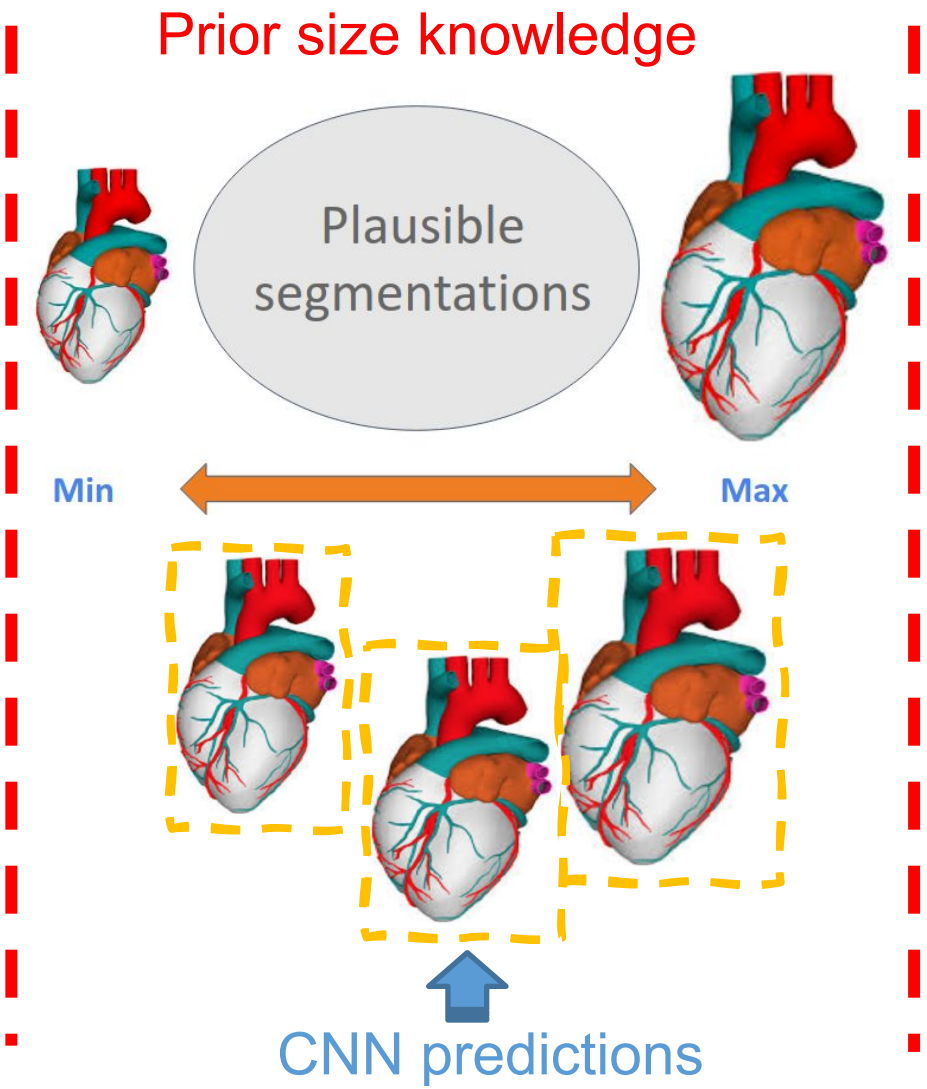
Inequality constraints

Prior size knowledge

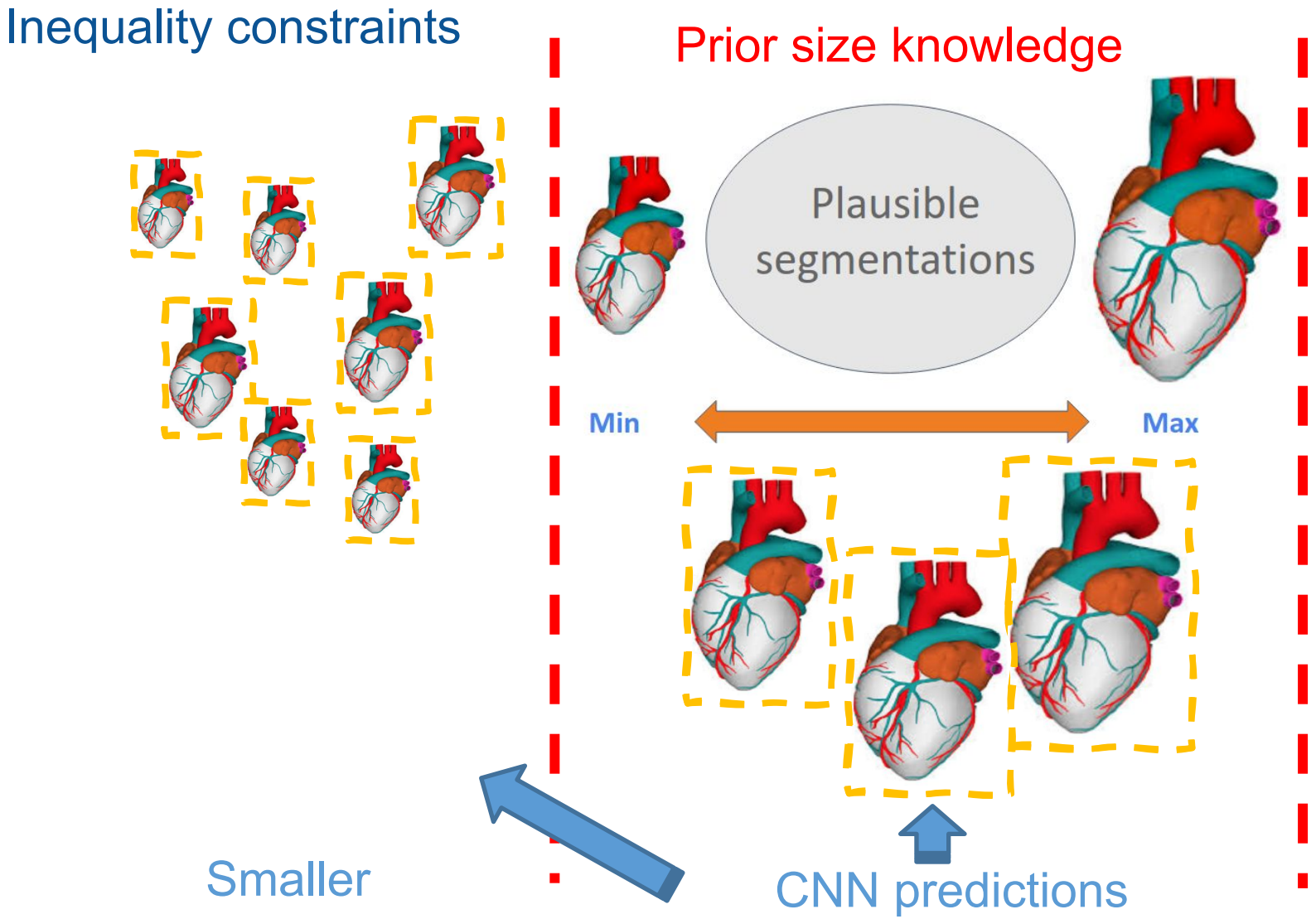


Constrained optimization (in CNNs)

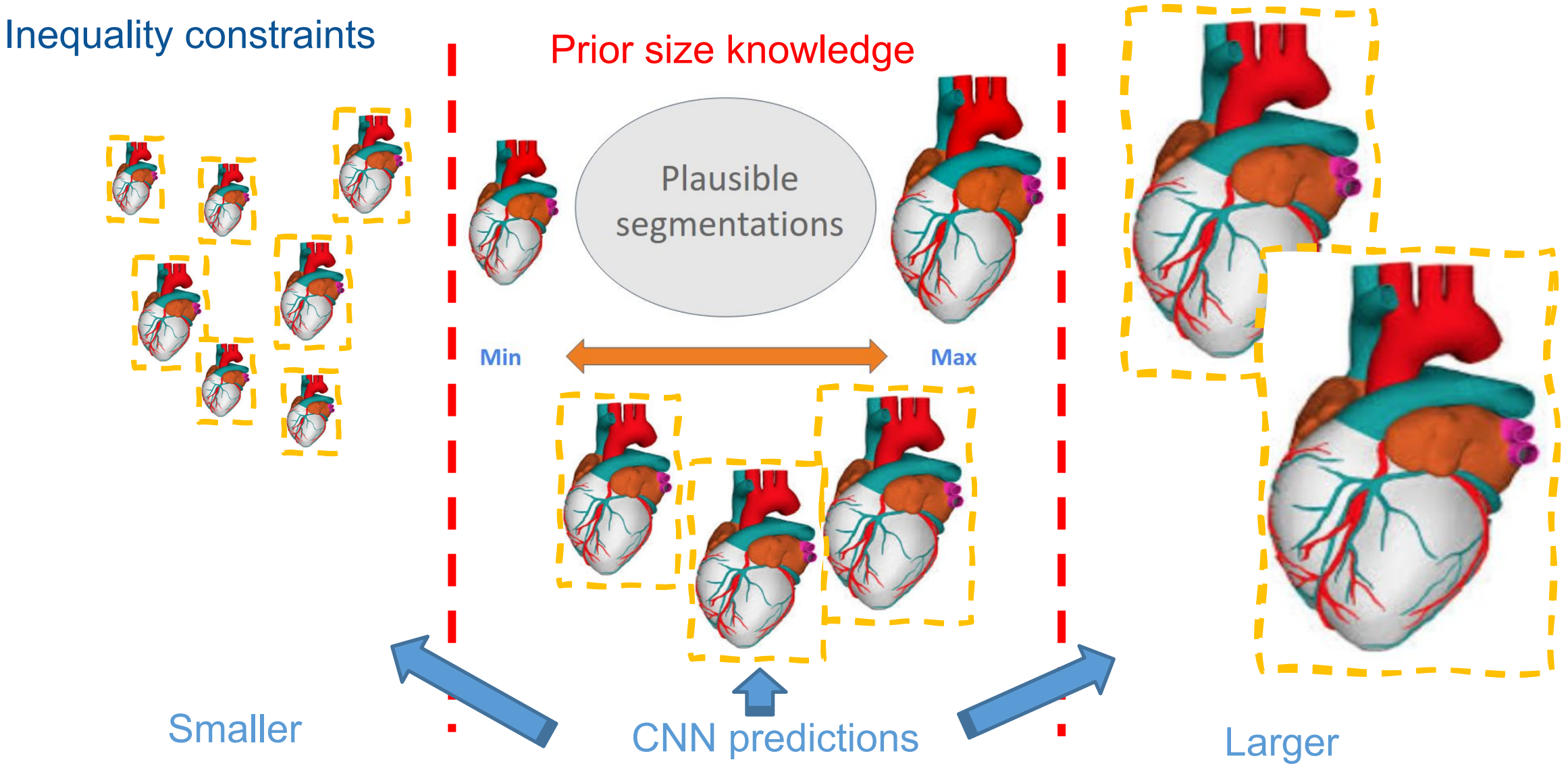
Inequality constraints



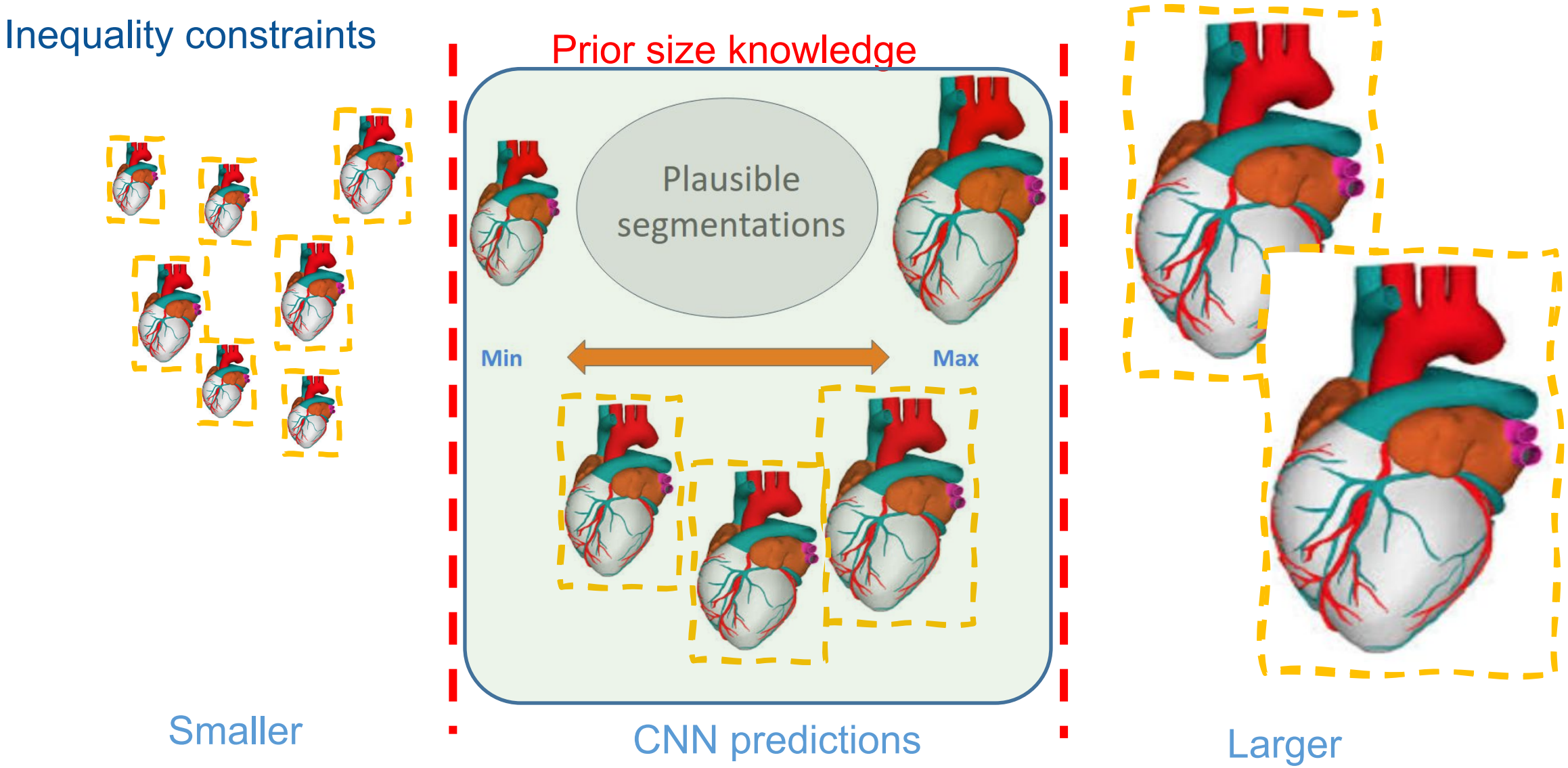
Constrained optimization (in CNNs)



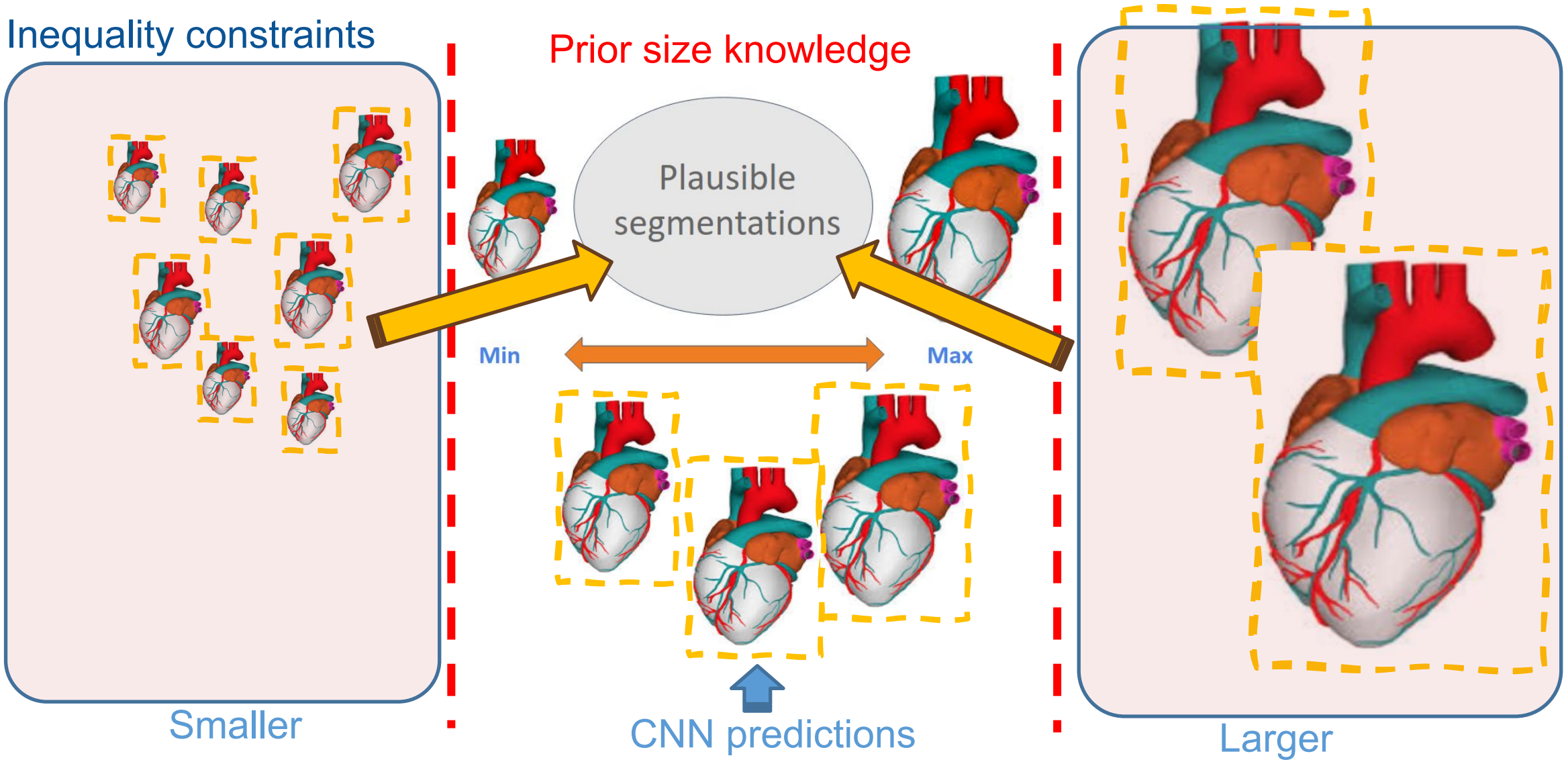
Constrained optimization (in CNNs)



Constrained optimization (in CNNs)

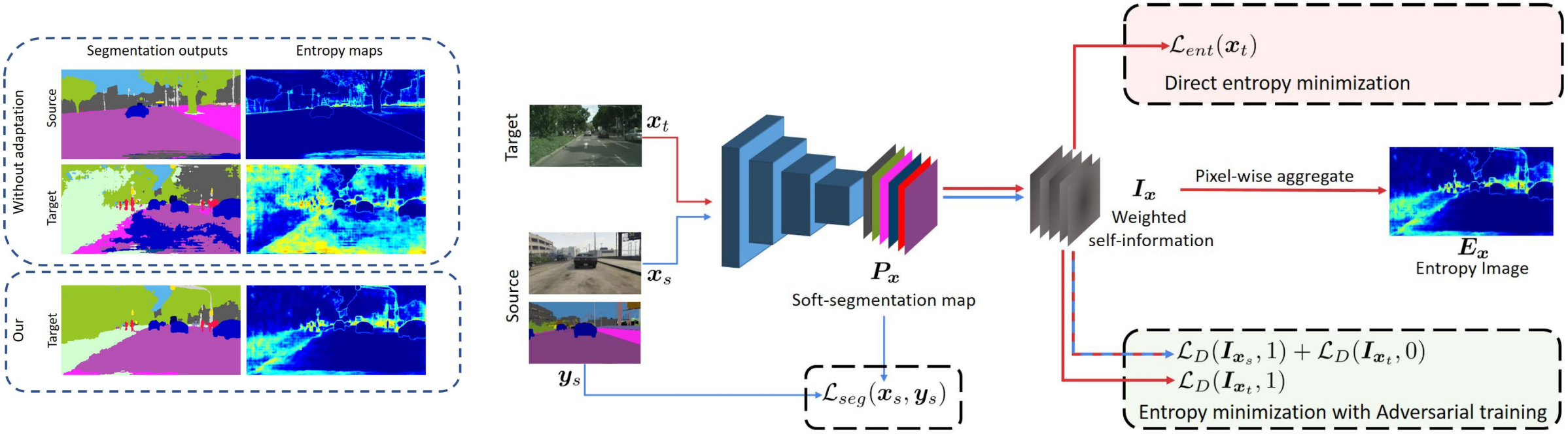


Constrained optimization (in CNNs)



Constrained optimization (in CNNs)

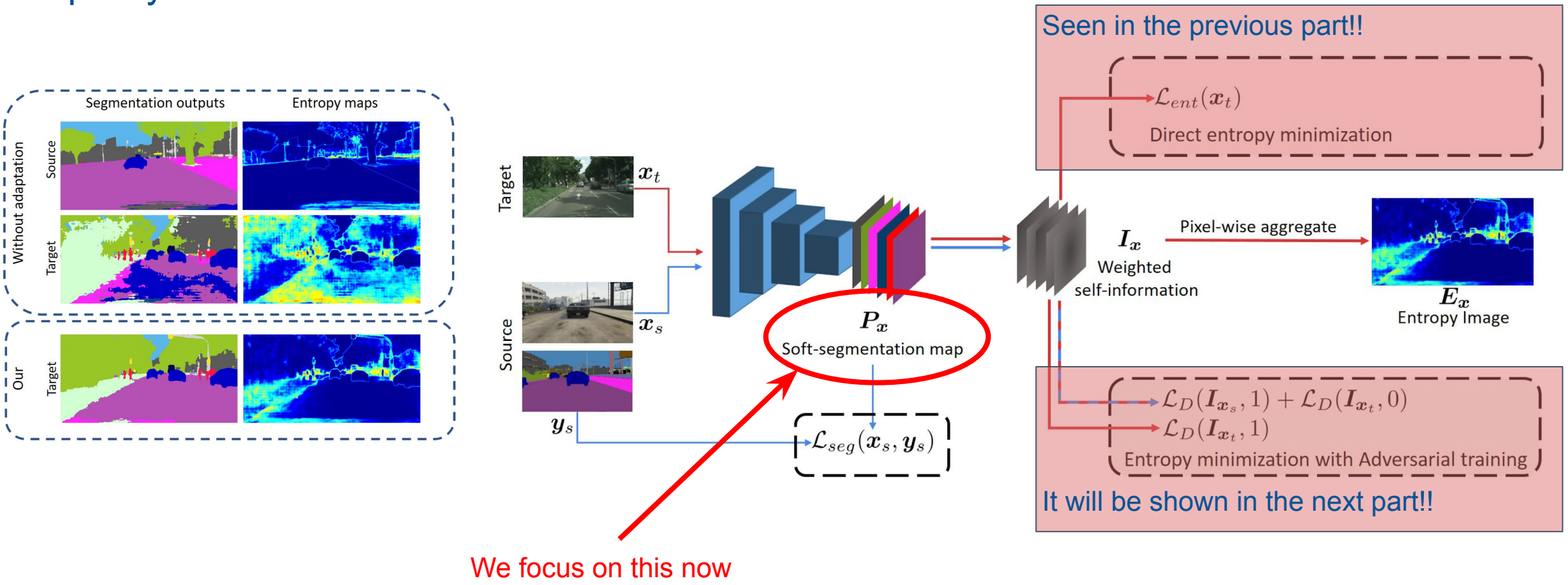
Inequality constraints



Images from [Vu et al., ADVENT: Adversarial Entropy Minimization for Domain Adaptation in Semantic Segmentation. CVPR'19]

Constrained optimization (in CNNs)

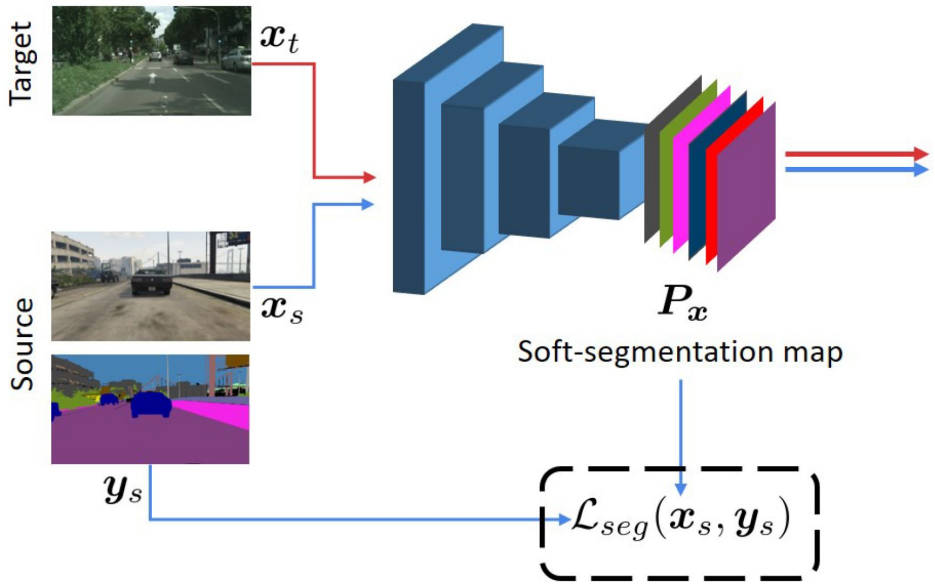
Inequality constraints



Images from [Vu et al., ADVENT: Adversarial Entropy Minimization for Domain Adaptation in Semantic Segmentation. CVPR'19]

Constrained optimization (in CNNs)

Inequality constraints



Class-ratio priors

$$\mathcal{L}_{cp}(x_t) = \sum_{c=1}^{C} \max\left(0, \mu p_s^{(c)} - \mathbb{E}_c(P_{x_t}^{(c)})\right)$$

It relaxes the class prior constraint

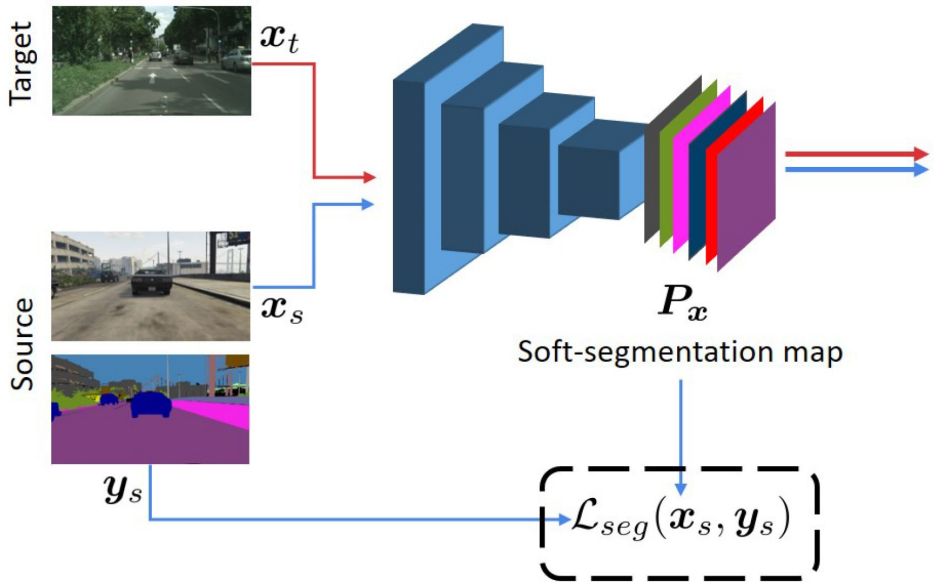
$\ell 1$-normalized histogram (source)

Estimated size on the prediction

Vu et al., ADVENT: Adversarial Entropy Minimization for Domain Adaptation in Semantic Segmentation. CVPR'19

Constrained optimization (in CNNs)

Inequality constraints



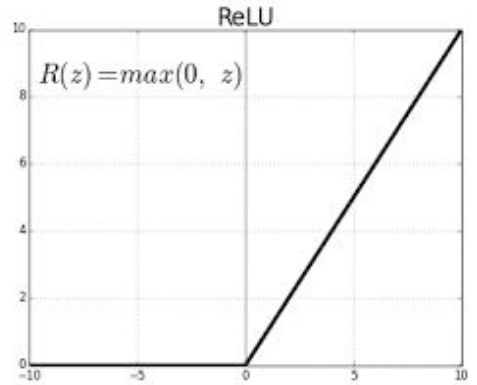
Class-ratio priors

$$\mathcal{L}_{cp}(x_t) = \sum_{c=1}^{C} \max\left(0, \mu p_s^{(c)} - \mathbb{E}_c(P_{x_t}^{(c)})\right)$$

It relaxes the class prior constraint

$\ell 1$-normalized histogram (source)

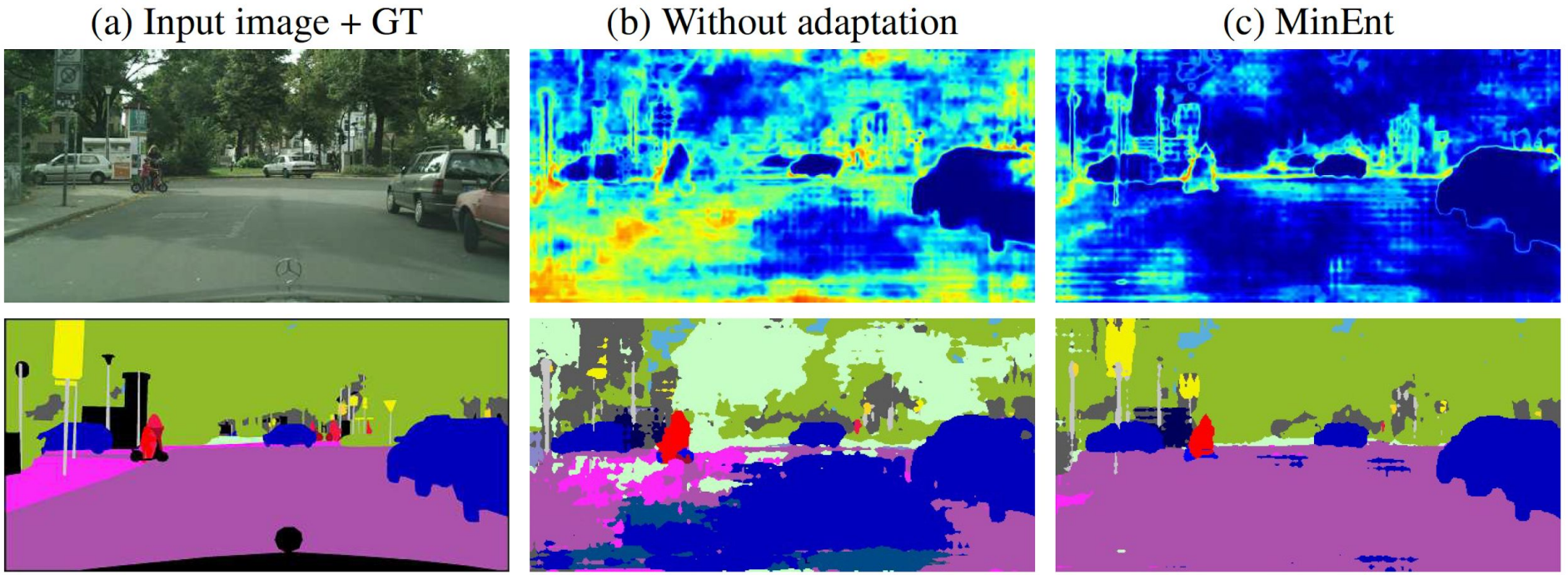
Estimated size on the prediction



Vu et al., ADVENT: Adversarial Entropy Minimization for Domain Adaptation in Semantic Segmentation. CVPR'19

Constrained optimization (in CNNs)

Inequality constraints



Images from [Vu et al., ADVENT: Adversarial Entropy Minimization for Domain Adaptation in Semantic Segmentation. CVPR'19]

Constrained optimization (in CNNs)

Inequality constraints

Information is given in the form of image-tags

Suppression $\sum_{p \in \Omega} s_\theta^{p,c} \leq 0 \quad \forall c \notin C$

"Person"

[Pathak et al., Constrained Convolutional Neural Networks for Weakly Supervised Segmentation, ICCV'15]

Constrained optimization (in CNNs)

Inequality constraints

Information is given in the form of image-tags

Inclusion (or existence)

$$\sum_{p \in \Omega} s_\theta^{p,c} \geq 1 \quad \forall c \in C$$

"Car"

[Pathak et al., Constrained Convolutional Neural Networks for Weakly Supervised Segmentation, ICCV'15]

Constrained optimization (in CNNs)

Inequality constraints

Information is given in the form of image-tags

Target Size
a > 1

$$\sum_{p \in \Omega} s_\theta^{p,c} \geq a \quad \forall c \in C$$

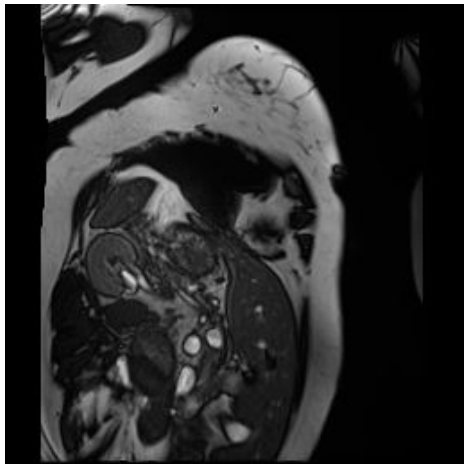
"Car"

[Pathak et al., Constrained Convolutional Neural Networks for Weakly Supervised Segmentation, ICCV'15]

Constrained optimization (in CNNs)

How we can benefit from this in the medical domain?

No cavity



Cavity

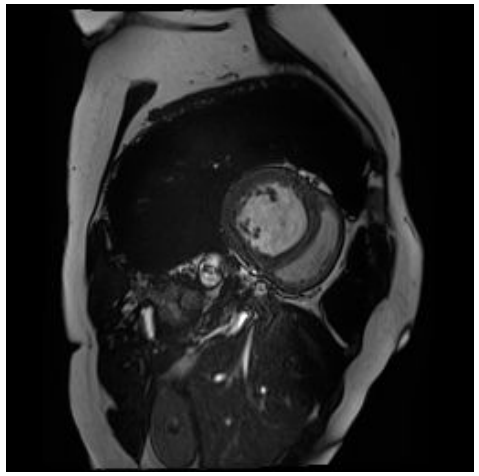


Image-tag information

$$\sum_{p \in \Omega} s_\theta^{p,c} \leq 0$$

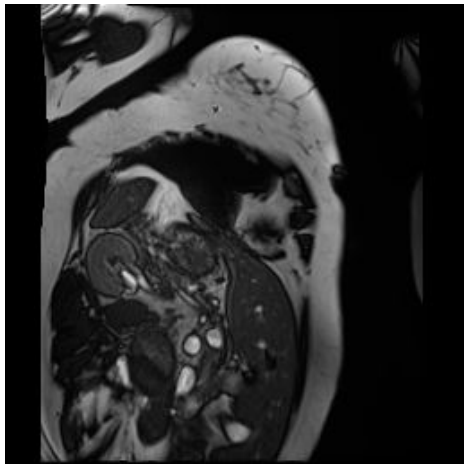
For negative image tags

[Kervadec et al., Constrained-CNN losses for weakly supervised segmentation. MedIA 2019]

Constrained optimization (in CNNs)

How we can benefit from this in the medical domain?

No cavity



Cavity

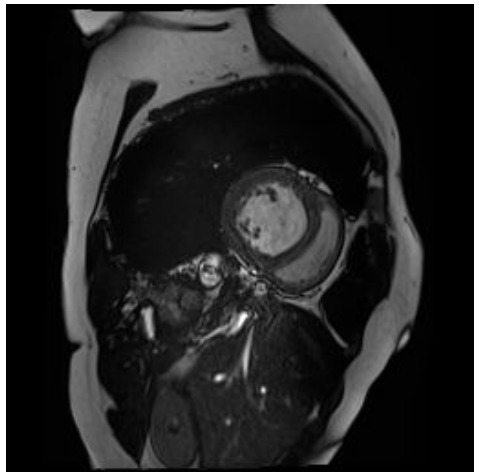
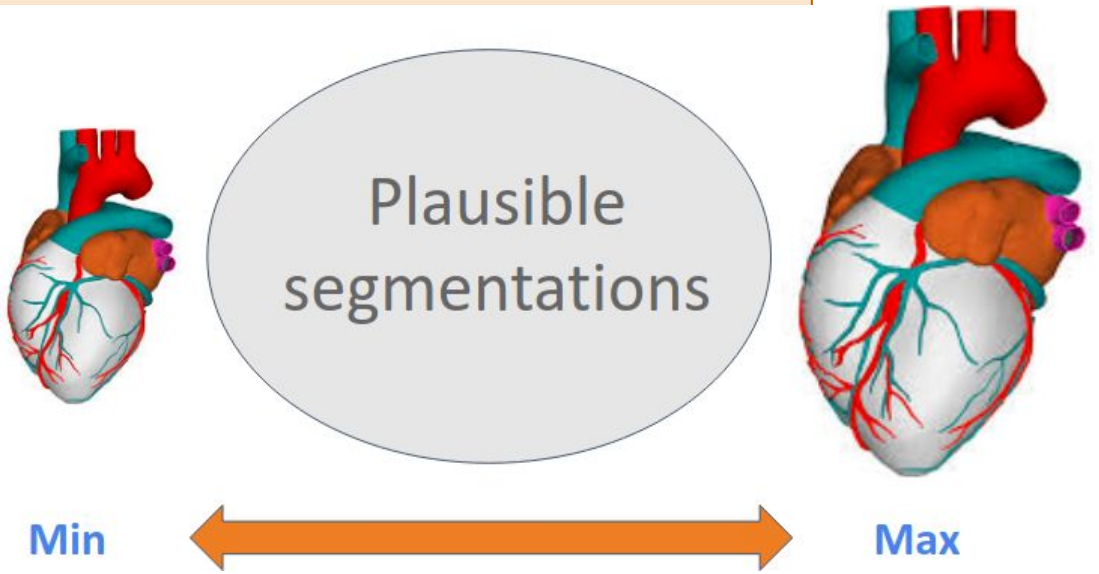


Image-tag information

$$\sum_{p \in \Omega} s_\theta^{p,c} \leq 0$$

For negative image tags



Size information

$$min \leq \sum_{p \in \Omega} s_\theta^{p,c} \leq max$$

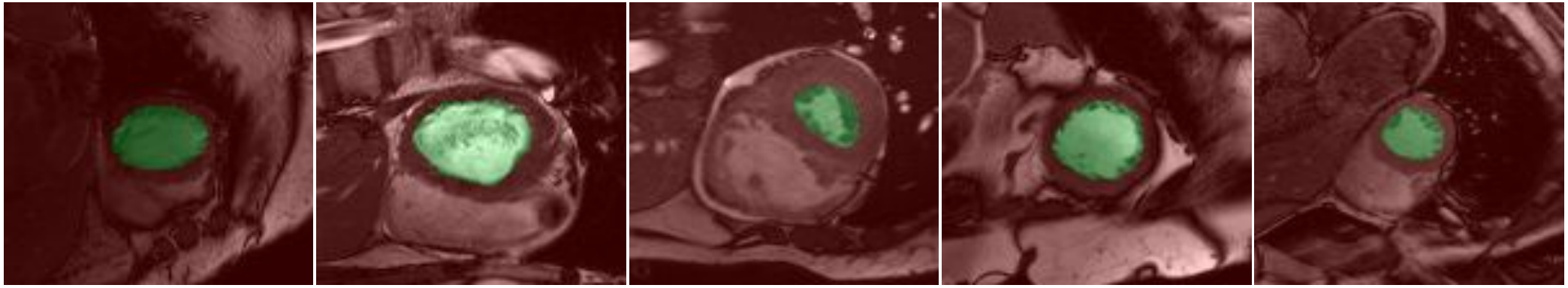
For positive image tags

[Kervadec et al., Constrained-CNN losses for weakly supervised segmentation. MedIA 2019]

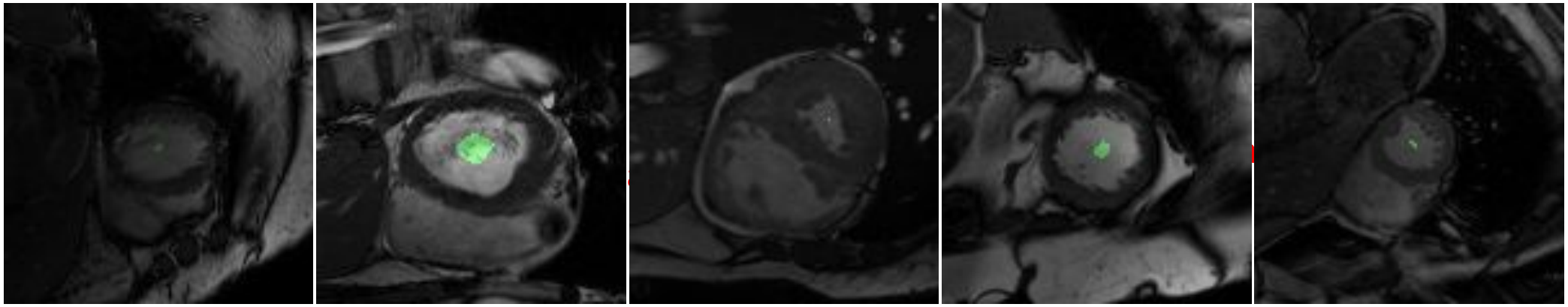
Constrained optimization (in CNNs)

Inequality constraints (e.g, L2 penalty)

Include some scribble/point annotations



Full annotations



Partial annotations for cross-entropy

[Kervadec et al., Constrained-CNN losses for weakly supervised segmentation. MedIA 2019]

Constrained optimization (in CNNs)

Inequality constraints (e.g, L2 penalty)

Objective

$$\min_{\boldsymbol{\theta}} \mathcal{H}(S) \quad \text{s.t} \quad a \leq \sum_{p \in \Omega} s_\theta^{p,c} \leq b \quad \Rightarrow \quad \mathcal{H}(S) + \lambda \mathcal{C}\left(V_S\right)$$

[Kervadec et al., Constrained-CNN losses for weakly supervised segmentation. MedIA 2019]

Constrained optimization (in CNNs)

Inequality constraints (e.g, L2 penalty)

Objective

$$\min_{\boldsymbol{\theta}} \mathcal{H}(S) \quad \text{s.t} \quad a \leq \sum_{p \in \Omega} s_\theta^{p,c} \leq b \quad \Rightarrow \quad \mathcal{H}(S) + \lambda \mathcal{C}\left(V_S\right)$$

$$\mathcal{H}(S) = -\sum_{p \in \mathcal{L}} log(s_\theta^p)$$

On annotated pixels

[Kervadec et al., Constrained-CNN losses for weakly supervised segmentation. MedIA 2019]

Constrained optimization (in CNNs)

Inequality constraints (e.g, L2 penalty)

Objective

$$\min_{\boldsymbol{\theta}} \mathcal{H}(S) \quad \text{s.t} \quad a \leq \sum_{p \in \Omega} s_\theta^{p,c} \leq b \quad \Rightarrow \quad \mathcal{H}(S) + \lambda \mathcal{C}(V_S)$$

$$\mathcal{H}(S) = -\sum_{p \in \mathcal{L}} log(s_\theta^p)$$

On annotated pixels

$$\mathcal{C}(V_S) = \begin{cases} (V_S - a)^2, & \text{if } V_S < a \\ (V_S - b)^2, & \text{if } V_S > b \\ 0, & \text{otherwise} \end{cases}$$

[Kervadec et al., Constrained-CNN losses for weakly supervised segmentation. MedIA 2019]

Constrained optimization (in CNNs)

Inequality constraints (e.g, L2 penalty)

Objective

$$\min_{\boldsymbol{\theta}} \mathcal{H}(S) \quad \text{s.t} \quad a \leq \sum_{p \in \Omega} s_\theta^{p,c} \leq b \quad \Rightarrow \quad \mathcal{H}(S) + \lambda \mathcal{C}\left(V_S\right)$$

Back-propagation

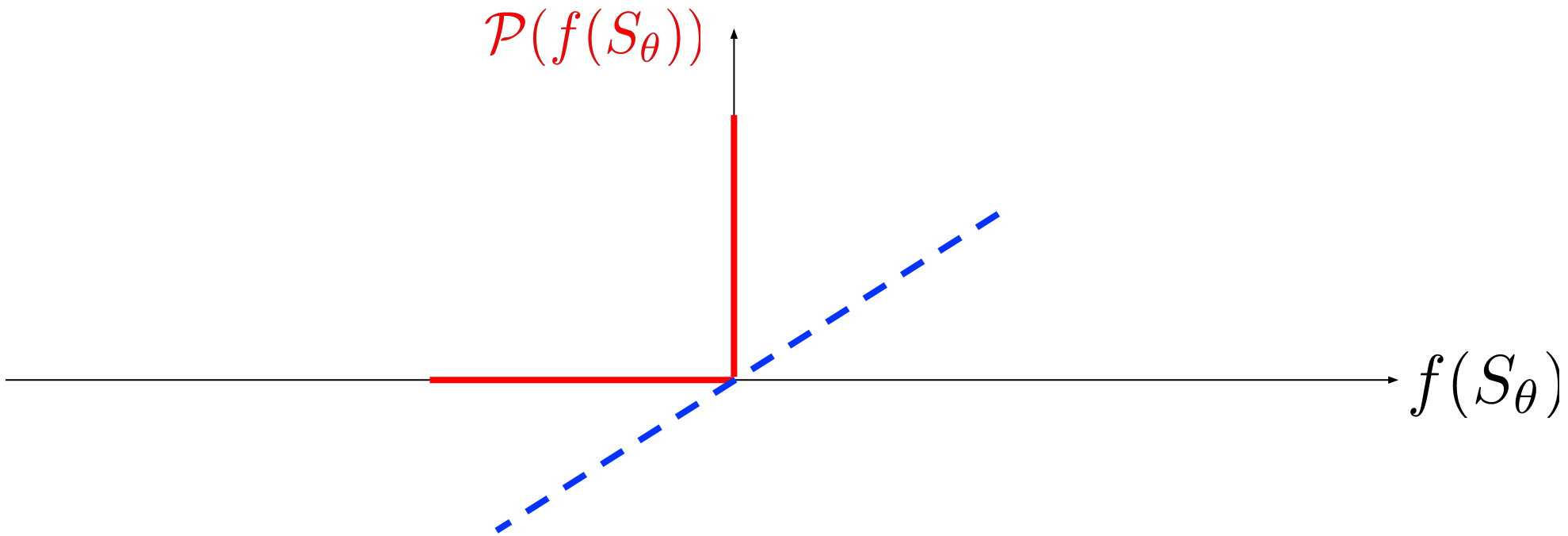
$$-\frac{\partial \mathcal{C}(V_S)}{\partial \boldsymbol{\theta}} \propto \begin{cases} (a - V_S)\, \frac{\partial S_p}{\partial \boldsymbol{\theta}}, & \text{if } V_S < a \\ (b - V_S)\, \frac{\partial S_p}{\partial \boldsymbol{\theta}}, & \text{if } V_S > b \\ 0, & \text{otherwise} \end{cases}$$

[Kervadec et al., Constrained-CNN losses for weakly supervised segmentation. MedIA 2019]

Constrained optimization (in CNNs)

Inequality constraints (Why not Lagrangian primal/dual?)

$$\mathcal{L}(S_\theta, \lambda) = \mathcal{E}(\theta) + \lambda f(S_\theta)$$

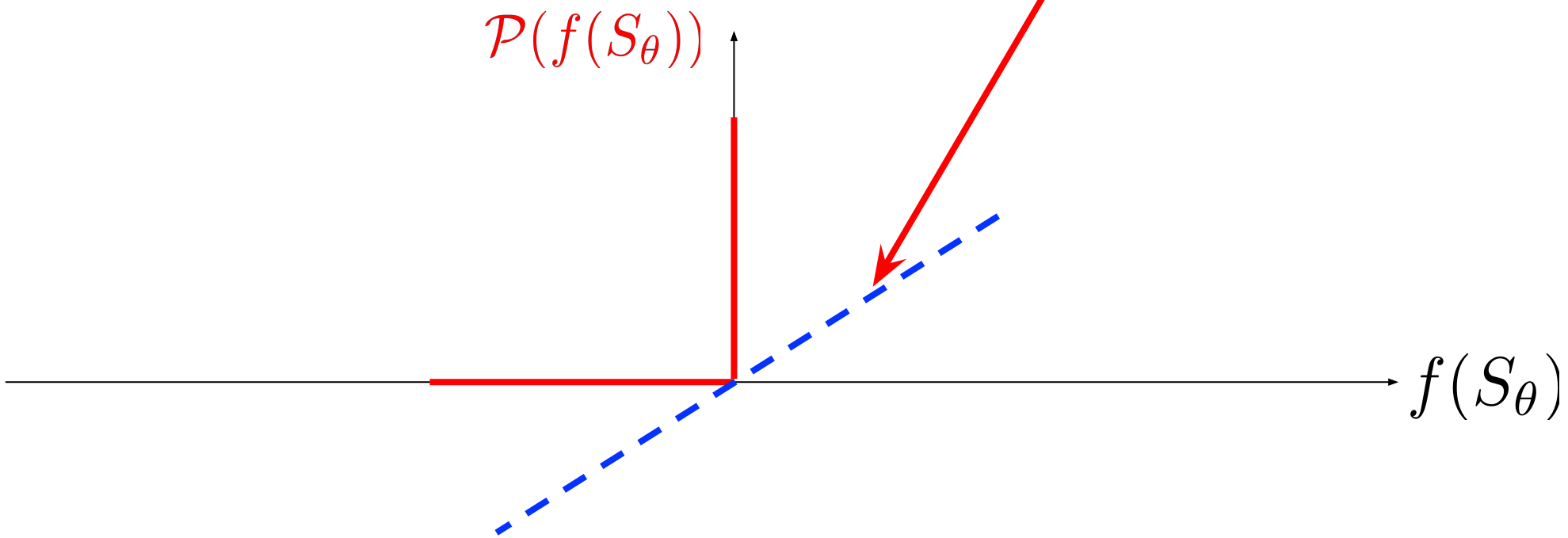


[Boyd and Vandenberghe, Convex optimization, 2004]

Constrained optimization (in CNNs)

Inequality constraints (Why not Lagrangian primal/dual?)

$$\mathcal{L}(S_\theta, \lambda) = \mathcal{E}(\theta) + \lambda f(S_\theta)$$

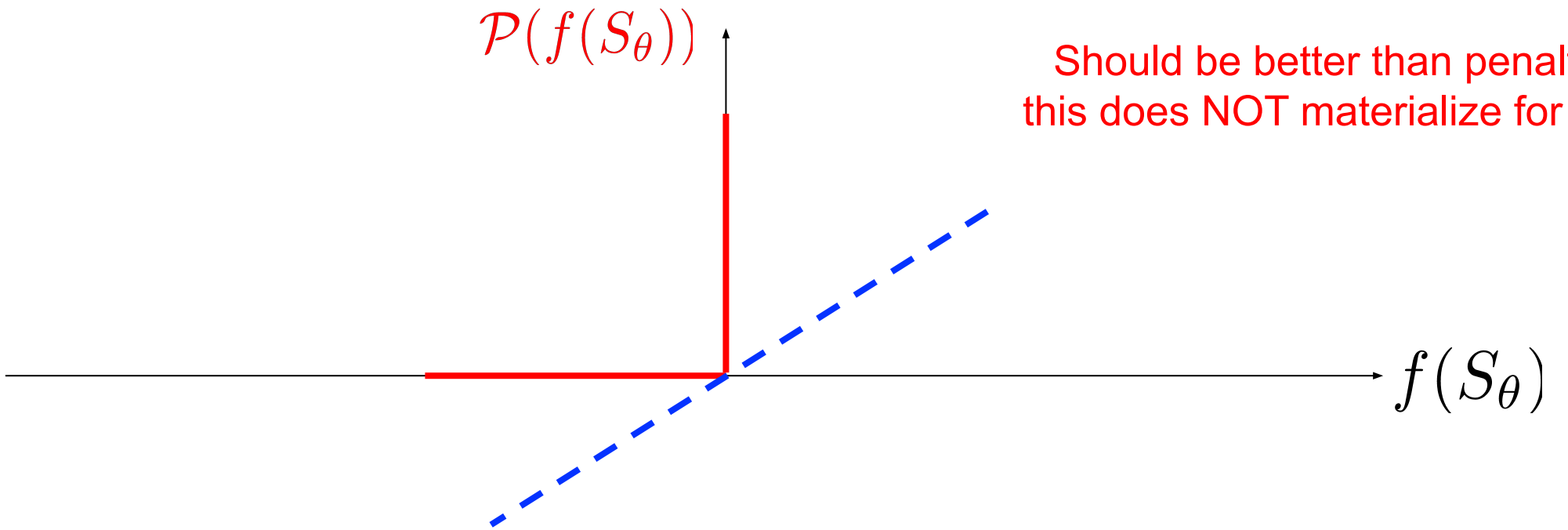


[Boyd and Vandenberghe, Convex optimization, 2004]

Constrained optimization (in CNNs)

Inequality constraints (Why not Lagrangian primal/dual?)

$$\mathcal{L}(S_\theta, \lambda) = \mathcal{E}(\theta) + \lambda f(S_\theta)$$



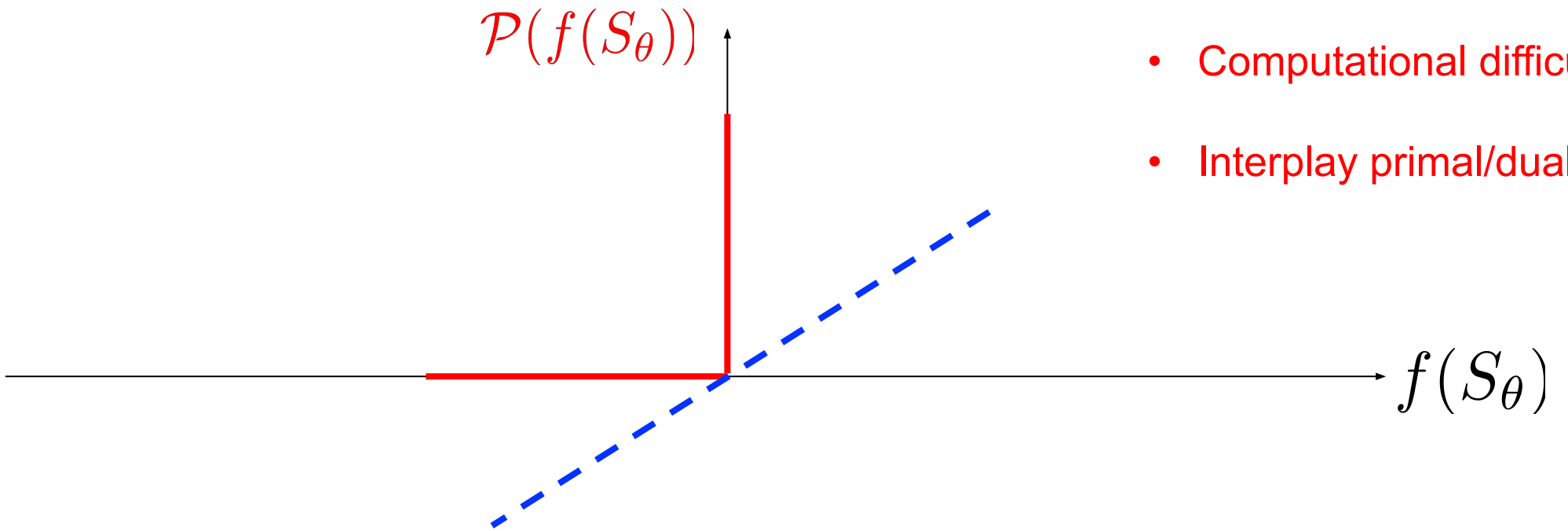
Should be better than penalties, but this does NOT materialize for CNNs!!!!!

[Boyd and Vandenberghe, Convex optimization, 2004]

Constrained optimization (in CNNs)

Inequality constraints (Why not Lagrangian primal/dual?)

$$\mathcal{L}(S_\theta, \lambda) = \mathcal{E}(\theta) + \lambda f(S_\theta)$$



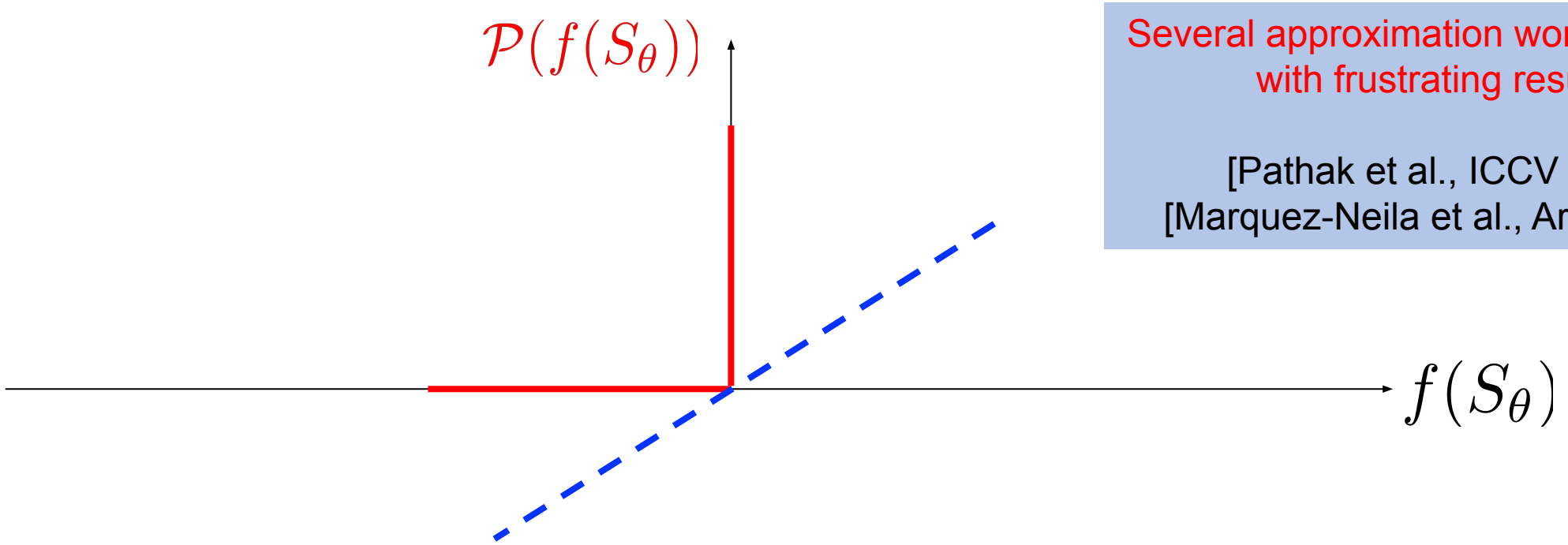
- Computational difficulty
- Interplay primal/dual

[Boyd and Vandenberghe, Convex optimization, 2004]

Constrained optimization (in CNNs)

Inequality constraints (Why not Lagrangian primal/dual?)

$$\mathcal{L}(S_\theta, \lambda) = \mathcal{E}(\theta) + \lambda f(S_\theta)$$



Several approximation works recently, with frustrating results:

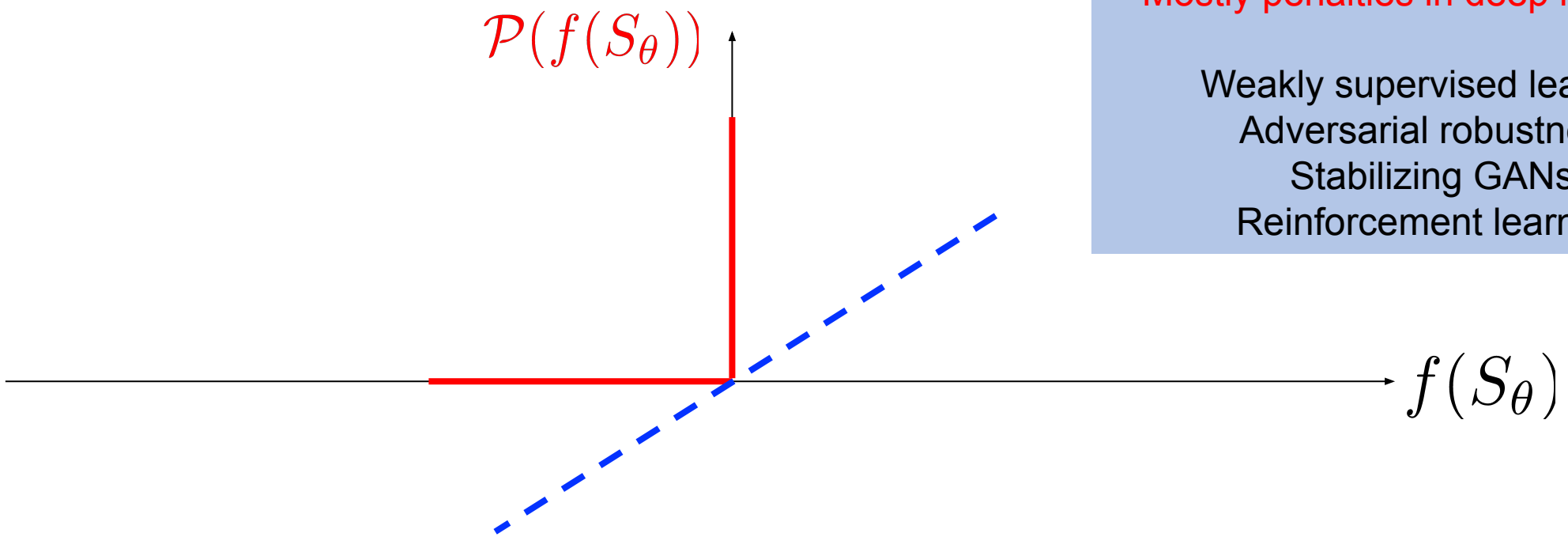
[Pathak et al., ICCV 2015]
[Marquez-Neila et al., ArXiv 2017]

[Boyd and Vandenberghe, Convex optimization, 2004]

Constrained optimization (in CNNs)

Inequality constraints (Why not Lagrangian primal/dual?)

$$\mathcal{L}(S_\theta, \lambda) = \mathcal{E}(\theta) + \lambda f(S_\theta)$$



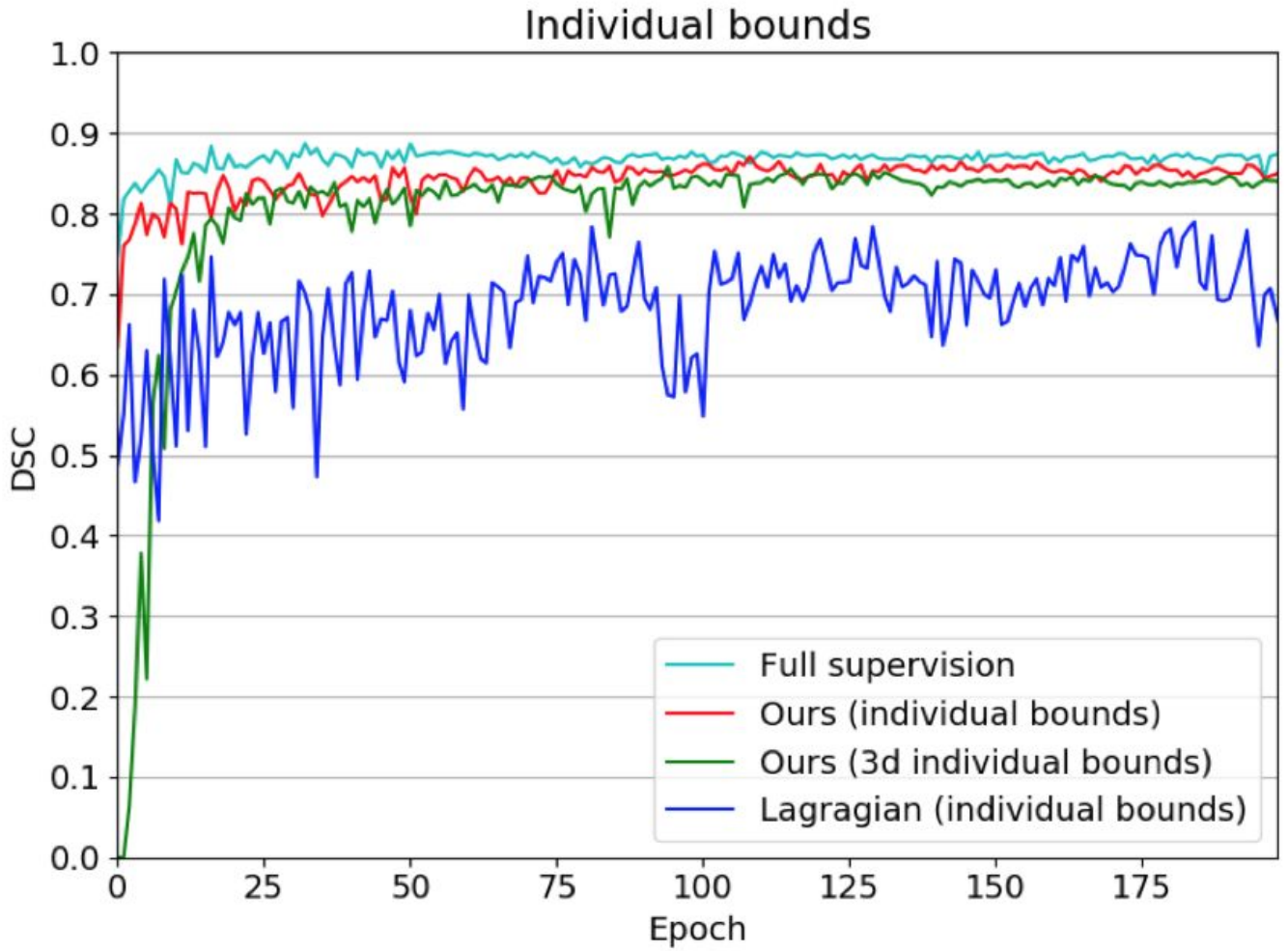
Mostly penalties in deep networks:

Weakly supervised learning
Adversarial robustness
Stabilizing GANs
Reinforcement learning

[Boyd and Vandenberghe, Convex optimization, 2004]

Constrained optimization (in CNNs)

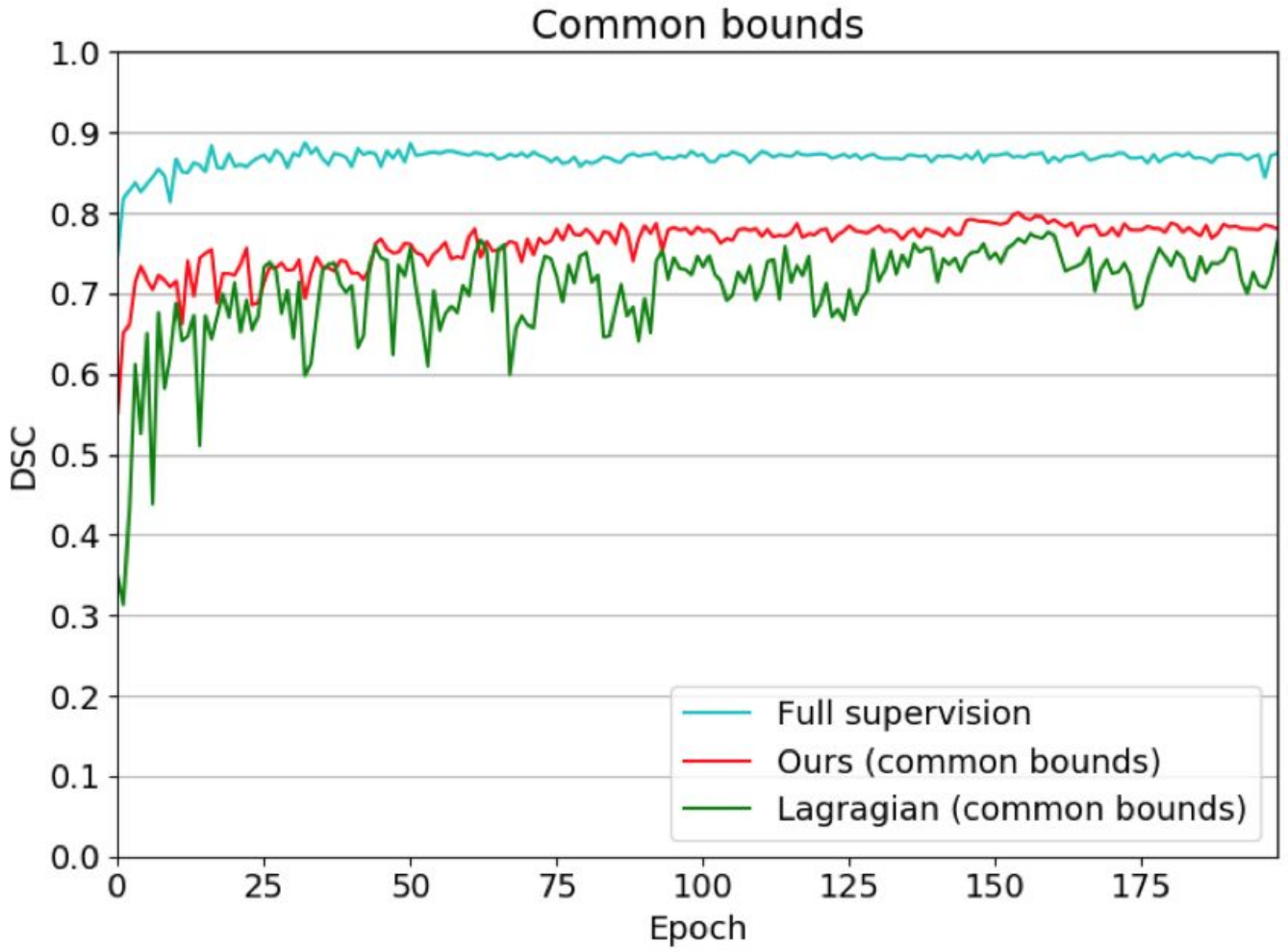
Inequality constraints (e.g, L2 penalty)



[Kervadec et al., Constrained-CNN losses for weakly supervised segmentation. MedIA 2019]

Constrained optimization (in CNNs)

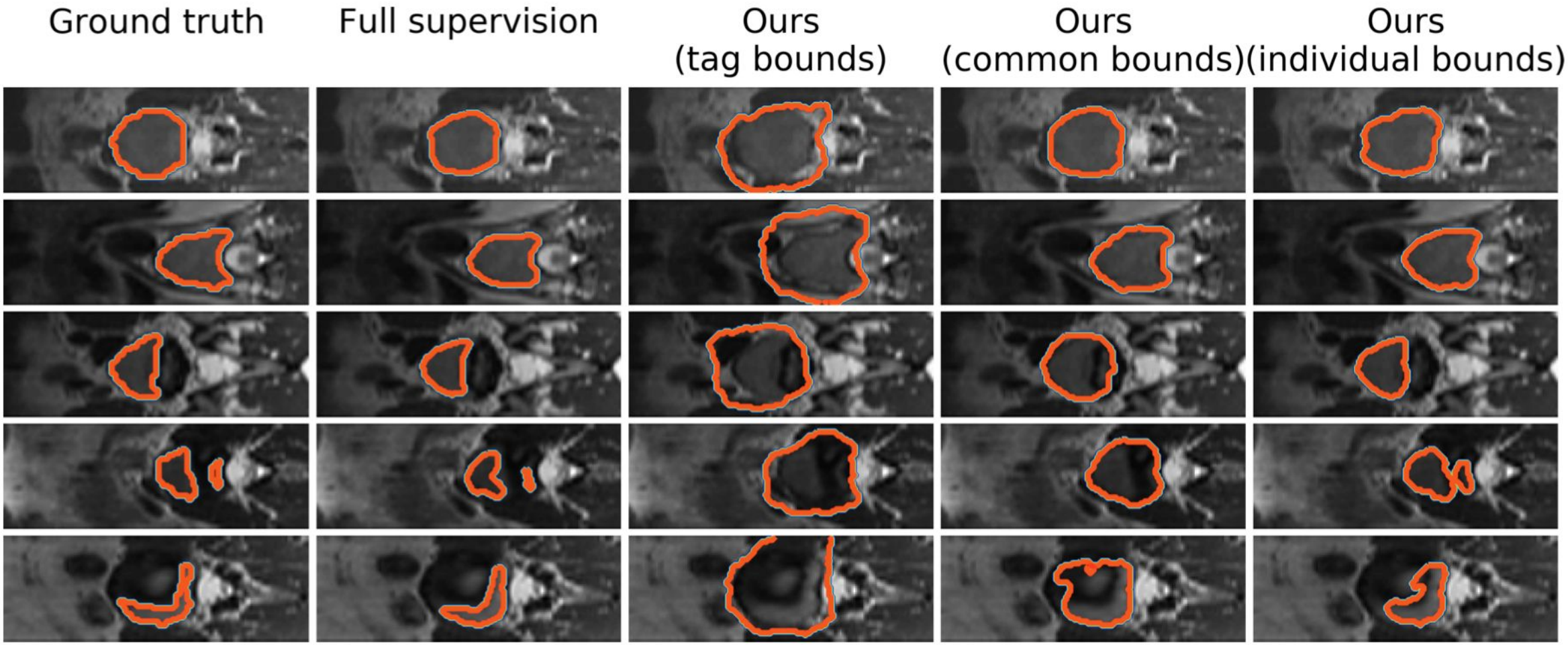
Inequality constraints (e.g, L2 penalty)



[Kervadec et al., Constrained-CNN losses for weakly supervised segmentation. MedIA 2019]

Constrained optimization (in CNNs)

Inequality constraints (e.g, L2 penalty)



[Kervadec et al., Constrained-CNN losses for weakly supervised segmentation. MedIA 2019]

Constrained optimization (in CNNs)

Inequality constraints (e.g, L2 penalty)

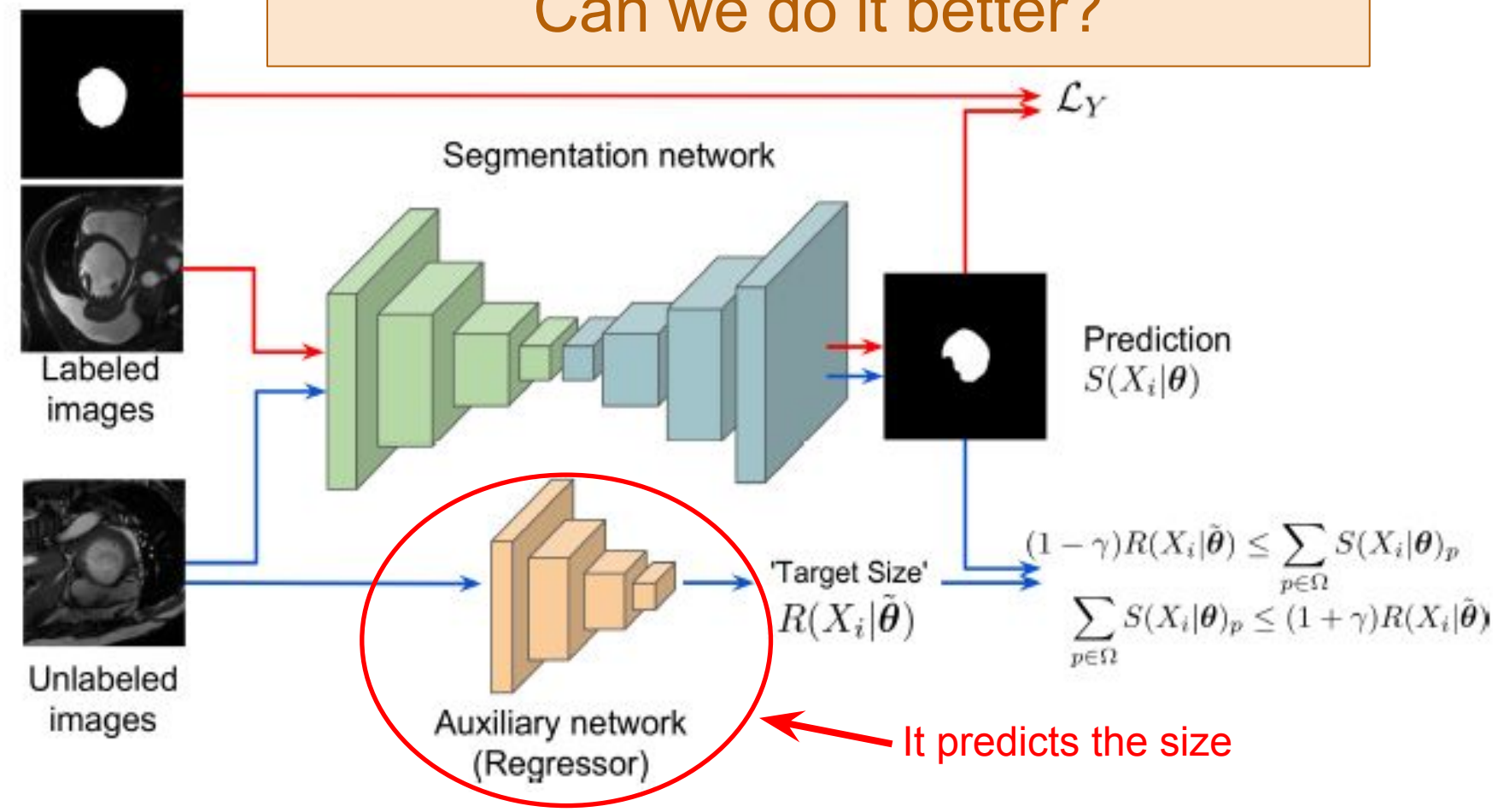
Can we do it better?

[Kervadec et al., Curriculum semi-supervised segmentation. MICCAI'19]

Constrained optimization (in CNNs)

Inequality constraints (e.g, L2 penalty)

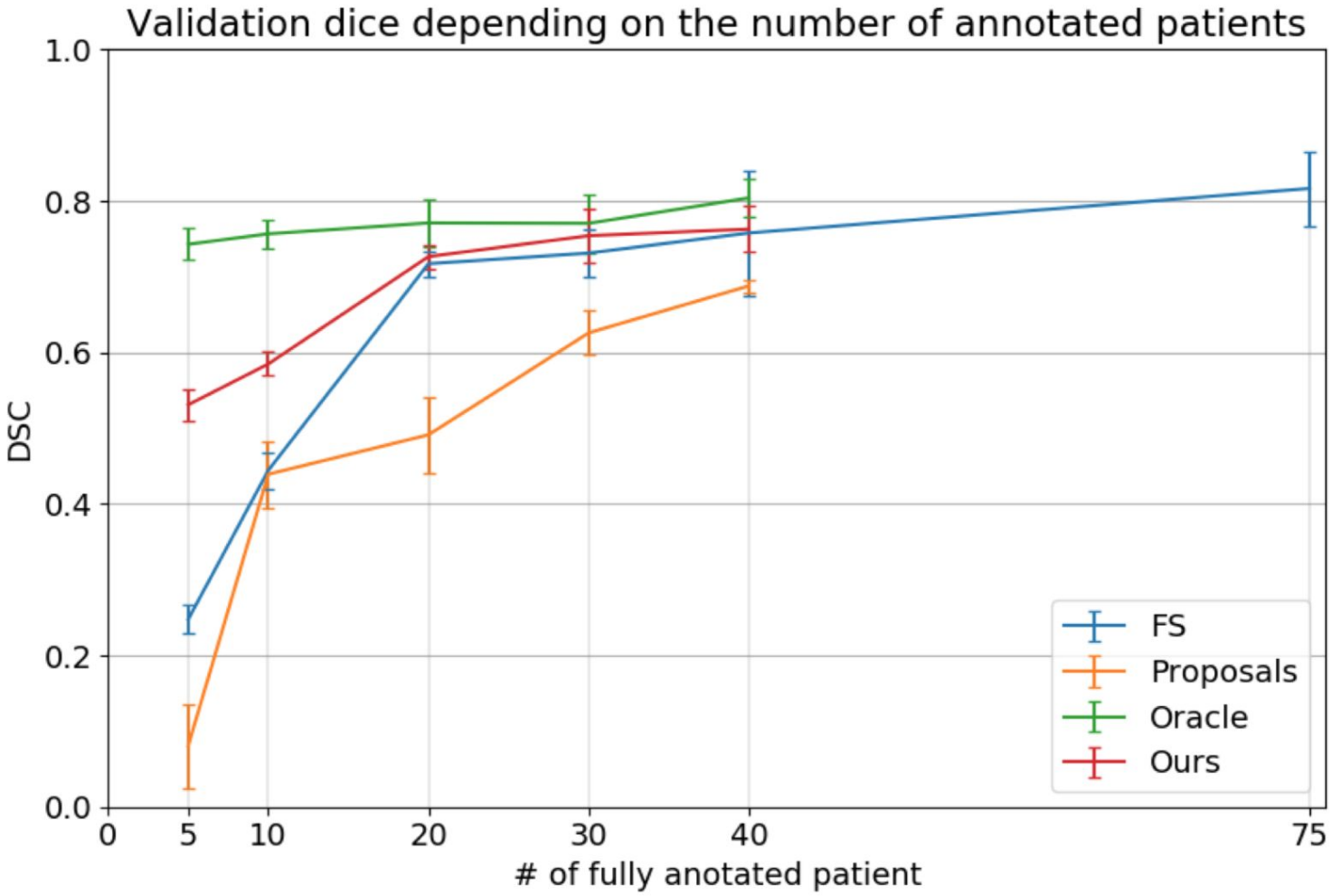
Can we do it better?



[Kervadec et al., Curriculum semi-supervised segmentation. MICCAI'19]

Constrained optimization (in CNNs)

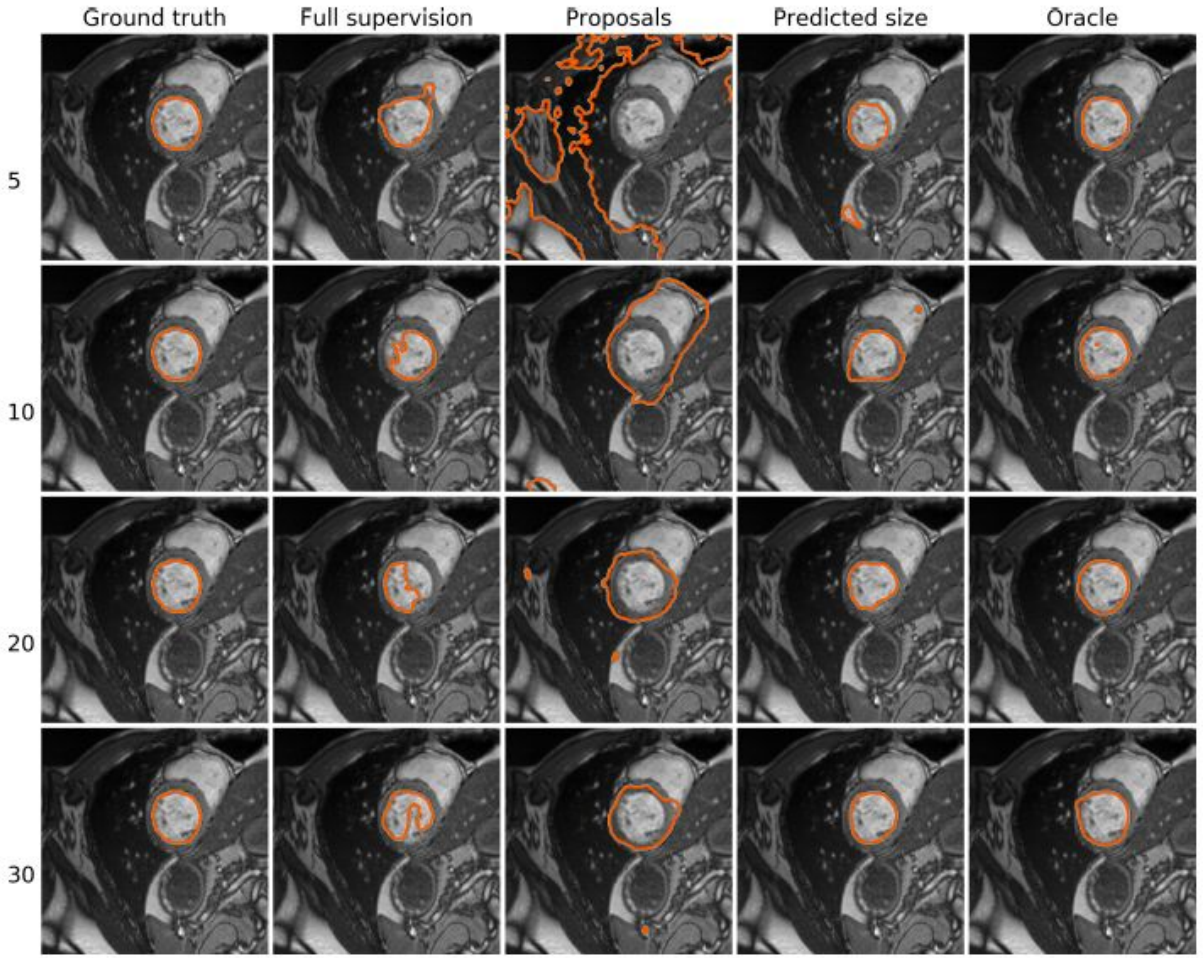
Inequality constraints (e.g, L2 penalty)



[Kervadec et al., Curriculum semi-supervised segmentation. MICCAI'19]

Constrained optimization (in CNNs)

Inequality constraints (e.g, L2 penalty)



[Kervadec et al., Curriculum semi-supervised segmentation. MICCAI'19]

Constrained optimization (in CNNs)

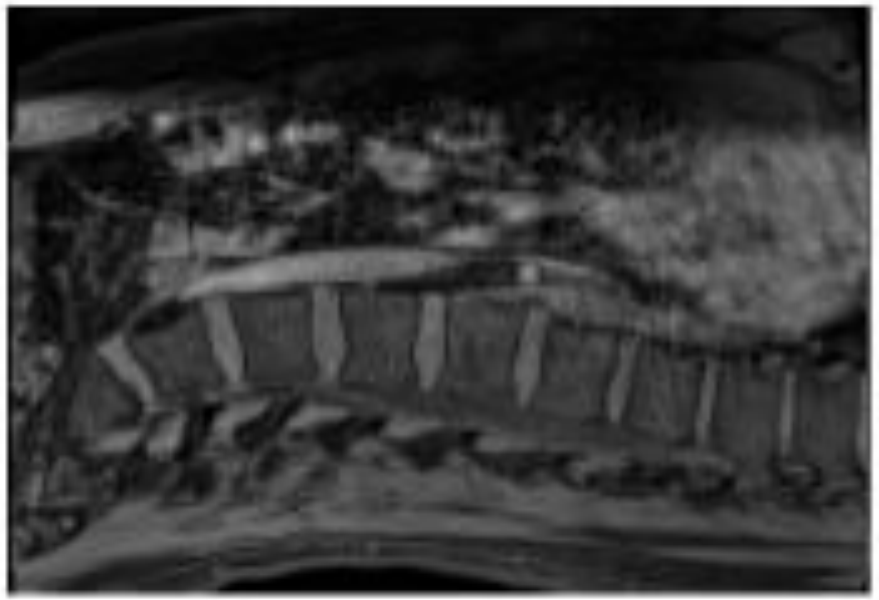
Inequality constraints (e.g, L2 penalty)

Also useful in Domain adaptation!

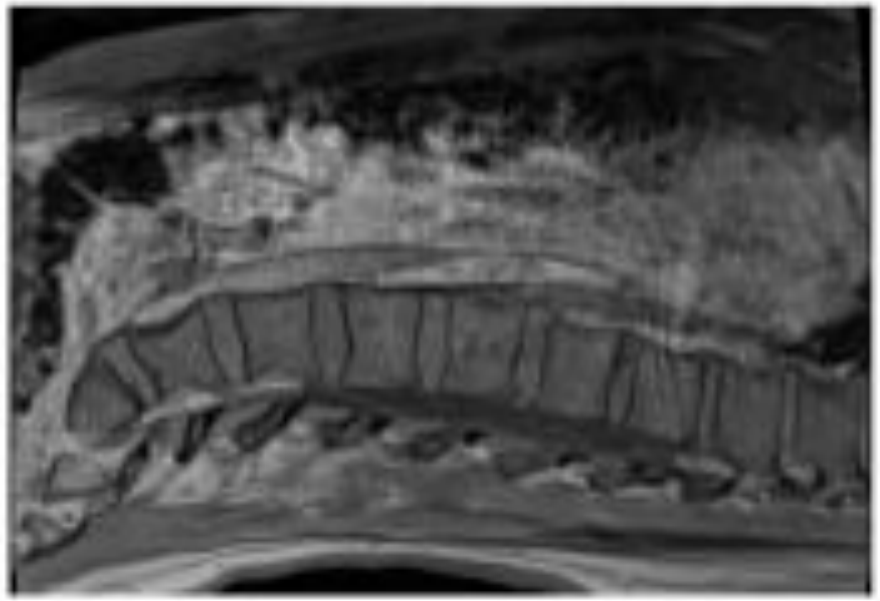
Constrained optimization (in CNNs)

Inequality constraints (e.g, L2 penalty)

Source **(labeled)** modality (WAT)



Target **(unlabeled)** modality (IP)

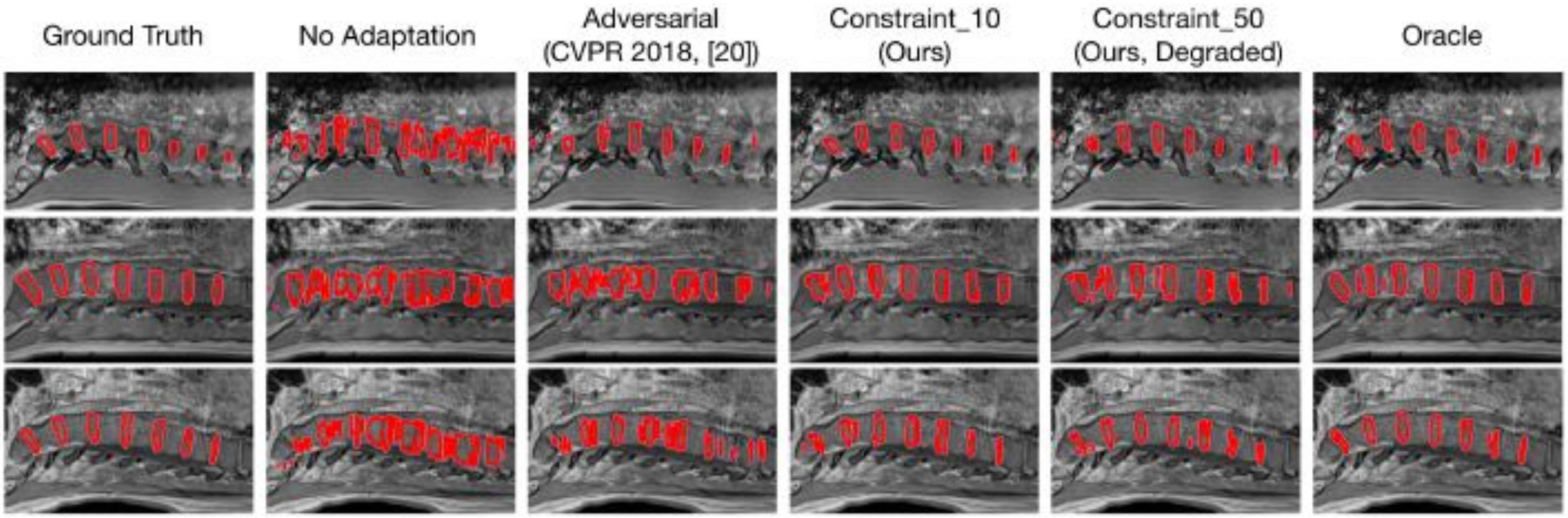


Dataset from IVD 2018 MICCAI Challenge

[Bateson et al., Constrained domain adaptation for segmentation. MICCAI'19]

Constrained optimization (in CNNs)

Inequality constraints (e.g, L2 penalty)

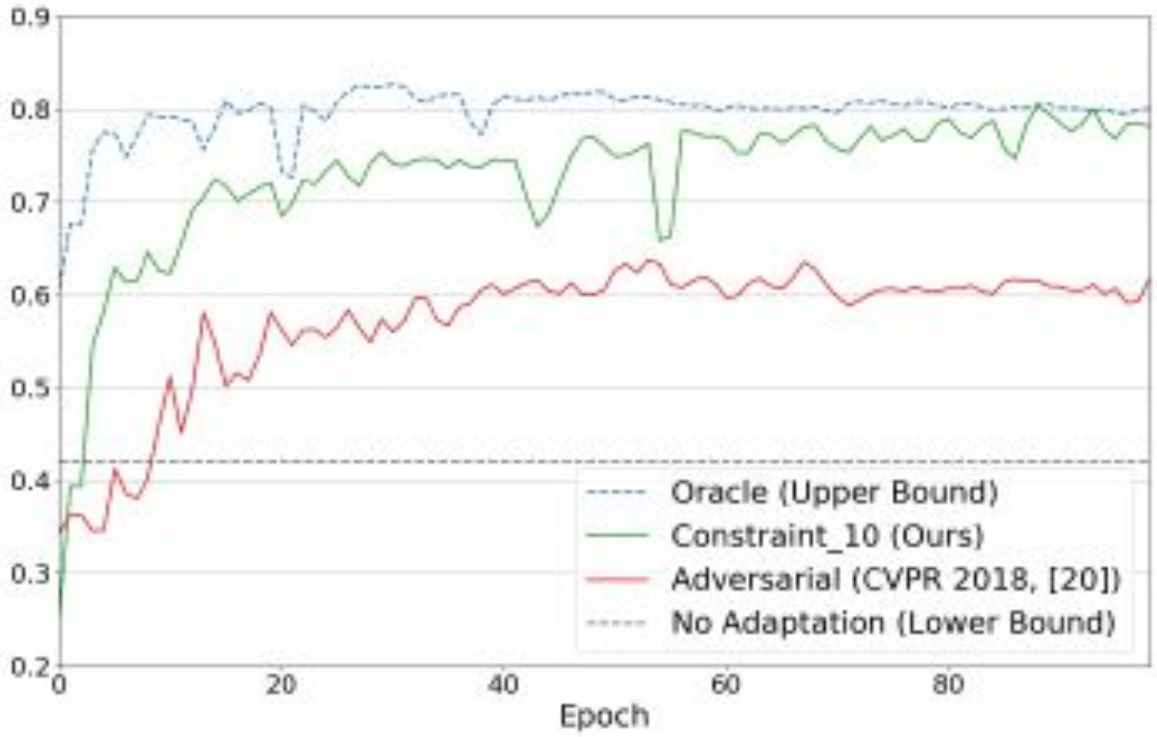


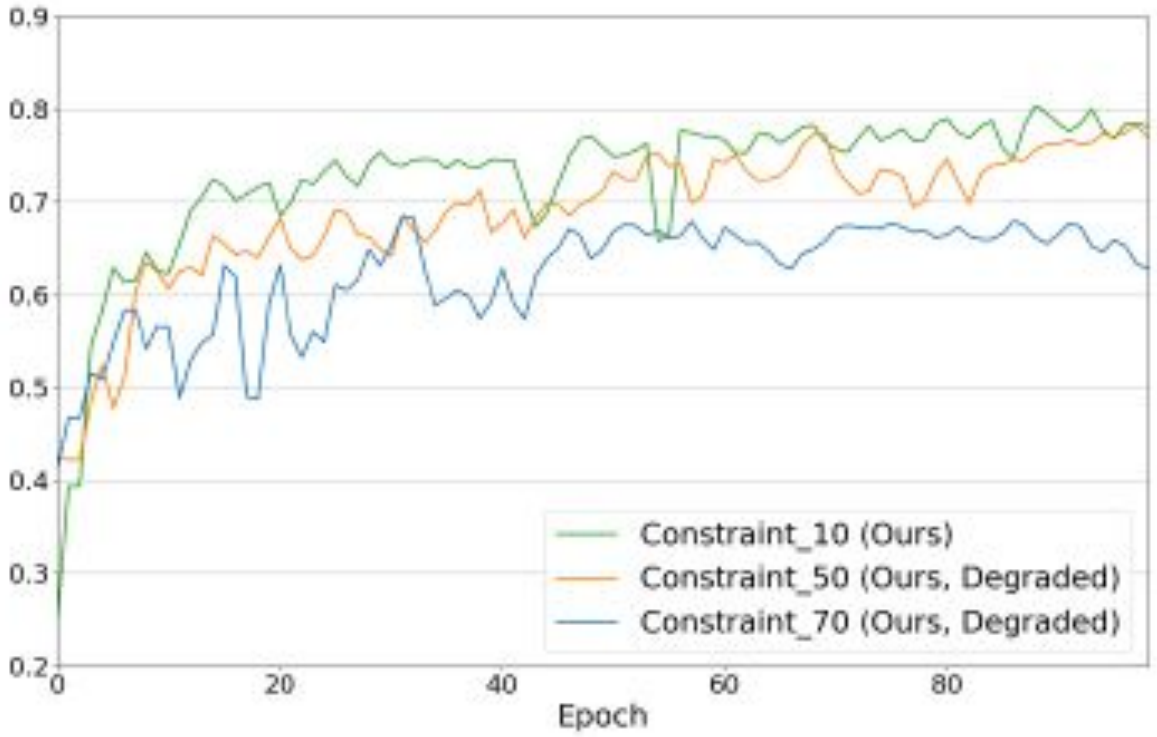
[Bateson et al., Constrained domain adaptation for segmentation. MICCAI'19]

Constrained optimization (in CNNs)

Inequality constraints (e.g, L2 penalty)

Sensitivity to bounds on the size constrain



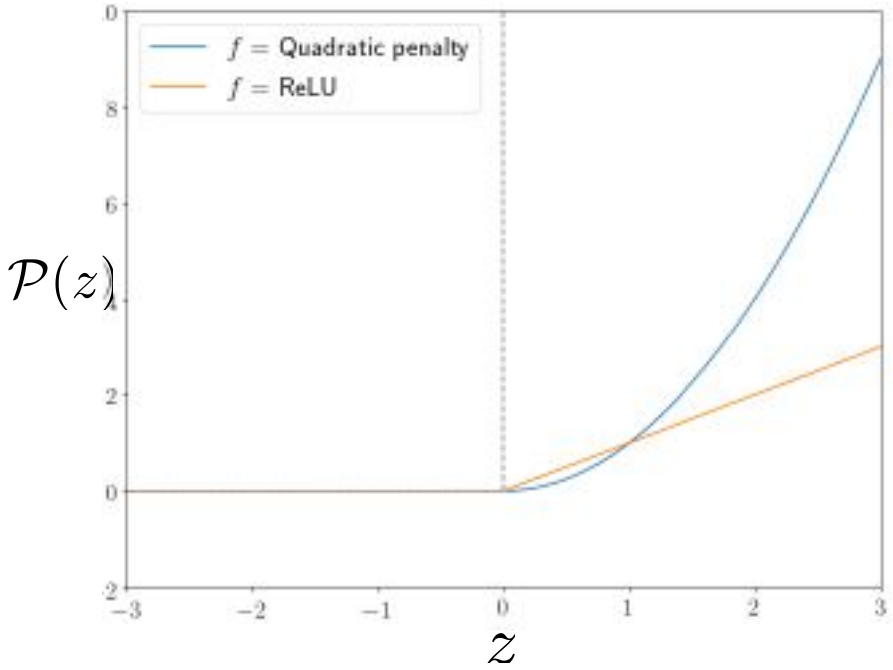


[Bateson et al., Constrained domain adaptation for segmentation. MICCAI'19]

Limitations of penalties

$$\min_{\theta} \mathcal{J}(\theta)$$

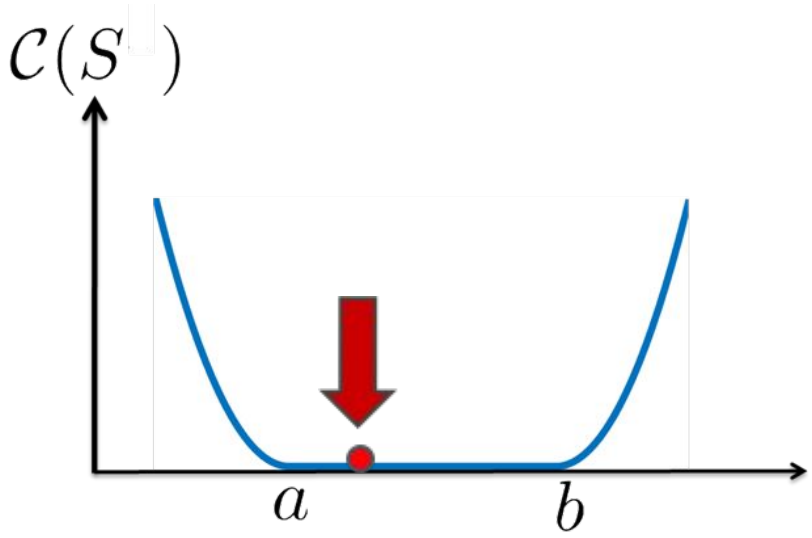
$$\mathcal{J}(\theta) = \mathcal{E}(\theta) + \lambda \mathcal{P}(f(S_\theta))$$



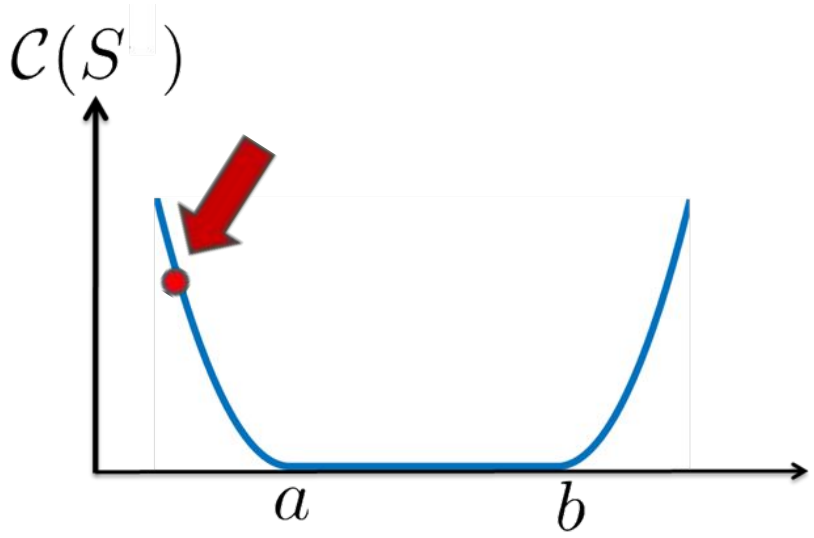
Problem: How to set the weight?
Multiples constraints make it worse?

Limitations of penalties

Multiple constraints



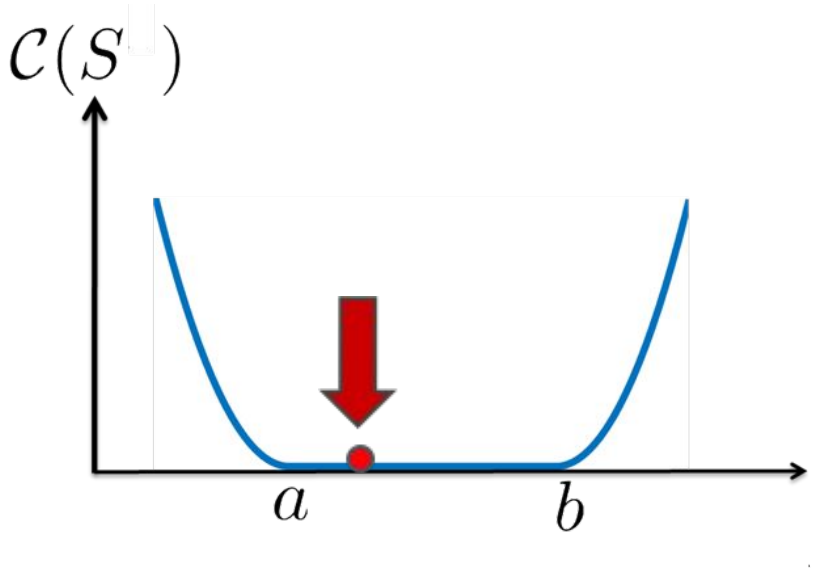
Constraint A satisfied



Constraint B violated

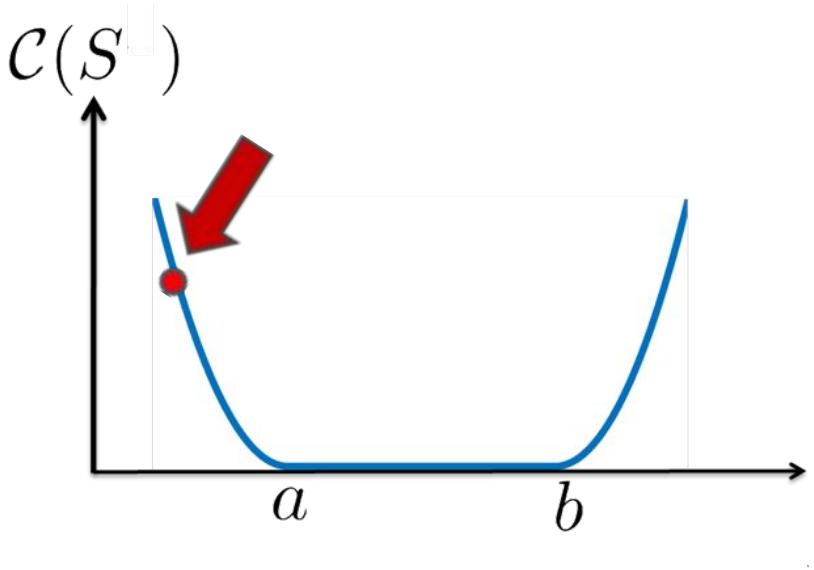
Limitations of penalties

Multiple constraints



Constraint A satisfied

Gradient is 0



Constraint B violated

Gradient is **NOT** 0

Log-barrier extensions: Approximates Lagrangian optimization but **NO** explicit dual steps

Lagrangian optimization can deal with these limitations:

- it finds automatically the optimal weights of the constraints.
- it acts as a barrier for satisfied constraints.
- it guarantees constraint satisfaction when feasible solutions exist.

[Kervadec et al., Log-barrier constrained CNNs, ArXiv 2019]

Log-barrier extensions: Approximates Lagrangian optimization but **NO** explicit dual steps

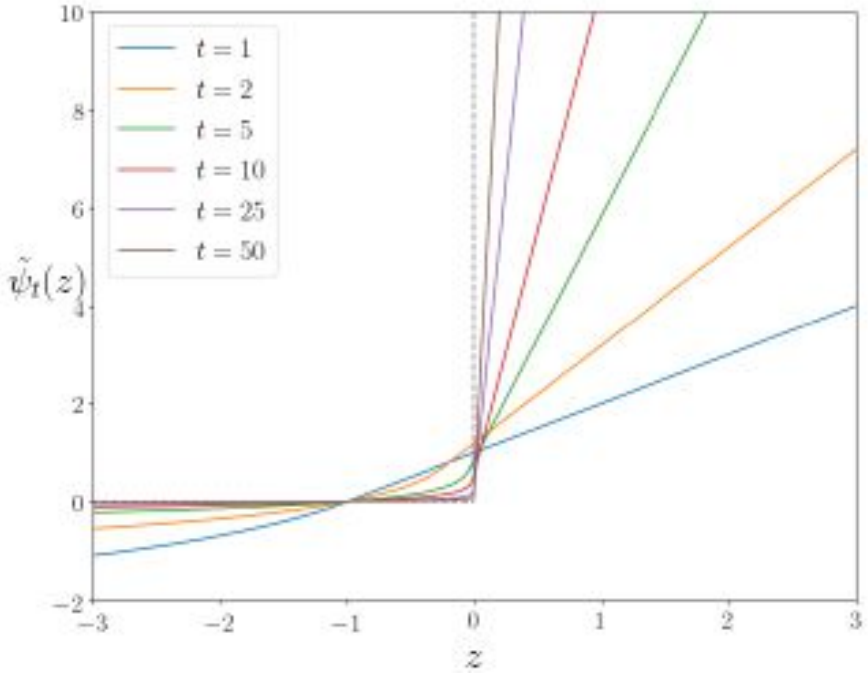
Lagrangian optimization can deal with these limitations:

- it finds automatically the optimal weights of the constraints.
- it acts as a barrier for satisfied constraints.
- it guarantees constraint satisfaction when feasible solutions exist.

$$\min_{\boldsymbol{\theta}} \mathcal{E}(\boldsymbol{\theta}) + \sum_{i=1}^{N} \tilde{\psi}_t \left(f_i(S_{\boldsymbol{\theta}}) \right) \quad \Rightarrow \quad \tilde{\psi}_t(z) = \begin{cases} -\frac{1}{t}\log(-z) & \text{if } z \leq -\frac{1}{t^2} \\ tz - \frac{1}{t}\log(\frac{1}{t^2}) + \frac{1}{t} & \text{otherwise} \end{cases}$$

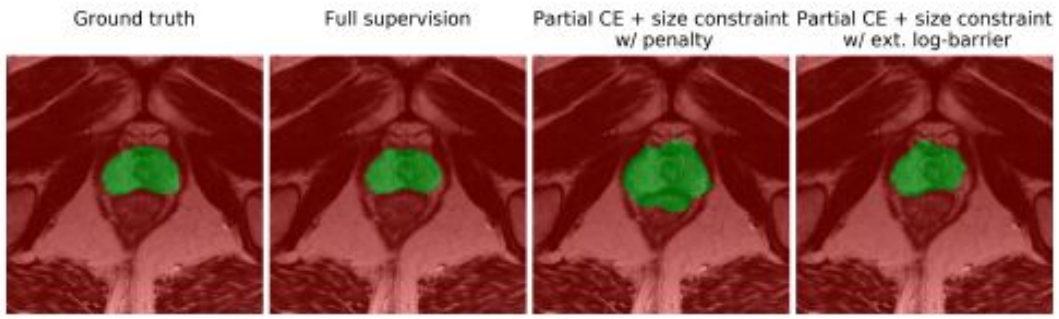
[Kervadec et al., Constrained Deep Networks: Lagrangian Optimization via Log-Barrier Extensions, ArXiv 2019]

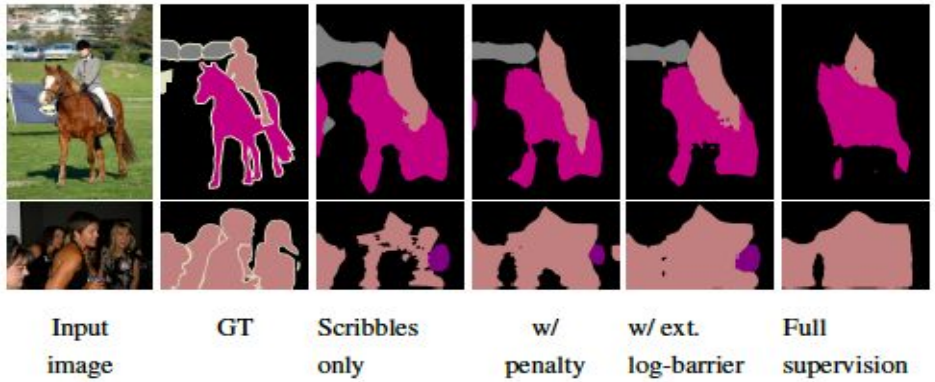
Log-barrier extensions: Approximates Lagrangian optimization but **NO** explicit dual steps



Promising results

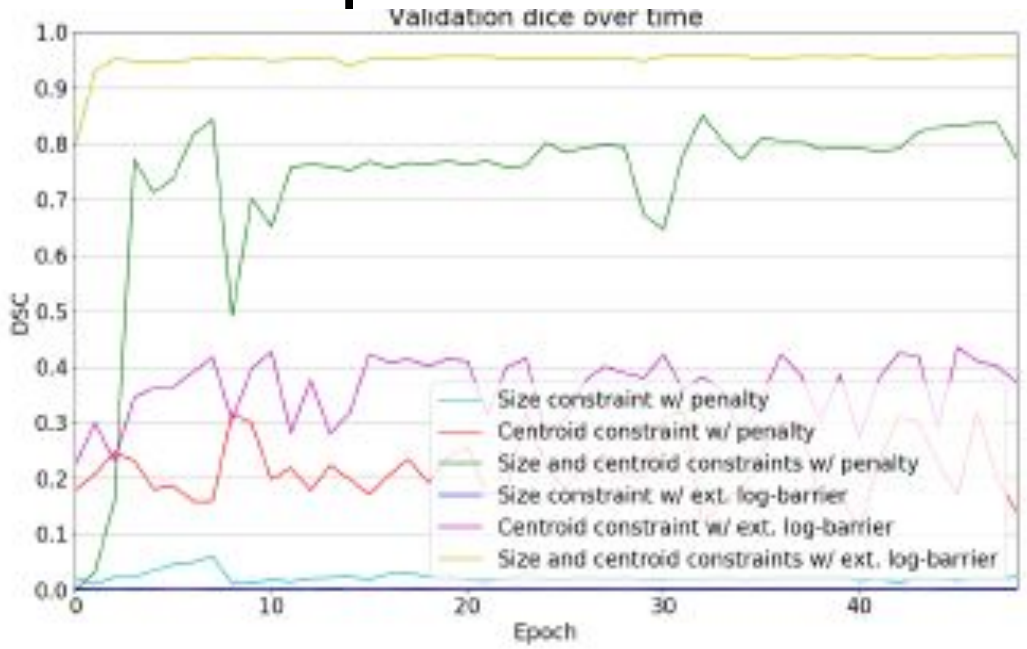
	Dataset	
Method	PROMISE12 (DSC)	VOC2012 (mIoU)
Partial cross-entropy	0.032 (0.015)	48.48 (14.88)
w/ penalty [12]	0.830 (0.057)	52.22 (14.94)
w/ extended log-barrier	**0.852** (0.038)	**53.40** (14.62)
Full supervision	0.891 (0.032)	59.87 (16.94)





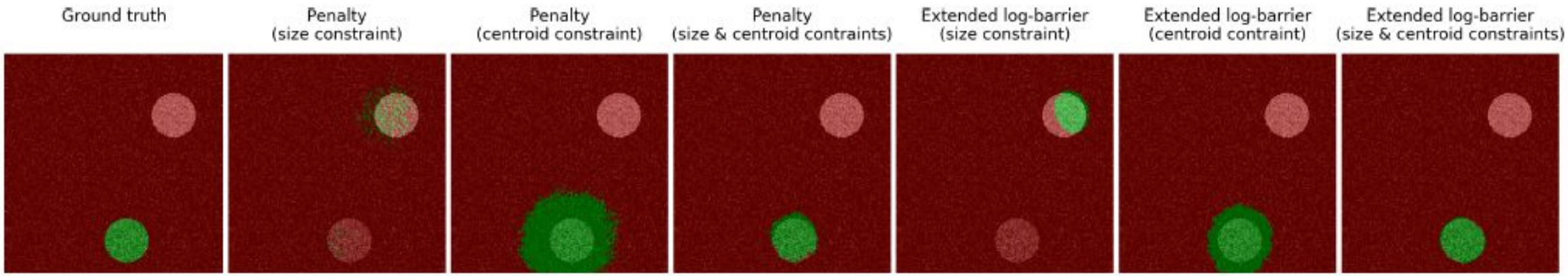
[Kervadec et al., Constrained Deep Networks: Lagrangian Optimization via Log-Barrier Extensions, ArXiv 2019]

Log-barrier extensions: Approximates Lagrangian optimization but **NO** explicit dual steps



Promising results

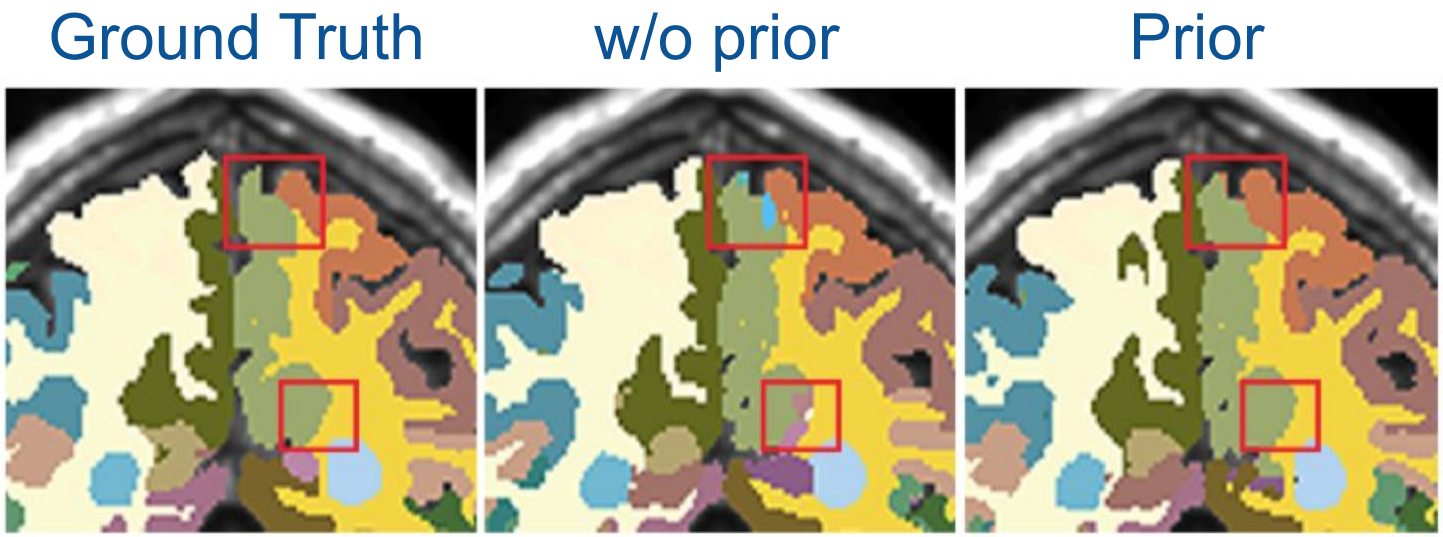
	Constraints		
Method	**Size**	**Centroid**	**Size & Centroid**
Penalty [12]	0.0601	0.3197	**0.8514**
Extended log-barrier	0.0018	0.4347	**0.9574**



[Kervadec et al., Constrained Deep Networks: Lagrangian Optimization via Log-Barrier Extensions, ArXiv 2019]

Other constraints (Semi-supervision)

Connectivity of anatomical structures



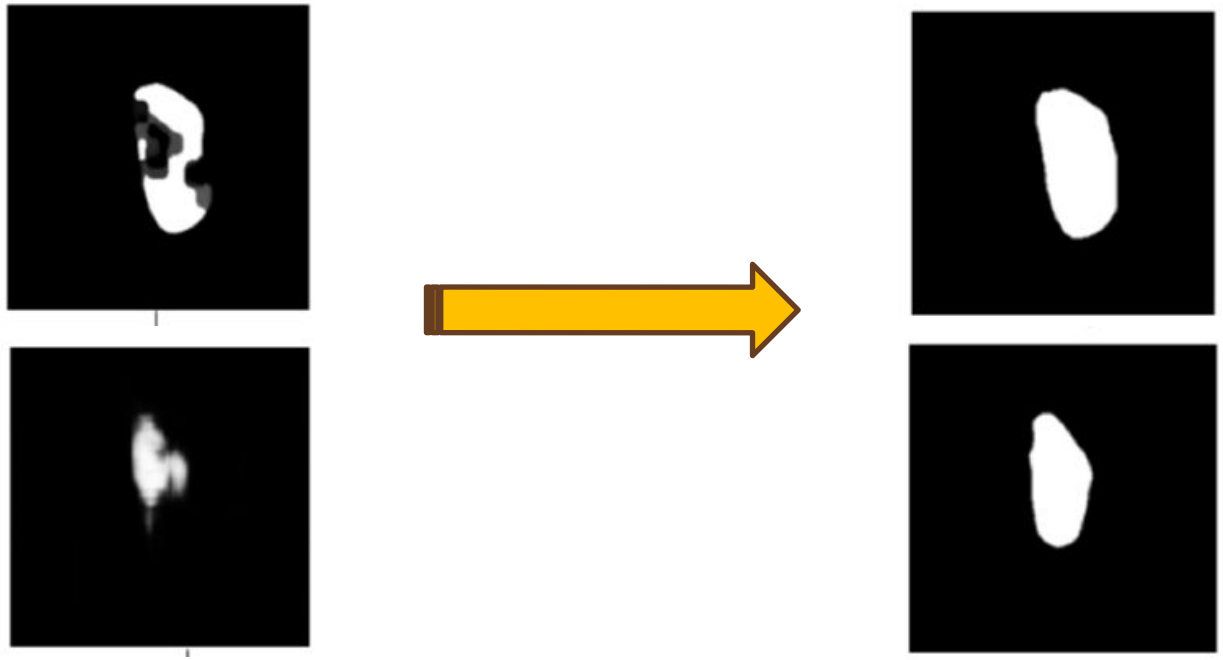
Images from [Ganaye et al. Semi-supervised learning for segmentation under semantic constraint. MICCAI'18]

Other constraints (Full-supervision)

Shape

Bad segmentations

Good segmentations

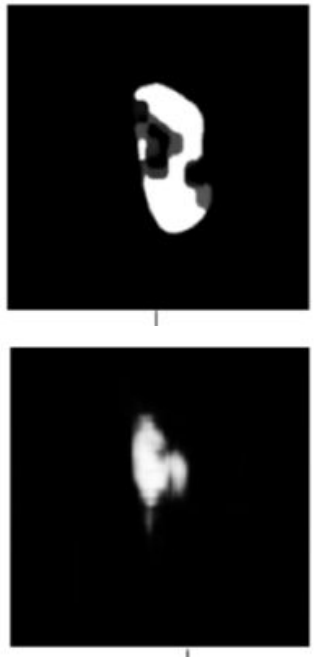


Ravishankar H, et al. Learning and incorporating shape models for semantic segmentation. MICCAI'17

Other constraints (Full-supervision)

Shape

Bad segmentations



Good segmentations

During training

Oktay et al., IEEE TMI'17

Post-processing

Larrazabal et al., MICCAI'19
Painchaud et al, MICCAI'19

Ravishankar H, et al. Learning and incorporating shape models for semantic segmentation. MICCAI'17

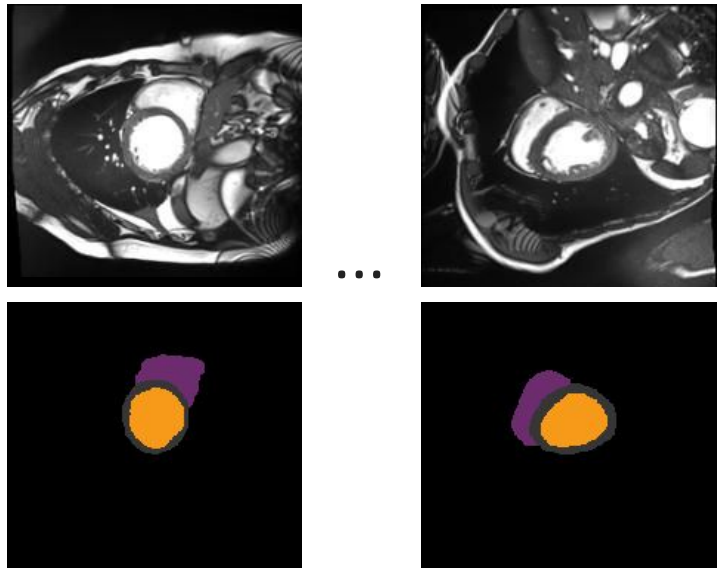
Take-home message

- Imposing constraints helps weakly-supervised segmentation learning by restricting plausible segmentations on unlabeled images
- Few constraints have been explored under low-labeled data regime
- Room for improvement (many opportunities)

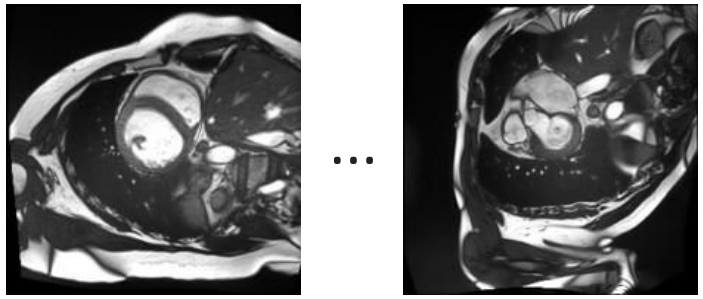
Adversarial learning methods for weakly-supervised segmentation

Learning with unlabeled images

Labeled images (few)

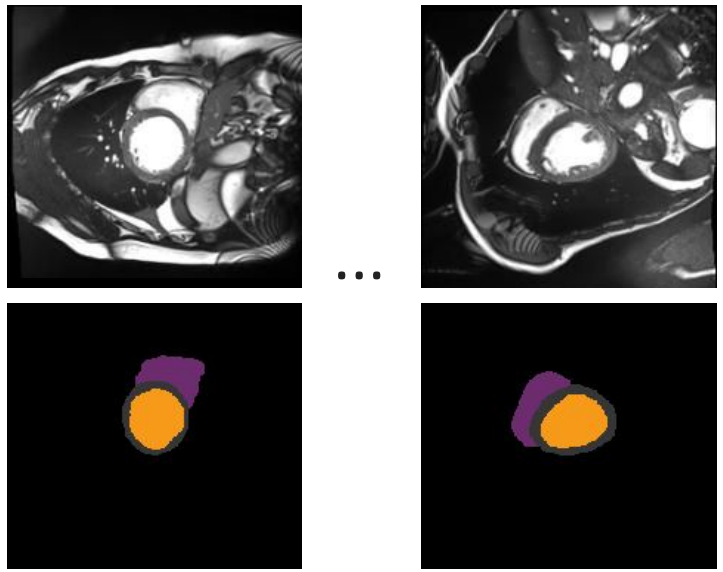


Unlabeled images (many)

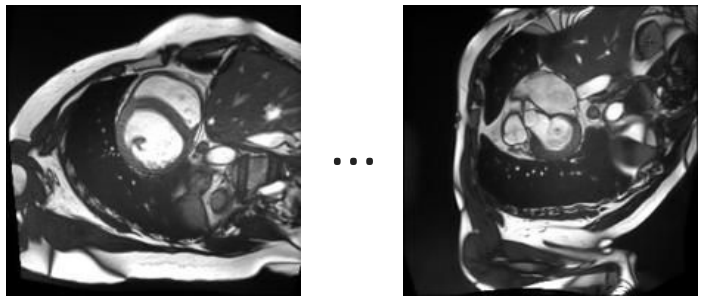


Learning with unlabeled images

Labeled images (few)



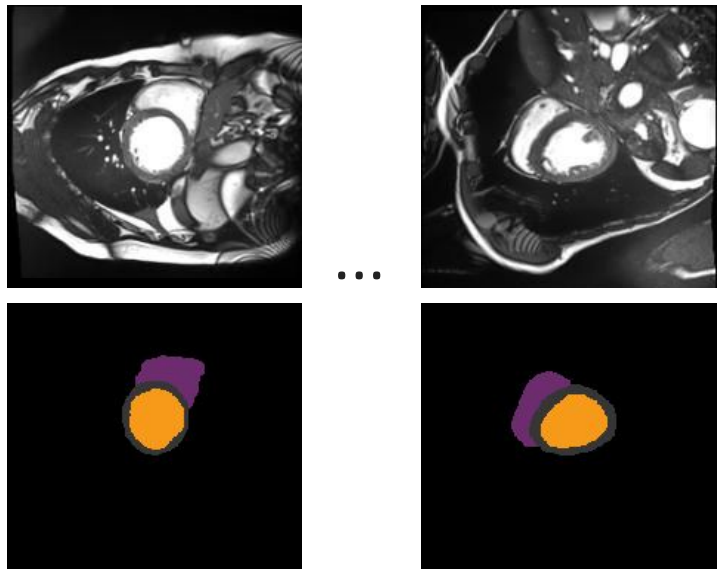
Unlabeled images (many)



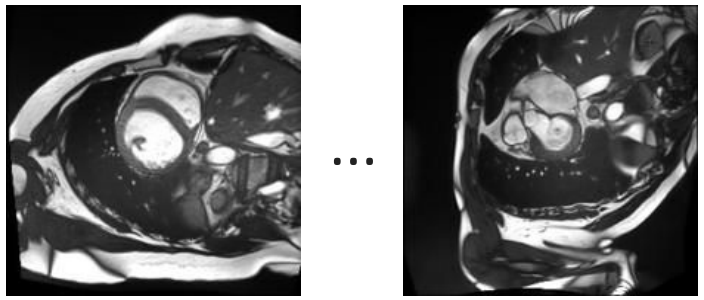
How can use this information for learning segmentation ?

Learning with unlabeled images

Labeled images (few)



Unlabeled images (many)



How can use this information for learning segmentation ?

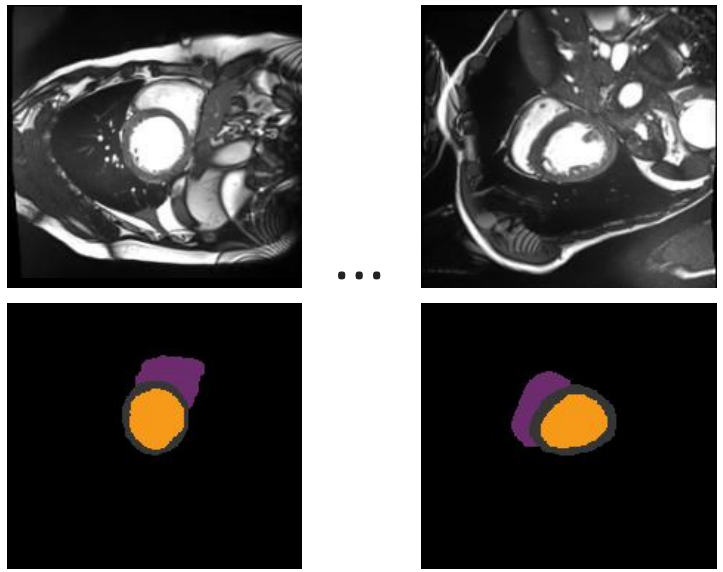
1) Knowledge-based priors:

- Size bounds, shape atlas, boundary smoothness, etc.
- Difficult to adapt to new domains or tasks

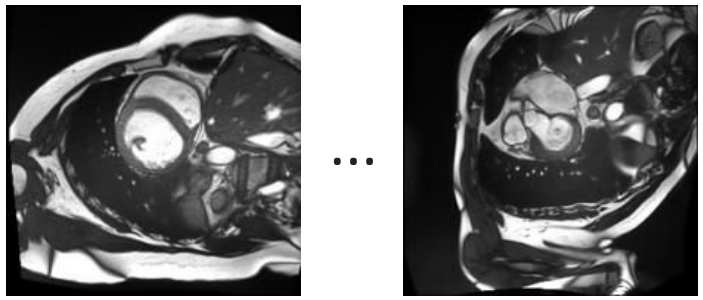
Presented in previous slides

Learning with unlabeled images

Labeled images (few)



Unlabeled images (many)



How can use this information for learning segmentation ?

1) Knowledge-based priors:

- Size bounds, shape atlas, boundary smoothness, etc.
- Difficult to adapt to new domains or tasks

Presented in previous slides

2) Adversarial learning:

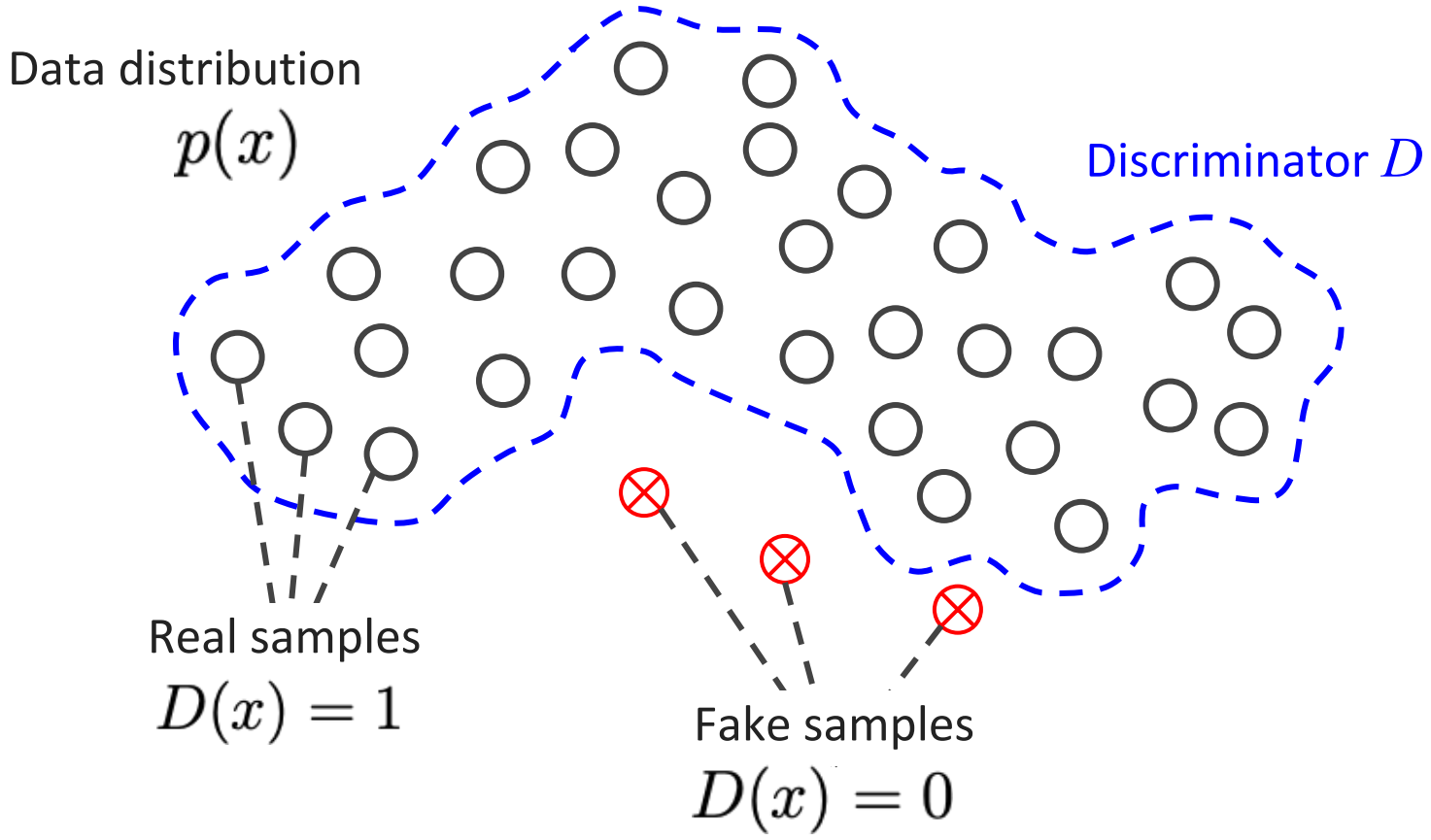
- Learn priors directly from <u>training data</u>
- Easily adapts to new domains or tasks

Discussed in next slides

Adversarial learning

Basic idea:

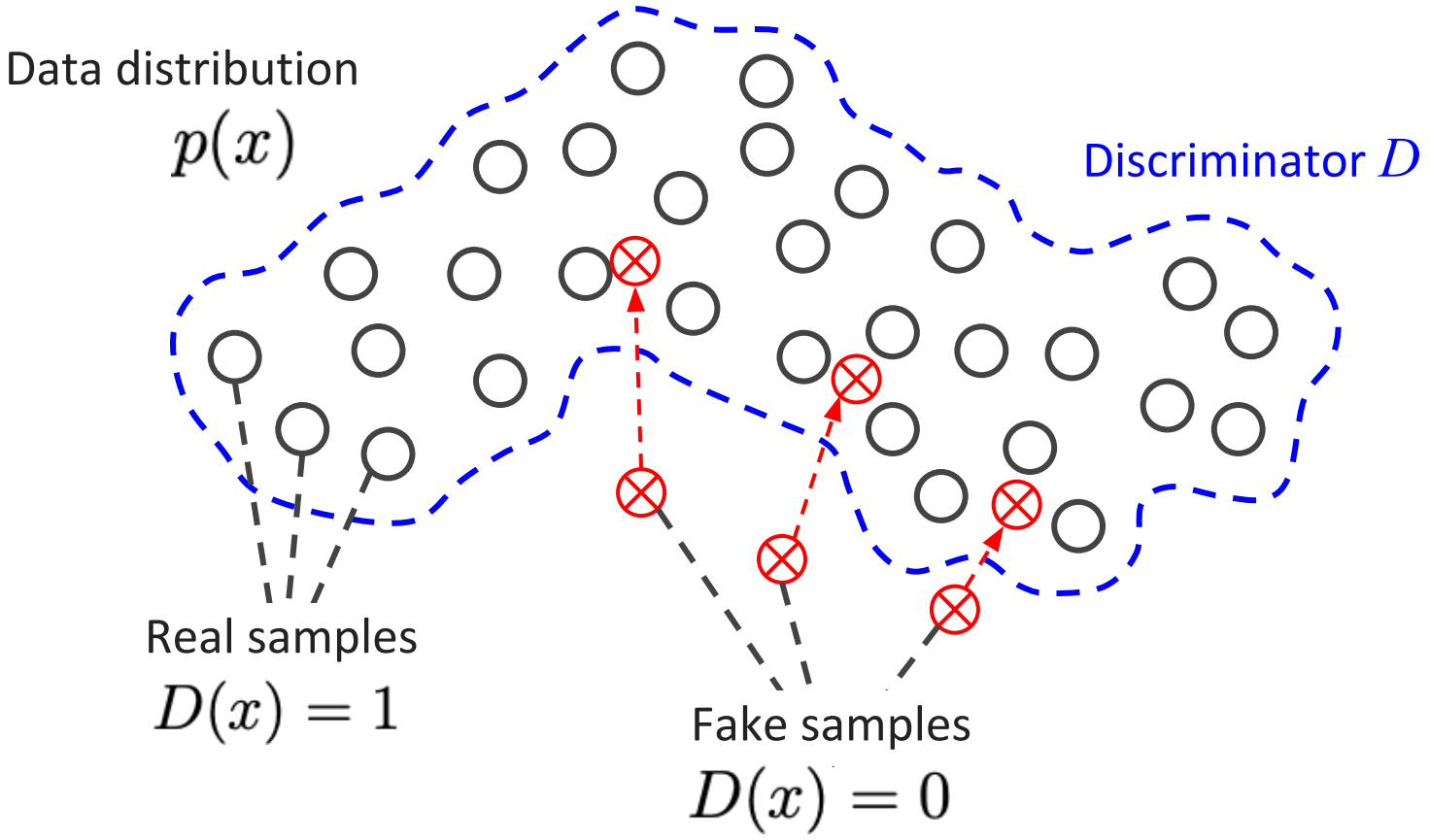
Learn the data distribution using a classifier (the discriminator)



Adversarial learning

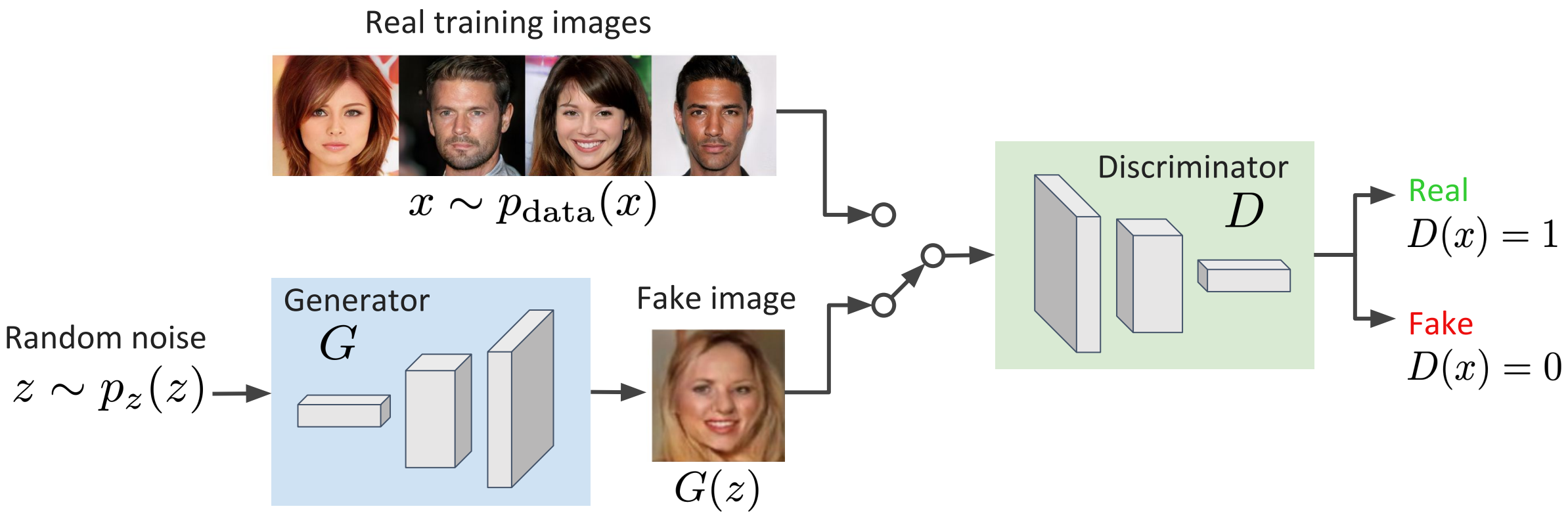
Basic idea:

Learn the data distribution using a classifier (the discriminator)

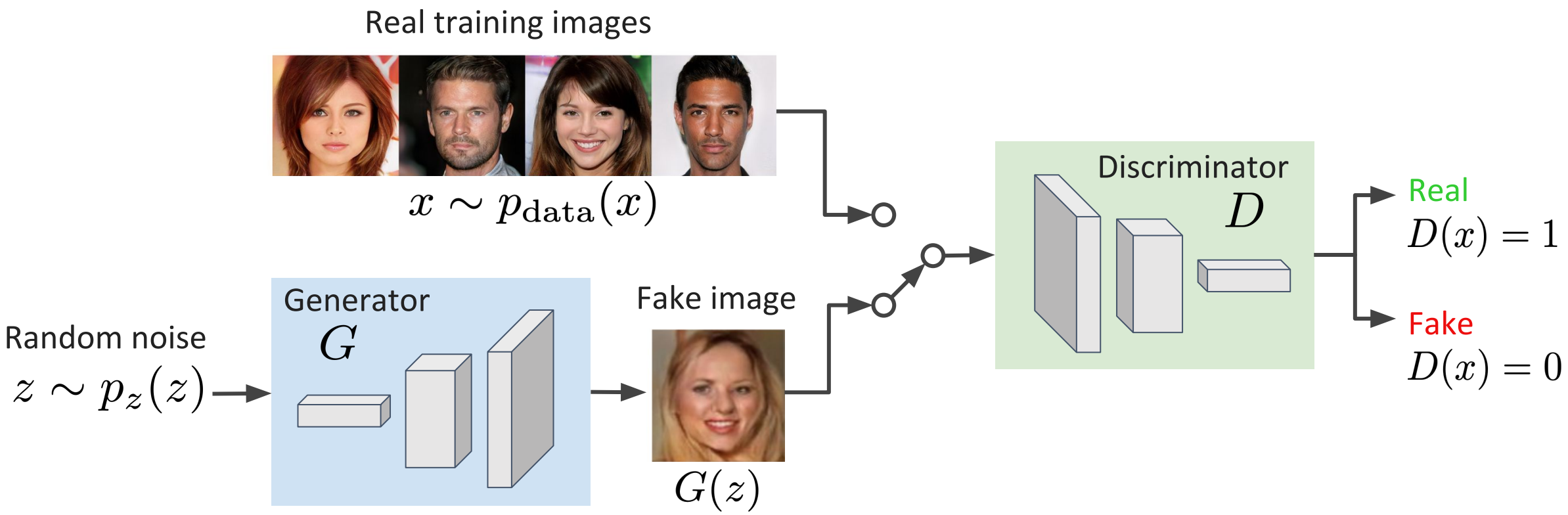


Objective: Generate samples in the distribution of real data

Generative adversarial network (GAN)

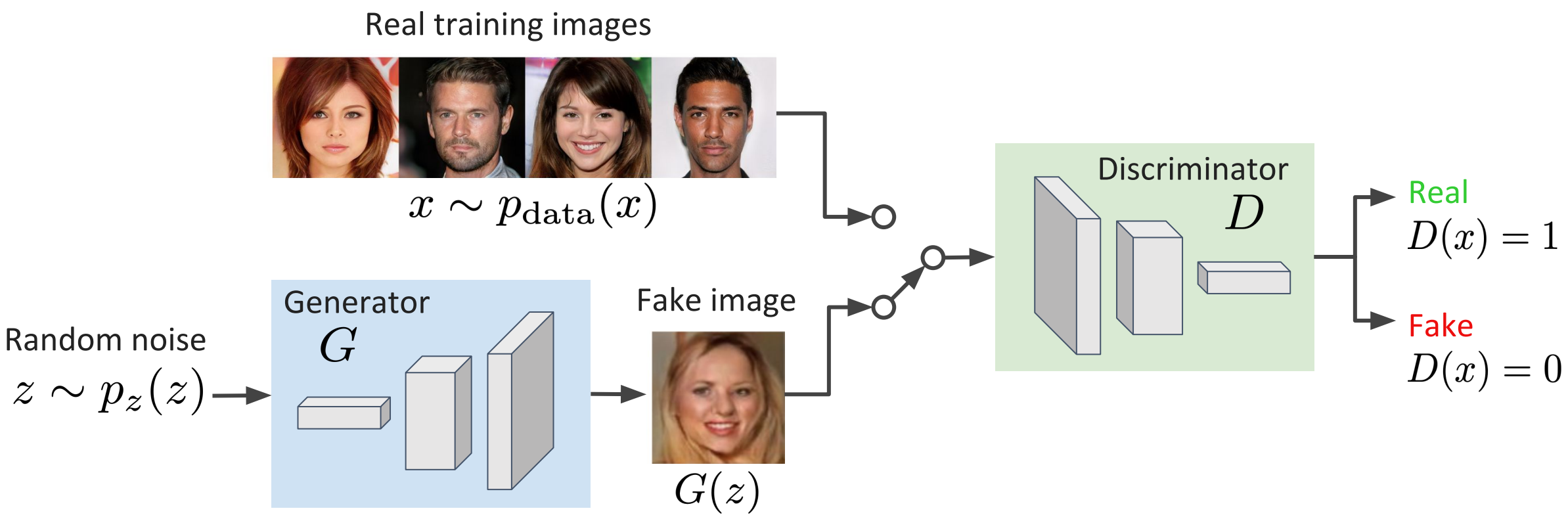


Generative adversarial network (GAN)



How to make sure that generated images look real ?

Generative adversarial network (GAN)



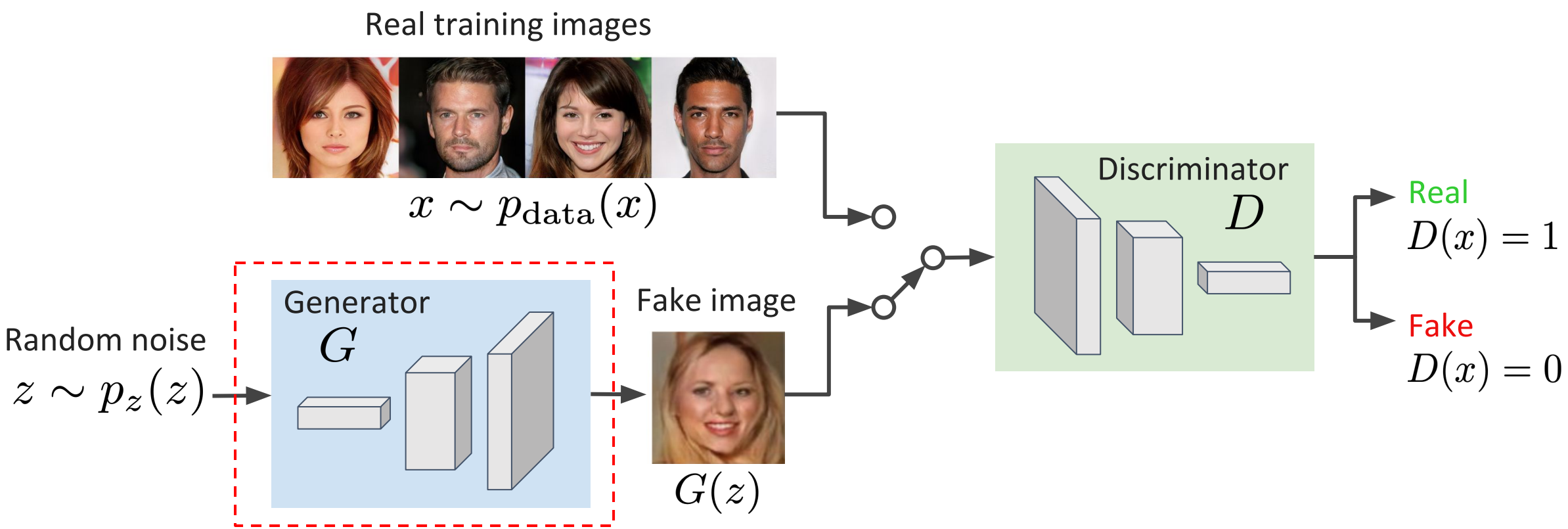
Training the discriminator (cross-entropy):

$$\max_D \; \mathbb{E}_{x \sim p_{\text{data}}(x)}\big[\log D(x)\big] + \mathbb{E}_{z \sim p_z(z)}\big[\log(1 - D(G(z))\big]$$

Output '1' for real images

Output '0' for generated images

Generative adversarial network (GAN)



Training the generator:

$$\min_G \ \mathbb{E}_{z \sim p_z(z)} \left[\log(1 - D(G(z)) \right]$$

Fool the discriminator into predicting ‘1’ for fake images

Generative adversarial network (GAN)

Real training images

$x \sim p_{\text{data}}(x)$

Random noise

$z \sim p_z(z)$

Generator

G

Fake image

$G(z)$

Discriminator

D

Real

$D(x) = 1$

Fake

$D(x) = 0$

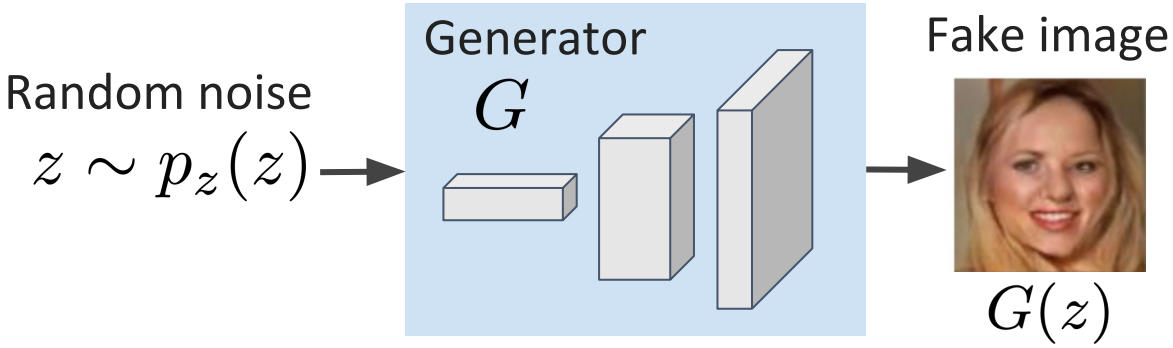
Training the whole architecture:

$$\min_G \max_D \; \mathbb{E}_{x \sim p_{\text{data}}(x)} \big[\log D(x) \big] \; + \; \mathbb{E}_{z \sim p_z(z)} \big[\log(1 - D(G(z)) \big]$$

Corresponds to a minimax problem (*more on this later...*)

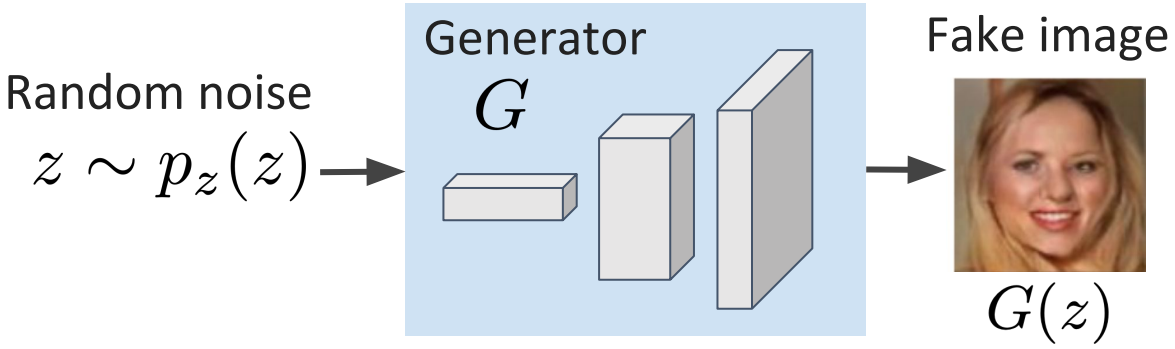
GANs for segmentation

GAN for image generation:

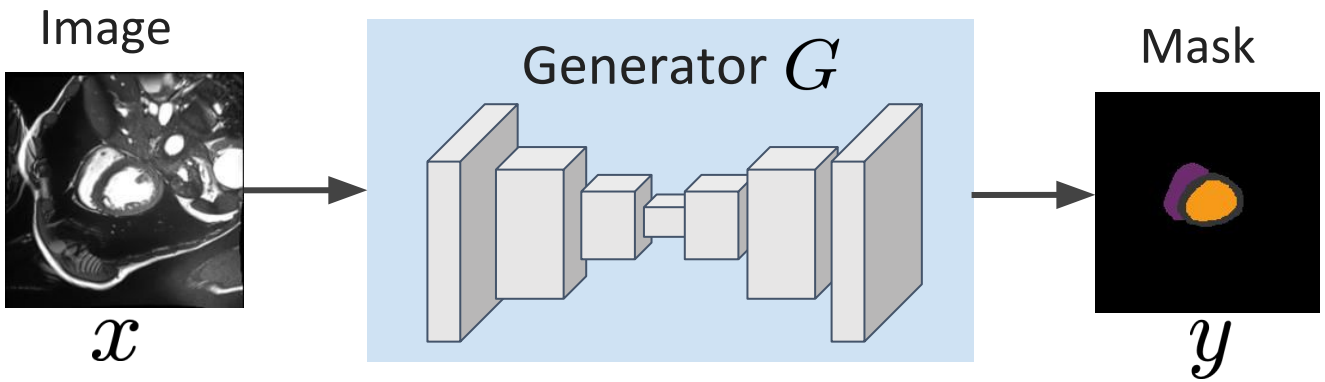


GANs for segmentation

GAN for image generation:

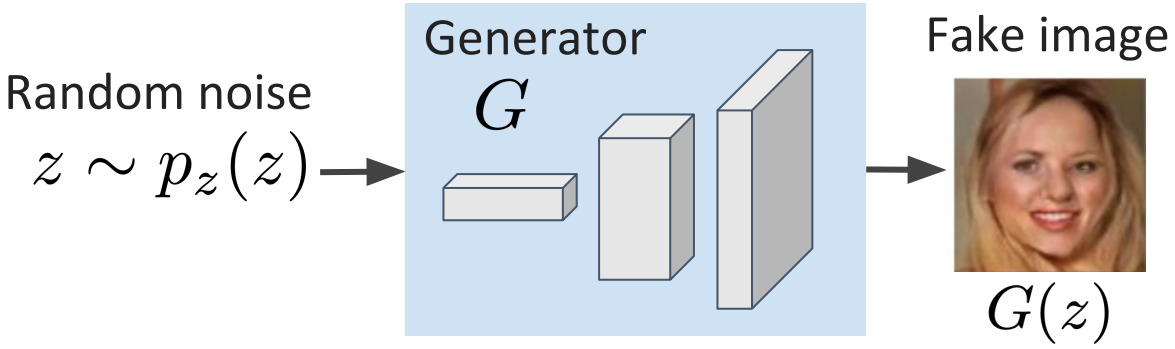


GAN for image segmentation:

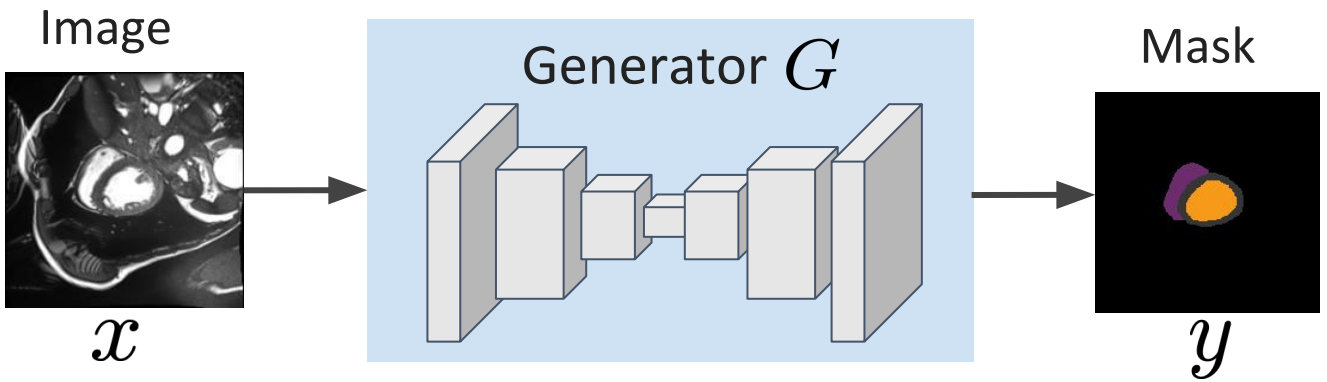


GANs for segmentation

GAN for image generation:



GAN for image segmentation:

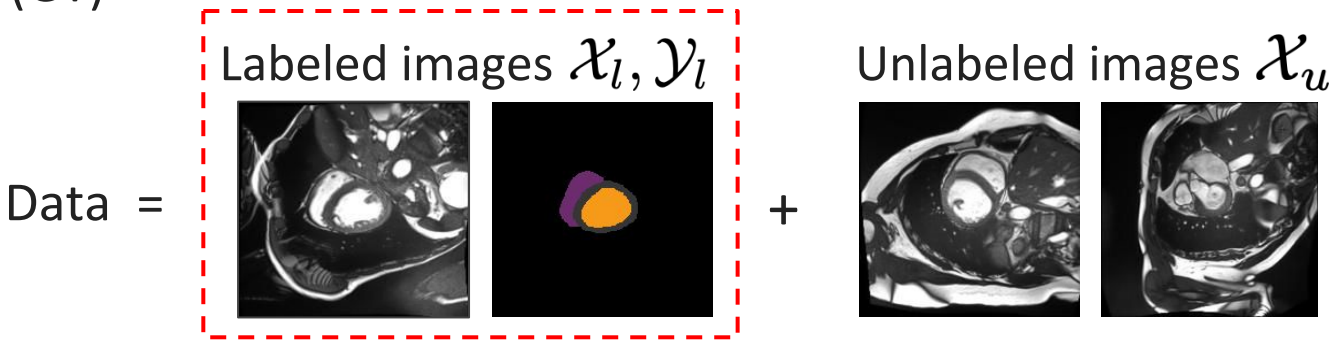


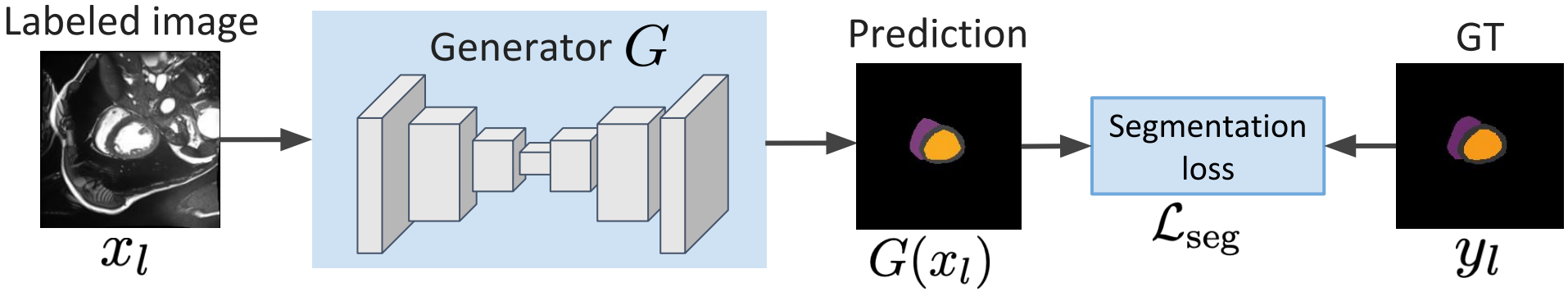
We are now modeling the distribution of *segmentation masks*

The generator is a segmentation network (encoder-decoder)

Adversarial semi-supervised segmentation

Basic idea: Learn to generate segmentation masks which can't be differentiated from ground-truth (GT)

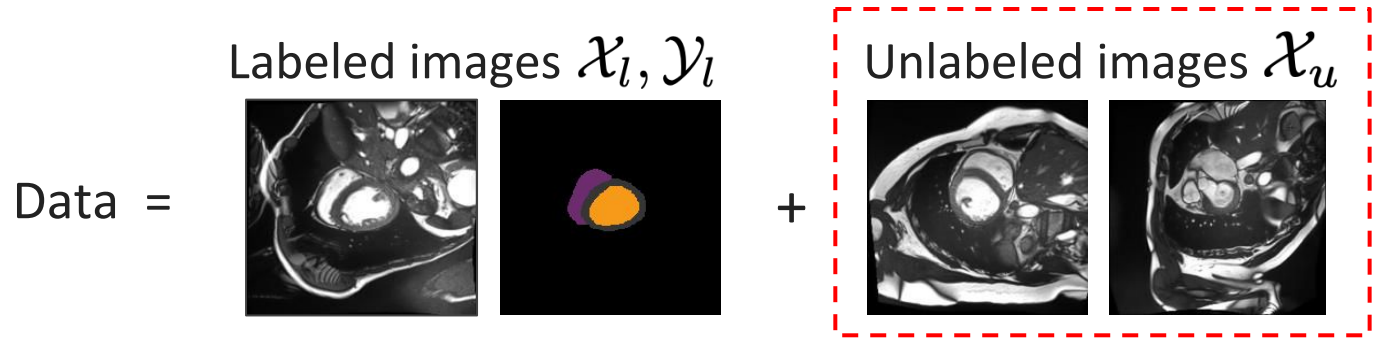


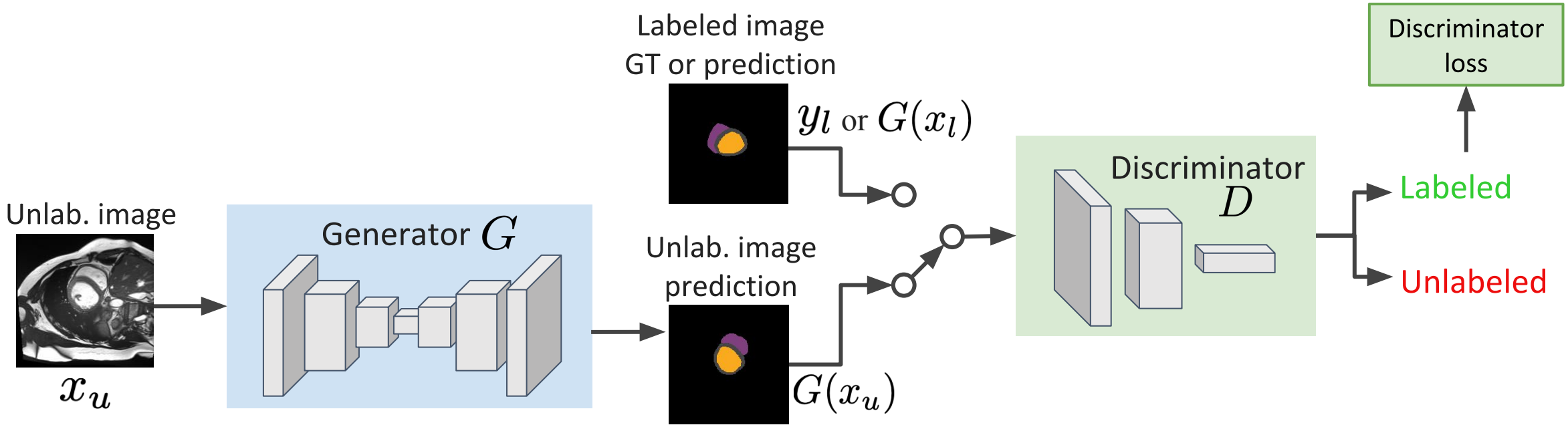


$$\mathcal{L}_{\text{sup}}(G) = \mathbb{E}_{(x_l,y_l)\sim\mathcal{X}_l,\mathcal{Y}_l}\left[\mathcal{L}_{\text{seg}}(G(x_l), y_l)\right]$$

Adversarial semi-supervised segmentation

Basic idea: Learn to generate segmentation masks which can't be differentiated from ground-truth (GT)



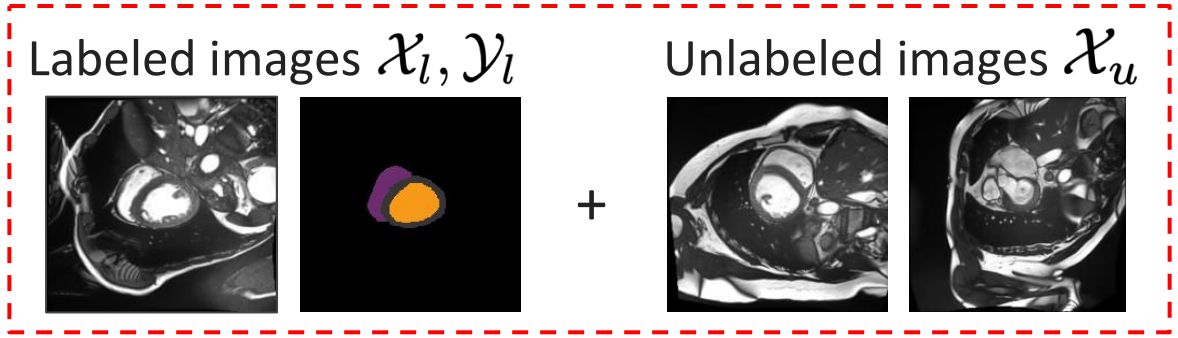


$$\mathcal{L}_{\text{adv}}(G, D) = \mathbb{E}_{x_u \sim \mathcal{X}_u}\left[\mathcal{L}_{\text{dis}}\big(D(G(x_u)), 0\big)\right] + \mathbb{E}_{x_l \sim \mathcal{X}_l}\left[\mathcal{L}_{\text{dis}}\big(D(G(x_l)), 1\big)\right]$$

Adversarial semi-supervised segmentation

Basic idea: Learn to generate segmentation masks which can't be differentiated from ground-truth (GT)

Data =



Both labeled and unlabeled:

Controls the trade-off between the two losses

$$\min_G \max_D \mathcal{L}(G,D) = \underbrace{\frac{1}{|\mathcal{X}_l|}\sum_{l=1}^{|\mathcal{X}_l|}\mathcal{L}_{\text{seg}}(G(x_l), y_l)}_{\text{Supervised loss}} - \underbrace{\frac{\lambda}{|\mathcal{X}_l|+|\mathcal{X}_u|}\left(\sum_{l=1}^{|\mathcal{X}_l|}\mathcal{L}_{\text{dis}}\big(D(G(x_l)),1\big) + \sum_{u=1}^{|\mathcal{X}_u|}\mathcal{L}_{\text{dis}}\big(D(G(x_u)),0\big)\right)}_{\text{Adversarial loss}}$$

Adversarial semi-supervised segmentation

Adversarial network for semi-supervised segmentation of histological images

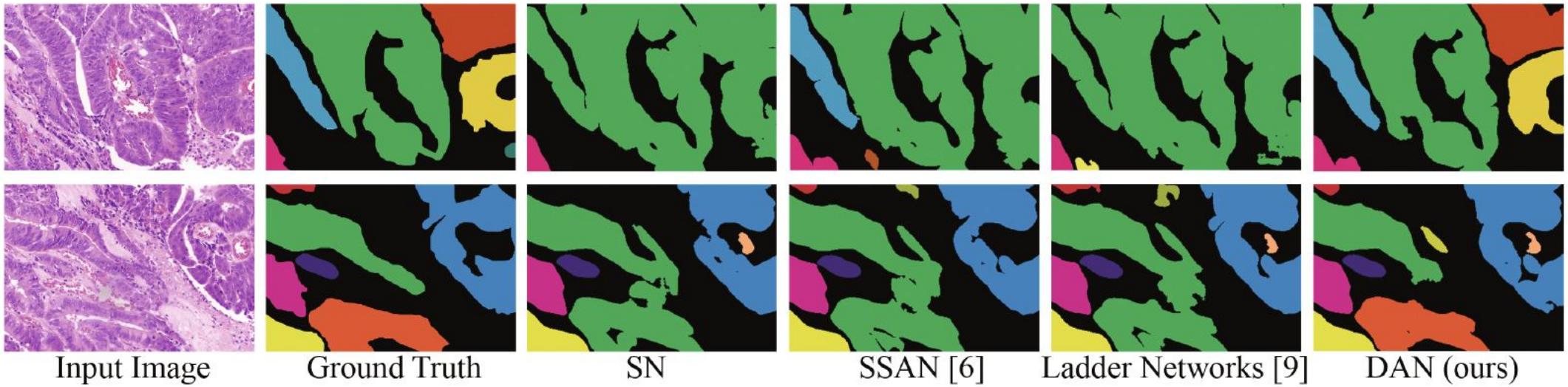
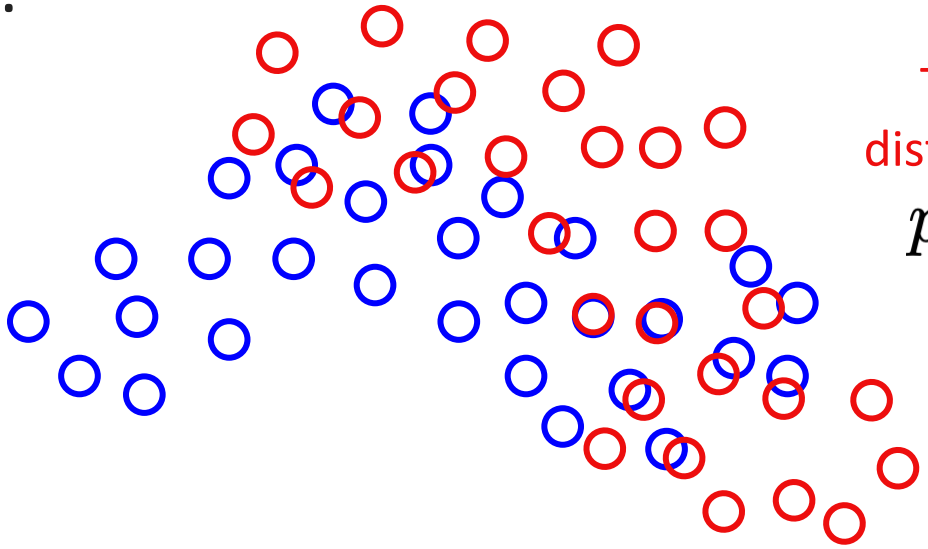


Image from Zhang, Y., et al. "Deep adversarial networks for biomedical image segmentation utilizing unannotated images." *Int. Conf. on Medical Image Computing and Computer-Assisted Intervention*. 2017.

Domain adaptation

Before adaptation:

Source distribution $p_S(x)$

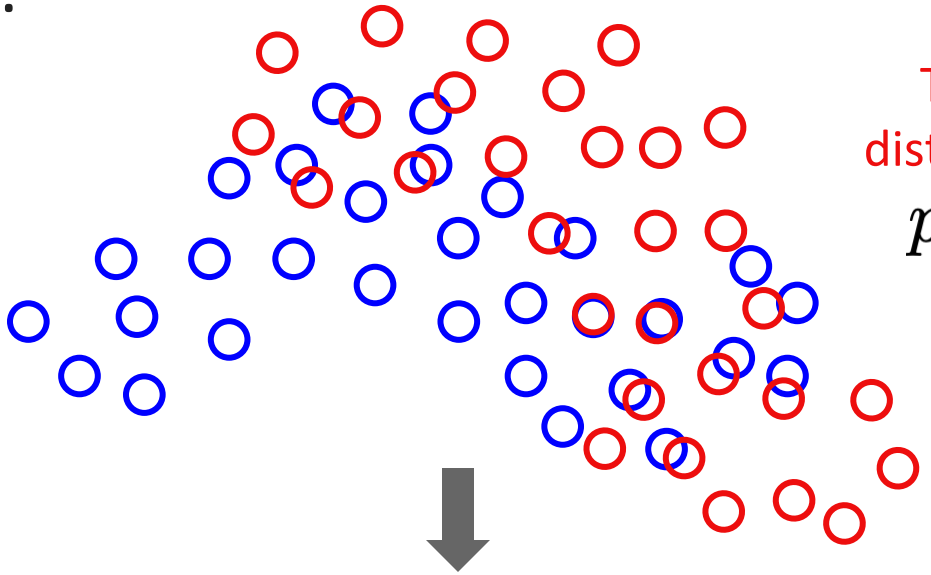


Target distribution $p_T(x)$

Domain adaptation

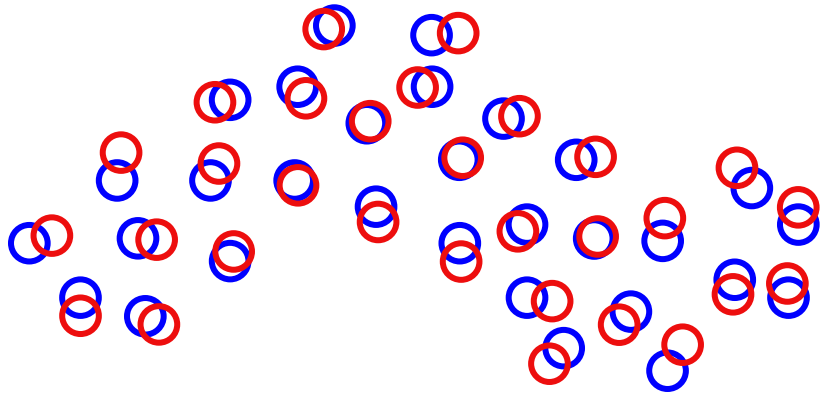
Before adaptation:

Source distribution $p_S(x)$



Target distribution $p_T(x)$

After adaptation:

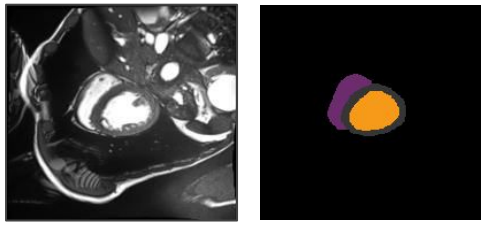


$$p_S(x) = p_T(x)$$

Adversarial domain adaptation

Data =

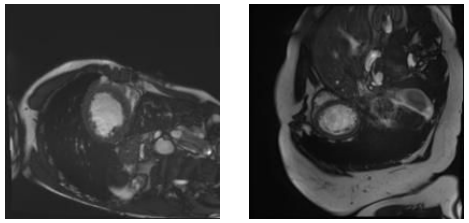
Labeled images $\mathcal{X}_S, \mathcal{Y}_S$



Source domain

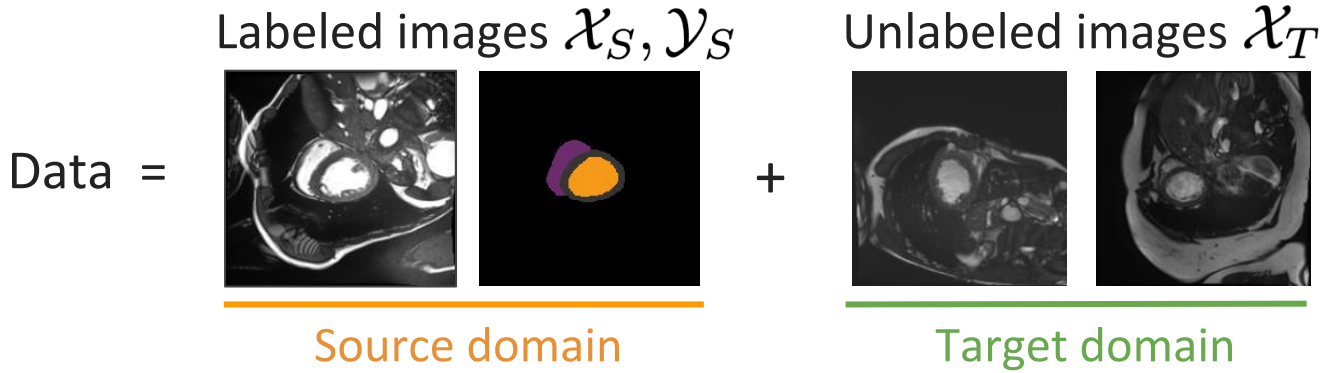
+

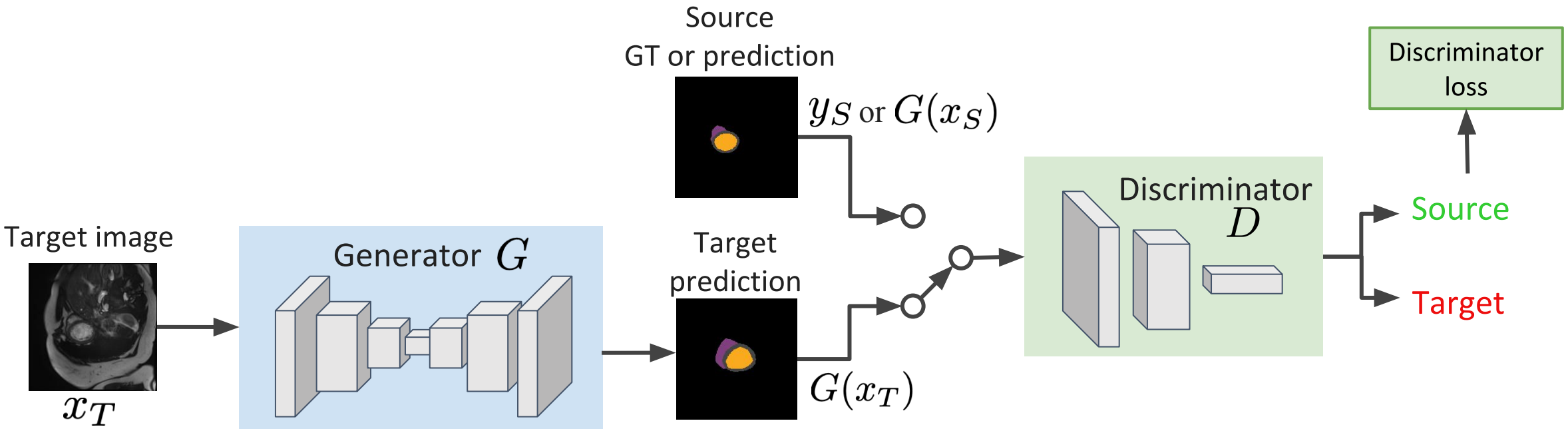
Unlabeled images $\mathcal{X}_T$



Target domain

Adversarial domain adaptation

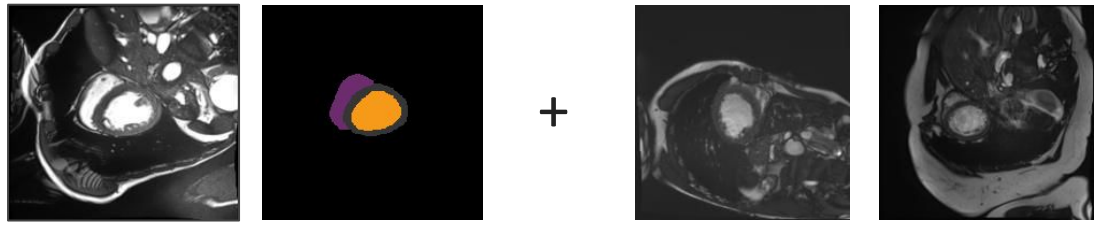




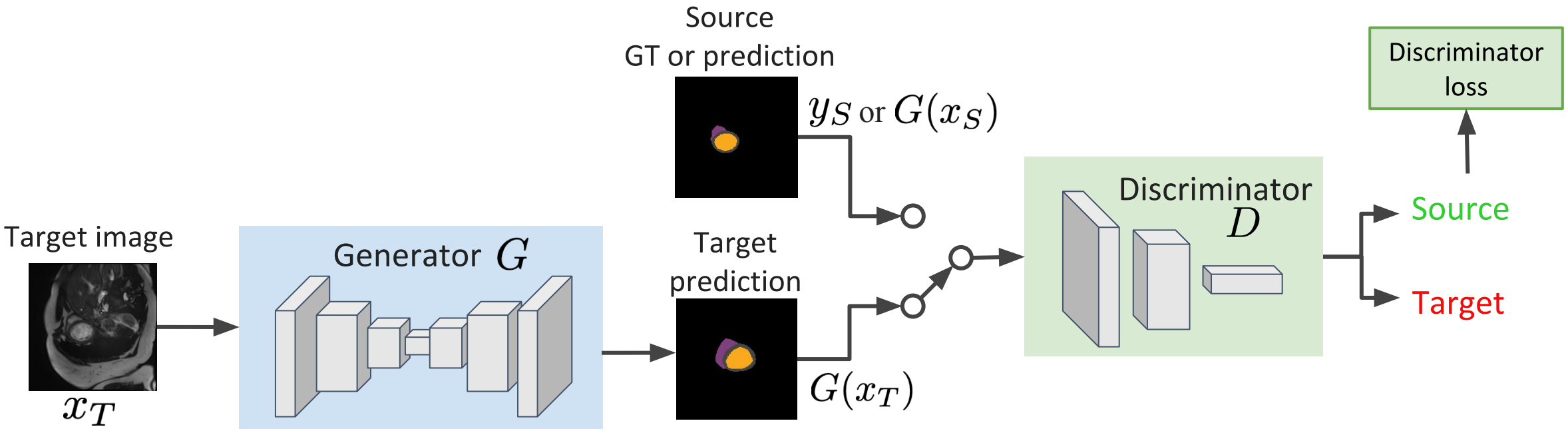
Tsai, Y.-H., et al. "Learning to adapt structured output space for semantic segmentation." IEEE Conf. on Computer Vision and Pattern Recognition. 2018.

Adversarial domain adaptation

Data = Labeled images x_S, y_S + Unlabeled images x_T



Source domain | Target domain

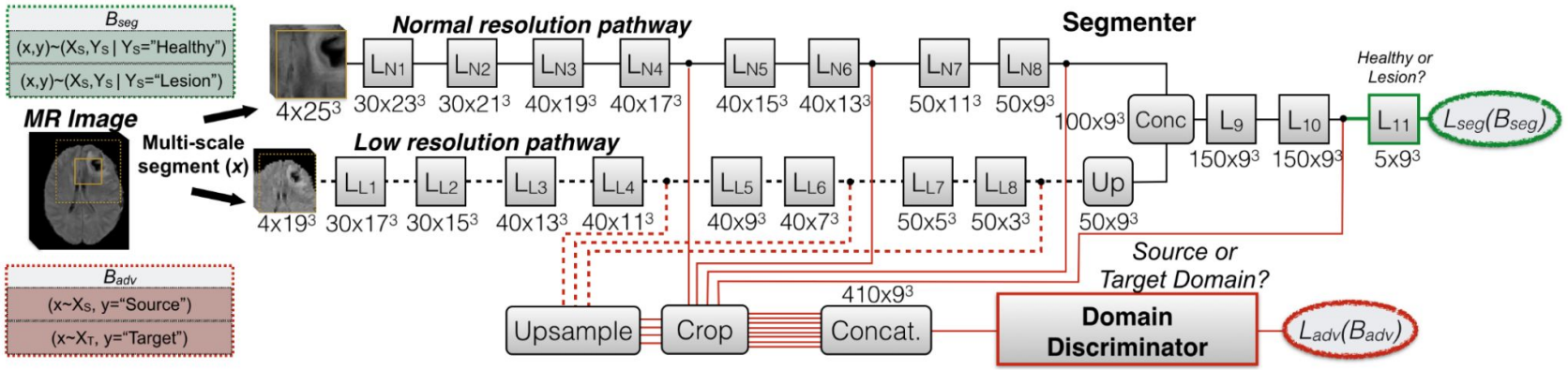


Like semi-supervised segmentation except target images are from a different domain

Tsai, Y.-H., et al. "Learning to adapt structured output space for semantic segmentation." IEEE Conf. on Computer Vision and Pattern Recognition. 2018.

Adversarial domain adaptation

Adversarial domain adaptation for brain lesion segmentation



Source domain (Database 1):

- GE, FLAIR, T2, MPRAGE, PD

Target domain (Database 2):

- SWI, FLAIR, T2, MPRAGE, PD

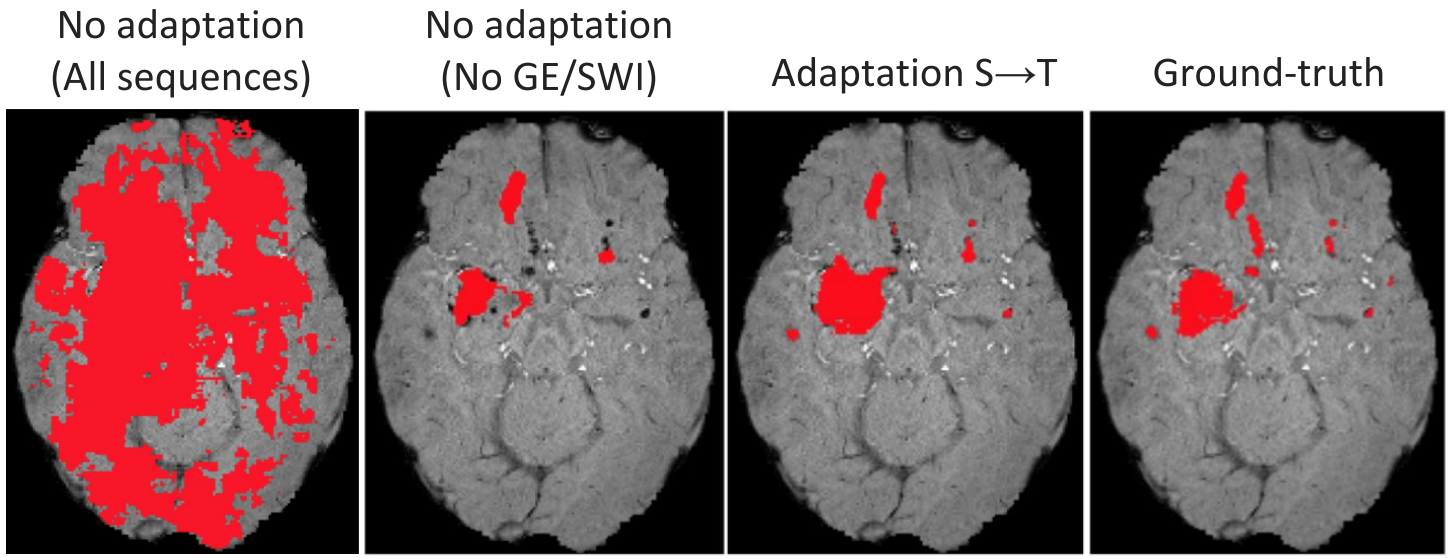
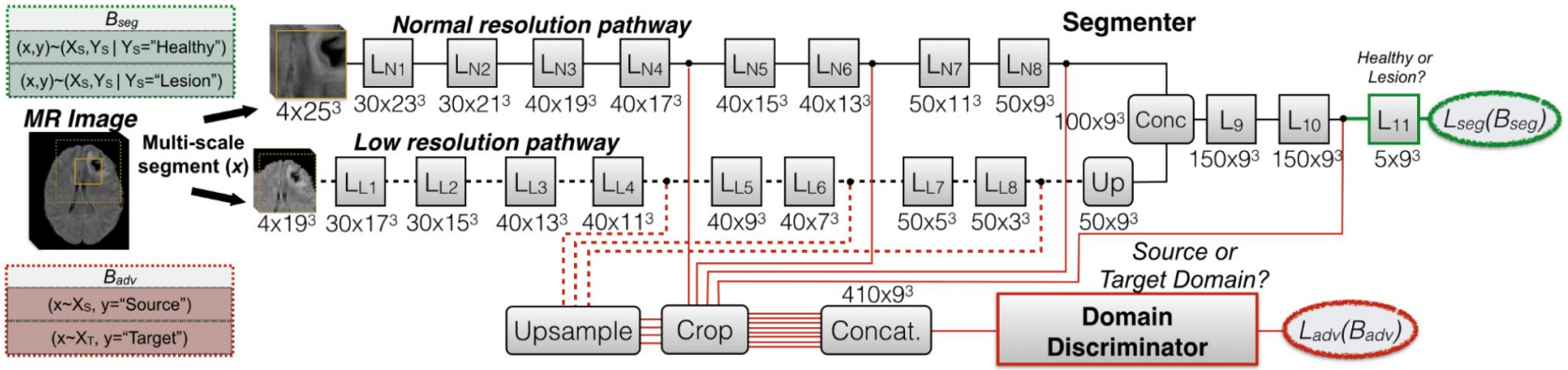


Image from Kamnitsas, K., et al. "Unsupervised domain adaptation in brain lesion segmentation with adversarial networks." *Int. Conf. on Information Processing in Medical Imaging*, 2017.

Adversarial domain adaptation

Adversarial domain adaptation for brain lesion segmentation



Adaptation done on *multi-scale feature representation*

Source domain (Database 1):

- GE, FLAIR, T2, MPRAGE, PD

Target domain (Database 2):

- SWI, FLAIR, T2, MPRAGE, PD

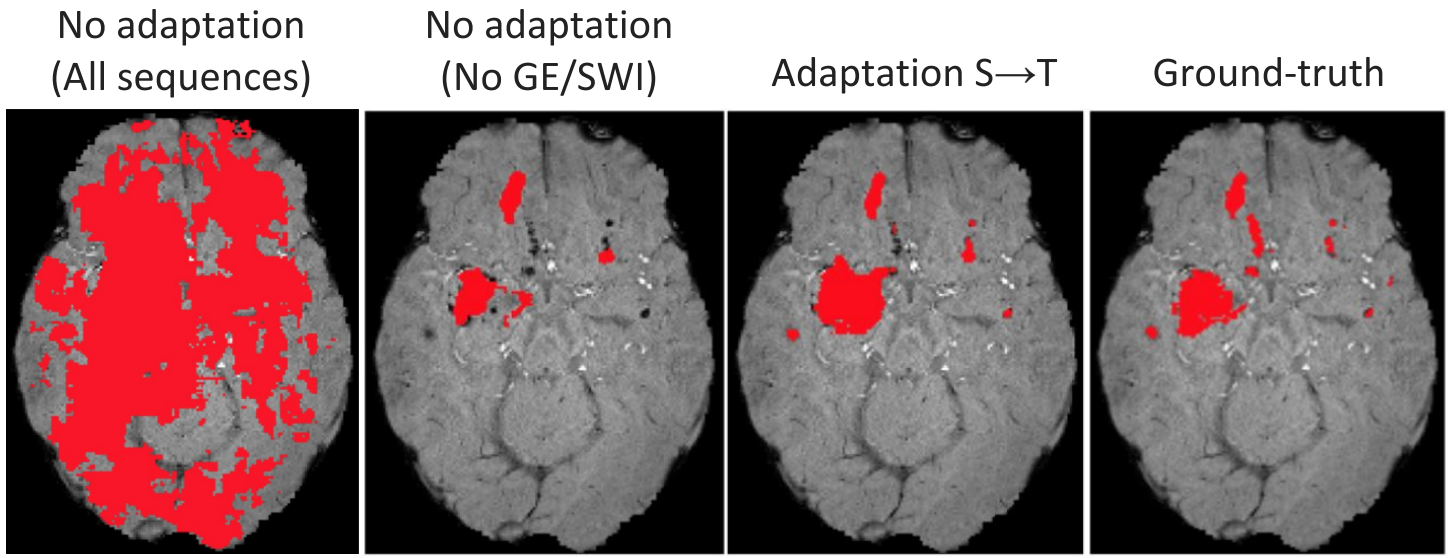


Image from Kamnitsas, K., et al. "Unsupervised domain adaptation in brain lesion segmentation with adversarial networks." *Int. Conf. on Information Processing in Medical Imaging*, 2017.

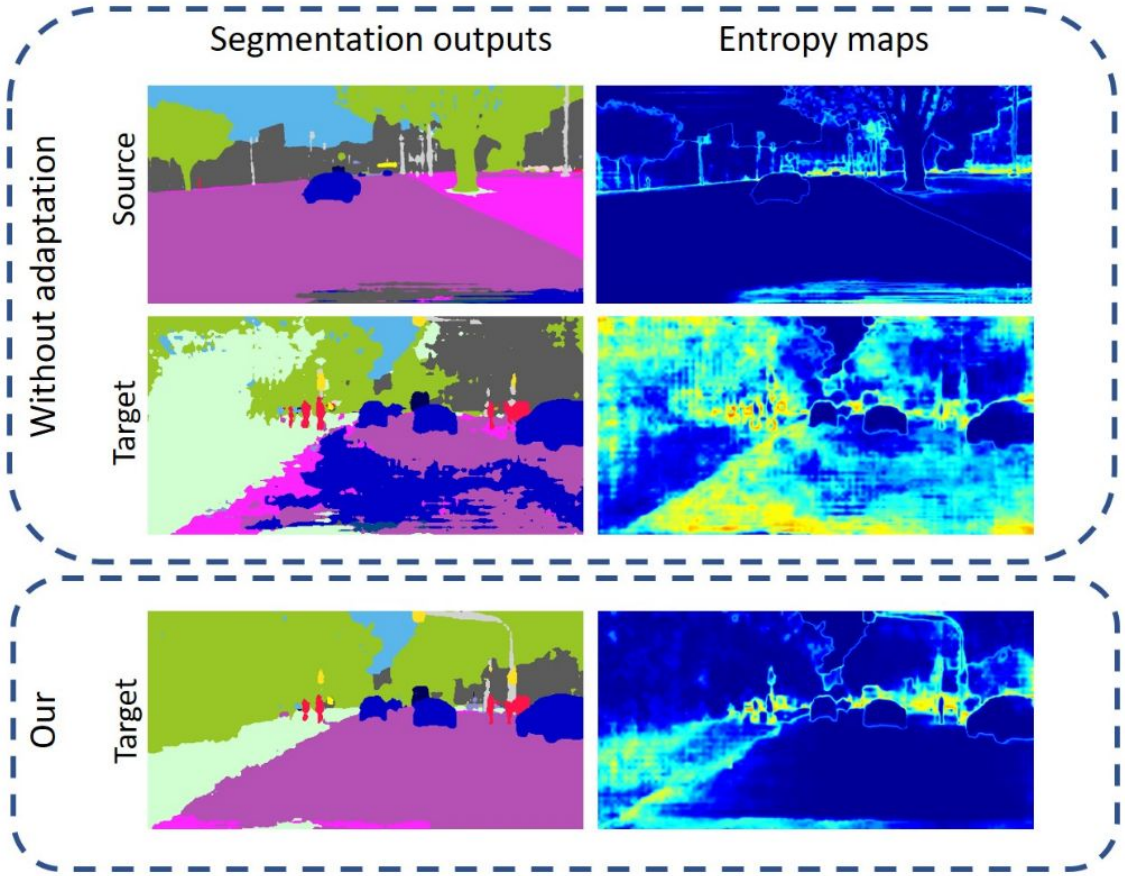
Adversarial domain adaptation

Adaptation on feature representation *versus* softmax output. What else ?

Vu, T.-H., et al. "Advent: Adversarial entropy minimization for domain adaptation in semantic segmentation." IEEE Conf. on Computer Vision and Pattern Recognition (CVPR). 2019.

Adversarial domain adaptation

Adaptation on feature representation *versus* softmax output. What else ?



Adversarial entropy minimization

- The discriminator must differentiate between source and target examples using the entropy spatial maps
- Forces the segmentation model to be consistent in its confidence across different semantic regions

Vu, T.-H., et al. "Advent: Adversarial entropy minimization for domain adaptation in semantic segmentation." IEEE Conf. on Computer Vision and Pattern Recognition (CVPR). 2019.

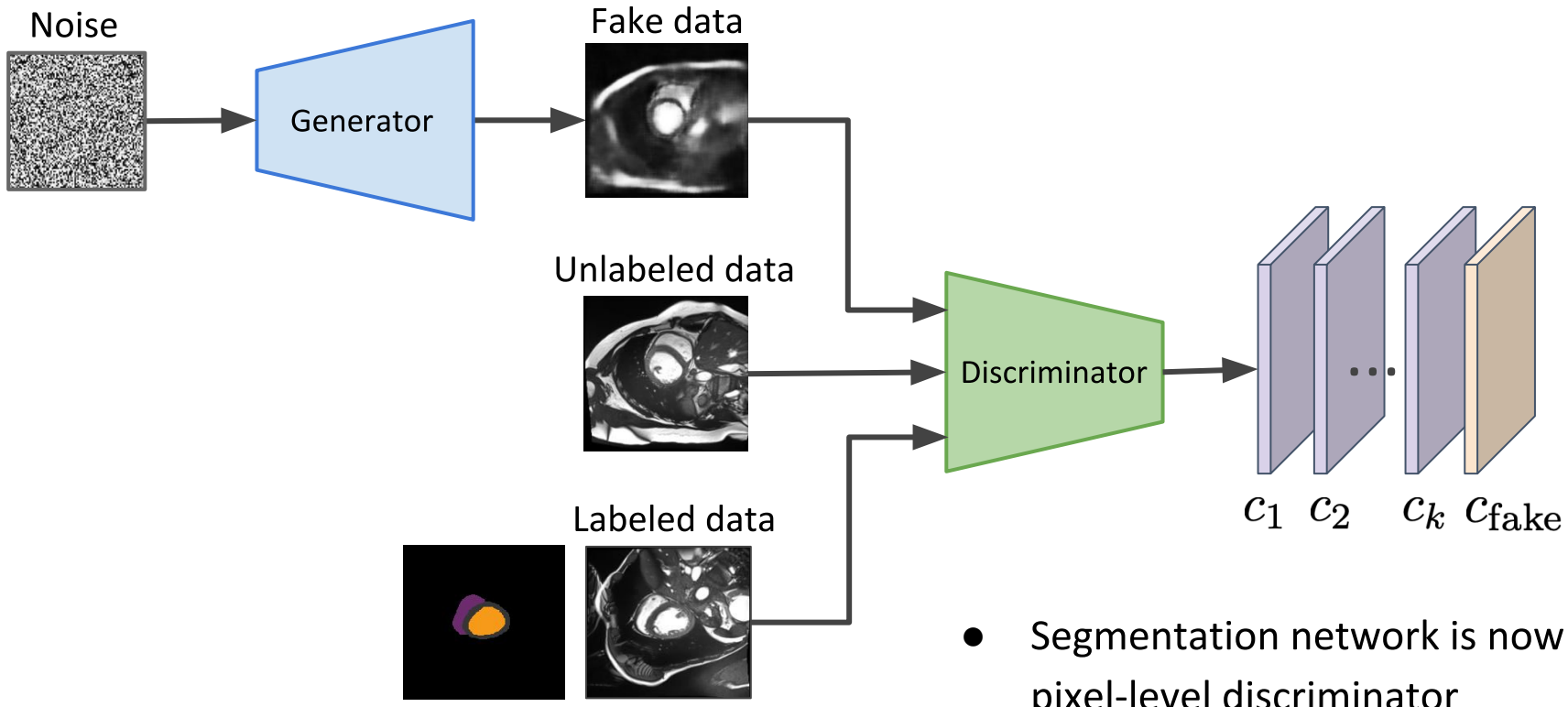
Semi-supervised segmentation with GANs

Can we use GAN-generated images to boost learning in a semi-supervised setting ?

Adapted from: Souly, N. et al.. "Semi supervised semantic segmentation using generative adversarial network." *IEEE Int. Conf. on Computer Vision*. 2017.

Semi-supervised segmentation with GANs

Can we use GAN-generated images to boost learning in a semi-supervised setting ?

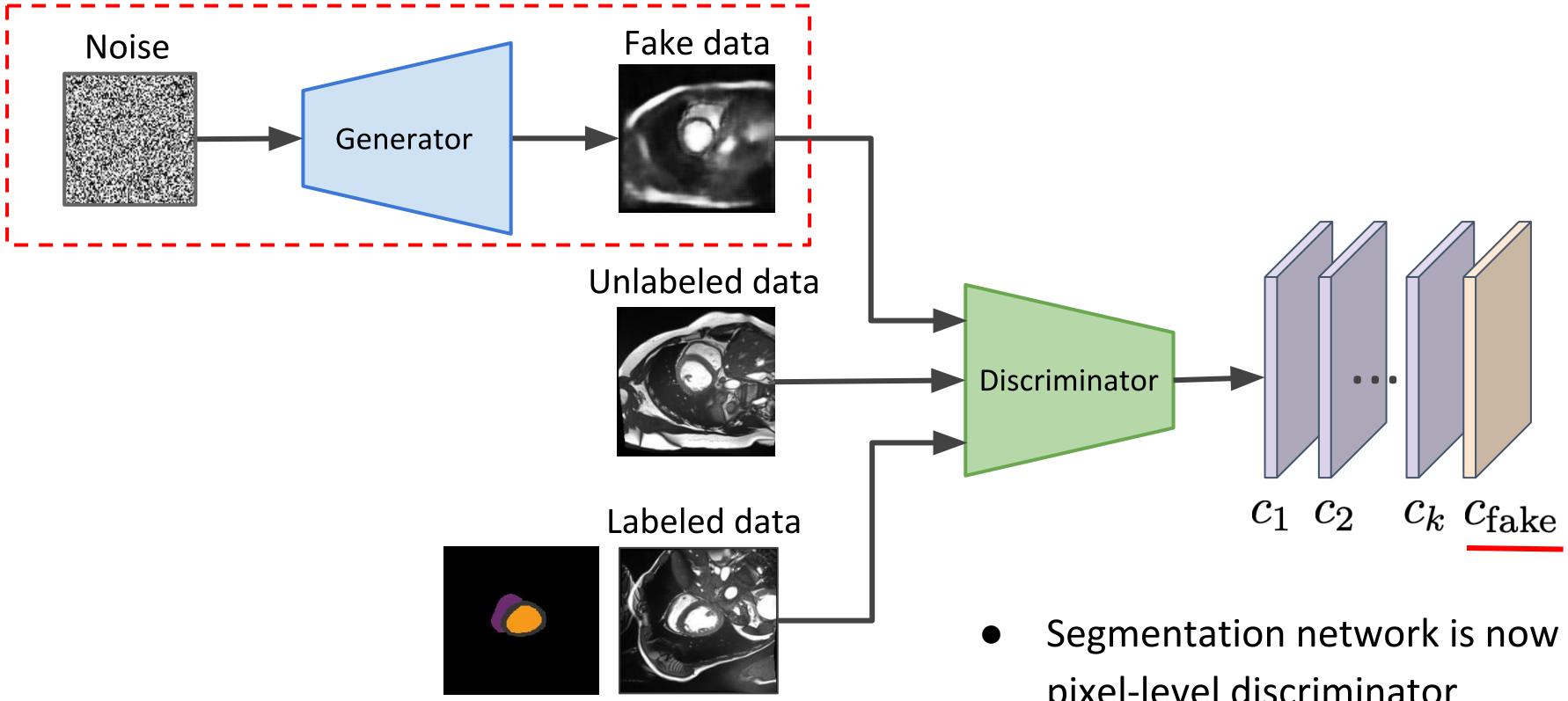


- Segmentation network is now considered as a pixel-level discriminator

Adapted from: Souly, N. et al.. "Semi supervised semantic segmentation using generative adversarial network." *IEEE Int. Conf. on Computer Vision*. 2017.

Semi-supervised segmentation with GANs

Can we use GAN-generated images to boost learning in a semi-supervised setting ?

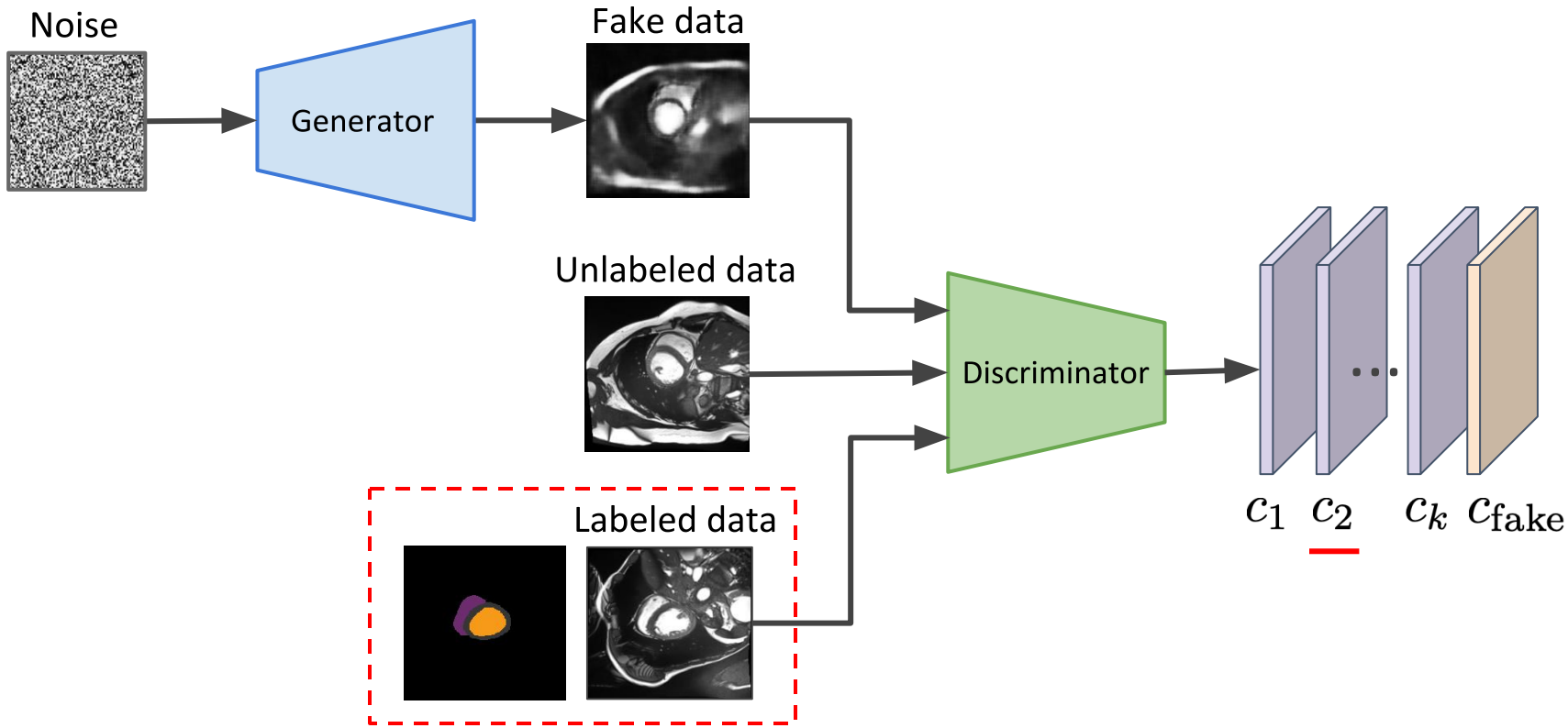


- Segmentation network is now considered as a pixel-level discriminator
- For each pixel, predicts the class label or an extra *fake* label

Adapted from: Souly, N. et al.. "Semi supervised semantic segmentation using generative adversarial network." *IEEE Int. Conf. on Computer Vision*. 2017.

Semi-supervised segmentation with GANs

Can we use GAN-generated images to boost learning in a semi-supervised setting ?



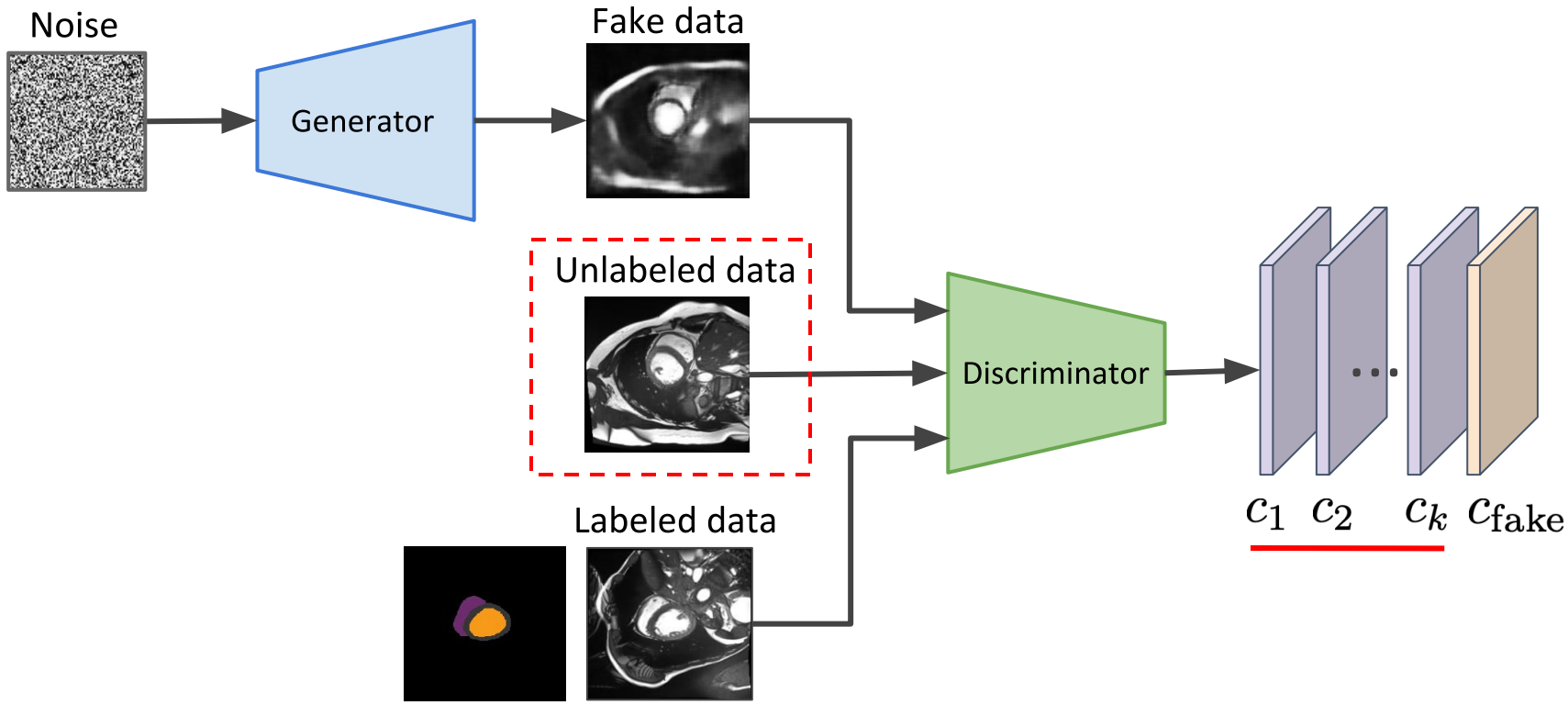
Labeled data: Predict the ground-truth label as in standard supervised segmentation

$$\mathcal{L}_{\text{sup}}(D) = \mathbb{E}_{(x,y)\sim p_{\text{data}}(x,y)}\left[-\sum_i \log p(Y_i = y_i \mid x)\right]$$

Adapted from: Souly, N. et al.. "Semi supervised semantic segmentation using generative adversarial network." *IEEE Int. Conf. on Computer Vision*. 2017.

Semi-supervised segmentation with GANs

Can we use GAN-generated images to boost learning in a semi-supervised setting ?



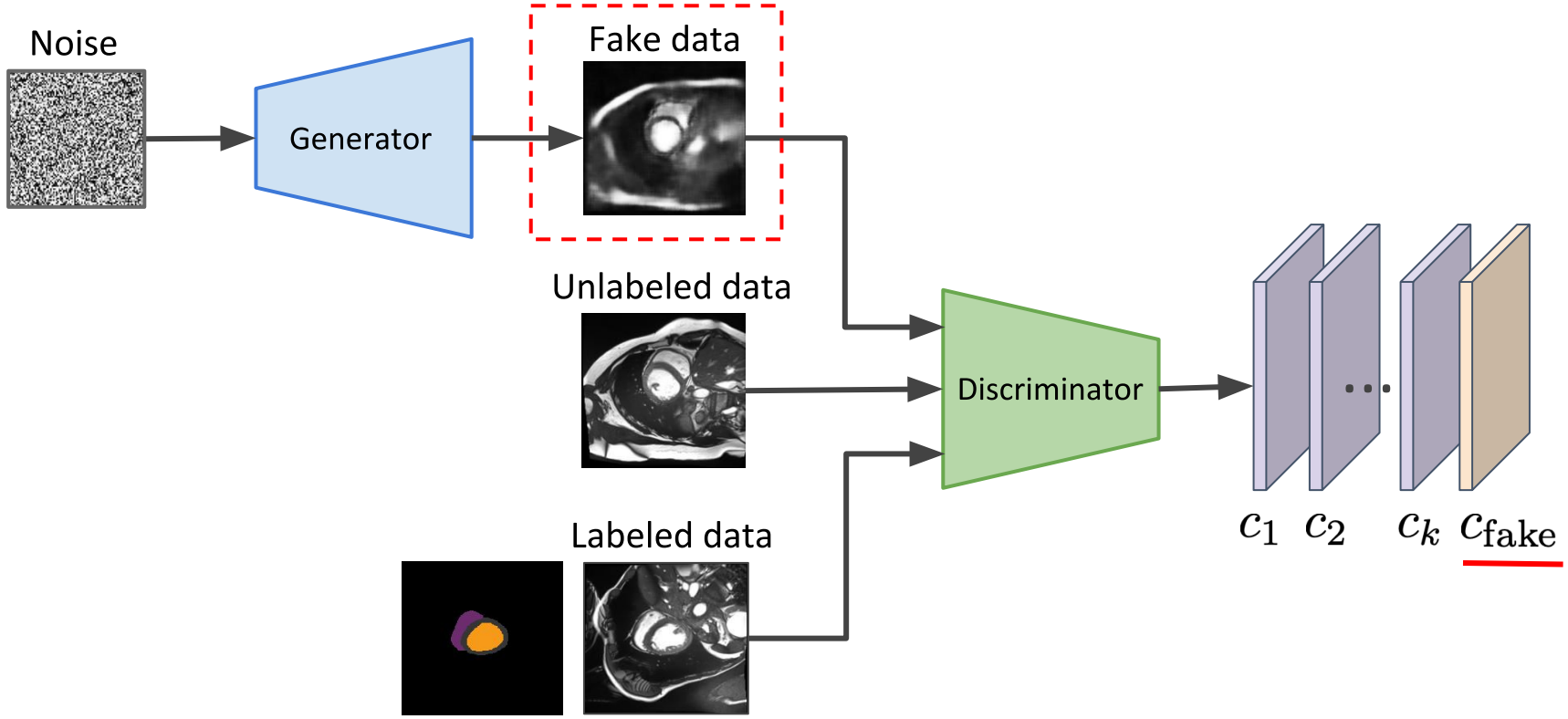
Unlabeled data: Predict the any label except fake

$$\mathcal{L}_{\text{unsup}}(D) = \mathbb{E}_{x \sim p_{\text{data}}(x)} \left[-\sum_i \log p(Y_i \neq \text{fake} \,|\, x) \right] = \mathbb{E}_{x \sim p_{\text{data}}(x)} \left[-\sum_i \log \left(1 - p(Y_i = \text{fake} \,|\, x)\right) \right]$$

Adapted from: Souly, N. et al.. "Semi supervised semantic segmentation using generative adversarial network." *IEEE Int. Conf. on Computer Vision*. 2017.

Semi-supervised segmentation with GANs

Can we use GAN-generated images to boost learning in a semi-supervised setting ?



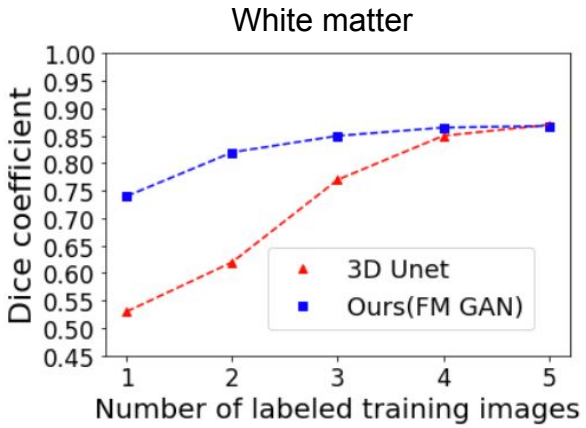
Fake data: Predict the label fake at <u>every pixel</u>

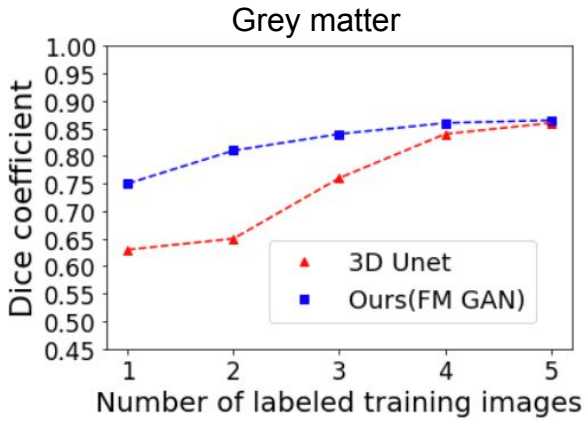
$$\mathcal{L}_{\text{fake}}(G, D) = \mathbb{E}_{z \sim p_z(z)} \left[-\sum_i \log p(Y_i = \text{fake} \,|\, G(z)) \right]$$

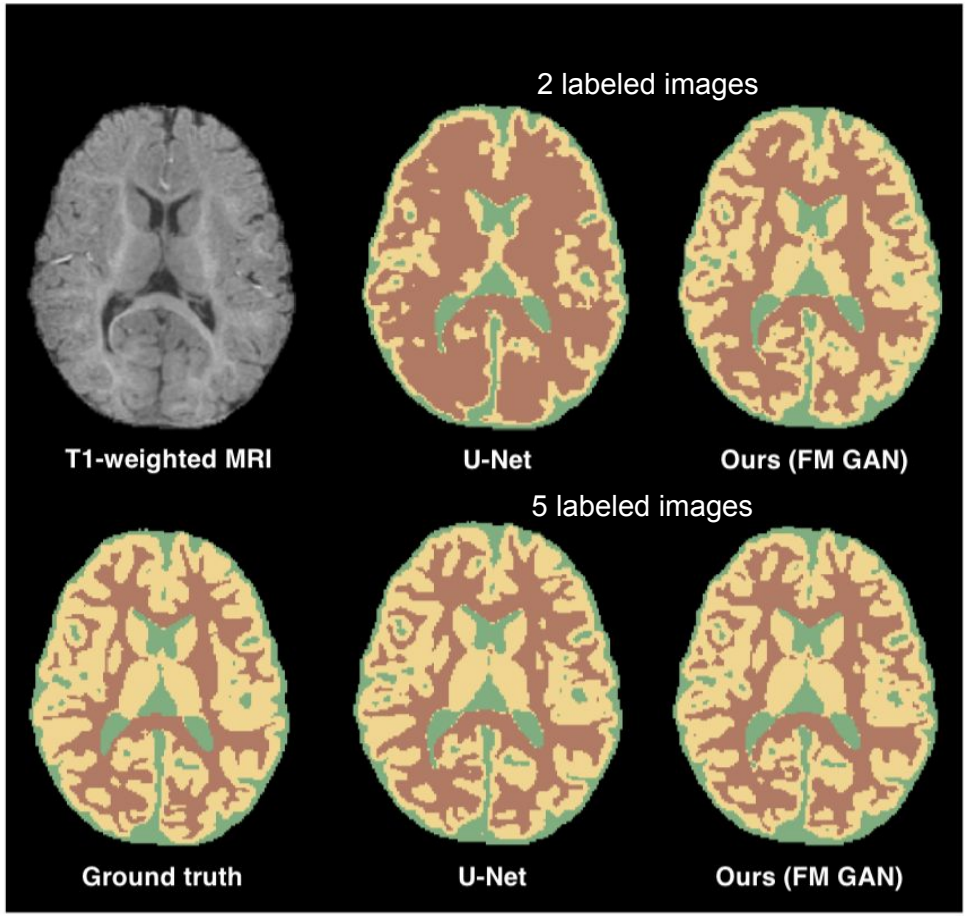
Adapted from: Souly, N. et al.. "Semi supervised semantic segmentation using generative adversarial network." *IEEE Int. Conf. on Computer Vision*. 2017.

Semi-supervised segmentation with GANs

Application to brain segmentation with very few training images







Mondal, A.K., Dolz, J. and Desrosiers, C. "Few-shot 3D multi-modal medical image segmentation using generative adversarial learning." *arXiv* (2018).

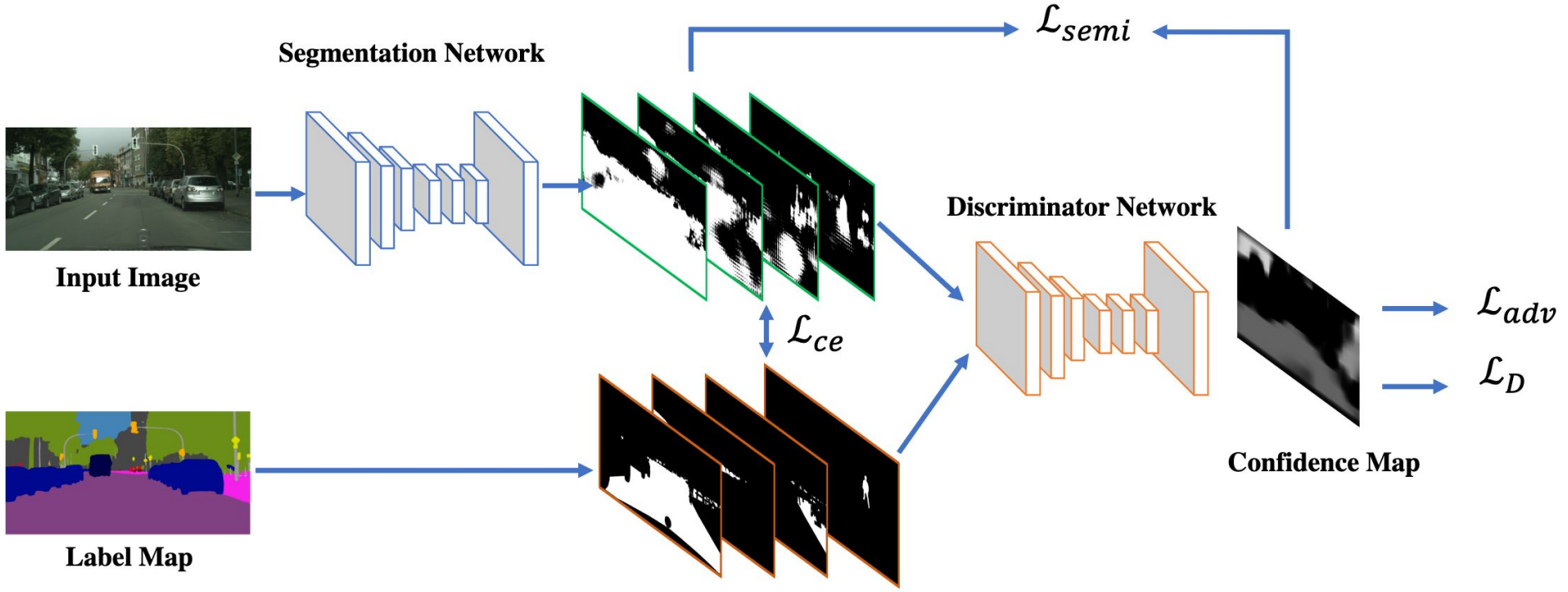
Adversarial model for self-training

Can we leverage discriminator predictions at the pixel-level ?

Hung, W.-C. et al. "Adversarial learning for semi-supervised semantic segmentation." *British Machine Vision Conference* (*BMVC*), 2018.

Adversarial model for self-training

Can we leverage discriminator predictions at the pixel-level ?

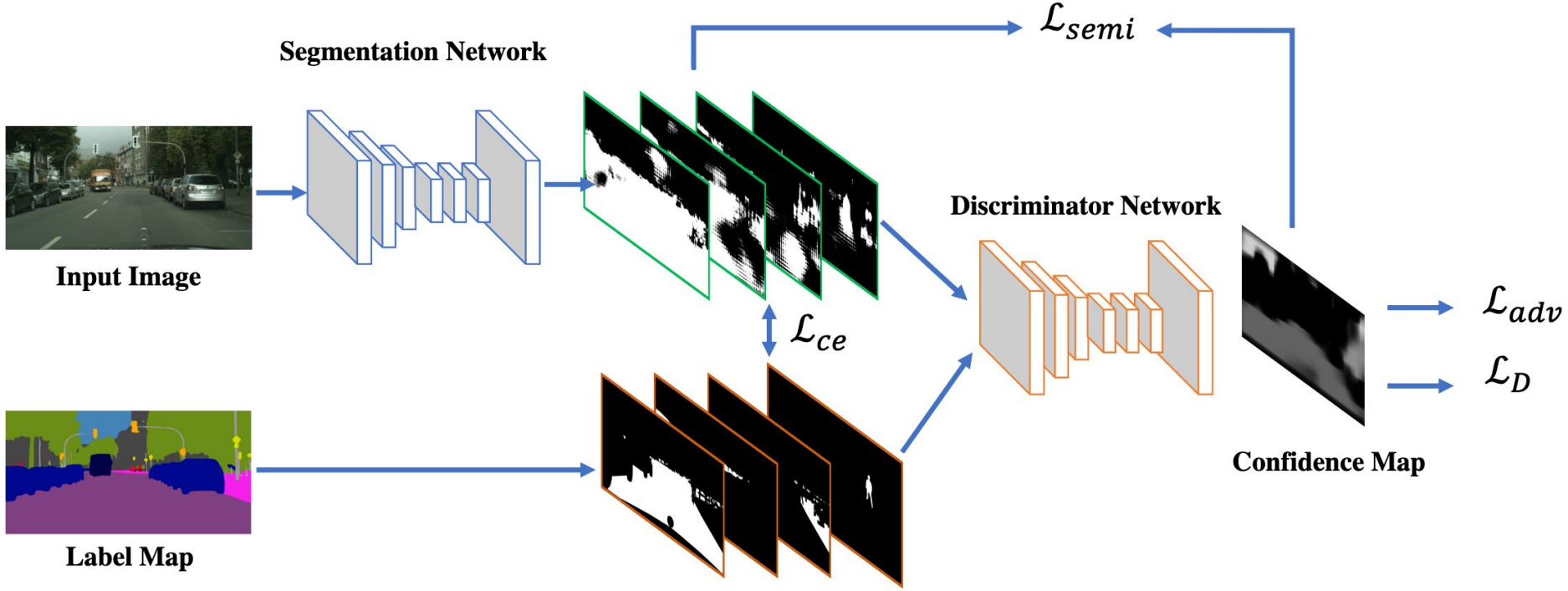


- The discriminator must discriminate between prediction and ground-truth (GT) at each pixel
- Consider the discriminator GT class probabilities as confidence scores
- Use high-confidence predictions on unlabeled images as pseudo-labels for self-training

Hung, W.-C. et al. "Adversarial learning for semi-supervised semantic segmentation." *British Machine Vision Conference* (*BMVC*), 2018.

Adversarial model for self-training

Can we leverage discriminator predictions at the pixel-level ?



Select pixels with confidence above a given threshold

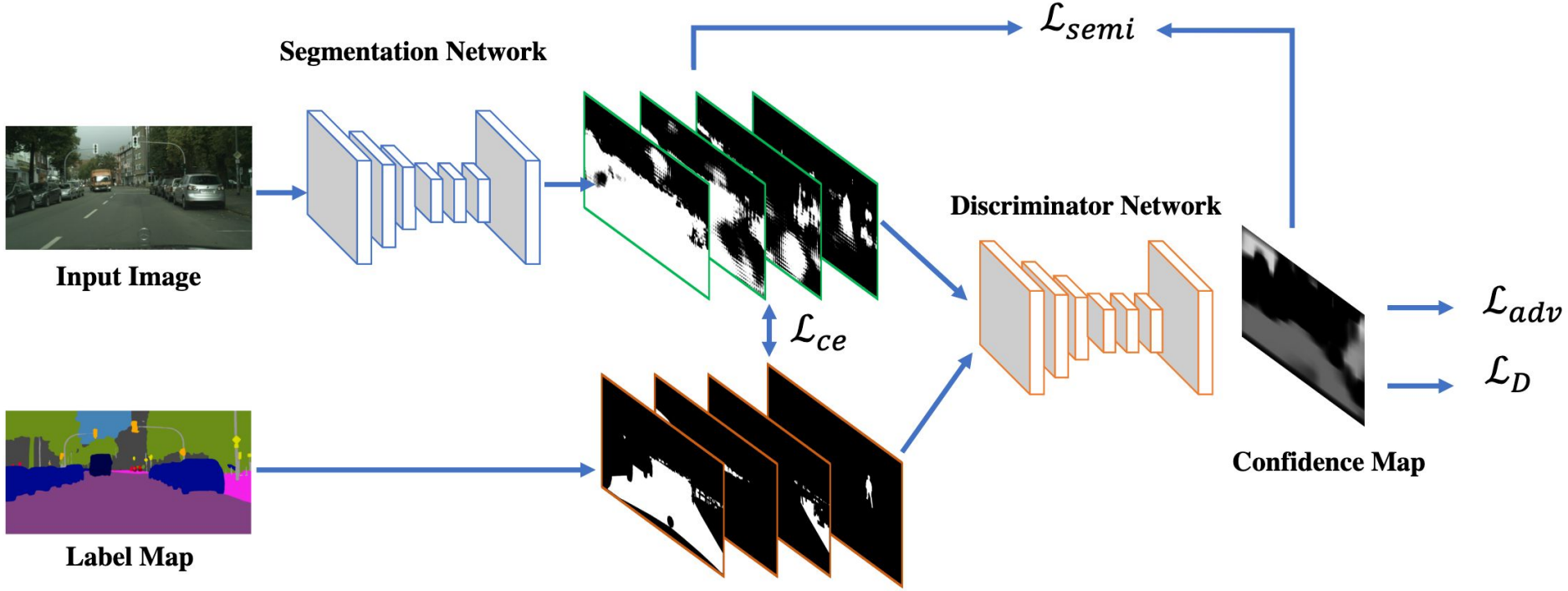
$$\mathcal{L}_{semi} = -\sum_{h,w}\sum_{c\in C} I(D(S(\mathbf{X}_n))^{(h,w)} > T_{semi}) \cdot \hat{\mathbf{Y}}_n^{(h,w,c)} \log(S(\mathbf{X}_n)^{(h,w,c)})$$

$$\hat{\mathbf{Y}}_n^{(h,w,c^*)} = 1 \text{ if } c^* = \arg\max_c S(\mathbf{X}_n)^{(h,w,c)}$$

Hung, W.-C. et al. "Adversarial learning for semi-supervised semantic segmentation." *British Machine Vision Conference* (*BMVC*), 2018.

Adversarial model for self-training

Can we leverage discriminator predictions at the pixel-level ?



Use class with highest probability as pseudo-label

$$\mathcal{L}_{semi} = -\sum_{h,w}\sum_{c\in C} I(D(S(\mathbf{X}_n))^{(h,w)} > T_{semi})\, \hat{\mathbf{Y}}_n^{(h,w,c)} \log(S(\mathbf{X}_n)^{(h,w,c)})$$

$$\hat{\mathbf{Y}}_n^{(h,w,c^*)} = 1 \text{ if } c^* = \arg\max_c S(\mathbf{X}_n)^{(h,w,c)}$$

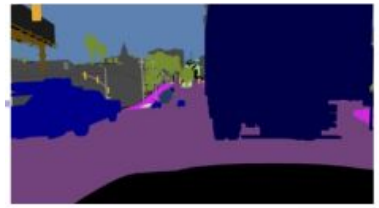
Hung, W.-C. et al. "Adversarial learning for semi-supervised semantic segmentation." *British Machine Vision Conference* (*BMVC*), 2018.

Cycle GANs for domain adaptation

How can we learn a model to segment target images without ground-truth ?

Source domain

Image



Ground-truth

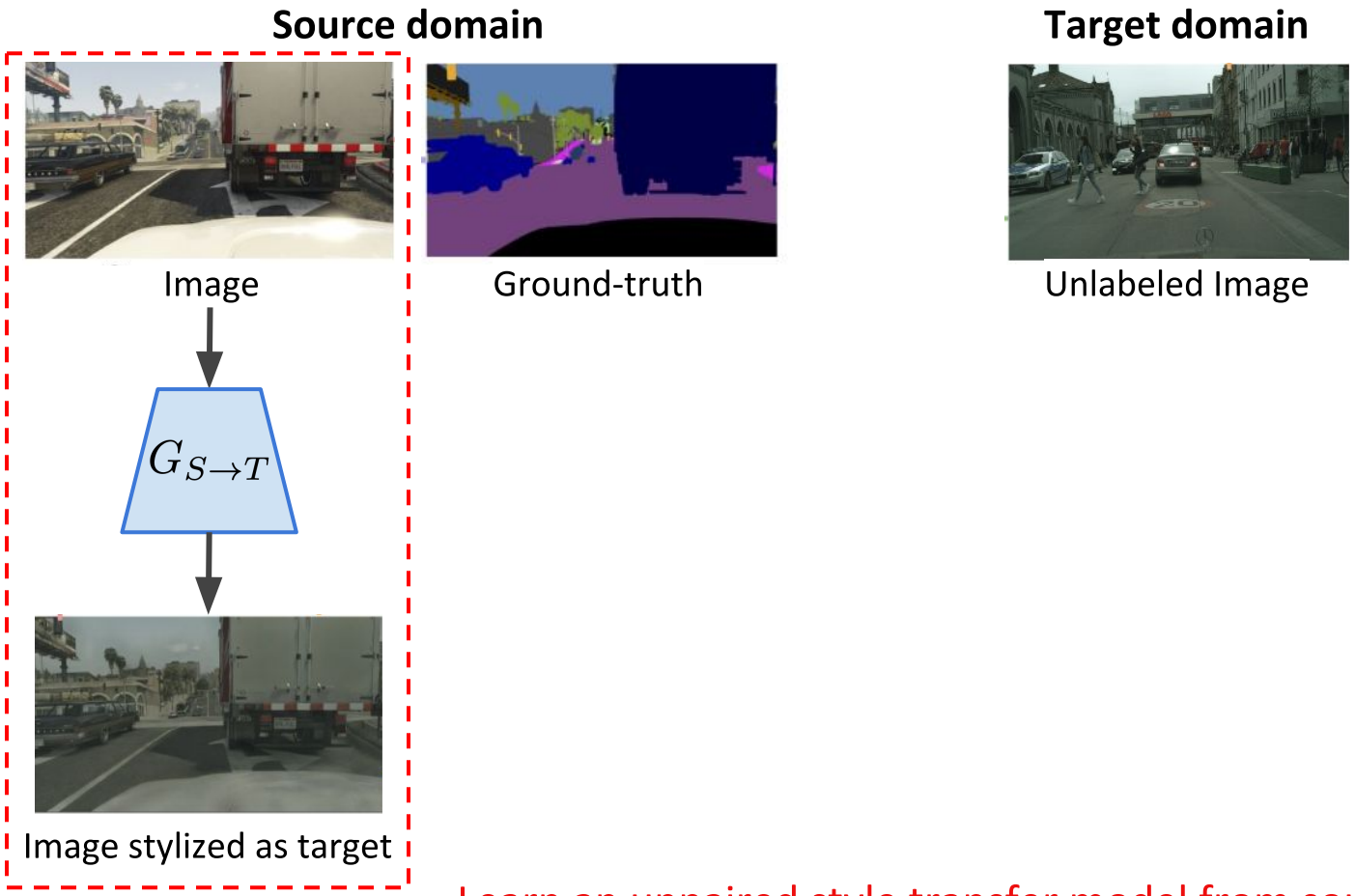
Target domain

Unlabeled Image

Hoffman, J., et al. "CyCADA: Cycle-Consistent Adversarial Domain Adaptation." *Int. Conf. on Machine Learning* (ICML). 2018.

Cycle GANs for domain adaptation

How can we learn a model to segment target images without ground-truth ?

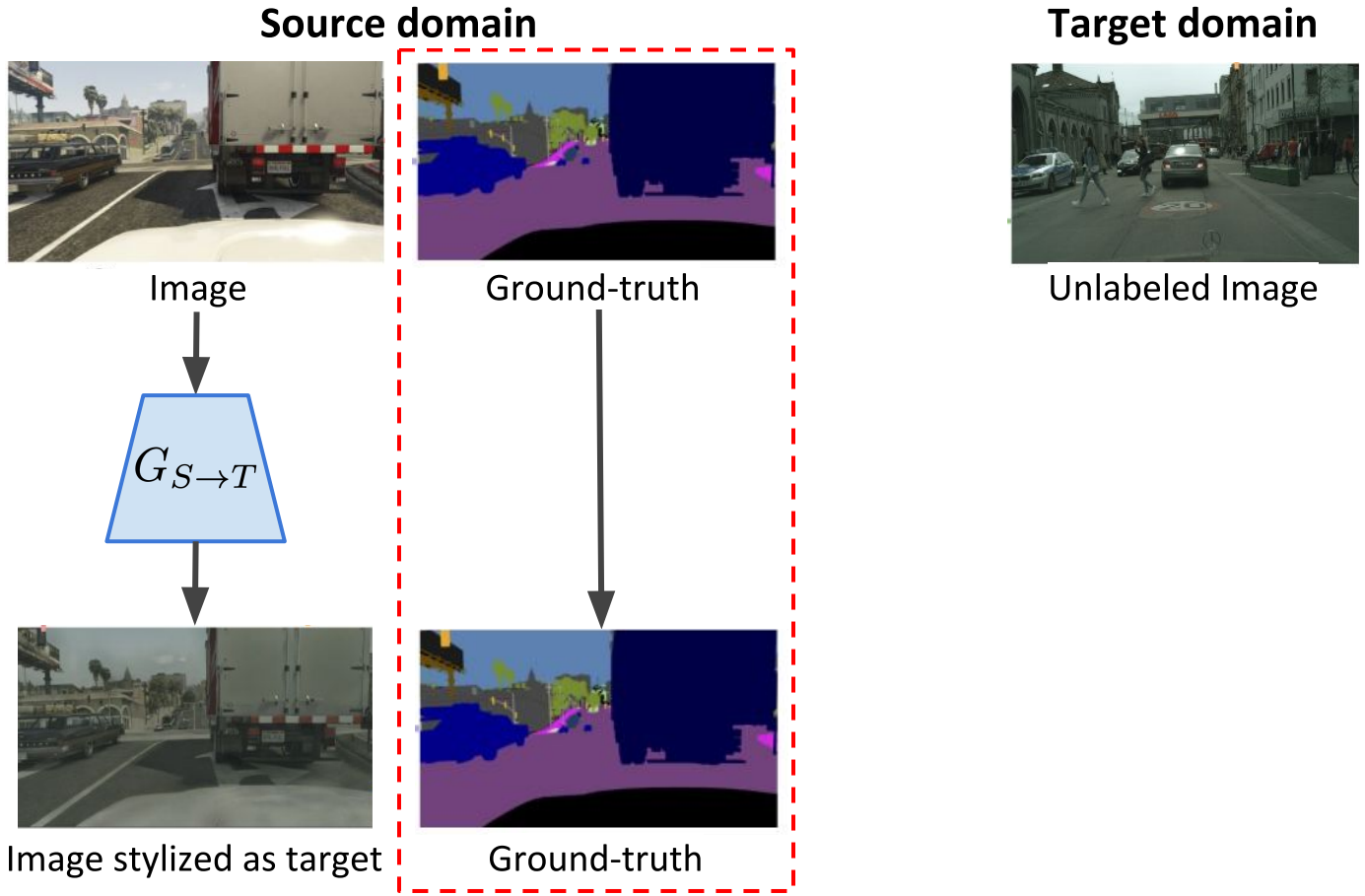


Learn an unpaired style transfer model from source to target domain

Hoffman, J., et al. "CyCADA: Cycle-Consistent Adversarial Domain Adaptation." *Int. Conf. on Machine Learning* (ICML). 2018.

Cycle GANs for domain adaptation

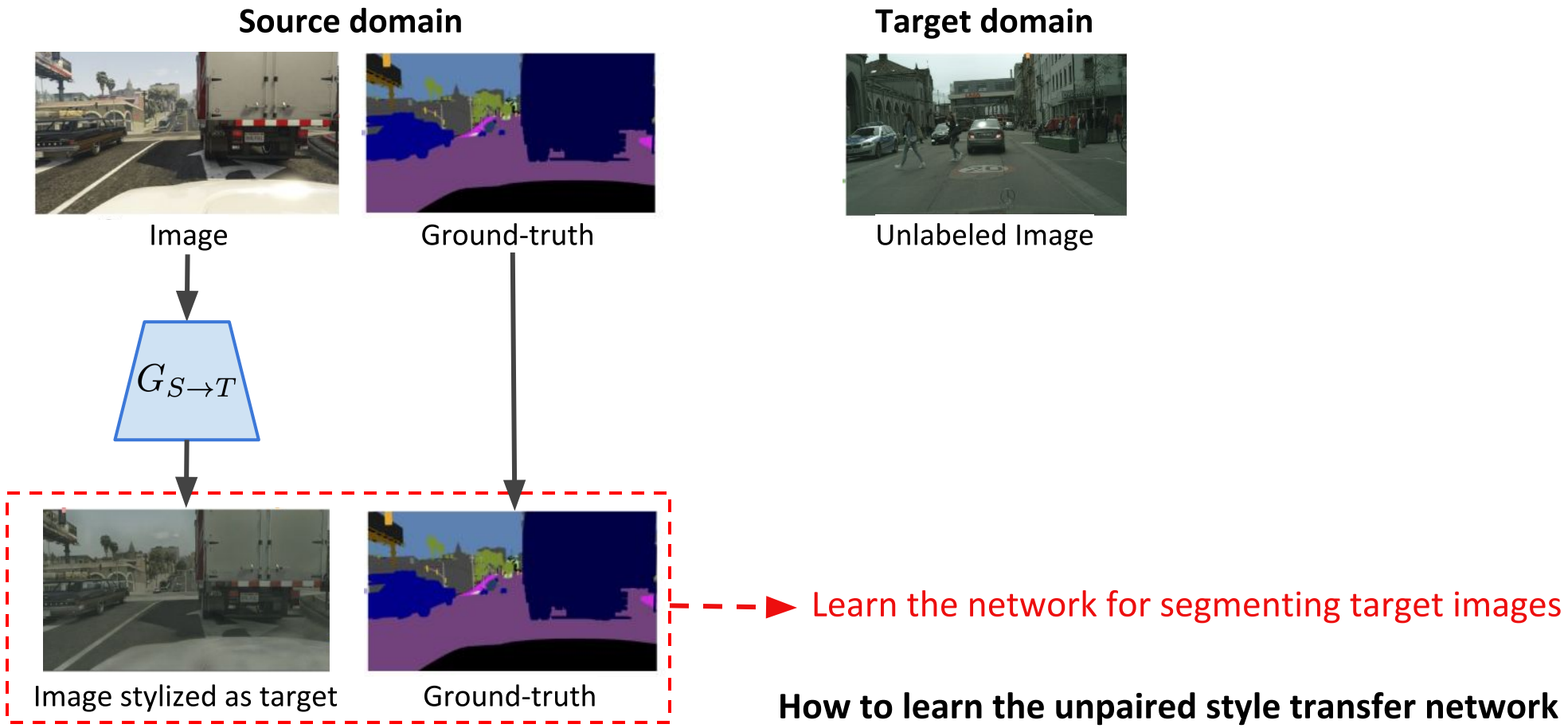
How can we learn a model to segment target images without ground-truth ?



Hoffman, J., et al. "CyCADA: Cycle-Consistent Adversarial Domain Adaptation." *Int. Conf. on Machine Learning* (ICML). 2018.

Cycle GANs for domain adaptation

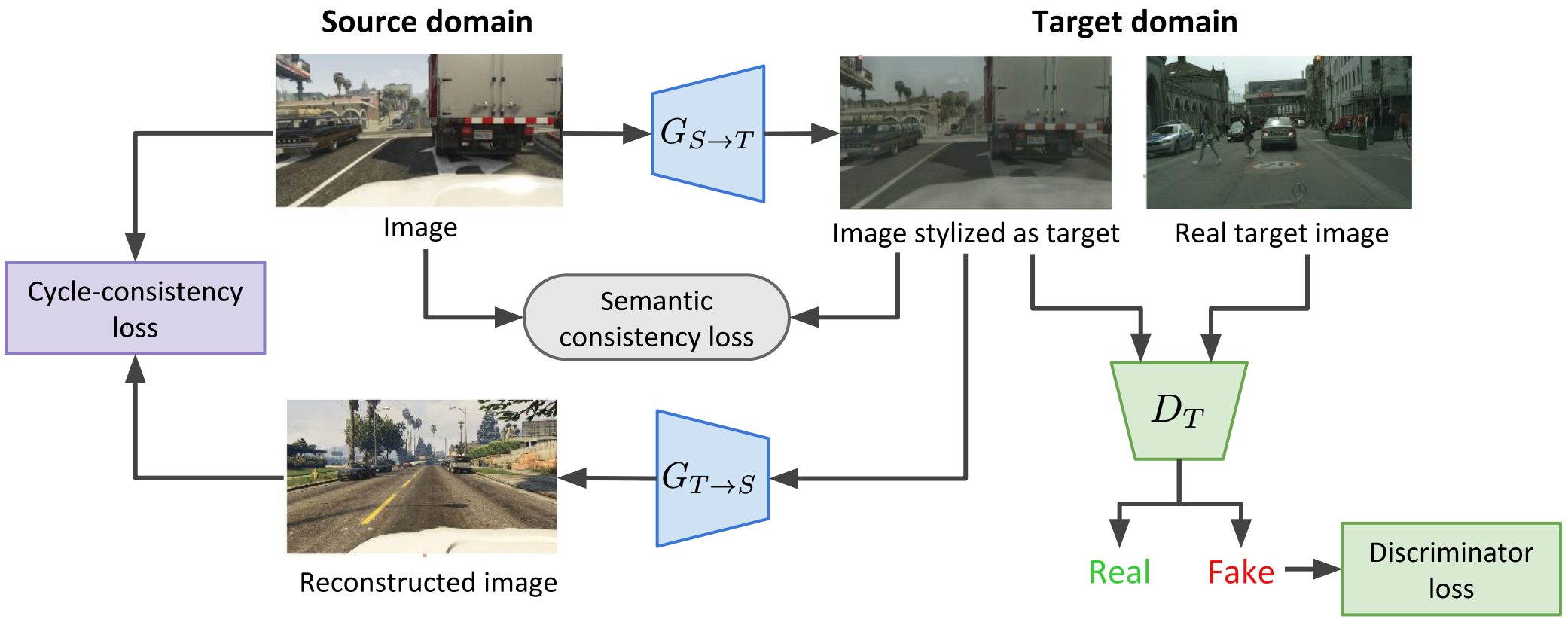
How can we learn a model to segment target images without ground-truth ?



Hoffman, J., et al. "CyCADA: Cycle-Consistent Adversarial Domain Adaptation." *Int. Conf. on Machine Learning* (ICML). 2018.

Cycle GANs for domain adaptation

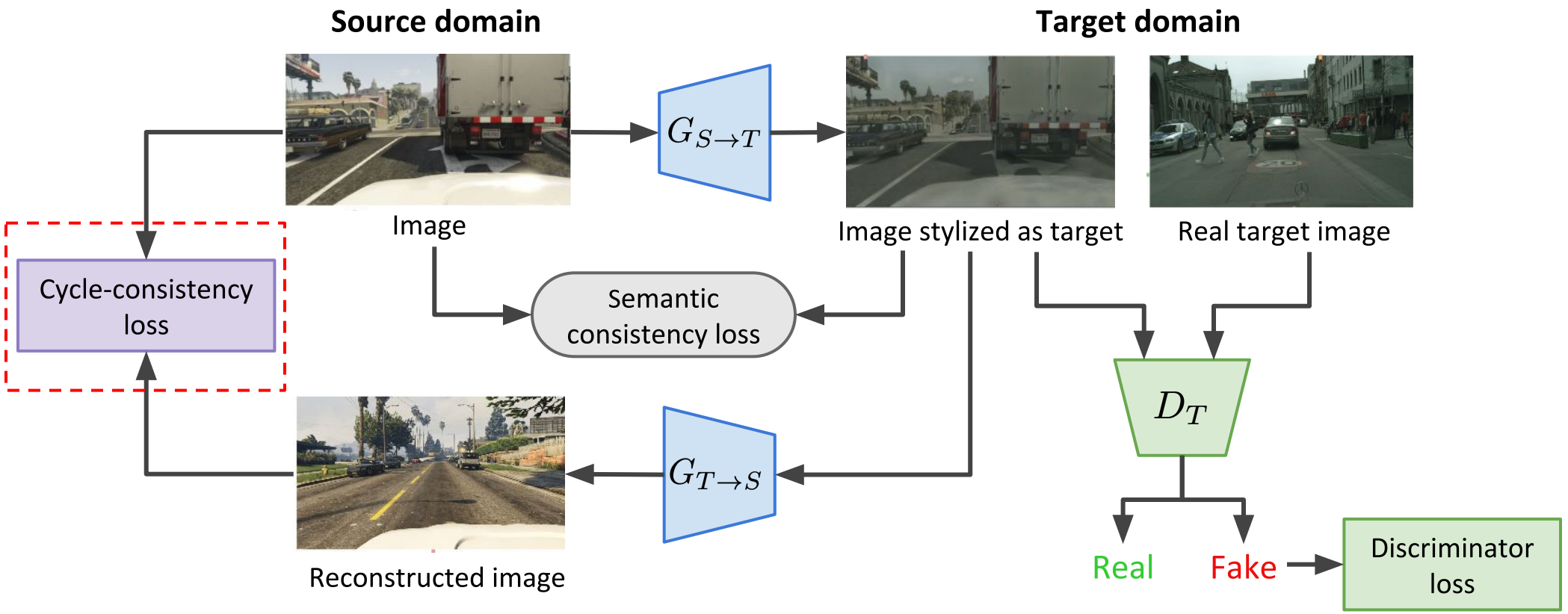
How can we learn a model to segment target images without ground-truth ?



Hung, W.-C. et al. "Adversarial learning for semi-supervised semantic segmentation." *British Machine Vision Conference* (*BMVC*), 2018.

Cycle GANs for domain adaptation

How can we learn a model to segment target images without ground-truth ?

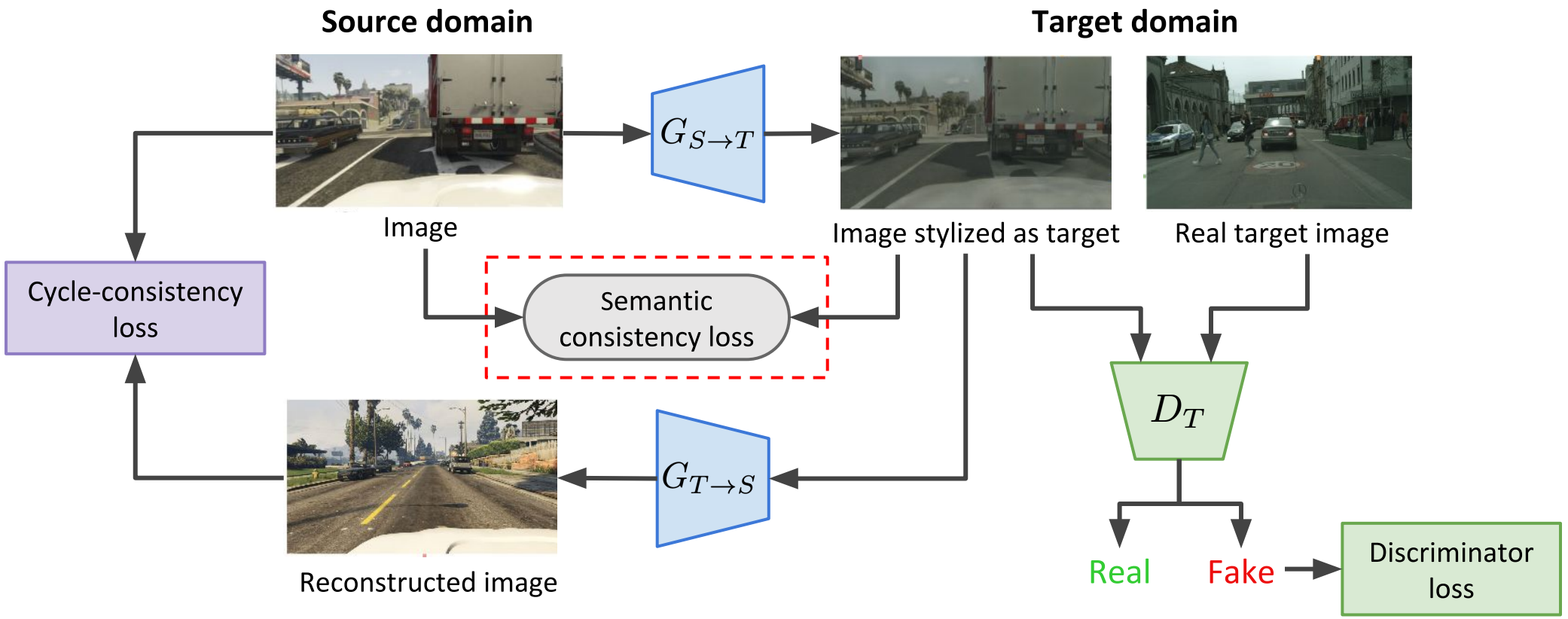


Cycle consistency loss: $L_{\text{cycle}}(G_{S\to T}, G_{T\to S}) = \mathbb{E}_{x\sim p_S(x)}\left[\|x - G_{T\to S}(G_{S\to T}(x))\|_1\right]$

Hoffman, J., et al. "CyCADA: Cycle-Consistent Adversarial Domain Adaptation." *Int. Conf. on Machine Learning* (ICML). 2018.

Cycle GANs for domain adaptation

How can we learn a model to segment target images without ground-truth ?



Semantic consistency loss: Segmentation for the source image and its stylized target version should be consistent

Hoffman, J., et al. "CyCADA: Cycle-Consistent Adversarial Domain Adaptation." *Int. Conf. on Machine Learning* (ICML). 2018.

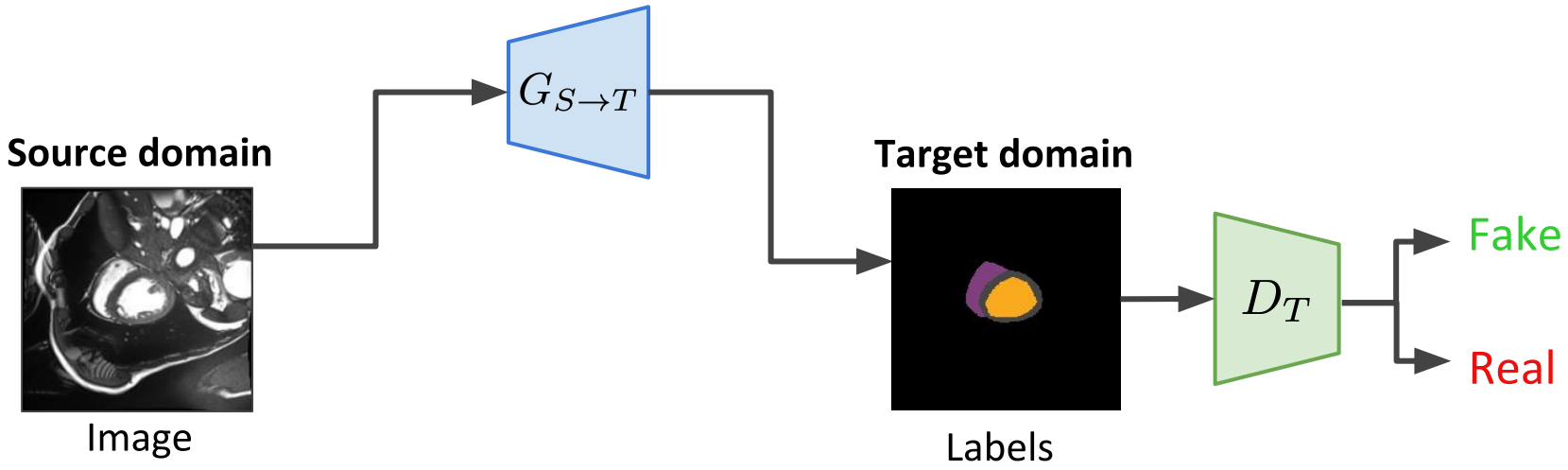
Cycle GANs for semi-supervised segmentation

How to apply the same idea when images are from a single domain (i.e., semi-supervised) ?

Mondal, A.K., et al. "Revisiting CycleGAN for semi-supervised segmentation." arXiv (2019).

Cycle GANs for semi-supervised segmentation

How to apply the same idea when images are from a single domain (i.e., semi-supervised) ?

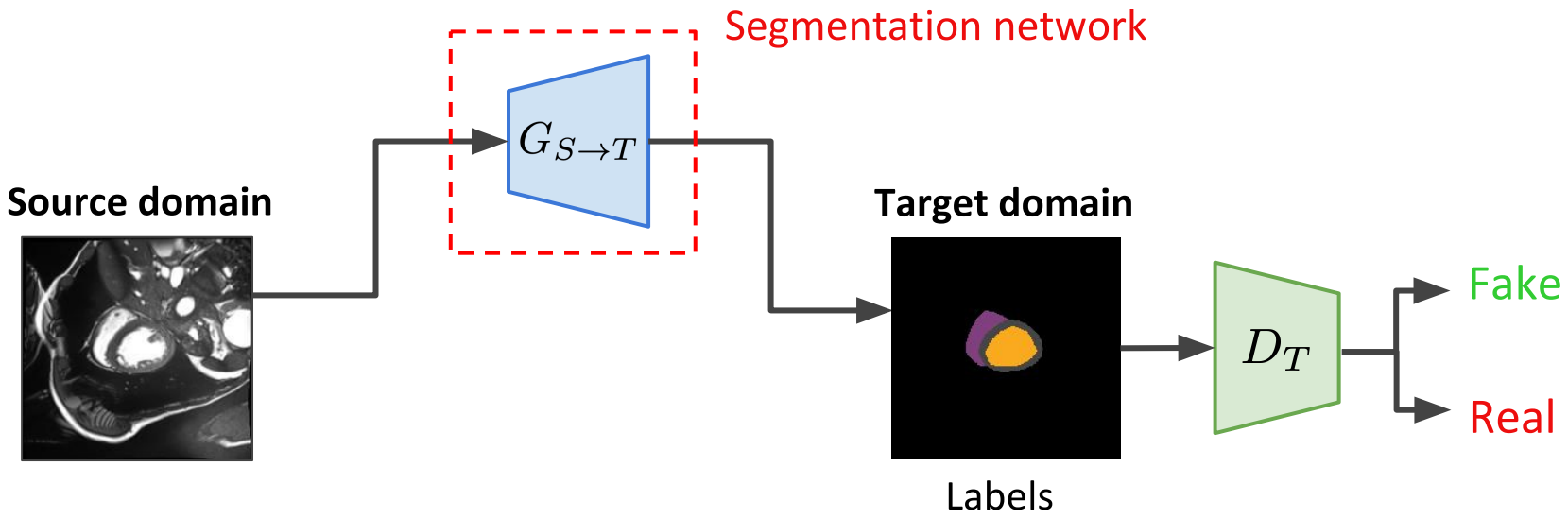


Standard adversarial model for semi-supervised segmentation, e.g. (Zhang et al., 2017)

Mondal, A.K., et al. "Revisiting CycleGAN for semi-supervised segmentation." arXiv (2019).

Cycle GANs for semi-supervised segmentation

How to apply the same idea when images are from a single domain (i.e., semi-supervised) ?

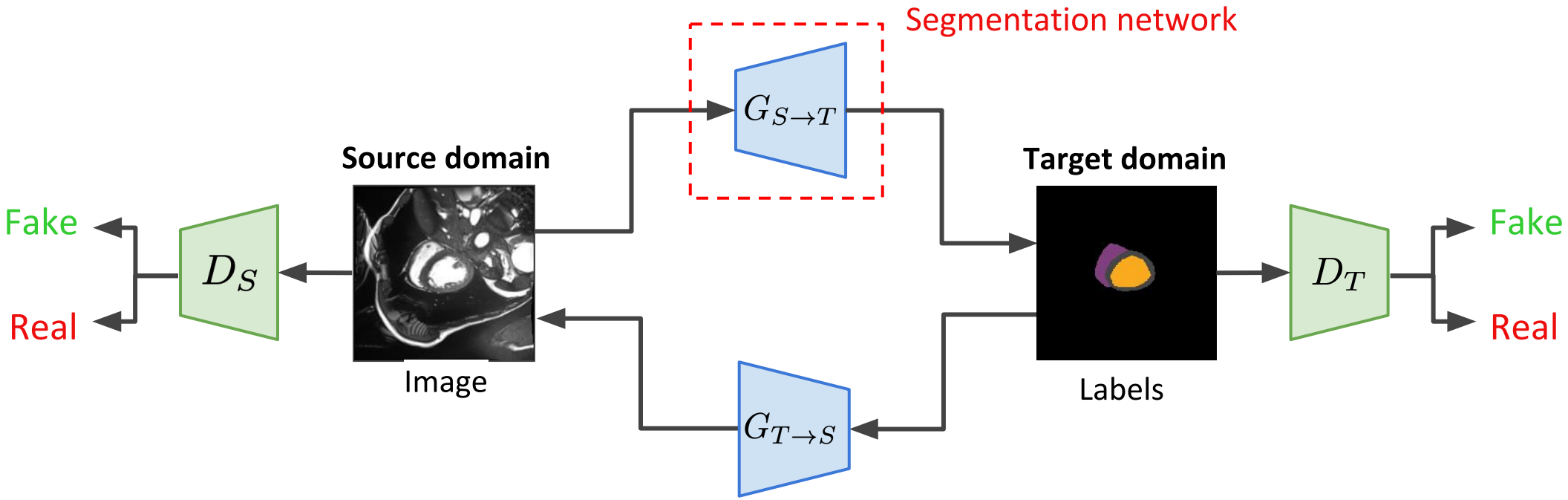


Standard adversarial model for semi-supervised segmentation, e.g. (Zhang et al., 2017)

Mondal, A.K., et al. "Revisiting CycleGAN for semi-supervised segmentation." arXiv (2019).

Cycle GANs for semi-supervised segmentation

How to apply the same idea when images are from a single domain (i.e., semi-supervised) ?



Intuition:

- Segmentations consistent with real data will lead to more plausible generated images
- Use image cycle consistency as additional loss for training with unlabeled images

Mondal, A.K., et al. "Revisiting CycleGAN for semi-supervised segmentation." arXiv (2019).

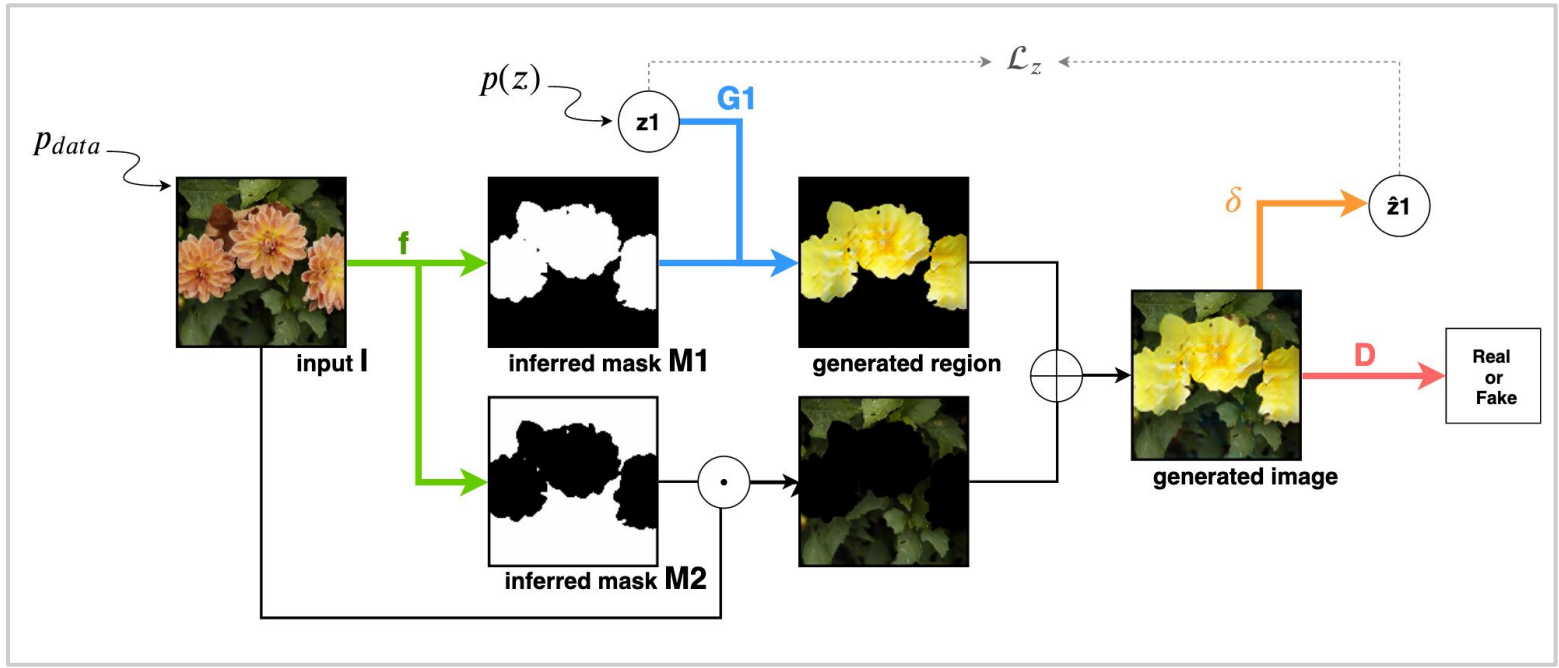
Cycle GANs for semi-supervised segmentation

Can we use adversarial learning to train a segmentation network without any labeled data ?

Chen, M., et al. "Unsupervised Object Segmentation by Redrawing." arXiv (2019).

GANs for unsupervised segmentation

Can we use adversarial learning to train a segmentation network without any labeled data ?



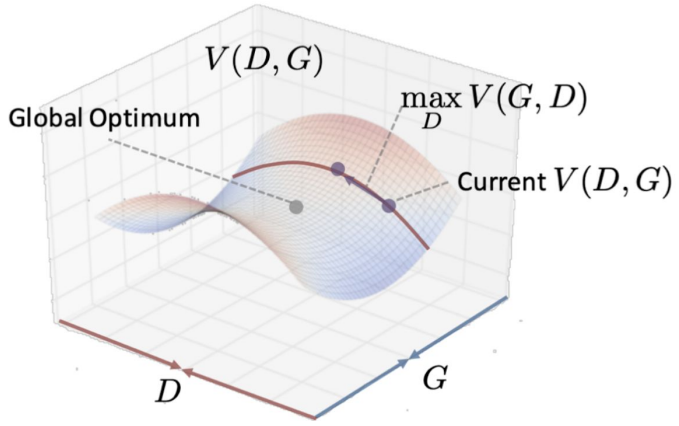
Basic idea:

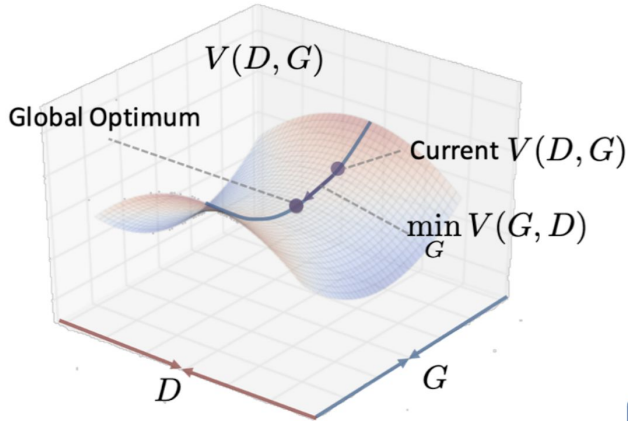
- Predict a segmentation mask for each unknown class
- Generate a fake image for each unknown class and compose them using respective masks
- Use a discriminator for enforcing composed image to be realistic

Chen, M., et al. "Unsupervised Object Segmentation by Redrawing." NeurIPS 2019

Challenges of adversarial learning

1) Unstable optimization of minimax problem

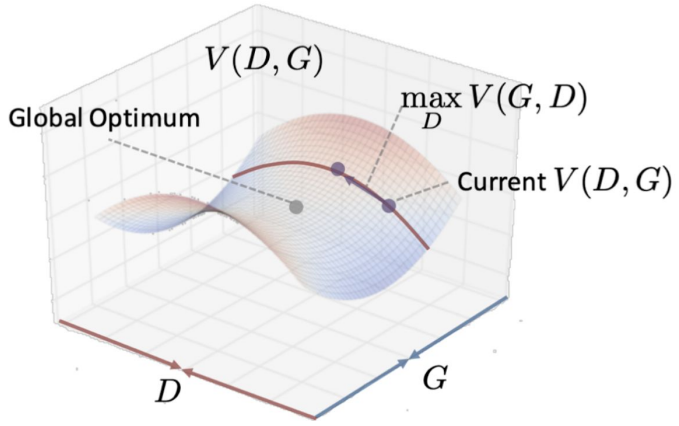


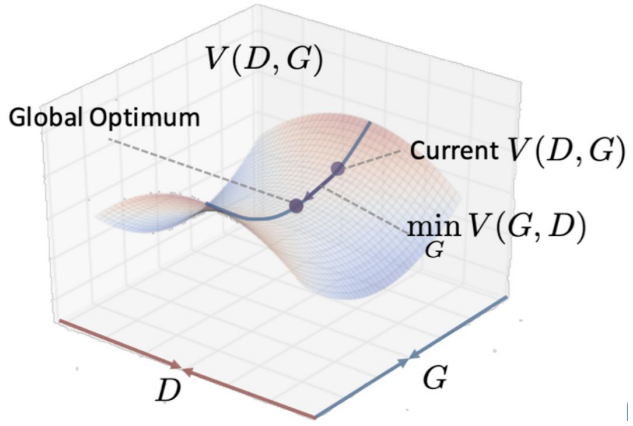


Ghosh, A., et al. "Multi-agent diverse generative adversarial networks." Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2018.
Chang, Mark. “Generative Adversarial Networks”, published online, 2016

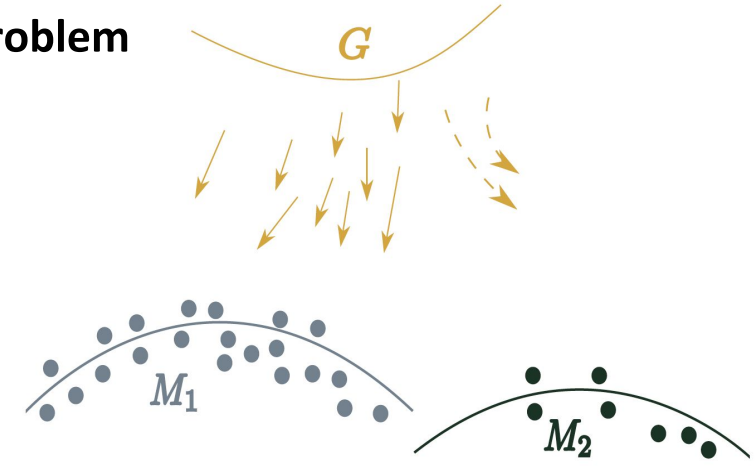
Challenges of adversarial learning

1) Unstable optimization of minimax problem





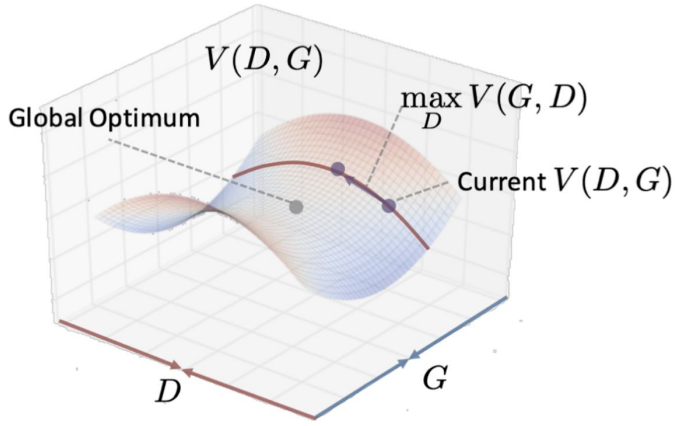
2) Mode collapse problem

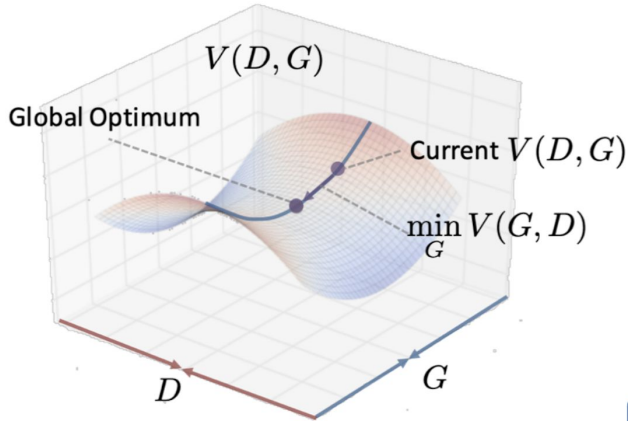


Ghosh, A., et al. "Multi-agent diverse generative adversarial networks." Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2018.
Chang, Mark. “Generative Adversarial Networks”, published online, 2016

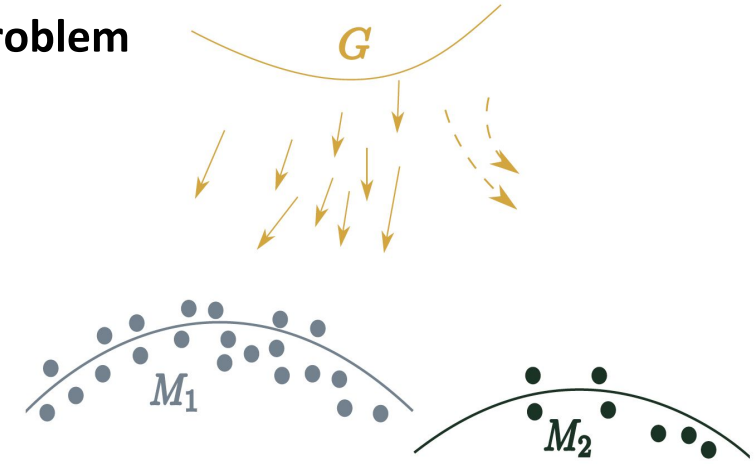
Challenges of adversarial learning

1) Unstable optimization of minimax problem





2) Mode collapse problem



Various solutions:

- Spectral normalization (Miyato *et al.*, 2018)
- Wasserstein GANs (Arjovsky *et al.*, 2017)
- Feature matching (Salimans *et al.*, 2016)
- etc.

Ghosh, A., et al. "Multi-agent diverse generative adversarial networks." Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2018.
Chang, Mark. “Generative Adversarial Networks”, published online, 2016

Concluding remarks

- Adversarial methods can potentially learn data priors without having to explicitly model them
- Enhances learning in a weakly-supervised setting by restricting plausible segmentations of partially-labeled or unlabeled images
- Helps adapt segmentation models across different data domains (e.g., acquisition modality or site)
- Not a silver bullet, can be very challenging at times
- Lots of exciting opportunities for future research

Thank you

Questions ?

References

[1] Bateson M, Kervadec H, Dolz J, Lombaert H, Ben Ayed I. Constrained domain adaptation for segmentation. In International Conference on Medical Image Computing and Computer-Assisted Intervention 2019 Oct 13 (pp. 326-334).

[2] Belkin M, Niyogi P, Sindhwani V. Manifold regularization: A geometric framework for learning from labeled and unlabeled examples. Journal of machine learning research. 2006;7(Nov):2399-434.

[3] Ben Ayed I, Wang M, Miles B, Garvin GJ. TRIC: Trust region for invariant compactness and its application to abdominal aorta segmentation. In International Conference on Medical Image Computing and Computer-Assisted Intervention 2014 Sep 14 (pp. 381-388).

[4] Boyd S, Parikh N, Chu E, Peleato B, Eckstein J. Distributed optimization and statistical learning via the alternating direction method of multipliers. Foundations and Trends® in Machine learning. 2011 Jul 26;3(1):1-22.

[5] Boykov Y, Veksler O, Zabih R. Fast approximate energy minimization via graph cuts. IEEE Transactions on Pattern Analysis and Machine Intelligence. 2001 Nov;23(11):1222-39.

[6] Carass A, Cuzzocreo JL, Han S, Hernandez-Castillo CR, Rasser PE, Ganz M, Beliveau V, Dolz J, Ben Ayed I, Desrosiers C, Thyreau B et al. Comparing fully automated state-of-the-art cerebellum parcellation from magnetic resonance images. NeuroImage. 2018 Dec 1;183:150-72.

[7] Chen LC, Papandreou G, Kokkinos I, Murphy K, Yuille AL. Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected CRFs. IEEE transactions on pattern analysis and machine intelligence. 2017 Apr 27;40(4):834-48.

[8] Chen M, Artières T, Denoyer L. Unsupervised Object Segmentation by Redrawing. In Advances in neural information processing systems 2019.

[9] Cordts M, Omran M, Ramos S, Rehfeld T, Enzweiler M, Benenson R, Franke U, Roth S, Schiele B. The cityscapes dataset for semantic urban scene understanding. InProceedings of the IEEE conference on computer vision and pattern recognition 2016 (pp. 3213-3223).

[10] Dolz J, Desrosiers C, Ben Ayed I. 3D fully convolutional networks for subcortical segmentation in MRI: A large-scale study. NeuroImage. 2018 Apr 15;170:456-70.

[11] Dou Q, Ouyang C, Chen C, Chen H, Glocker B, Zhuang X, Heng PA. PnP-AdaNet: Plug-and-play adversarial domain adaptation network with a benchmark at cross-modality cardiac segmentation. IEEE Access, 7:99065–99076, 2019.

[12] Fechter T, Adebahr S, Baltas D, Ben Ayed I, Desrosiers C, Dolz J. Esophagus segmentation in CT via 3D fully convolutional neural network and random walk. Medical physics. 2017 Dec 1;44(12):6341-52.

[13] Ganaye PA, Sdika M, Benoit-Cattin H. Semi-supervised learning for segmentation under semantic constraint. In International Conference on Medical Image Computing and Computer-Assisted Intervention 2018 Sep 16 (pp. 595-602).

[14] Ghosh A, Kulharia V, Namboodiri VP, Torr PH, Dokania PK. Multi-agent diverse generative adversarial networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition 2018 (pp. 8513-8521).

[15] Grandvalet Y, Bengio Y. Semi-supervised learning by entropy minimization. In Advances in neural information processing systems 2005 (pp. 529-536).

[16] Hoffman J, Tzeng E, Park T, Zhu JY, Isola P, Saenko K, Efros AA, Darrell T. Cycada: Cycle-consistent adversarial domain adaptation. In International Conference on Machine Learning (ICML). 2018

References

[17] Hung WC, Tsai YH, Liou YT, Lin YY, Yang MH. Adversarial learning for semi-supervised semantic segmentation. In the British Machine Vision Conference (BMVC) 2018.

[18] Ji Z, Shen Y, Ma C, Gao M. Scribble-Based Hierarchical Weakly Supervised Learning for Brain Tumor Segmentation. In International Conference on Medical Image Computing and Computer-Assisted Intervention 2019 Oct 13 (pp. 175-183).

[19] Jia Z, Huang X, Eric I, Chang C, Xu Y. Constrained deep weak supervision for histopathology image segmentation. IEEE Transactions on Medical Imaging. 2017 Jul 7;36(11):2376-88.

[20] Kervadec H, Dolz J, Tang M, Granger E, Boykov Y, Ben Ayed I. Constrained-CNN losses for weakly supervised segmentation. Medical image analysis. 2019 May 1;54:88-99.

[21] Kervadec H, Dolz J, Yuan J, Desrosiers C, Granger E, Ben Ayed I. Constrained Deep Networks: Lagrangian Optimization via Log-Barrier Extensions. arXiv preprint arXiv:1904.04205. 2019 Apr 8.

[22] Kervadec H, Dolz J, Granger E, Ben Ayed I. Curriculum semi-supervised segmentation. In International Conference on Medical Image Computing and Computer-Assisted Intervention 2019 Oct 13 (pp. 568-576).

[23] Krähenbühl P, Koltun V. Efficient inference in fully connected crfs with gaussian edge potentials. In Advances in neural information processing systems 2011 (pp. 109-117).

[24] Krause A, Perona P, Gomes RG. Discriminative clustering by regularized information maximization. In Advances in neural information processing systems 2010 (pp. 775-783).

[25] Larrazabal AJ, Martinez C, Ferrante E. Anatomical Priors for Image Segmentation via Post-Processing with Denoising Autoencoders. In International Conference on Medical Image Computing and Computer-Assisted Intervention 2019.

[26] Lee DH. Pseudo-label: The simple and efficient semi-supervised learning method for deep neural networks. In Workshop on Challenges in Representation Learning, ICML 2013 Jun 21 (Vol. 3, p. 2).

[27] Lin D, Dai J, Jia J, He K, Sun J. Scribblesup: Scribble-supervised convolutional networks for semantic segmentation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition 2016 (pp. 3159-3167).

[28] Marin D, Tang M, Ayed IB, Boykov Y. Beyond Gradient Descent for Regularized Segmentation Losses. In IEEE conference on Computer Vision and Pattern Recognition (CVPR) 2019.

[29] Mondal AK, Dolz J, Desrosiers C. Few-shot 3D multi-modal medical image segmentation using generative adversarial learning. arXiv preprint arXiv:1810.12241. 2018 Oct 29.

[30] Mondal AK, Agarwal A, Dolz J, Desrosiers C. Revisiting CycleGAN for semi-supervised segmentation. arXiv preprint arXiv:1908.11569. 2019 Aug 30.

[31] Njeh I, Sallemi L, Ayed IB, Chtourou K, Lehericy S, Galanaud D, Hamida AB. 3D multimodal MRI brain glioma tumor and edema segmentation: a graph cut distribution matching approach. Computerized Medical Imaging and Graphics. 2015 Mar 1;40:108-19.

[32] Oktay O, Ferrante E, Kamnitsas K, Heinrich M, Bai W, Caballero J, Cook SA, De Marvao A, Dawes T, O'Regan DP, Kainz B. Anatomically constrained neural networks (ACNNs): application to cardiac image enhancement and segmentation. IEEE transactions on medical imaging. 2017 Sep 26;37(2):384-95.

[33] Pathak D, Krahenbuhl P, Darrell T. Constrained convolutional neural networks for weakly supervised segmentation. In Proceedings of the IEEE international conference on computer vision 2015 (pp. 1796-1804).

[34] Painchaud N, Skandarani Y, Judge T, Bernard O, Lalande A, Jodoin PM. Cardiac MRI Segmentation with Strong Anatomical Guarantees. In International Conference on Medical Image Computing and Computer-Assisted Intervention 2019 Oct 13 (pp. 632-640).

References

[35] Qu H, Wu P, Huang Q, Yi J, Riedlinger GM, De S, Metaxas DN. Weakly Supervised Deep Nuclei Segmentation using Points Annotation in Histopathology Images. In International Conference on Medical Imaging with Deep Learning 2019 May 24 (pp. 390-400).

[36] Ravishankar H, Venkataramani R, Thiruvenkadam S, Sudhakar P, Vaidya V. Learning and incorporating shape models for semantic segmentation. In International Conference on Medical Image Computing and Computer-Assisted Intervention 2017 Sep 10 (pp. 203-211).

[37] Souly N, Spampinato C, Shah M. Semi supervised semantic segmentation using generative adversarial network. InProceedings of the IEEE International Conference on Computer Vision 2017 (pp. 5688-5696).

[38] Tang M, Perazzi F, Djelouah A, Ben Ayed I, Schroers C, Boykov Y. On regularized losses for weakly-supervised CNN segmentation. In Proceedings of the European Conference on Computer Vision (ECCV) 2018 (pp. 507-522).

[39] Tsai YH, Hung WC, Schulter S, Sohn K, Yang MH, Chandraker M. Learning to adapt structured output space for semantic segmentation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition 2018 (pp. 7472-7481).

[40] Vu TH, Jain H, Bucher M, Cord M, Pérez P. Advent: Adversarial entropy minimization for domain adaptation in semantic segmentation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition 2019 (pp. 2517-2526).

[41] Wang L, Nie D, Li G, Puybareau É, Dolz J, Zhang Q, Wang F, Xia J, Wu Z, Chen J, Thung KH et al. Benchmark on automatic 6-month-old infant brain segmentation algorithms: the iSeg-2017 challenge. IEEE transactions on medical imaging. 2019 Feb 27.

[42] Weston J, Ratle F, Mobahi H, Collobert R. Deep learning via semi-supervised embedding. In Neural Networks: Tricks of the Trade 2012 (pp. 639-655). Springer, Berlin, Heidelberg.

[43] Zhang Y, Yang L, Chen J, Fredericksen M, Hughes DP, Chen DZ. Deep adversarial networks for biomedical image segmentation utilizing unannotated images. In International Conference on Medical Image Computing and Computer-Assisted Intervention 2017 Sep 10 (pp. 408-416).

[44] Zhang Y, David P, Gong B. Curriculum domain adaptation for semantic segmentation of urban scenes. In Proceedings of the IEEE International Conference on Computer Vision 2017 (pp. 2020-2030).

[45] Zhang Y, David P, Foroosh H, Gong B. A Curriculum Domain Adaptation Approach to the Semantic Segmentation of Urban Scenes. IEEE transactions on pattern analysis and machine intelligence. 2019 Mar 6.

[46] Zhou Y, Li Z, Bai S, Wang C, Chen X, Han M, Fishman E, Yuille A. Prior-aware Neural Network for Partially-Supervised Multi-Organ Segmentation. In Proceedings of the IEEE International Conference on Computer Vision (ICCV) 2019

[47] Zhu Q, Du B, Yan P. Boundary-weighted Domain Adaptive Neural Network for Prostate MR Image Segmentation. arXiv preprint arXiv:1902.08128. 2019 Feb 21.

[48] Zhu X, Ghahramani Z, Lafferty JD. Semi-supervised learning using gaussian fields and harmonic functions. In Proceedings of the 20th International conference on Machine learning (ICML-03) 2003 (pp. 912-919).

[49] Zou Y, Yu Z, Vijaya Kumar BV, Wang J. Unsupervised domain adaptation for semantic segmentation via class-balanced self-training. In Proceedings of the European Conference on Computer Vision (ECCV) 2018 (pp. 289-305).

[50] Zou Y, Yu Z, Liu X, Kumar BV, Wang J. Confidence Regularized Self-Training. In Proceedings of the IEEE International Conference on Computer Vision (ICCV) 2019