

Montreal, July 4–8

DLMI2022

Weakly Supervised CNN Segmentation:
Models and Optimization

Ismail Ben Ayed
Jose Dolz

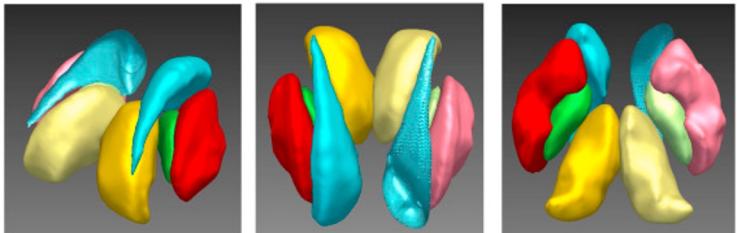
Why weak supervision is interesting?

Deep CNNs are dominating computer vision

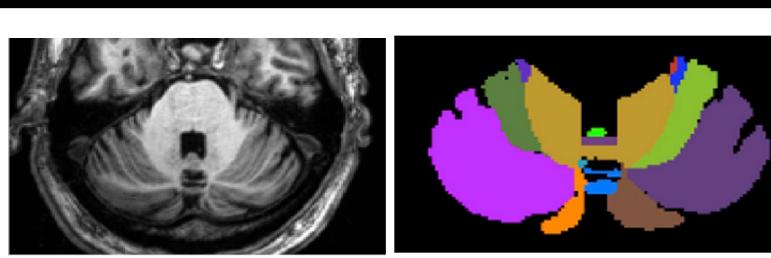
e.g., semantic segmentation



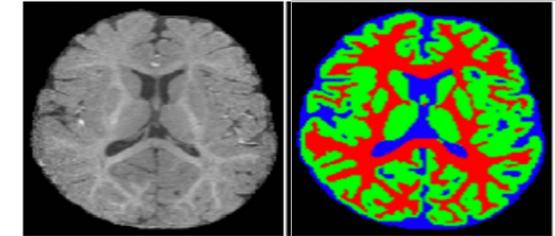
... and medical image analysis



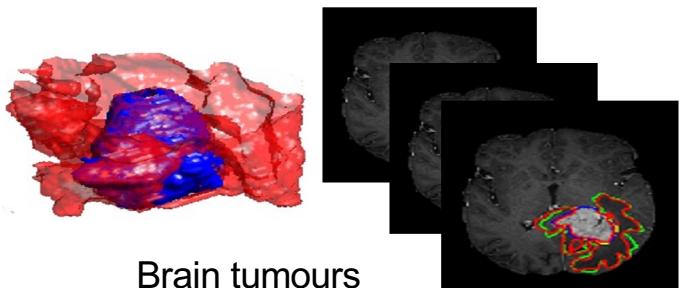
Subcortical structures
(Dolz et al., Neuroimage 2018)



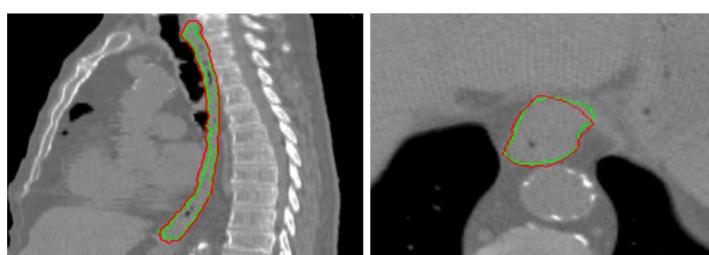
Cerebellum parcellation
(Carass et al., Neuroimage 2018)



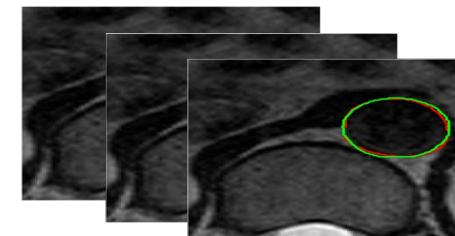
Brain tissues (6-month infant)
(Li et al., TMI 2019)



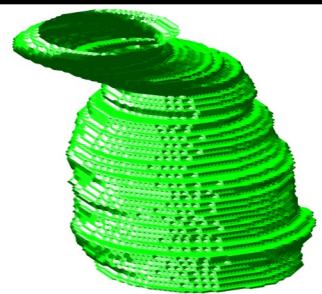
Brain tumours
(Njeh et al., CMIG 2015)



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



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

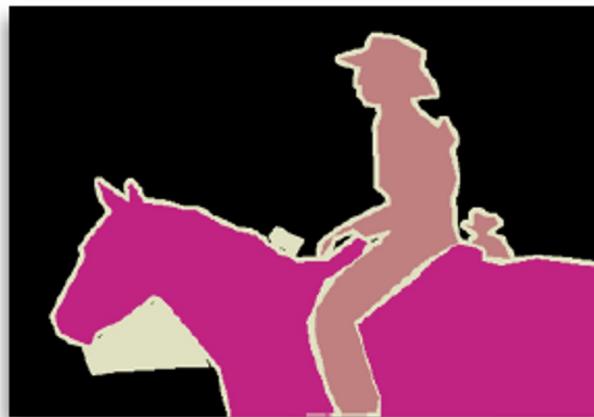


But, massive and dense annotations are not always available

Full supervision



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



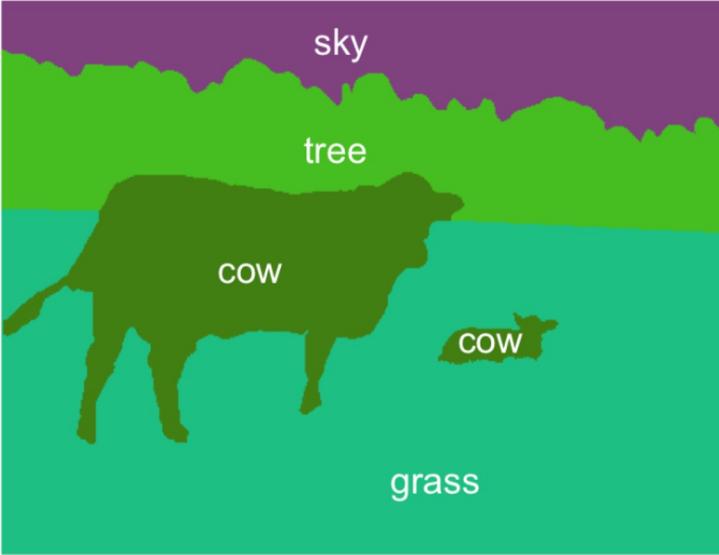
Weak supervision
(e.g., image-level tags)



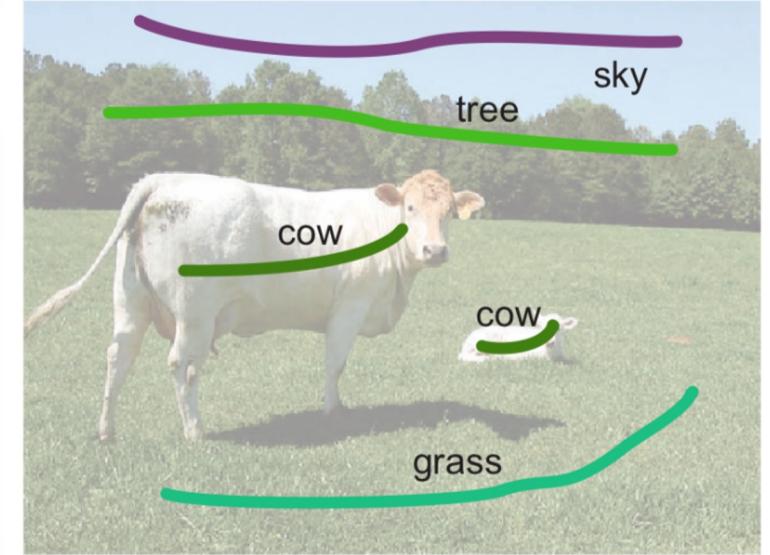
- 1s per label per image
- Scalable for large numbers of labels

person
horse
background

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

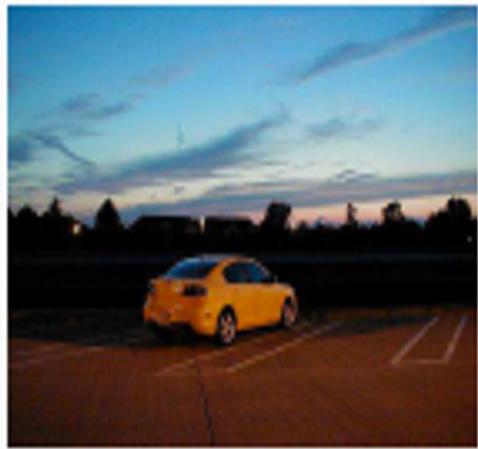


Full annotations



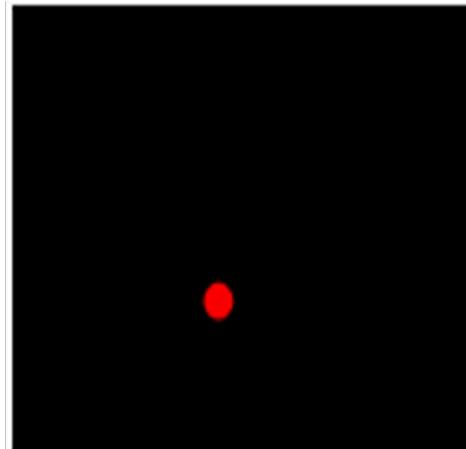
Semi-supervised

Forms of semi/weak supervision: Examples in segmentation

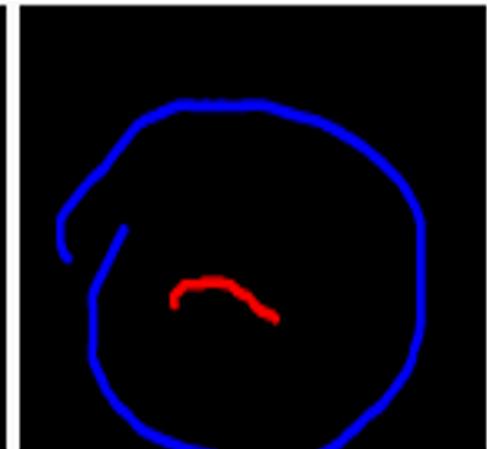


Car
Parking
Sky
No person

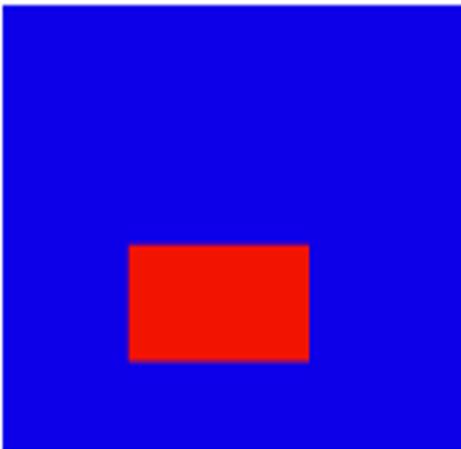
Image tags



points

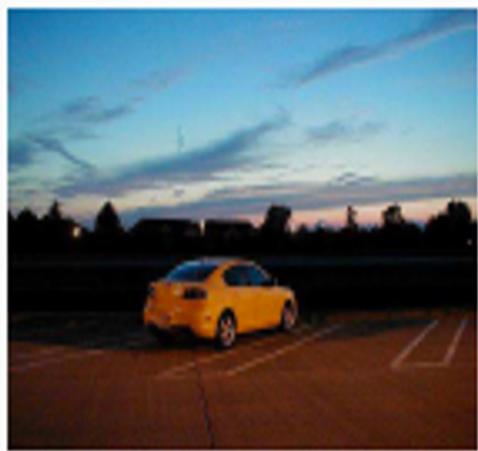


scribbles



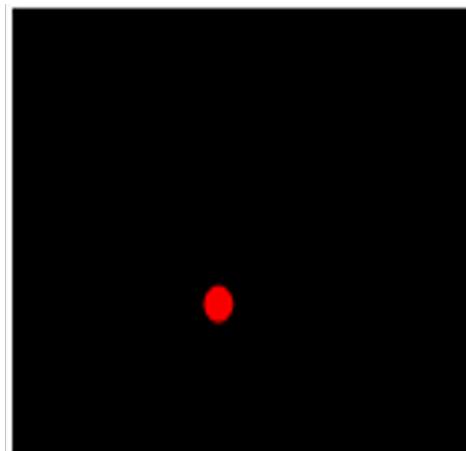
boxes

Forms of semi/weak supervision: Examples in segmentation

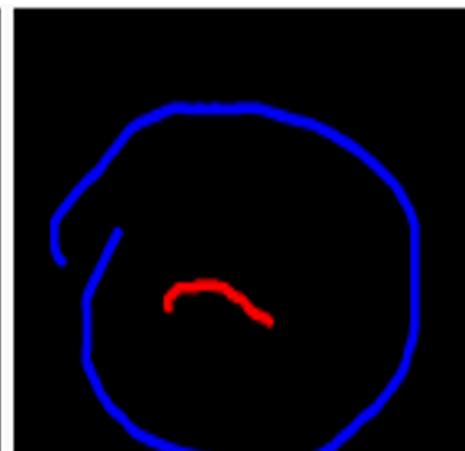


Car
Parking
Sky
No person

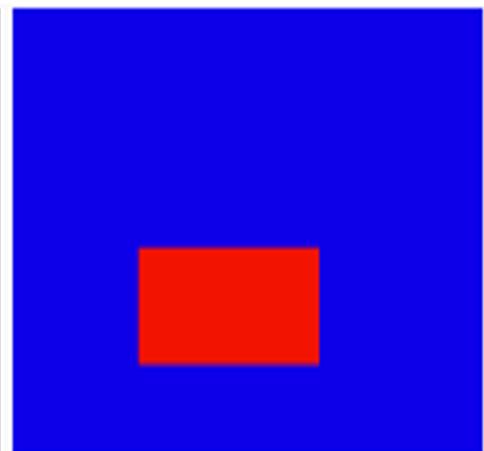
Image tags



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]

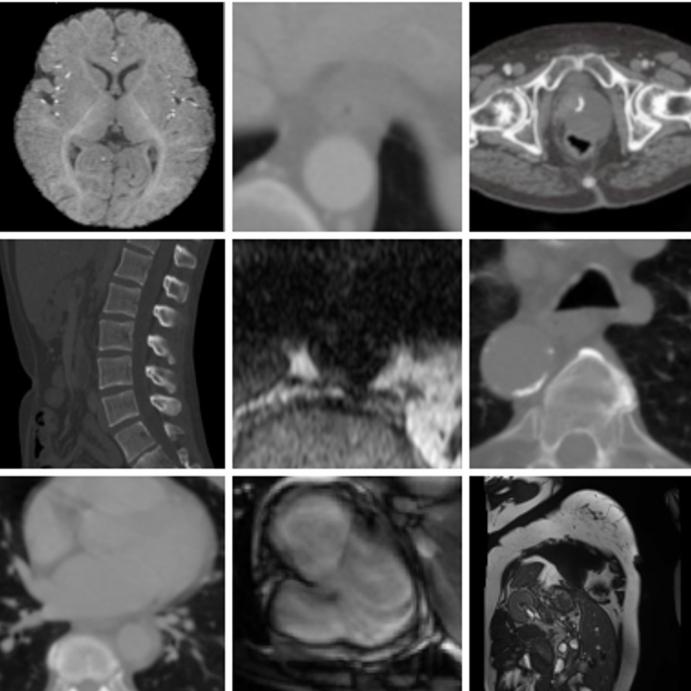
[Rajchl et al., TMI 2017]
[Bai et al., MICCAI 2017]
[Kervadec et al., Media]

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?

Select all images with
esophagus
Click verify once there are none left.



VERIFY

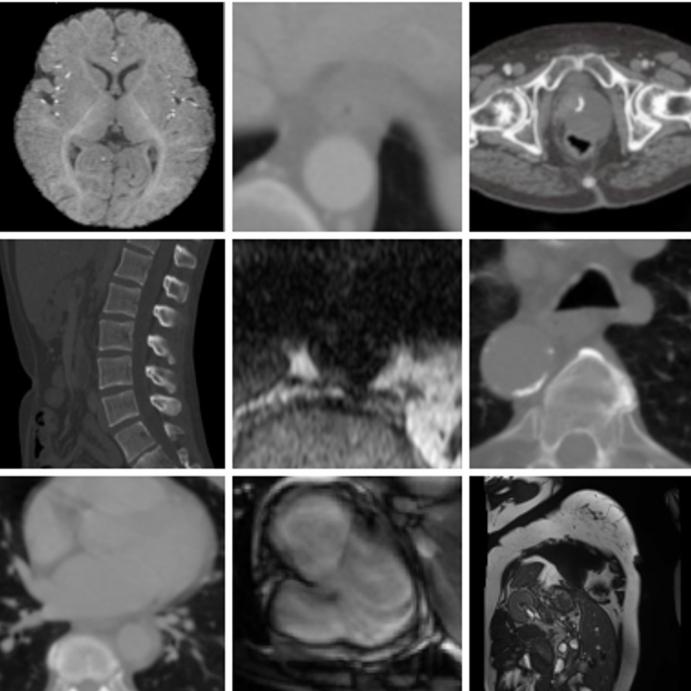
9

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?

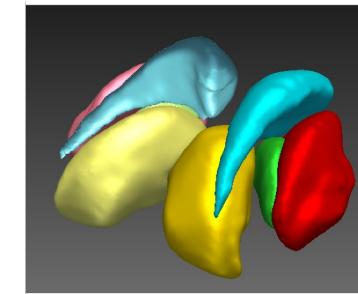
Select all images with
esophagus
Click verify once there are none left.



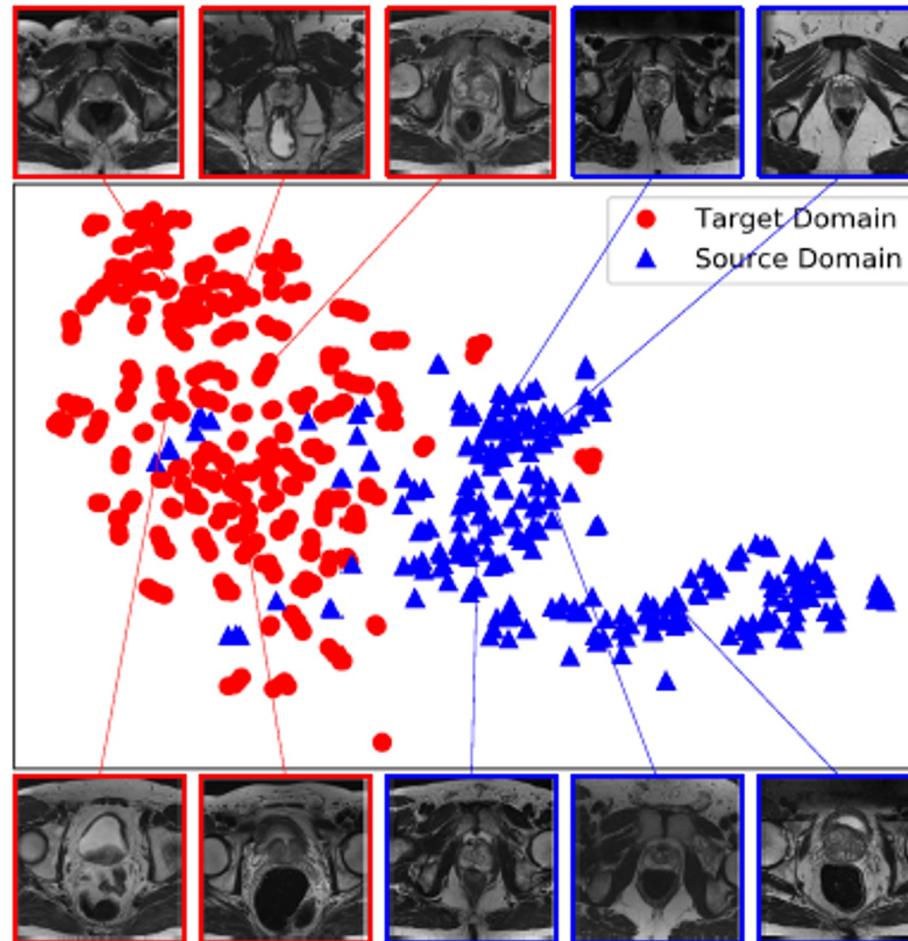
VERIFY

Icons: C, Headphones, i

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

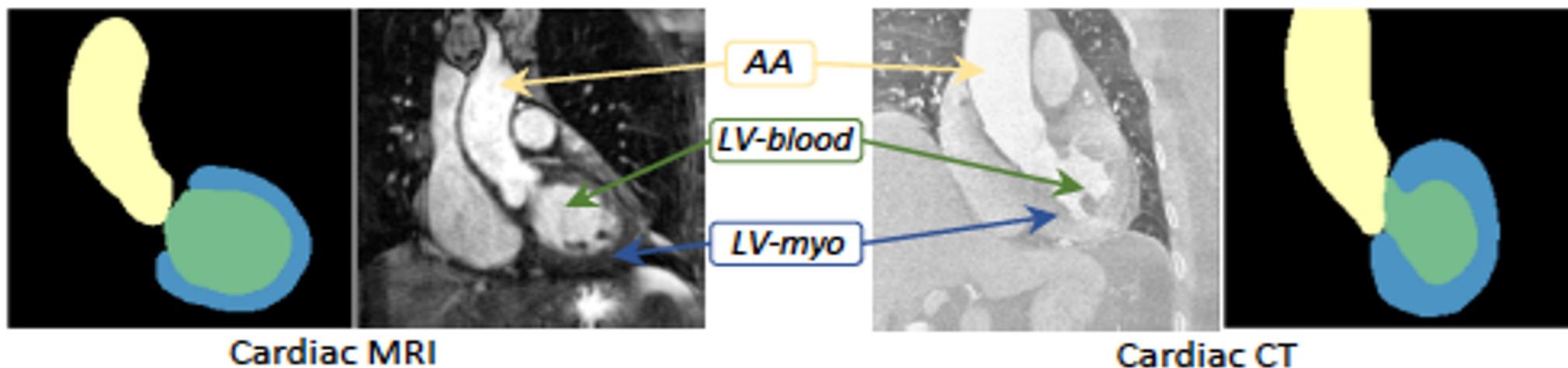


Domain shifts make things worse (even with full annotations in one domain)



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

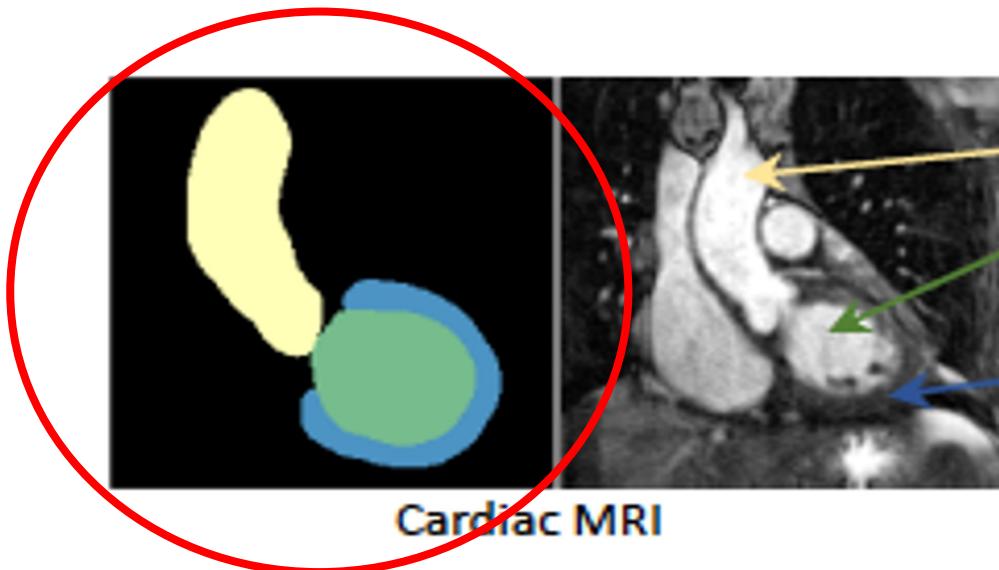
Domain shifts: within and across modalities



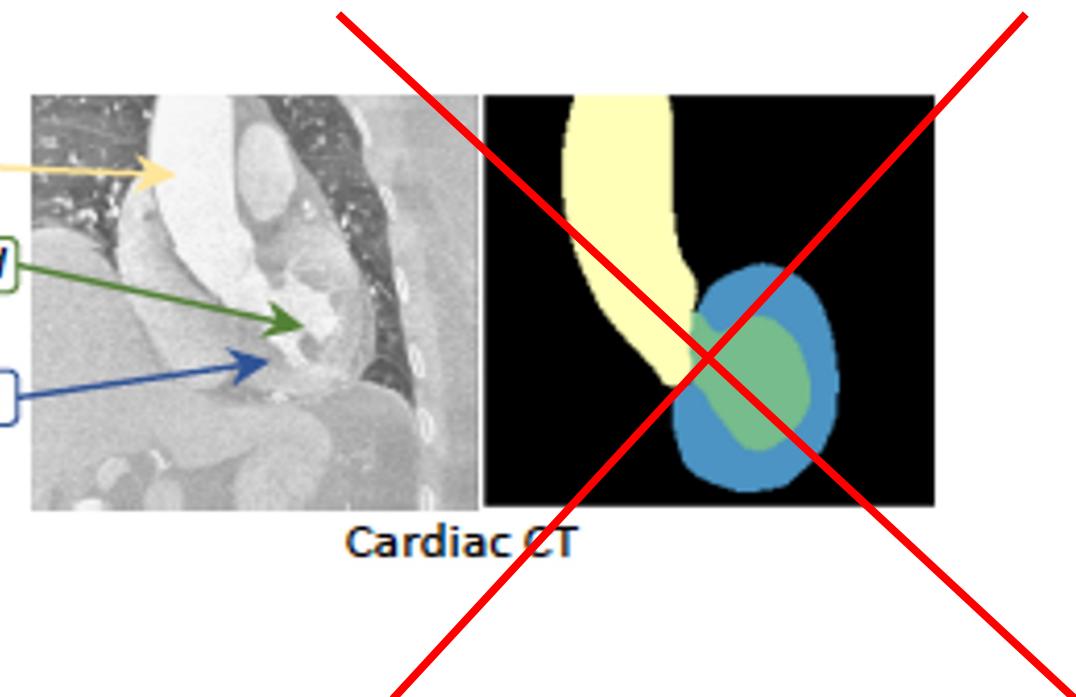
[Images from Dou et al., PnP-AdaNet: Plug-and-play adversarial domain adaptation network with a benchmark at cross-modality cardiac segmentation ArXiv 2018]

Unsupervised domain adaptation

We have labels for
the source domain

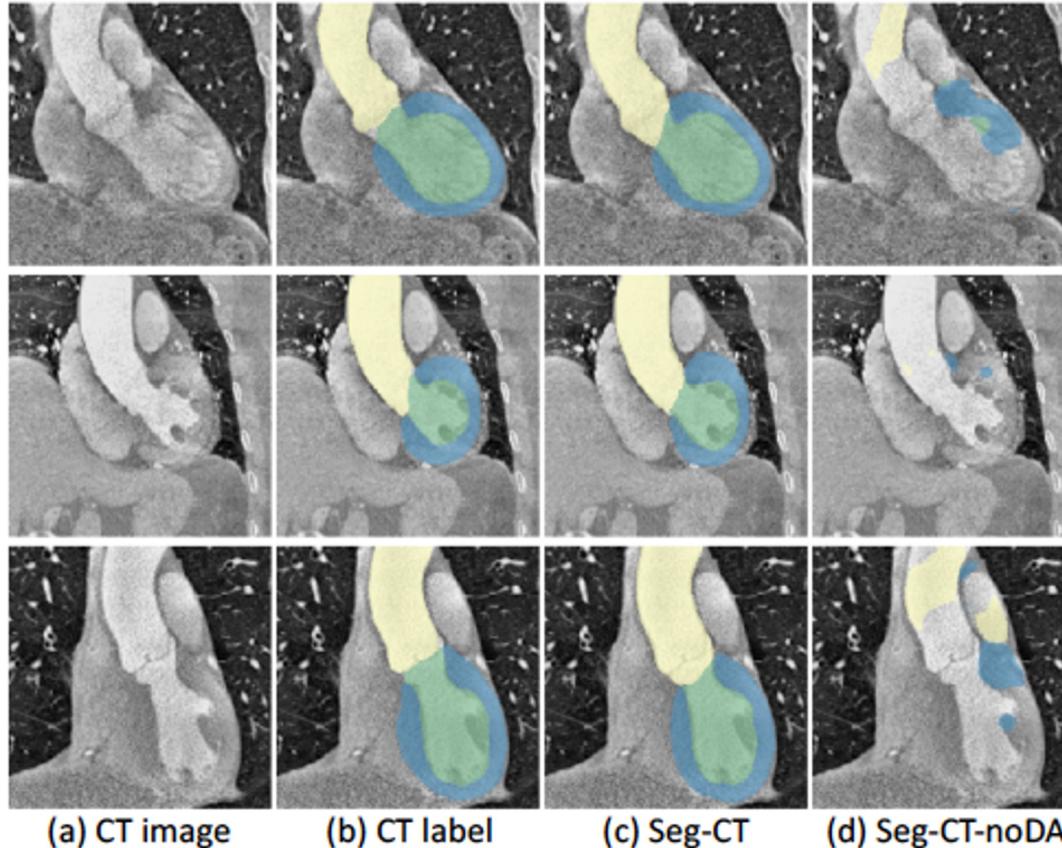


No labels for the target



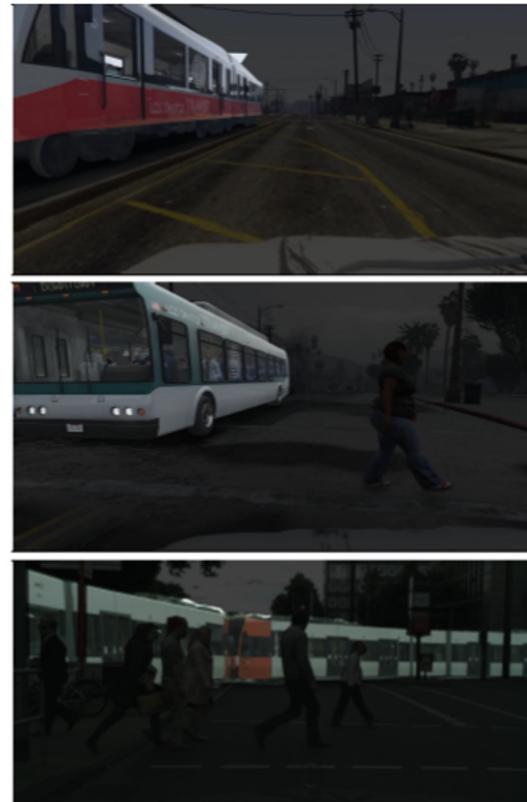
[Images from Dou et al., PnP-AdaNet: Plug-and-play adversarial domain adaptation network with a benchmark at cross-modality cardiac segmentation ArXiv 2018]

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

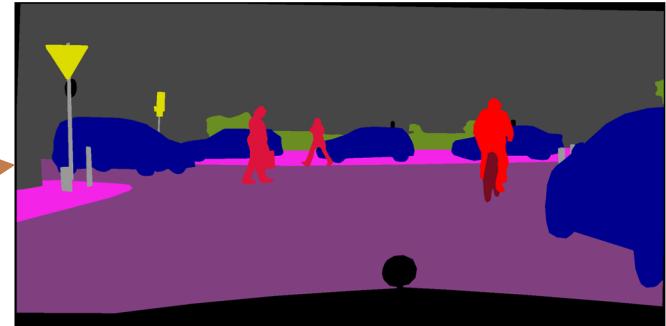
“train”
Cityscapes

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

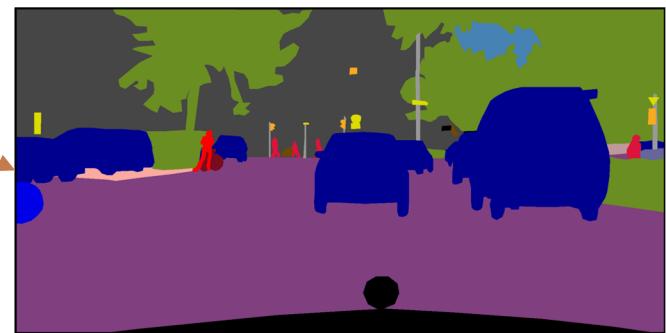
A lot of interest in vision as well:
Domain shifts are *everywhere* BUT we cannot label *everywhere*



road	sidewalk	building	wall	fence	pole	traffic light	traffic sign	vegetation	terrain
sky	person	rider	car	truck	bus	train	motorcycle	bicycle	unlabeled

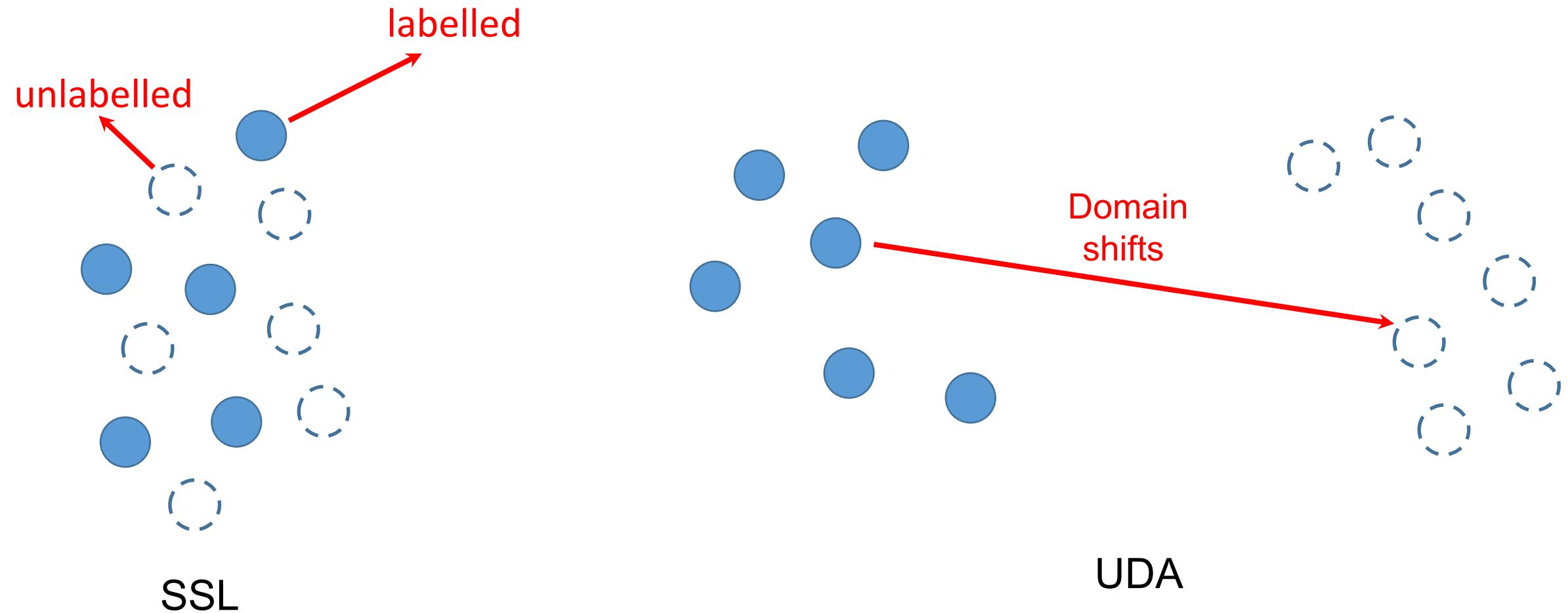


Frankfurt

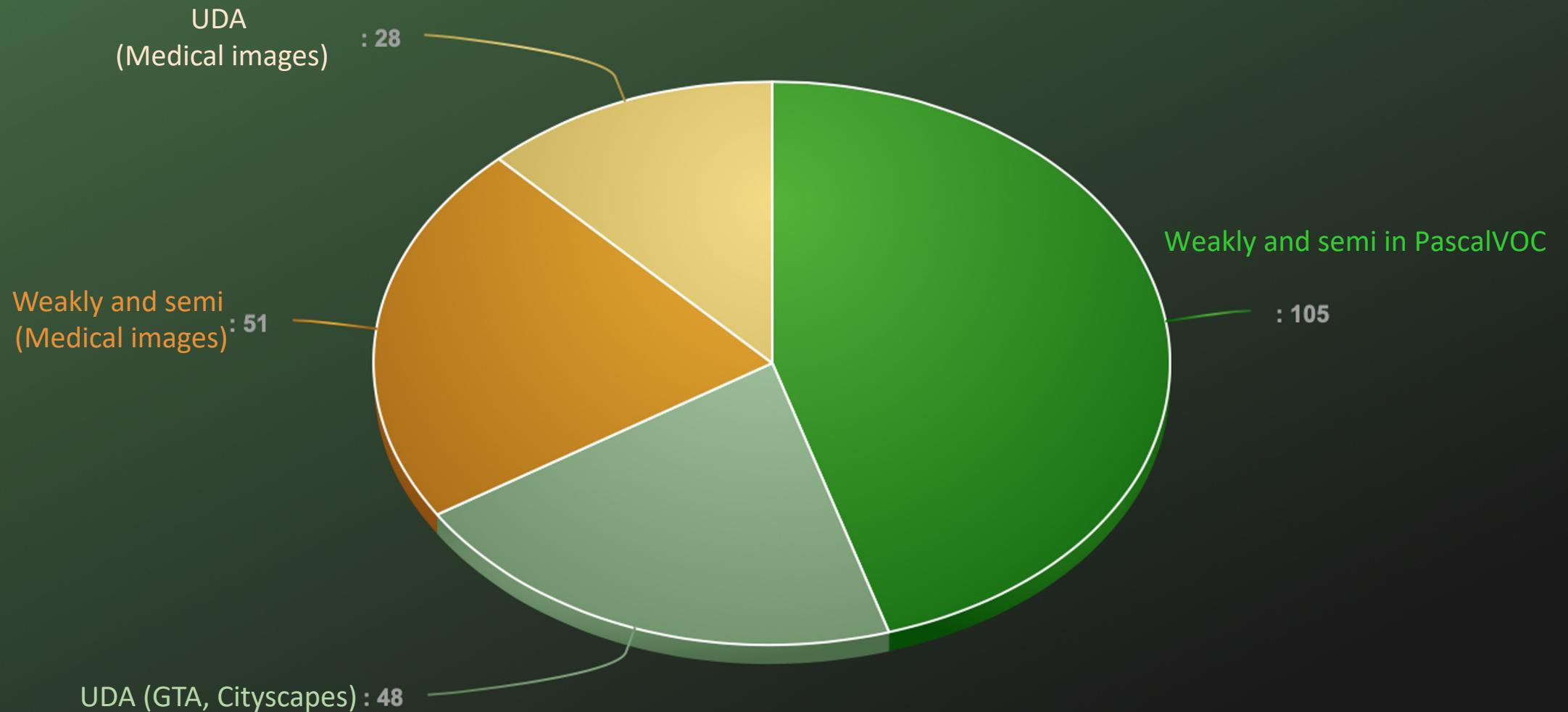


Zurich

$UDA = SSL + \text{domain shift}$



Surprisingly in medical image analysis, we are behind



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

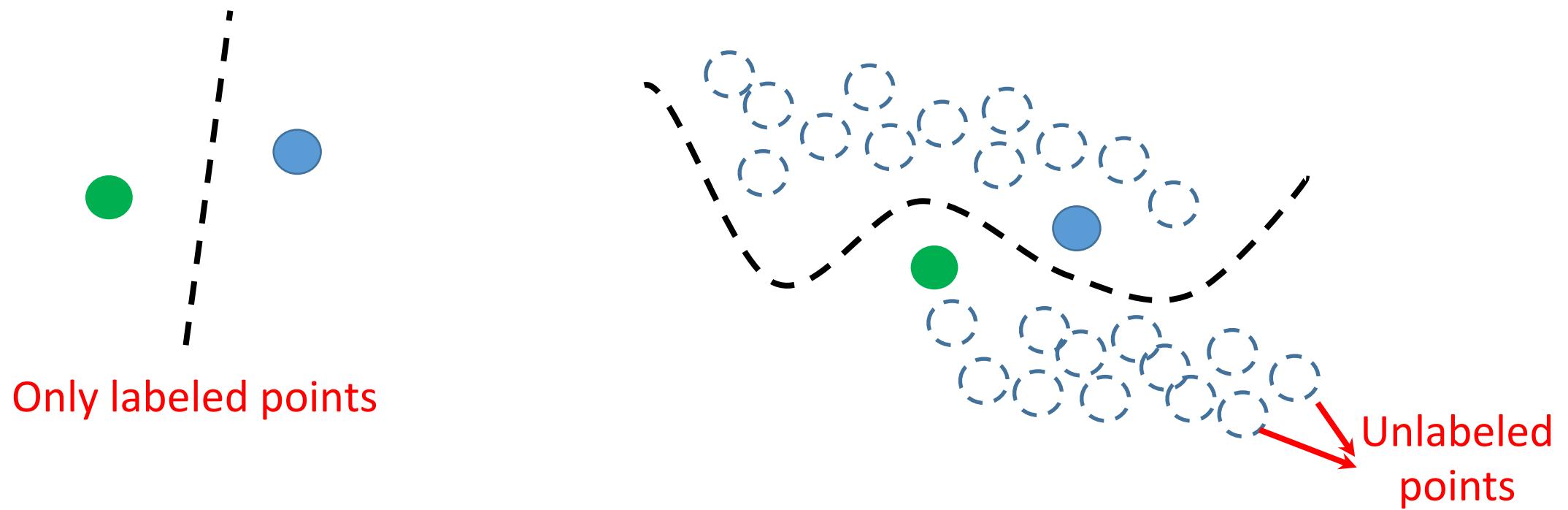
- Structure-driven priors: *Regularization (Part 1)*
- Knowledge- and data-driven priors (Part 2)

Part 1

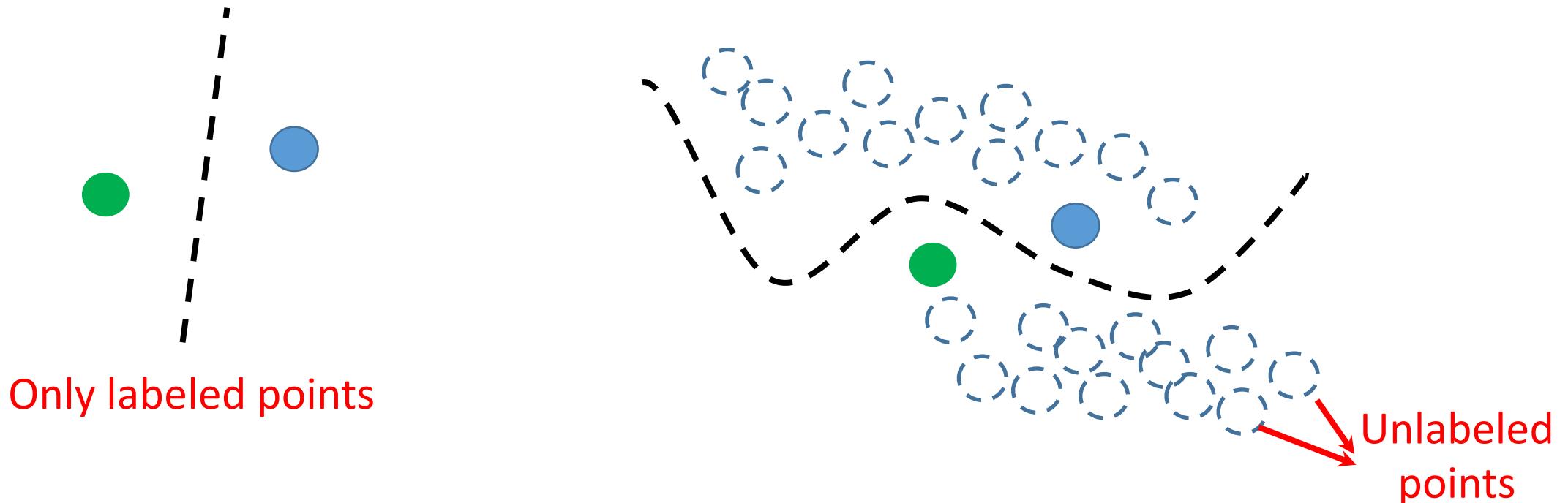
Regularization

Laplacian (and CRFs)

Semi-supervised learning (general form)

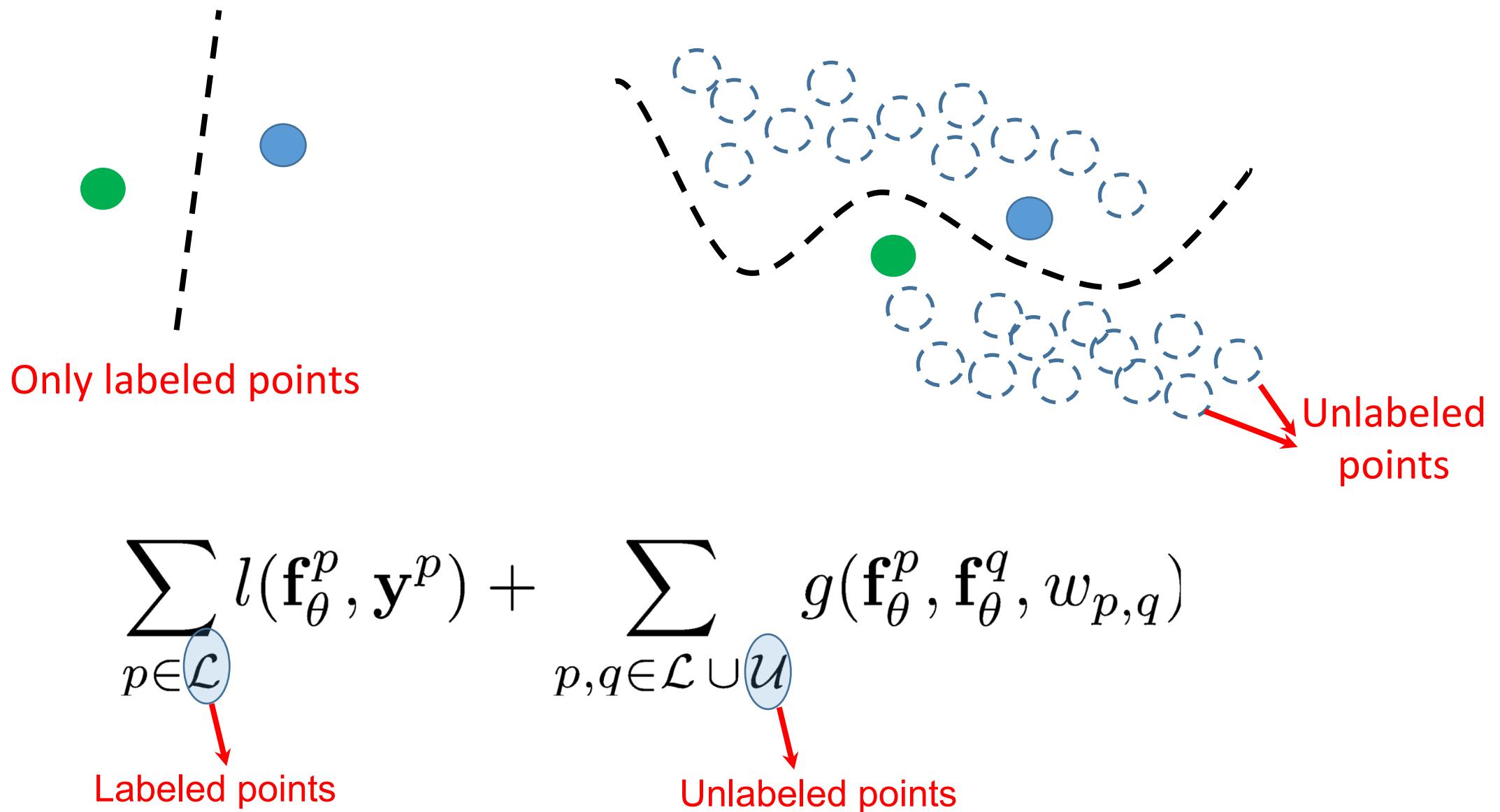


Semi-supervised learning (general form)



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

Semi-supervised learning (general form)



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

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

Labels
(binary simplex vectors)

The diagram illustrates a neural network architecture. It starts with three yellow input nodes at the bottom left. These nodes are fully connected to four blue hidden nodes in the middle row. These hidden nodes are in turn fully connected to four green output nodes on the right. The output nodes represent simplex probability vectors, specifically softmax outputs. To the right of the output layer, a vertical brace groups the four output nodes together, and a bracket groups this entire set with the expression $\mathbf{f}_{\theta}^p = \mathbf{s}_{\theta}^p \in [0, 1]^K$. Above the output layer, the individual output nodes are labeled $f_{\theta}^{p,1}, f_{\theta}^{p,2}, \dots, f_{\theta}^{p,C}$.

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

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

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

**Labels
(binary simplex vectors)**

e.g.: Laplacian

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

The diagram illustrates the general form of a semi-supervised learning loss function. It consists of two main parts: a supervised loss term and an unlabeled loss term. The supervised loss term is a sum over labeled points p in set L, where each term is the loss function l applied to the feature vector f_theta^p and its corresponding label y^p. The unlabeled loss term is a sum over pairs of points p and q in the union of sets L and U, where each term is a function g applied to the feature vectors f_theta^p and f_theta^q, and a weight w_{p,q}. Red arrows point from the labels y^p and the weight w_{p,q} to their respective terms in the equation. Red text provides examples for the loss function l (cross-entropy) and the function g (Laplacian), along with descriptions of the labels and weights.

Semi-supervised learning (general form)

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

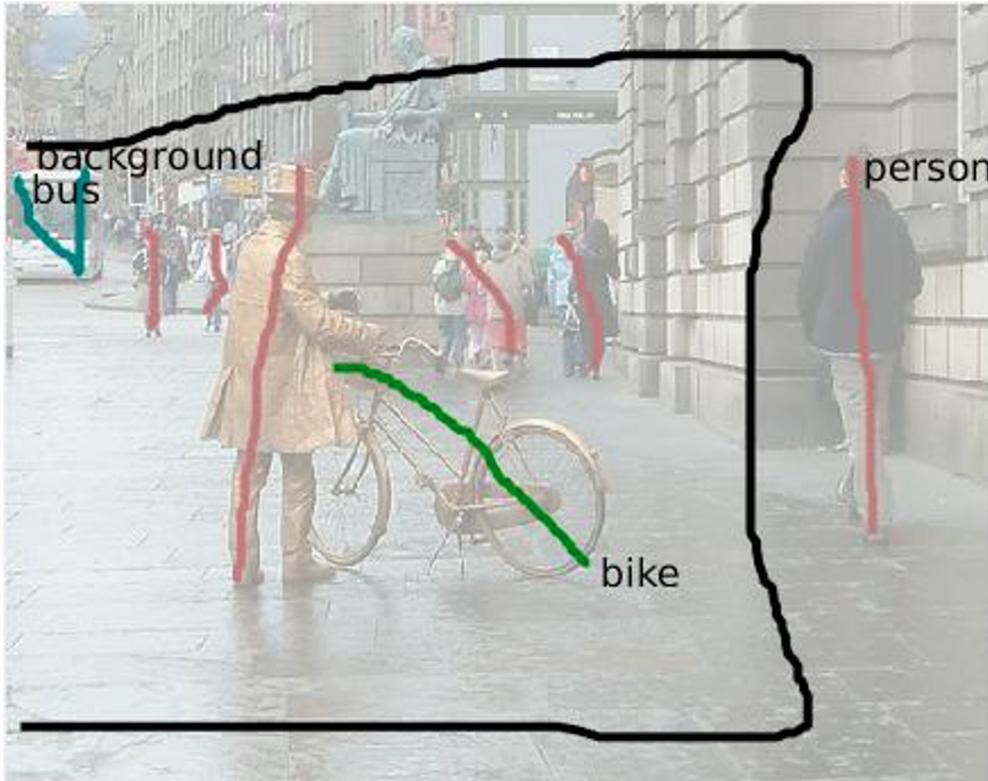
Diagram illustrating the general form of semi-supervised learning loss:

- The first term, $\sum_{p \in \mathcal{L}} l(\mathbf{f}_\theta^p, \mathbf{y}^p)$, represents the supervised loss for labeled examples. It is annotated with "e.g.: cross-entropy" and "e.g.: simplex probability vectors (*softmax outputs of the network*)".
- The second term, $\sum_{p,q \in \mathcal{L} \cup \mathcal{U}} g(\mathbf{f}_\theta^p, \mathbf{f}_\theta^q, w_{p,q})$, represents the unlabeled loss or regularization term. It is enclosed in a light blue oval and annotated with "e.g.: Laplacian" and the formula $w_{p,q} \|\mathbf{f}_\theta^p - \mathbf{f}_\theta^q\|^2$.
- Annotations for the supervised loss term:
 - "Labels (binary simplex vectors)" points to \mathbf{y}^p .
 - " $p \in \mathcal{L}$ " points to the index p in the summation.
 - " $l(\mathbf{f}_\theta^p, \mathbf{y}^p)$ " points to the loss function.

- [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

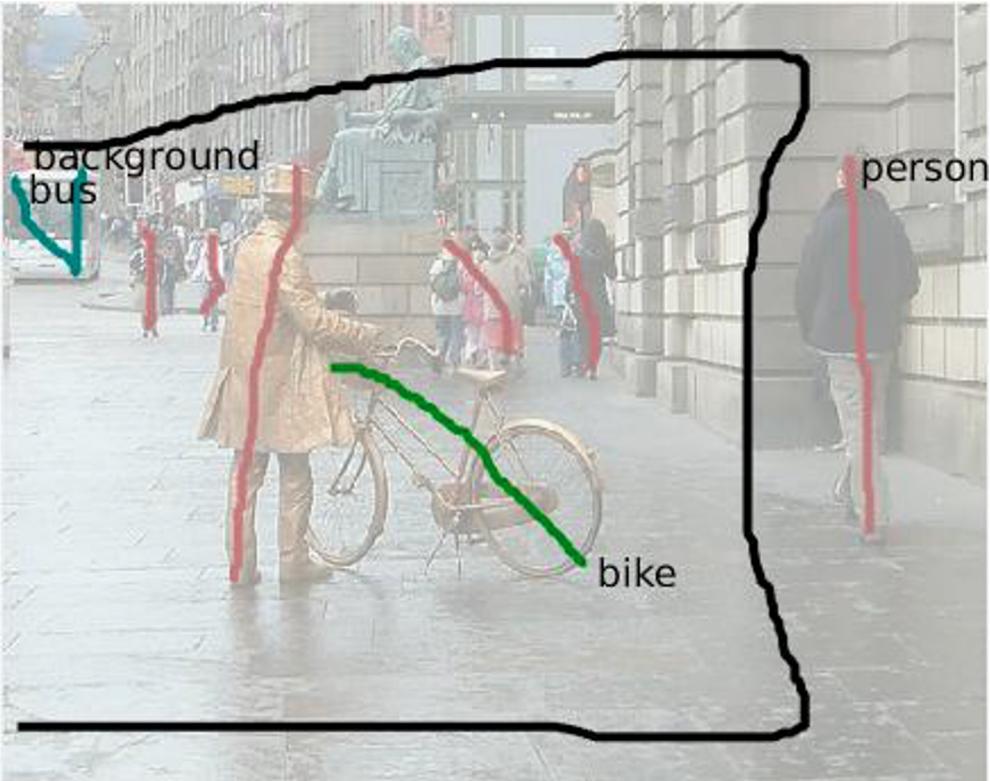
$$\min_{\theta} \sum_{p \in \mathcal{L}} l(\mathbf{y}^p, \mathbf{s}_{\theta}^p) + \sum_{p,q \in \mathcal{L} \cup \mathcal{U}} w_{p,q} \|\mathbf{s}_{\theta}^p - \mathbf{s}_{\theta}^q\|^2$$



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

Semi-supervision loss for segmentation

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



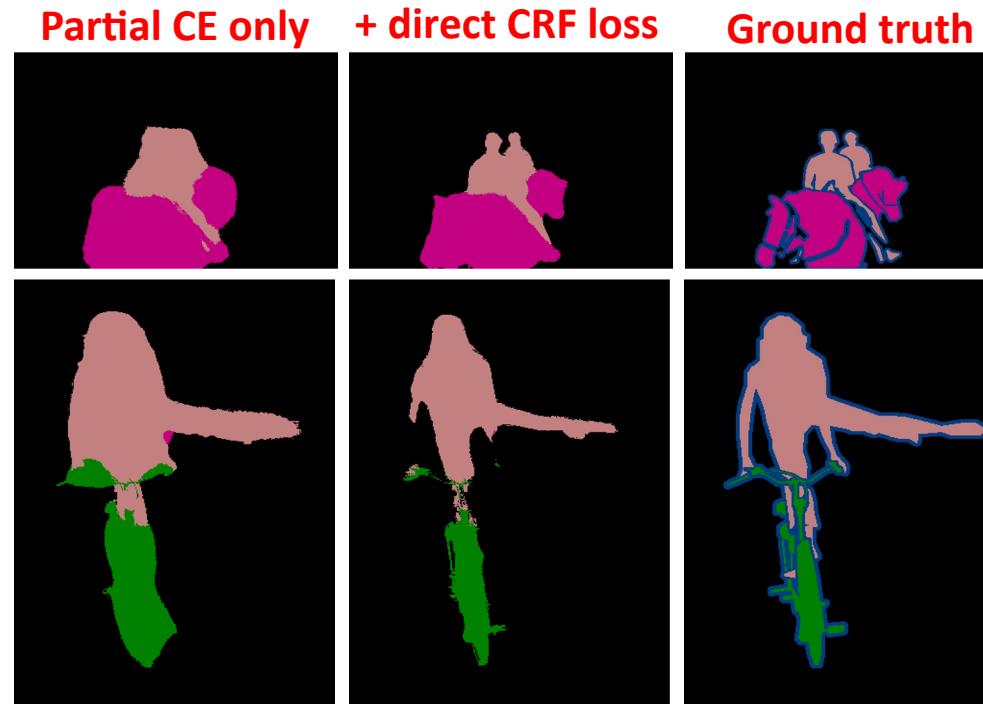
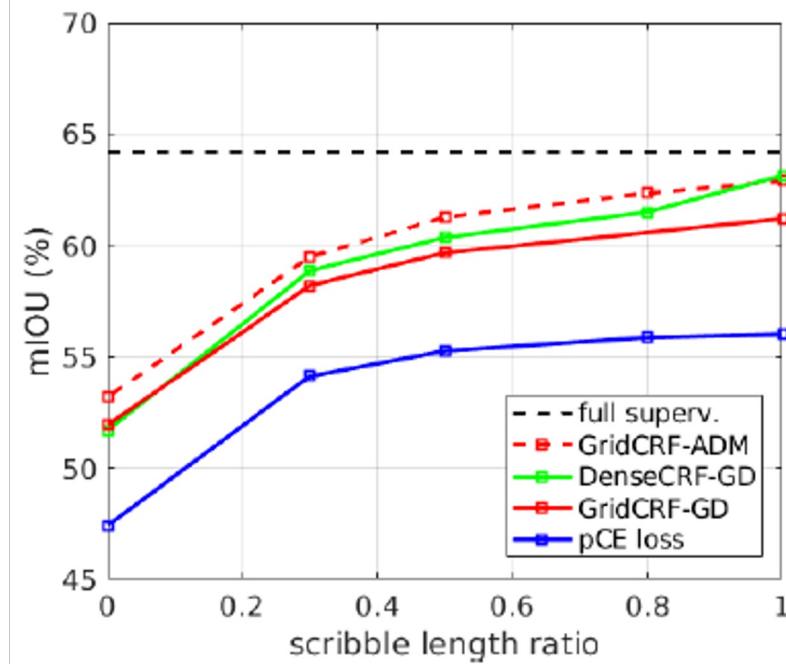
On the vertices of the simplex (binary variables), this is exactly the Potts model in Conditional Random Fields (e.g., Dense CRFs)!

Semi-supervision loss for segmentation

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

↓

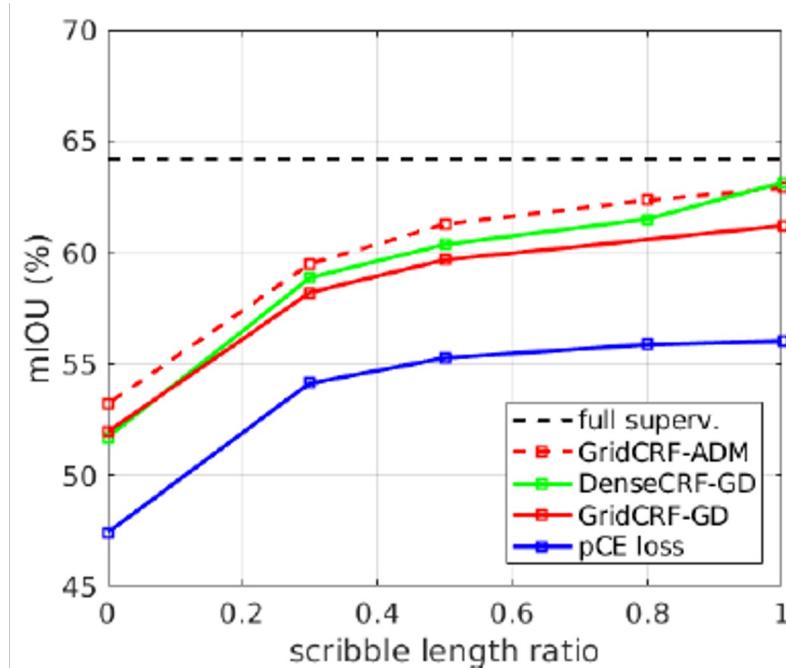
SGD



[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

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



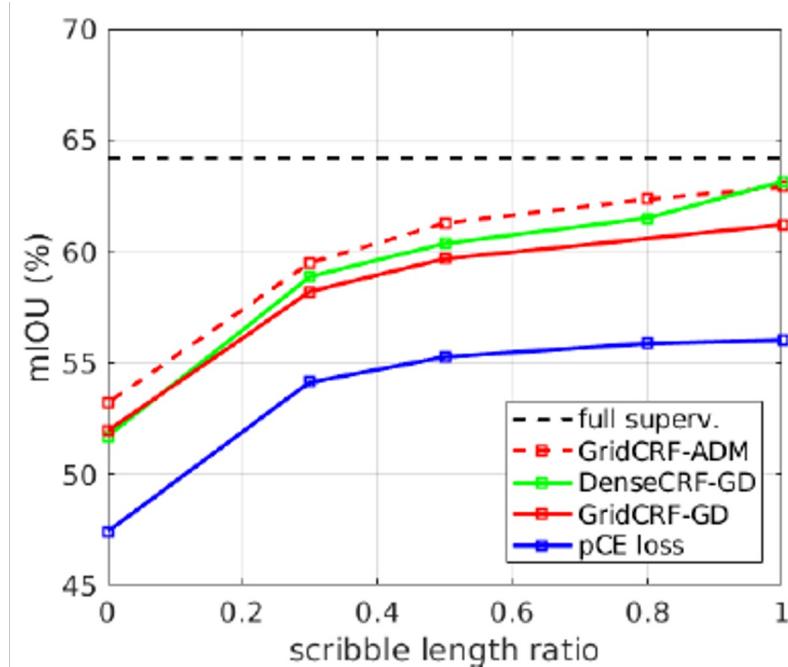
The exciting part in this plot:

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]

Semi-supervision loss for segmentation

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

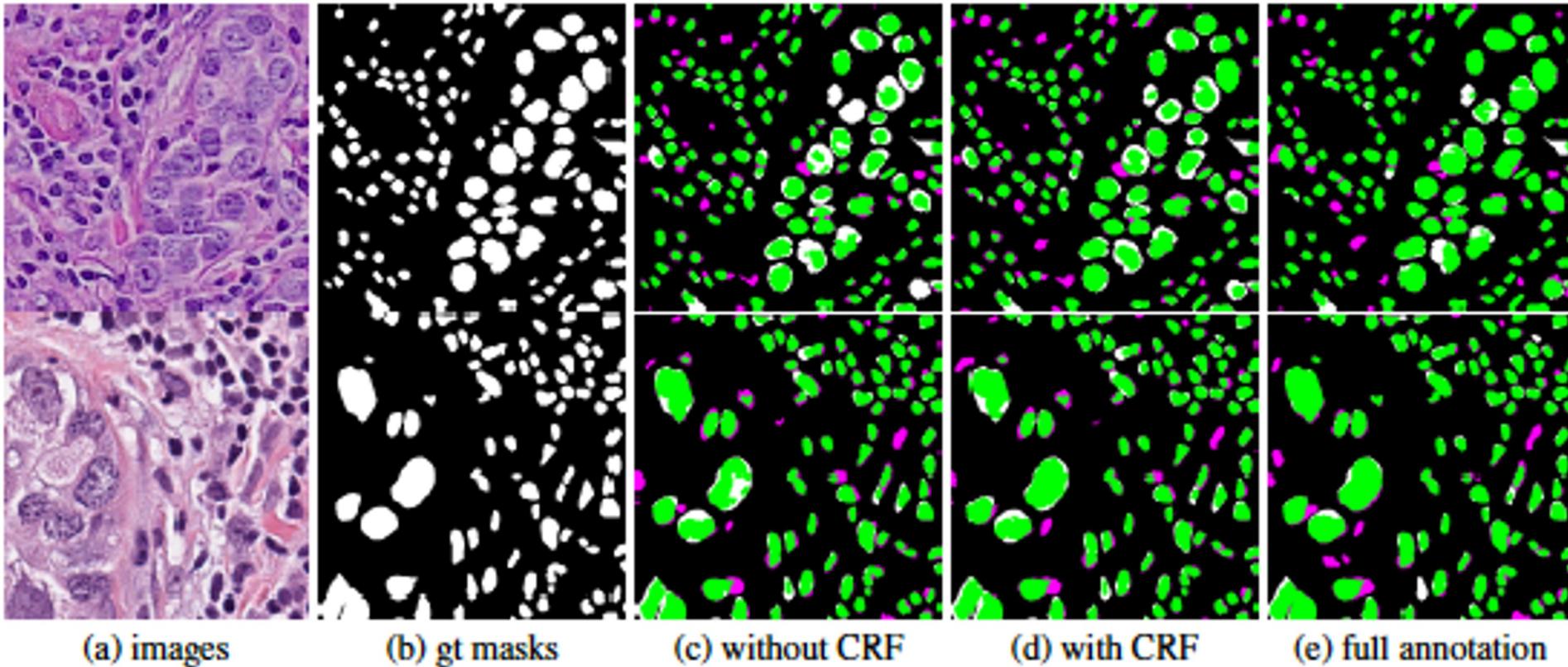


The disturbing part (for those who know classical CRFs):
Dense CRF is not supposed to be better than grid CRF

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

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

entropy

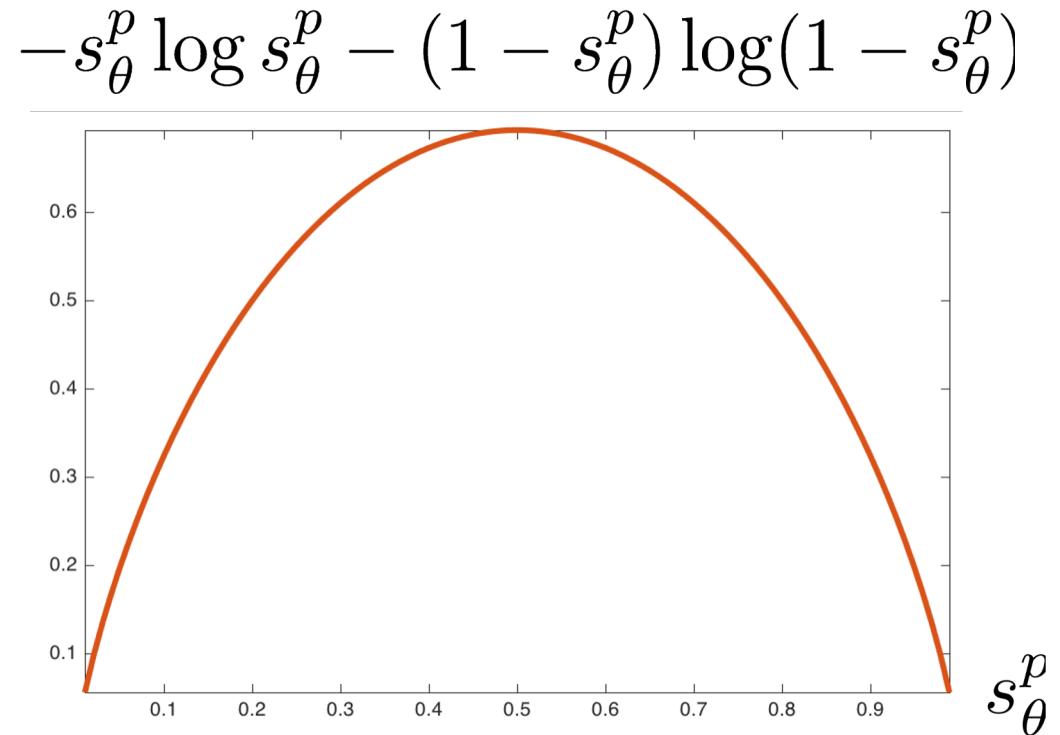
Entropy minimization for SSL

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

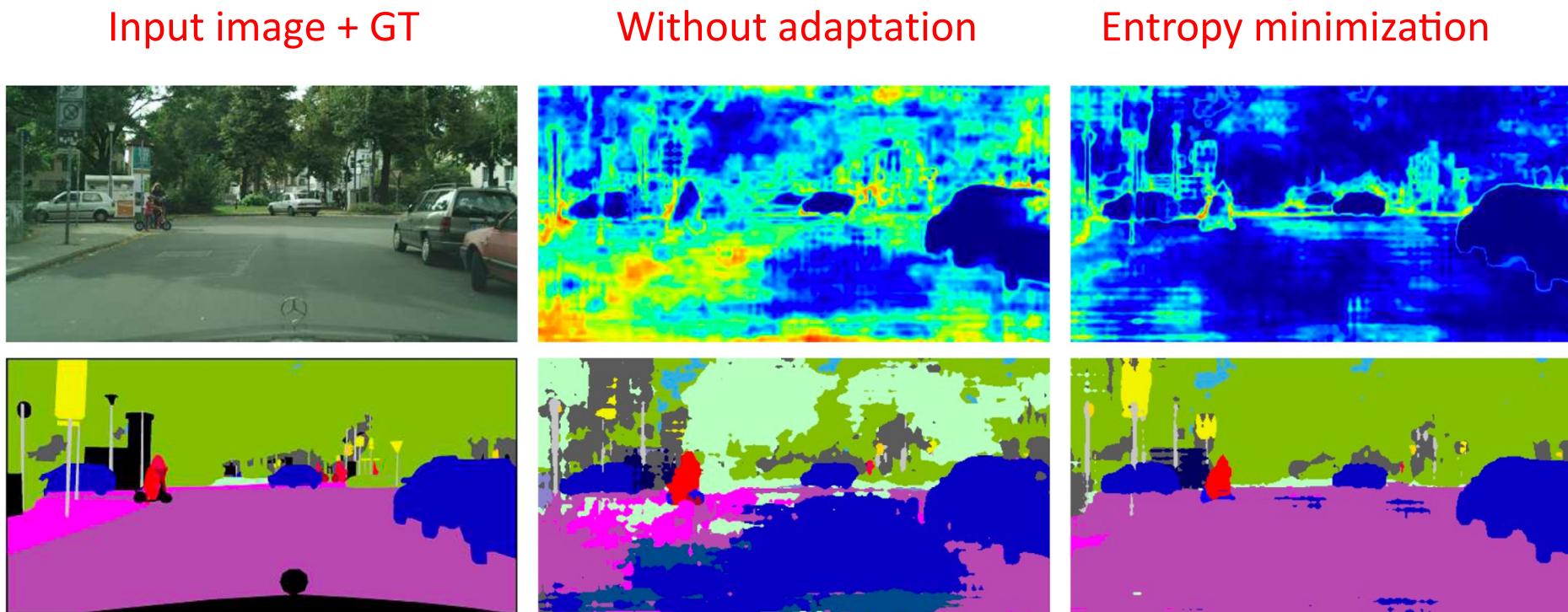
Shannon Entropies: “unsupervised cross-entropies (with unknown labels)”

- 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)

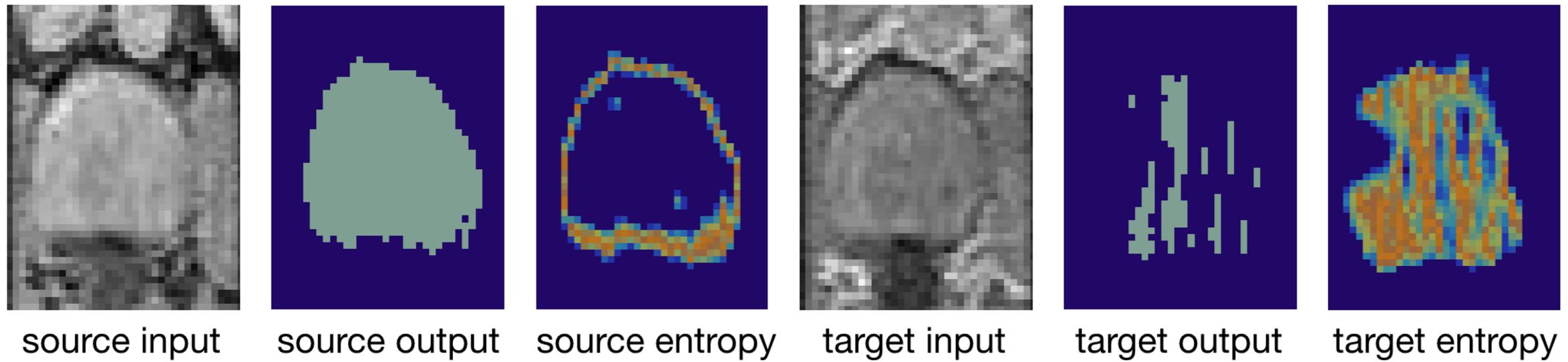


Entropy minimization for UDA

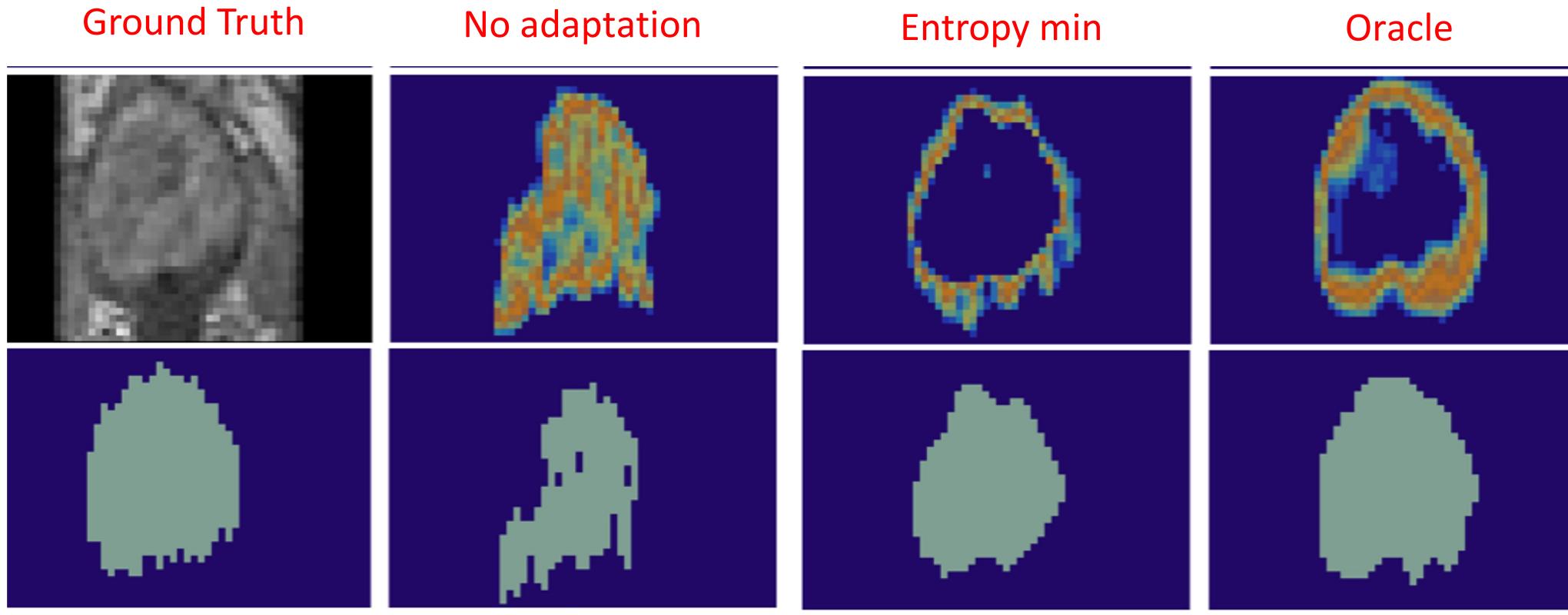


Images from Vu et al., ADVENT: Adversarial Entropy Minimization for Domain Adaptation in Semantic Segmentation, CVPR 2019

Entropy minimization for UDA

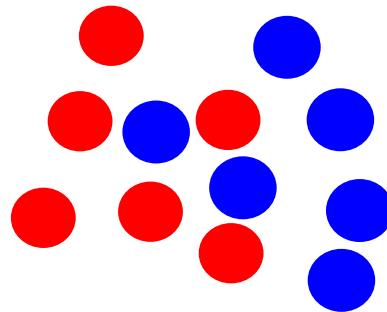


Entropy minimization for UDA

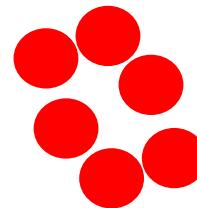


Images from Bateson et al., Source-relaxed domain adaptation for segmentation, MICCAI 2020

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



*High entropy
(low confidence)*



*Low entropy
(high confidence)*

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

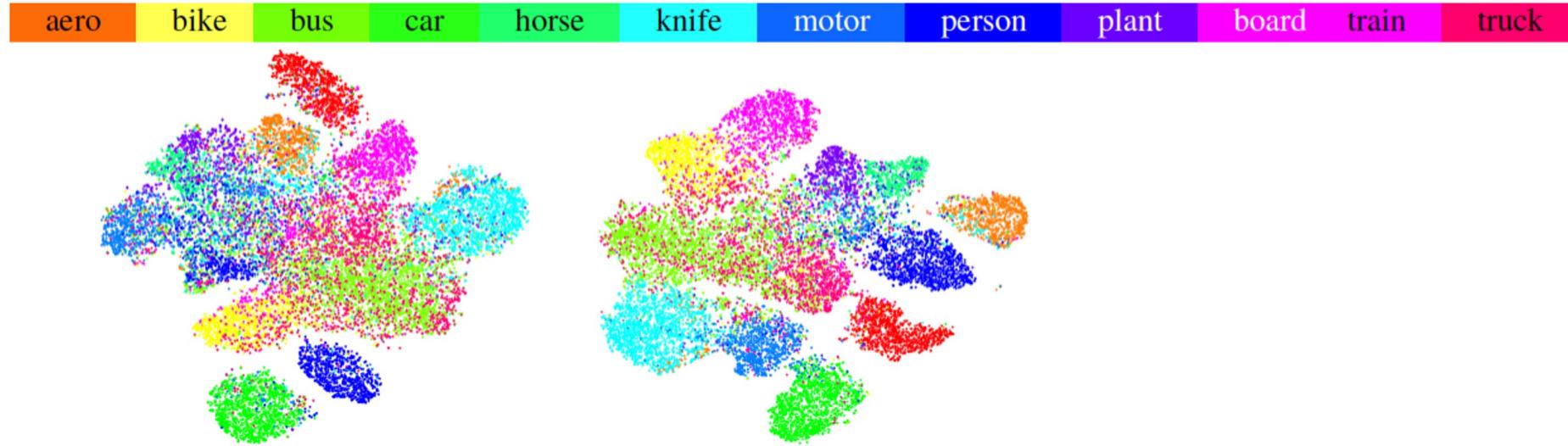
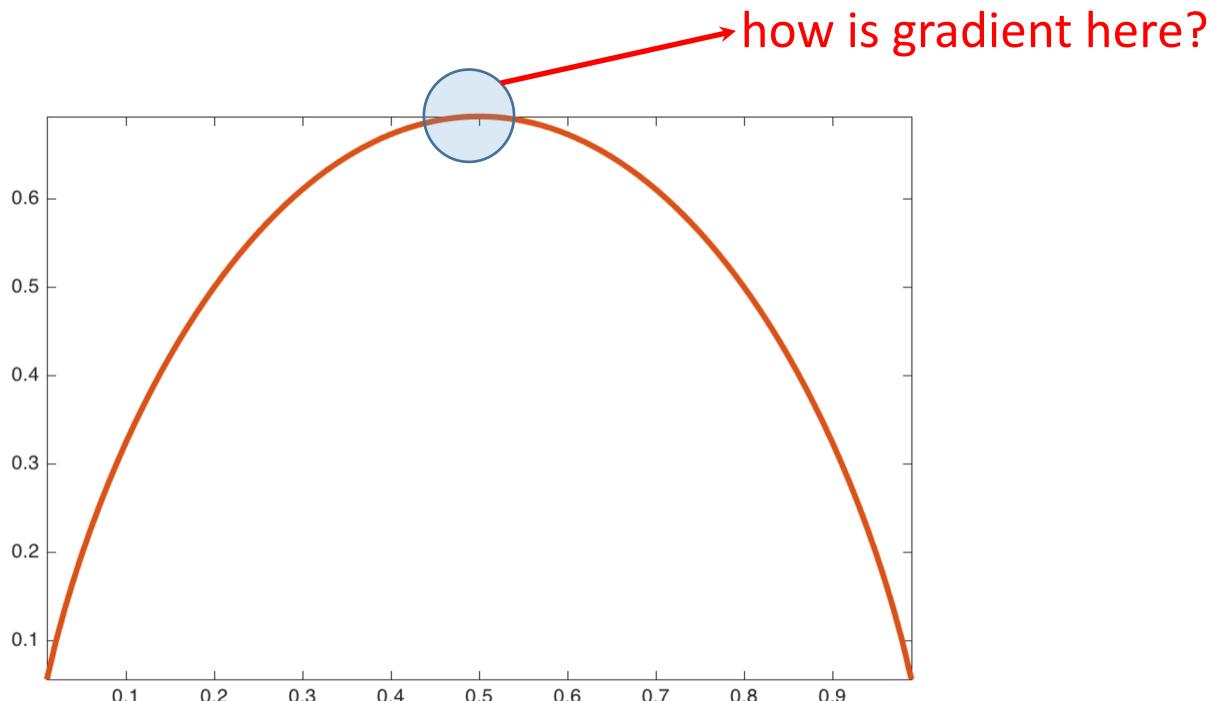
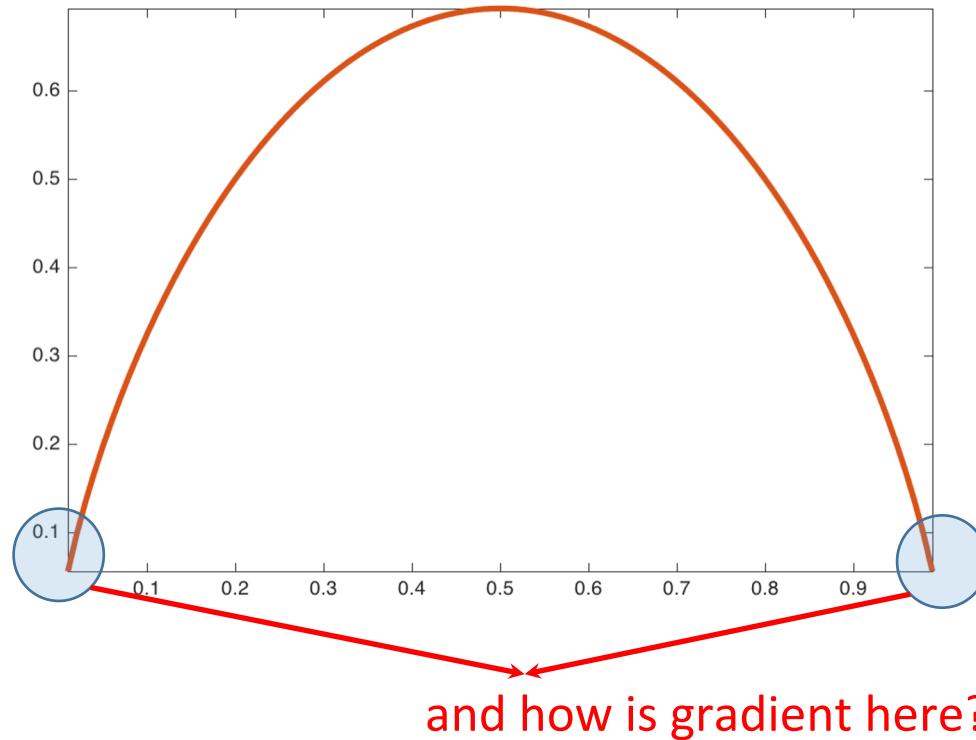


Image classification UDA on VisDA17 data set: Feature visualization for source model (left) and *min-entropy (lower bound on Shannon)* minimization (right) - equivalent to self training (clarified in the next slide)

Difficulty of optimizing entropy

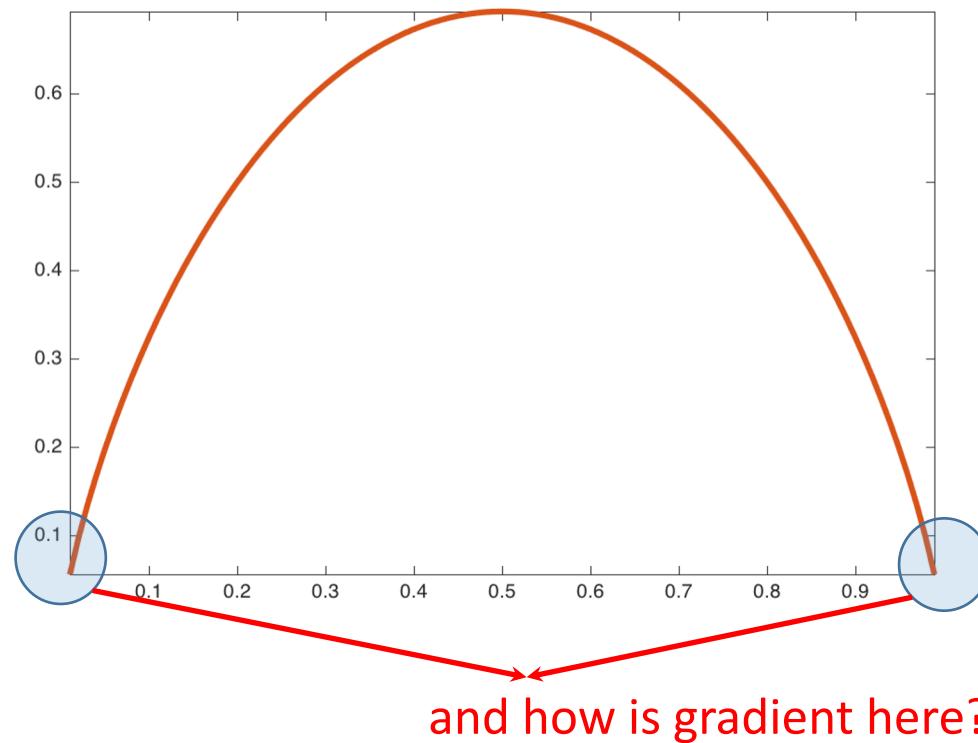


Difficulty of optimizing entropy

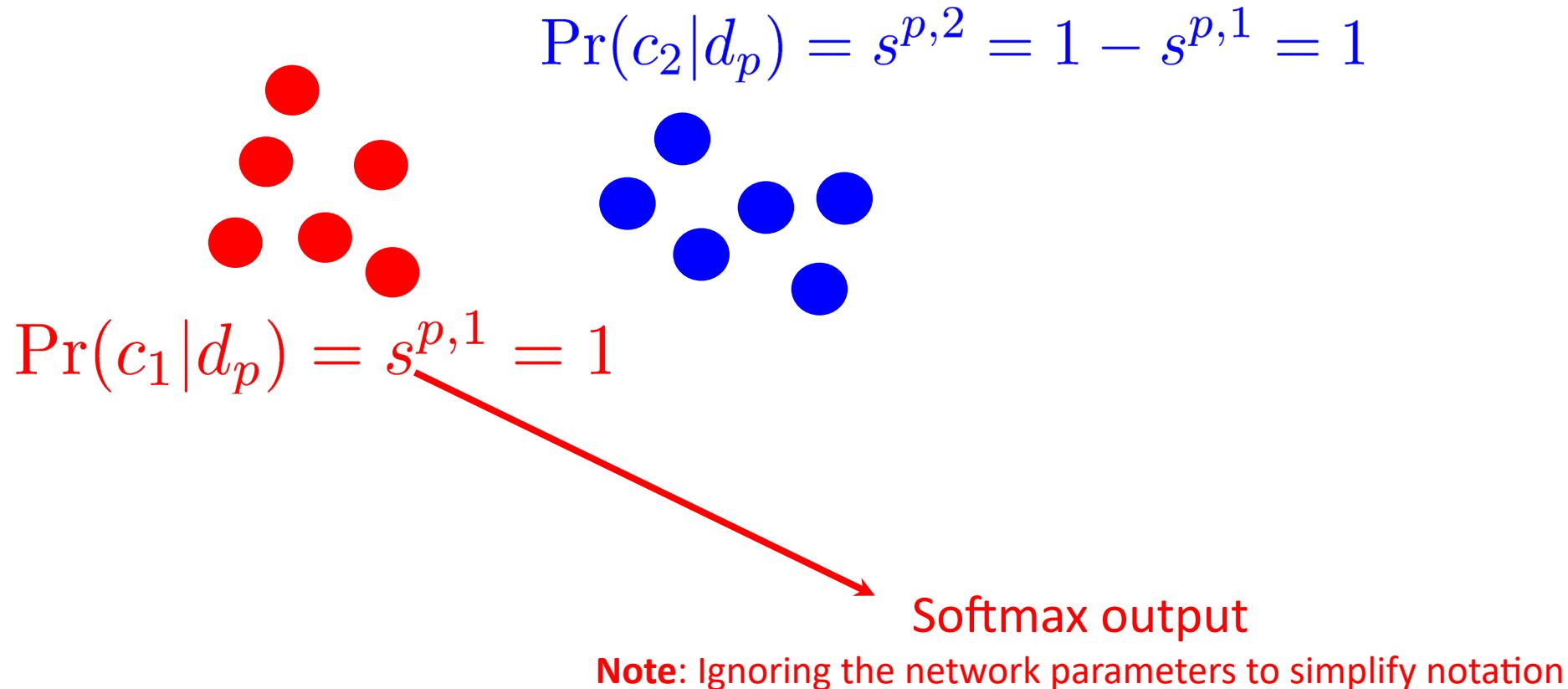


Difficulty of optimizing entropy

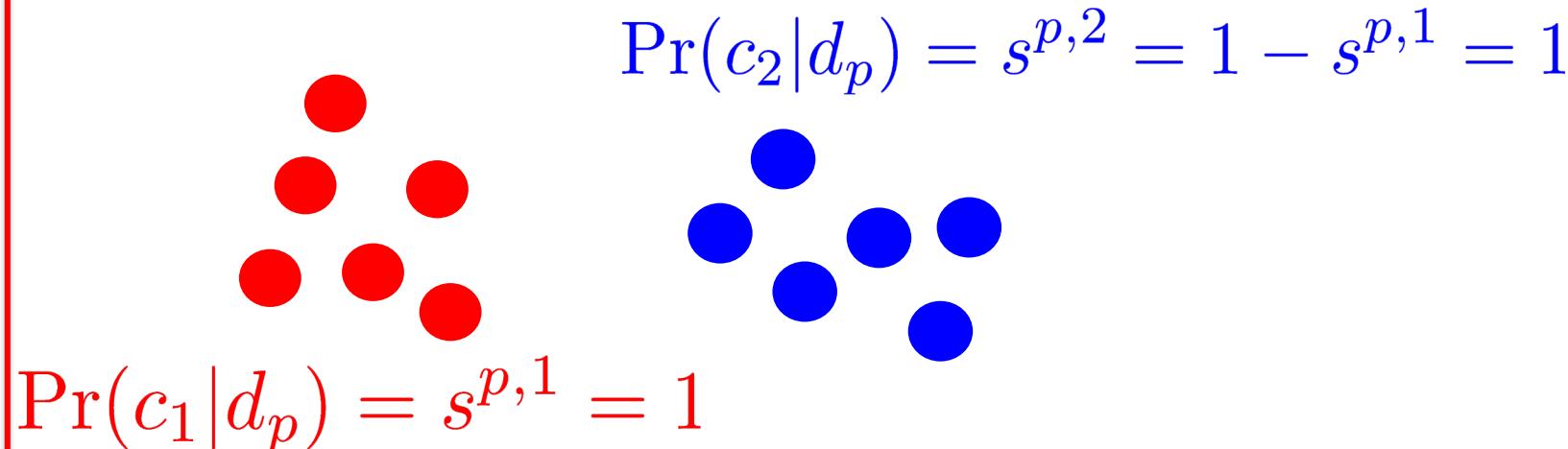
Typically we add other cues to facilitate optimization and avoid trivial solutions
(more on this later)



Avoiding the trivial solutions of entropy minimization

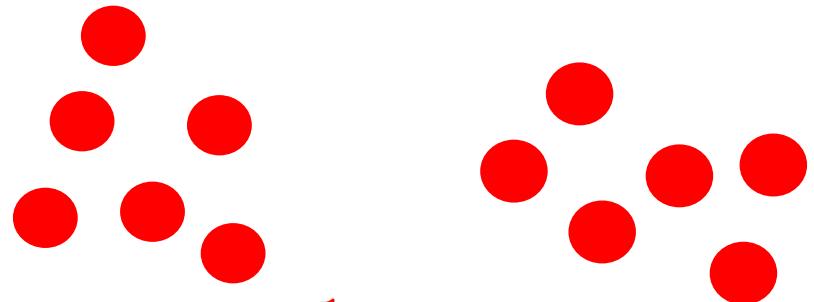


Avoiding the trivial solutions of entropy minimization

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

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

Min entropy
(max confidence)

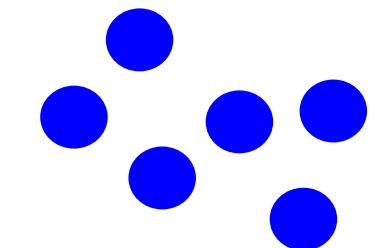
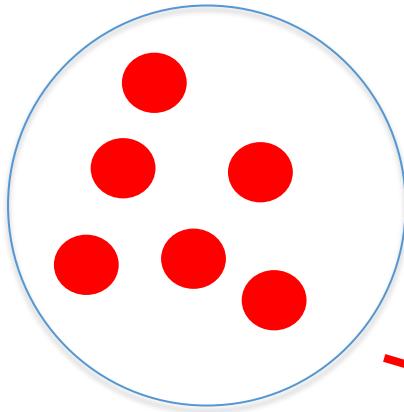
Avoiding the trivial solutions of entropy minimization



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

This bad solution also has a minimum entropy!!!

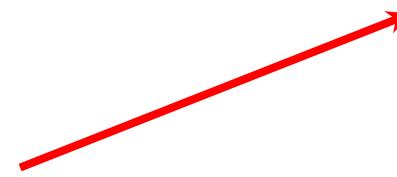
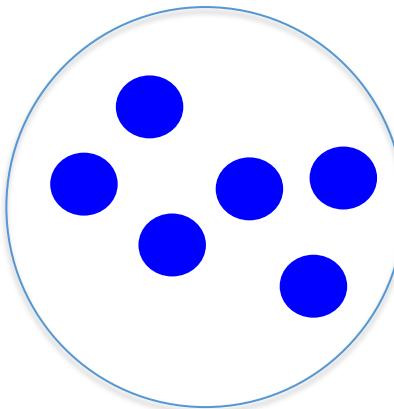
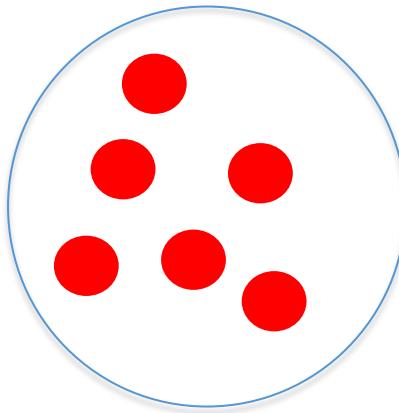
Avoiding the trivial solutions of entropy minimization



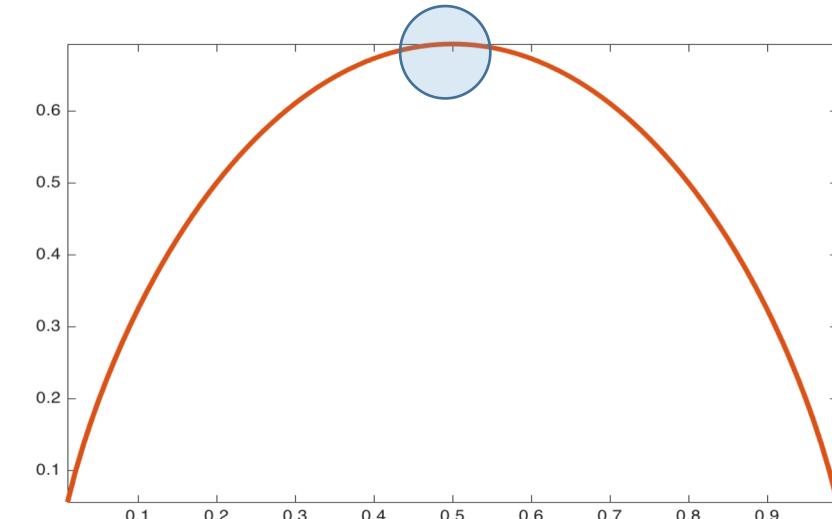
Marginal probabilities of the labels
-- *Class proportion*
-- *Region size (normalized) in segmentation*

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

Avoiding the trivial solutions of entropy minimization



Balanced solution maximizes the entropy of label marginal



$\Pr(c_1)$

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

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

$MI = \text{Entropy}$ (label marginal) – Entropy (posterior)

Standard and old in clustering, e.g.:

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

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

$$MI = \textcolor{red}{\text{Entropy}(\text{label marginal}) - Entropy(\text{posteriors})}$$

Up to a constant

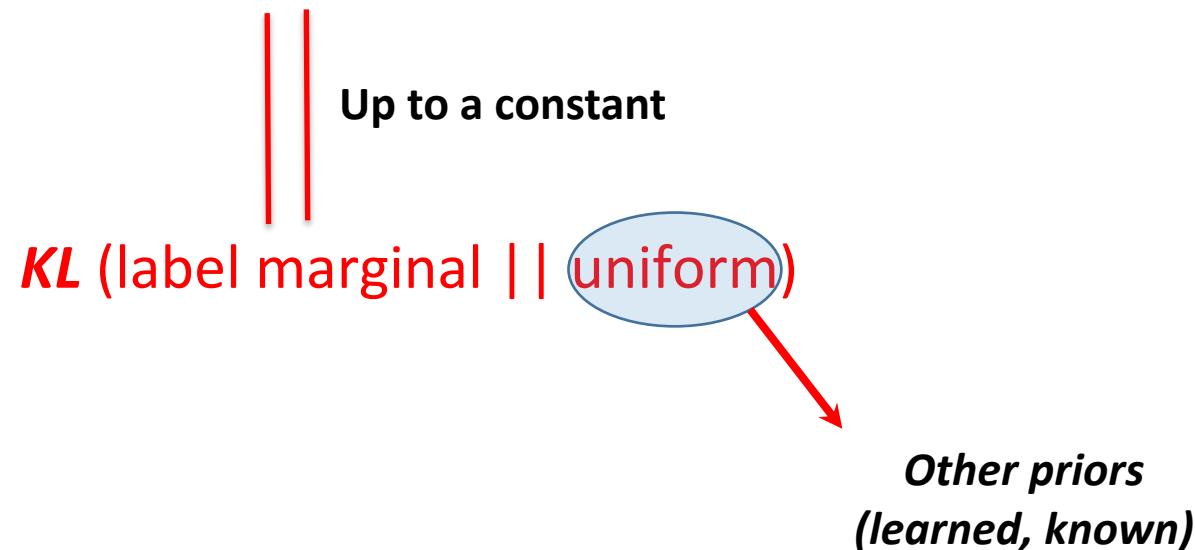
$$\textcolor{red}{KL(\text{label marginal} || \text{uniform})}$$

Standard and old in clustering:

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

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

$$MI = \textcolor{red}{\text{Entropy}} \text{ (label marginal)} - \textcolor{red}{\text{Entropy}} \text{ (posteriors)}$$

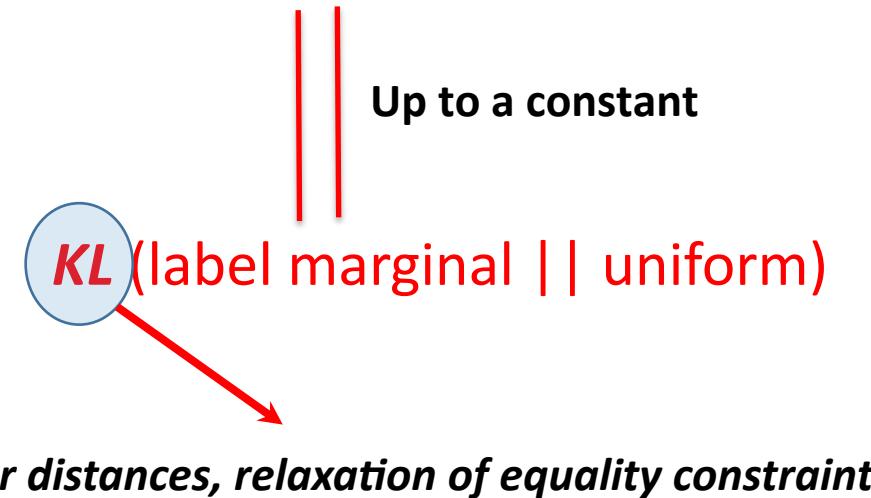


Standard and old in clustering:

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

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

$$MI = \textcolor{red}{Entropy}(\text{label marginal}) - \textcolor{red}{Entropy}(\text{posteriors})$$



Standard and old in clustering:

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

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

Semi-supervised learning, e.g.

[Berthelot et al., NeurIPS'19]
[Kervadec et al., Media'19]

Few-shot learning, e.g.,

[Boudiaf et al., NeurIPS'20]
[Dhillon et al., ICLR'20]

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

Unsupervised domain adaptation, e.g.,

Liang et al., ICML'20
Bateson et al., MICCAI'20

Deep clustering
&

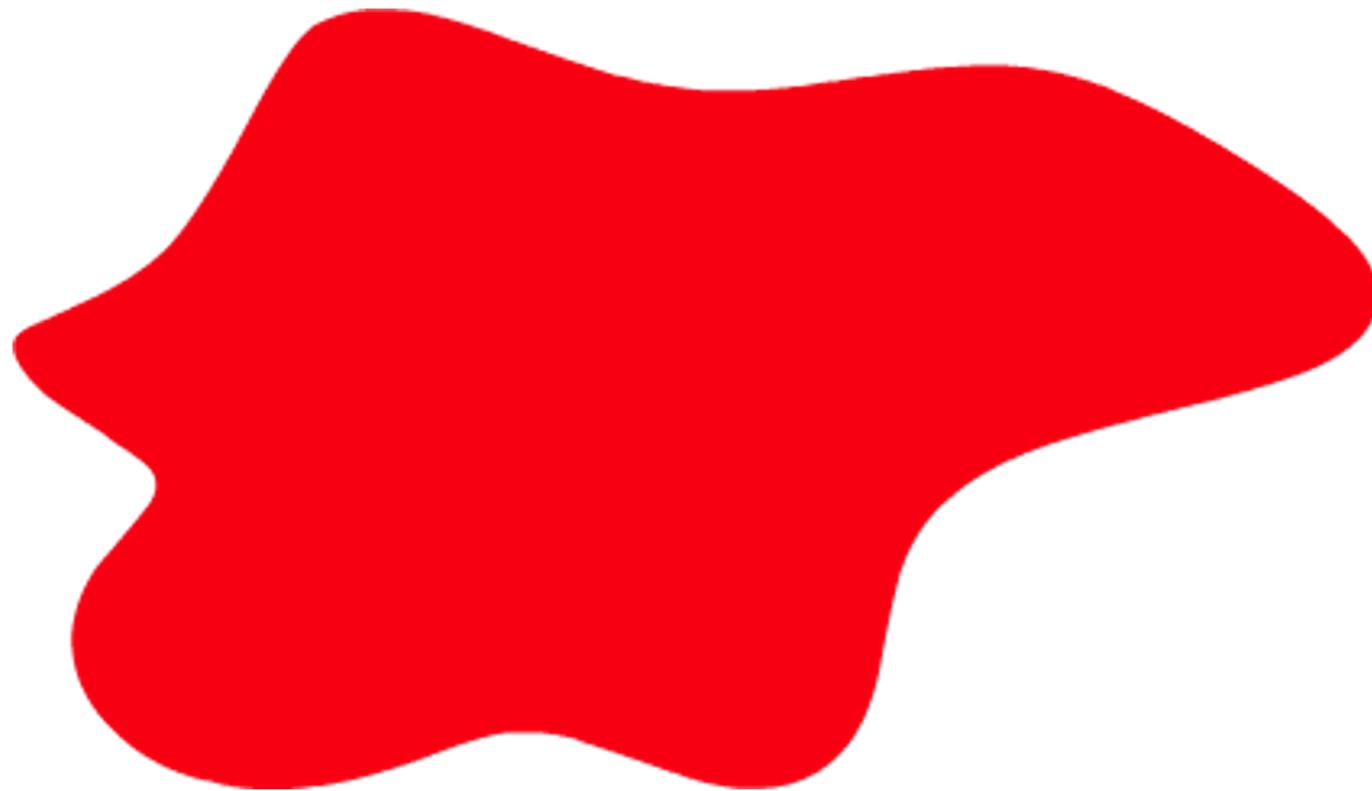
Unsupervised Representation Learning, e.g.,

Asano et al., ICLR'20
Jabi et al., TPAMI'20

Constrained CNNs

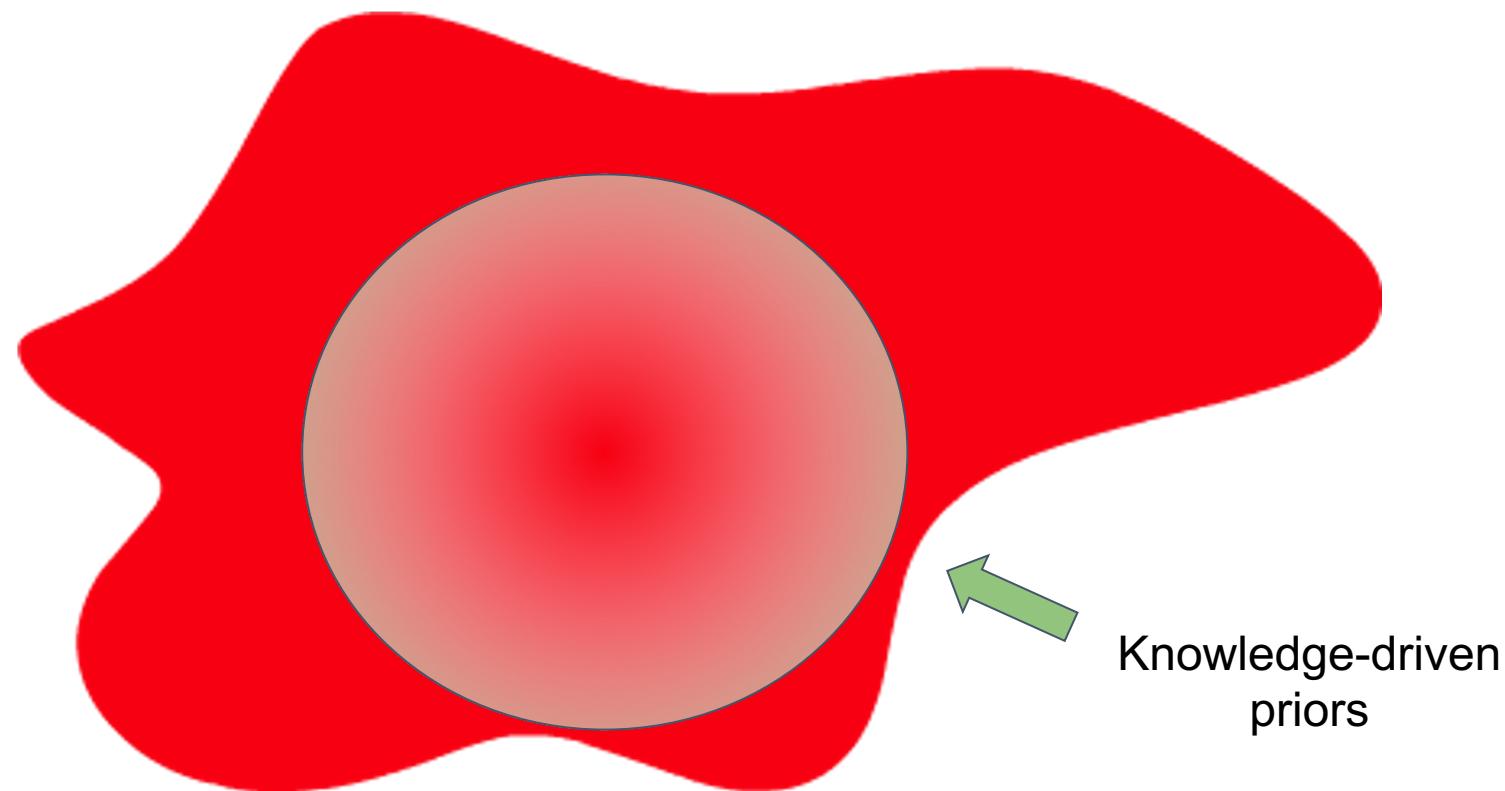
Constrained optimization (in CNNs)

Knowledge vs data driven priors



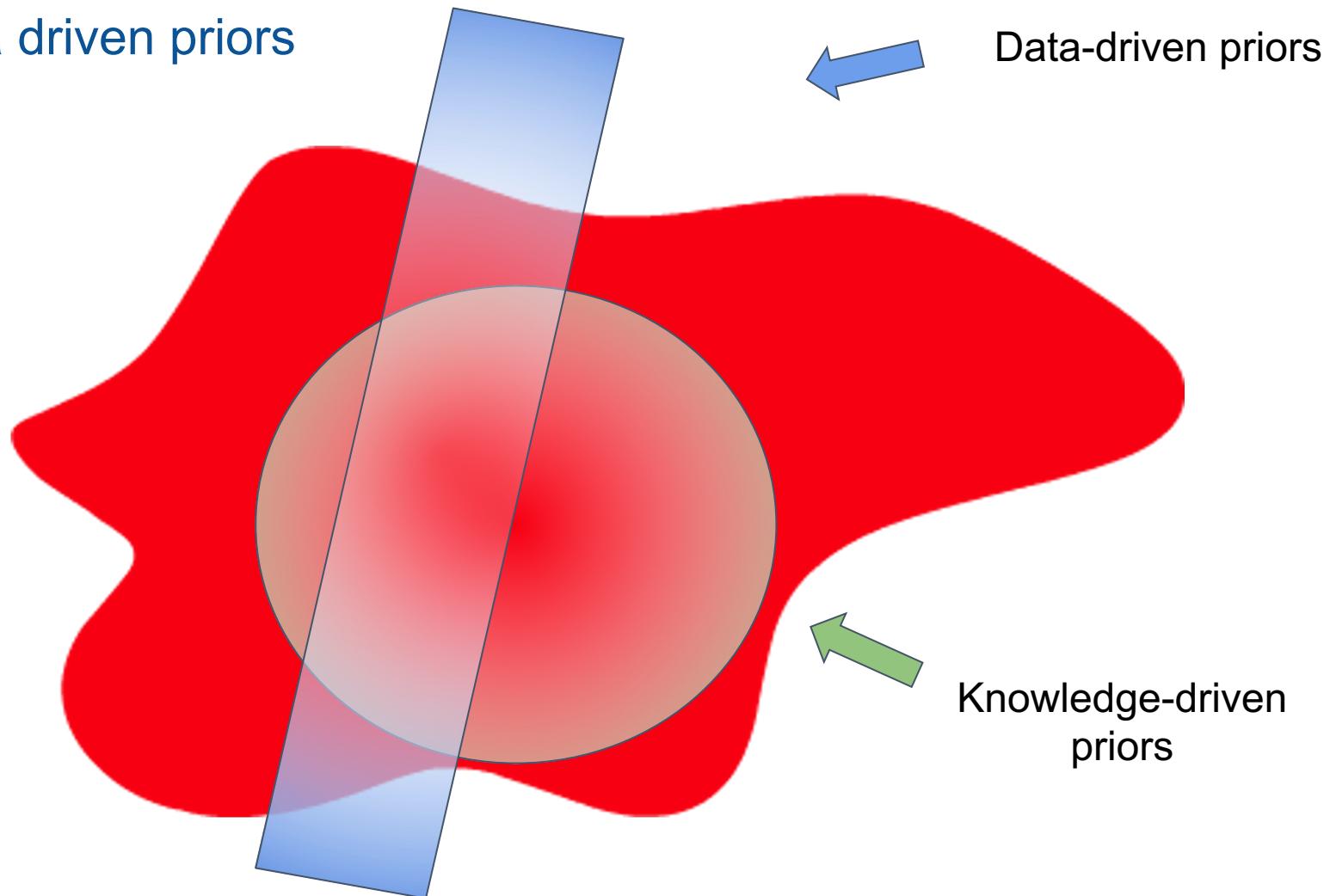
Constrained optimization (in CNNs)

Knowledge vs data driven priors



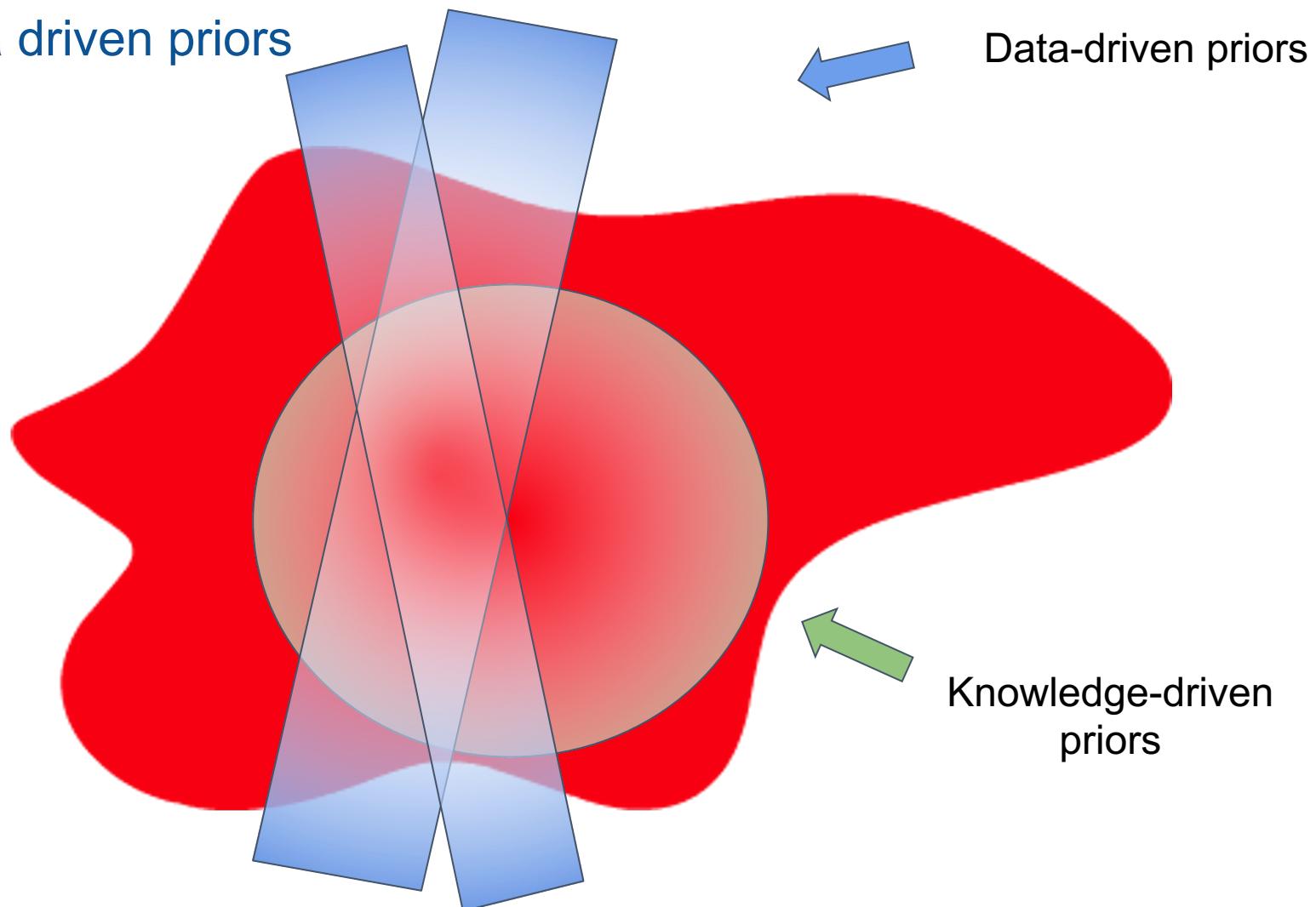
Constrained optimization (in CNNs)

Knowledge vs data driven priors



Constrained optimization (in CNNs)

Knowledge vs data driven priors

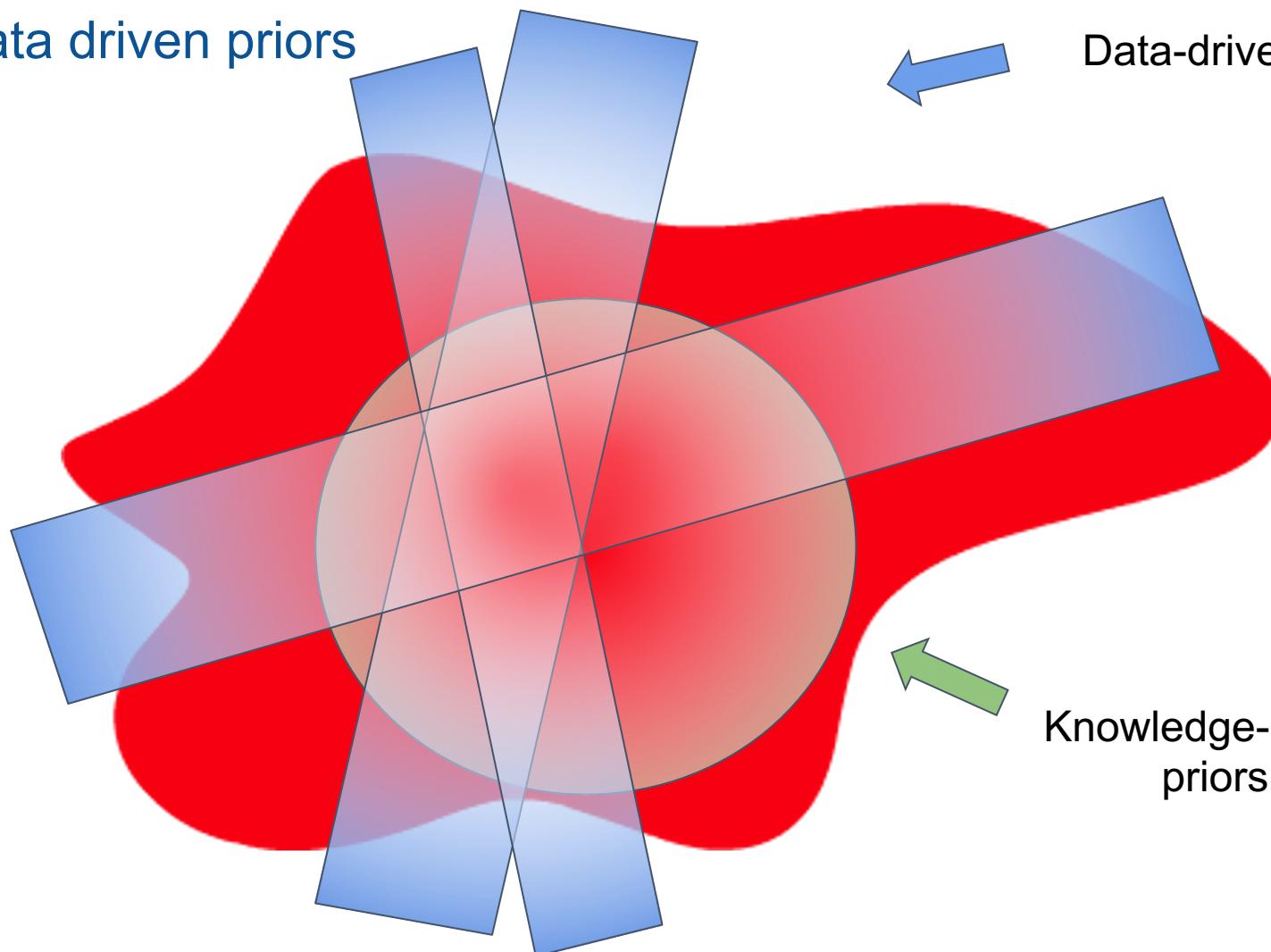


Constrained optimization (in CNNs)

Knowledge vs data driven priors

Data-driven priors

Knowledge-driven
priors



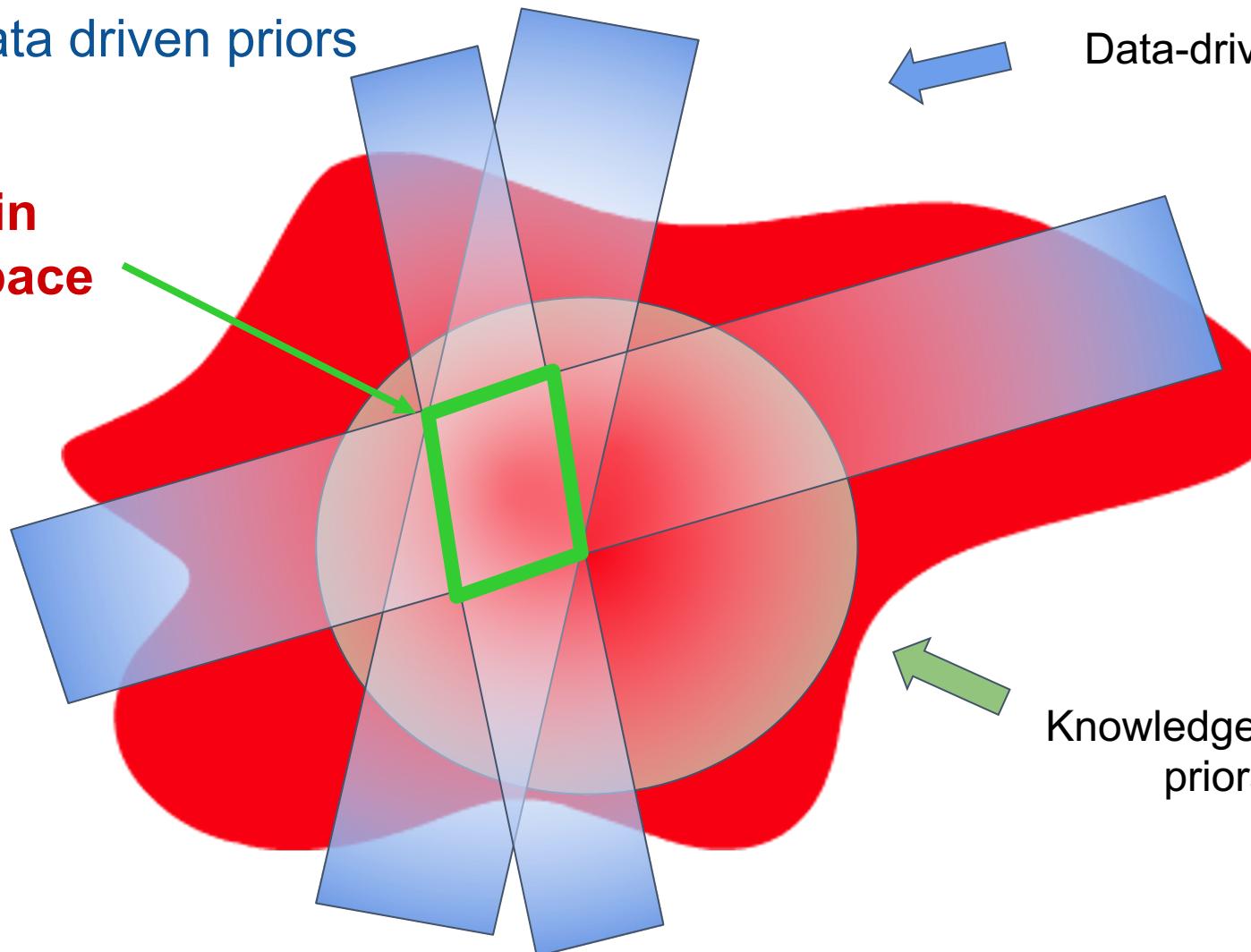
Constrained optimization (in CNNs)

Knowledge vs data driven priors

**Both constrain
the search space**

Data-driven priors

Knowledge-driven
priors

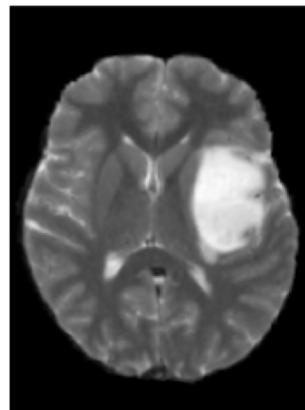


Data-driven priors (cues)

Image tags



Person
Bike



Tumor

Original
Image

Image tags

- Pathak et al., Constrained convolutional neural networks for weakly supervised segmentation, ICCV 2015
- Kervadec et al., Constrained-CNN losses for weakly supervised segmentation, Media 2019.

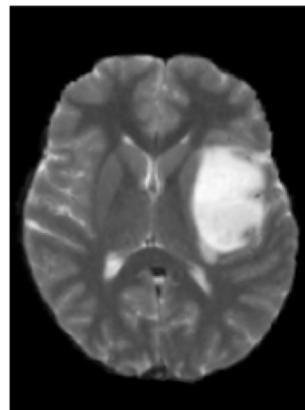
Data-driven priors (cues)

Image tags

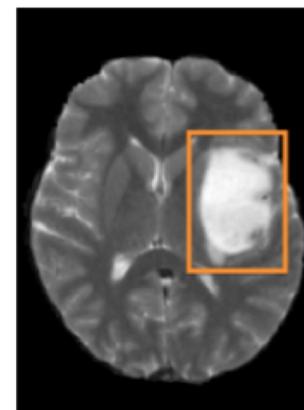
Bounding boxes



Person
Bike



Tumor



Original
Image

Image tags

Bounding
boxes

- Pathak et al., Constrained convolutional neural networks for weakly supervised segmentation, ICCV 2015
- Kervadec et al., Constrained-CNN losses for weakly supervised segmentation, Media 2019.

Data-driven priors (cues)

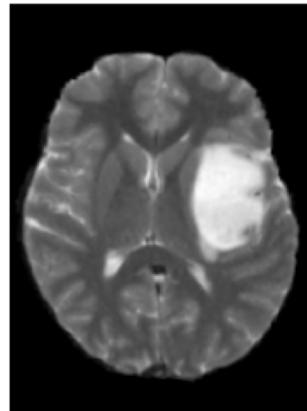
Image tags

Bounding boxes

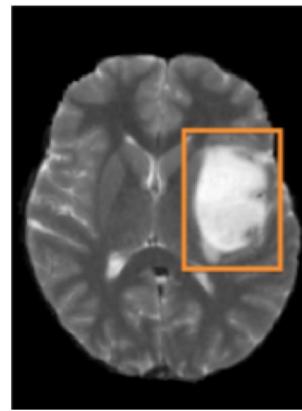
Scribbles



Person
Bike

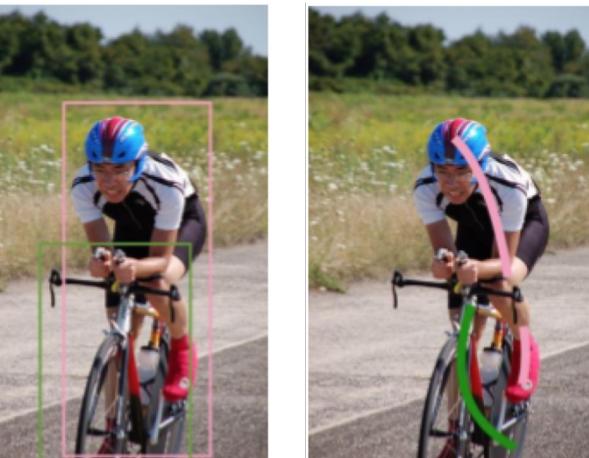


Tumor

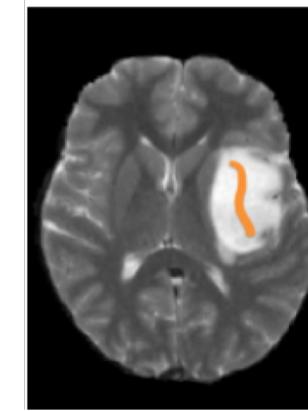


Original
Image

Image tags



Bounding
boxes



Scribbles

- Pathak et al., Constrained convolutional neural networks for weakly supervised segmentation, ICCV 2015
- Kervadec et al., Constrained-CNN losses for weakly supervised segmentation, Media 2019.

Data-driven priors (cues)

Image tags

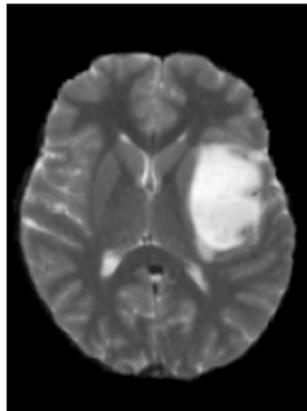
Bounding boxes

Scribbles

Points



Person
Bike

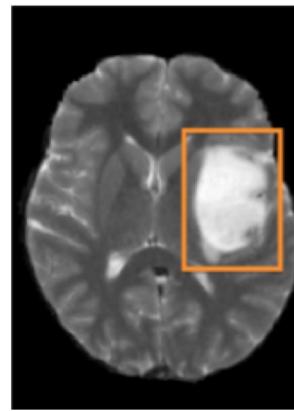


Tumor

Original
Image



Bounding
boxes



Scribbles



Points

- Pathak et al., Constrained convolutional neural networks for weakly supervised segmentation, ICCV 2015
- Kervadec et al., Constrained-CNN losses for weakly supervised segmentation, Media 2019.

Data-driven priors (cues)

Another data-driven priors

Image captions



A boy jumping on a skateboard

Extreme points

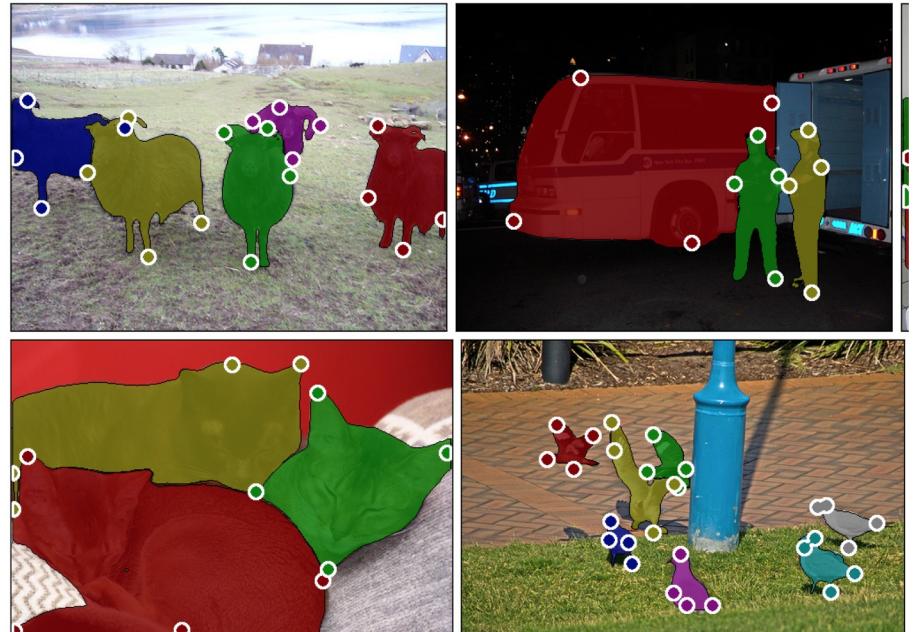


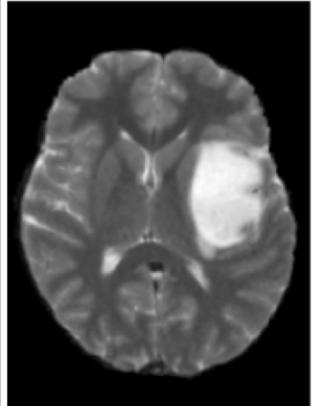
Image from Maninis et al, CVPR'18

- Maninis et al. Deep extreme cut: From extreme points to object segmentation. CVPR 2018

From global cues to pixel labels



Person
Bike



Tumor

Original
Image

Image tags



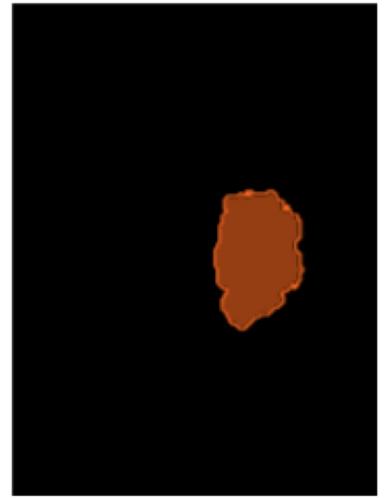
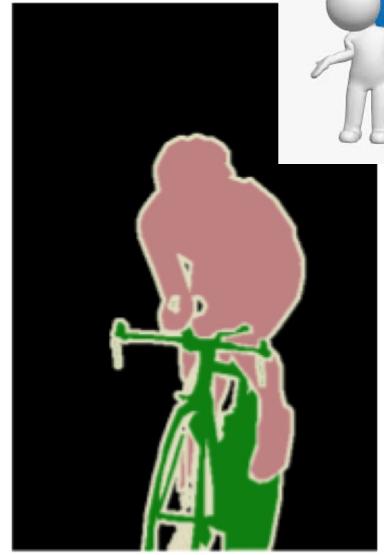
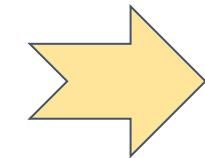
Bounding
boxes



Scribbles

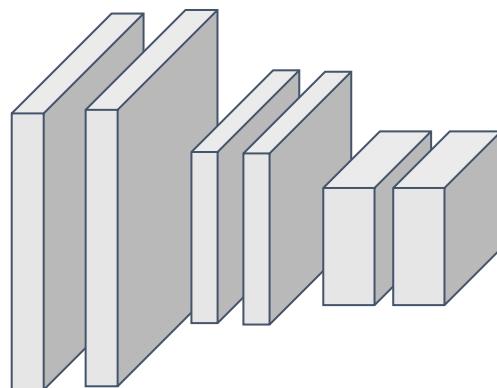
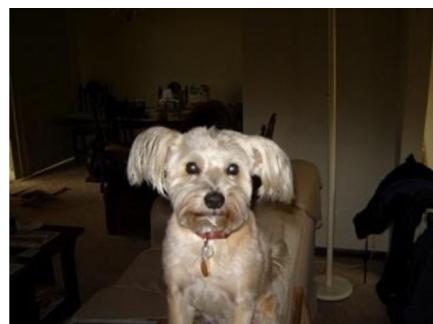


Points



From global cues to pixel labels

Step 1: Get a classification CNN



Convolutional layers



FC Layers

Class scores

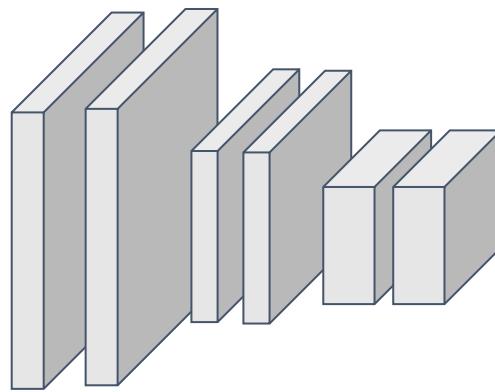
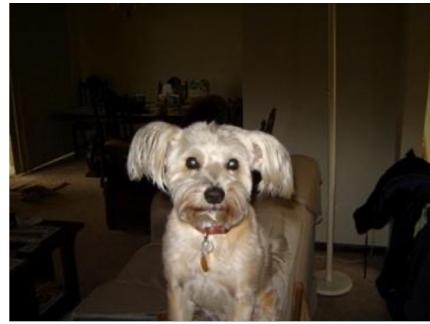
Cat

Dog

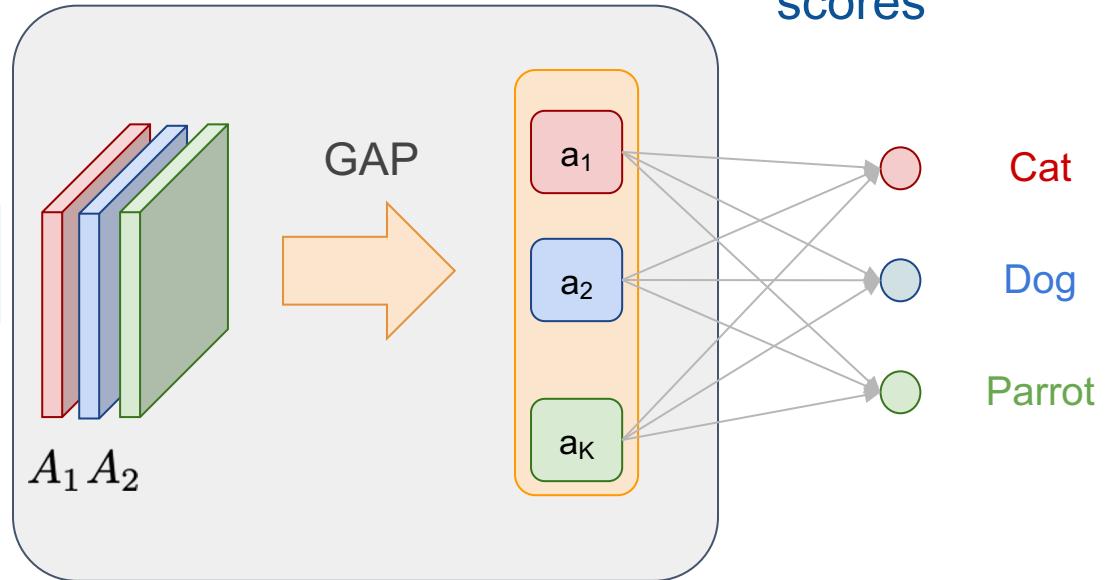
Parrot

From global cues to pixel labels

Step 2: Modify the last layers



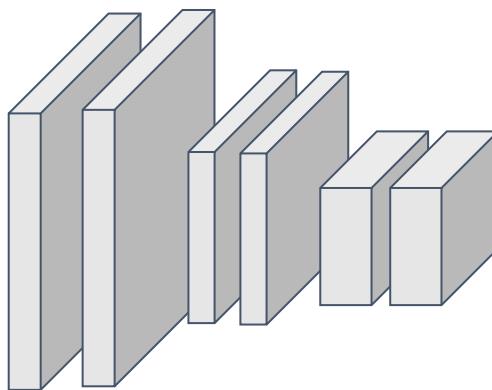
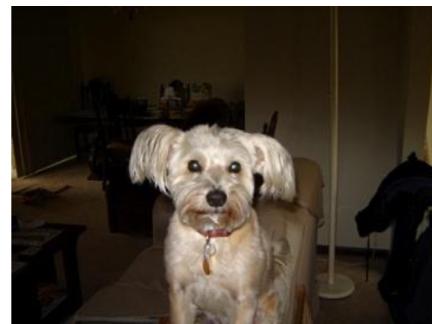
Convolutional layers



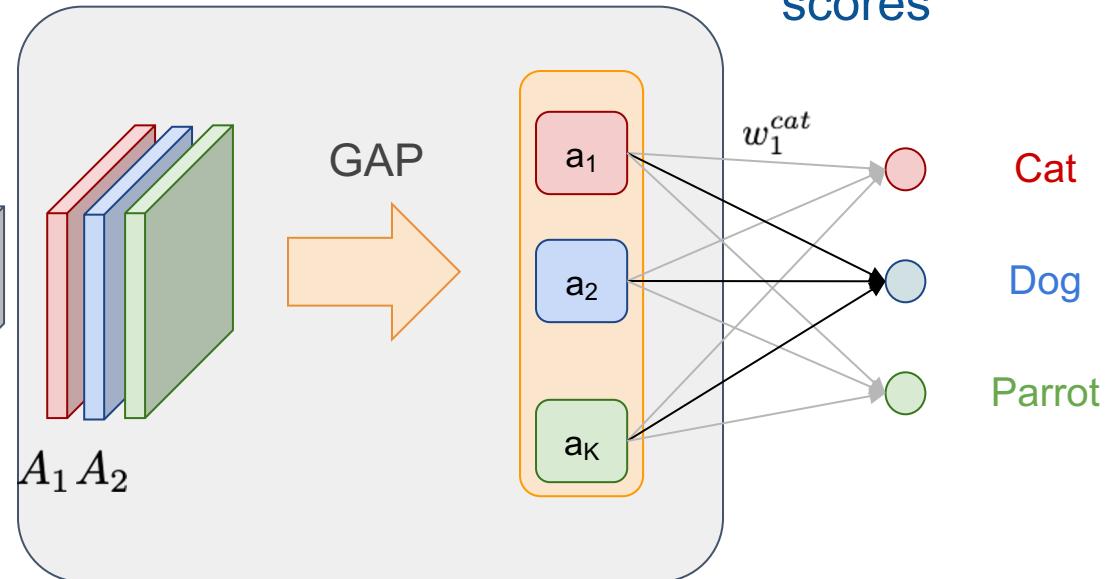
$$GAP(A_k) = a_k = \frac{1}{|N|} \sum_{x,y} A_k(x,y)$$

From global cues to pixel labels

Step 2: Modify the last layers



Convolutional layers



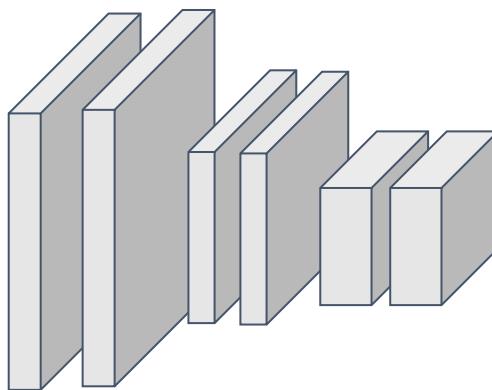
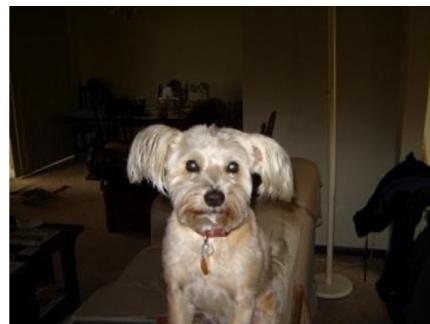
$$GAP(A_k) = a_k = \frac{1}{|N|} \sum_{x,y} A_k(x,y)$$

Class score
(logits)

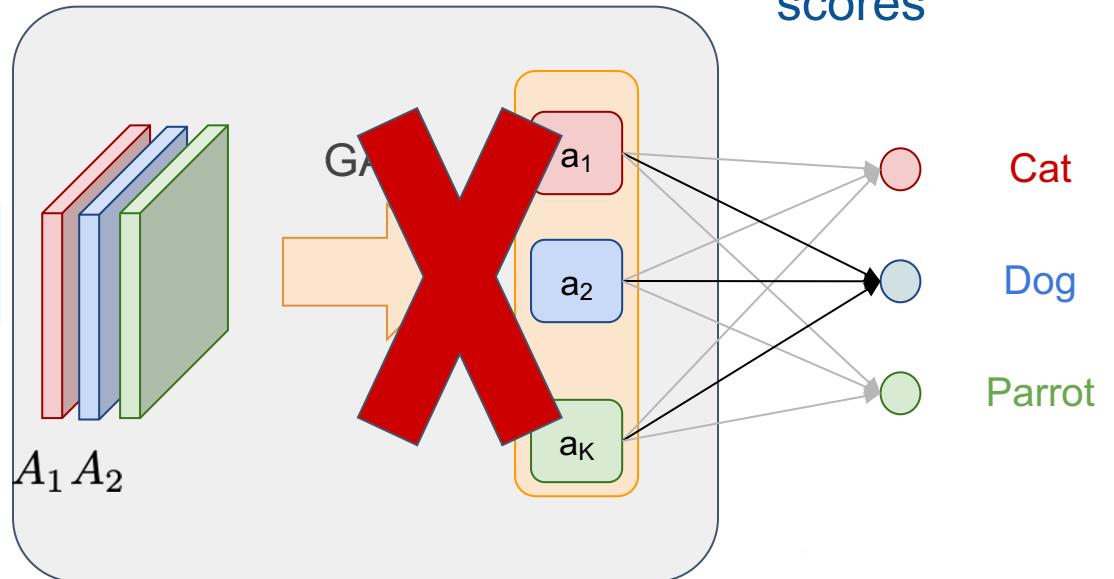
$$S_c = \sum_k w_k^c a_k = \frac{1}{N} \sum_k w_k^c \sum_{x,y} a_k(x,y)$$

From global cues to pixel labels

Step 3: Get the CAMs



Convolutional layers



$$CAM_{Dog}(x, y) = \sum_k w_k^{Dog} A_k(x, y) =$$



From global cues to pixel labels

Mushroom



Penguin



Teapot



- Zhou et al., Learning deep features for discriminative localization. CVPR 2016

From global cues to pixel labels

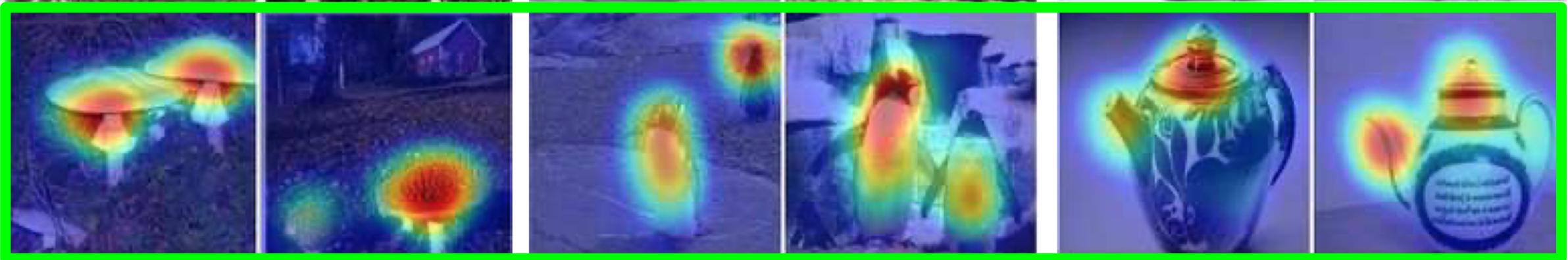
Mushroom



Penguin



Teapot



These activations maps can be used as **pseudo-masks**

From global cues to pixel labels

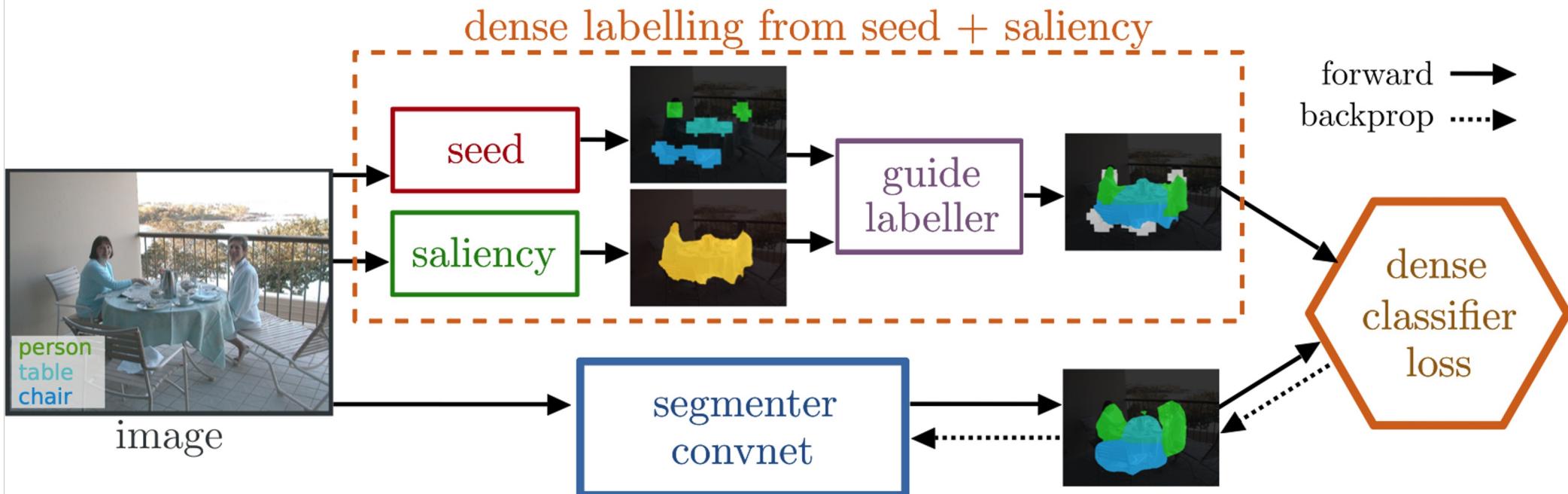
Problem: they focus only on highly discriminative regions



From global cues to pixel labels

Problem: they focus only on highly discriminative regions

Incorporate saliency maps

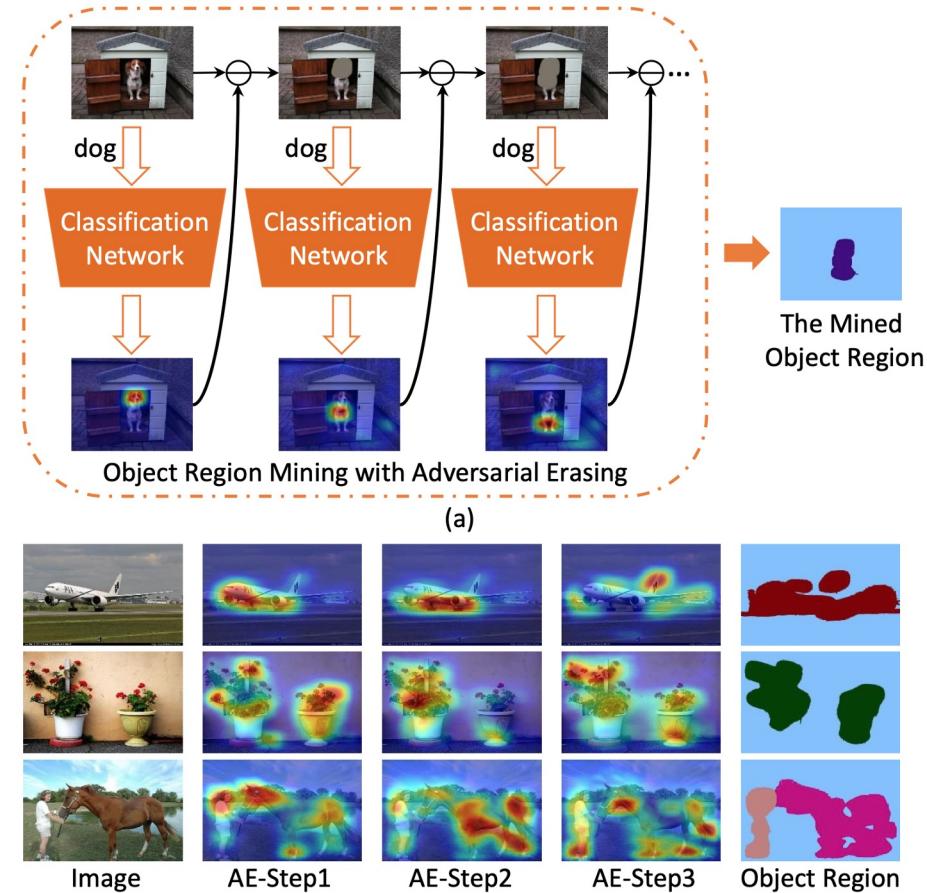


- Oh et al. Exploiting Saliency for Object Segmentation from Image Level Labels. CVPR 2017
- Fan et al. Learning Integral Objects With Intra-Class Discriminator for Weakly-Supervised Semantic Segmentation. CVPR 2020

From global cues to pixel labels

Problem: they focus only on highly discriminative regions

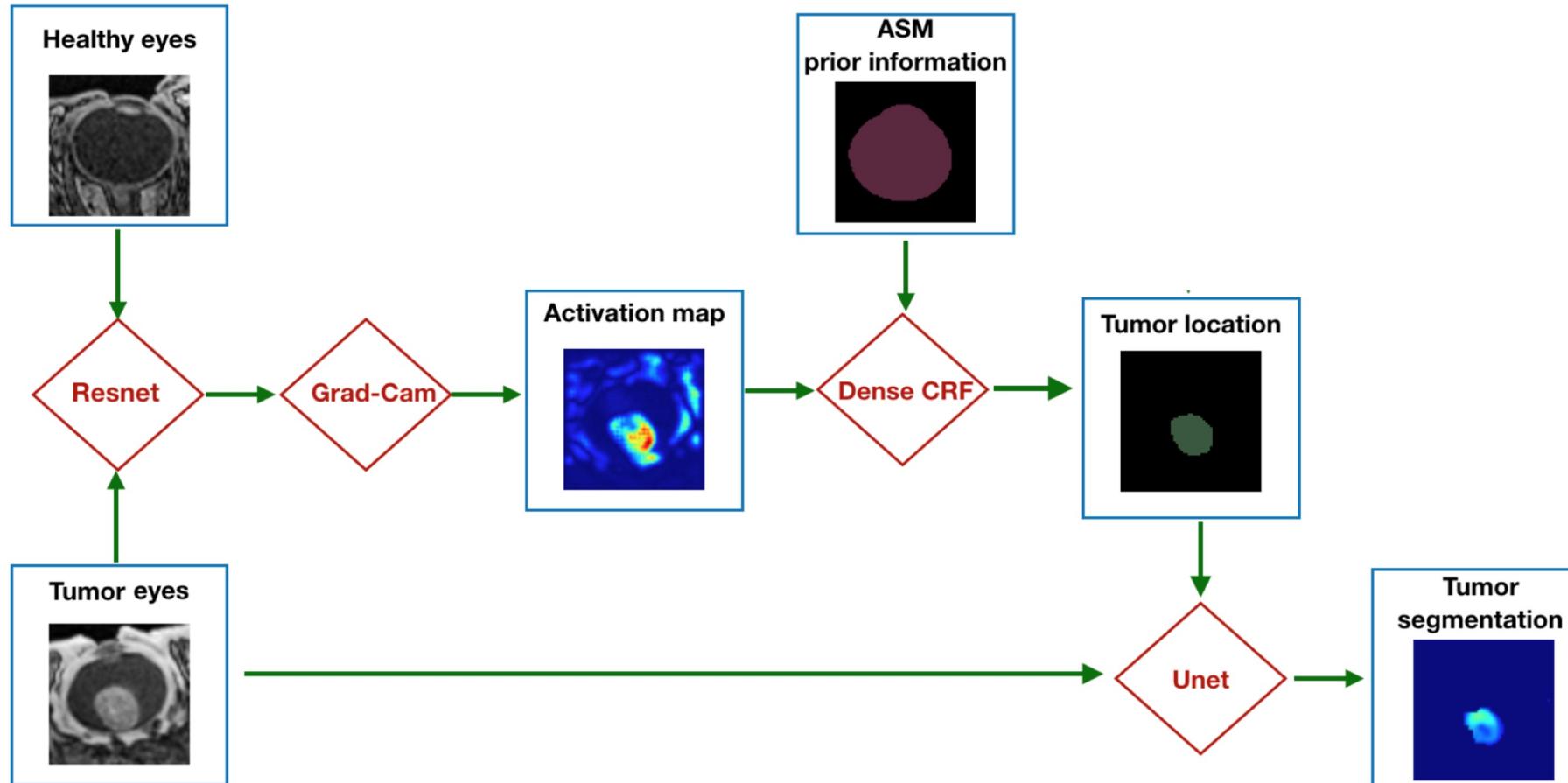
Region mining



- Wei et al. Object Region Mining with Adversarial Erasing: A Simple Classification to Semantic Segmentation Approach. CVPR 2017
- Wang et al. Weakly-Supervised Semantic Segmentation by Iteratively Mining Common Object Features. CVPR 2018

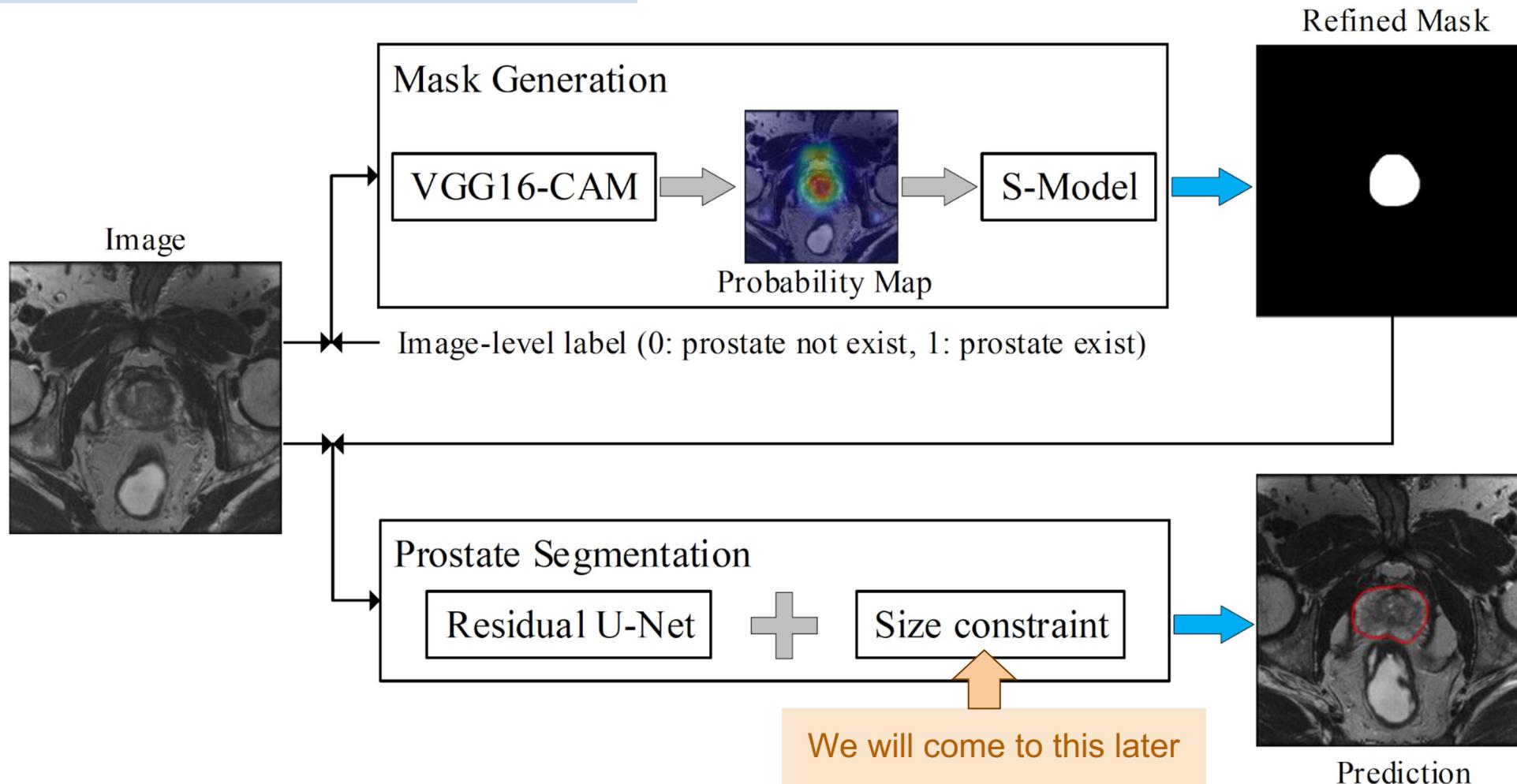
From global cues to pixel labels

CAMs in the medical domain



From global cues to pixel labels

CAMs in the medical domain



Constrained optimization (in CNNs)

Knowledge-driven priors

Common priors in natural images

Target Size



- Pathak et al., Constrained convolutional neural networks for weakly supervised segmentation, ICCV 2015
- Xu et al., Learning to Segment Under Various Forms of Weak Supervision, CVPR 2015
- Zhang et al., Curriculum Domain Adaptation for Semantic Segmentation of Urban Scenes. ICCV'17

Constrained optimization (in CNNs)

Knowledge-driven priors

Common priors in natural images

Target Location



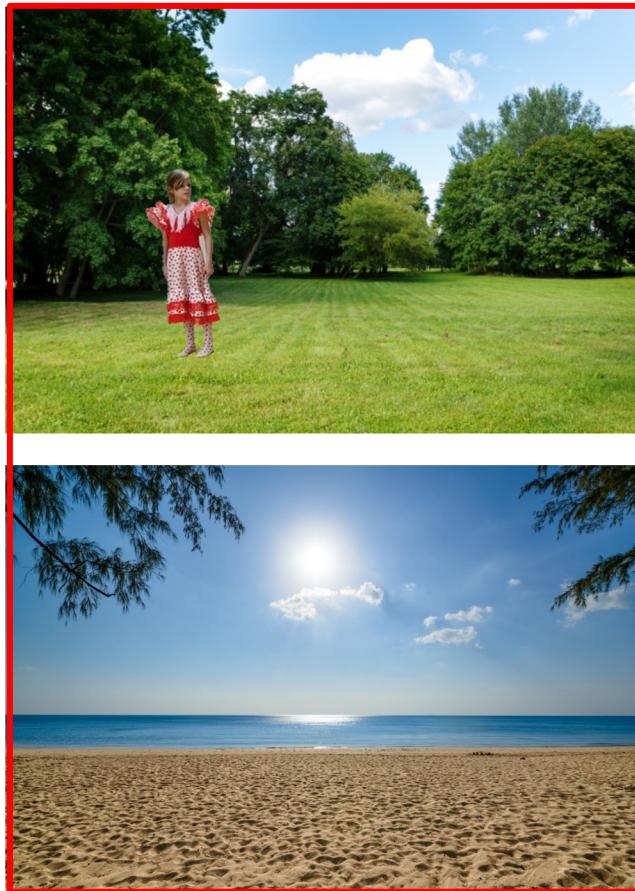
- Remez et al. Learning to segment via cut-and-paste. ECCV 2018
- Georgakis et al Synthesizing training data for object detection in indoor scenes. RSS 2017

Constrained optimization (in CNNs)

Knowledge-driven priors

Common priors in natural images

Number of instances



Constrained optimization (in CNNs)

Knowledge-driven priors

Common priors in natural images

Contrast
Foreground/Background

Saliency



Images from Hou et al, CVPR'17

- Hou et al. Deeply supervised salient object detection with short connections. CVPR 2017
- Li et al. Instance-level salient object segmentation. CVPR 2017

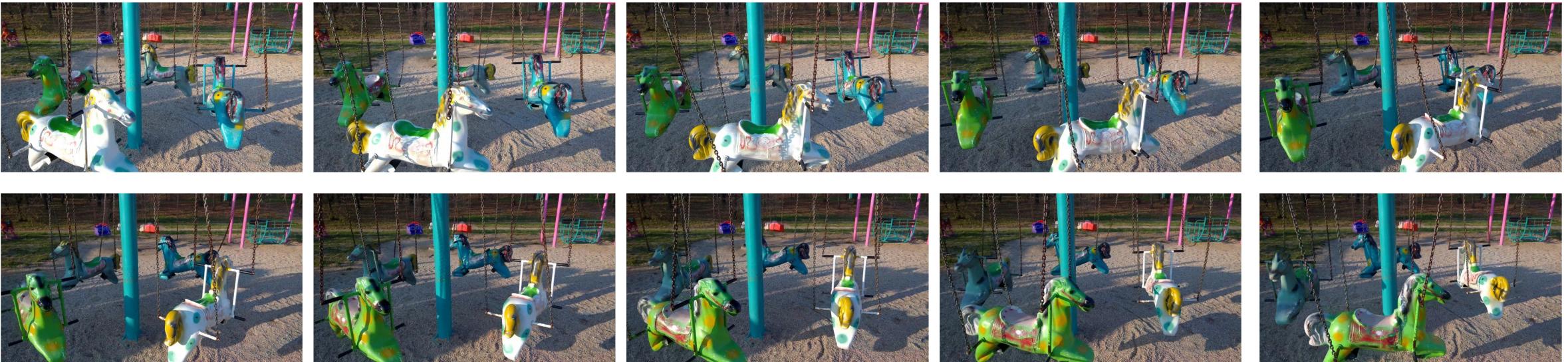
Constrained optimization (in CNNs)

Knowledge-driven priors

Common priors in natural images

Motion

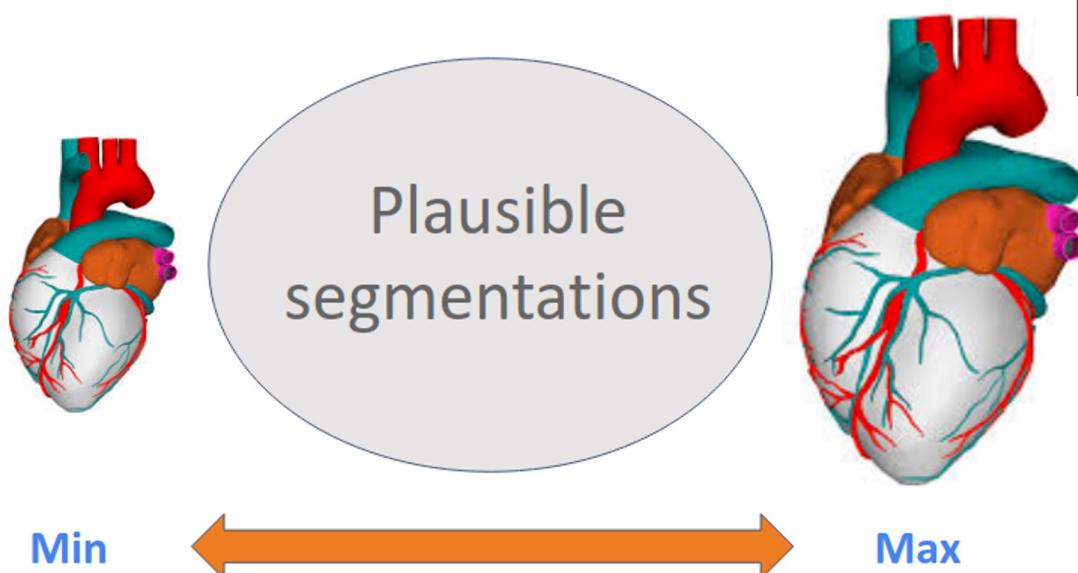
Images from the DAVIS Challenge Dataset



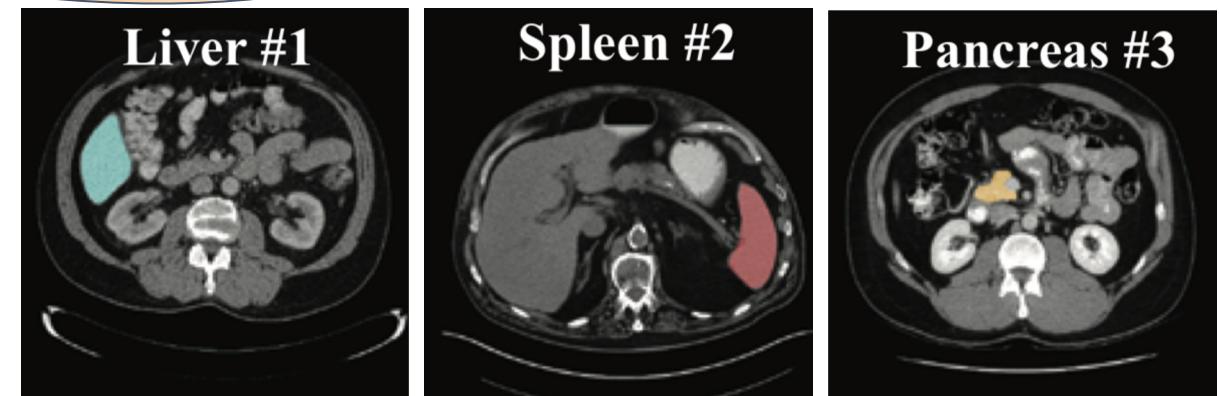
- Tokmakov et al. Weakly-supervised semantic segmentation using motion cues. ECCV 2016
- Pathak et al. Learning features by watching objects move. CVPR 2017

Constrained optimization (in CNNs)

Knowledge-driven priors



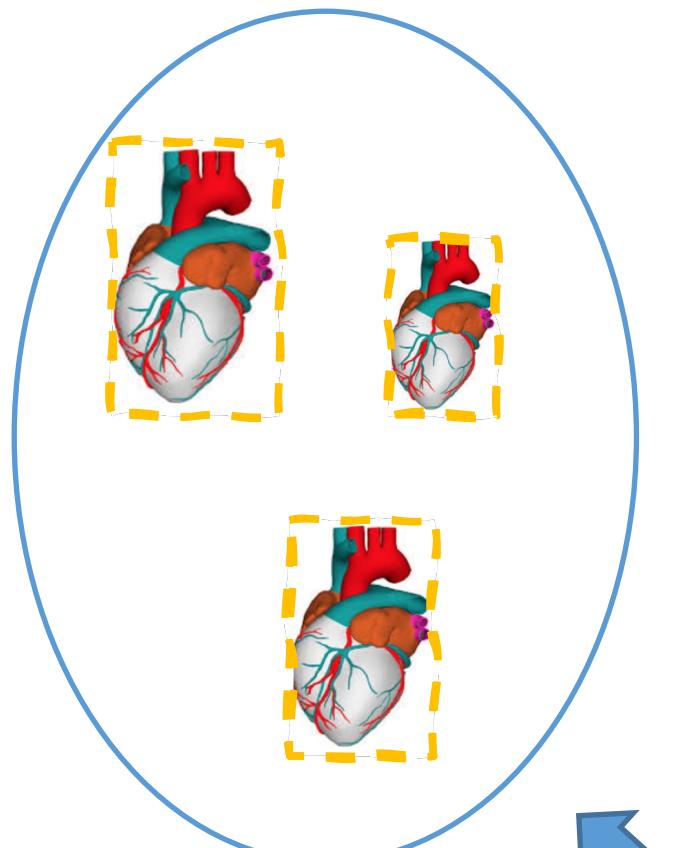
What about priors in the medical domain?



Partial labeled data
(exploit target relationships)

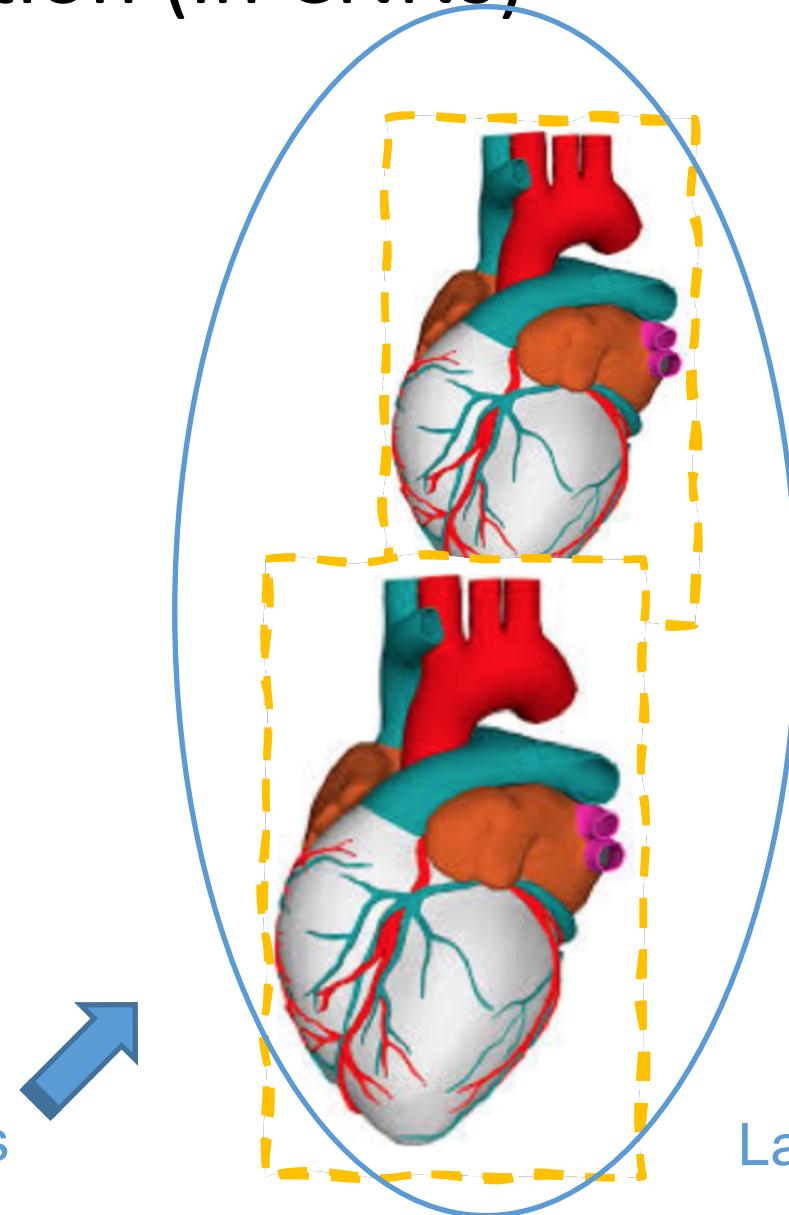
Constrained optimization (in CNNs)

Equality constraints



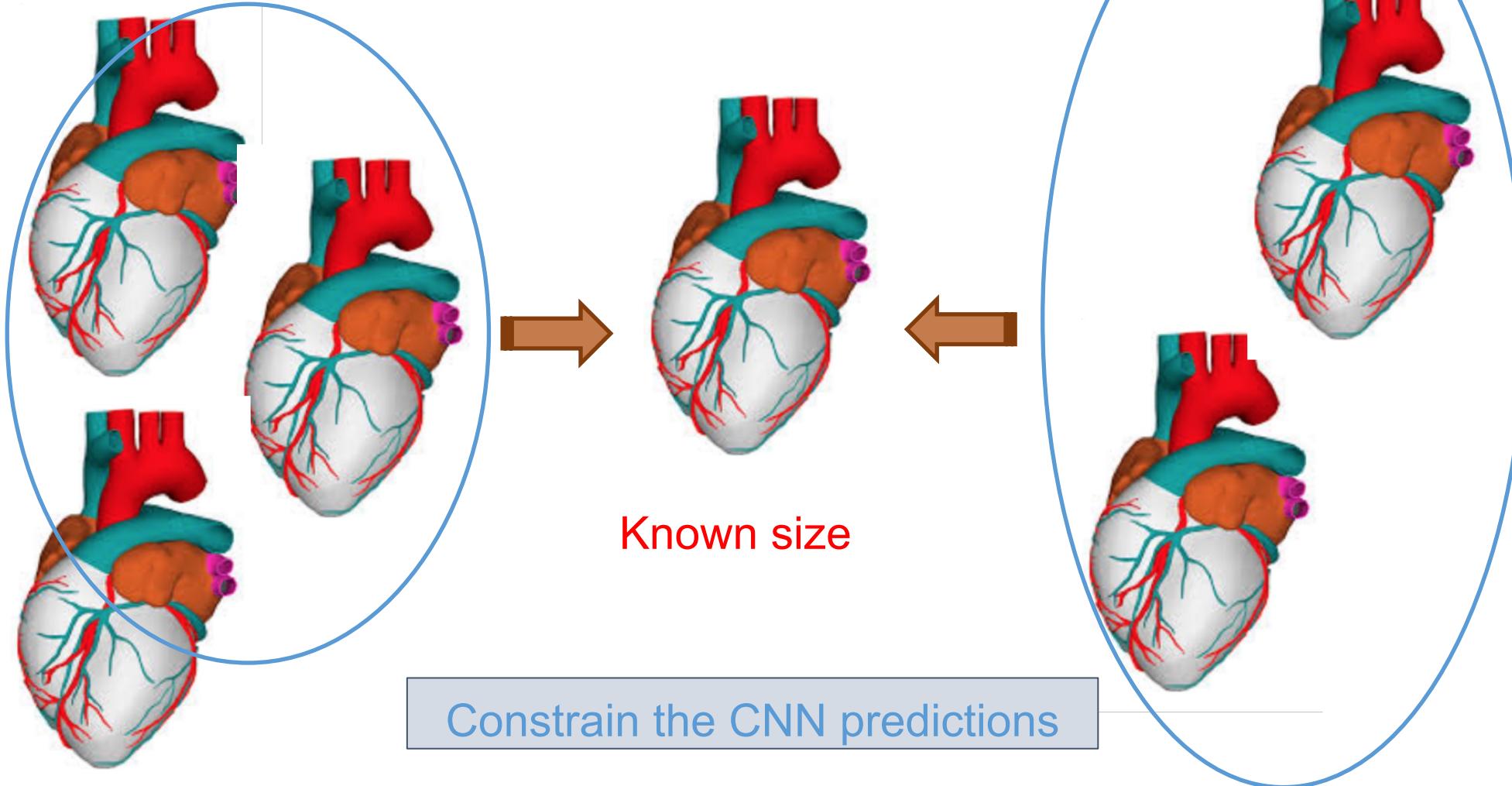
Known size

CNN predictions



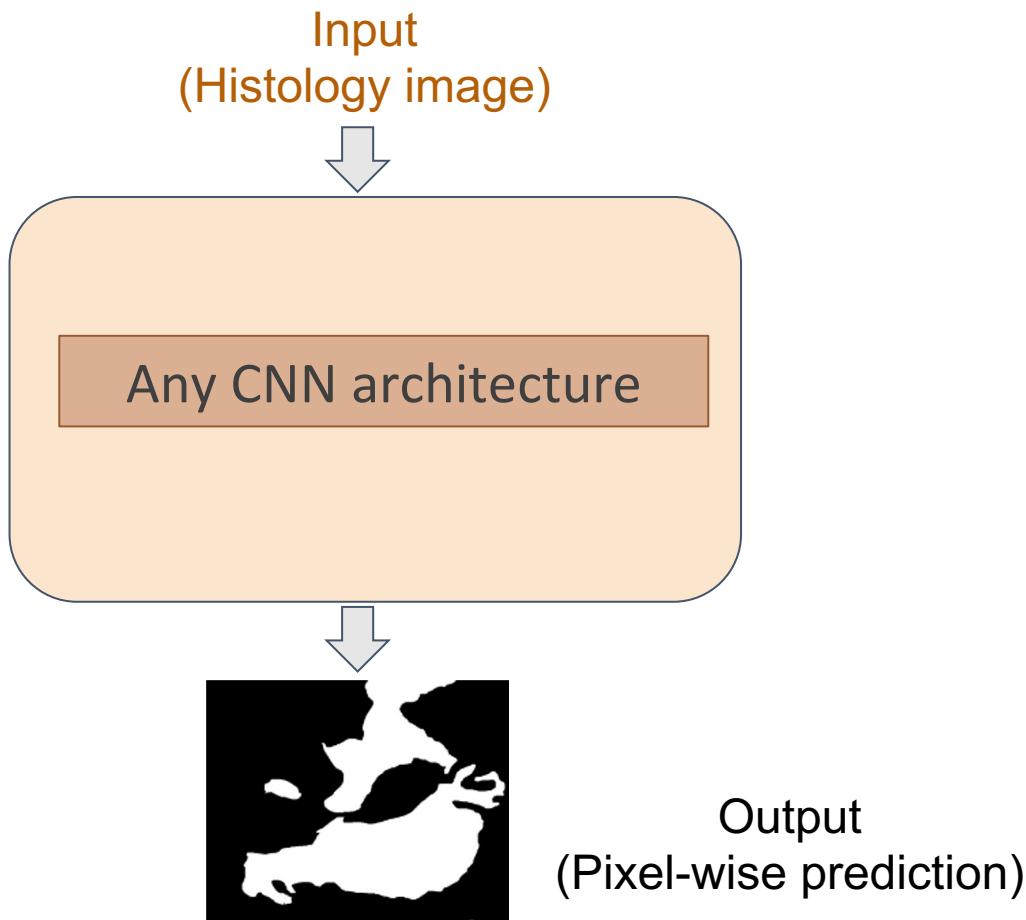
Constrained optimization (in CNNs)

Equality constraints



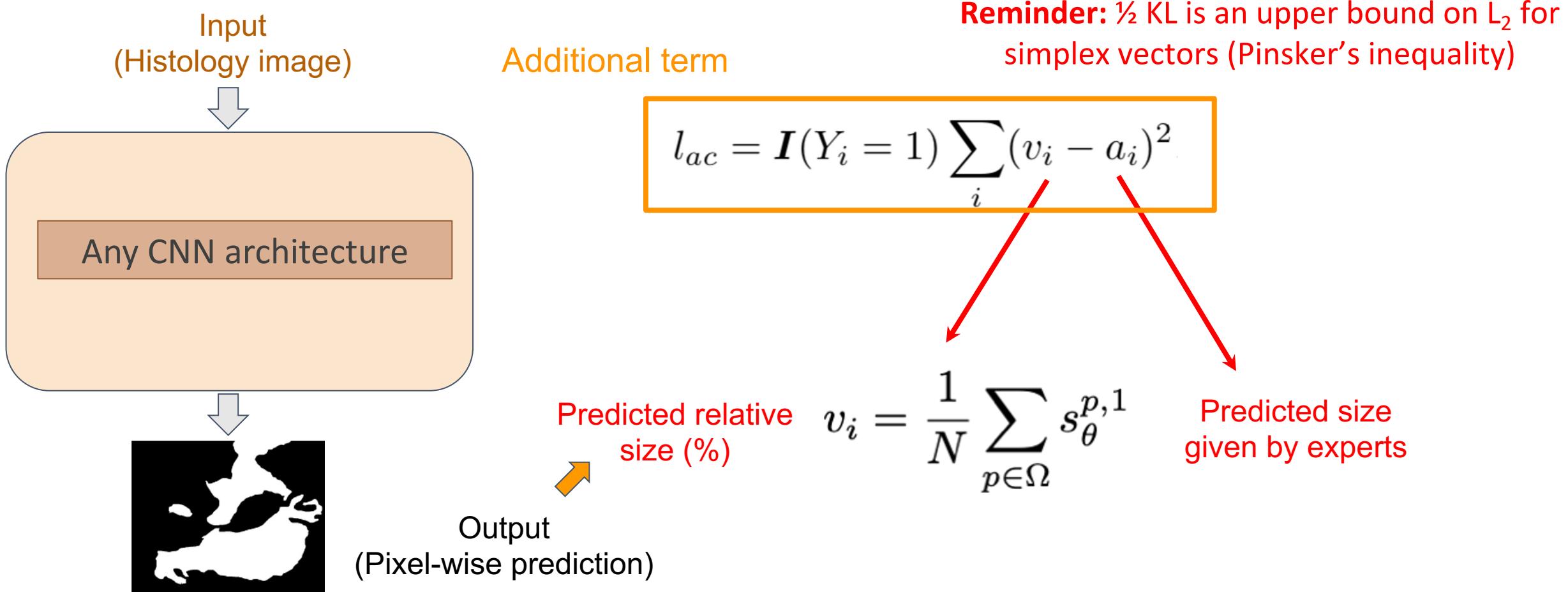
Constrained optimization (in CNNs)

Equality constraints (e.g, L2 penalty)



Constrained optimization (in CNNs)

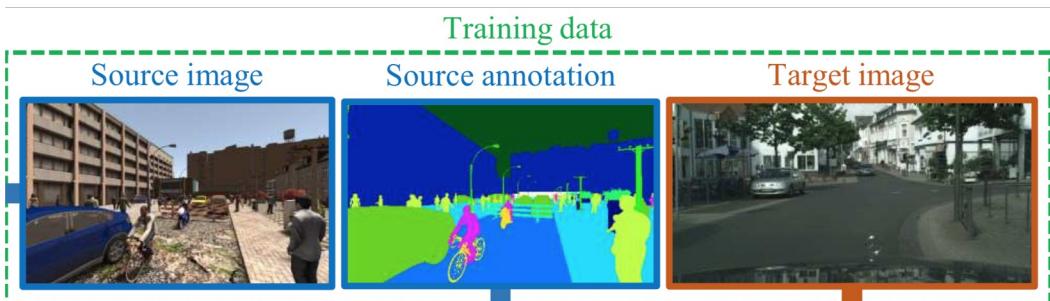
Equality constraints (e.g, L2 penalty)



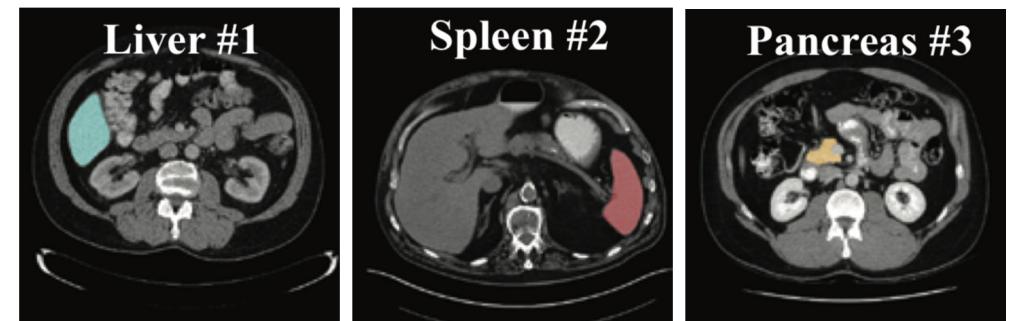
Constrained optimization (in CNNs)

Equality constraints (e.g, KL)

Unsupervised domain
adaptation



Partially labeled data



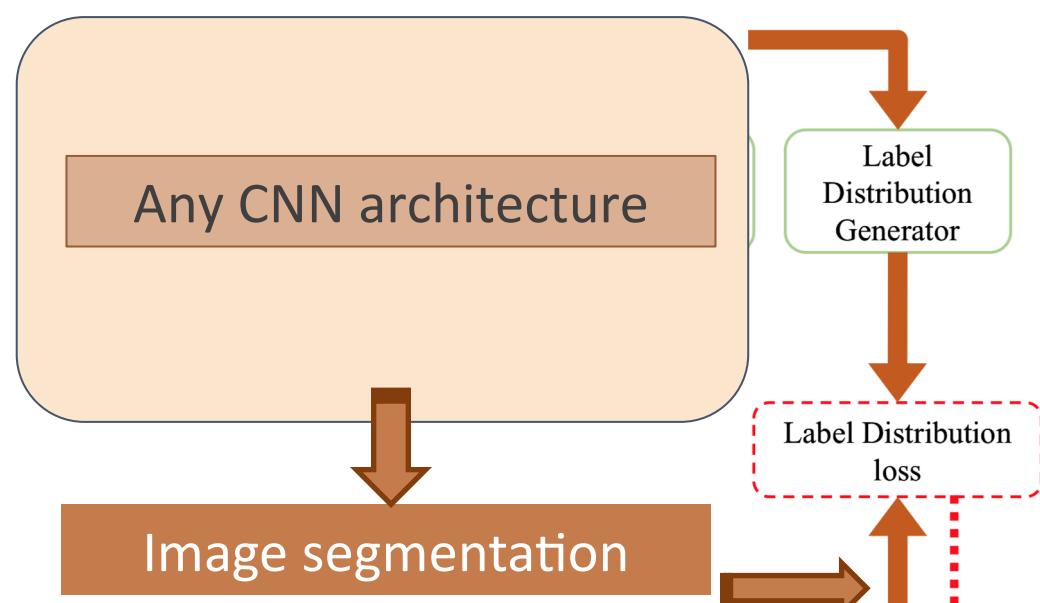
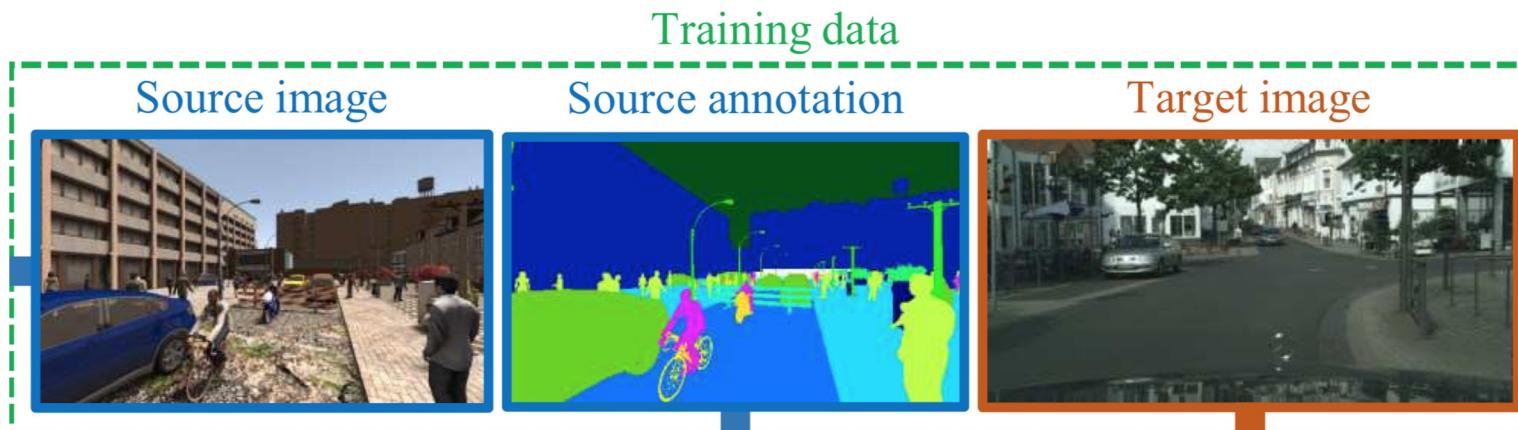
Constrained optimization (in CNNs)

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



Constrained optimization (in CNNs)

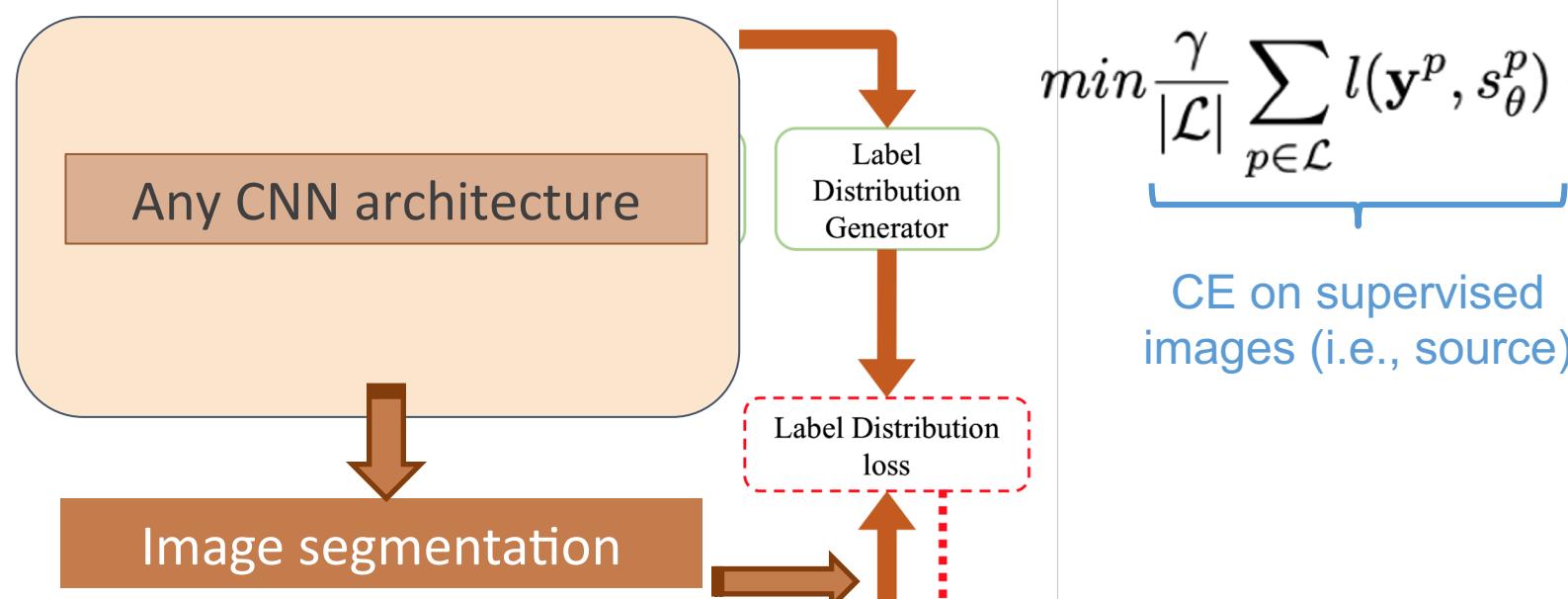
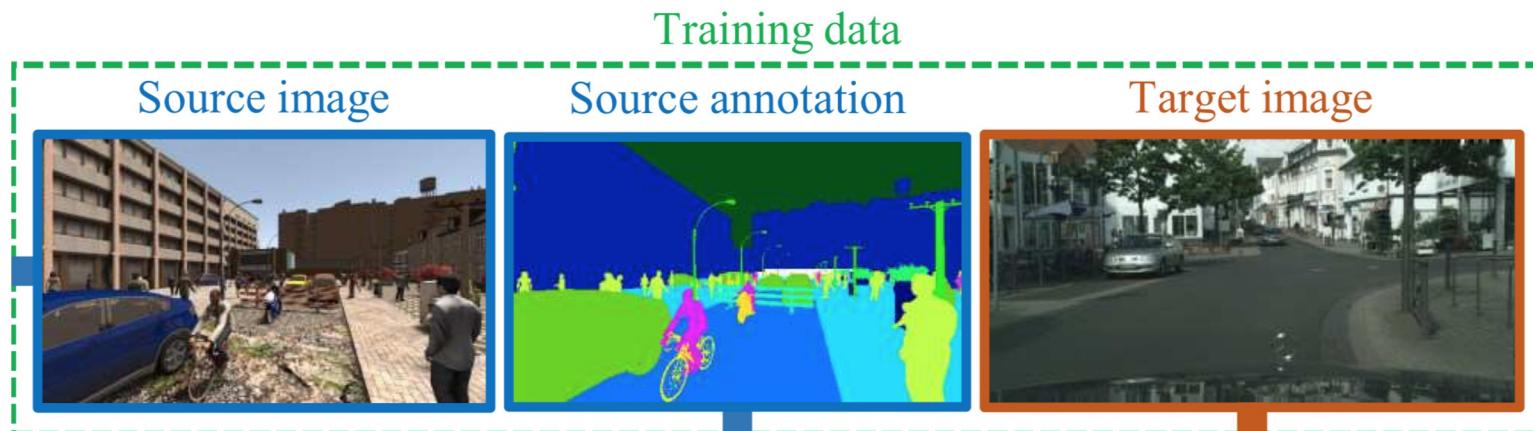
Equality constraints (e.g, KL): Curriculum DA



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

Constrained optimization (in CNNs)

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

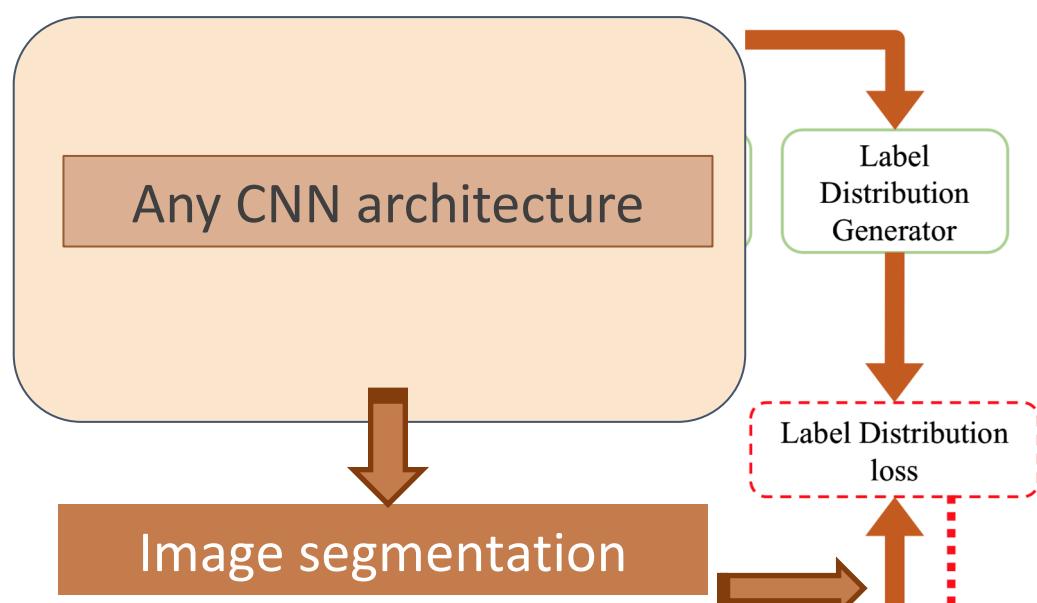
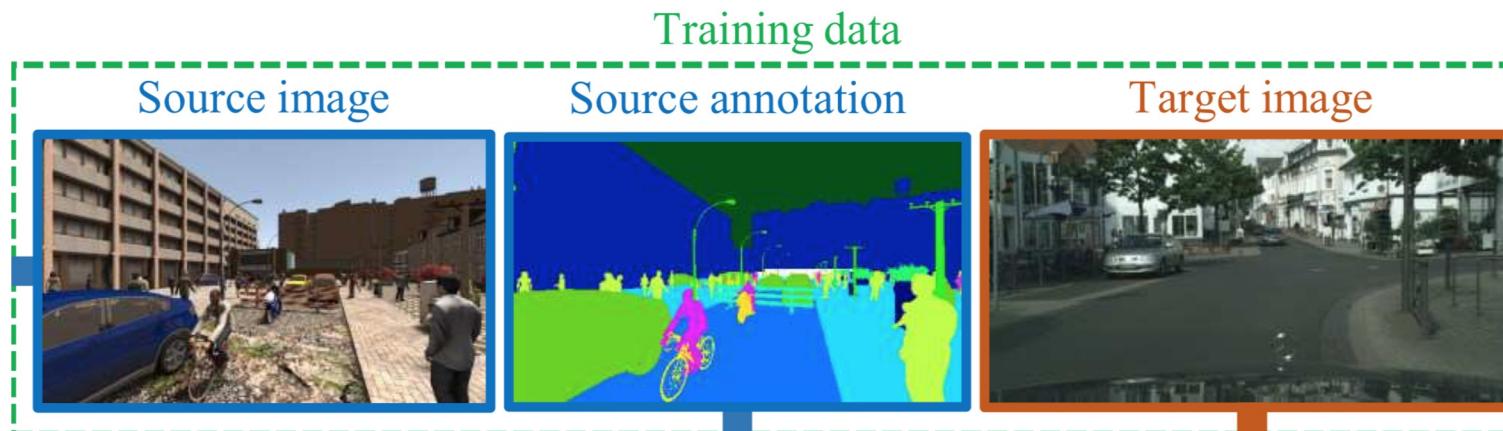


$$\min \frac{\gamma}{|\mathcal{L}|} \sum_{p \in \mathcal{L}} l(\mathbf{y}^p, s_\theta^p) + \frac{1 - \gamma}{|\mathcal{U}|} \sum_{q \in \mathcal{U}} \sum_k \mathbf{C}(a^{q,k}, \hat{a}^{q,k})$$

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

Constrained optimization (in CNNs)

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



$$\min \frac{\gamma}{|\mathcal{L}|} \sum_{p \in \mathcal{L}} l(\mathbf{y}^p, s_\theta^p) + \underbrace{\frac{1 - \gamma}{|\mathcal{U}|} \sum_{q \in \mathcal{U}} \sum_k \mathbf{C}(a^{q,k}, \hat{a}^{q,k})}_{\text{CE on supervised images (i.e., source)}}$$

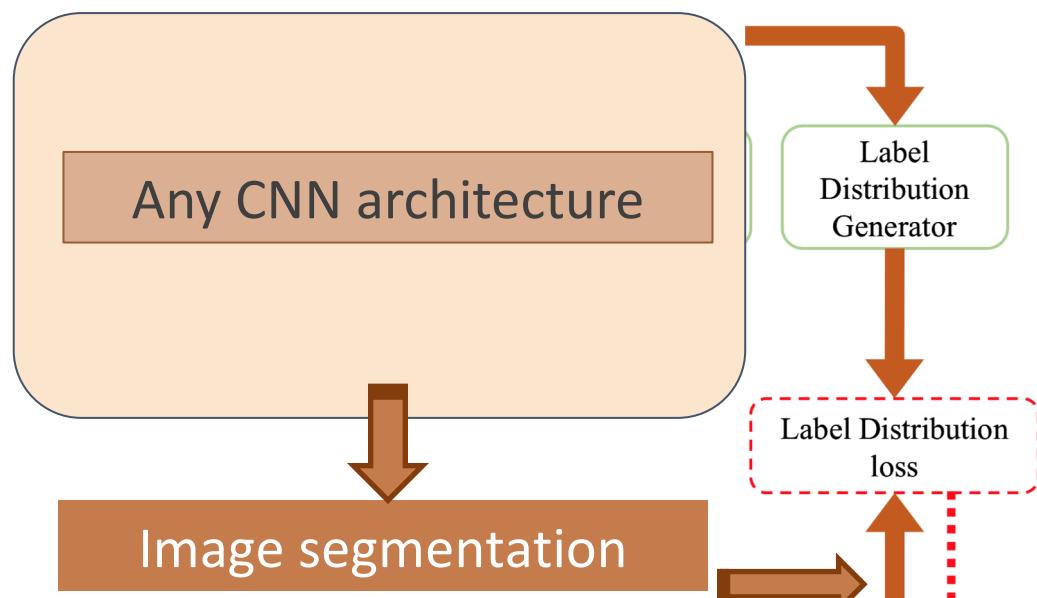
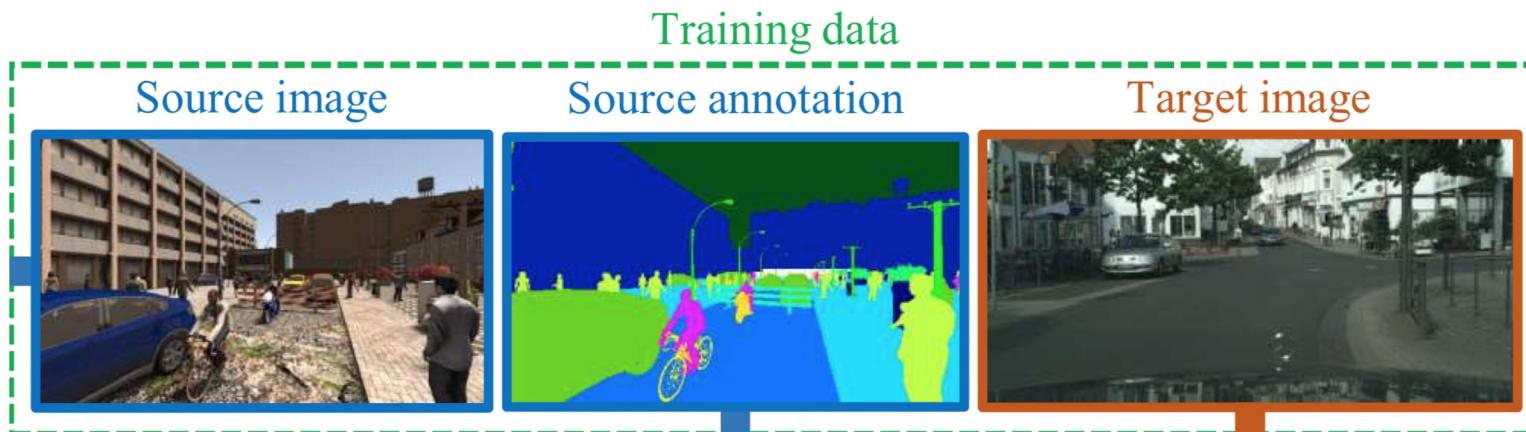
CE on supervised
images (i.e., source)

Additional term

$$\frac{1 - \gamma}{|\mathcal{U}|} \sum_{q \in \mathcal{U}} \sum_k \mathbf{C}(a^{q,k}, \hat{a}^{q,k})$$

Constrained optimization (in CNNs)

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



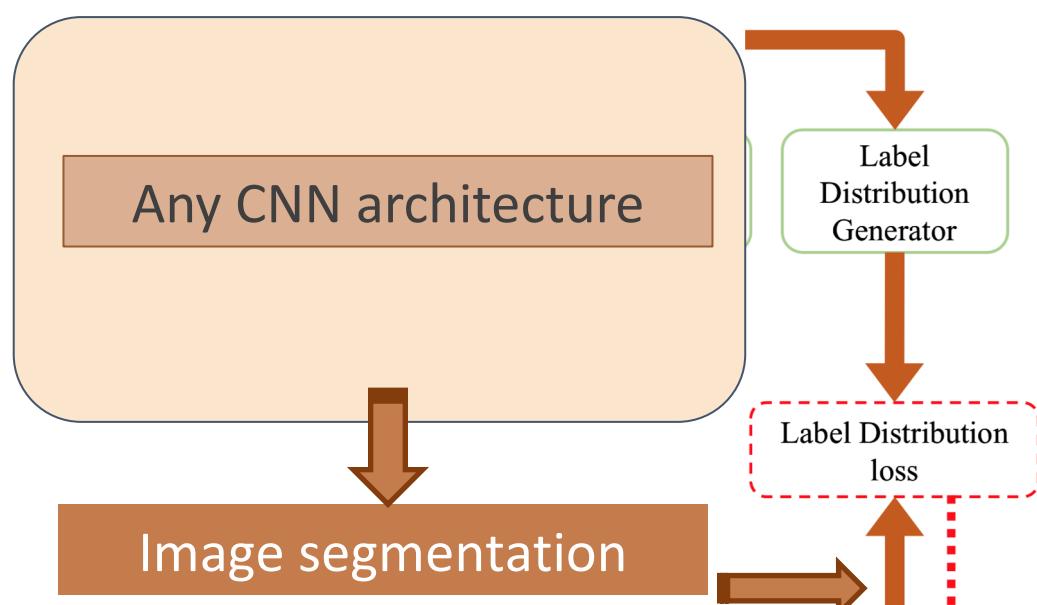
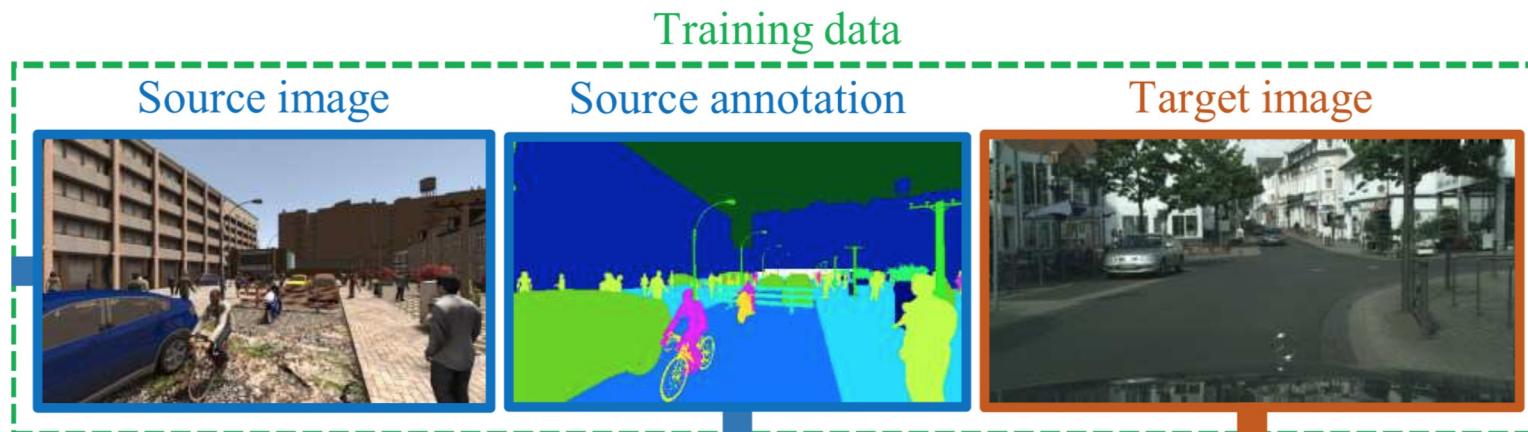
$$\min \frac{\gamma}{|\mathcal{L}|} \sum_{p \in \mathcal{L}} l(\mathbf{y}^p, s_\theta^p) + \frac{1 - \gamma}{|\mathcal{U}|} \sum_{q \in \mathcal{U}} \sum_k \mathbf{C}(a^{q,k}, \hat{a}^{q,k})$$

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

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

Constrained optimization (in CNNs)

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



$$\min \frac{\gamma}{|\mathcal{L}|} \sum_{p \in \mathcal{L}} l(\mathbf{y}^p, s_\theta^p) + \underbrace{\frac{1 - \gamma}{|\mathcal{U}|} \sum_{q \in \mathcal{U}} \sum_k \mathbf{C}(a^{q,k}, \hat{a}^{q,k})}_{\text{CE on supervised images (i.e., source)}}$$

Additional term

$$\frac{1 - \gamma}{|\mathcal{U}|} \sum_{q \in \mathcal{U}} \sum_k \mathbf{C}(a^{q,k}, \hat{a}^{q,k})$$

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

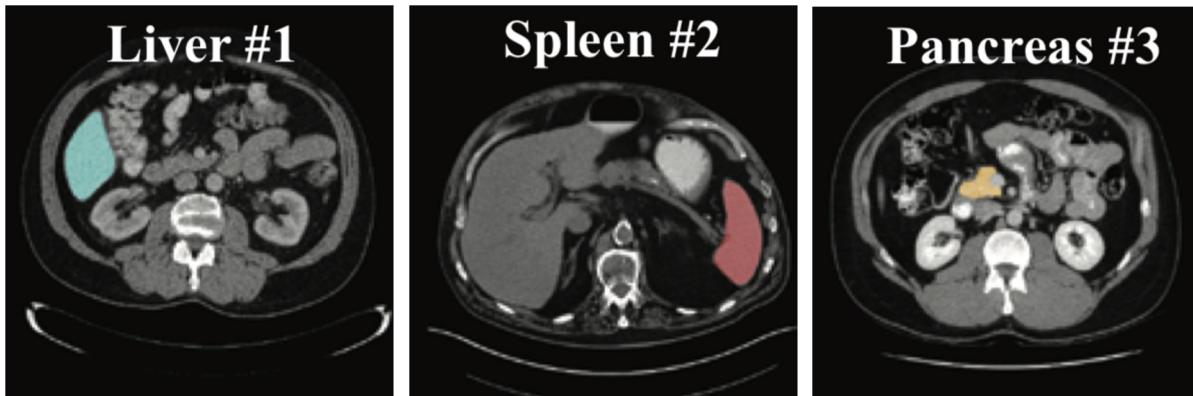
$$\mathbf{C}(\mathbf{a}^q, \hat{\mathbf{a}}^q) = H(\mathbf{a}^q) + KL(\mathbf{a}^q, \hat{\mathbf{a}}^q)$$

Predicted size

From predicted image

Constrained optimization (in CNNs)

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



Constrained optimization (in CNNs)

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

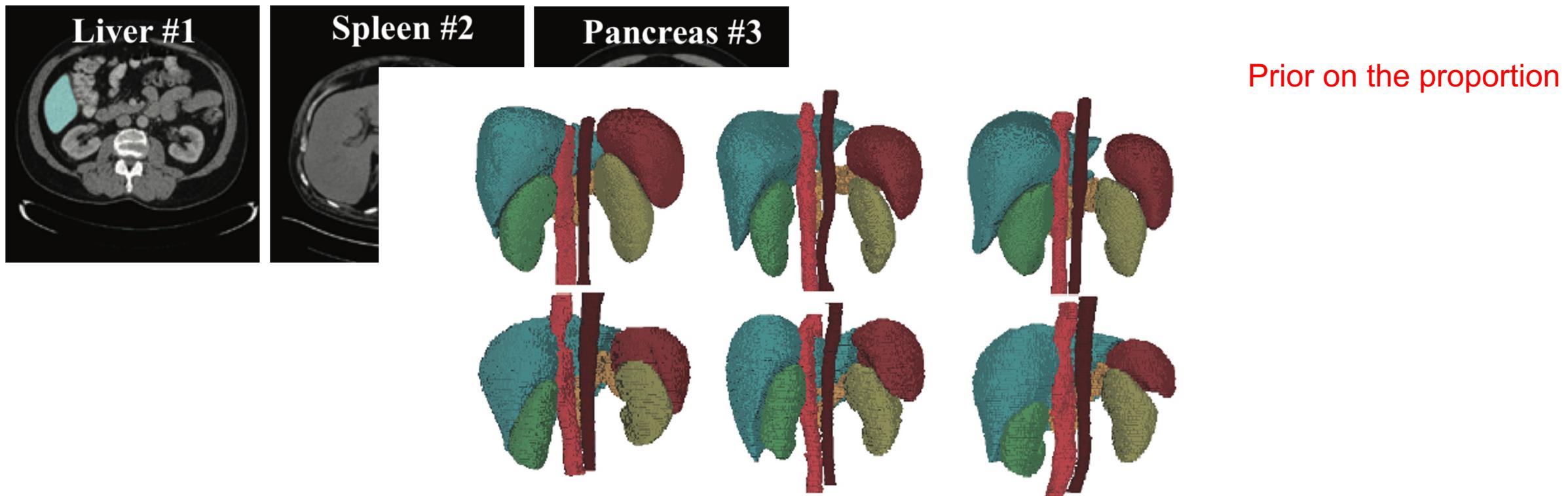
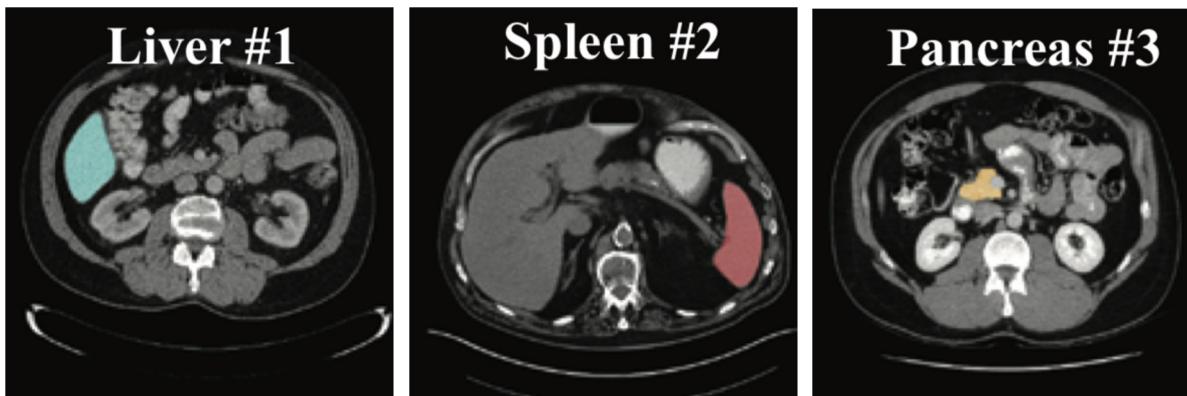


Figure 1. 3D Visualization of several abdominal organs (liver, spleen, left kidney, right kidney, aorta, inferior vena cava) to show the similarity of patient-wise abdominal organ size distributions.

Constrained optimization (in CNNs)

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



Main objective:

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

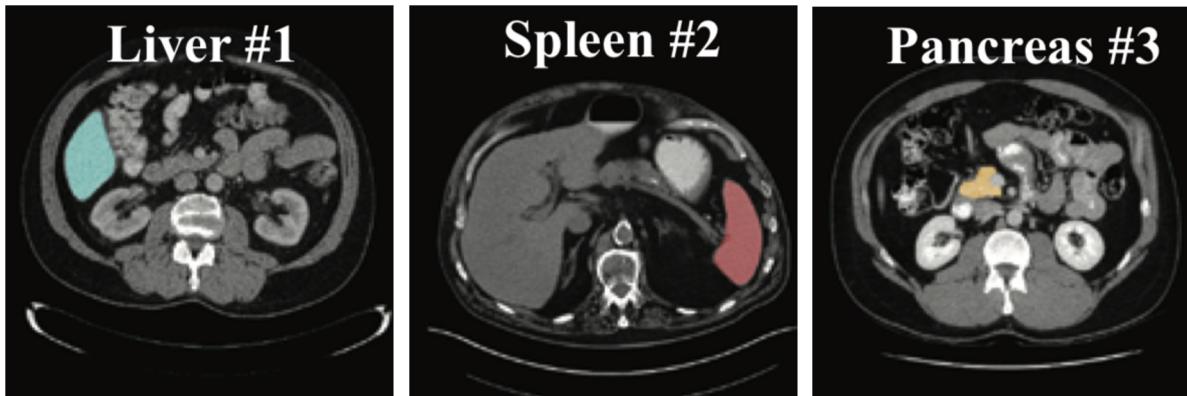
Fully labeled images

Partially labeled images

Prior-aware loss

Constrained optimization (in CNNs)

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



Prior-aware loss

Averaged predicted distribution

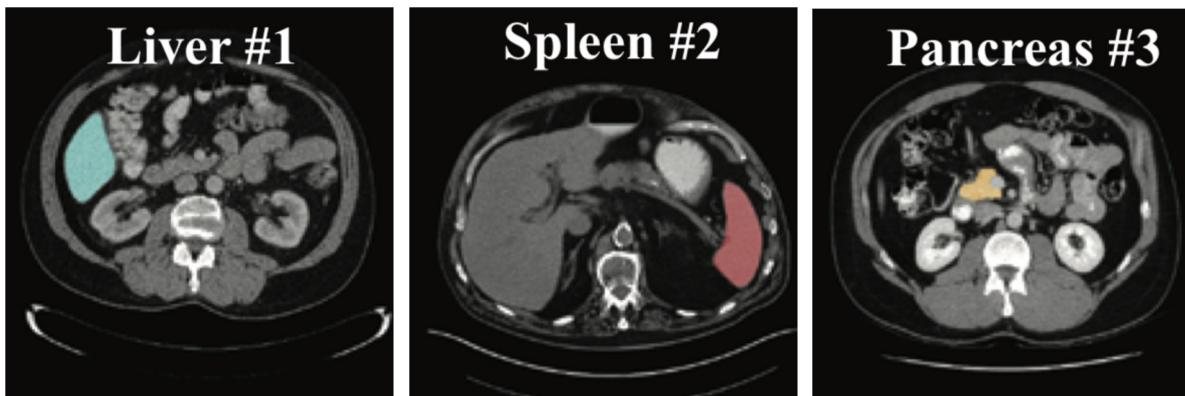
$$\hat{\mathbf{p}} = \frac{1}{N} \sum_{p \in \mathcal{P}} \mathbf{s}_{\theta}^p$$

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

On partially labeled images

Constrained optimization (in CNNs)

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



Prior-aware loss

Embed prior knowledge

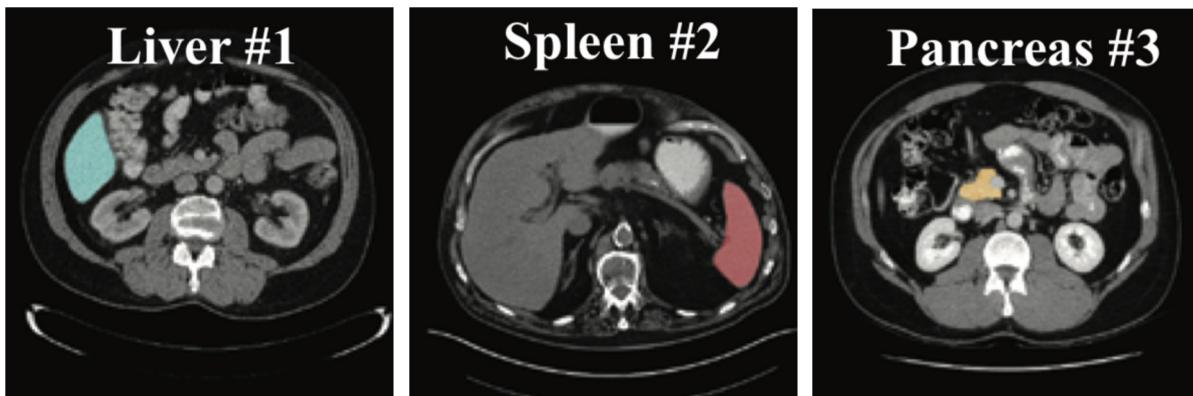
$$KL(\mathbf{q}|\hat{\mathbf{p}})$$



Real label distribution Average predicted distribution

Constrained optimization (in CNNs)

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



KL can be expanded

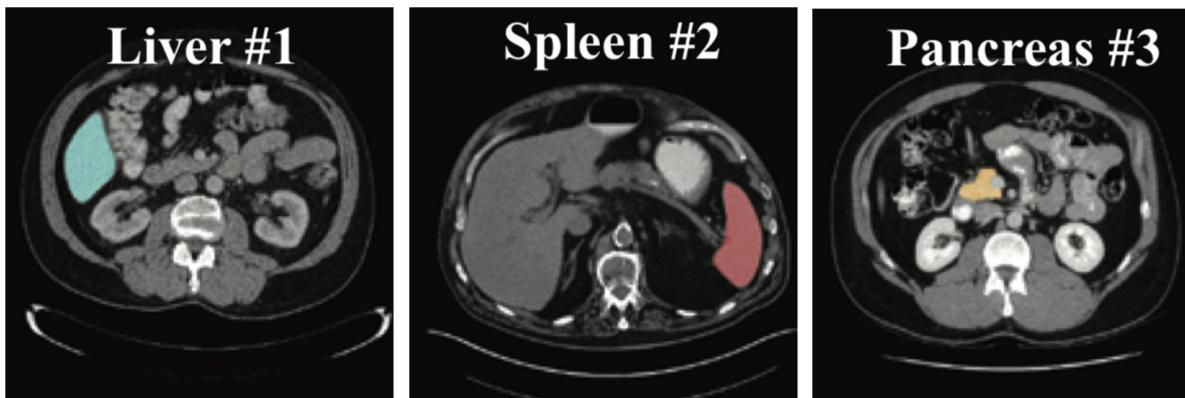
$$\sum_c KL(q^c|\hat{p}^c) = - \sum_c (q^c \log \hat{p}^c + ((1 - q^c) \log(1 - \hat{p}^c)) + const$$

Prior-aware loss

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

Constrained optimization (in CNNs)

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



Prior-aware loss

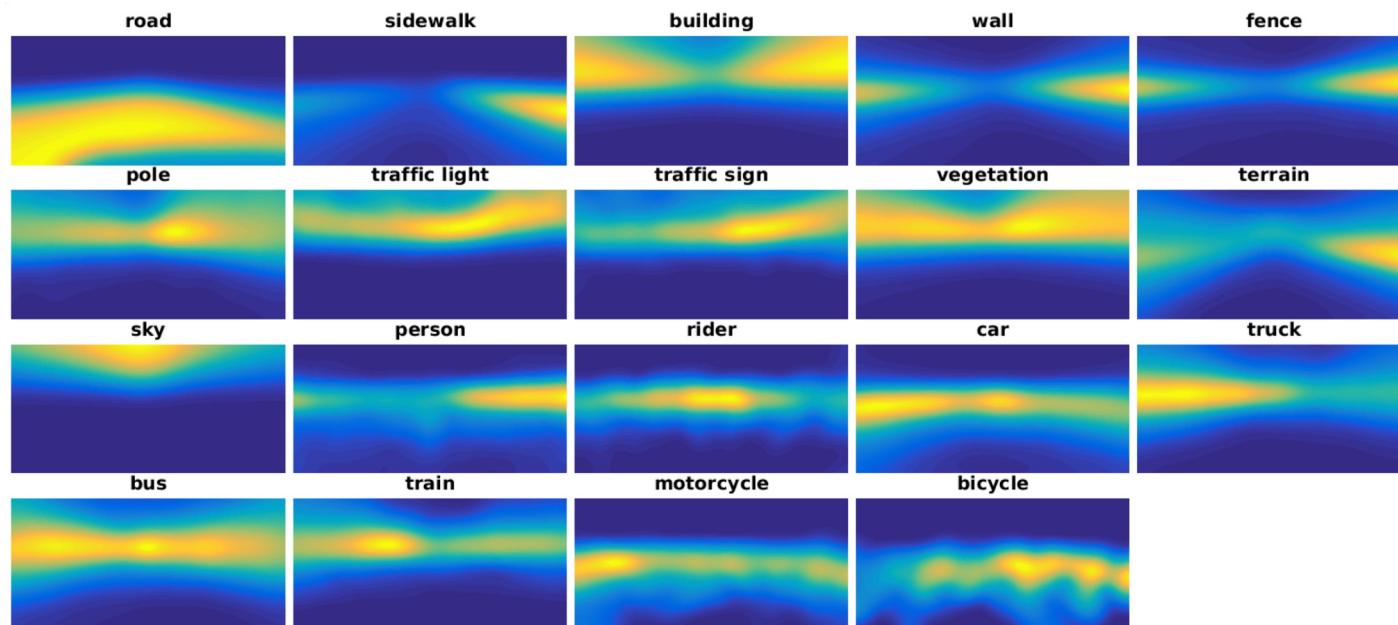
$$-\sum_{c=0}^{|K|} \left\{ q^c \log \frac{1}{N} \sum_{p \in \mathcal{P}} \mathbf{s}_{\theta}^{p,c} + (1 - q^c) \log \left(1 - \frac{1}{N} \sum_{p \in \mathcal{P}} \mathbf{s}_{\theta}^{p,c} \right) \right\} + const$$

KL can be expanded

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

Constrained optimization (in CNNs)

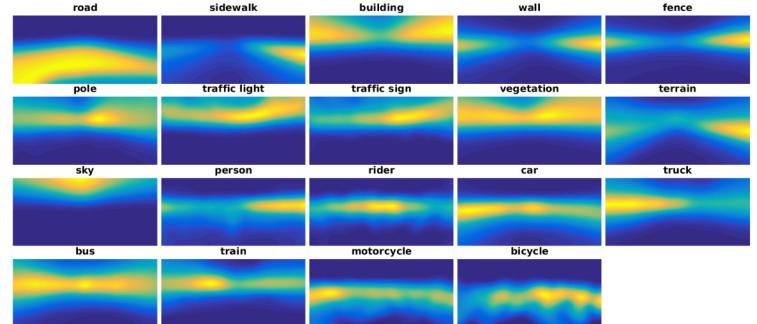
Equality constraints (at pixel-level)



Spatial priors on GTA5

Constrained optimization (in CNNs)

Equality constraints (at pixel-level)

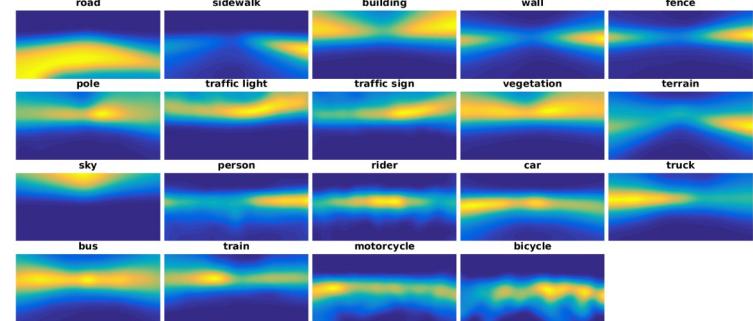
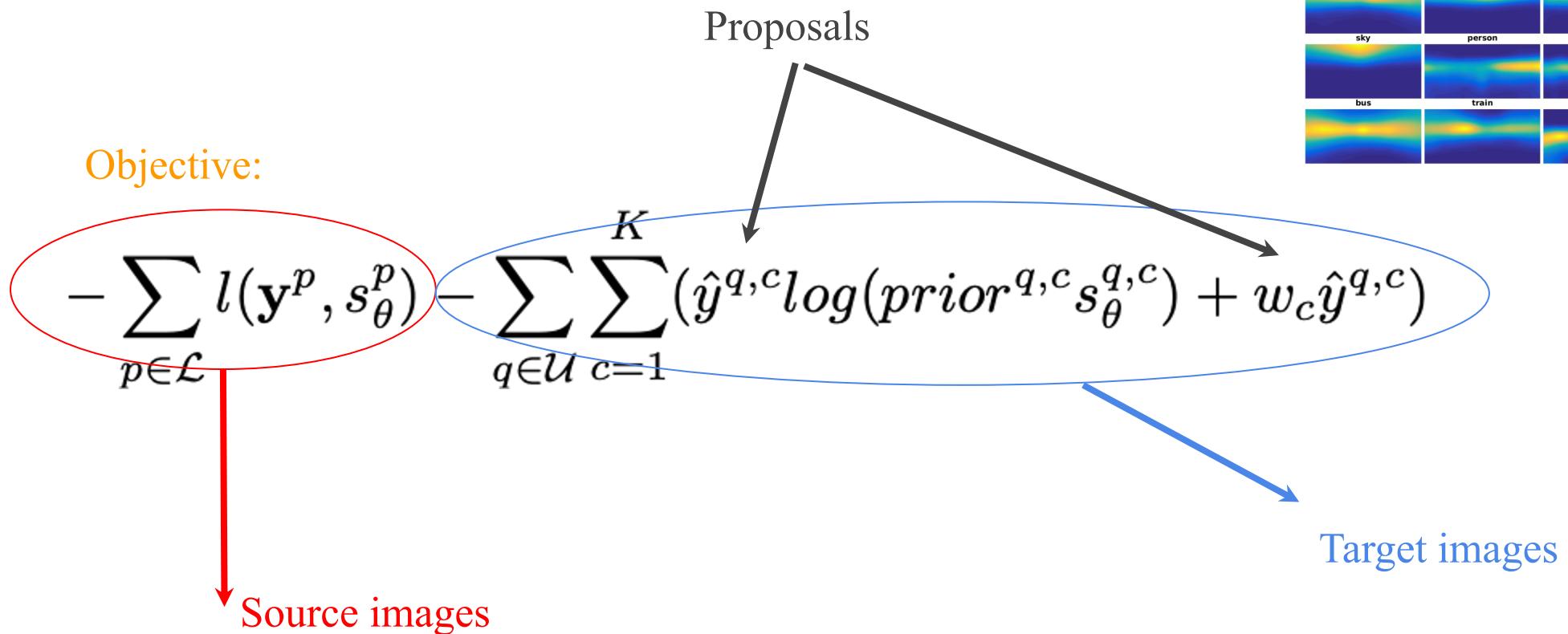


Objective:

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

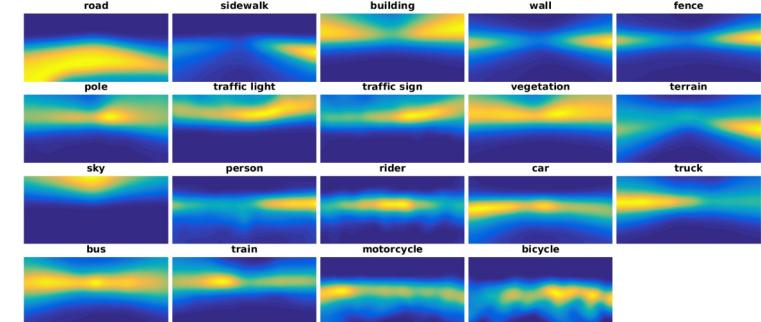
Constrained optimization (in CNNs)

Equality constraints (at pixel-level)



Constrained optimization (in CNNs)

Equality constraints (at pixel-level)



Objective:

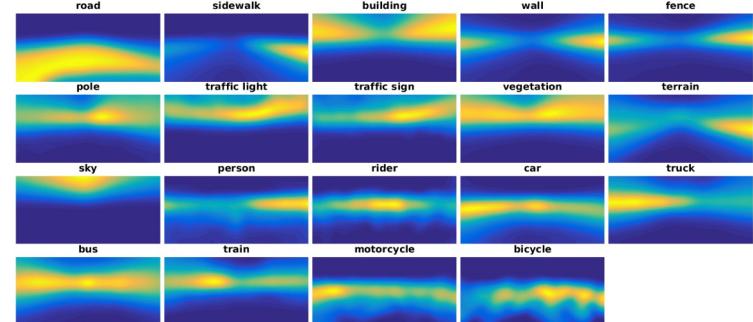
$$-\sum_{p \in \mathcal{L}} l(\mathbf{y}^p, s_\theta^p) - \sum_{q \in \mathcal{U}} \sum_{c=1}^K (\hat{y}^{q,c} \log(prior^{q,c} s_\theta^{q,c}) + w_c \hat{y}^{q,c})$$

This becomes two KL

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

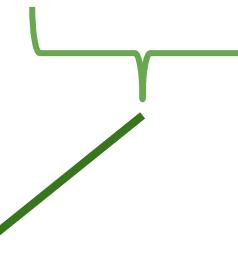
Constrained optimization (in CNNs)

Equality constraints (at pixel-level)



Objective:

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

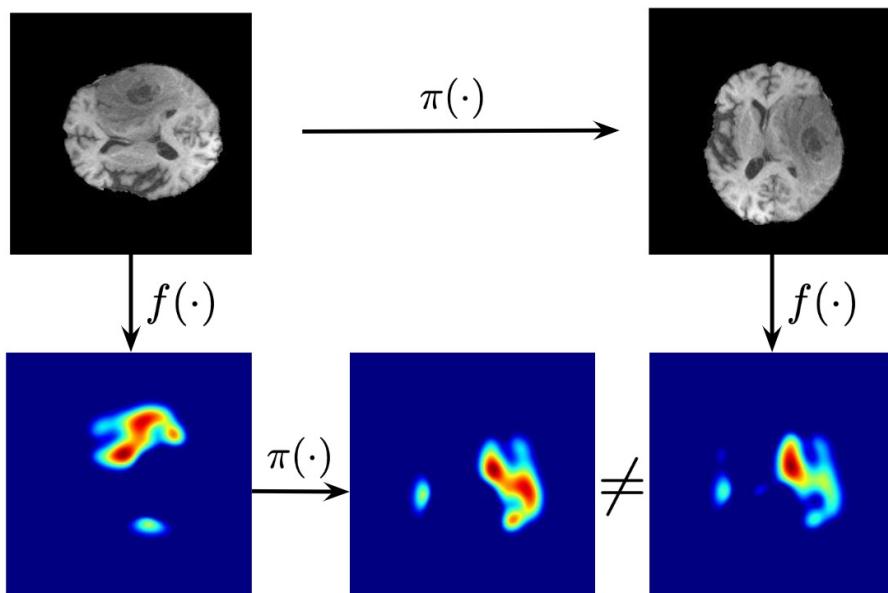


Weights the proposals

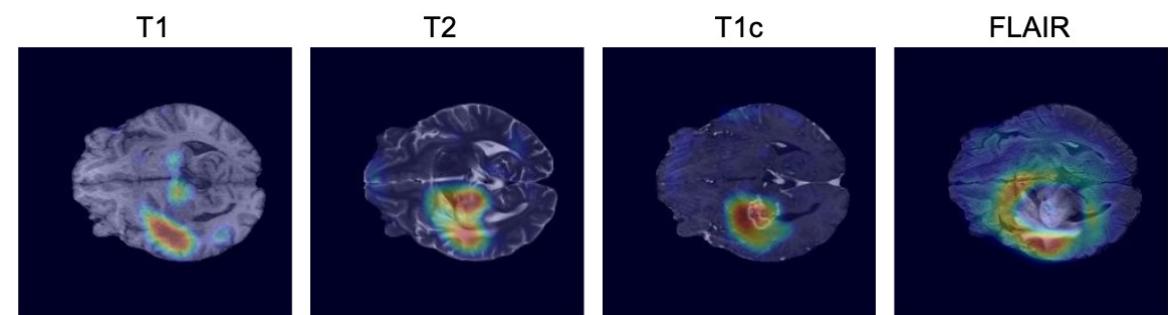
Constrained optimization (in CNNs)

Equality constraints (at pixel-level)

Imposing Consistency across image modalities



CAMs not equivariant to spatial
transformations



CAMs not consistent across modalities

Constrained optimization (in CNNs)

Equality constraints (at pixel-level)

Consistency regularization term on the CAMs and between modalities

Same-modality equivariant
constraints

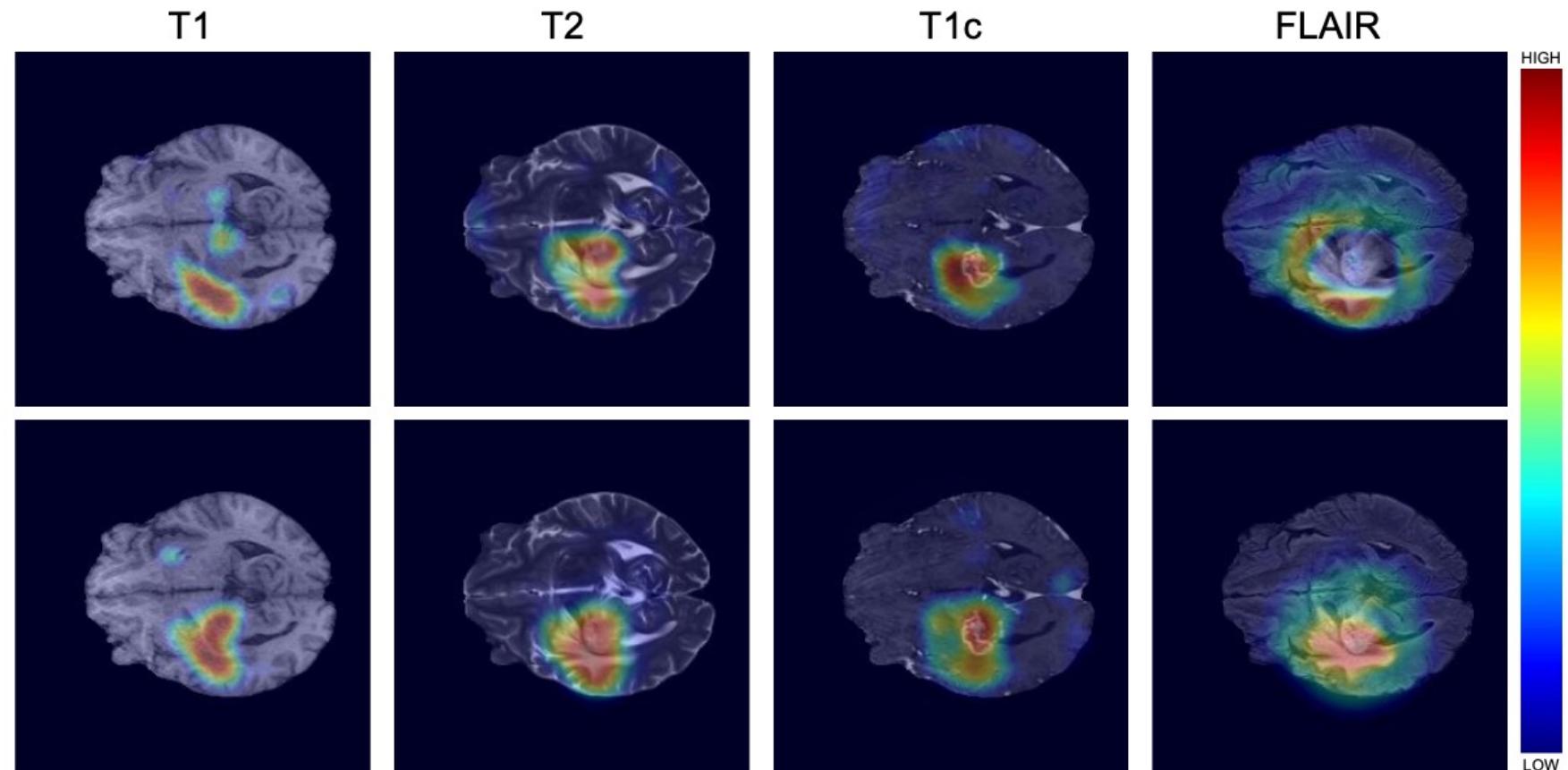
Cross-modality equivariant
constraints

$$\mathcal{L}_{class} + l_{reg}(T(f(M_m)), f(T(M_m))) + l_{reg}(T(f(M_m)), f(T(M_n)))$$

Constrained optimization (in CNNs)

Equality constraints (at pixel-level)

Baseline

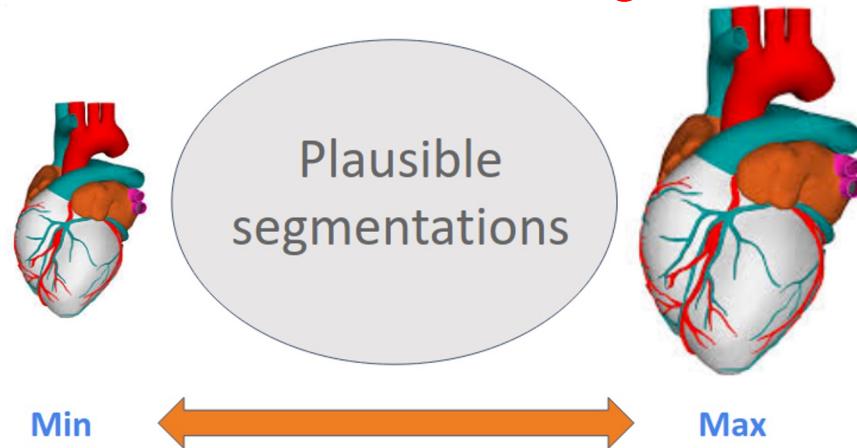


Baseline + equivariant
constraints

Constrained optimization (in CNNs)

Inequality constraints

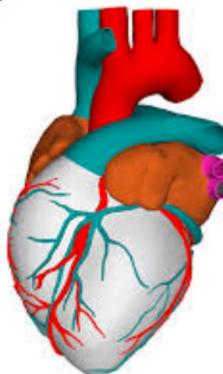
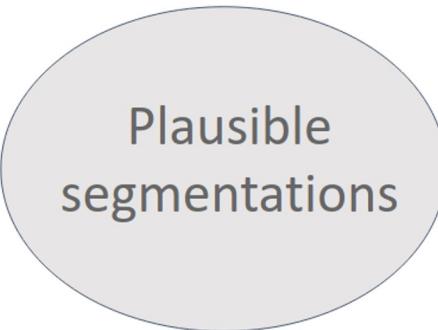
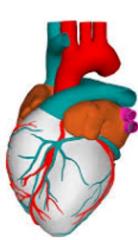
Prior size knowledge



Constrained optimization (in CNNs)

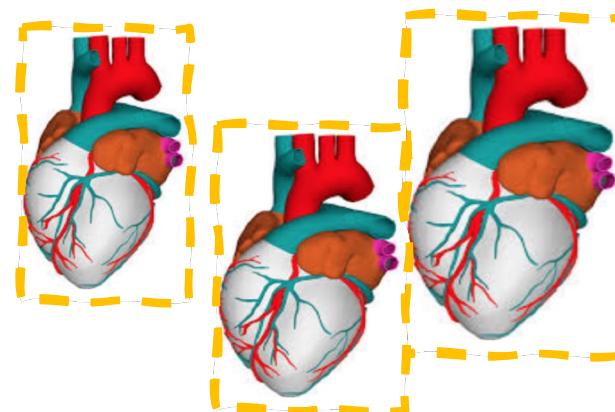
Inequality constraints

Prior size knowledge



Min

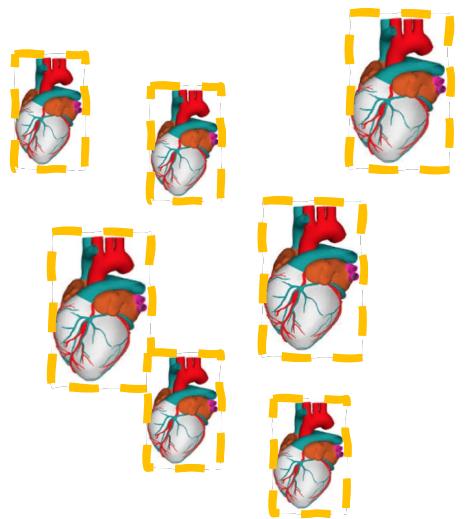
Max



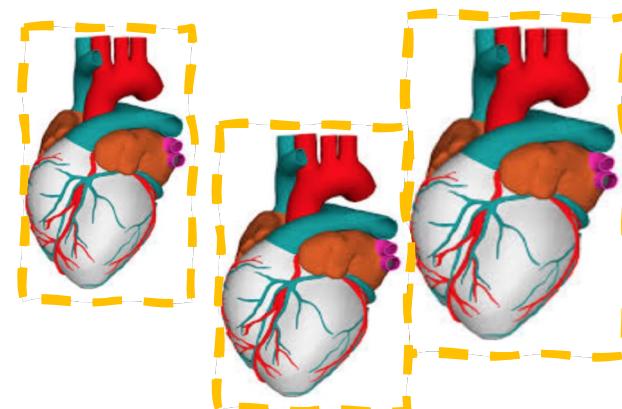
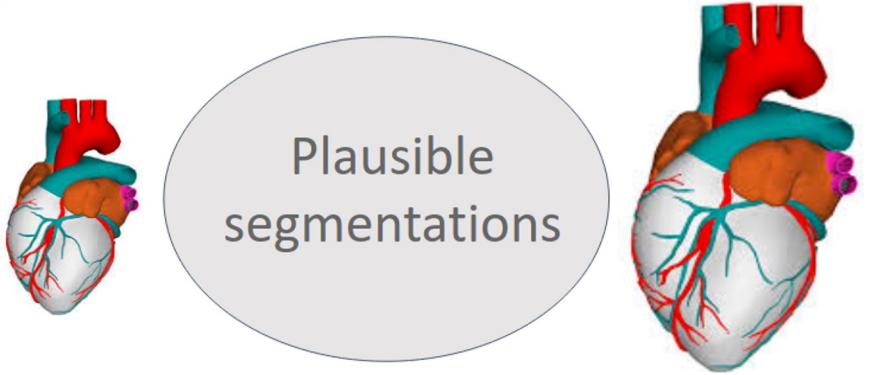
CNN predictions

Constrained optimization (in CNNs)

Inequality constraints



Prior size knowledge

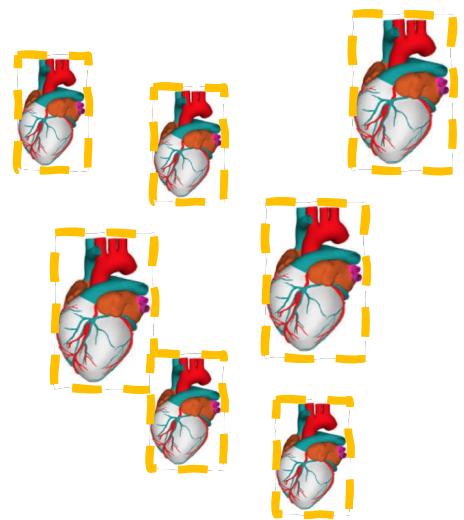


Smaller

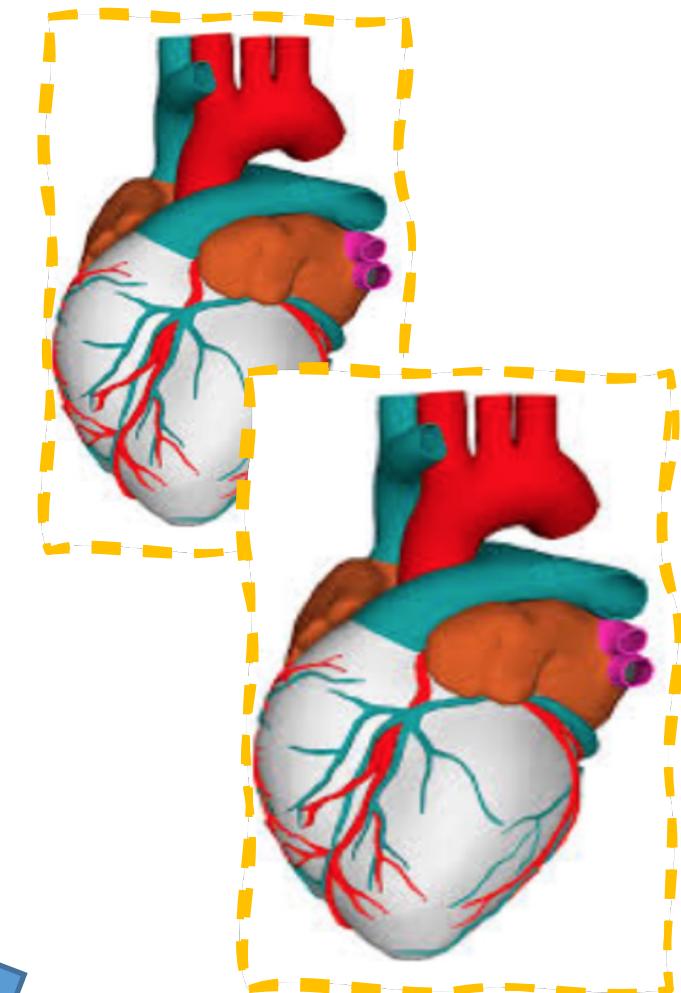
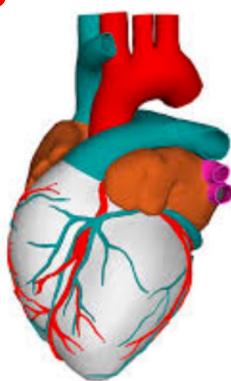
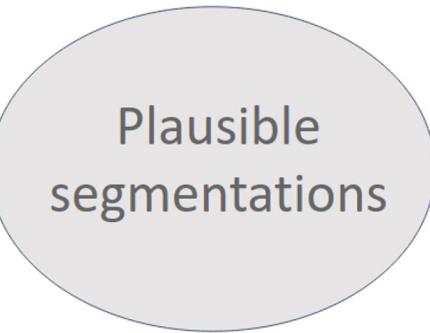
CNN predictions

Constrained optimization (in CNNs)

Inequality constraints



Prior size knowledge



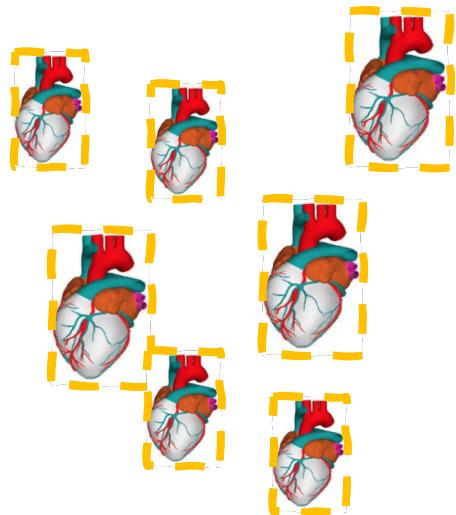
Smaller

CNN predictions

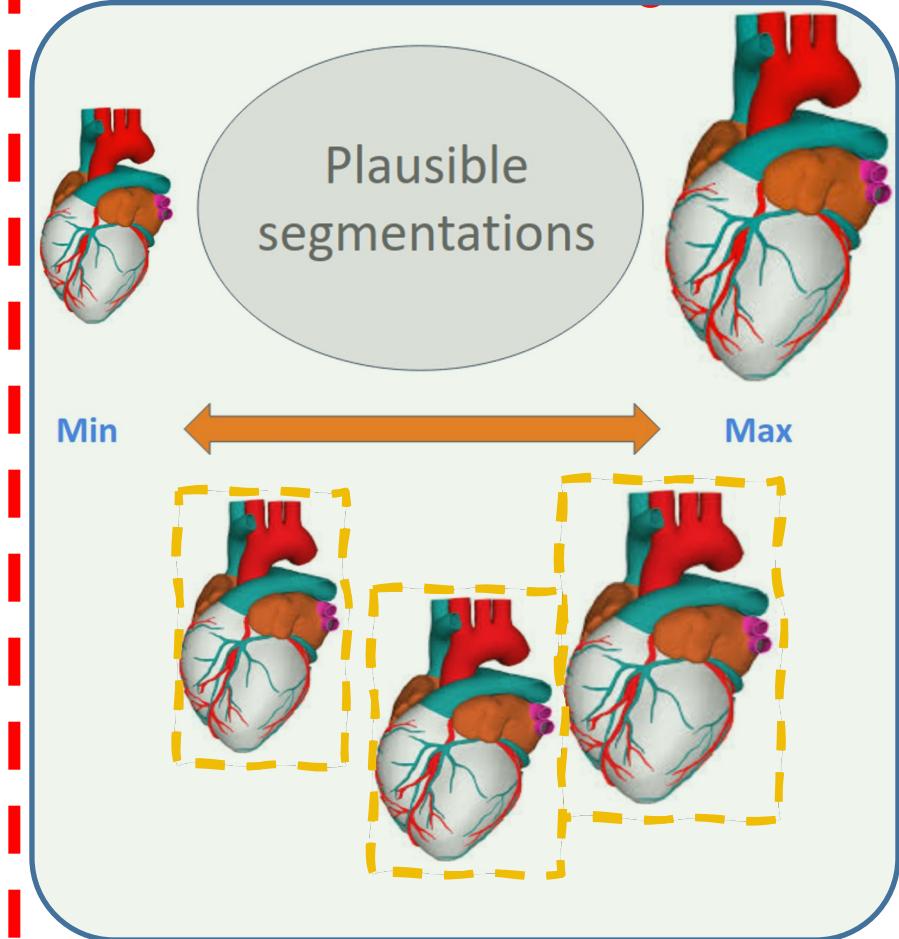
Larger

Constrained optimization (in CNNs)

Inequality constraints

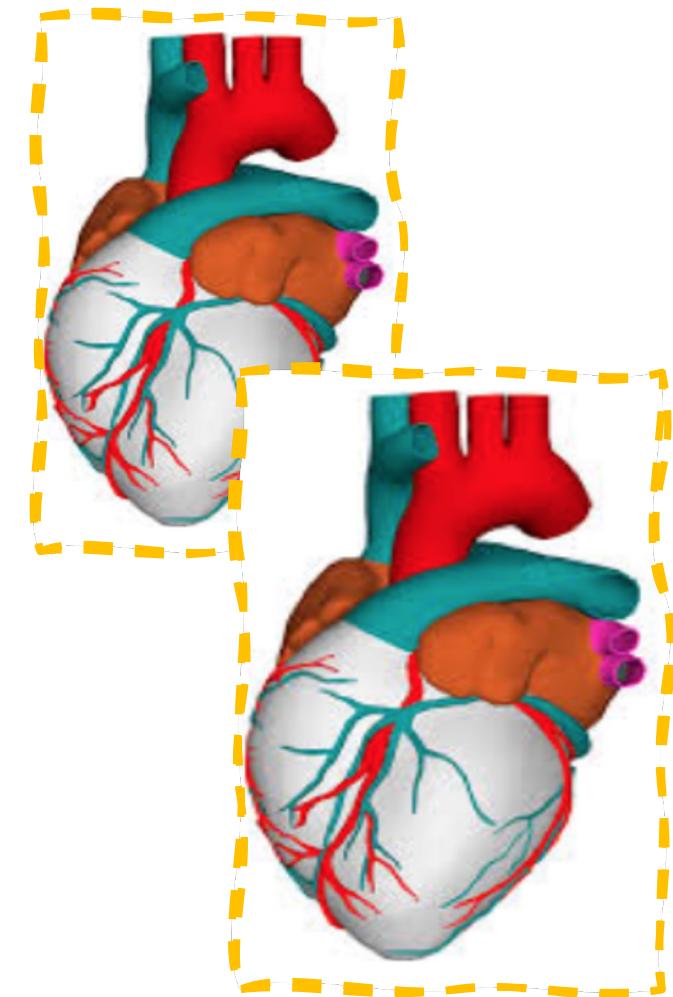


Prior size knowledge



Smaller

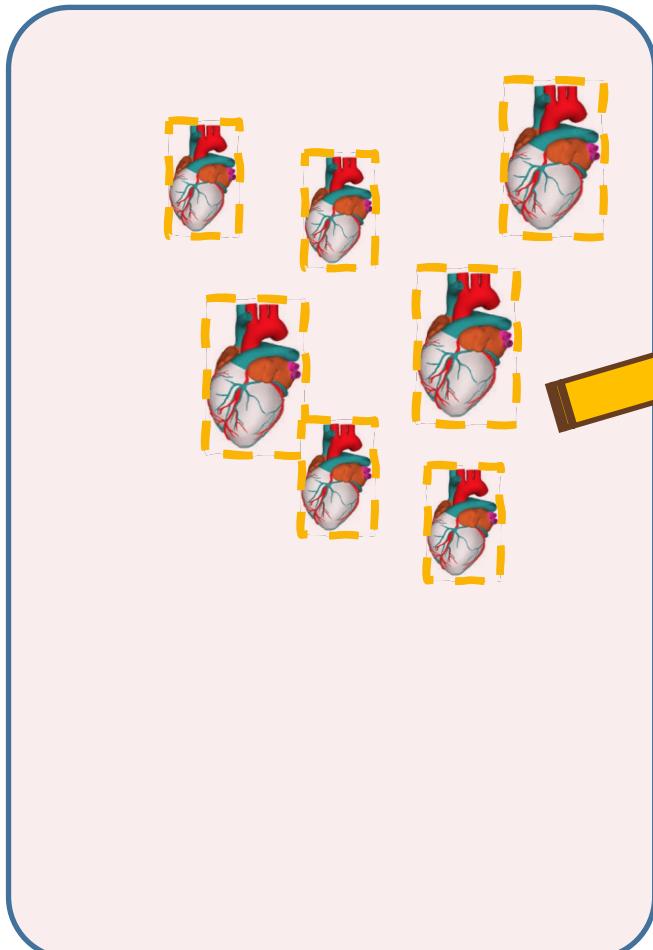
CNN predictions



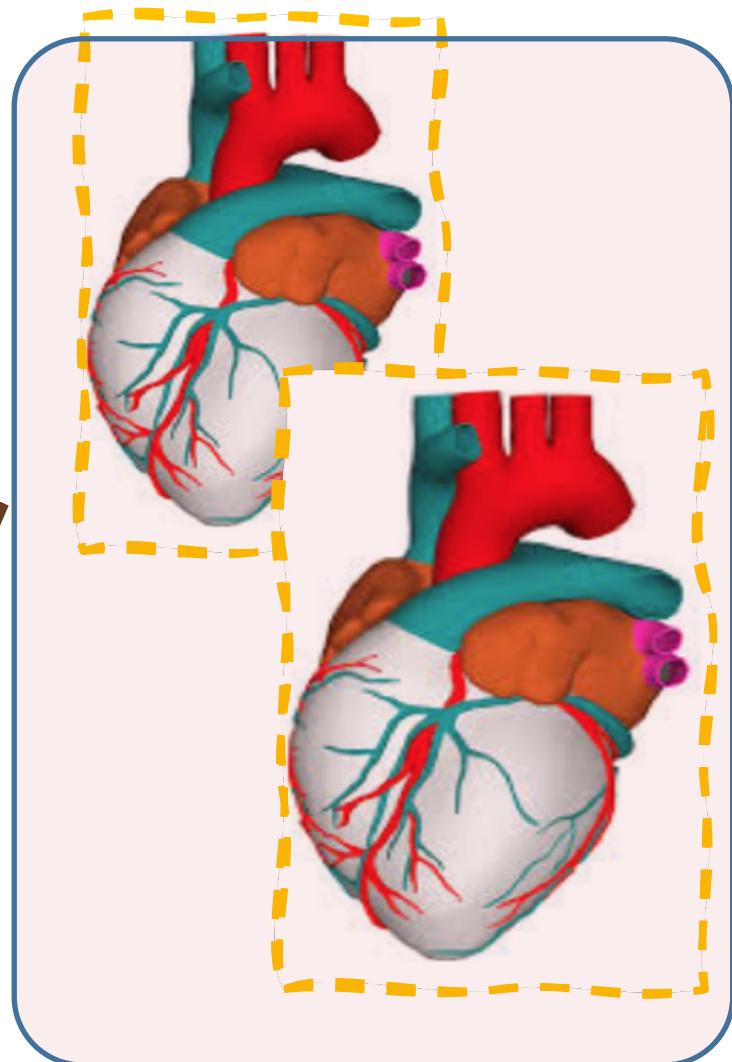
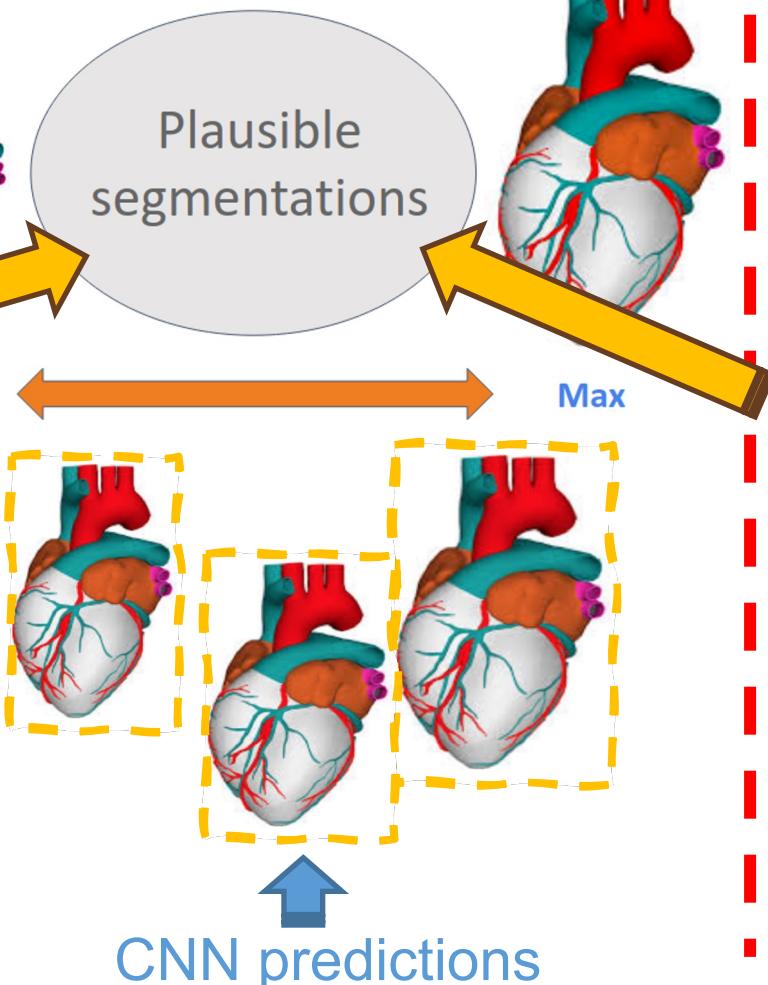
Larger

Constrained optimization (in CNNs)

Inequality constraints



Prior size knowledge

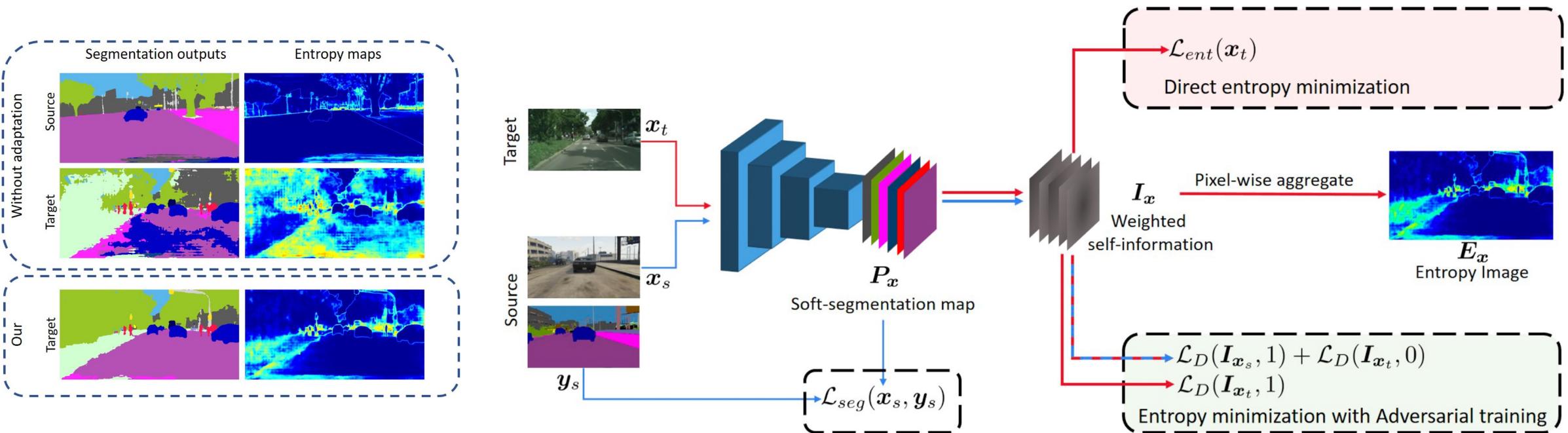


Smaller

Larger

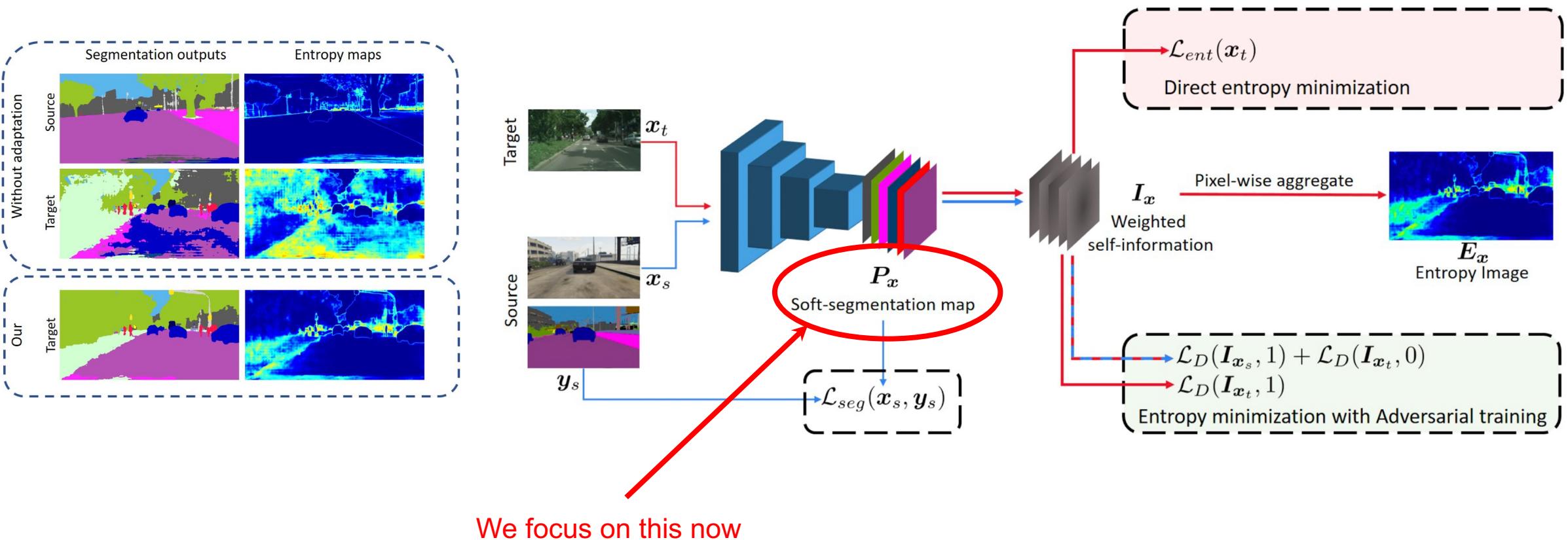
Constrained optimization (in CNNs)

Inequality constraints



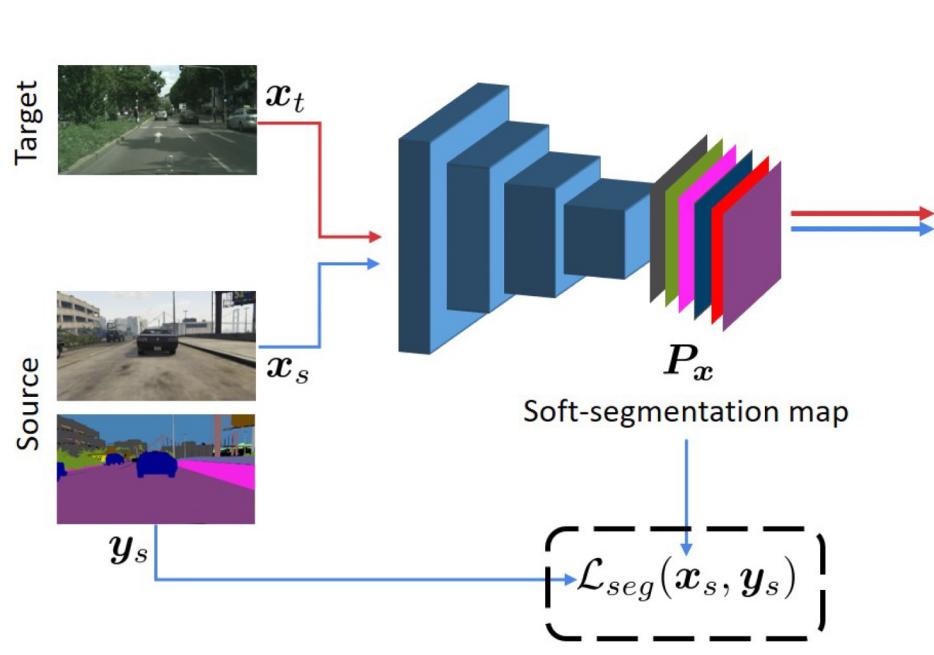
Constrained optimization (in CNNs)

Inequality constraints



Constrained optimization (in CNNs)

Inequality constraints



Class-ratio priors

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

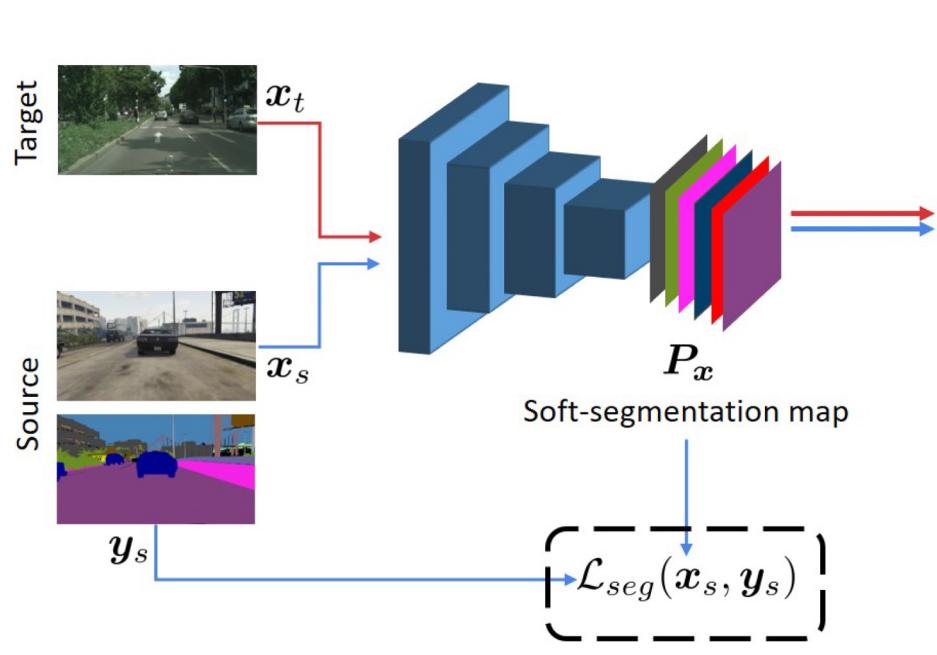
It relaxes the class prior constraint

$\ell 1$ -normalized histogram (source)

Estimated size on the prediction

Constrained optimization (in CNNs)

Inequality constraints



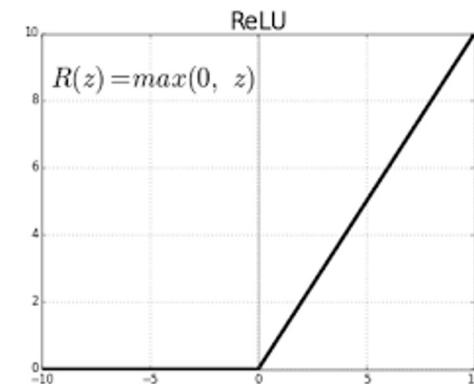
Class-ratio priors

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

It relaxes the class prior constraint

Estimated size on the prediction

$\ell 1$ -normalized histogram (source)



Constrained optimization (in CNNs)

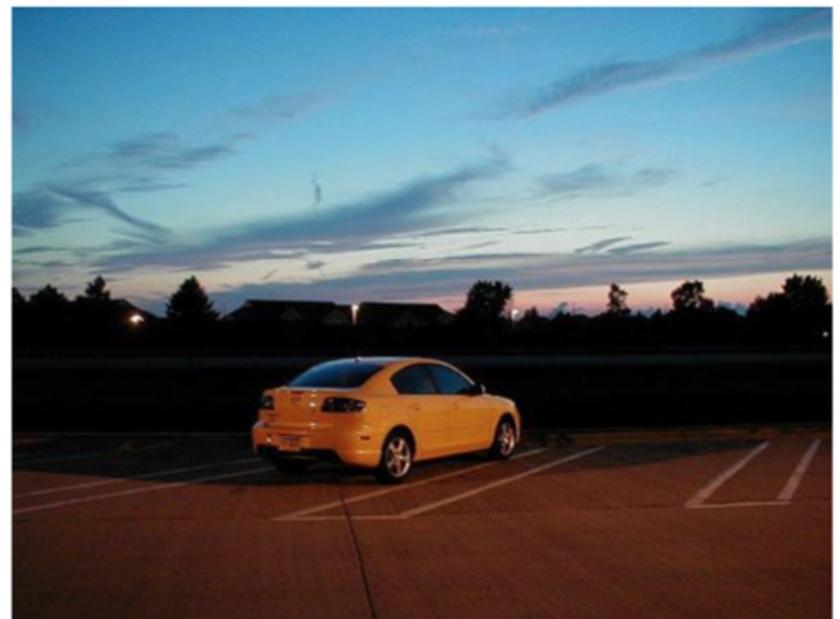
Inequality constraints

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”



Constrained optimization (in CNNs)

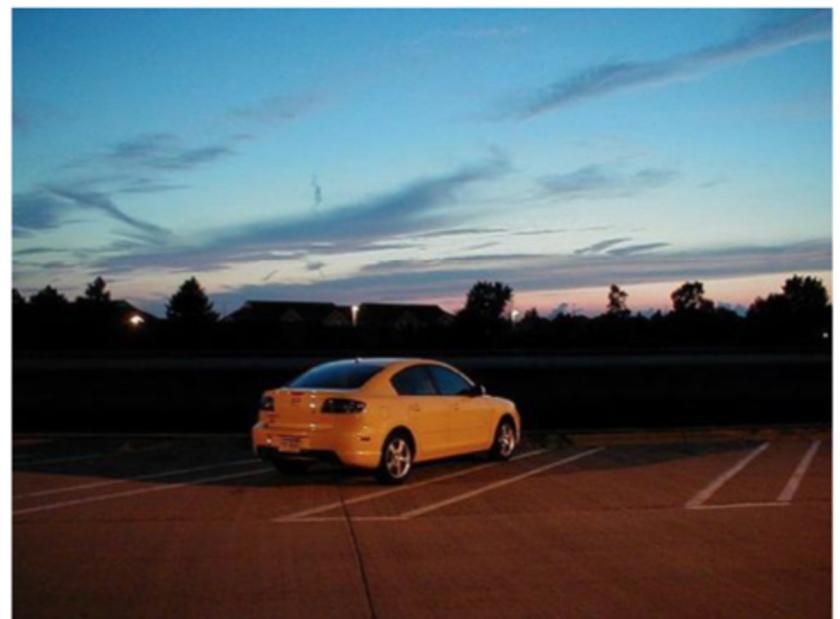
Inequality constraints

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”



Constrained optimization (in CNNs)

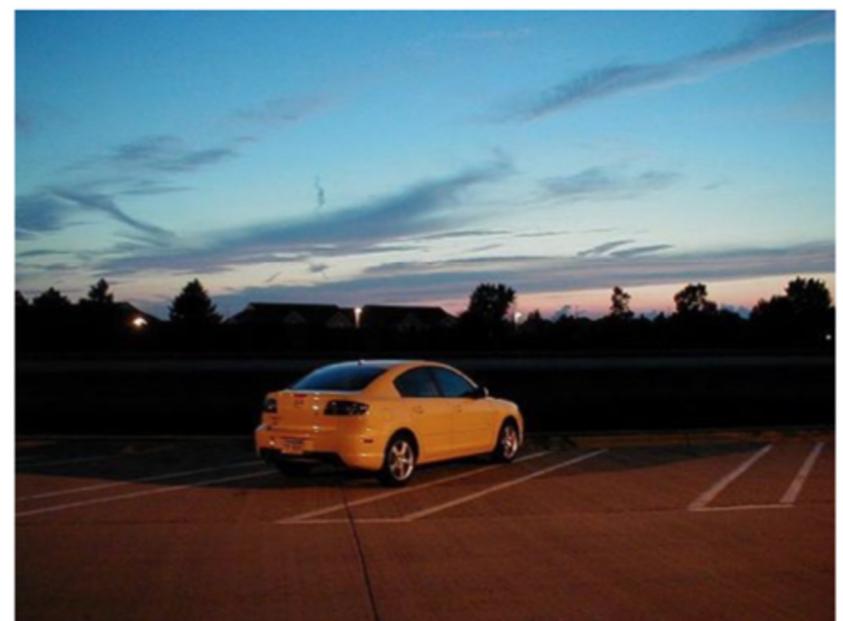
Inequality constraints

Information is given in
the form of image-tags

Target Size
 $a > 1$

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

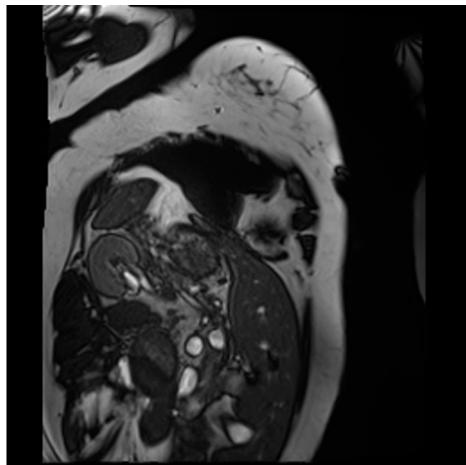
“Car”



Constrained optimization (in CNNs)

How we can benefit from this in the medical domain?

No cavity



Cavity

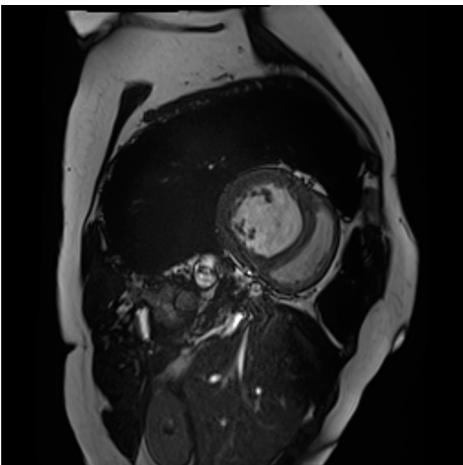


Image-tag information

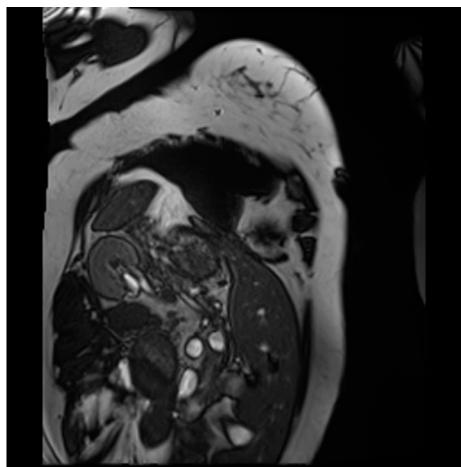
$$\sum_{p \in \Omega} s_{\theta}^{p,c} \leq 0$$

For negative image tags

Constrained optimization (in CNNs)

How we can benefit from this in the medical domain?

No cavity



Cavity

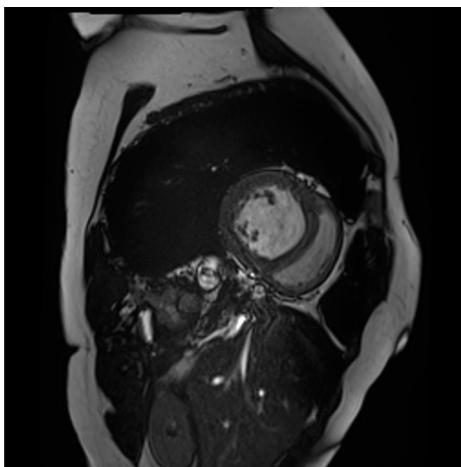
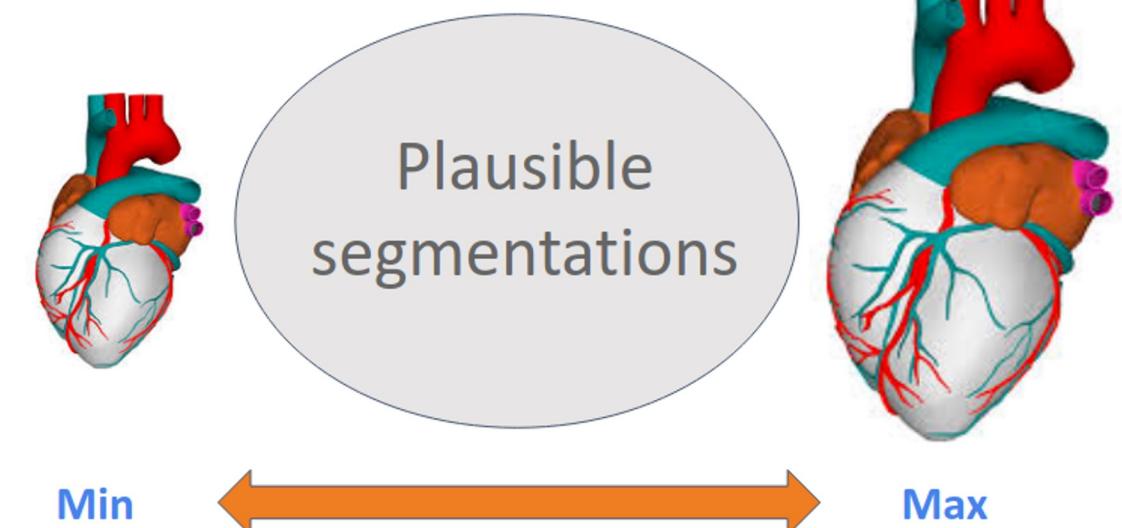


Image-tag information

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

For negative image tags

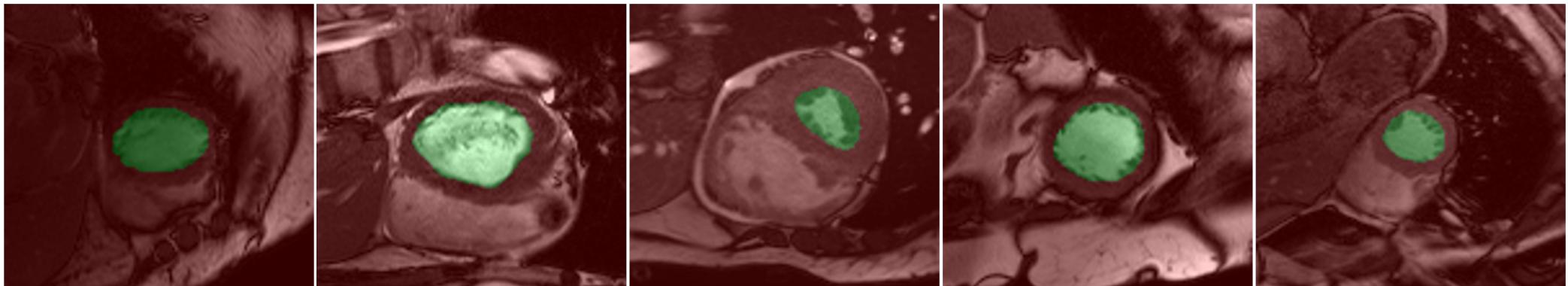


$$\min \leq \sum_{p \in \Omega} s_{\theta}^{p,c} \leq \max$$

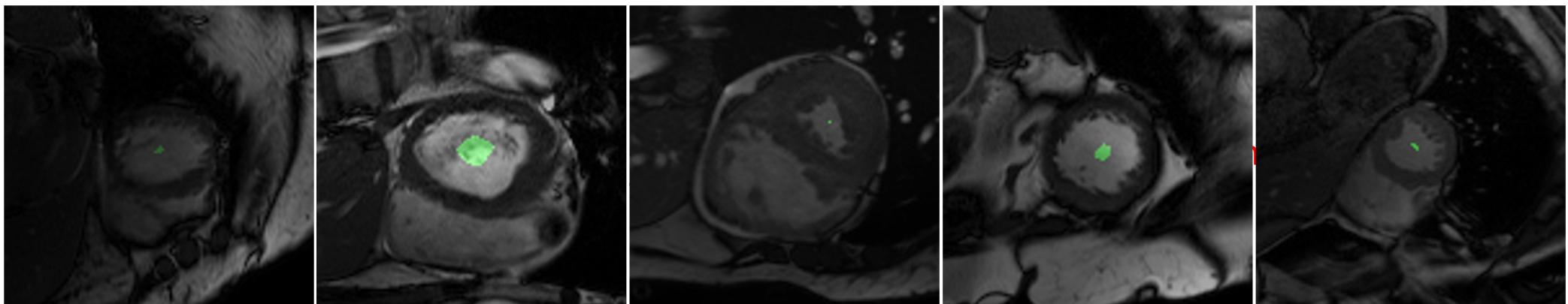
For positive image tags

Constrained optimization (in CNNs)

Inequality constraints (e.g, L2 penalty)



Full annotations



Partial annotations for cross-entropy

Constrained optimization (in CNNs)

Inequality constraints (e.g, L2 penalty)

Objective

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

Constrained optimization (in CNNs)

Inequality constraints (e.g, L2 penalty)

Objective

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

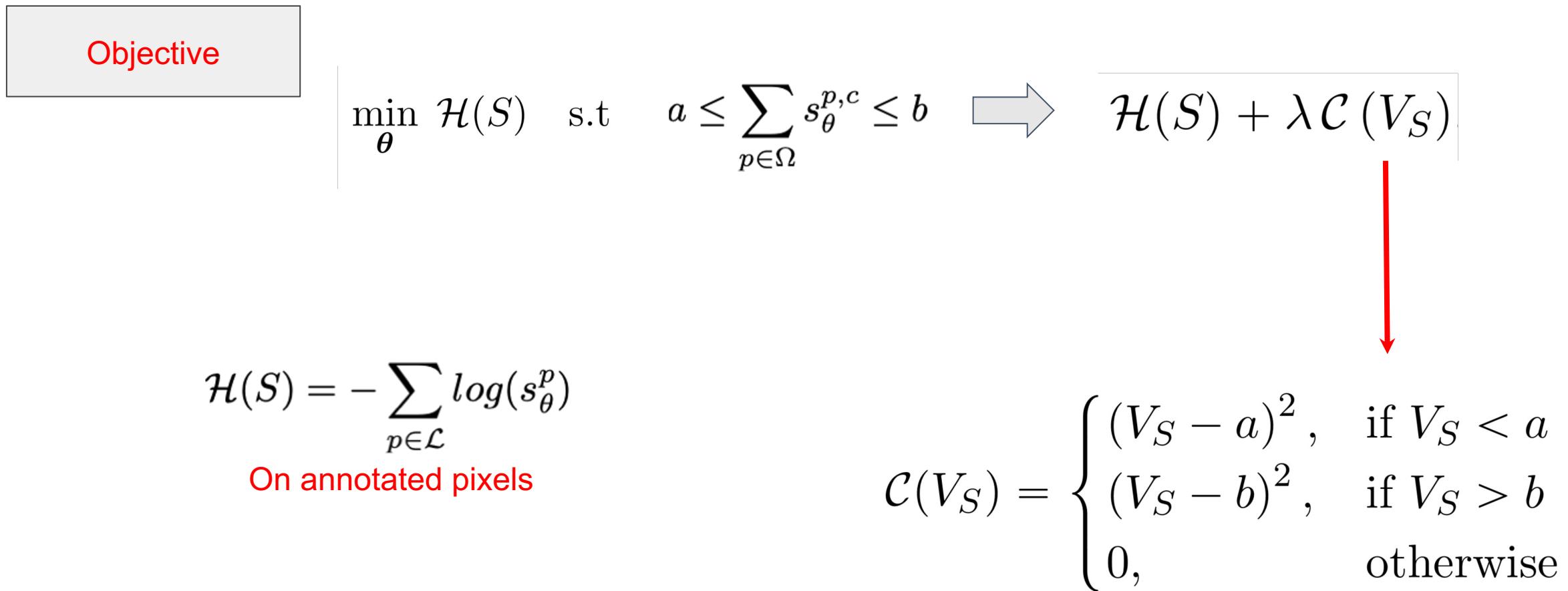
$\mathcal{H}(S) = - \sum_{p \in \mathcal{L}} \log(s_{\theta}^p)$

On annotated pixels

The diagram illustrates the derivation of the constrained optimization objective. It starts with the 'Objective' box containing the equation $\min_{\theta} \mathcal{H}(S)$. This is followed by a constraint $a \leq \sum_{p \in \Omega} s_{\theta}^{p,c} \leq b$, which is then combined with the objective function to form the final expression $\mathcal{H}(S) + \lambda \mathcal{C}(V_S)$. Below this, the formula for $\mathcal{H}(S)$ is given as $\mathcal{H}(S) = - \sum_{p \in \mathcal{L}} \log(s_{\theta}^p)$, with the note 'On annotated pixels' underneath. Red arrows indicate the flow from the definition of $\mathcal{H}(S)$ to its inclusion in the main equation, and from the 'On annotated pixels' note to the term $\sum_{p \in \mathcal{L}} \log(s_{\theta}^p)$.

Constrained optimization (in CNNs)

Inequality constraints (e.g, L2 penalty)



Constrained optimization (in CNNs)

Inequality constraints

But we can do more than simply the size

Shape moment $\mu_{p,q}^{(k)}(s_{\theta}) := \sum_{i \in \Omega} s_{\theta}^{(i,k)} x_{(i)}^p y_{(i)}^q,$

Central moment $\bar{\mu}_{p,q}^{(k)} := \sum_{i \in \Omega} s_{\theta}^{(i,k)} \left(x_{(i)} - \frac{\mu_{1,0}^{(k)}}{\mu_{0,0}^{(k)}} \right)^p \left(y_{(i)} - \frac{\mu_{0,1}^{(k)}}{\mu_{0,0}^{(k)}} \right)^q.$

Constrained optimization (in CNNs)

Inequality constraints

But we can do more than simply the size

Shape moment $\mu_{p,q}^{(k)}(s_{\theta}) := \sum_{i \in \Omega} s_{\theta}^{(i,k)} x_{(i)}^p y_{(i)}^q,$

Central moment $\bar{\mu}_{p,q}^{(k)} := \sum_{i \in \Omega} s_{\theta}^{(i,k)} \left(x_{(i)} - \frac{\mu_{1,0}^{(k)}}{\mu_{0,0}^{(k)}} \right)^p \left(y_{(i)} - \frac{\mu_{0,1}^{(k)}}{\mu_{0,0}^{(k)}} \right)^q.$

Constrained optimization (in CNNs)

Inequality constraints

But we can do more than simply the size

From shape and central moment

Volume

$$\mathfrak{V}^{(k)}(s_{\theta}) := \mu_{0,0}^{(k)}(s_{\theta}).$$

Centroid

$$\mathfrak{C}^{(k)}(s_{\theta}) := \left(\frac{\mu_{1,0}^{(k)}(s_{\theta})}{\mu_{0,0}^{(k)}(s_{\theta})}, \frac{\mu_{0,1}^{(k)}(s_{\theta})}{\mu_{0,0}^{(k)}(s_{\theta})} \right).$$

Length

$$\mathfrak{L}^{(k)}(s_{\theta}) := \sum_{i,j \in \mathcal{G}_{\Omega}} |s_{\theta}^{(i,k)} - s_{\theta}^{(j,k)}| L_{\Omega,i,j}.$$

Laplacian

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)

Thank you!