ÉTS
Le génie pour l'industrie

Erasmus MC
University Medical Center Rotterdam

UC SANTA CRUZ

Learning with Limited Supervision

Yuyin Zhou (Yan Wang)
Ismail Ben Ayed
Jose Dolz
Christian Desrosiers
Marleen de Bruijne
Hoel Kervadec

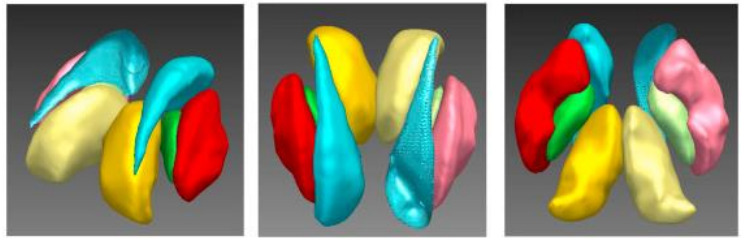
Why are we doing this tutorial at MICCAI?
(for the 4th time)

Deep CNNs are dominating computer vision

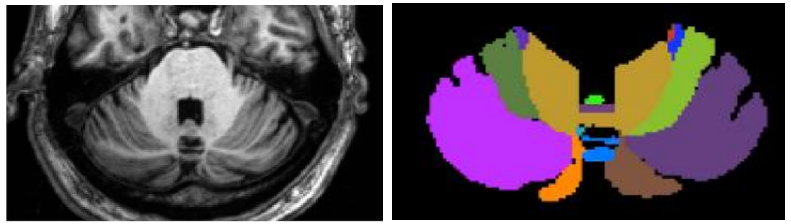
e.g., semantic segmentation

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

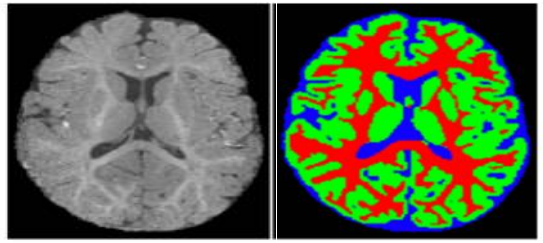
… and medical image analysis



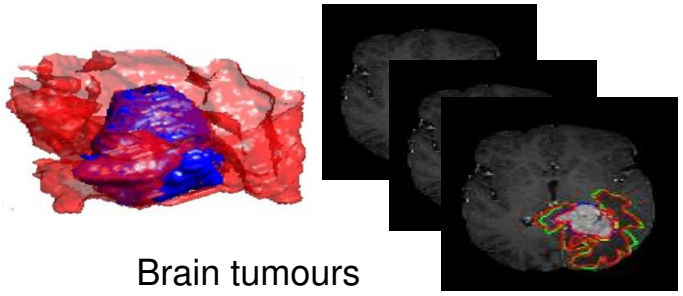
Subcortical structures
(Dolz et al., Neuroimage 2018)



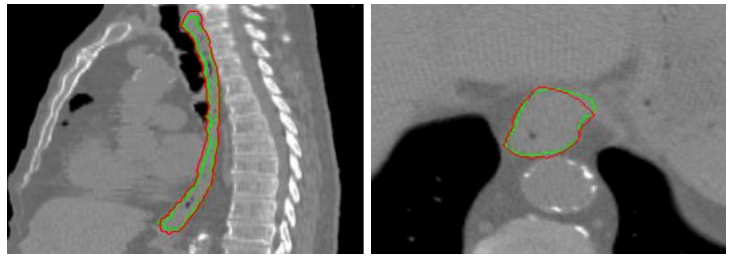
Cerebellum parcellation
(Carass et al., Neuroimage 2018)



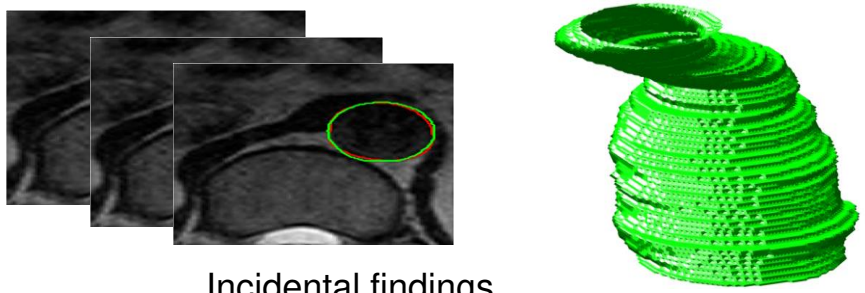
Brain tissues (6-month infant)
(Li et al., TMI 2019)



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

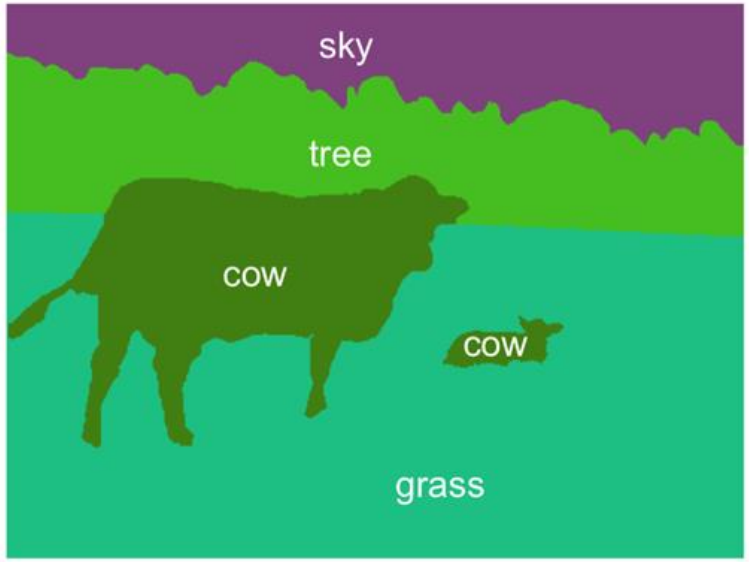
- more than 1h per image (even several hours for a medical image)
- Bottleneck for learning at large scale

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

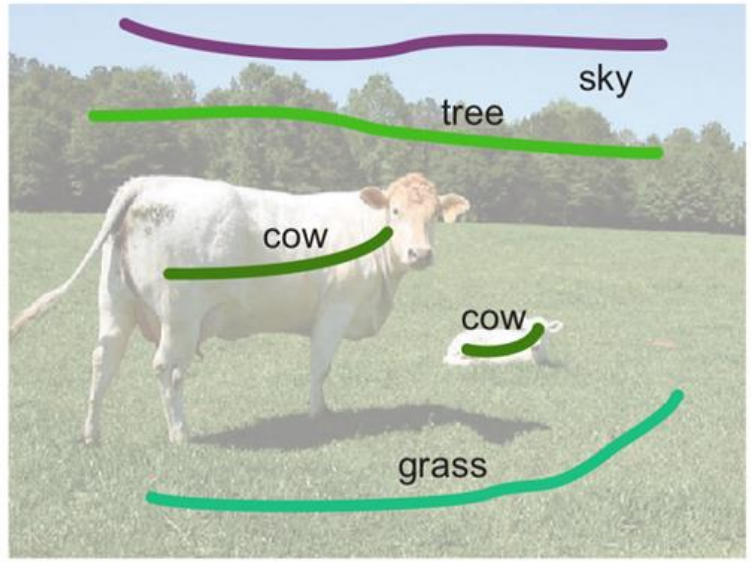
- 1s per label per image
- Scalable for large numbers of labels

person
horse
background

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



Full annotations



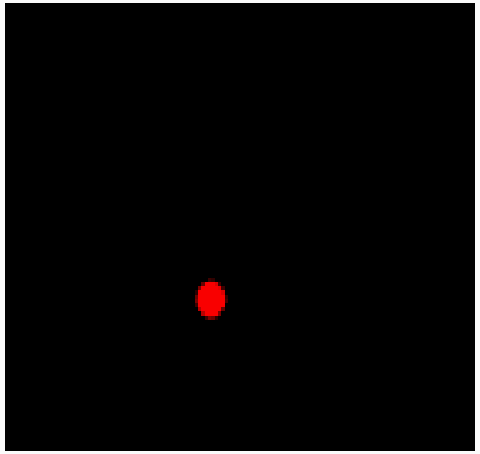
Semi-supervised

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

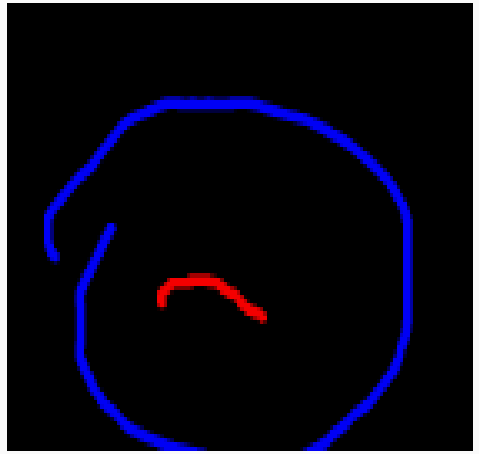
Forms of semi/weak supervision: Examples in segmentation

Car
Parking
Sky
No person

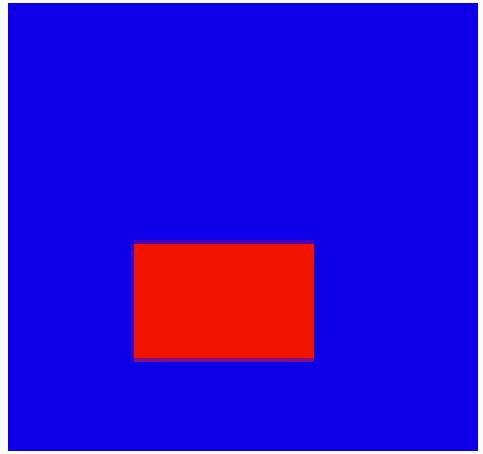
Image tags



points



scribbles

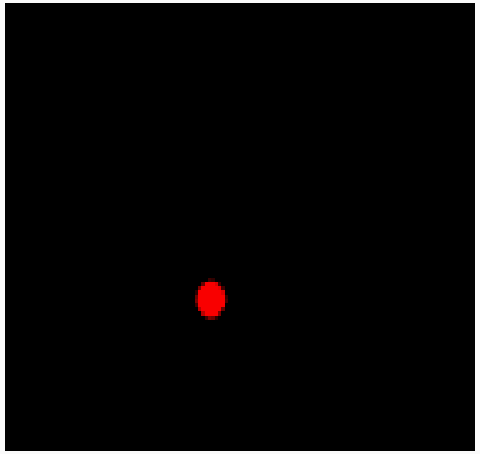


boxes

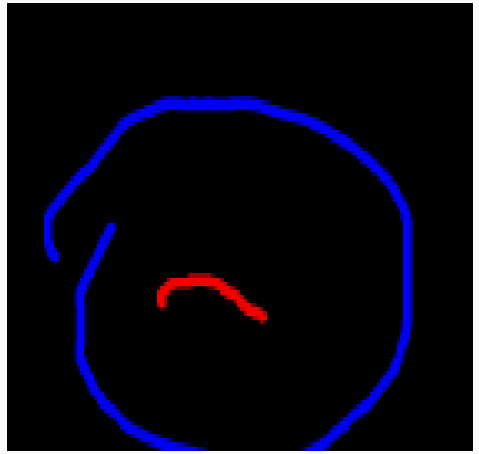
Forms of semi/weak supervision: Examples in segmentation

Car
Parking
Sky
No person

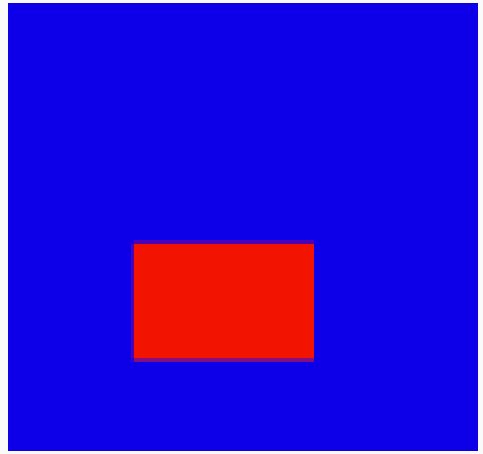
Image tags



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 2019]
[Kervadec et al., MIDL 2020]

Full annotations are much more problematic in medical imaging

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

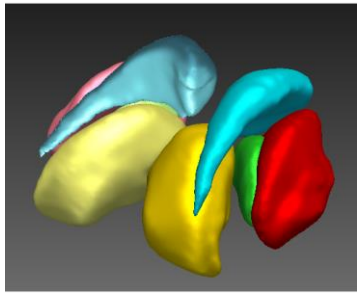
Crowdsourcing?

Full annotations are much more problematic in medical imaging

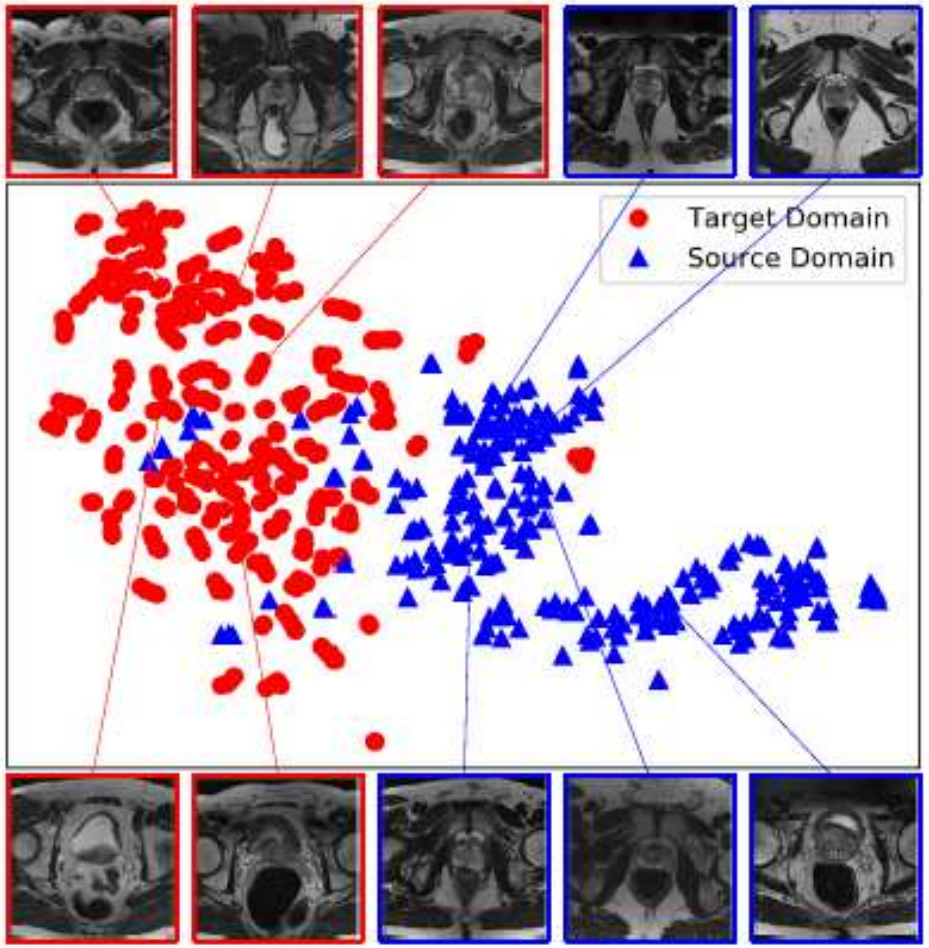
Not anywhere close to the 10k images of Pascal VOC and the 5k of Cityskapes

Crowdsourcing?

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

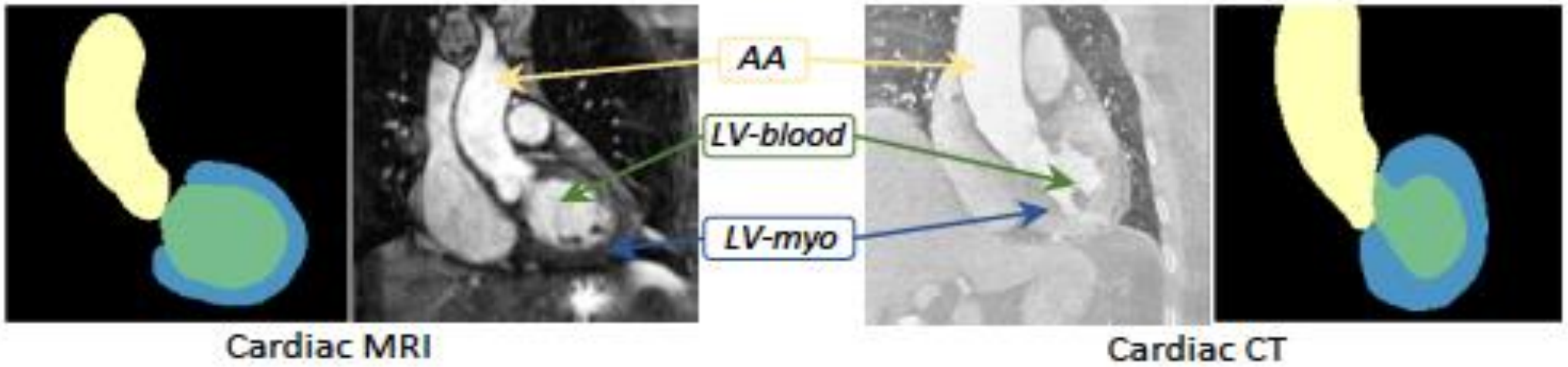


Domain shifts make things worse (even with full annotations in one domain)



[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

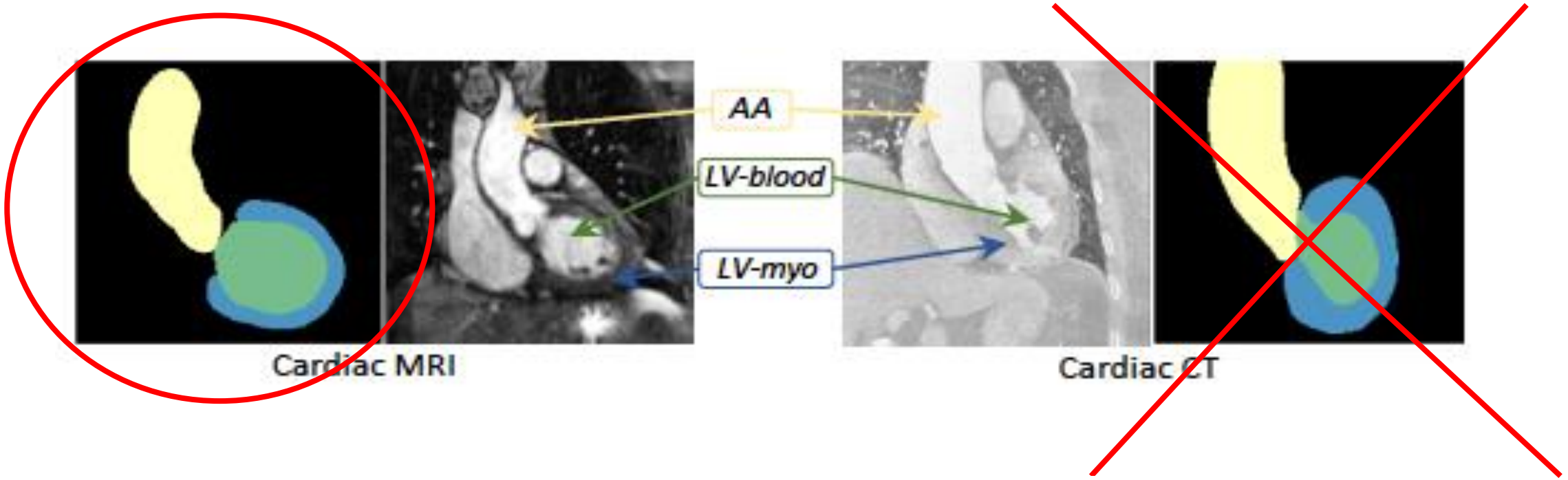


[Images from Dou et al., PnP-AdaNet: Plug-and-play adversarial domain adaptation network with a benchmark at cross-modality cardiac segmentation ArXiv 2018]

Unsupervised domain adaptation

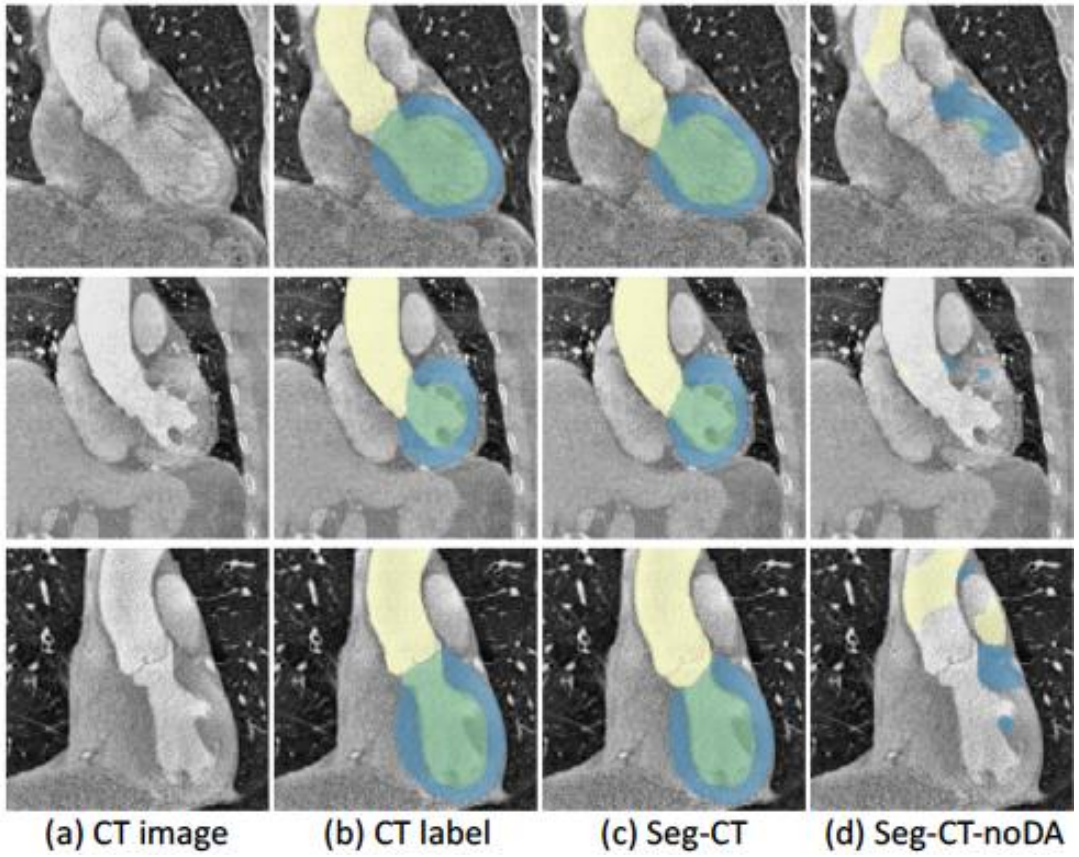
We have labels for the source domain

No labels for the target



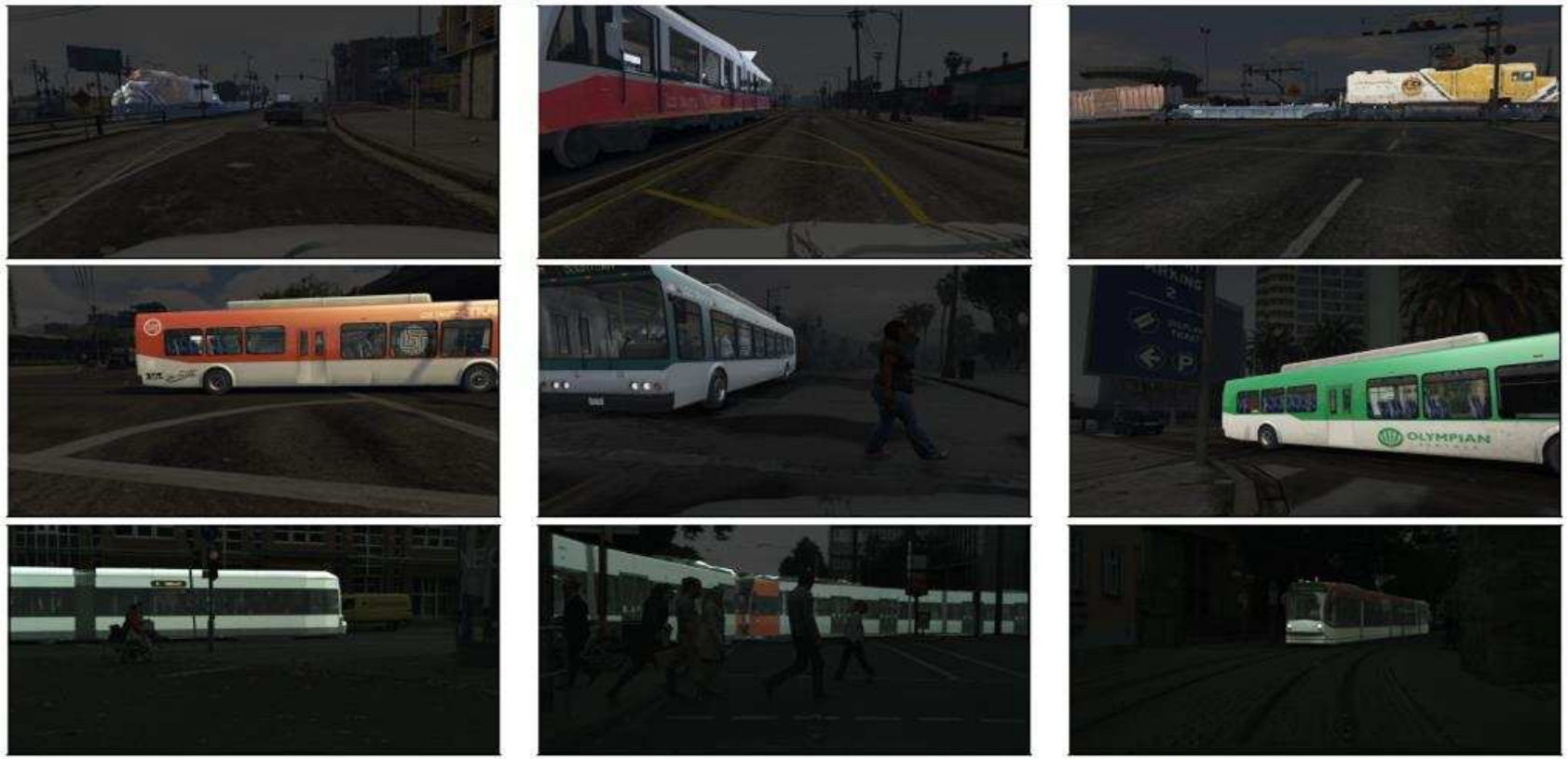
[Images from Dou et al., PnP-AdaNet: Plug-and-play adversarial domain adaptation network with a benchmark at cross-modality cardiac segmentation ArXiv 2018]

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*



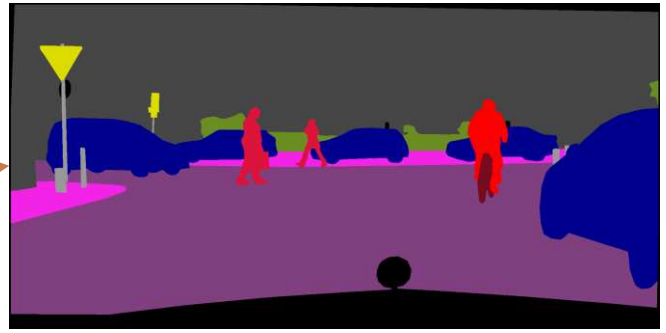
"train"
GTA

"bus"
GTA

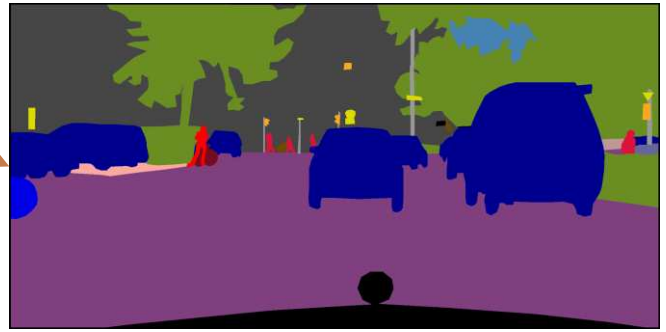
"train"
Cityscapes

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*



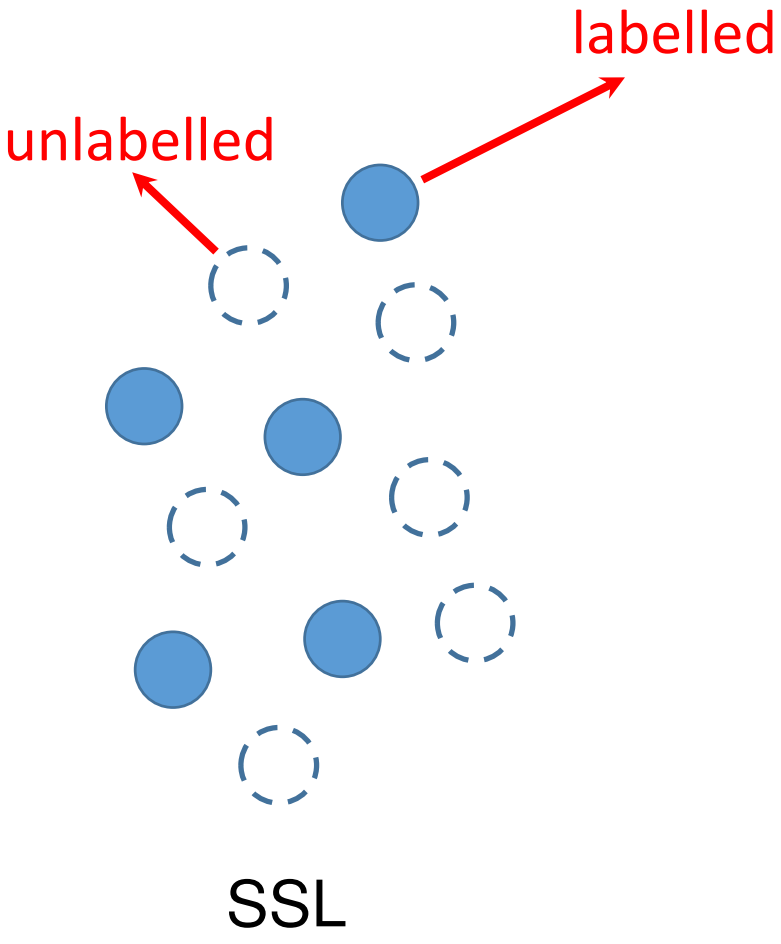
Frankfurt



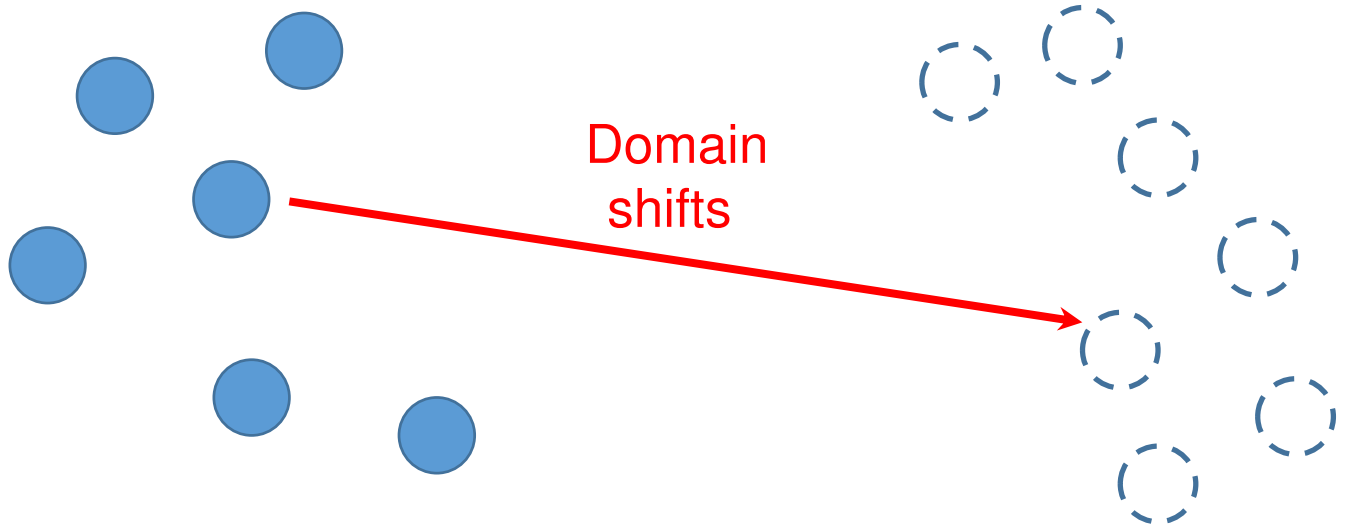
Zurich

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

UDA = SSL + domain shift

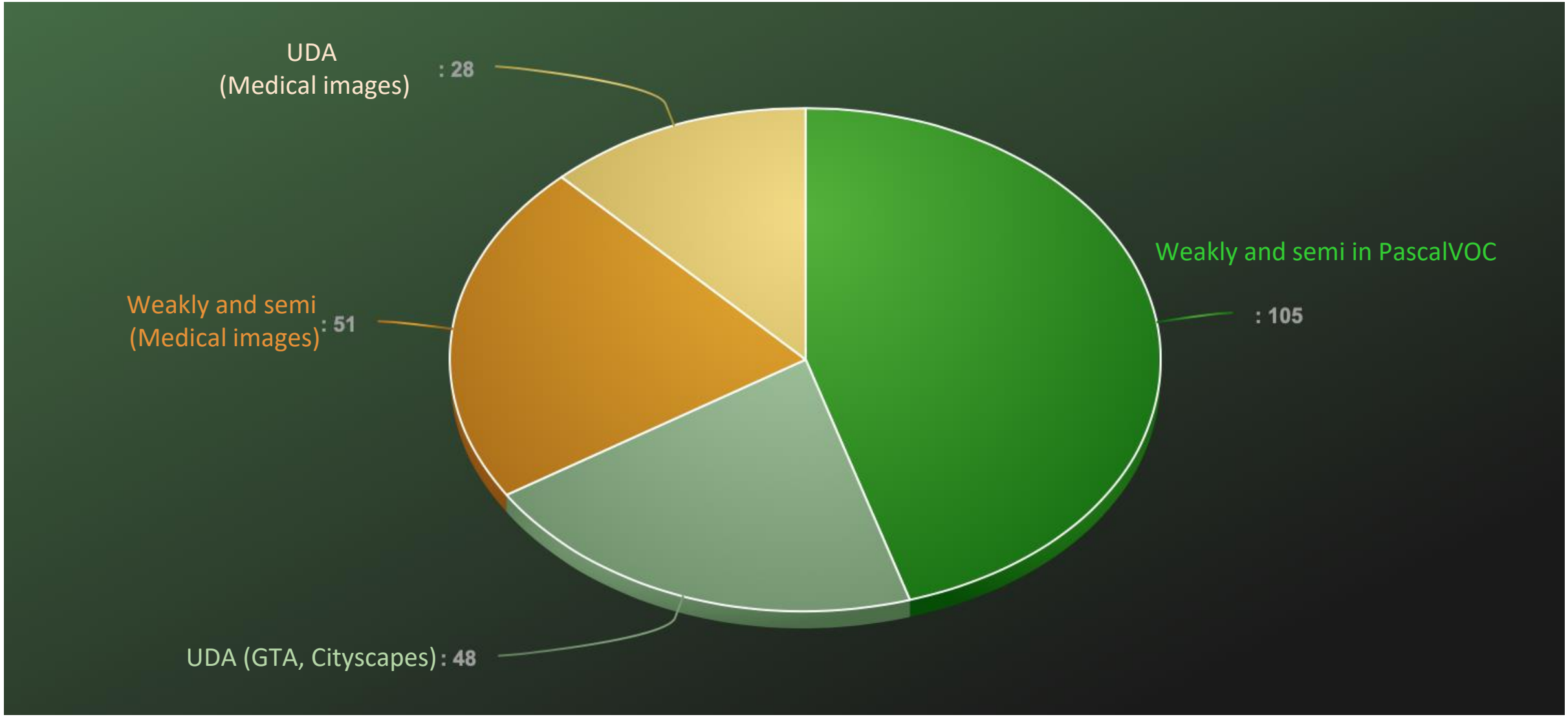


SSL



UDA

An increasing interest in medical image analysis as well



[Numbers compiled in 2020]

and there is few-shot learning (mostly in vision; a few recent works in medical imaging)

Few-shot tasks at testing time

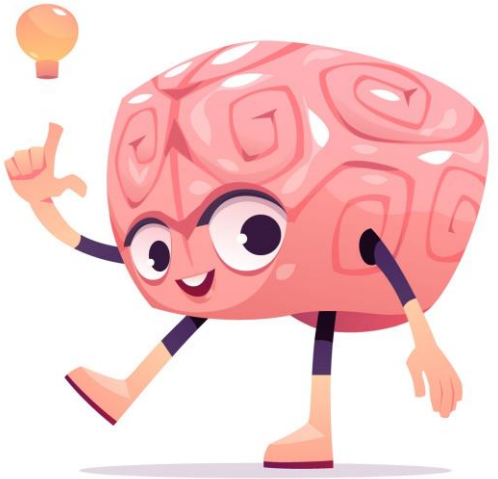
Learn from a few examples per **new** class

Classify these

and there is few-shot learning (mostly in vision, a few recent works in medical imaging)

-- **Humans recognize easily** with few examples

-- Modern ML generalize very poorly



Why it is interesting:

Available data sets represent small sub-domains of the world

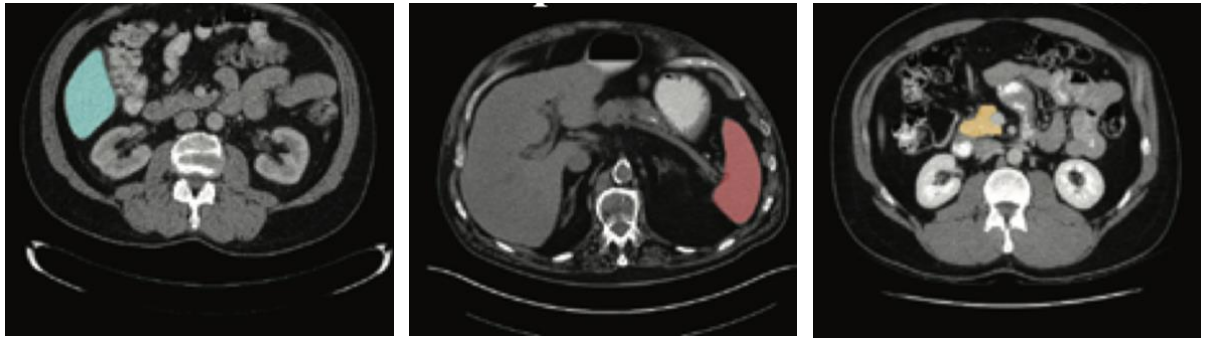
Cityscapes (5k images; 1.5h per image):

Urban scenes, less than 30 classes

The problem is even more motivated for dense predictions like segmentation

New classes, but with few examples

It is relevant in medical imaging as well (e.g., transfer learning in multi-organ segmentation)

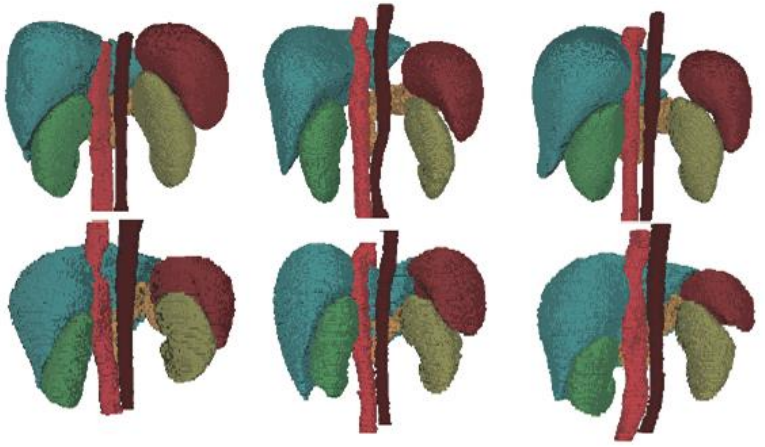


Liver Spleen Pancreas

You could think of it as a meaningful/advanced transfer learning

(should be better than transferring knowledge from “cars” to “spleens”!)

up to 13 classes in abdominal-imaging data sets



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

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

- Structure-driven priors: *Regularization (Part 1)*
- Knowledge-driven priors (e.g., anatomy): *Constraints (Part 2)*
- Data-driven priors *(Part 3)*
- *Hands on implementations in PyTorch*

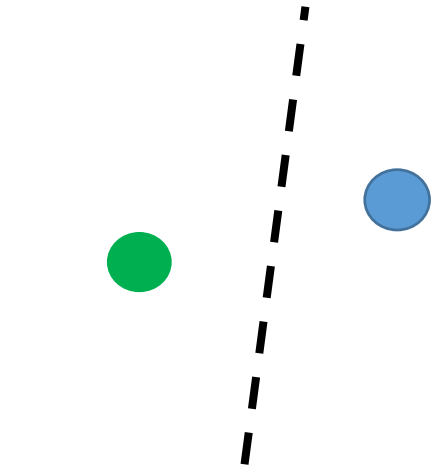
Learning with Limited Supervision

Yuyin Zhou (Yan Wang)
Ismail Ben Ayed
Jose Dolz
Christian Desrosiers
Marleen de Bruijne
Hoel Kervadec

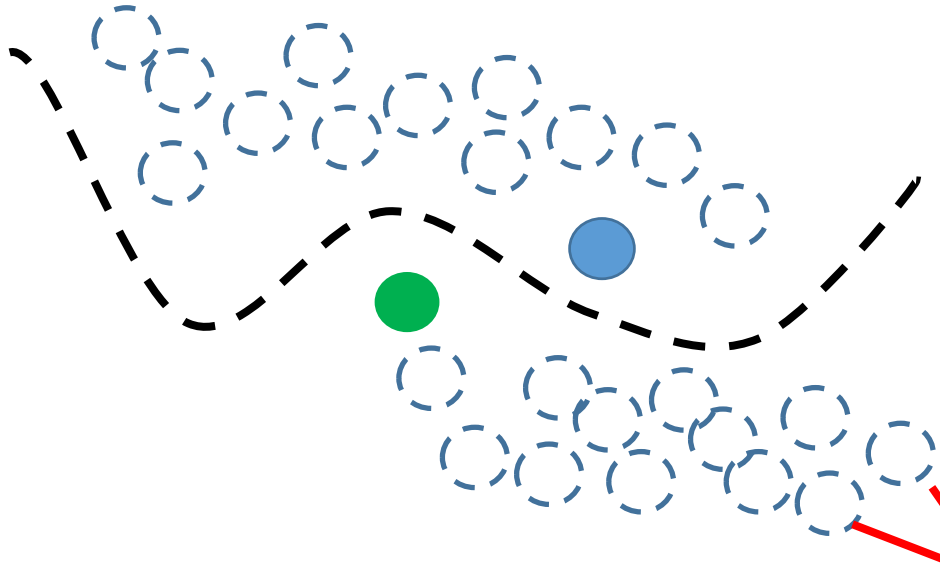
Regularization

Laplacian regularization & CRFs

Semi-supervised learning (general form)

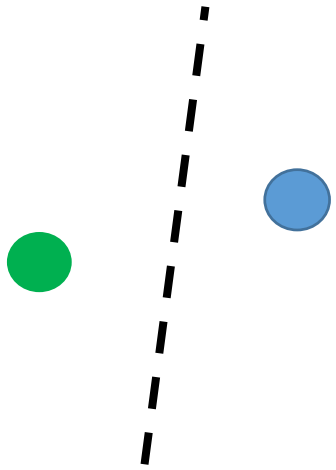


Only labeled points

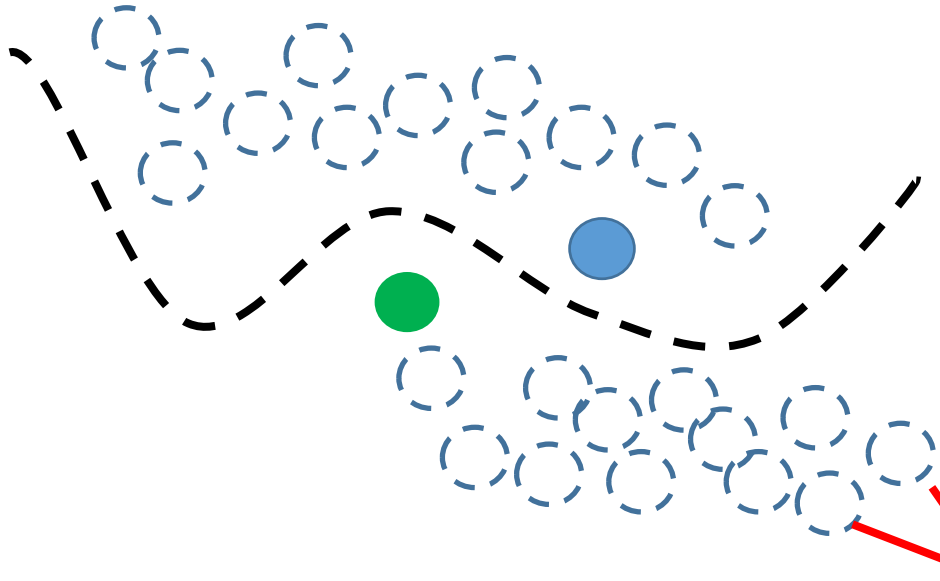


Unlabeled points

Semi-supervised learning (general form)



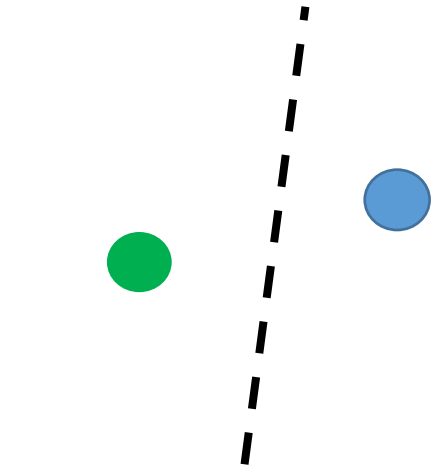
Only labeled points



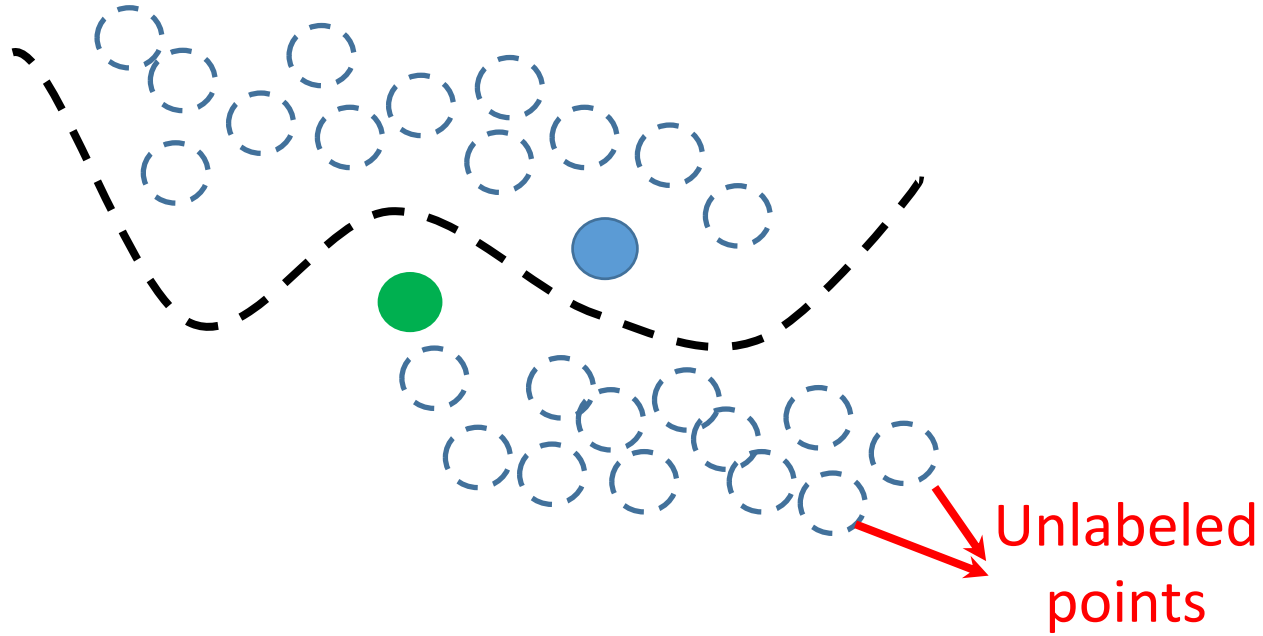
Unlabeled points

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

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

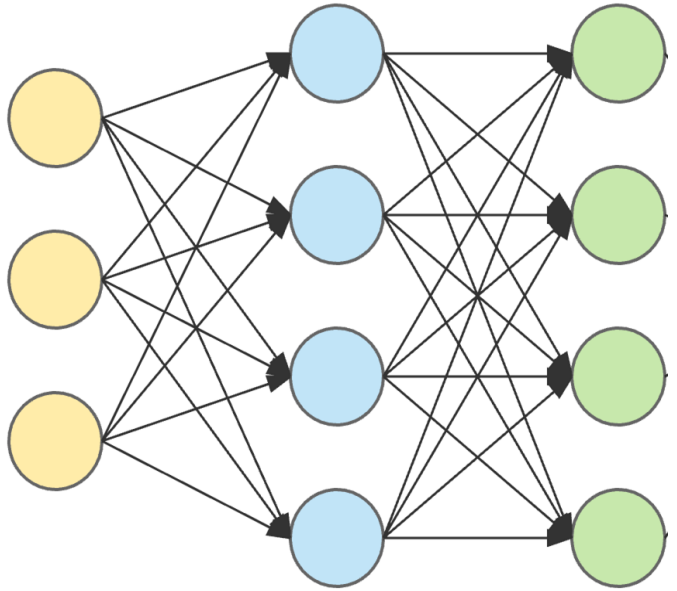
Semi-supervised learning (general form)

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

Labels

cross-entropy

softmax outputs



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

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

Semi-supervised learning (general form)

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

cross-entropy

Labels

softmax outputs of the network

e.g.: **Laplacian**

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

Semi-supervised learning (general form)

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

cross-entropy

Labels

softmax outputs

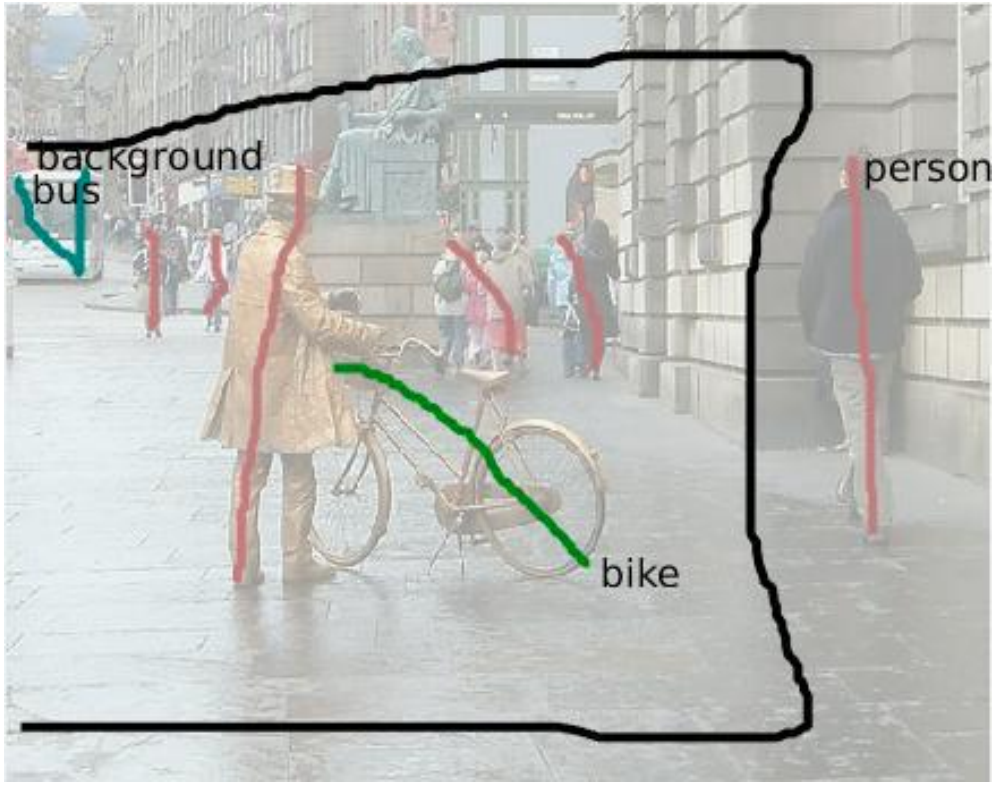
e.g.: **Laplacian**

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

- [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

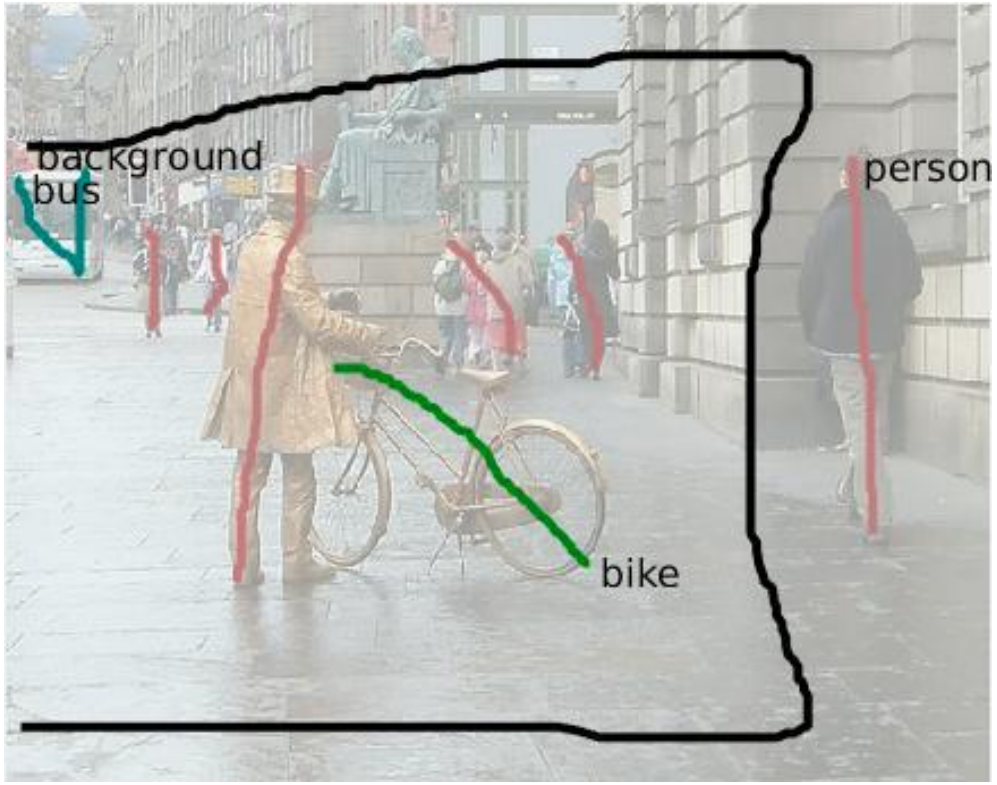
$$\min_{\theta} \sum_{p \in \mathcal{L}} l(\mathbf{y}^p, \mathbf{s}_\theta^p) + \sum_{p,q \in \mathcal{L} \cup \mathcal{U}} w_{p,q} \|\mathbf{s}_\theta^p - \mathbf{s}_\theta^q\|^2$$



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

Semi-supervision loss for segmentation

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



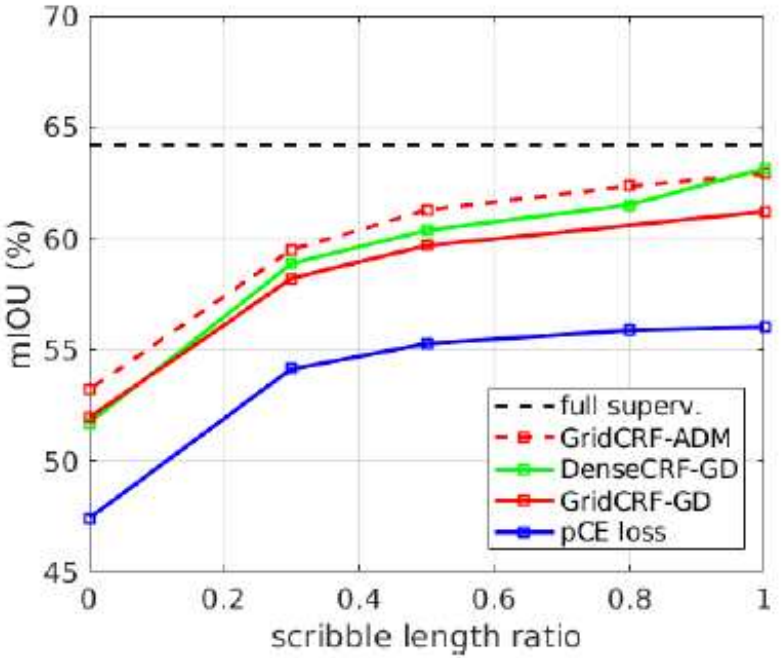
On the vertices of the simplex (binary variables), this is exactly the popular Potts model in CRFs

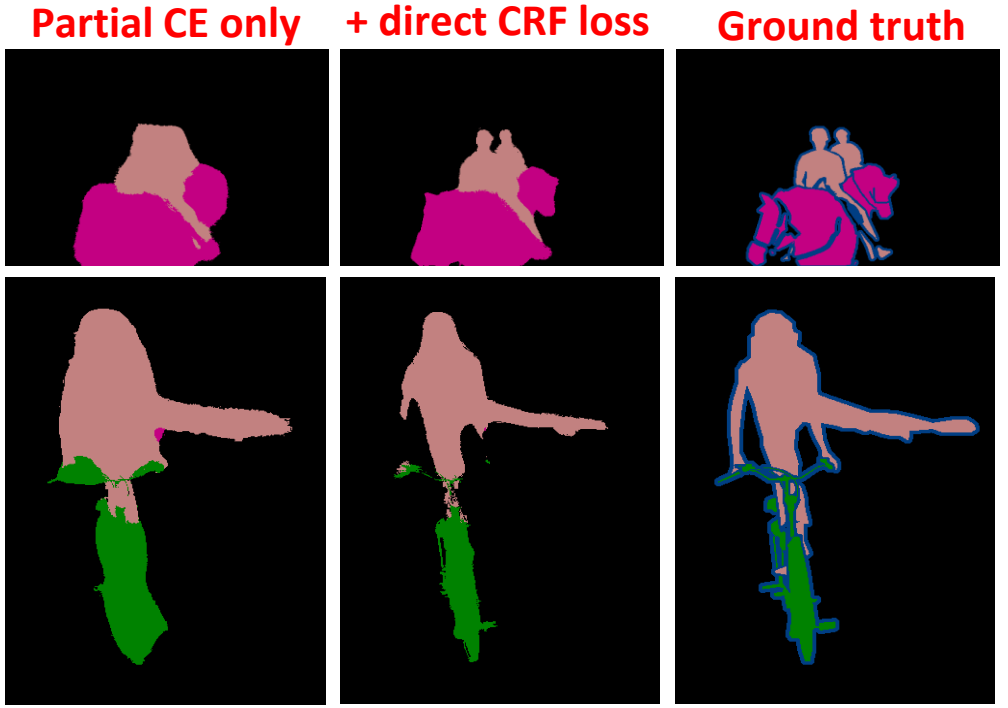
[Tang et al., On regularized losses for weakly supervised segmentation, ECCV 2018]

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

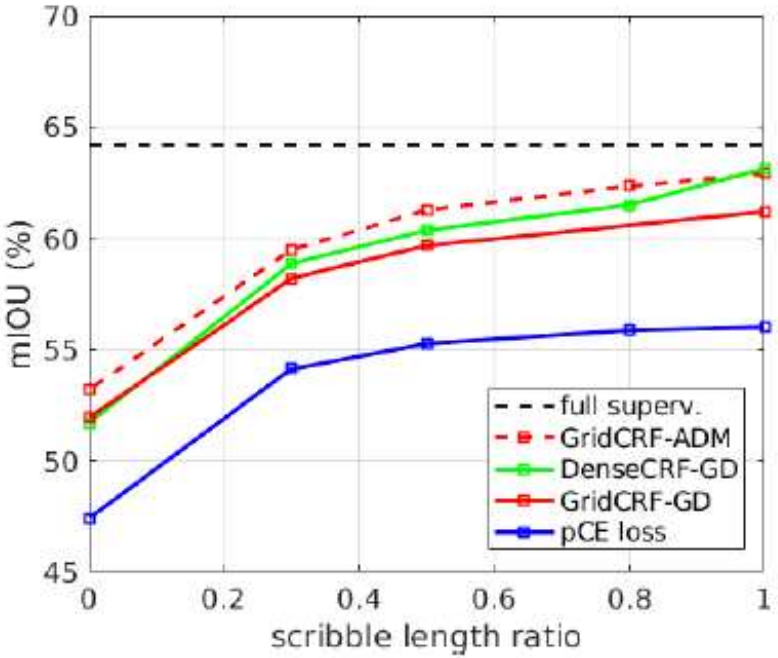




[Tang et al., On regularized losses for weakly supervised segmentation, ECCV 2018]
[Marin et al., Beyond gradient descent for regularized segmentation losses, CVPR 2019]

Semi-supervision loss for segmentation

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

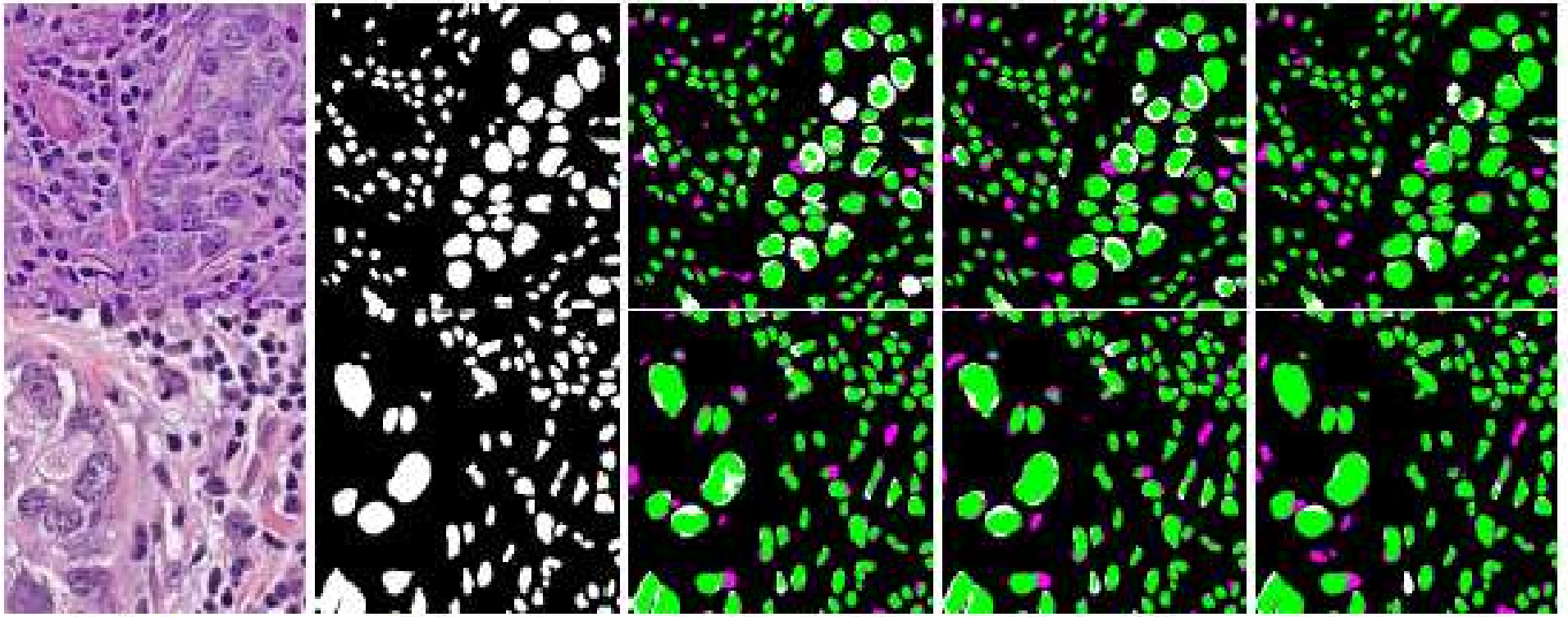


97.6% of full supervision performance with **3%** of the labels!

[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]

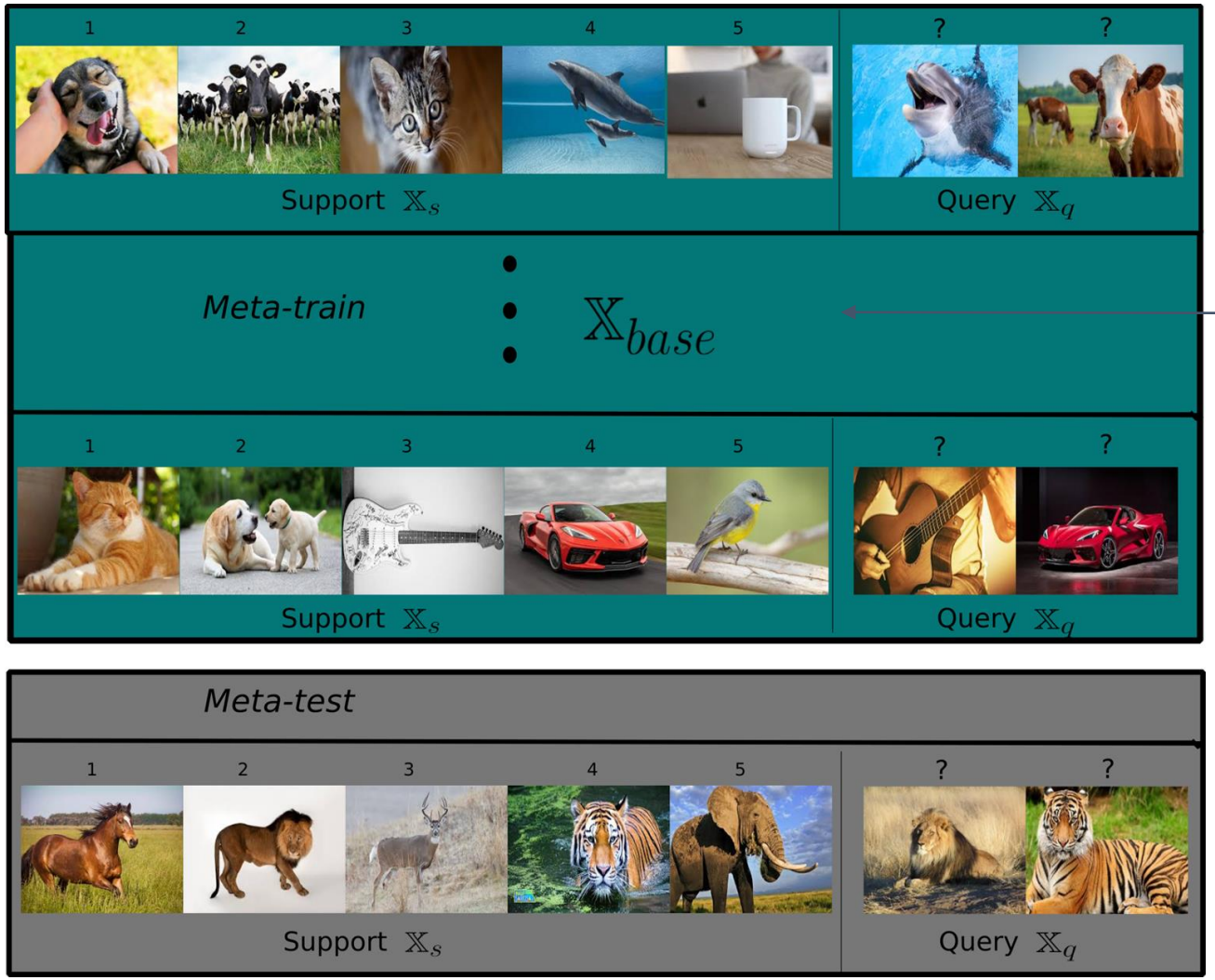
Laplacian-regularized few-shot learning

Ziko et al., ICML 2020

A very large body of recent works, mostly based on:

Meta-learning

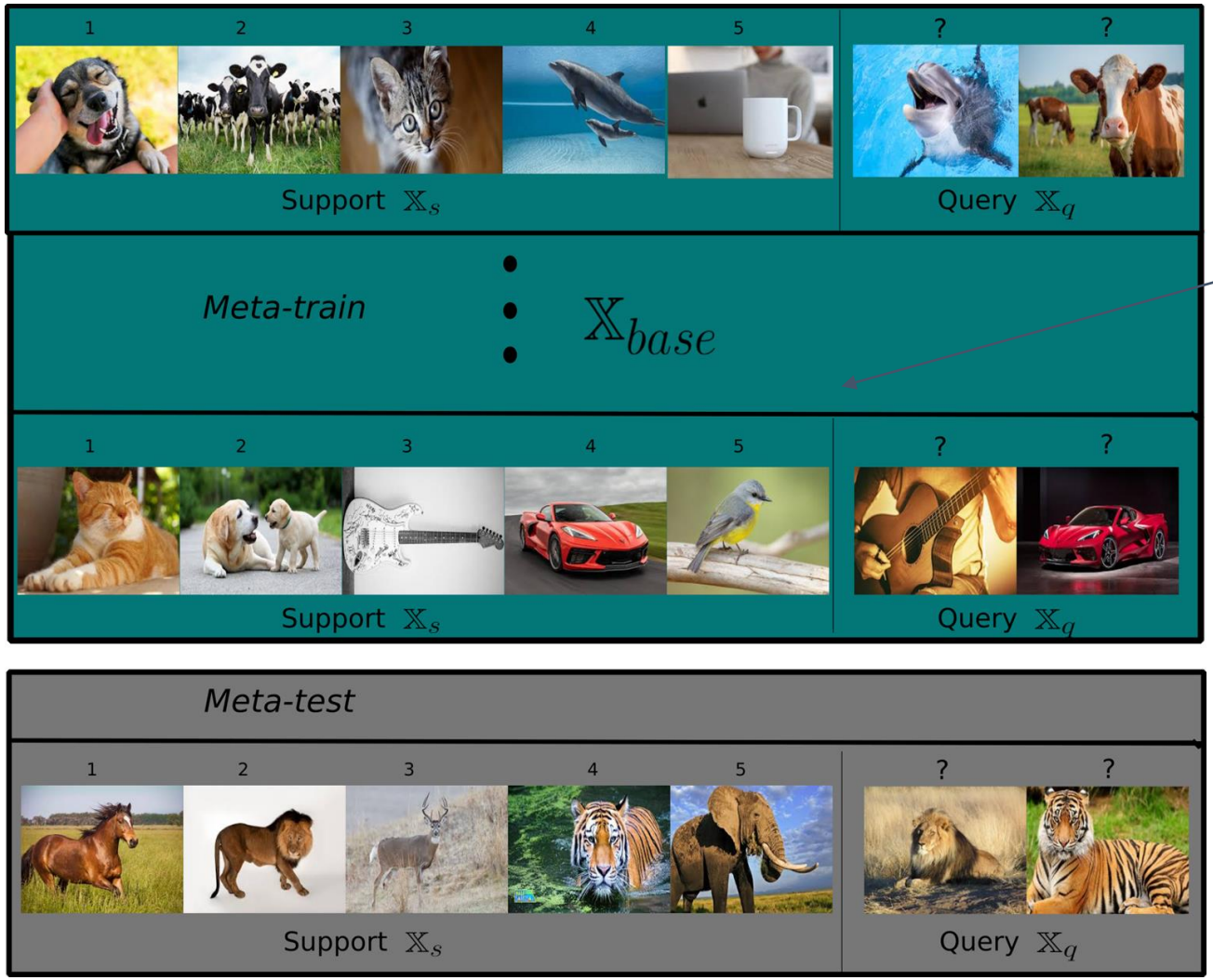
Meta-Learning or “learning-to-learn”



Base training with enough labeled data

(base classes ***different from the*** test classes)

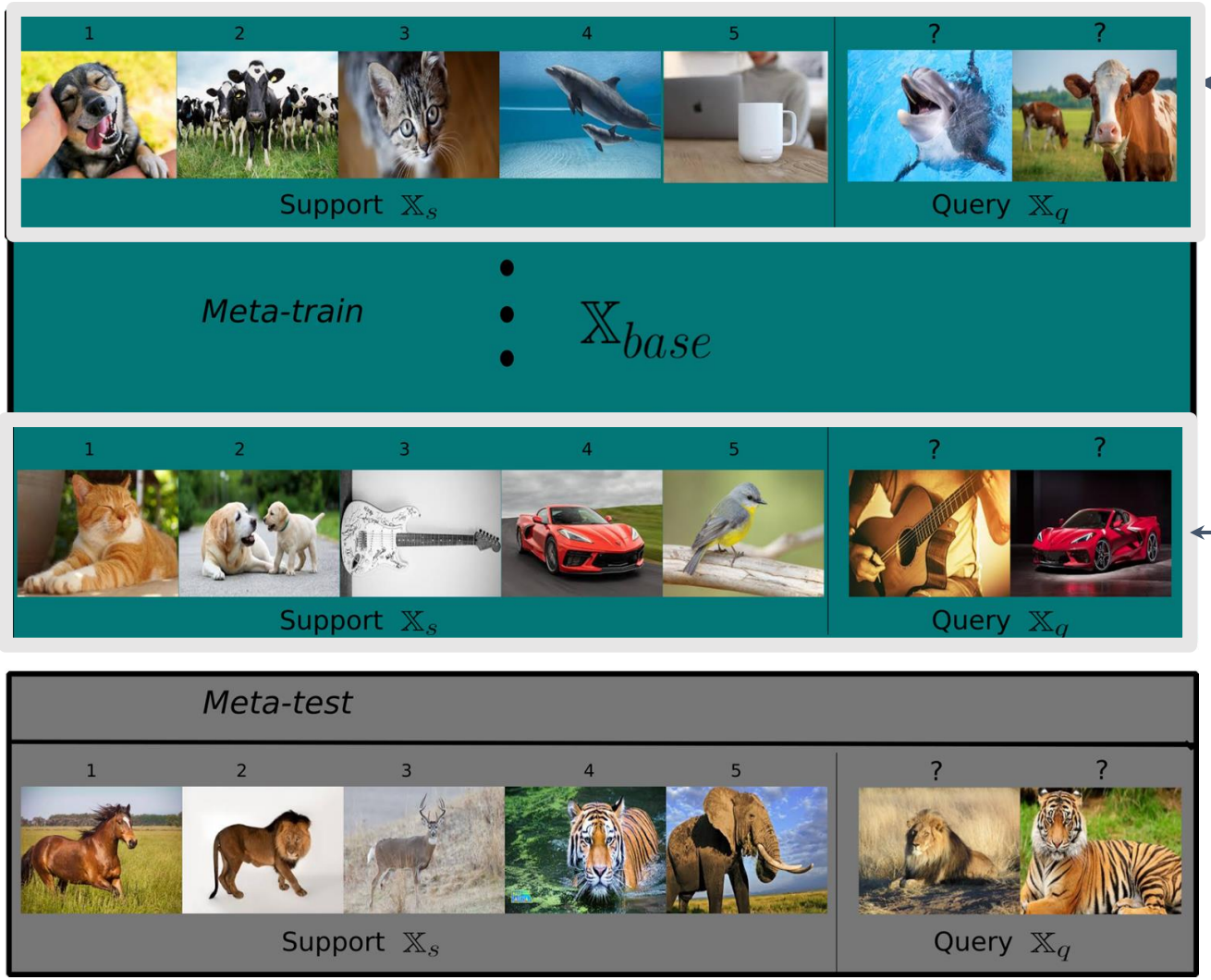
Meta-Learning or “learning-to-learn”



learn initial model

$$f_\theta$$

Meta-Learning or “learning-to-learn”



Create artificial episodes for ***episodic training***

Vinyal et al, (Neurips ‘16),
Snell et al, (Neurips ‘17),
Sung et al, (CVPR ‘ 18),
Finn et al, (ICML‘ 17),
Ravi et al, (ICLR‘ 17),
Lee et al, (CVPR‘ 19),
Hu et al, (ICLR ‘20),
Ye et al, (CVPR ‘20), . . .

Taking a few steps backwards

Simple transfer-learning baselines and good classical regularizers outperform convoluted meta-learning approaches

[Chen et al., ICLR'19]; [Tian et al., ECCV'20]
[Dhillon et al., ICLR'20]; [Ziko et al., ICML'20]; [Boudiaf et al., NeurIPS'20]

Taking a few steps backwards

Simple transfer-learning baselines and good classical regularizers outperform convoluted meta-learning approaches

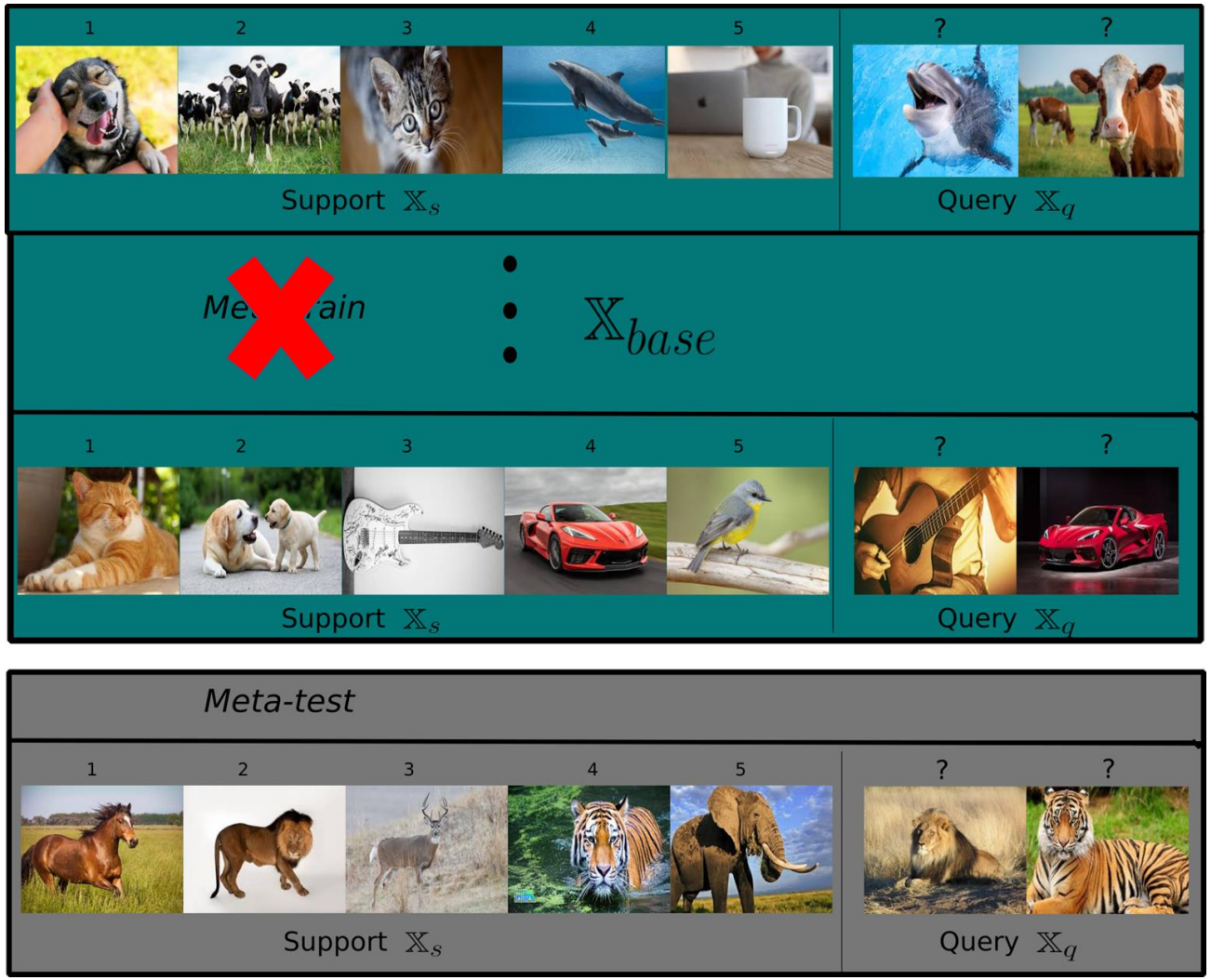
[Chen et al., ICLR'19]; [Tian et al., ECCV'20]
[Dhillon et al., ICLR'20]; [Ziko et al., ICML'20]; [Boudiaf et al., NeurIPS'20]

Entropy regularization

Laplacian regularization

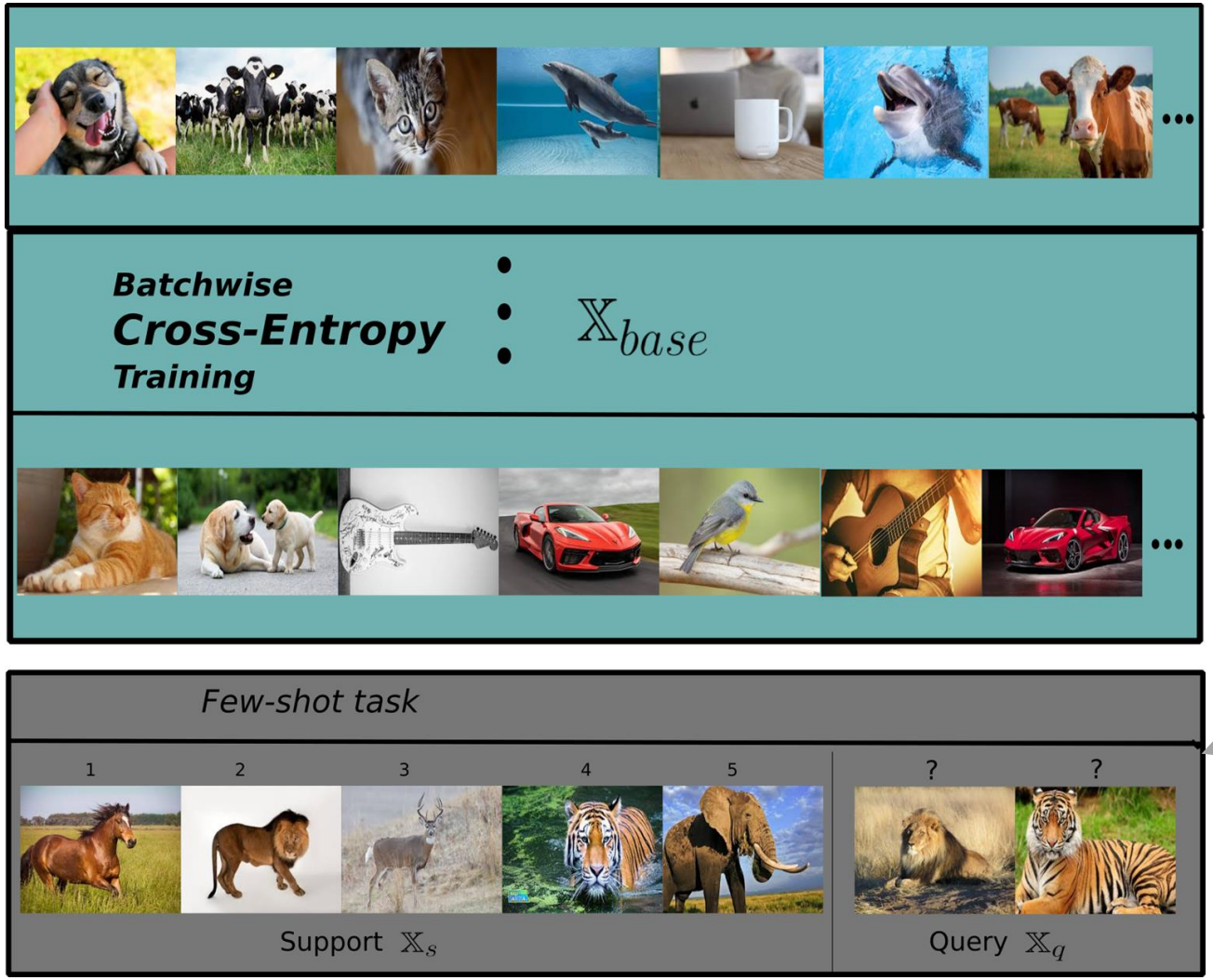
Mutual-information regularization

Baseline Framework



No need to
meta-train

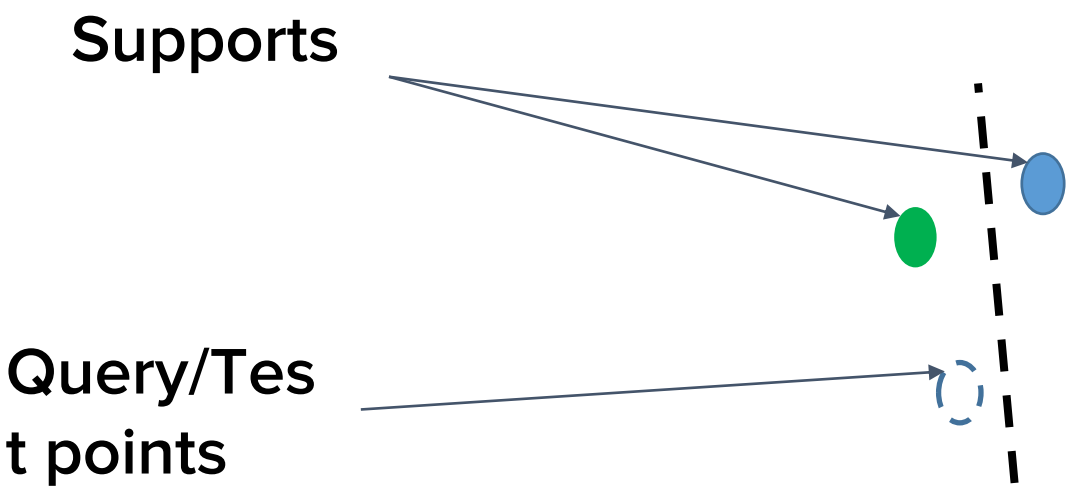
Baseline Framework



Conventional training

Types of inference: Ind. vs. Trans.

Inductive inference



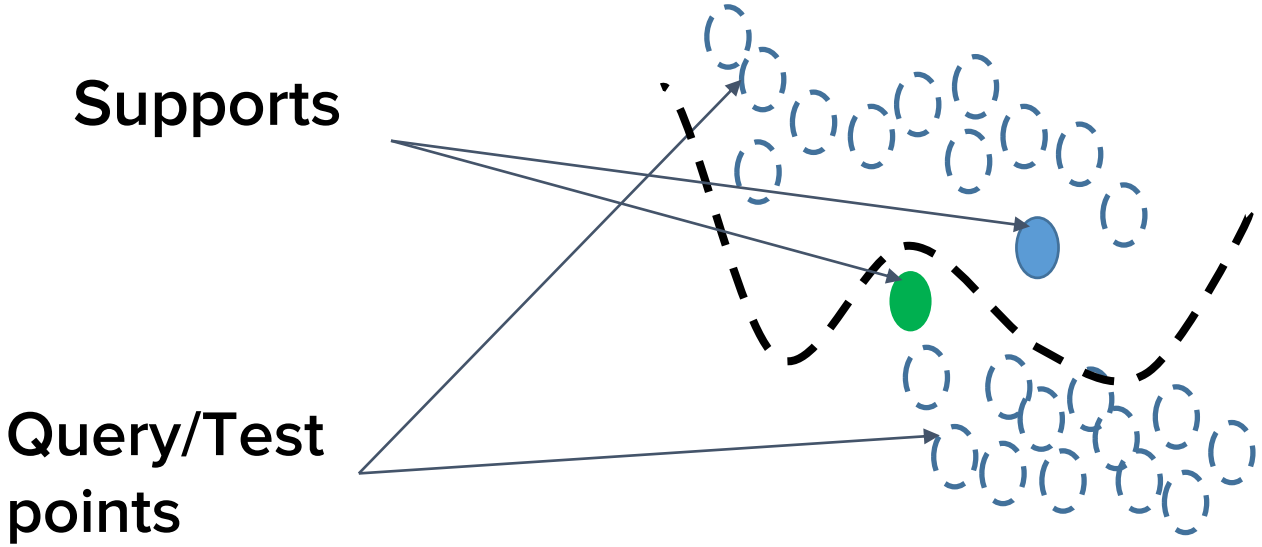
Examples

Vinayls et al., NeurIPS'16
(Attention mechanism)

Snell et al., NeurIPS'17
(Prototypical nets)

Transductive inference

(This is very relevant to segmentation: more on this later)



Examples

Dhillon et al., ICLR'20 (entropy)

Ziko et al., ICML'20 (Laplacian)

Predict for all test points, instead of one at a time

LaplacianShot objective for inference

$$\mathcal{E}(\mathbf{Y}) = \mathcal{N}(\mathbf{Y}) + \frac{\lambda}{2}\mathcal{L}(\mathbf{Y})$$

$$\mathcal{N}(\mathbf{Y}) = \sum_{q=1}^{N}\sum_{c=1}^{C} y_{q,c}\, d(\mathbf{x}_q - \mathbf{m}_c)$$

[Ziko et al., Laplacian regularized few-shot learning, ICML 2020]

LaplacianShot objective for inference

$$\mathcal{E}(\mathbf{Y}) = \mathcal{N}(\mathbf{Y}) + \frac{\lambda}{2}\mathcal{L}(\mathbf{Y})$$

$$\mathcal{N}(\mathbf{Y}) = \sum_{q=1}^{N}\sum_{c=1}^{C} y_{q,c}\, d(\mathbf{x}_q - \mathbf{m}_c)$$

Mean of the support samples

Feature embedding

$$\mathbf{x}_q = f_\theta(x_q)$$

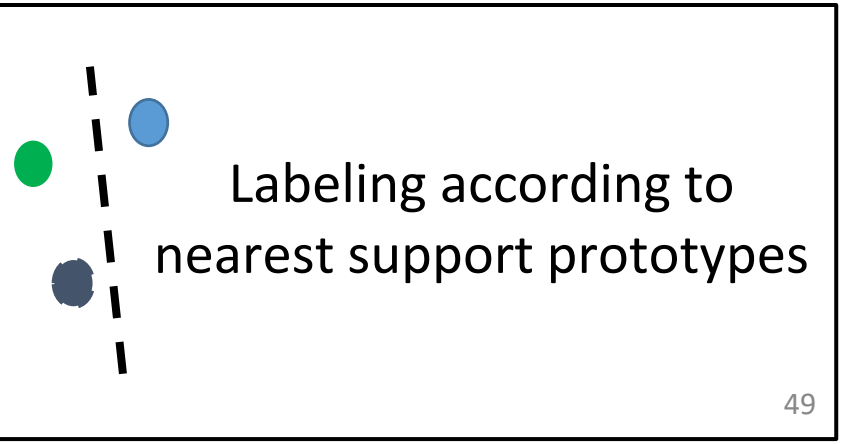
Label assignments for queries

[Ziko et al., Laplacian regularized few-shot learning, ICML 2020]

LaplacianShot objective for inference

$$\mathcal{E}(\mathbf{Y}) = \mathcal{N}(\mathbf{Y}) + \frac{\lambda}{2}\mathcal{L}(\mathbf{Y})$$

$$\mathcal{N}(\mathbf{Y}) = \sum_{q=1}^{N}\sum_{c=1}^{C} y_{q,c}\, d(\mathbf{x}_q - \mathbf{m}_c)$$

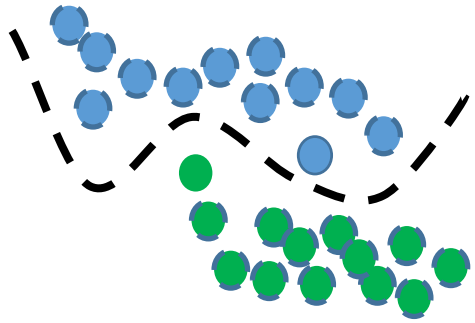


[Ziko et al., Laplacian regularized few-shot learning, ICML 2020]

LaplacianShot objective for inference

$$\mathcal{E}(\mathbf{Y}) = \mathcal{N}(\mathbf{Y}) + \frac{\lambda}{2}\mathcal{L}(\mathbf{Y})$$

$$\sum_{q,p} w(\mathbf{x}_q, \mathbf{x}_p)\|\mathbf{y}_q - \mathbf{y}_p\|^2$$

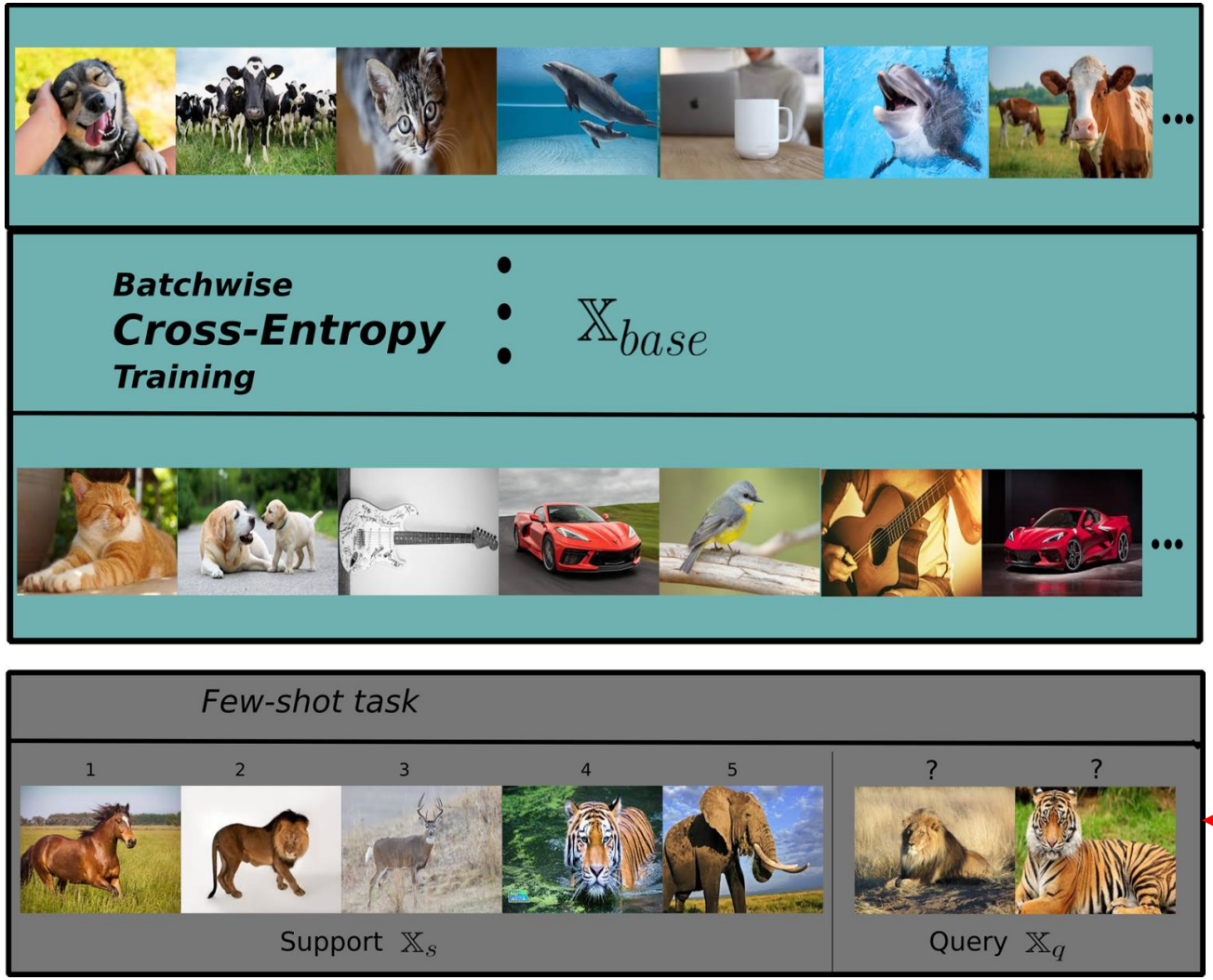


Laplacian regularization:

Encourages nearby points to have similar assignments

[Ziko et al., Laplacian regularized few-shot learning, ICML 2020]

Experiments



Conventional training on base classes

LaplacianShot during inference

Results (Mini-ImageNet)

Methods	Network	1-shot	5-shot
MAML [Finn et al., 2017]	ResNet-18	49.61 ± 0.92	65.72 ± 0.77
Chen [Chen et al., 2019]	ResNet-18	51.87 ± 0.77	75.68 ± 0.63
RelationNet [Sung et al., 2018]	ResNet-18	52.48 ± 0.86	69.83 ± 0.68
MatchingNet [Vinyals et al., 2016]	ResNet-18	52.91 ± 0.88	68.88 ± 0.69
ProtoNet [Snell et al., 2017]	ResNet-18	54.16 ± 0.82	73.68 ± 0.65
Gidaris [Gidaris and Komodakis, 2018]	ResNet-15	55.45 ± 0.89	70.13 ± 0.68
SNAIL[Mishra et al., 2018]	ResNet-15	55.71 ± 0.99	68.88 ± 0.92
AdaCNN [Munkhdalai et al., 2018]	ResNet-15	56.88 ± 0.62	71.94 ± 0.57
TADAM [Oreshkin et al., 2018]	ResNet-15	58.50 ± 0.30	76.70 ± 0.30
CAML [Jiang et al., 2019]	ResNet-12	59.23 ± 0.99	72.35 ± 0.71
TPN [Yanbin et al., 2019]	ResNet-12	59.46	75.64
TEAM [Qiao et al., 2019]	ResNet-18	60.07	75.90
MTL [Sun et al., 2019]	ResNet-18	61.20 ± 1.80	75.50 ± 0.80
VariationalFSL [Zhang et al., 2019]	ResNet-18	61.23 ± 0.26	77.69 ± 0.17
Transductive tuning [Dhillon et al., 2020]	ResNet-12	62.35 ± 0.66	74.53 ± 0.54
MetaoptNet[Lee et al., 2019]	ResNet-18	62.64 ± 0.61	78.63 ± 0.46
SimpleShot [Wang et al., 2019]	ResNet-18	63.10 ± 0.20	79.92 ± 0.14
CAN+T [Hou et al., 2019]	ResNet-12	67.19 ± 0.55	80.64 ± 0.35
LaplacianShot (ours)	ResNet-18	**72.11** ± 0.19	**82.31** ± 0.14

Results (Mini-ImageNet)

Methods	Network	1-shot	5-shot
MAML [Finn et al., 2017]	ResNet-18	49.61 ± 0.92	65.72 ± 0.77
Chen [Chen et al., 2019]	ResNet-18	51.87 ± 0.77	75.68 ± 0.63
RelationNet [Sung et al., 2018]	ResNet-18	52.48 ± 0.86	69.83 ± 0.68
MatchingNet [Vinyals et al., 2016]	ResNet-18	52.91 ± 0.88	68.88 ± 0.69
ProtoNet [Snell et al., 2017]	ResNet-18	54.16 ± 0.82	73.68 ± 0.65
Gidaris [Gidaris and Komodakis, 2018]	ResNet-15	55.45 ± 0.89	70.13 ± 0.68
SNAIL[Mishra et al., 2018]	ResNet-15	55.71 ± 0.99	68.88 ± 0.92
AdaCNN [Munkhdalai et al., 2018]	ResNet-15	56.88 ± 0.62	71.94 ± 0.57
TADAM [Oreshkin et al., 2018]	ResNet-15	58.50 ± 0.30	76.70 ± 0.30
CAML [Jiang et al., 2019]	ResNet-12	59.23 ± 0.99	72.35 ± 0.71
TPN [Yanbin et al., 2019]	ResNet-12	59.46	75.64
TEAM [Qiao et al., 2019]	ResNet-18	60.07	75.90
MTL [Sun et al., 2019]	ResNet-18	61.20 ± 1.80	75.50 ± 0.80
VariationalFSL [Zhang et al., 2019]	ResNet-18	61.23 ± 0.26	77.69 ± 0.17
Transductive tuning [Dhillon et al., 2020]	ResNet-12	62.35 ± 0.66	74.53 ± 0.54
MetaoptNet[Lee et al., 2019]	ResNet-18	62.64 ± 0.61	78.63 ± 0.46
SimpleShot [Wang et al., 2019]	ResNet-18	63.10 ± 0.20	79.92 ± 0.14
CAN+T [Hou et al., 2019]	ResNet-12	67.19 ± 0.55	80.64 ± 0.35
LaplacianShot (ours)	ResNet-18	**72.11** ± 0.19	**82.31** ± 0.14

!!!!!!

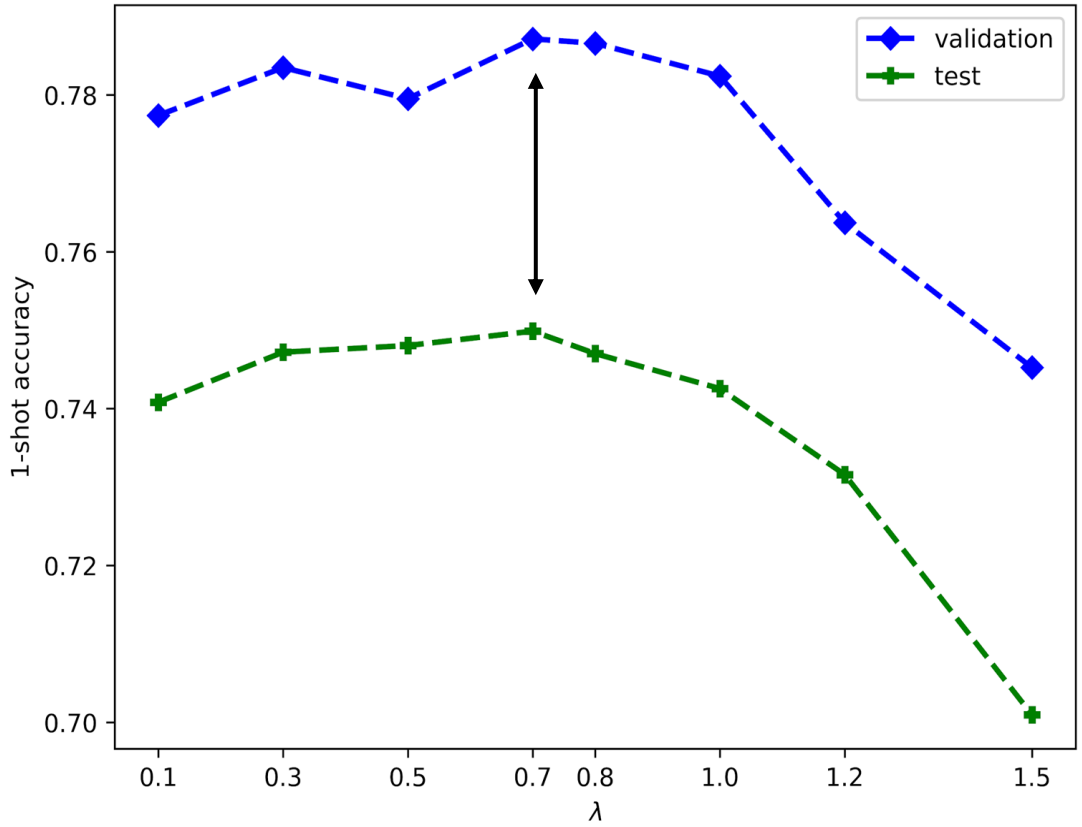
Results (Mini-ImageNet)

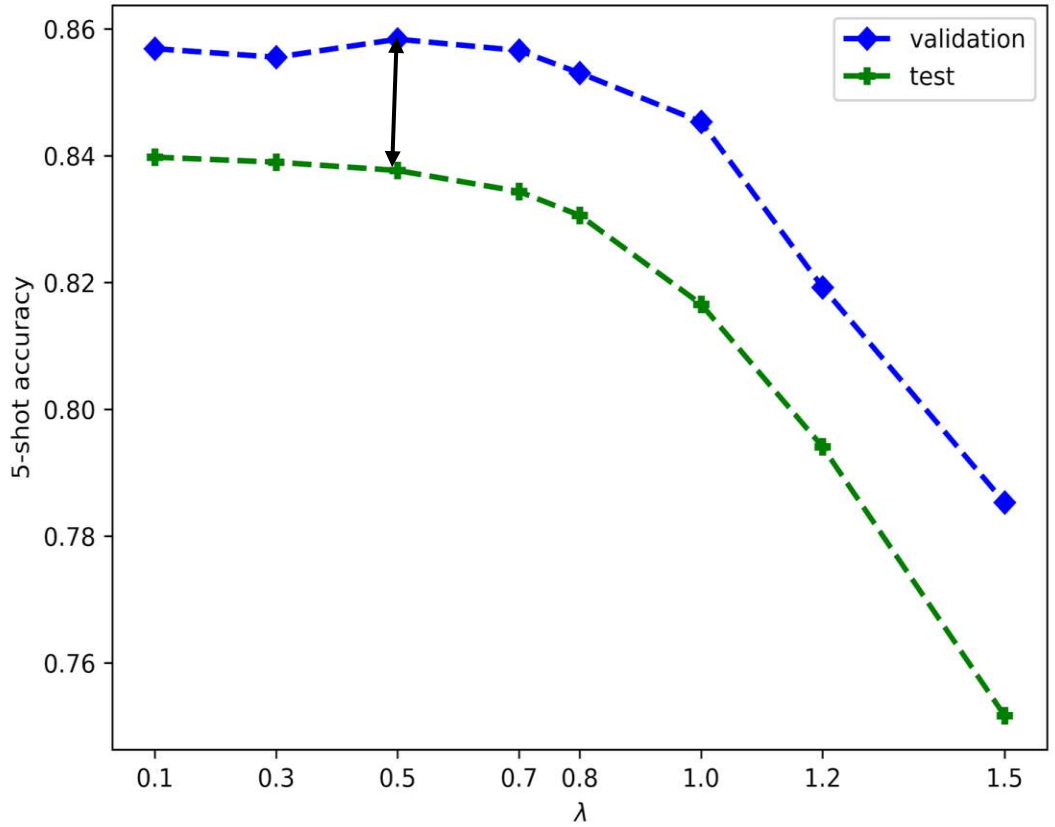
Methods	Network	1-shot	5-shot
Qiao [Qiao et al., 2018]	WRN	59.60 ± 0.41	73.74 ± 0.19
LEO [Rusu et al., 2019]	WRN	61.76 ± 0.08	77.59 ± 0.12
ProtoNet [Snell et al., 2017]	WRN	62.60 ± 0.20	79.97 ± 0.14
CC+rot[Gidaris et al., 2019]	WRN	62.93 ± 0.45	79.87 ± 0.33
MatchingNet [Vinyals et al., 2016]	WRN	64.03 ± 0.20	76.32 ± 0.16
FEAT [Ye et al., 2020]	WRN	65.10 ± 0.20	81.11 ± 0.14
Transductive tuning [Dhillon et al., 2020]	WRN	65.73 ± 0.68	78.40 ± 0.52
SimpleShot [Wang et al., 2019]	WRN	65.87± 0.20	82.09 ± 0.14
SIB [Hu et al., 2020]	WRN	70.0 ± 0.6	79.2 ± 0.4
BD-CSPN [Liu et al., 2019]	WRN	70.31 ± 0.93	81.89 ± 0.60
LaplacianShot (ours)	WRN	**74.86** ± 0.19	**84.13** ± 0.14

Results (Tiered-ImageNet)

Methods	Network	1-shot	5-shot
MetaoptNet[Lee et al., 2019]	ResNet-18	65.99 ± 0.72	81.56 ± 0.53
SimpleShot [Wang et al., 2019]	ResNet-18	69.68 ± 0.22	84.56 ± 0.16
CAN+T [Hou et al., 2019]	ResNet-12	73.21 ± 0.58	84.93 ± 0.38
LaplacianShot (ours)	ResNet-18	**78.98** ± 0.21	**86.39** ± 0.16
Meta SGD [Li et al., 2017]	WRN	62.95 ± 0.03	79.34 ± 0.06
LEO [Rusu et al., 2019]	WRN	66.33 ± 0.05	81.44 ± 0.09
FEAT [Ye et al., 2020]	WRN	70.41 ± 0.23	84.38 ± 0.16
CC+rot[Gidaris et al., 2019]	WRN	70.53 ± 0.51	84.98 ± 0.36
SimpleShot [Wang et al., 2019]	WRN	70.90 ± 0.22	85.76 ± 0.15
Transductive tuning [Dhillon et al., 2020]	WRN	73.34 ± 0.71	85.50 ± 0.50
BD-CSPN [Liu et al., 2019]	WRN	78.74 ± 0.95	86.92 ± 0.63
LaplacianShot (ours)	WRN	**80.18** ± 0.21	**87.56**± 0.15

Choosing λ





Average Inference time

Methods	Network	inference time (s)
SimpleShot [Wang et al., 2019]	WRN	0.009
Transductive tuning [Dhillon et al., 2020]	WRN	20.7
LaplacianShot	WRN	0.012

Transductive

LaplacianShot Takeaways

-- **SOTA results** without bell and whistles

-- Simple **constrained graph clustering** works very well

-- **No** network **fine-tuning**, neither **meta-learning**

-- **Model agnostic**

-- **Fast** transductive inference: almost inductive time

Regularization

Entropy

Entropy minimization for SSL

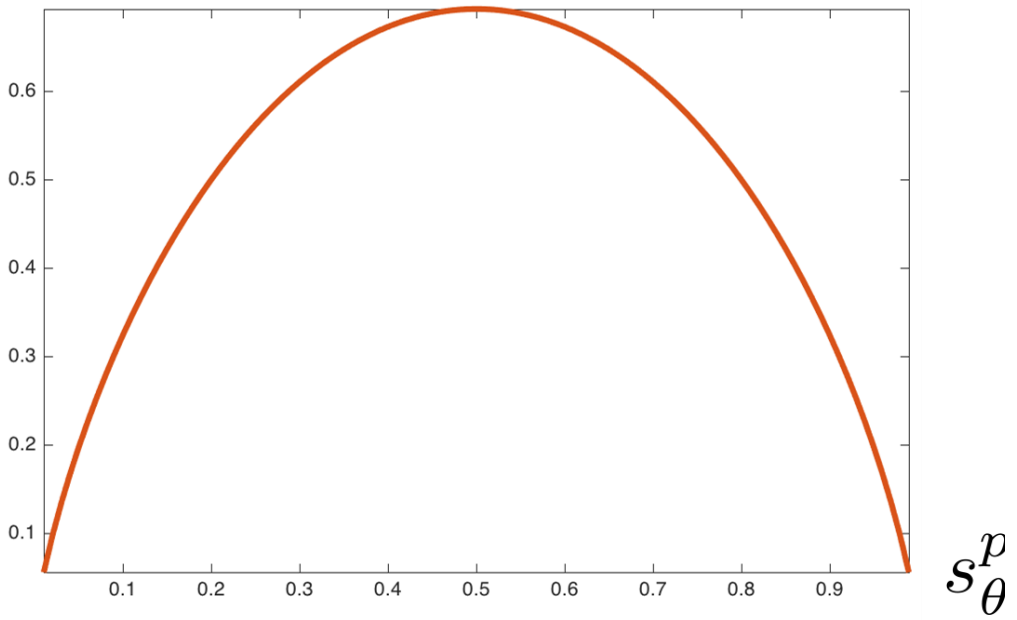
$$\min_{\theta} -\sum_{p\in\mathcal{L}}\sum_{c=1}^{C} y^{p,c}\log s_\theta^{p,c} \;\; -\sum_{p\in\mathcal{U}}\sum_{c=1}^{C} s_\theta^{p,c}\log s_\theta^{p,c}$$

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

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

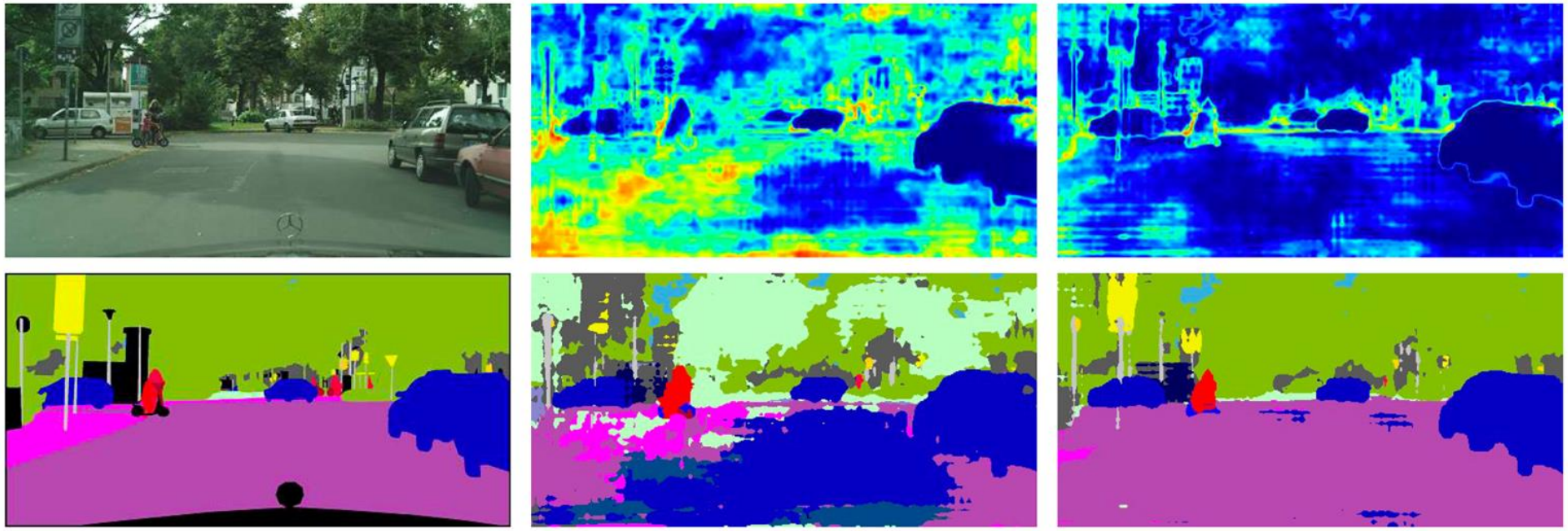
Effect of the entropy (why is it good for SSL?):
It makes the predictions confident (like cross-entropy)

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



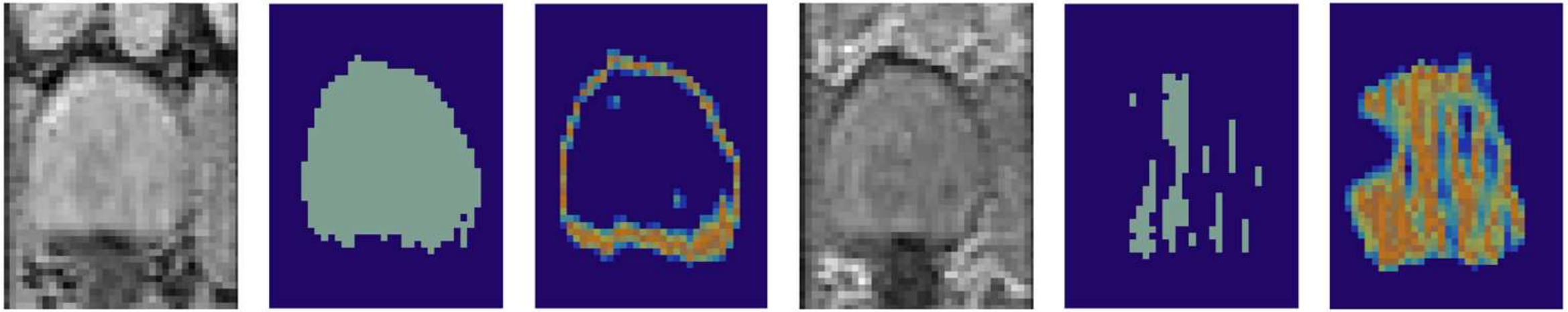
Entropy minimization for UDA

Input image + GT | Without adaptation | Entropy minimization



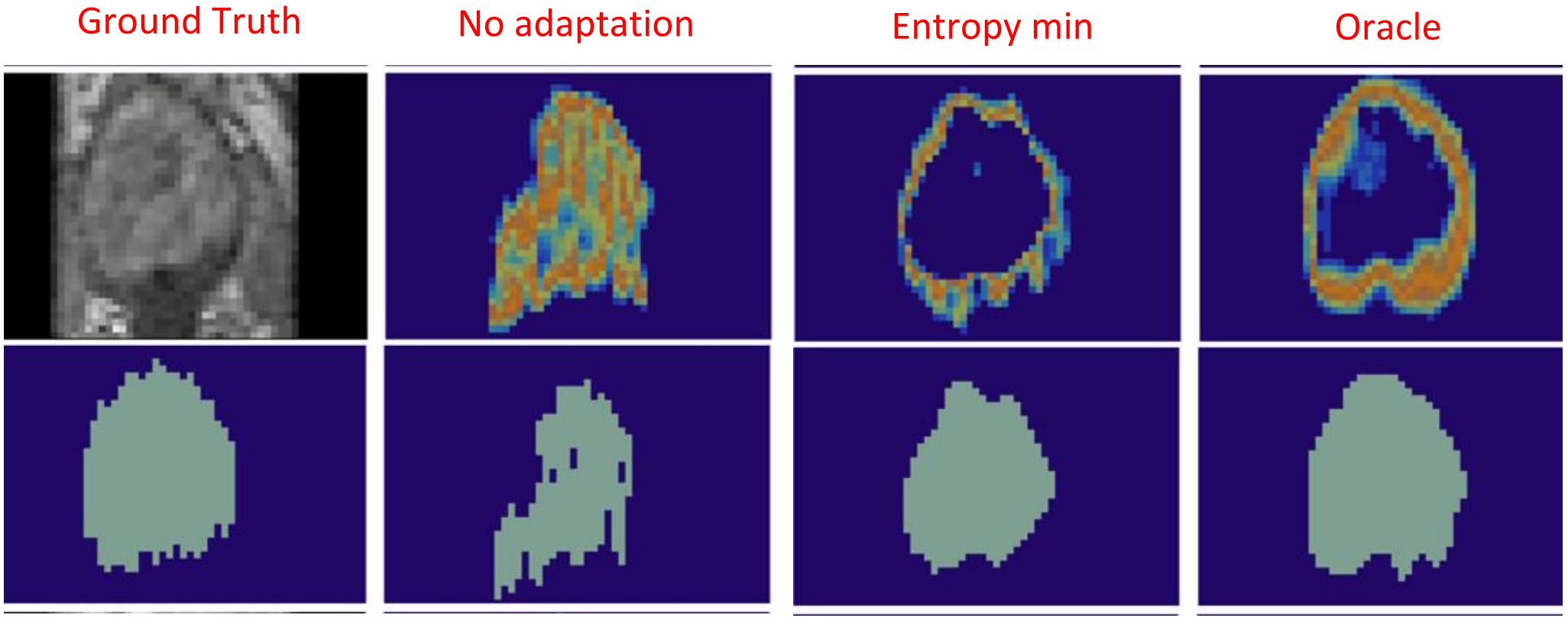
Images from Vu et al., ADVENT: Adversarial Entropy Minimization for Domain Adaptation in Semantic Segmentation, CVPR 2019

Entropy minimization for UDA



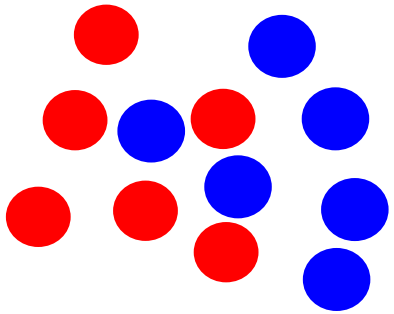
Images from Bateson et al., Source-relaxed domain adaptation for segmentation, MICCAI 2020

Entropy minimization for UDA

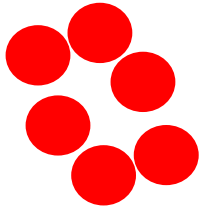


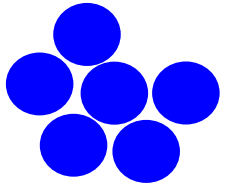
Images from Bateson et al., Source-relaxed domain adaptation for segmentation, MICCAI 2020

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



High entropy
(low confidence)





Low entropy
(high confidence)

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

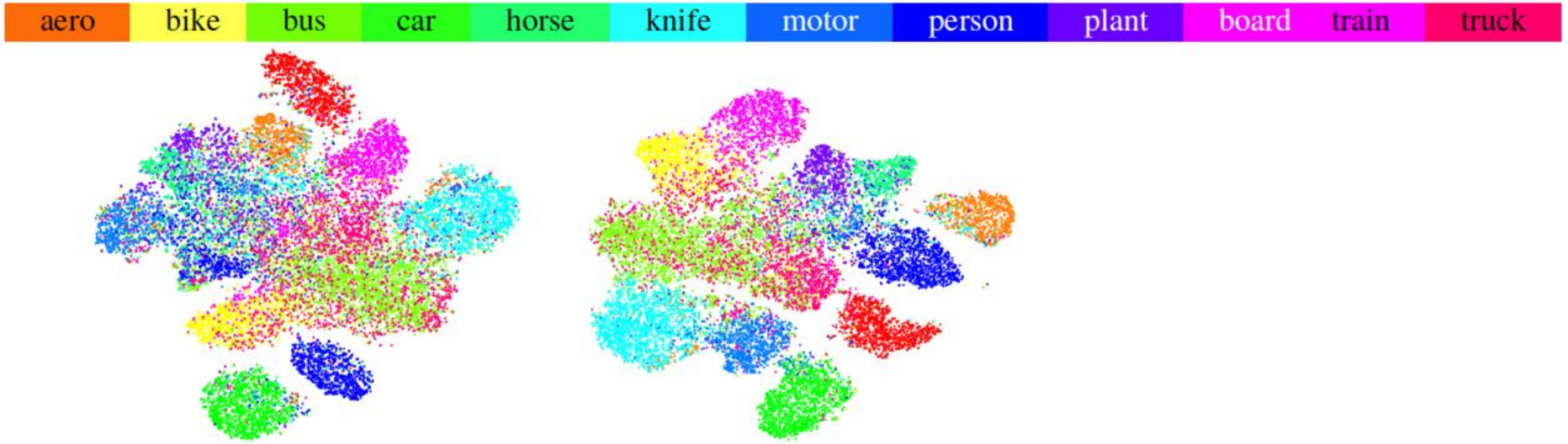
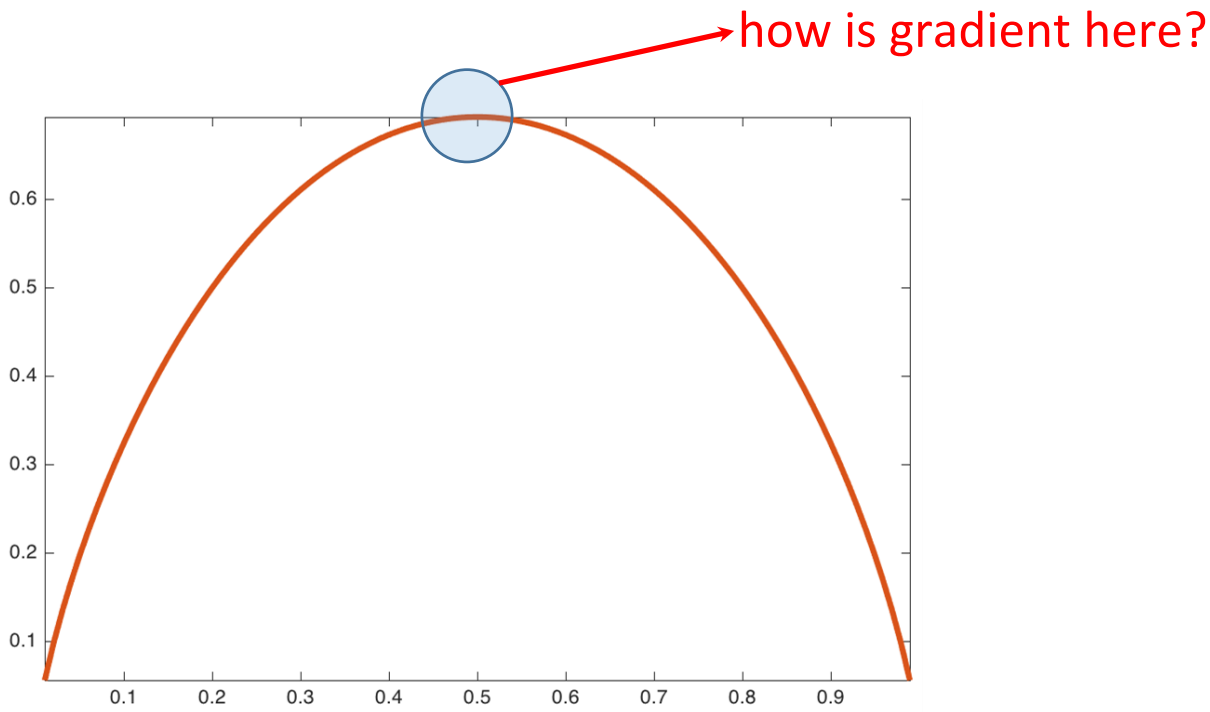


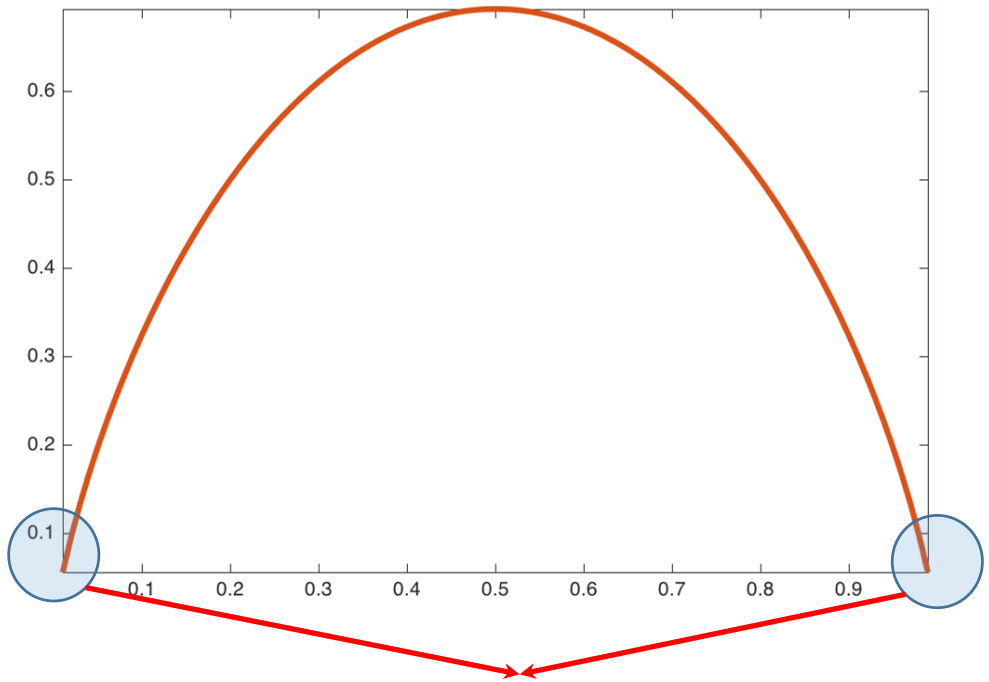
Image classification UDA on VisDA17 data set: Feature visualization for source model (left) and *min-entropy (lower bound on Shannon)* minimization (right) - equivalent to self training (clarified in the next slide)

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

Difficulty of optimizing entropy

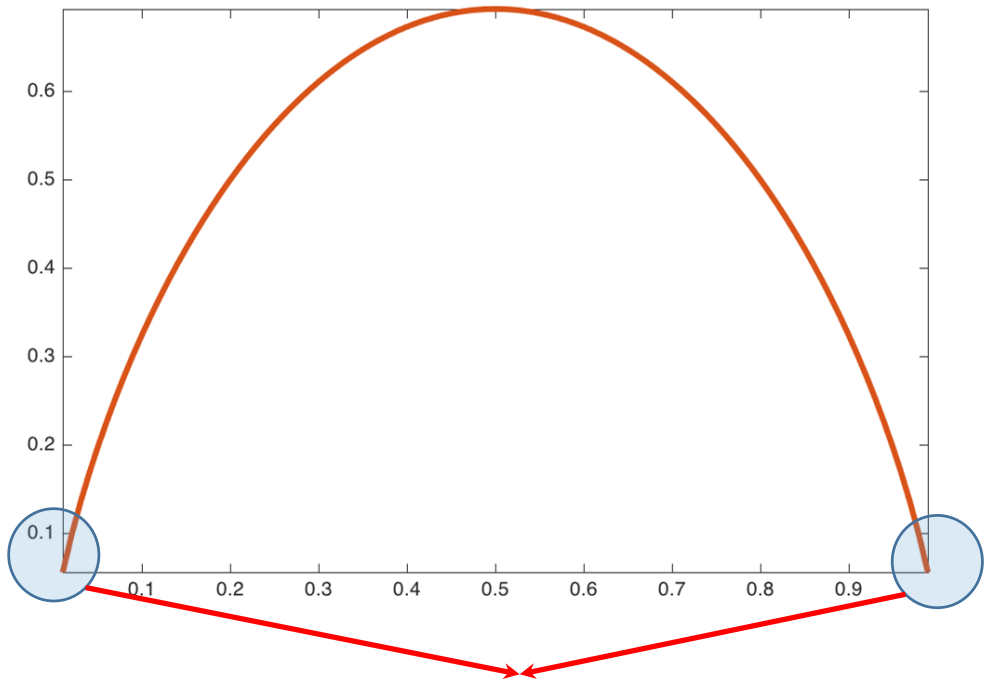


Difficulty of optimizing entropy



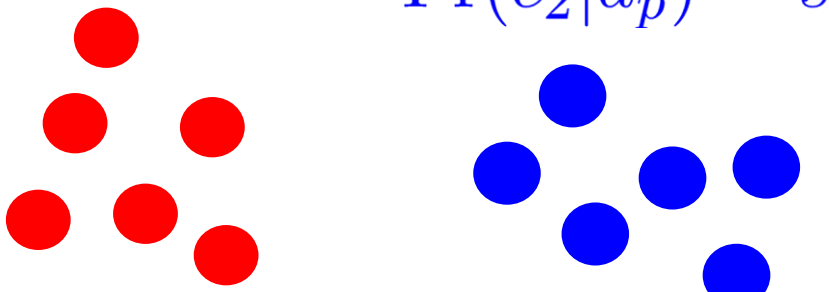
Difficulty of optimizing entropy

Typically we add other cues to facilitate optimization and avoid trivial solutions



Avoiding the trivial solutions of entropy minimization

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

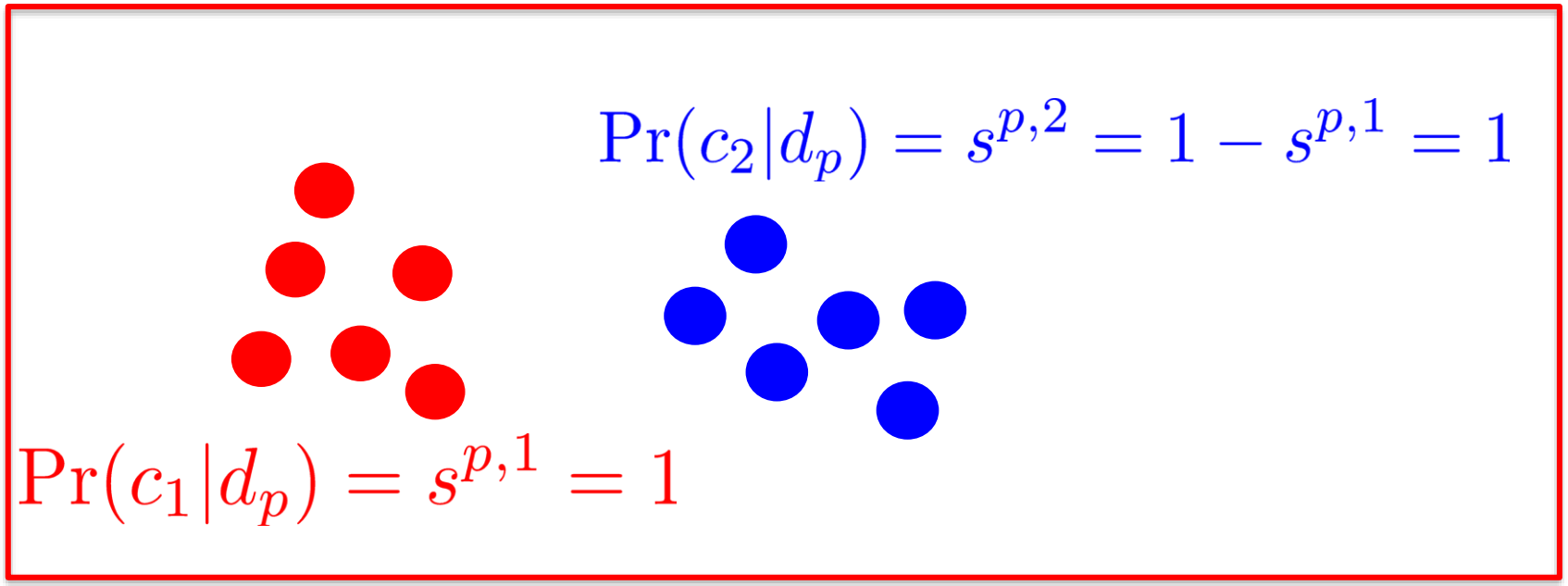


$$\Pr(c_1|d_p) = s^{p,1} = 1$$

Softmax output

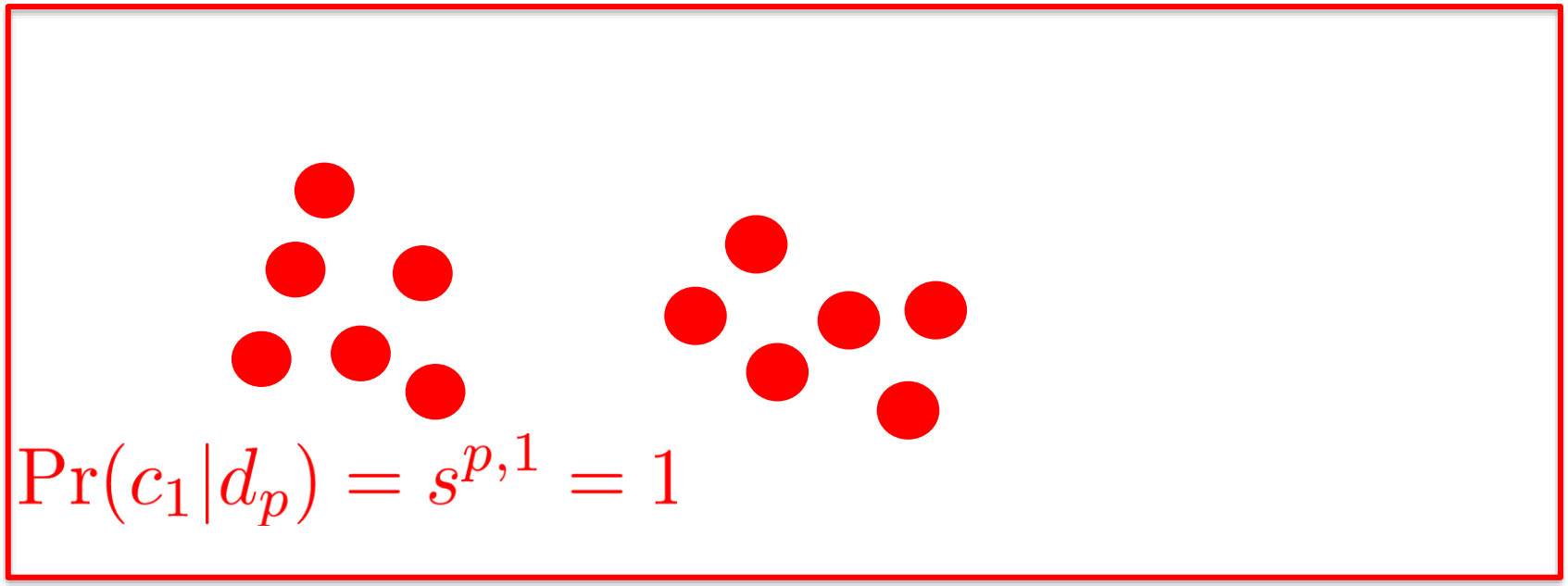
Note: Ignoring the network parameters to simplify notation

Avoiding the trivial solutions of entropy minimization



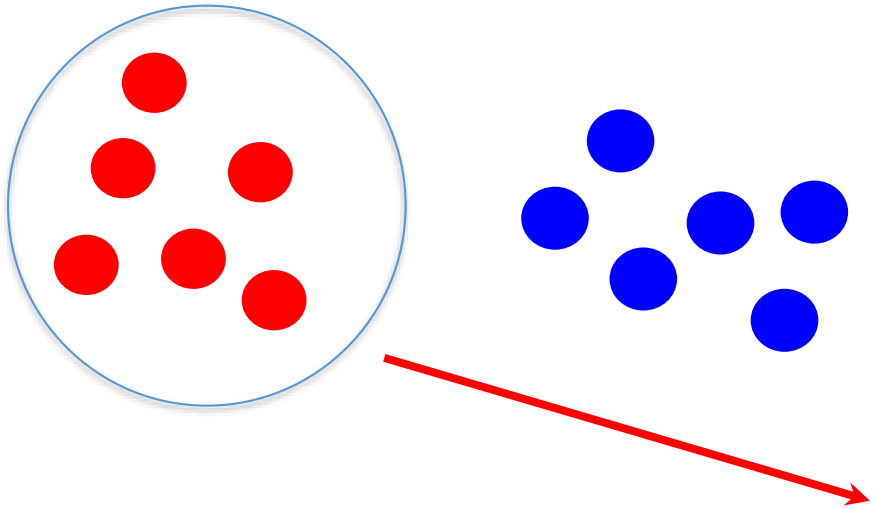
Min entropy
(max confidence)

Avoiding the trivial solutions of entropy minimization



This bad solution also has a minimum entropy!!!

Avoiding the trivial solutions of entropy minimization



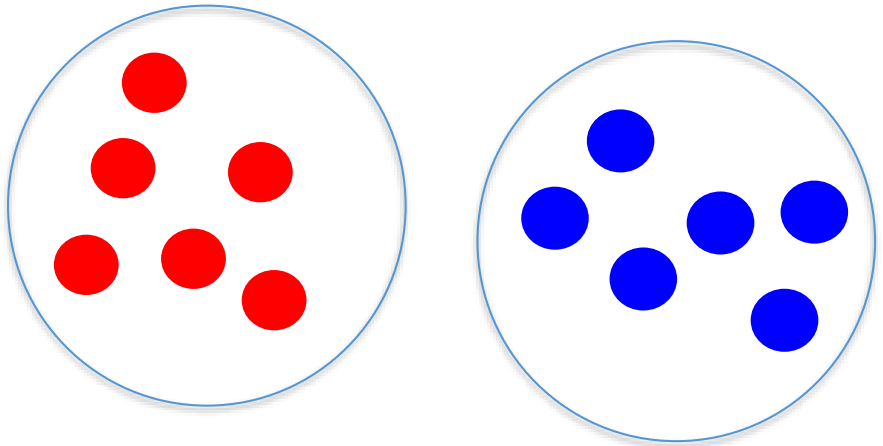
Marginal probabilities of the labels

-- Class proportion

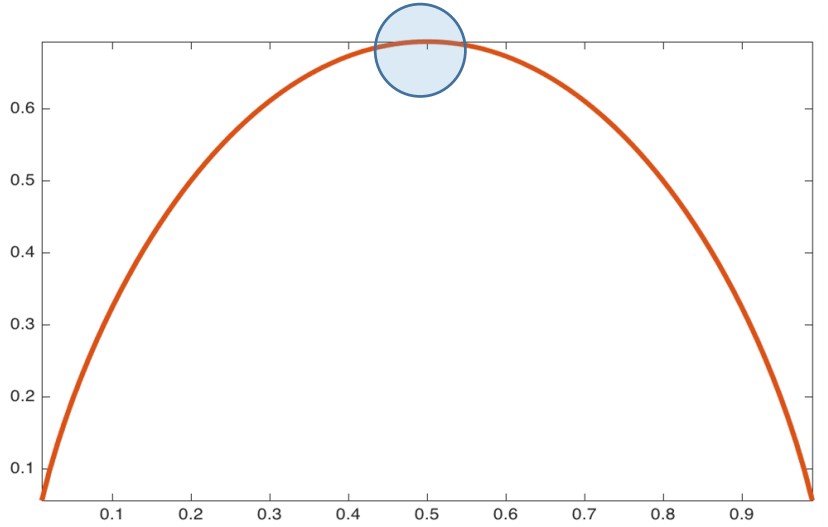
-- Region size (normalized) in segmentation

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

Avoiding the trivial solutions of entropy minimization



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



$\Pr(c_1)$

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

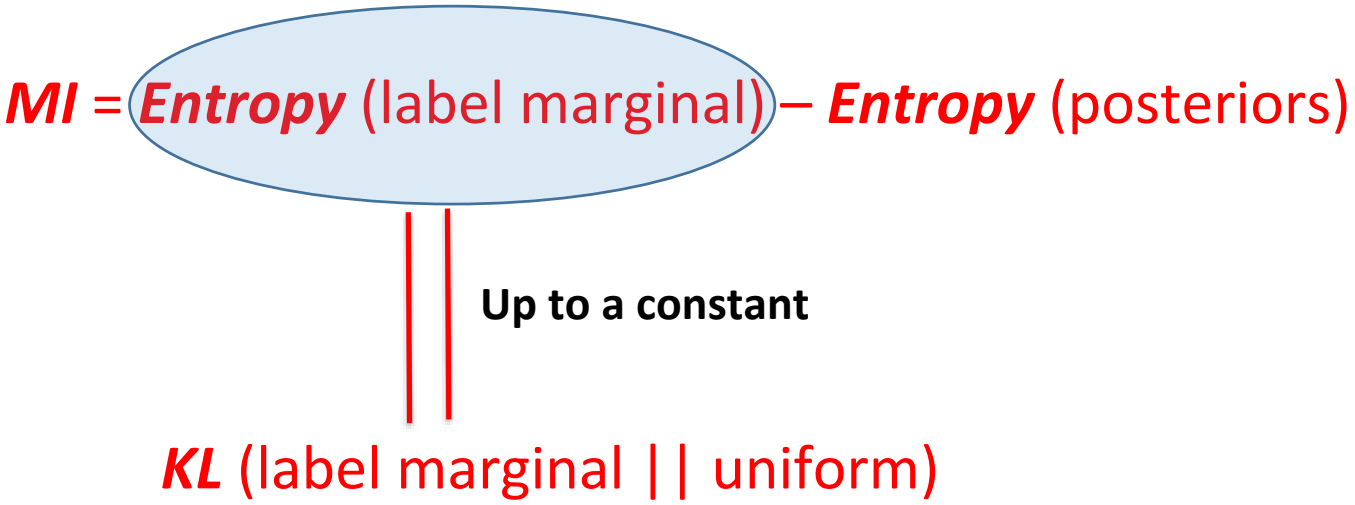
$$I(X,Y) = H(Y) - H(Y/X)$$

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

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

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

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

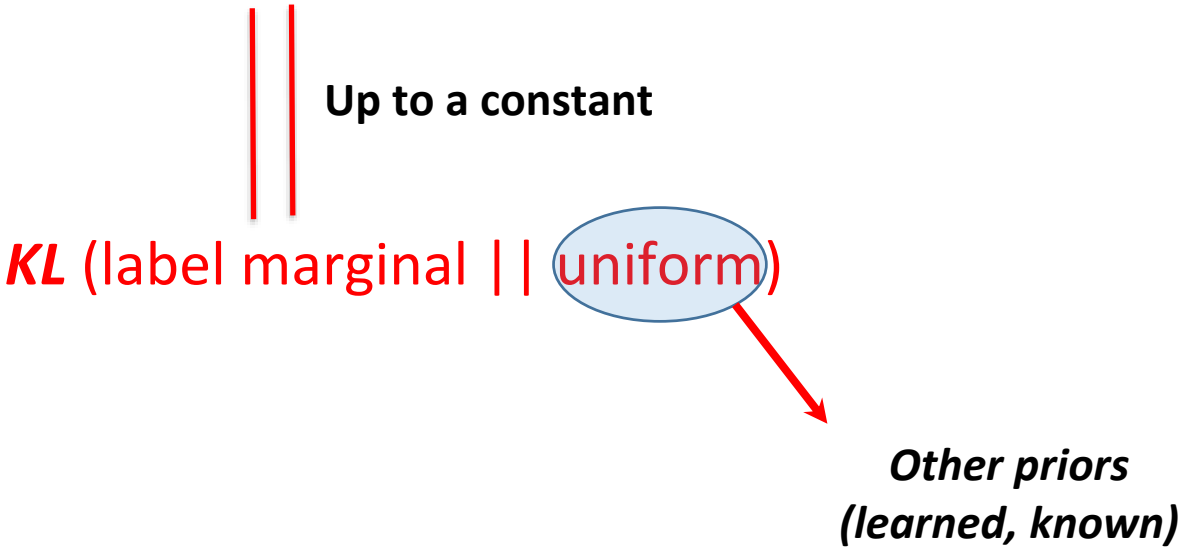


Standard and old in discriminative clustering:

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

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

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

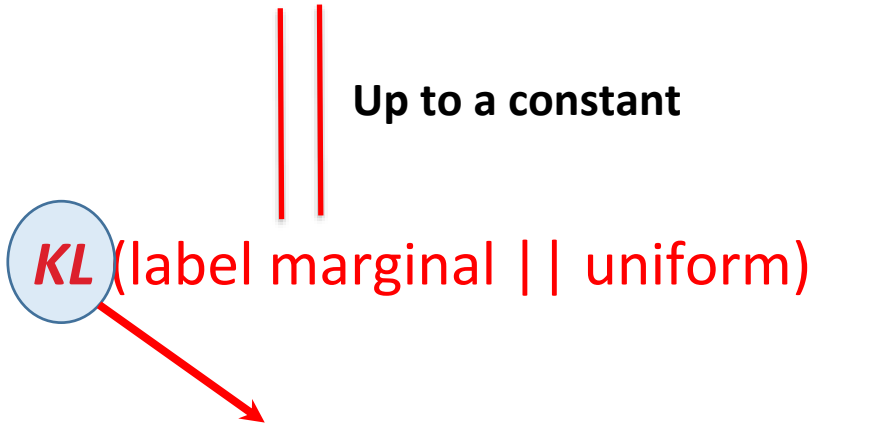


Standard and old in discriminative clustering:

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

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

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



Standard and old in discriminative clustering:

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

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

Semi-supervised learning, e.g.

[Berthelot et al., NeurIPS'19]

[Kervadec et al., MedIA'19]

Few-shot learning, e.g.,

[Boudiaf et al., NeurIPS'20]

[Dhillion et al., ICLR'20]

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

Unsupervised domain adaptation, e.g.,

Liang et al., ICML'20

Bateson et al., MICCAI'20

Deep clustering

&

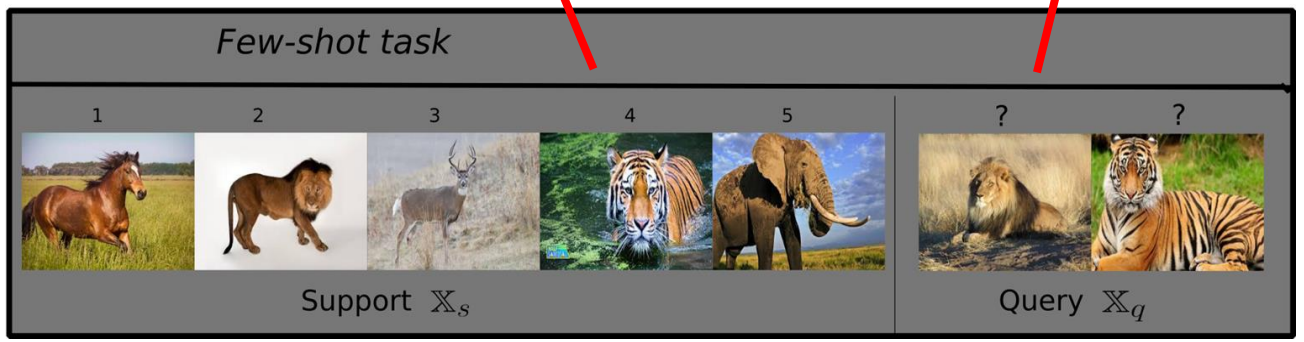
Unsupervised Representation Learning, e.g.,

Asano et al., ICLR'20

Jabi et al., TPAMI'20

Transductive information maximization (TIM) for few-shot learning

Cross-Entropy (over support samples) + MI (over queries)

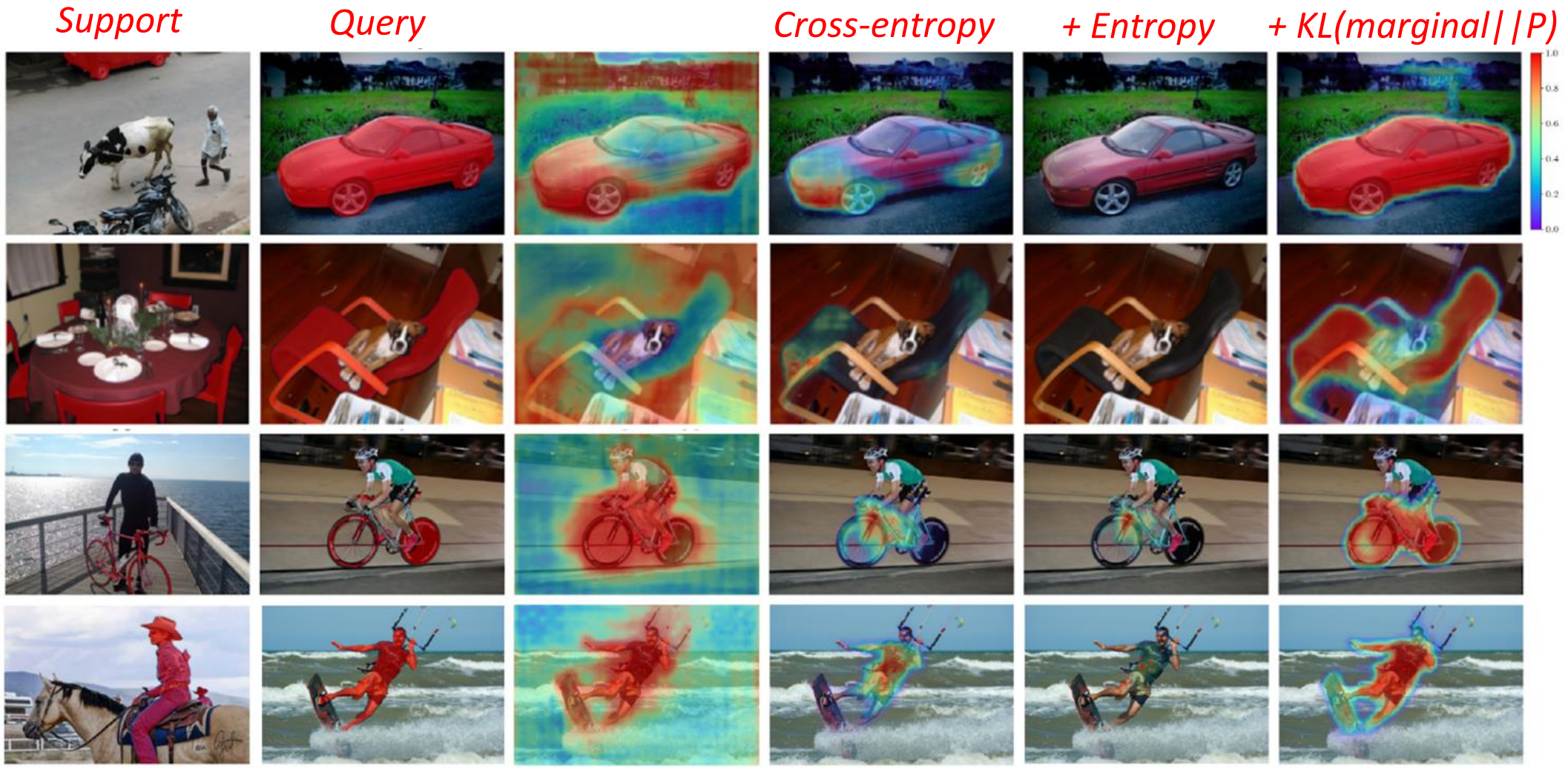


TIM inference:
Gradient optimization w.r.t classifier (last-layer) weights

Boudiaf et al., Transductive information maximization for few-shot learning, NeurIPS 2020

Few-shot segmentation

A good transductive inference is all you need?



Boudiaf et al., Few-shot segmentation without meta-learning: A good transductive inference is all you need? CVPR 2021

SOTA performance by a good margin

Method	Backbone	1 shot					5 shot				
		Fold-0	Fold-1	Fold-2	Fold-3	Mean	Fold-0	Fold-1	Fold-2	Fold-3	Mean
OSLSM [29] (BMVC'18)	VGG-16	33.6	55.3	40.9	33.5	40.8	35.9	58.1	42.7	39.1	43.9
co-FCN [25] (ICLRW'18)		36.7	50.6	44.9	32.4	41.1	37.5	50.0	44.1	33.9	41.4
AMP [30] (ICCV'19)		41.9	50.2	46.7	34.7	43.4	41.8	55.5	50.3	39.9	46.9
PANet [37] (ICCV'19)		42.3	58.0	51.1	41.2	48.1	51.8	64.6	59.8	46.5	55.7
FWB [23] (ICCV'19)		47.0	59.6	52.6	48.3	51.9	50.9	62.9	56.5	50.1	55.1
SG-One [42] (TCYB'20)		40.2	58.4	48.4	38.4	46.3	41.9	58.6	48.6	39.4	47.1
CRNet [19] (CVPR'20)		-	-	-	-	55.2	-	-	-	-	58.5
FSS-1000 [17] (CVPR'20)		-	-	-	-	-	37.4	60.9	46.6	42.2	56.8
RPMM [21] (ECCV'20)		47.1	65.8	50.6	48.5	53.0	50.0	66.5	51.9	47.6	54.0
CANet [41] (CVPR'19)	ResNet50	52.5	65.9	51.3	51.9	55.4	55.5	67.8	51.9	53.2	57.1
PGNet [40] (ICCV'19)		56.0	66.9	50.6	50.4	56.0	57.7	68.7	52.9	54.6	58.5
CRNet [19] (CVPR'20)		-	-	-	-	55.7	-	-	-	-	58.8
SimPropNet [10] (IJCAI'20)		54.9	67.3	54.5	52.0	57.2	57.2	68.5	58.4	56.1	60.0
LTM [39] (MMMM'20)		52.8	69.6	53.2	52.3	57.0	57.9	69.9	56.9	57.5	60.6
RPMM [38] (ECCV'20)		55.2	66.9	52.6	50.7	56.3	56.3	67.3	54.5	51.0	57.3
PPNet [21] (ECCV'20)*		47.8	58.8	53.8	45.6	51.5	58.4	67.8	64.9	56.7	62.0
PFENet [33] (TPAMI'20)		**61.7**	**69.5**	55.4	**56.3**	**60.8**	63.1	70.7	55.8	57.9	61.9
RePRI (ours)		60.2	67.0	**61.7**	47.5	59.1	**64.5**	**70.8**	**71.7**	**60.3**	**66.8**
Oracle-RePRI	ResNet50	72.4	78.0	77.1	65.8	73.3	75.1	80.8	81.4	74.4	77.9
FWB [23] (ICCV'19)	ResNet101	51.3	64.5	56.7	52.2	56.2	54.9	67.4	62.2	55.3	59.9
DAN [36] (ECCV'20)		54.7	68.6	57.8	51.6	58.2	57.9	69.0	60.1	54.9	60.5
PFENet [33] (TPAMI'20)		**60.5**	**69.4**	54.4	**55.9**	**60.1**	62.8	70.4	54.9	57.6	61.4
RePRI (ours)		59.6	68.6	**62.2**	47.2	59.4	**66.2**	**71.4**	**67.0**	**57.7**	**65.6**

Few-shot tasks with varying structure

(e.g., number of support examples)

	PASCAL-5^i			COCO-20^i		
Method	1-S	5-S	10-S	1-S	5-S	10-S
RPMM [38]	56.3	57.3	57.6	30.6	35.5	33.1
PFENet [33]	**60.8**	61.9	62.1	**35.8**	39.0	39.7
RePRI (ours)	59.1	**66.8**	**68.2**	34.0	**42.1**	**44.4**

Few-shot tasks with varying structure

(e.g., number of support examples)

Method	PASCAL-5^i			COCO-20^i		
	1-S	5-S	10-S	1-S	5-S	10-S
RPMM [38]	56.3	57.3	57.6	30.6	35.5	33.1
PFENet [33]	**60.8**	61.9	62.1	**35.8**	39.0	39.7
RePRI (ours)	59.1	**66.8**	**68.2**	34.0	**42.1**	**44.4**

Performance saturates despite the larger number of support examples

Experiments with domain shifts

		COCO → PASCAL	
Method	Backbone	1 shot	5 shot
RPMM [38]		49.6	53.8
PFENet [33]	ResNet50	61.1	63.4
RePRI (ours)		**63.2**	**67.7**

A final note on the link between self-training and entropy-min

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

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

A final note on the link between self-training and entropy-min

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

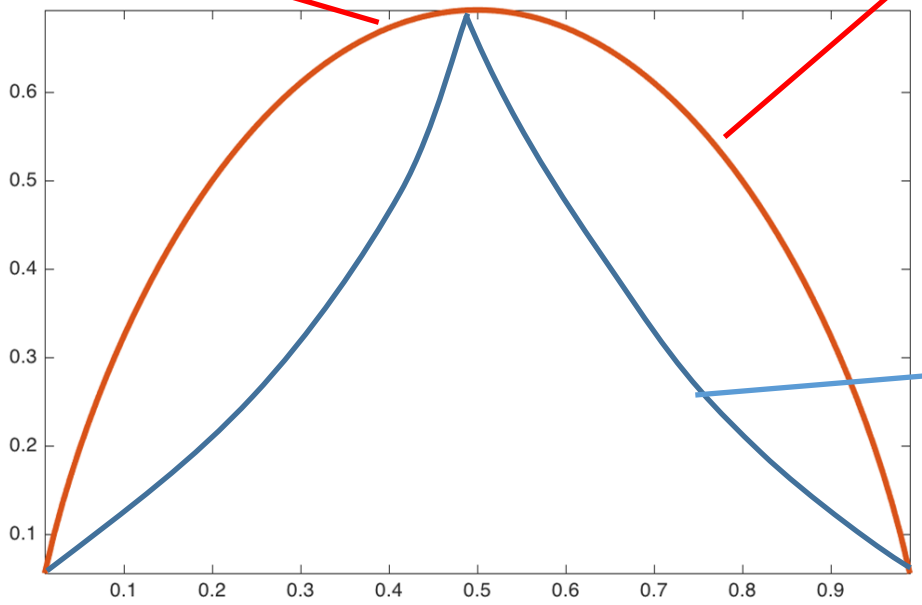
Or equivalently (re-writing without pseudo-labels):
Min-entropy (a lower bound on Shannon entropy)

- 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

A final note on the link between self-training and entropy-min

Small gradients for non-confident predictions

Shannon entropy



Min-entropy

Dolz et al., Teach me to segment with mixed supervision: Confident students becomes masters, IPMI 2021

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 also competitive for UDA/SSL)

GTA5 to Cityscapes adaptation

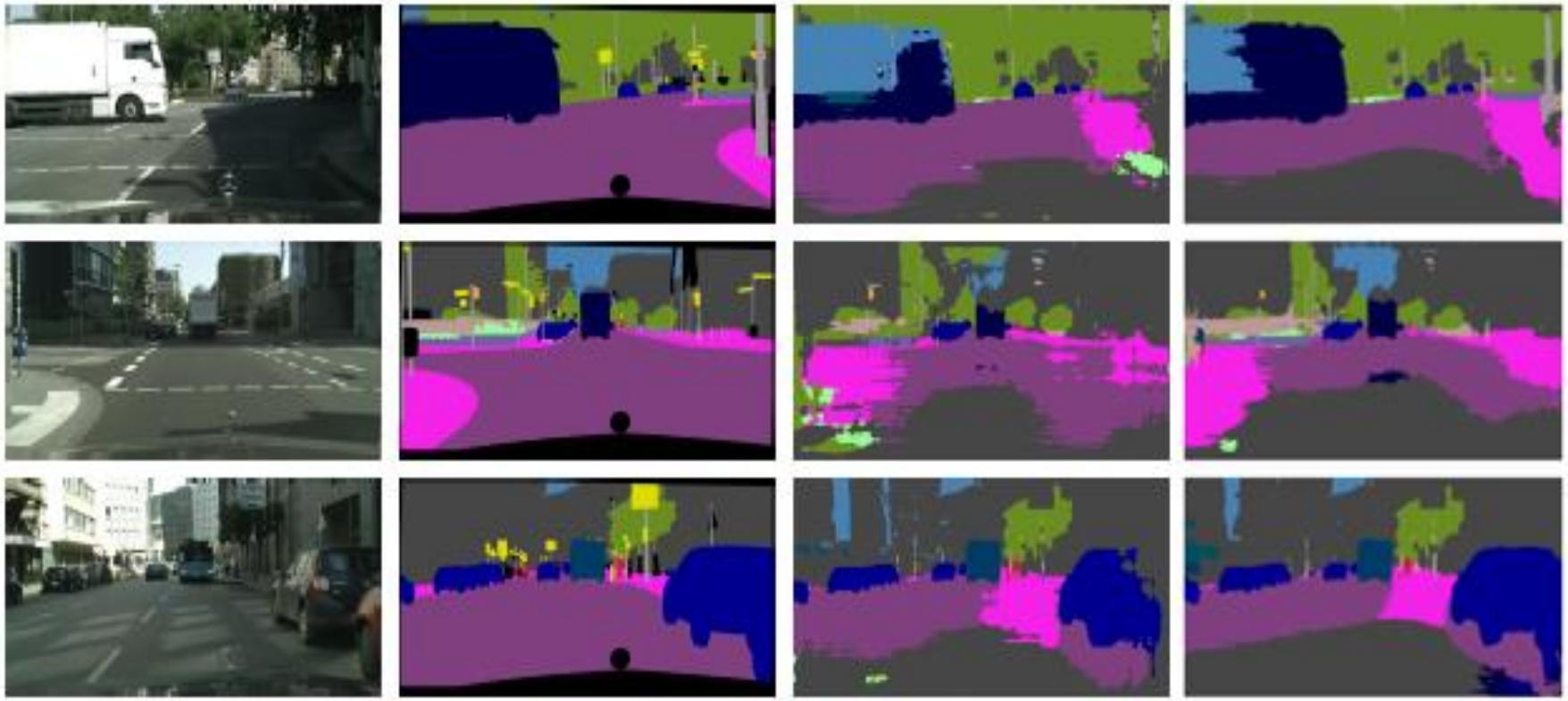


image GT Source model Self-Training

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