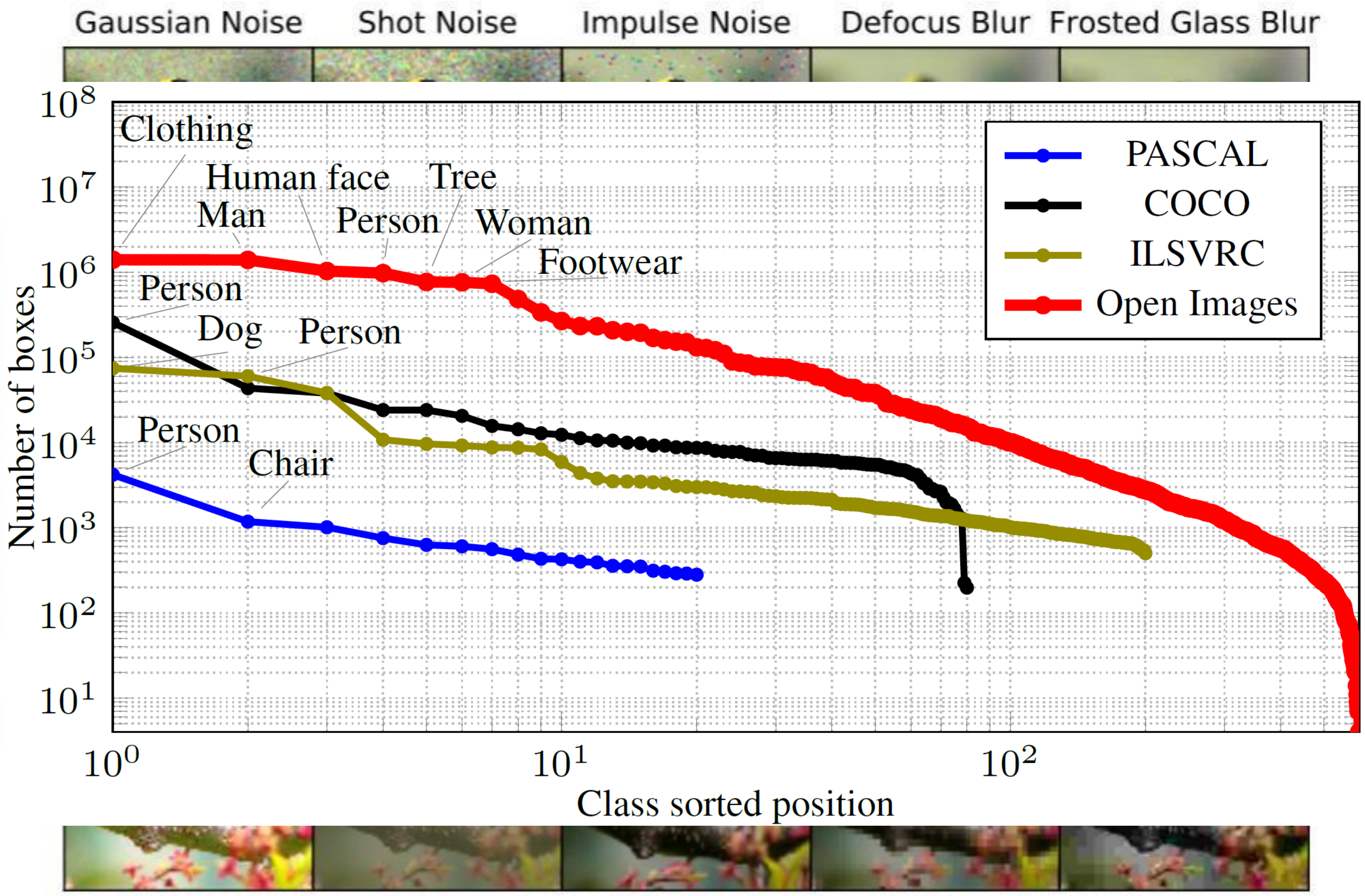
Learning with Imperfect Labels and Visual Data

Zhiding Yu August 23rd 2020

New Frontiers for Learning with Limited Labels or Data

Challenge I - Real World Data Are Imperfect

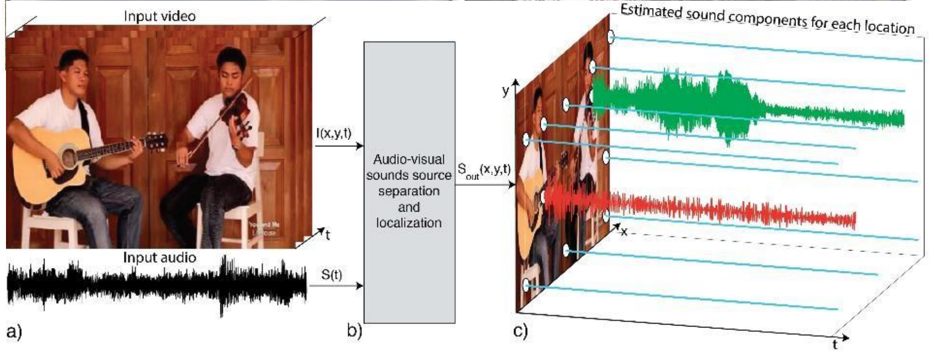
- **Domain gap**
- **Data bias**
- **Data noise**
- Can be *Long tail*
- Can contain *Occlusions*
- Can be *Cluttered*
- Can be *Ambiguous*
- Can be *Multisensory*
- …

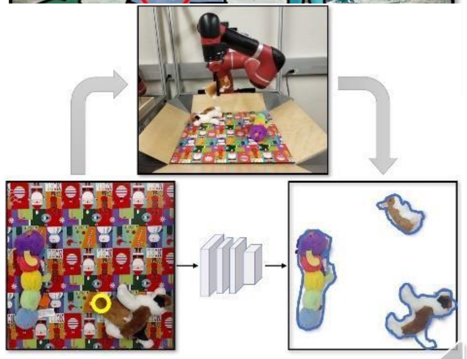


Challenge II – Real World Labels Are Imperfect

- **Inexact** (indirect) supervision
- **Incomplete** (limited) supervision
- **Inaccurate** (noisy) supervision
- Can be *Versatile*
- Can be *Multimodal*
- Can be *Sequential*
- Can be *Sparse*
- Can be *Interactive*

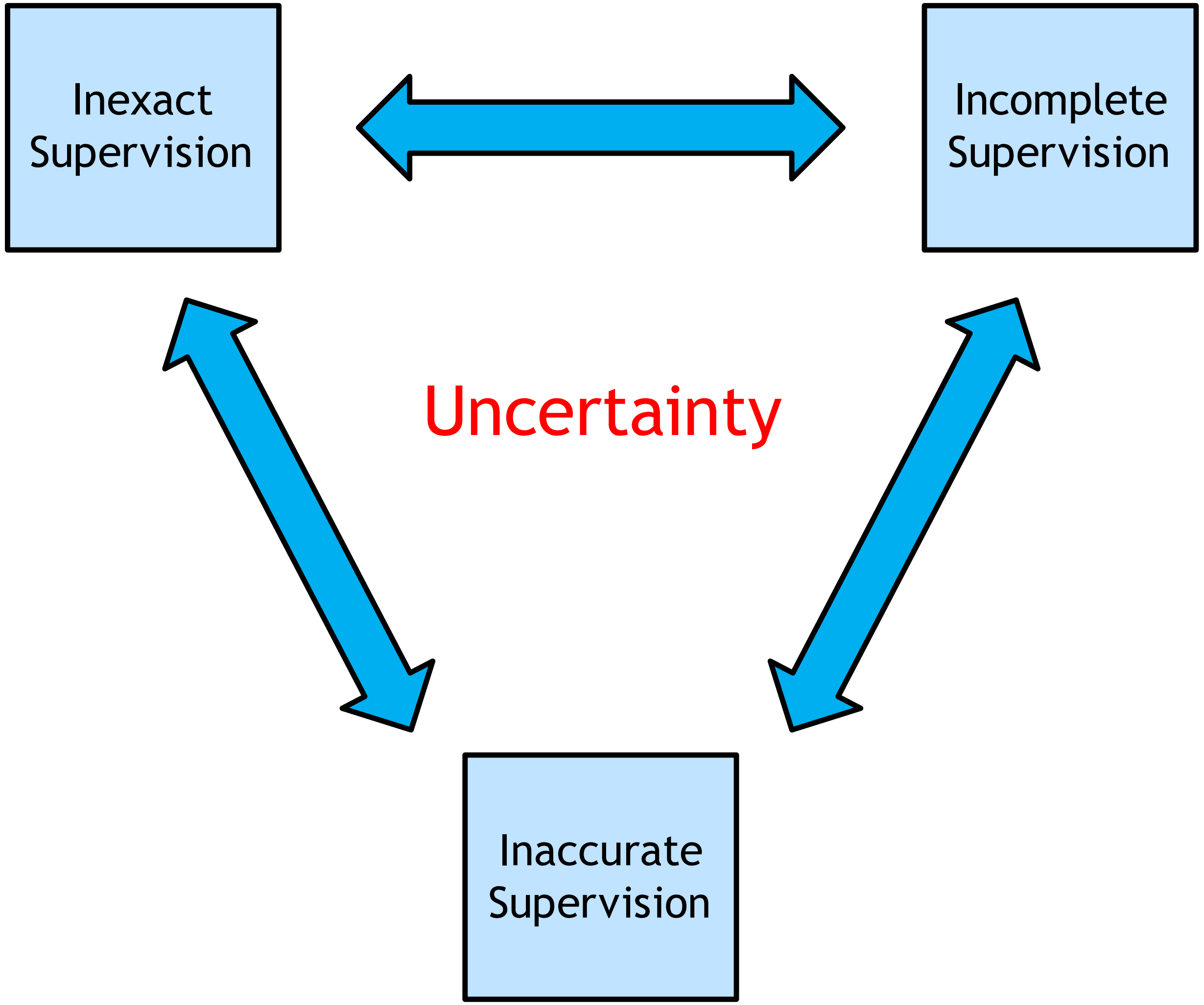
…



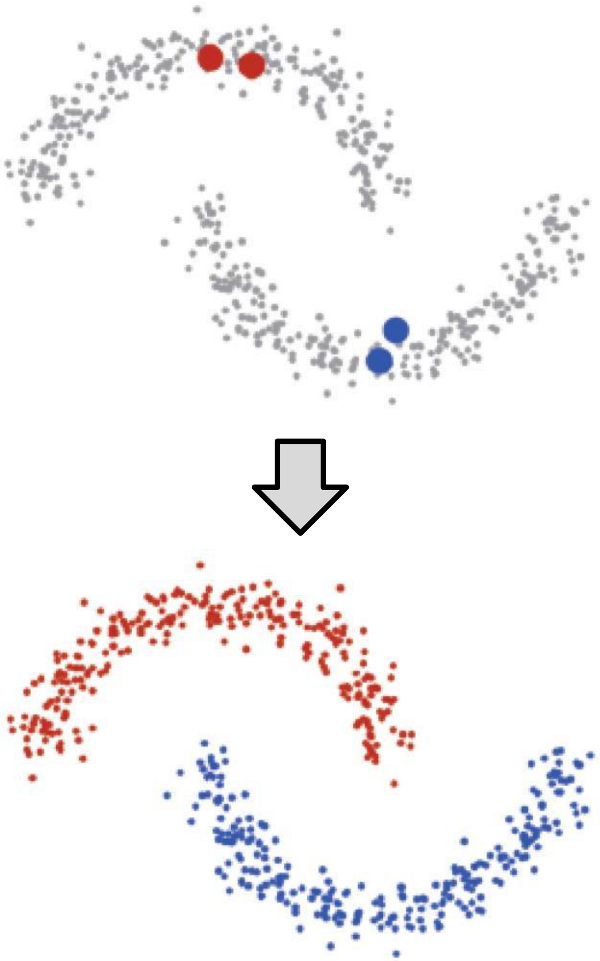


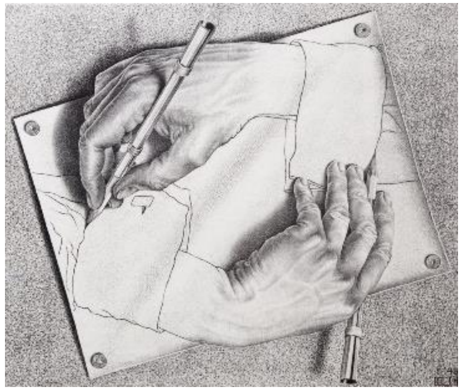
Goal and Challenge

- The ability to improve generalization by learning continuously on new data
- The ability to leverage diverse forms of weak supervision, or self-improving on unlabeled data
- The ability to overcome the uncertainty (ill-posed nature) due to the lack of constraints under partial supervision

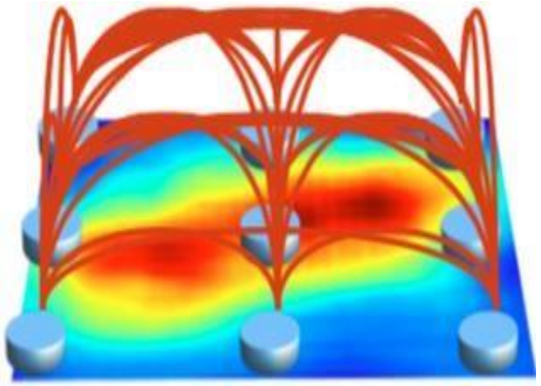


Overcoming Uncertainties from Imperfect Labels

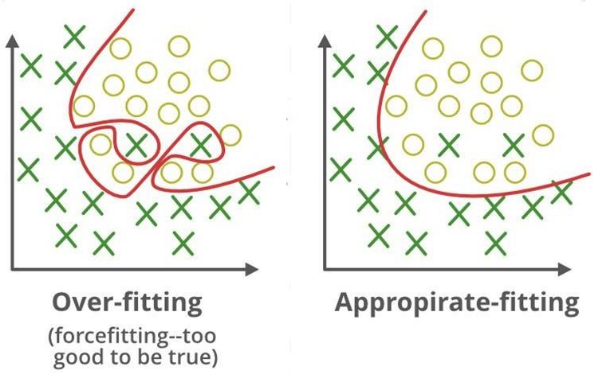




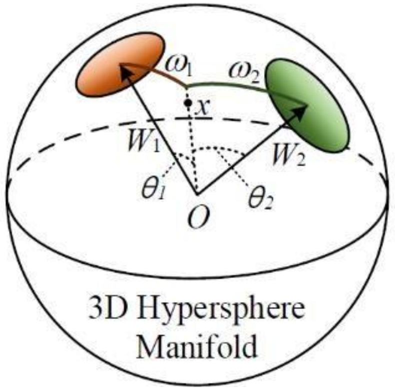
Self-supervision



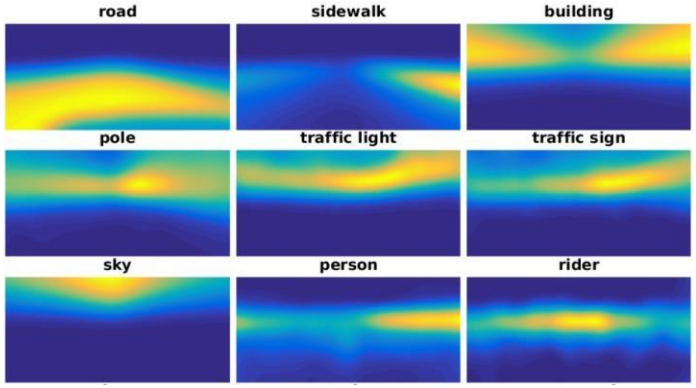
Structuredness



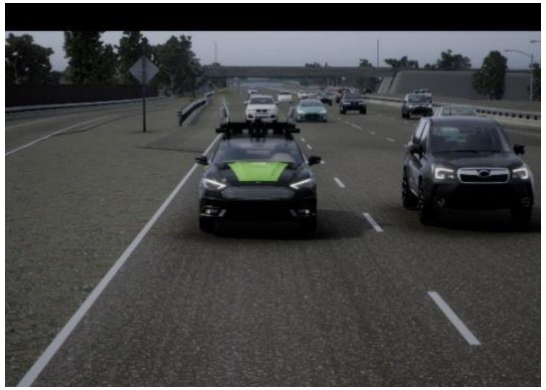
Regularization



Inductive Bias



Domain Prior



Synthetic Data

Part I: Learning on Unlabeled Data with Proxy-Label Approaches

Semi-Supervised Learning with Self-Ensembling

Labeled data $\mathcal{L} = \{(x_i, y_i)\}_{i=1}^{L}$, unlabeled data $\mathcal{U} = \{x_i\}_{i=L+1}^{N}$, $L \ll N$.

Learn $f : \mathcal{X} \to [0,1]^K$ parameterized by $\theta \in \Theta$

$$\min_{\theta} \sum_{i=1}^{L} \ell(f(x_i; \theta), y_i) + \lambda R(\theta, \mathcal{L}, \mathcal{U}),$$

where R is the regularization term leveraging unlabeled data.

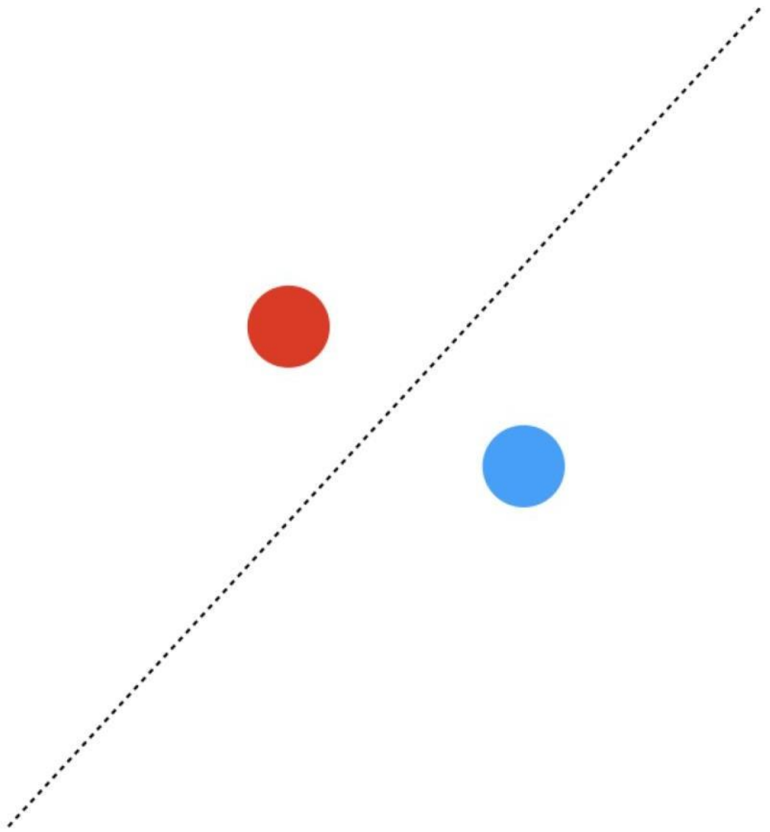
Core idea: Perturbing the input slightly and requiring that the prediction remains the same.

Define R as a consistency loss

$$R_C(\theta, \mathcal{L}, \mathcal{U}) = \sum_{i=1}^{N} \mathbb{E}_{\xi', \xi}\, d\big(\tilde{f}(x_i; \theta', \xi'), f(x_i; \theta, \xi)\big)$$

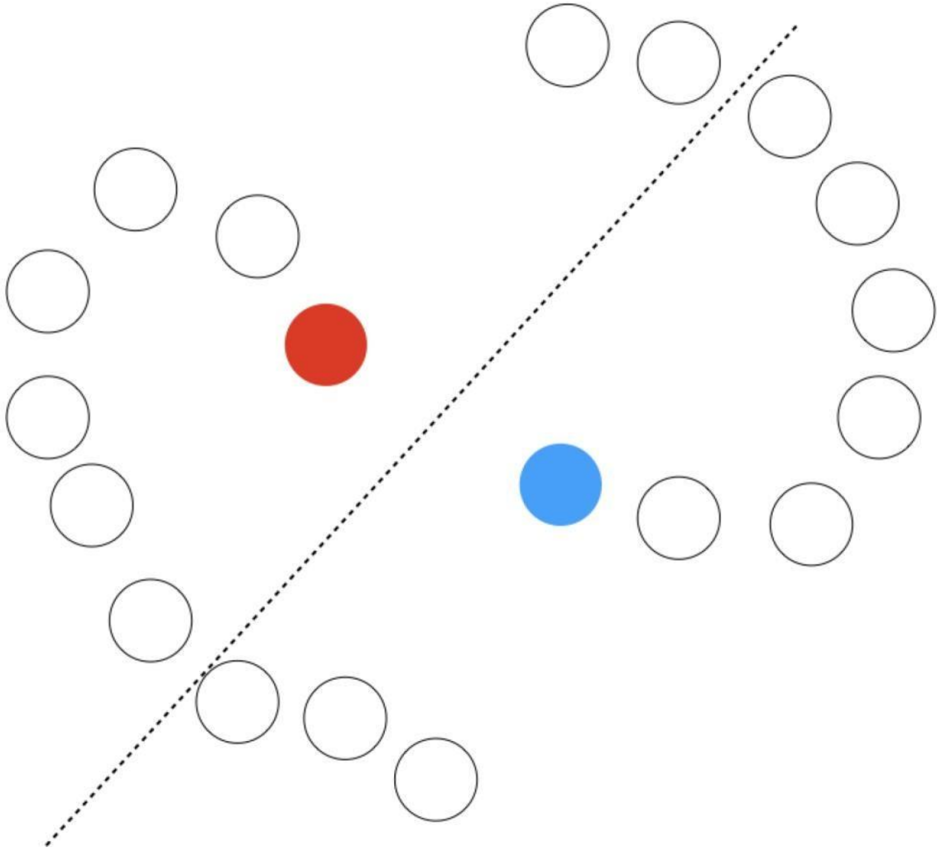
Luo et al., Smooth Neighbors on Teacher Graphs for Semi-supervised Learning, CVPR18

How Does Self-Ensembling Work?



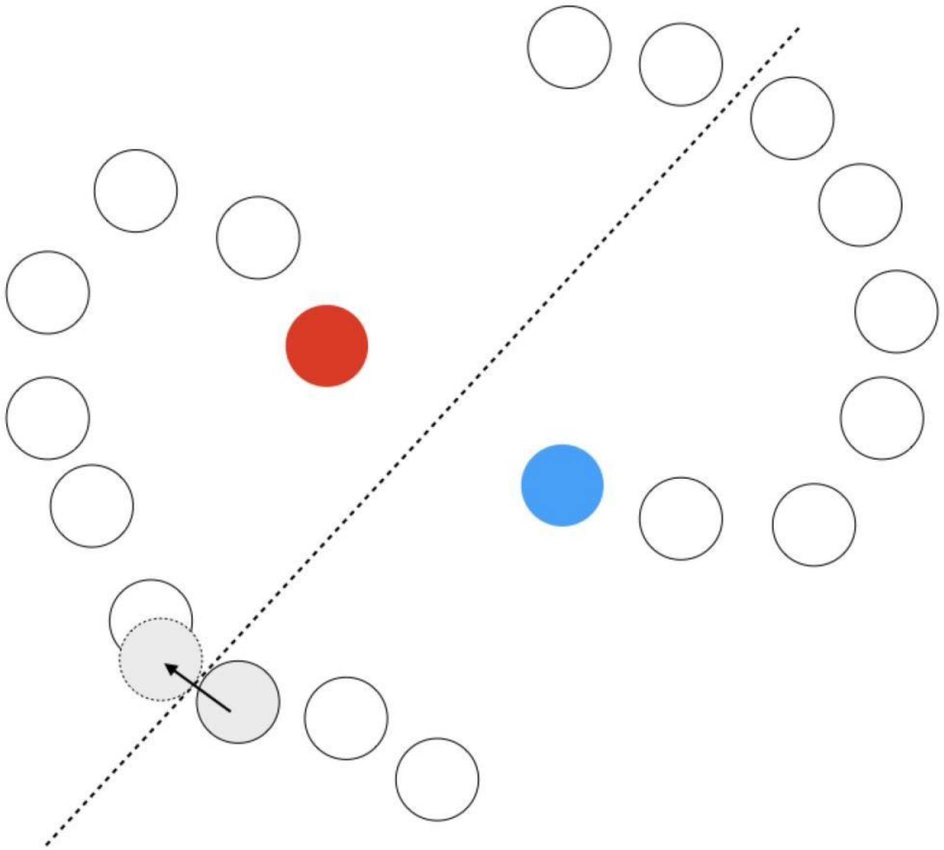
Luo et al., Smooth Neighbors on Teacher Graphs for Semi-supervised Learning, CVPR18

How Does Self-Ensembling Work?



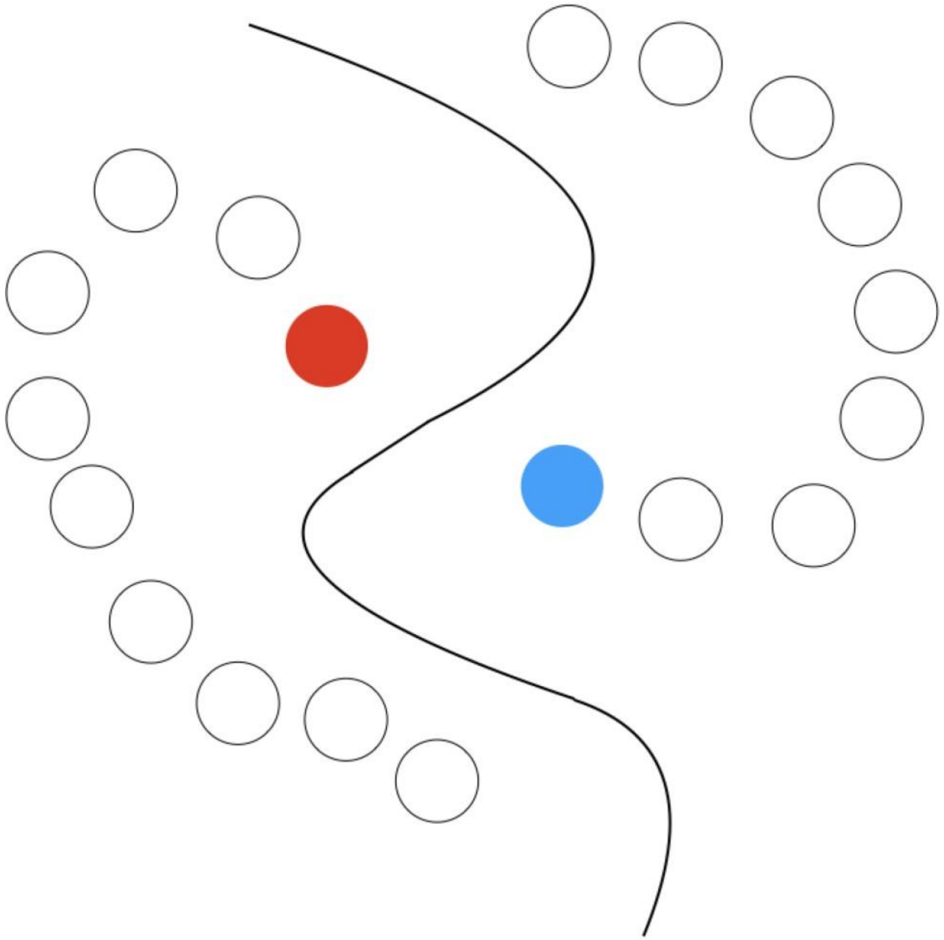
Luo et al., Smooth Neighbors on Teacher Graphs for Semi-supervised Learning, CVPR18

How Does Self-Ensembling Work?



Luo et al., Smooth Neighbors on Teacher Graphs for Semi-supervised Learning, CVPR18

How Does Self-Ensembling Work?



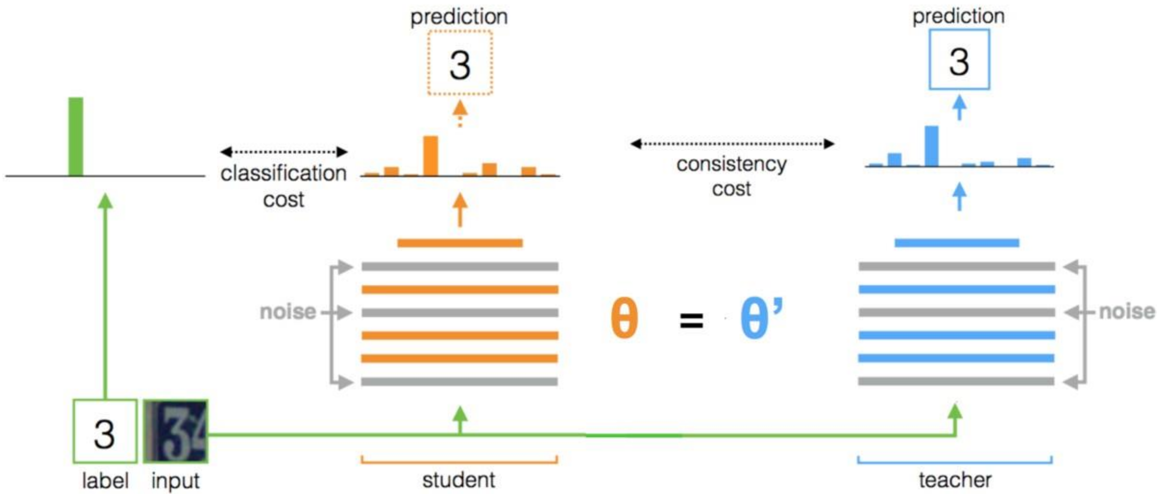
Luo et al., Smooth Neighbors on Teacher Graphs for Semi-supervised Learning, CVPR18

Self-Ensembling Methods

Π Model

Use a noisy teacher $\tilde{f}$ to alleviate the bias in targets with $\theta' = \theta$

$$R_C(\theta, \mathcal{L}, \mathcal{U}) = \sum_{i=1}^{N} \mathbb{E}_{\xi', \xi} \, \|\tilde{f}(x_i; \theta, \xi') - f(x_i; \theta, \xi)\|^2,$$

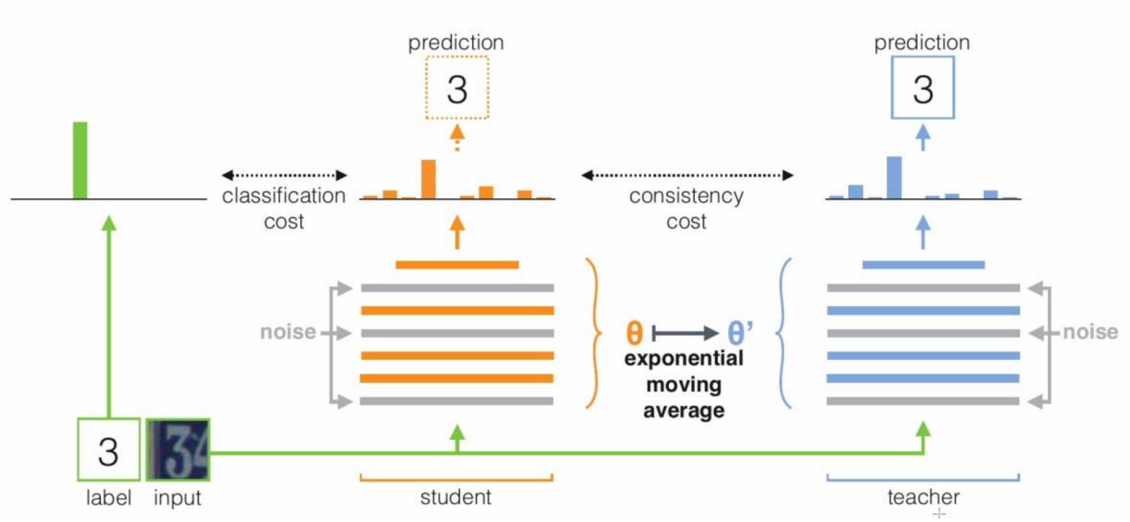


Temporal Ensembling for Semi-Supervised Learning, ICLR16

Mean Teacher Model

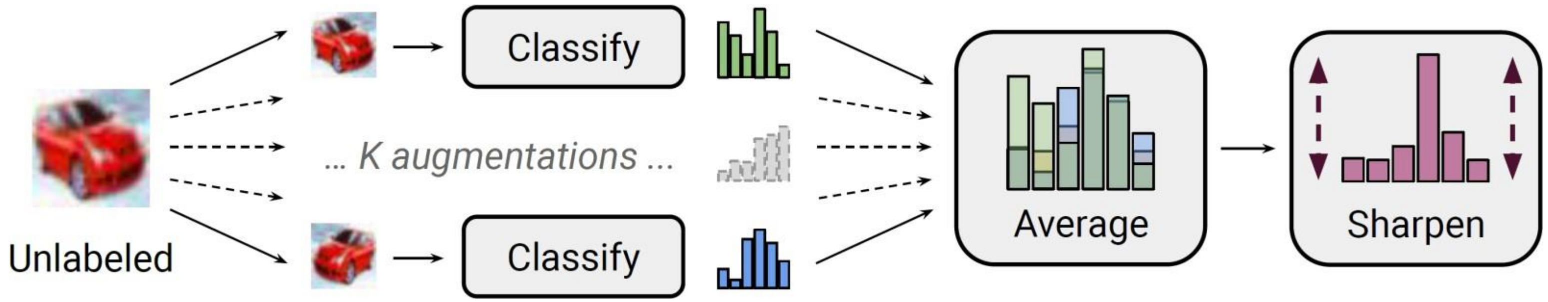
A better (more accurate) teacher using the EMA of parameters θ

$$\theta'_t \leftarrow \alpha \theta'_{t-1} + (1 - \alpha)\theta_t$$



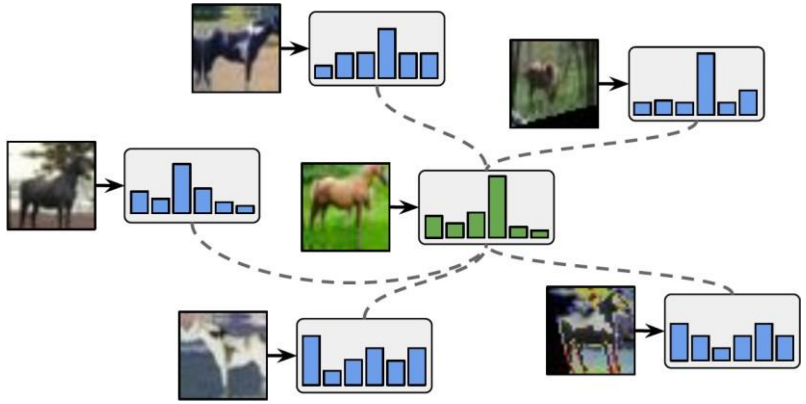
Mean Teachers Are Better Role Models: Weight-averaged Consistency Targets Improve Semi-supervised Deep Learning Results, NIPS17

Semi-Sup Learning with Pseudo-Labels

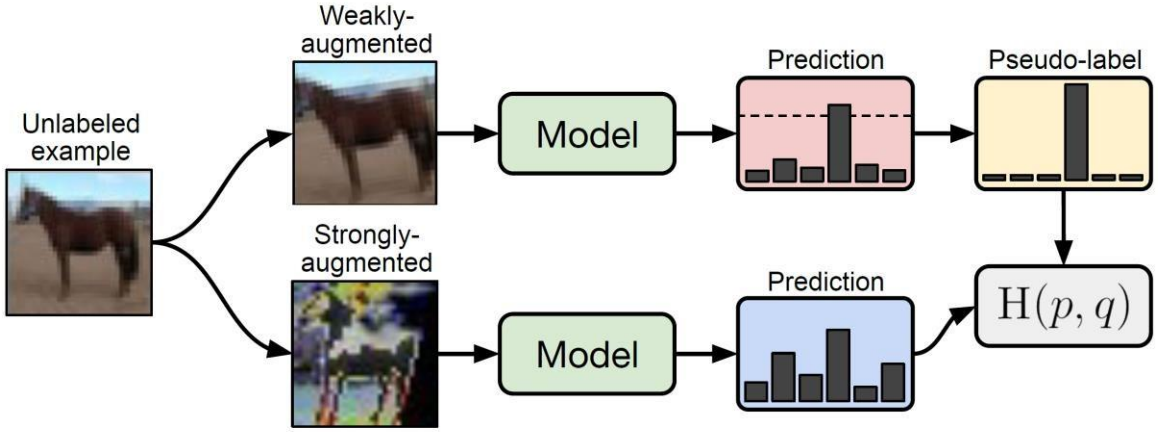


MixMatch

ReMixMatch



FixMatch



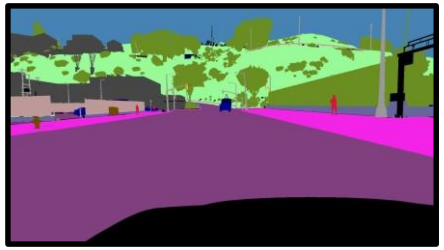
Berthelot et al., MixMatch: A Holistic Approach to Semi-supervised Learning, NeurIPS19
Berthelot et al., ReMixMatch: Semi-supervised Learning with Distribution Alignment and Augmentation Anchoring, ICLR20
Sohn et al., FixMatch: Simplifying Semi-supervised Learning with Consistency and Confidence, arXiv:2001.07685

Unsupervised Domain Adaptation (UDA)

Image classification

Car

Semantic segmentation



Source Domain (Labeled)

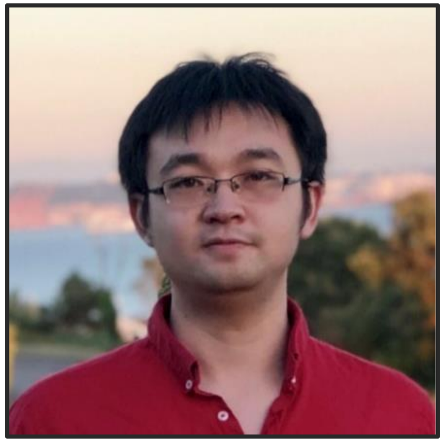
Adaptation

Adaptation

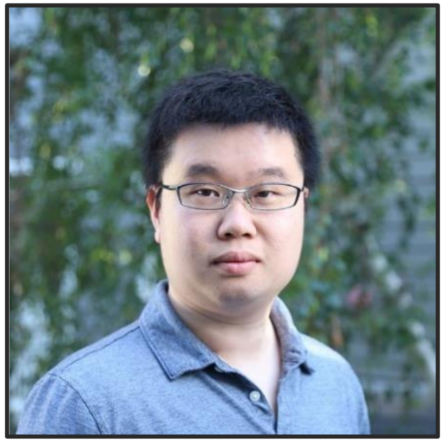
Target Domain (Unlabeled)

Domain Adaptation for Semantic Segmentation via Class-Balanced Self-Training (ECCV18)

Confidence Regularized Self-Training (ICCV19)

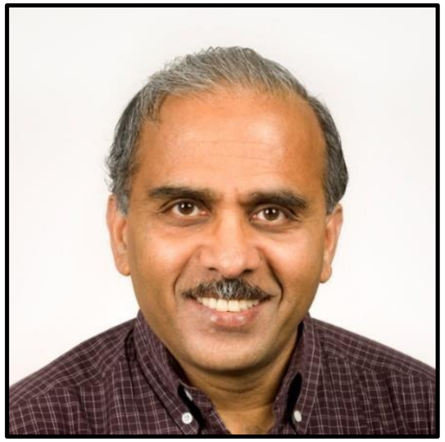


Yang Zou, CMU

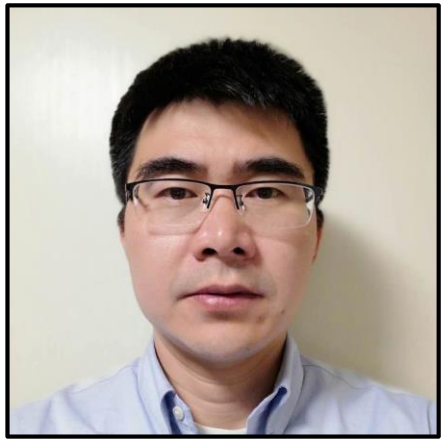


Zhiding Yu, NVIDIA

Xiaofeng Liu, CMU



Vijayakumar Bhagavatula, CMU



Jinsong Wang, GM

Class-Balanced Self-Training (CBST)

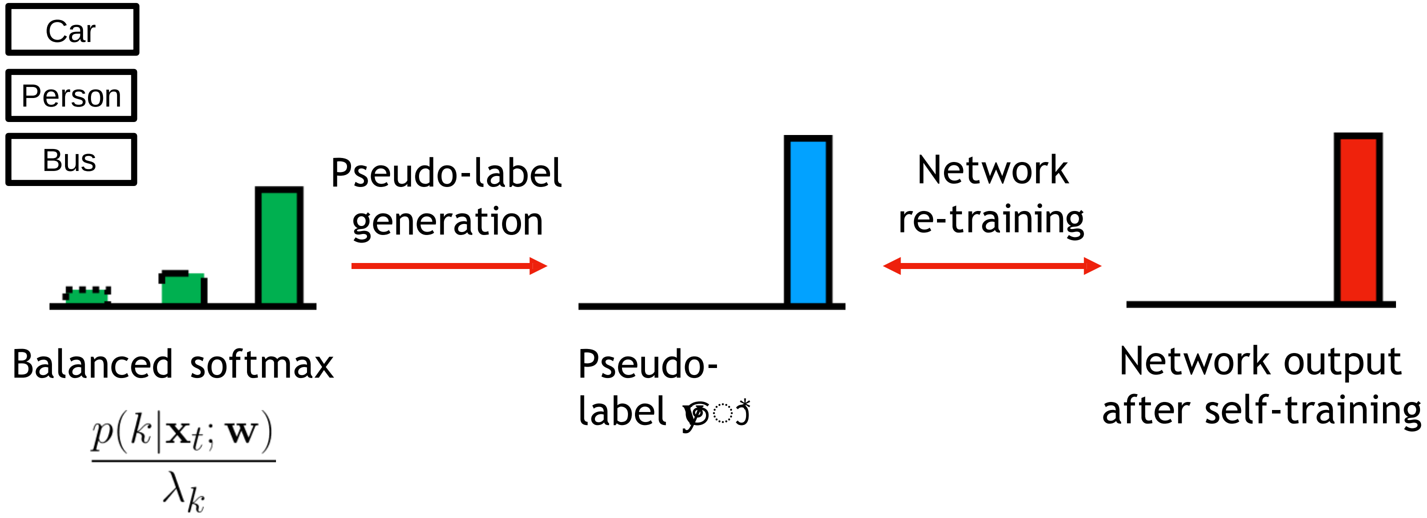
$$\min_{\mathbf{w},\hat{\mathbf{Y}}_T} \mathcal{L}_{CB}(\mathbf{w},\hat{\mathbf{Y}}_T) = -\sum_{s\in S}\sum_{k=1}^{K} y_s^{(k)} \log p(k|\mathbf{x}_s;\mathbf{w}) - \sum_{t\in T}\sum_{k=1}^{K} \hat{y}_t^{(k)} \log \frac{p(k|\mathbf{x}_t;\mathbf{w})}{\lambda_k}$$

$$s.t.\ \hat{\mathbf{y}}_t = (\hat{y}_t^{(1)}, \ldots, \hat{y}_t^{(K)}) \in \Delta^{K-1} \cup \{\mathbf{0}\},\ \forall t$$

$$\lambda_k > 0$$

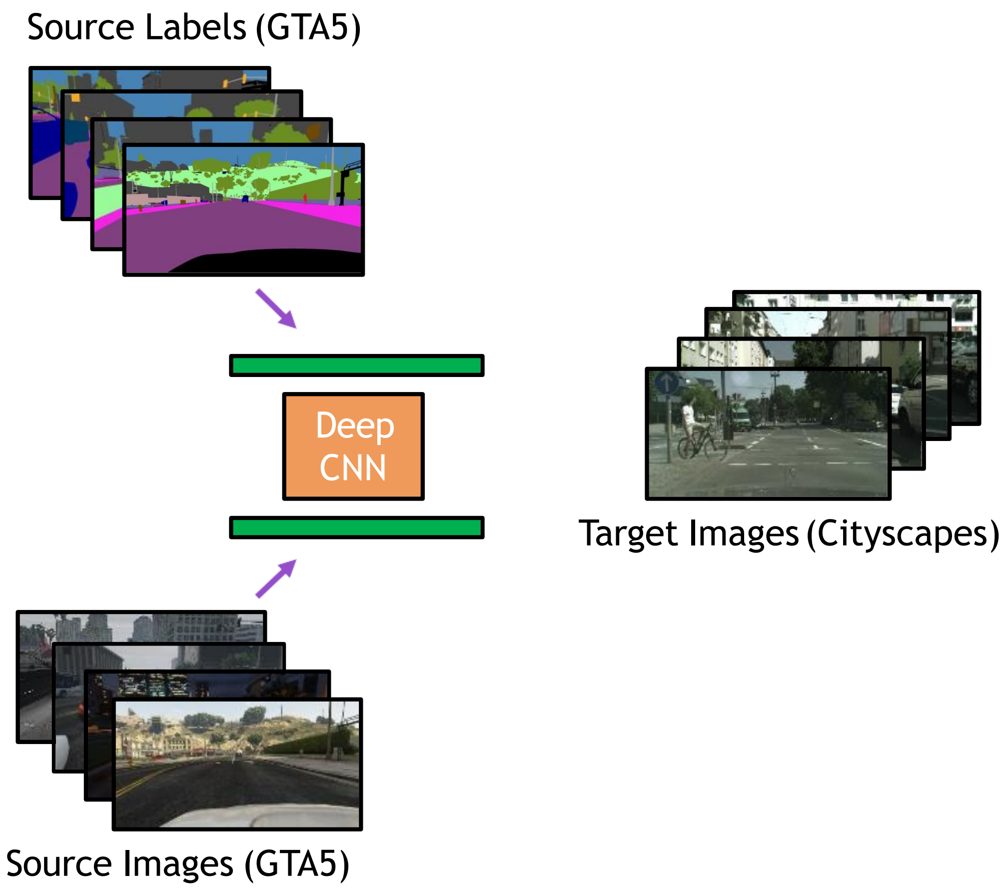
where: $\mathbf{x}$: input sample $\mathbf{p}$: class predication vector $\mathbf{y}$: label vector $\hat{\mathbf{y}}$: pseudo-label vector
$\mathbf{w}$: network parameters s: source sample index t: target sample index $\triangle^{K-1}$: probability simplex

$$\hat{y}_t^{(k)*} = \begin{cases} 1, \text{ if } k = \arg\max_k \{\frac{p(k|\mathbf{x}_t;\mathbf{w})}{\lambda_k}\} \\ \quad \text{and} \quad p(k|\mathbf{x}_t;\mathbf{w}) > \lambda_k \\ 0, \text{ otherwise} \end{cases}$$

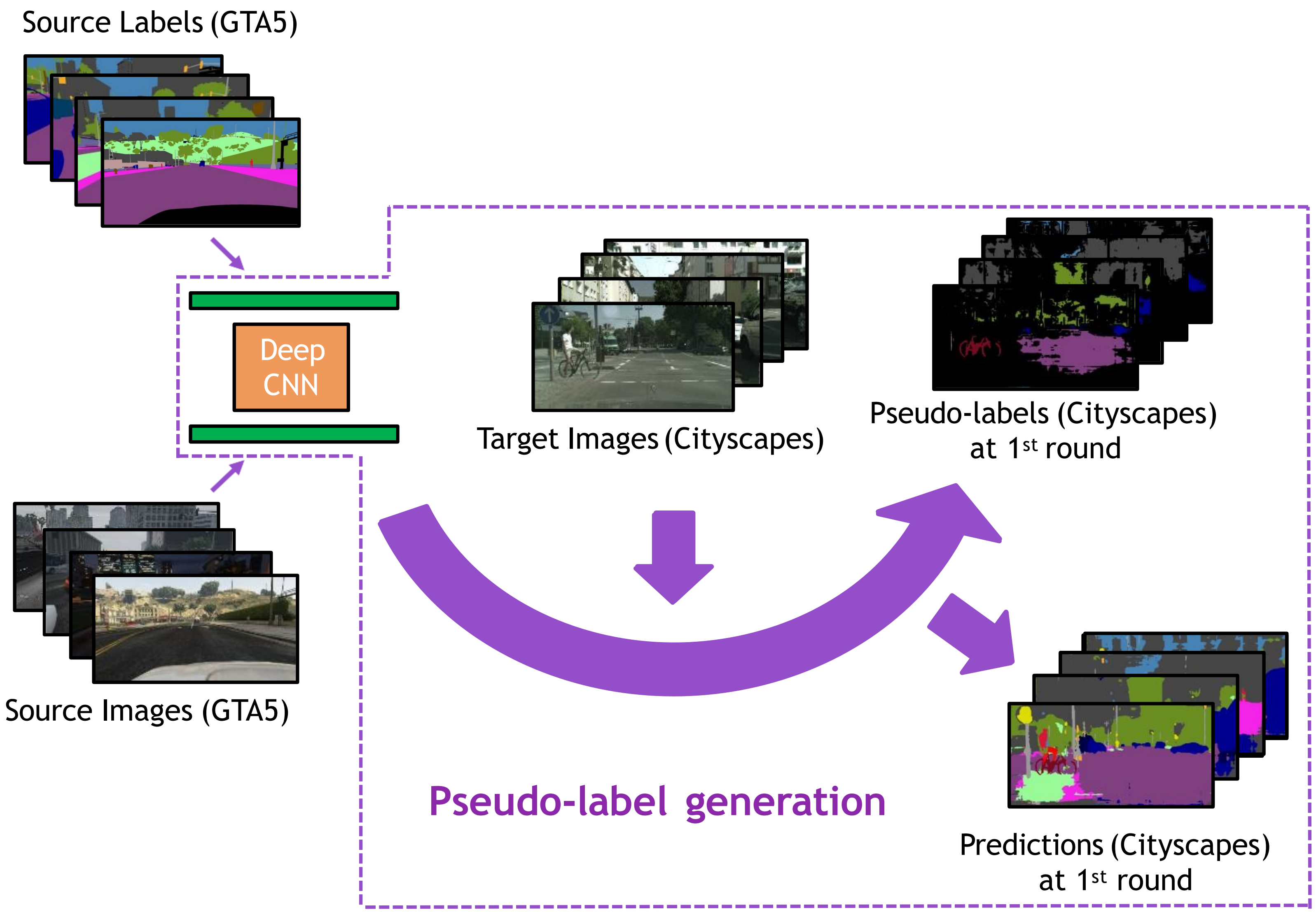


Yang et al., Domain Adaptation for Semantic Segmentation via Class-Balanced Self-Training, ECCV18

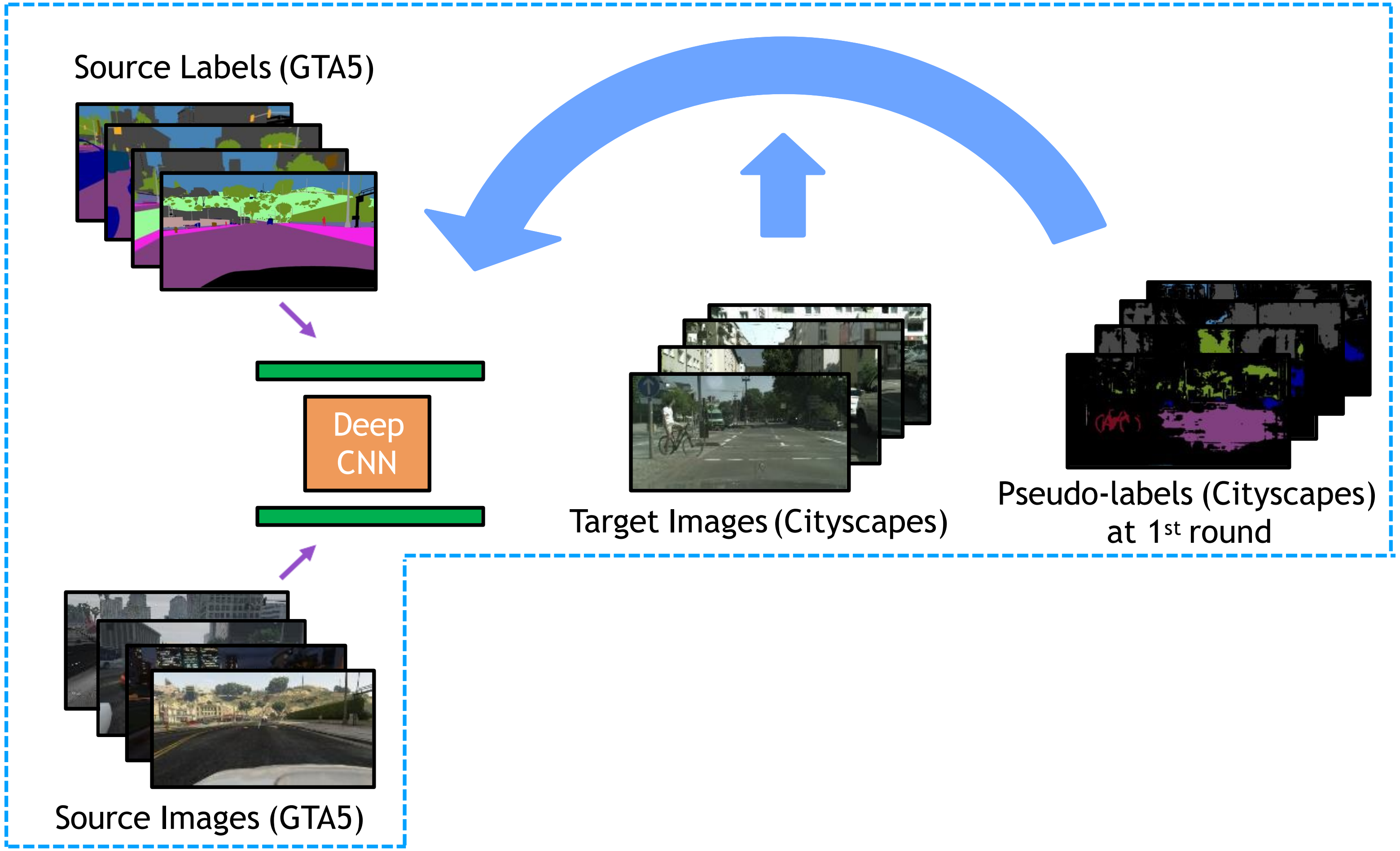
Domain Adaptation through Deep Self-Training



Domain Adaptation through Deep Self-Training

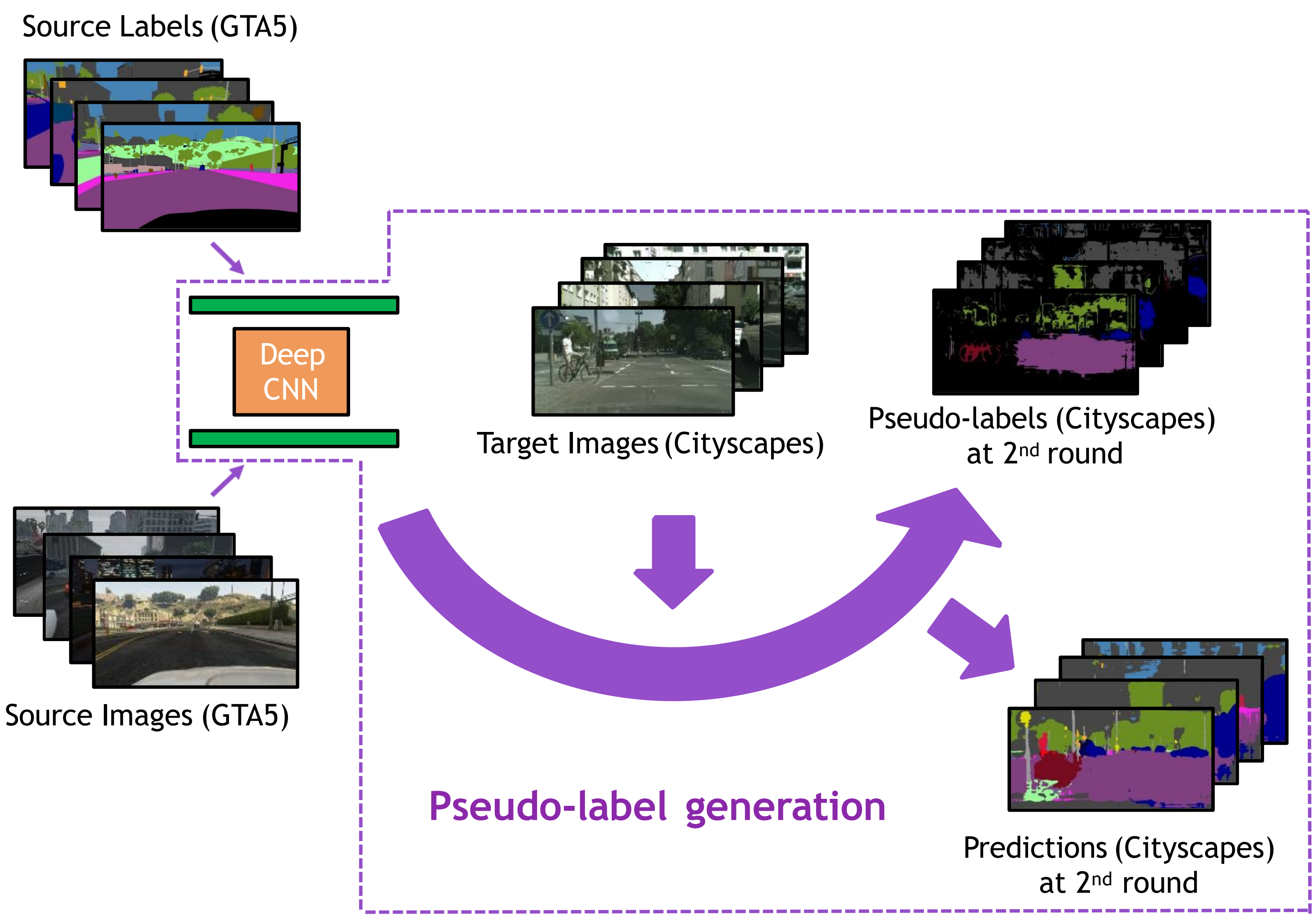


Domain Adaptation through Deep Self-Training

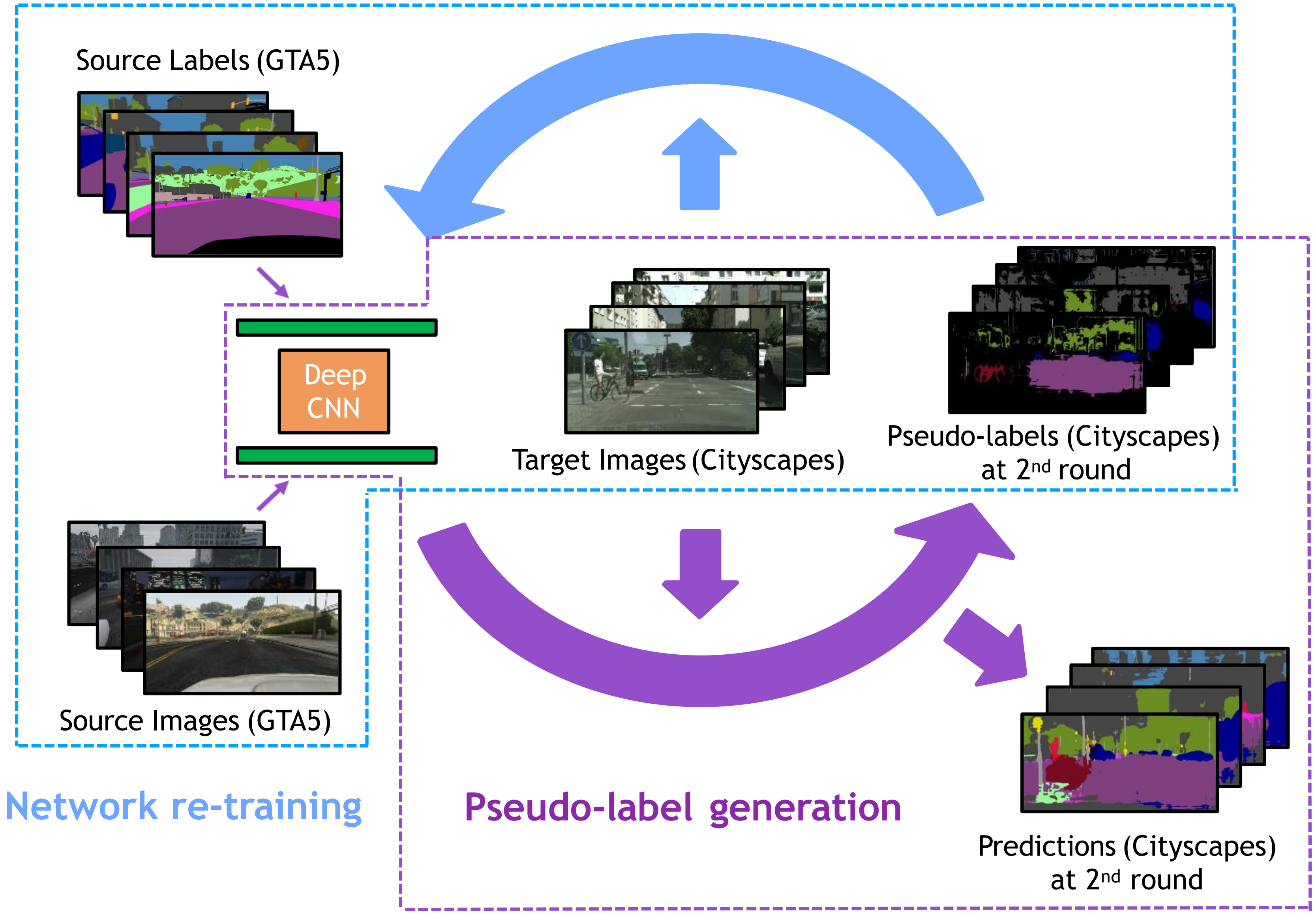


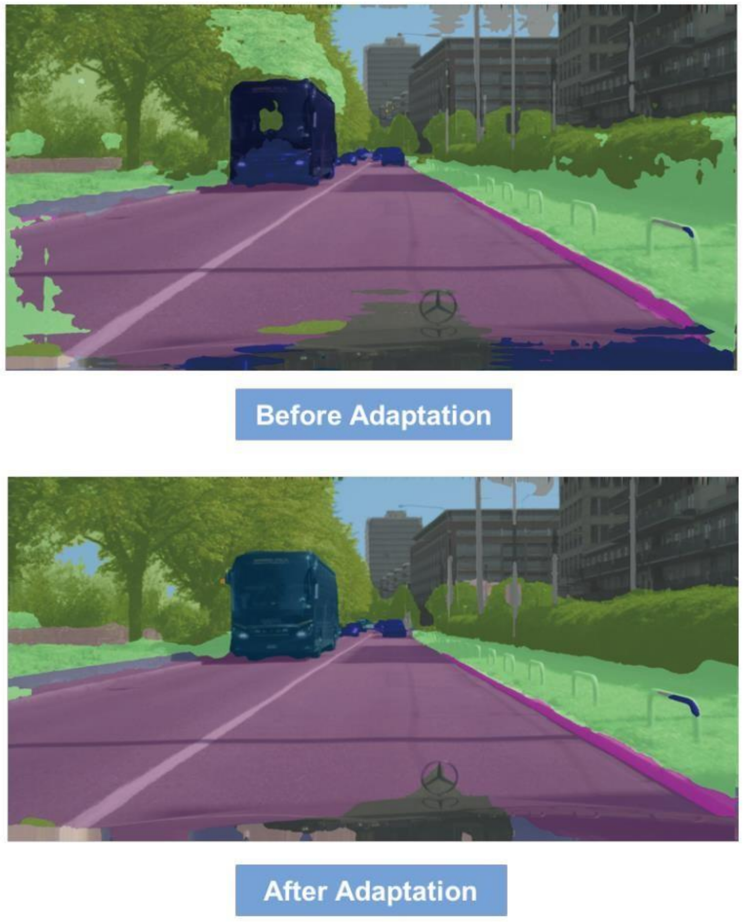
Network re-training

Domain Adaptation through Deep Self-Training

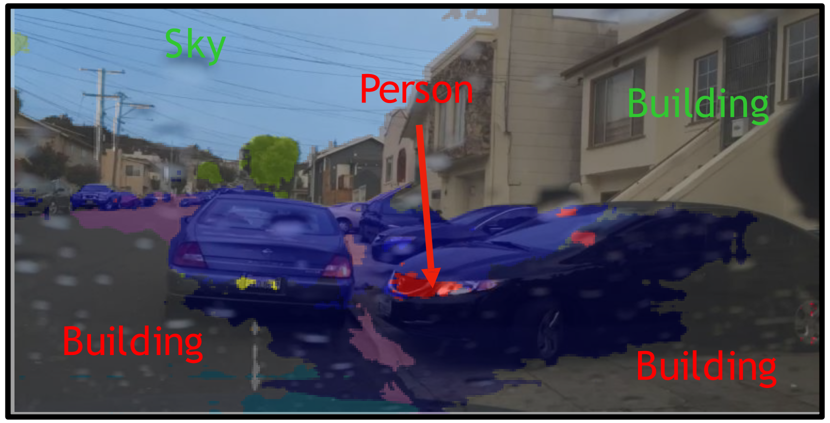


Domain Adaptation through Deep Self-Training

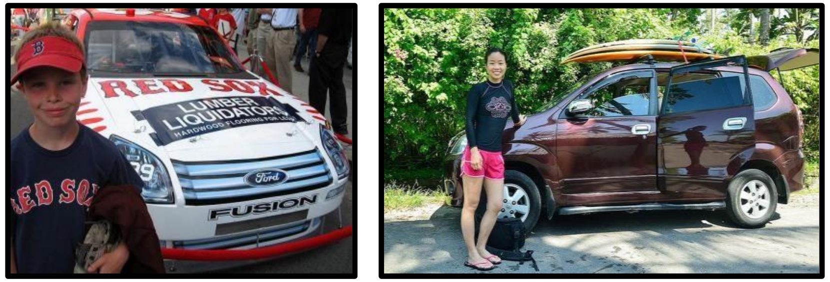




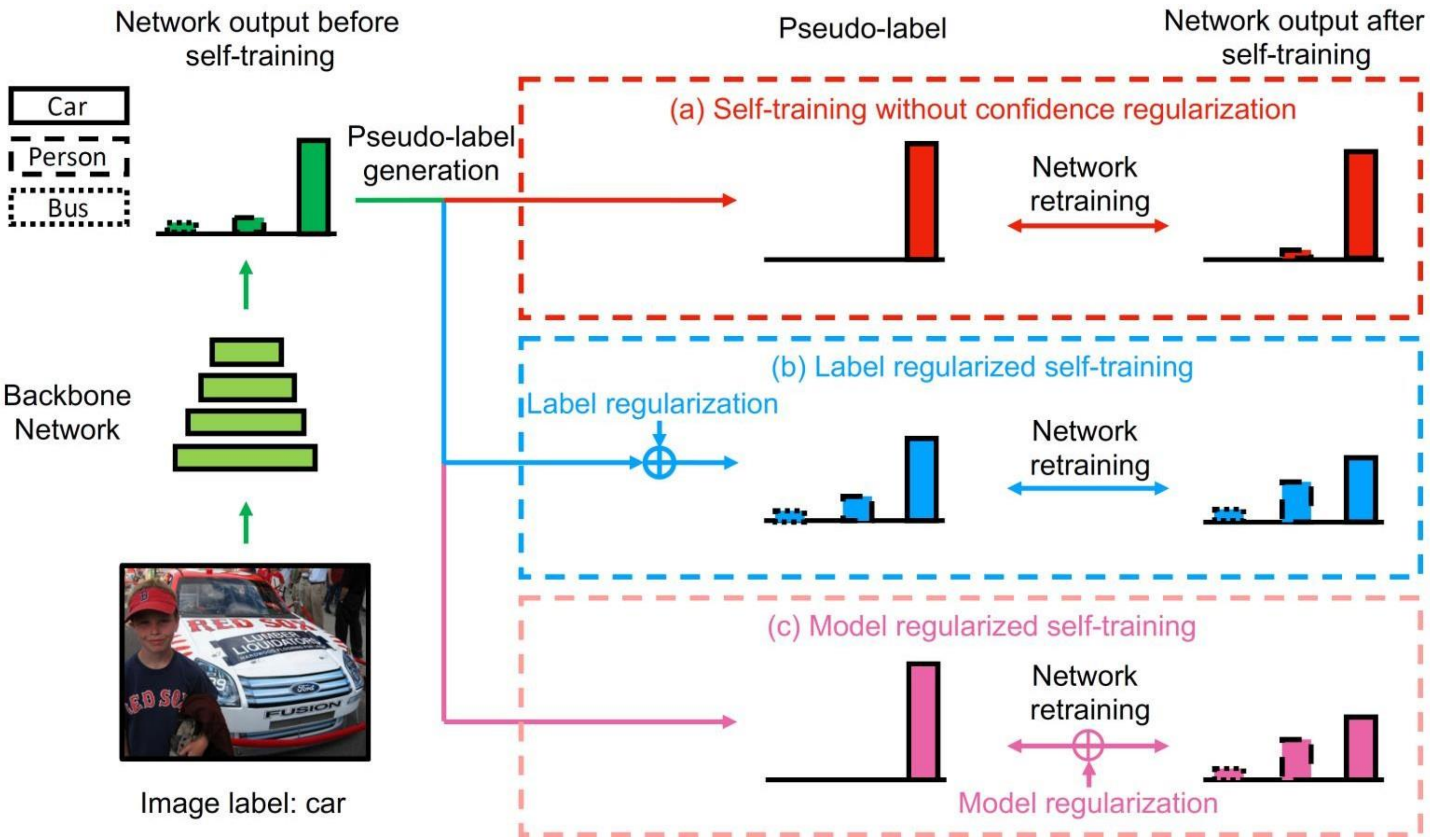
Confidence Regularized Self-Training (CRST)



Sample from BDD100K
Green: Correct Red: Misclassified

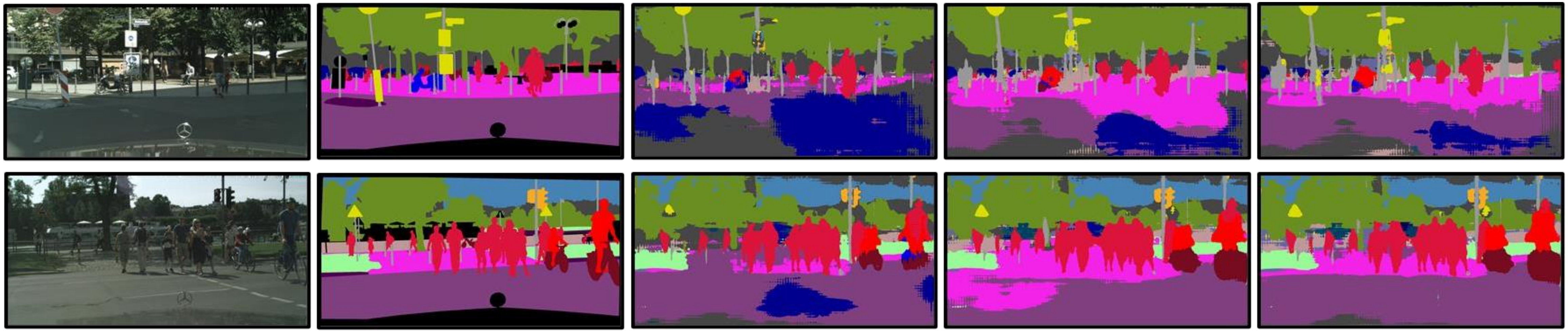


Samples from VisDA-17 (With label "Car")



Yang et al., Confidence-Regularized Self-Training, ICCV19

Qualitative Results: GTA5 -> Cityscapes



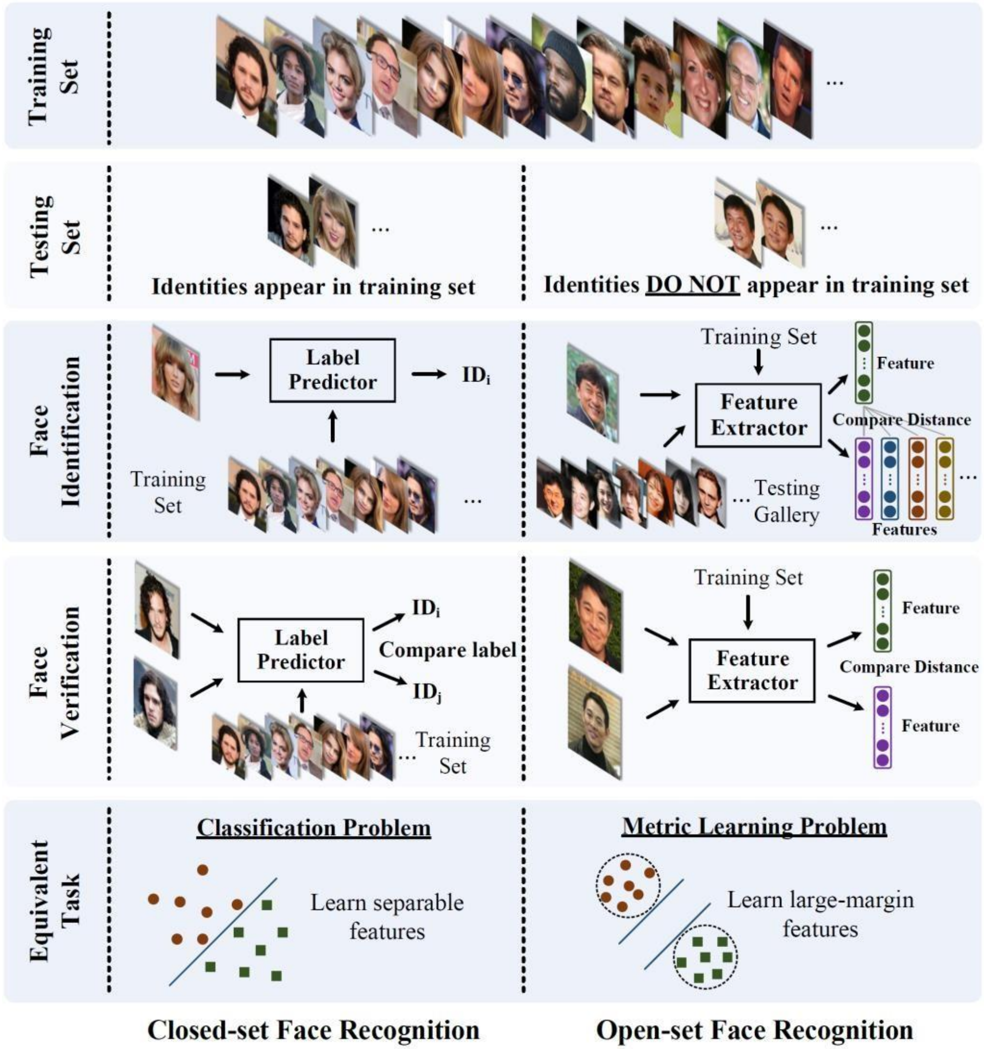
Original Image | Ground Truth | Source Model | CBST | CRST

Person Re-ID: An Open Set Recognition Task

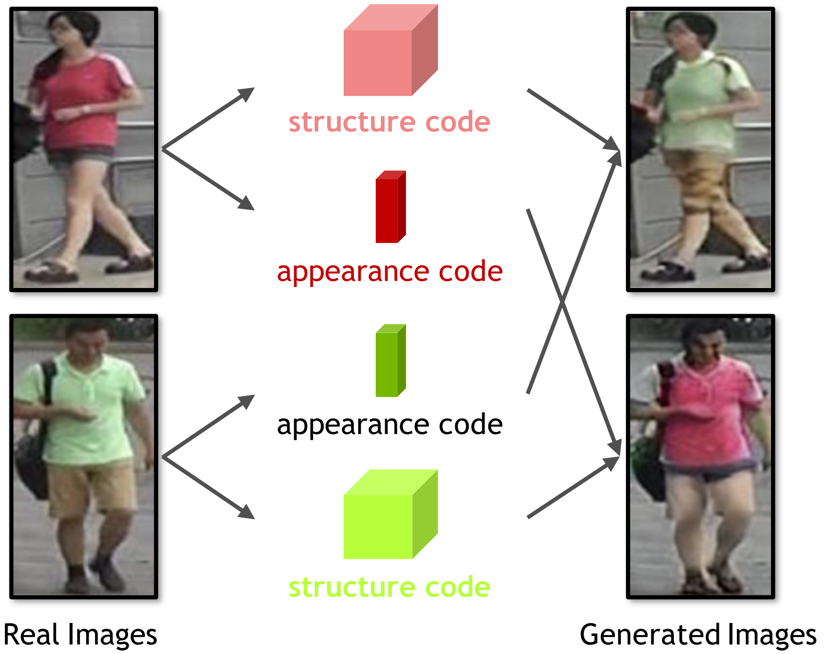
neural network
(resnet-50)

cross-entropy loss,
triplet loss, …

discriminative network features/embeddings



Cross-Domain Degeneration of DG-Net



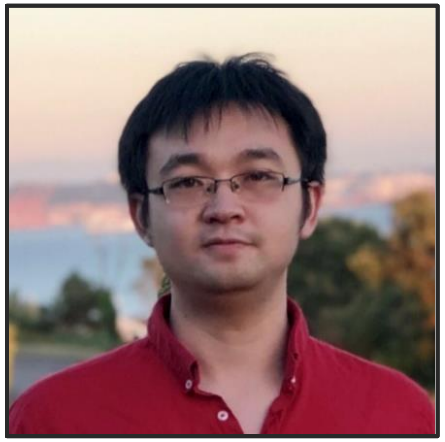
Methods	Market-1501		DukeMTMC-reID	
	Rank@1	mAP	Rank@1	mAP
Verif-Identif [55]	79.5	59.9	68.9	49.3
DCF [22]	80.3	57.5	-	-
SSM [2]	82.2	68.8	-	-
SVDNet [38]	82.3	62.1	76.7	56.8
PAN [57]	82.8	63.4	71.6	51.5
GLAD [47]	89.9	73.9	-	-
HA-CNN [24]	91.2	75.7	80.5	63.8
MLFN [4]	90.0	74.3	81.0	62.8
Part-aligned [37]	91.7	79.6	84.4	69.3
PCB [39]	93.8	81.6	83.3	69.2
Mancs [43]	93.1	82.3	84.9	71.8
DeformGAN [34]	80.6	61.3	-	-
LSRO [56]	84.0	66.1	67.7	47.1
Multi-pseudo [17]	85.8	67.5	76.8	58.6
PT [27]	87.7	68.9	78.5	56.9
PN-GAN [31]	89.4	72.6	73.6	53.2
FD-GAN [10]	90.5	77.7	80.0	64.5
Ours	**94.8**	**86.0**	**86.6**	**74.8**

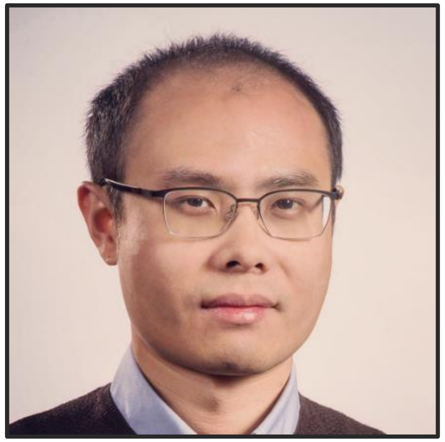
Appearance Space	Structure Space
clothing/shoes color, texture and style, other id-related cues, etc.	body size, hair, carrying, pose, background, position, viewpoint, etc.

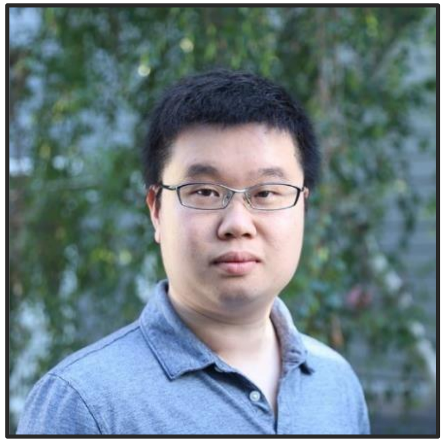
Metric	Market→Duke	Duke→Market	MSMT→Market	Market→MSMT	MSMT→Duke	Duke→MSMT
Rank@1	42.6	56.1	61.8	17.1	61.9	20.6
mAP	24.3	26.8	33.6	5.4	40.7	6.4

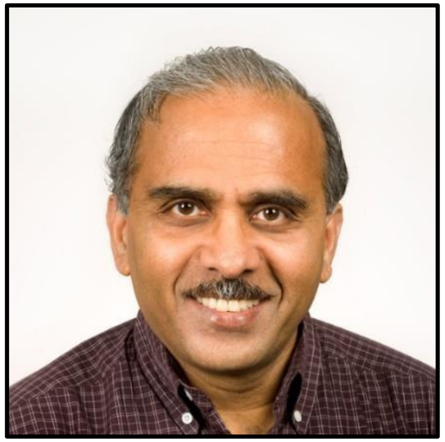
Zheng et al., Joint Discriminative and Generative Learning for Person Re-identification, CVPR19

Joint Disentangling and Adaptation for Cross-Domain Person Re-Identification (ECCV20)



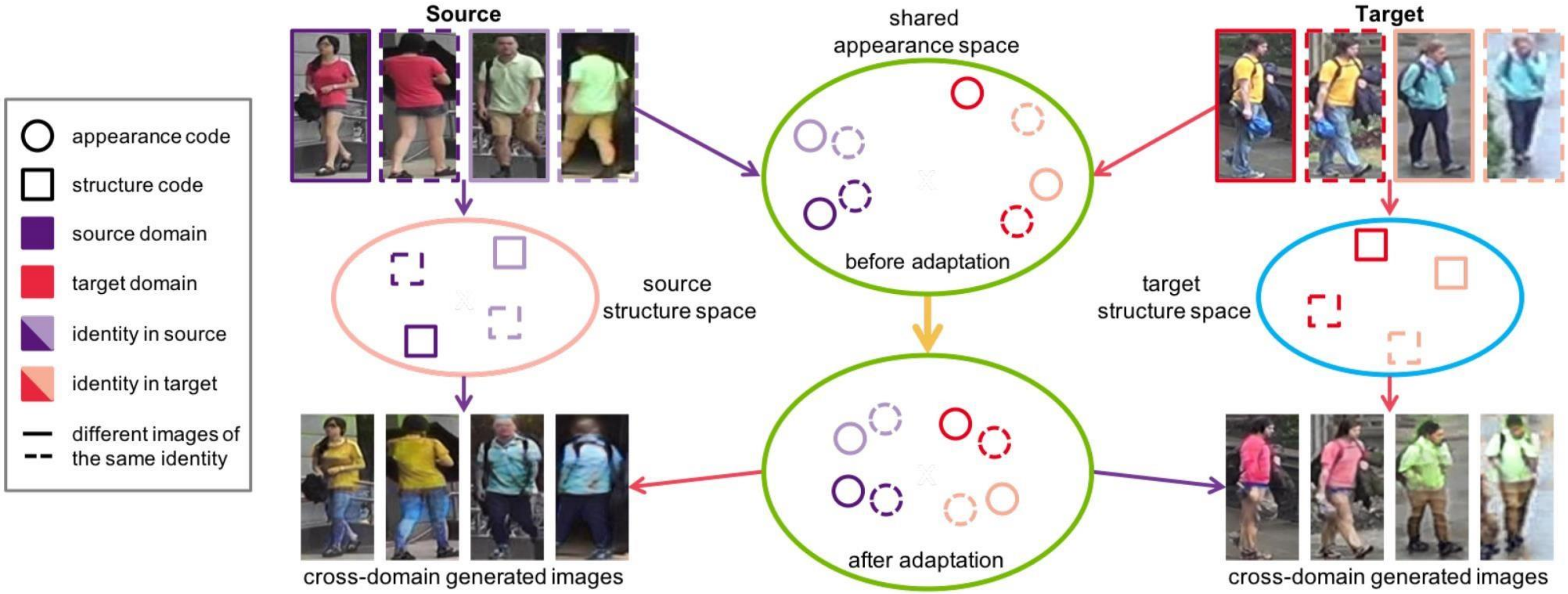






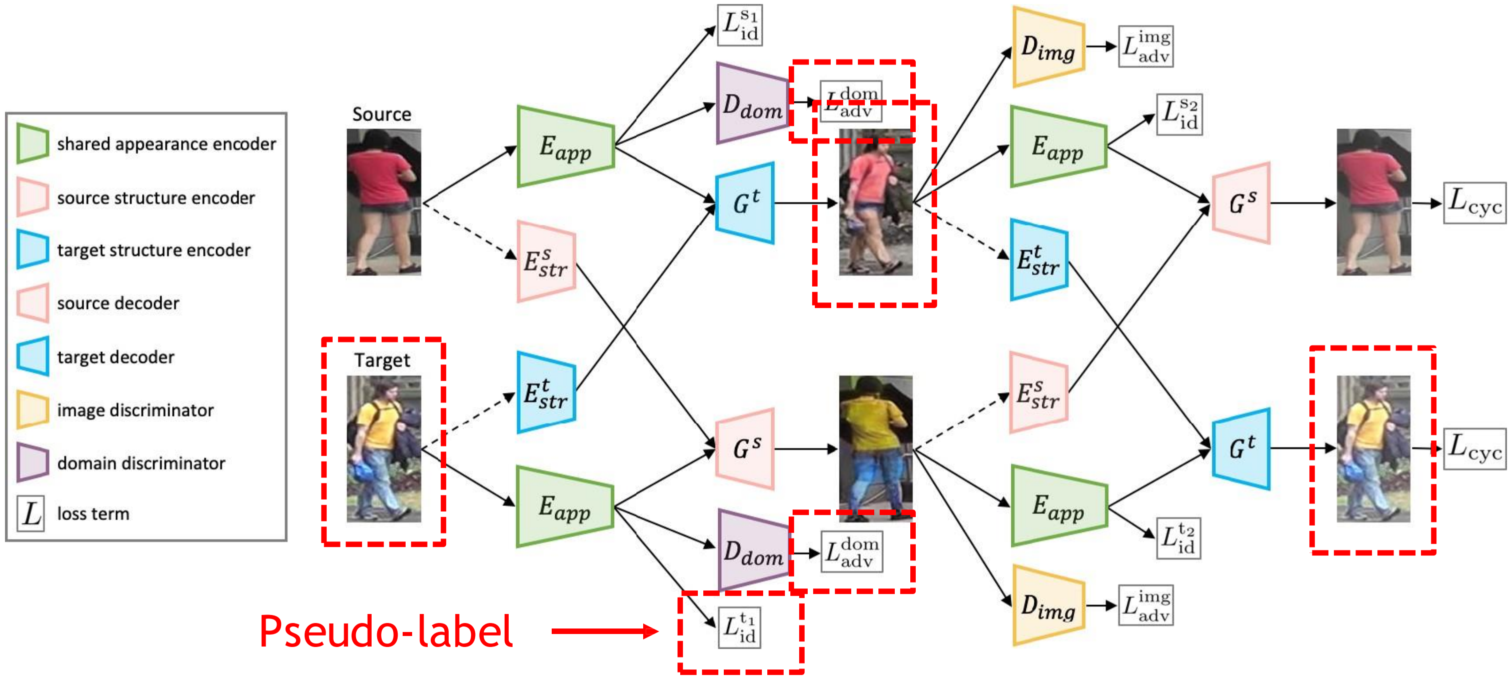
Yang Zou, CMU/NVIDIA
Xiaodong Yang, NVIDIA
Zhiding Yu, NVIDIA
Vijayakumar Bhagavatula, CMU
Jan Kautz, NVIDIA

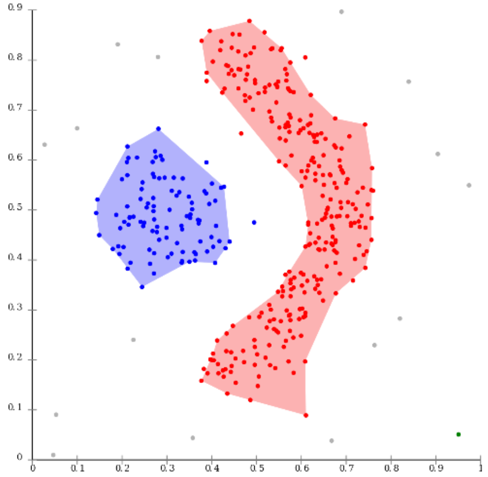
Person Re-ID as An Open Set UDA Problem

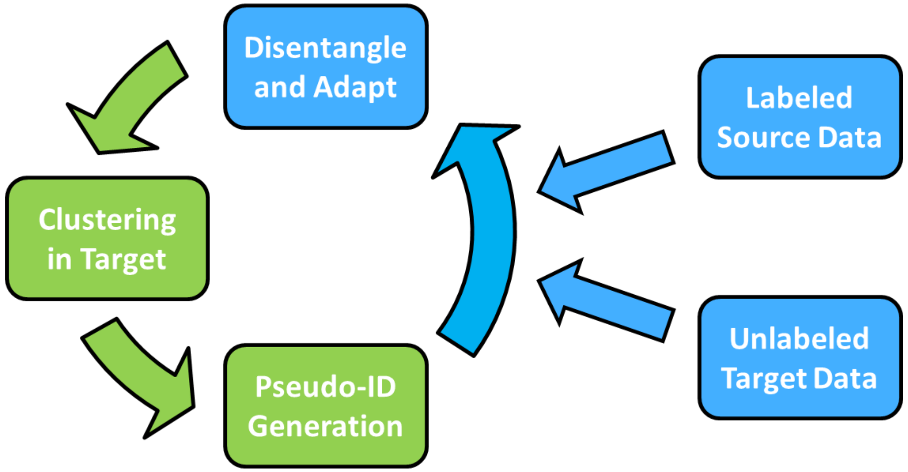


Yang et al., Joint Disentangling and Adaptation for Cross-Domain Person Re-Identification, ECCV20

DG-Net++: Joint Disentangling and Adaptation







- Cycle consistency as self-supervision to disentangle shared representation and reduce domain gap
- Adversarial loss to enforce global distribution matching
- DBSCAN as a clustering-based proxy-label approach to give pseudo-label
- Iterative disentangle & self-training

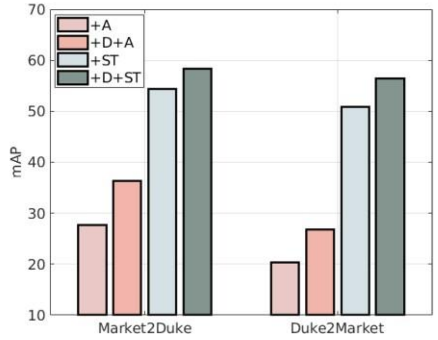
Quantitative Results and Ablation Studies

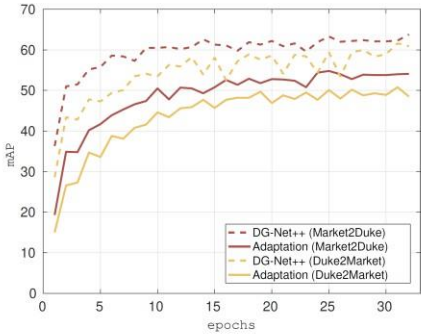
Methods	Market-1501 → DukeMTMC-reID				DukeMTMC-reID → Market-1501			
	Rank@1	Rank@5	Rank@10	mAP	Rank@1	Rank@5	Rank@10	mAP
SPGAN [6]	41.1	56.6	63.0	22.3	51.5	70.1	76.8	22.8
AIDL [51]	44.3	59.6	65.0	23.0	58.2	74.8	81.1	26.5
MMFA [31]	45.3	59.8	66.3	24.7	56.7	75.0	81.8	27.4
HHL [64]	46.9	61.0	66.7	27.2	62.2	78.8	84.0	31.4
CAL [38]	55.4	-	-	36.7	64.3	-	-	34.5
ARN [30]	60.2	73.9	79.5	33.4	70.3	80.4	86.3	39.4
ECN [65]	63.3	75.8	80.4	40.4	75.1	87.6	91.6	43.0
PDA [29]	63.2	77.0	82.5	45.1	75.2	86.3	90.2	47.6
CR-GAN [4]	68.9	80.2	84.7	48.6	77.7	89.7	92.7	54.0
IPL [42]	68.4	80.1	83.5	49.0	75.8	89.5	93.2	53.7
SSG [11]	73.0	80.6	83.2	53.4	80.0	90.0	92.4	58.3
DG-Net++	**78.9**	**87.8**	**90.4**	**63.8**	**82.1**	**90.2**	**92.7**	**61.7**

Methods	Market-1501 → MSMT17				DukeMTMC-reID → MSMT17			
	Rank@1	Rank@5	Rank@10	mAP	Rank@1	Rank@5	Rank@10	mAP
PTGAN [52]	10.2	-	24.4	2.9	11.8	-	27.4	3.3
ENC [65]	25.3	36.3	42.1	8.5	30.2	41.5	46.8	10.2
SSG [11]	31.6	-	49.6	13.2	32.2	-	51.2	13.3
DG-Net++	**48.4**	**60.9**	**66.1**	**22.1**	**48.8**	**60.9**	**65.9**	**22.1**

Methods	MSMT17 → Market-1501				MSMT17 → DukeMTMC-reID			
	Rank@1	Rank@5	Rank@10	mAP	Rank@1	Rank@5	Rank@10	mAP
PAUL [55]	68.5	-	-	40.1	72.0	-	-	53.2
DG-Net++	**83.1**	**91.5**	**94.3**	**64.6**	**75.2**	**73.6**	**86.9**	**58.2**

Methods	Market-1501 → DukeMTMC-reID				DukeMTMC-reID → Market-1501			
	Rank@1	Rank@5	Rank@10	mAP	Rank@1	Rank@5	Rank@10	mAP
Baseline	37.4	52.4	58.4	19.3	39.7	57.9	64.3	15.0
+A+ST	71.4	81.8	85.7	57.5	75.7	86.4	90.1	57.1
+D	44.5	60.6	66.7	24.2	50.1	68.0	73.9	26.8
+D+A	53.2	68.7	73.8	36.3	52.2	70.7	77.0	28.6
+D+ST	74.2	82.8	86.5	58.4	78.0	87.1	90.3	56.5
+D+A+ST	**78.9**	**87.8**	**90.4**	**63.8**	**82.1**	**90.2**	**92.7**	**61.7**

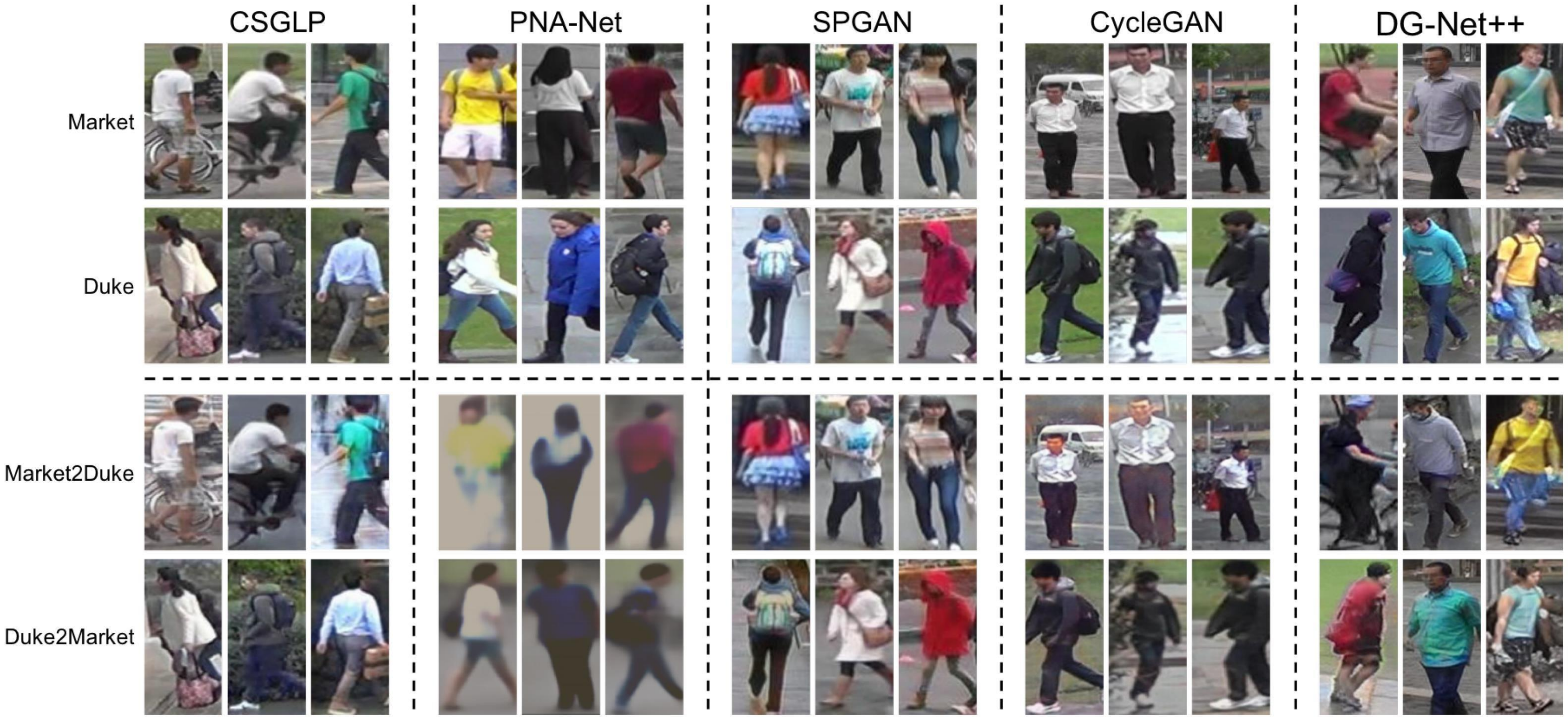




Method	Metric	Market→Duke	Duke→Market	MSMT→Market	Market→MSMT	MSMT→Duke	Duke→MSMT
DG-Net [59]	Rank@1	42.6	56.1	61.8	17.1	61.9	20.6
	mAP	24.3	26.8	33.6	5.4	40.7	6.4
DG-Net++	Rank@1	**78.9** (+36.3)	**82.1** (+26.0)	**83.1** (+21.3)	**48.4** (+31.3)	**75.2** (+13.3)	**48.8** (+28.2)
	mAP	**63.8** (+39.5)	**61.7** (+34.9)	**64.6** (+31.0)	**22.1** (+16.7)	**58.2** (+17.5)	**22.1** (+15.7)

Yang et al., Joint Disentangling and Adaptation for Cross-Domain Person Re-Identification, ECCV20

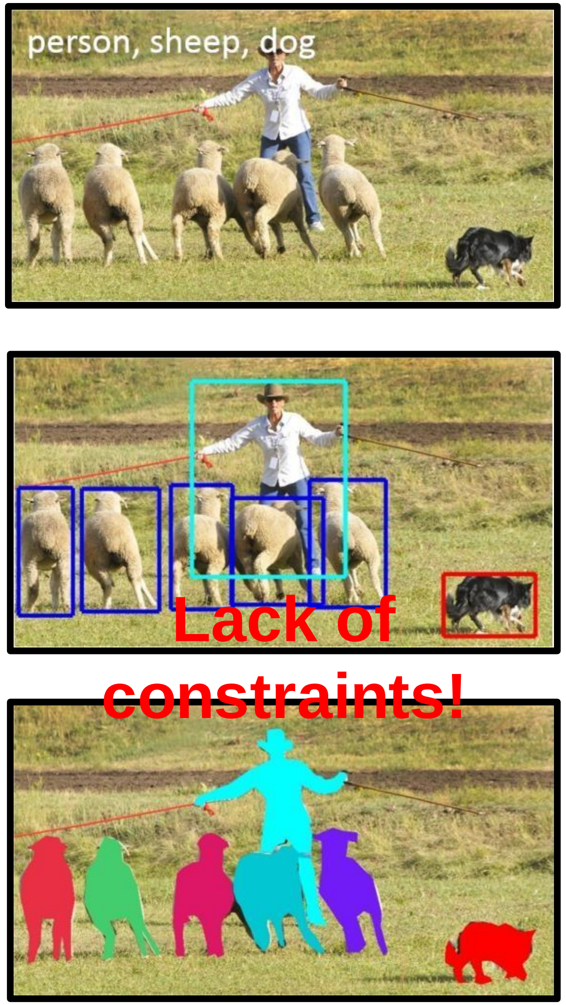
Qualitative Results

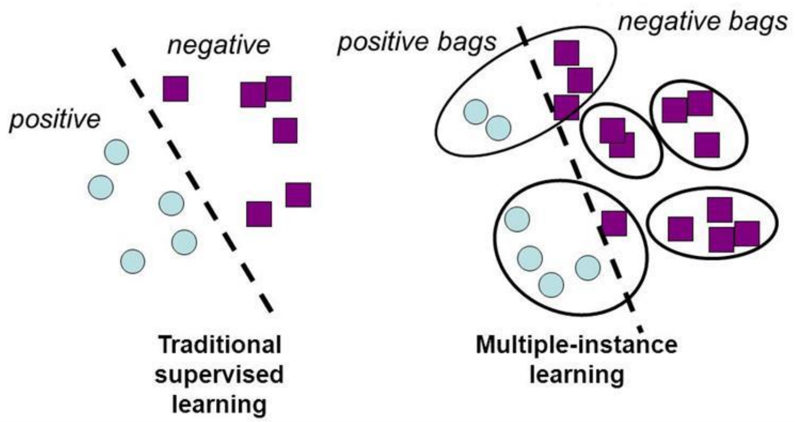


Yang et al., Joint Disentangling and Adaptation for Cross-Domain Person Re-Identification, ECCV20

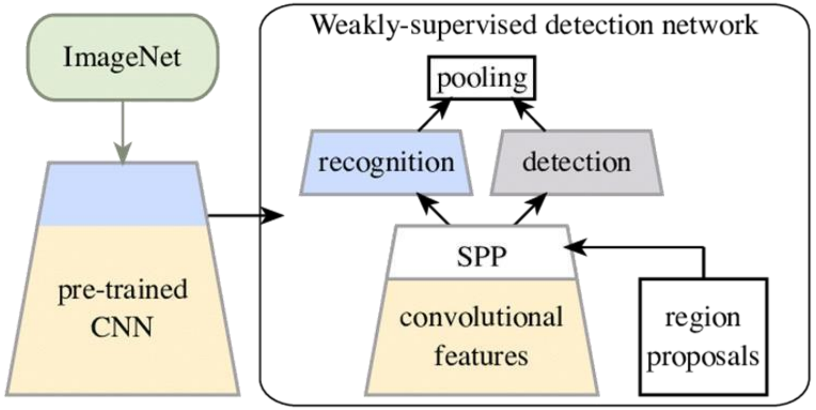
Part II: Unifying Levels of Supervision with Proxy-Label Approaches

Weakly Supervised Object Detection

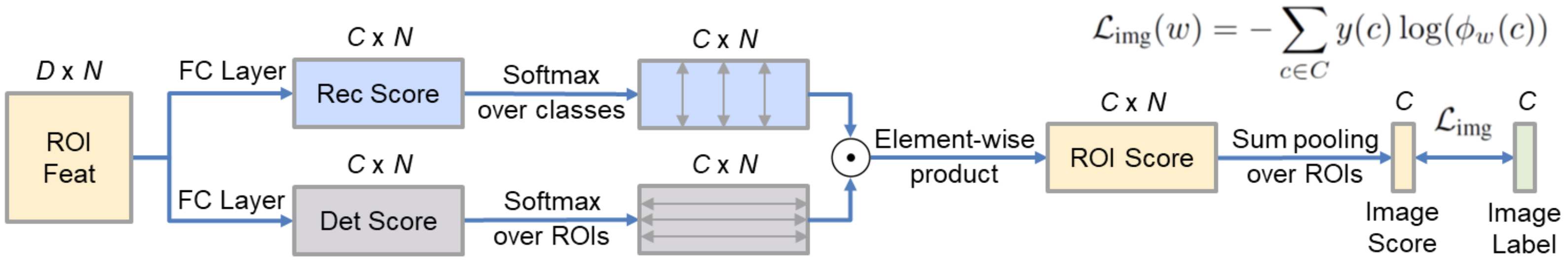




Concept of multiple-instance-learning

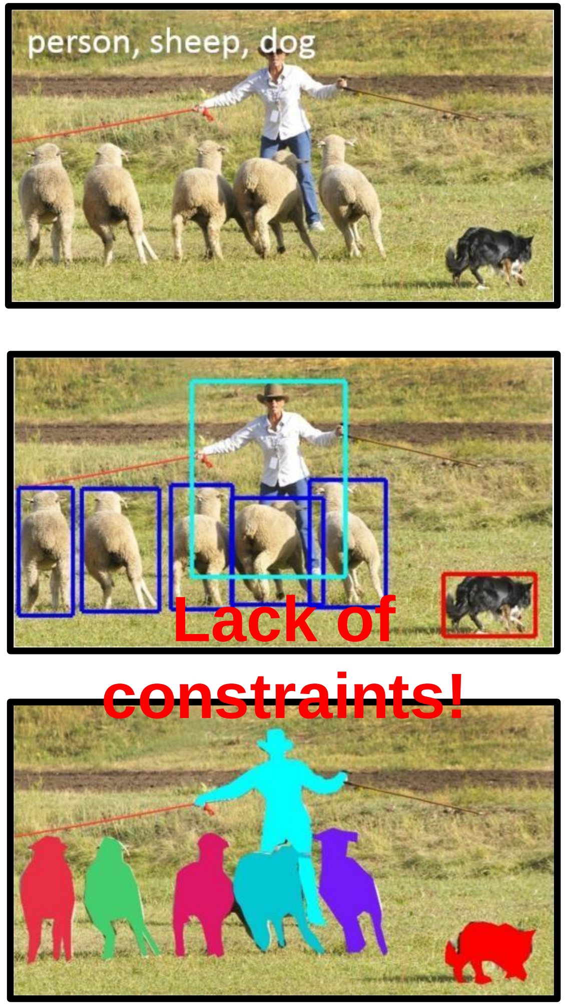


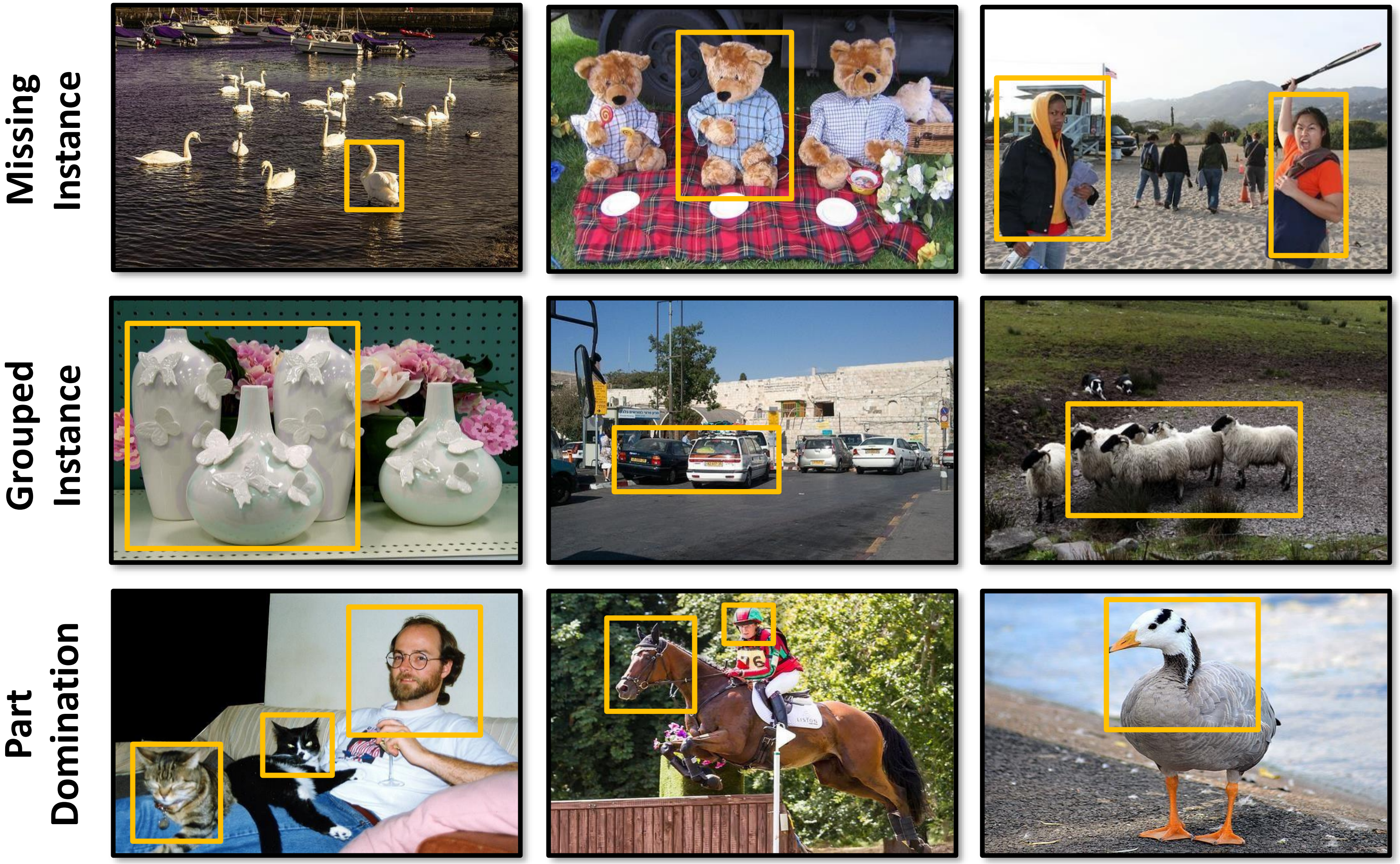
Overview of WSDDN



Bilen et al., Weakly-Supervised Deep Detection Network, CVPR16

Weakly Supervised Object Detection



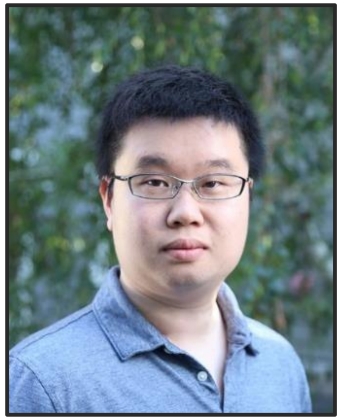


Bilen et al., Weakly-Supervised Deep Detection Network, CVPR16

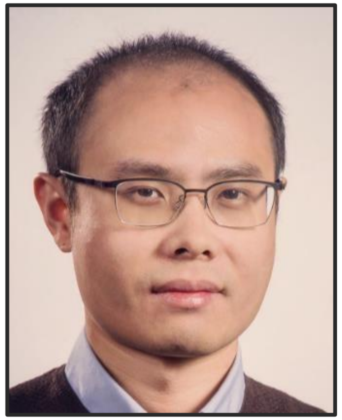
Instance-aware, Context-focused, and Memory-efficient Weakly Supervised Object Detection (CVPR20)

UFO^2: A Unified Framework towards Omni-supervised Object Detection (ECCV20)

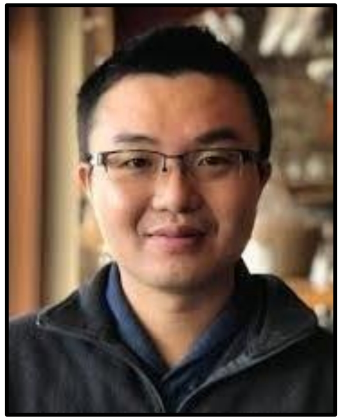
Zhongzheng Ren
UIUC/NVIDIA



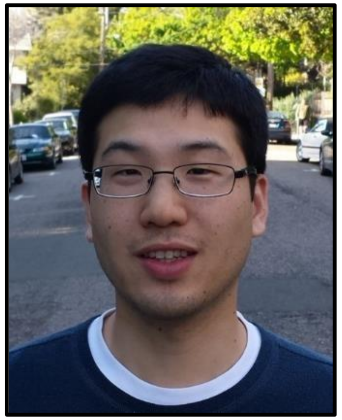
Zhiding Yu
NVIDIA



Xiaodong Yang
NVIDIA



Ming-Yu Liu
NVIDIA

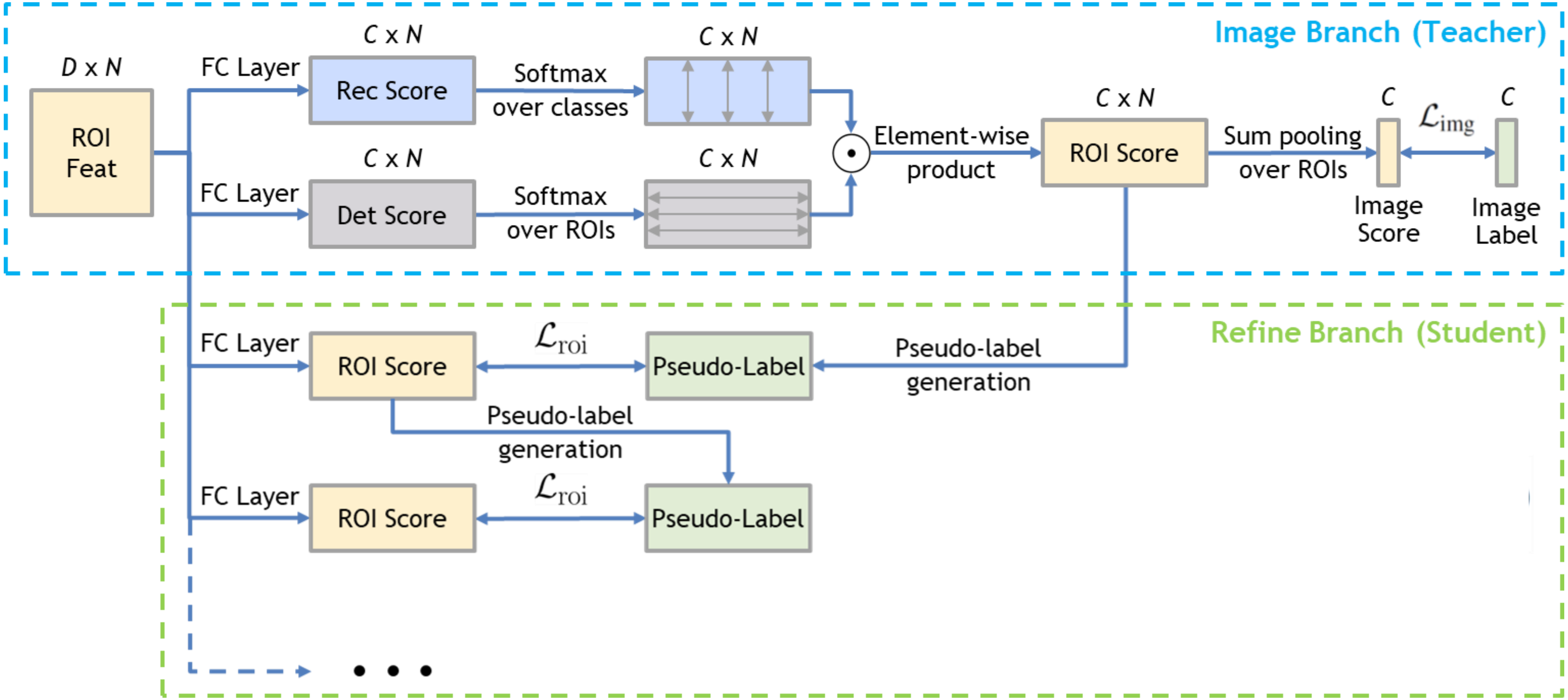


Yong Jae Lee
UC Davis

Alex Schwing
UIUC

Jan Kautz
NVIDIA

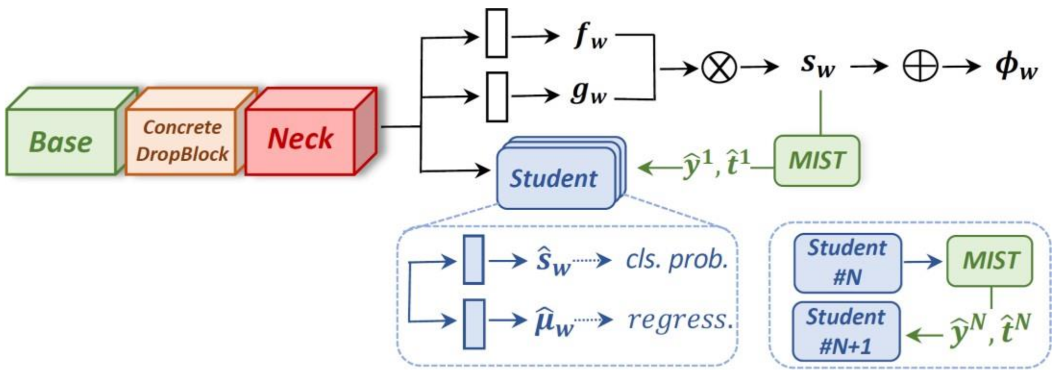
Online Instance Classifier Refinement (OICR)



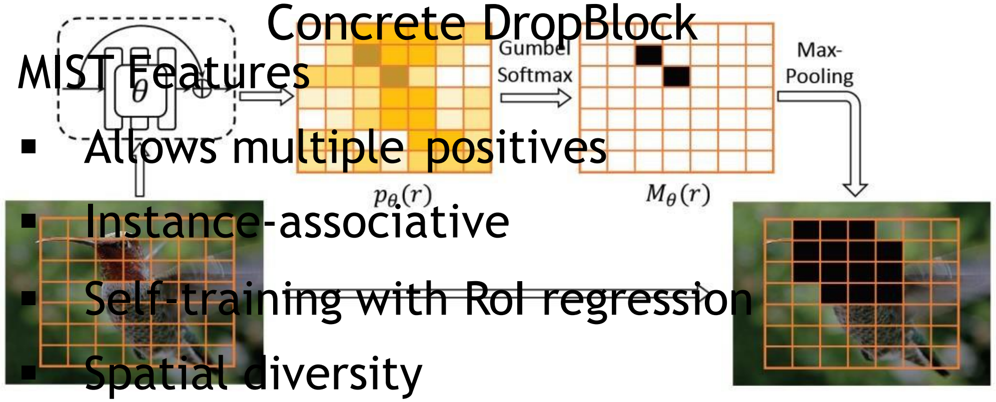
Tang et al., Multiple Instance Detection Network with Online Instance Classifier Refinement, CVPR17

Multiple Instance Self-Training (MIST)

MIST Framework



Concrete DropBlock



MIST Features

- Allows multiple positives
- Instance-associative
- Self-training with RoI regression
- Spatial diversity

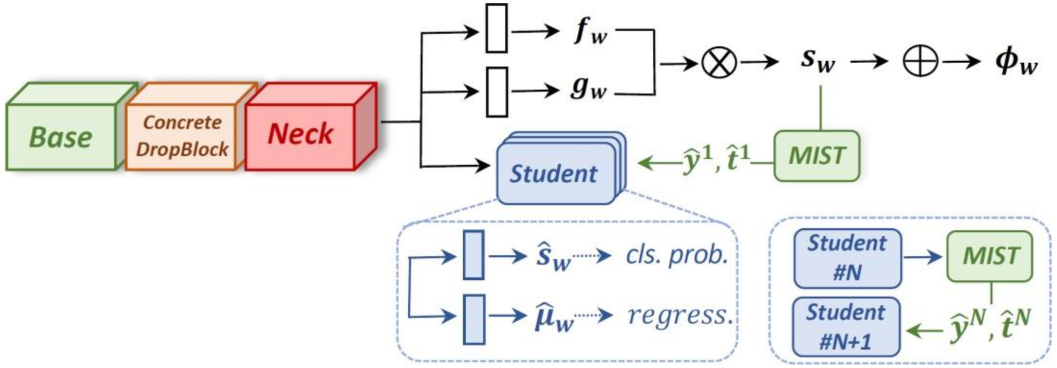
Data-Split Metrics	07 trainval CorLoc	07 test Det.	12 trainval CorLoc	12 test Det.
Baseline [45]*	60.8	42.5	-	-
+ PCL [44]	62.7	43.5	63.2	40.6
+ MIST w/o Reg.	62.9	48.3	65.1	-
+ MIST	**64.9**	**51.4**	**66.7**	-
+ Img Spa.-Dropout	64.3	51.1	65.9	-
+ ROI Spa.-Dropout	66.8	52.4	67.3	-
+ DropBlock [14]	67.1	52.9	68.4	-
+ Concrete DropBlock	**68.8**	**54.9**	**70.9**	**52.1**

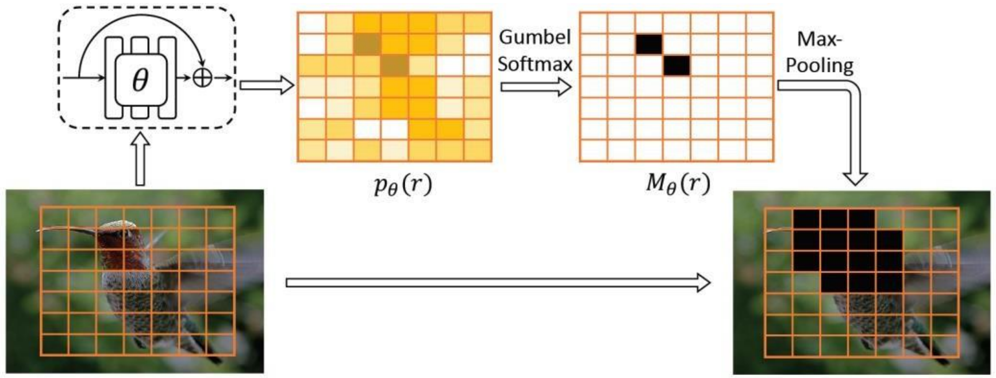
Methods	Val-AP	Val-AP$_{50}$	Test-AP	Test-AP$_{50}$
Fast R-CNN	18.9	38.6	19.3	39.3
Faster R-CNN	21.2	41.5	21.5	42.1
WSDDN [5]	-	-	-	11.5
WCCN [9]	-	-	-	12.3
PCL [45]	8.5	19.4		
C-MIDN [12]	9.6	21.4	-	-
WSOD2 [61]	10.8	22.7	-	-
Diba *et al.* [10]+SSD	-	-	-	13.6
OICR [46]+Ens+FRCNN	7.7	17.4	-	-
Ge *et al.* [13]+FRCNN	8.9	19.3	-	-
PCL [45]+Ens.+FRCNN	9.2	19.6	-	-
Ours (single-model)	**11.4**	**24.3**	**12.1**	**24.8**

Ren et al., Instance-aware, Context-focused, and Memory-efficient Weakly Supervised Object Detection, CVPR20

Multiple Instance Self-Training (MIST)

MIST Framework





Algorithm 1 Multiple Instance Self-Training

Input: Image I, class label y, proposals R, threshold τ, percentage p

Output: Pseudo boxes $\hat{R}^1$

Feed I into model; get ROI scores s

for ground-truth class c **do**

  $R(c)_{sorted} \leftarrow$ SORT$(s(c, *))$ //sort ROIs by scores of class c

  $R'(c) \leftarrow$ top p percent of $R(c)_{sorted}$

  $\hat{R}(c) \leftarrow r'_0$ // save first region (top-scoring) $r'_0 \in R'$

  for i in $\{2 \ldots |R'(c)|\}$ **do** // start from the second highest

    APPEND$(\hat{R}(c), r'_i)$ **if** IoU$(r'_i, \hat{r}_j) < \tau$, $\forall\, \hat{r}_j \in \hat{R}(c)$

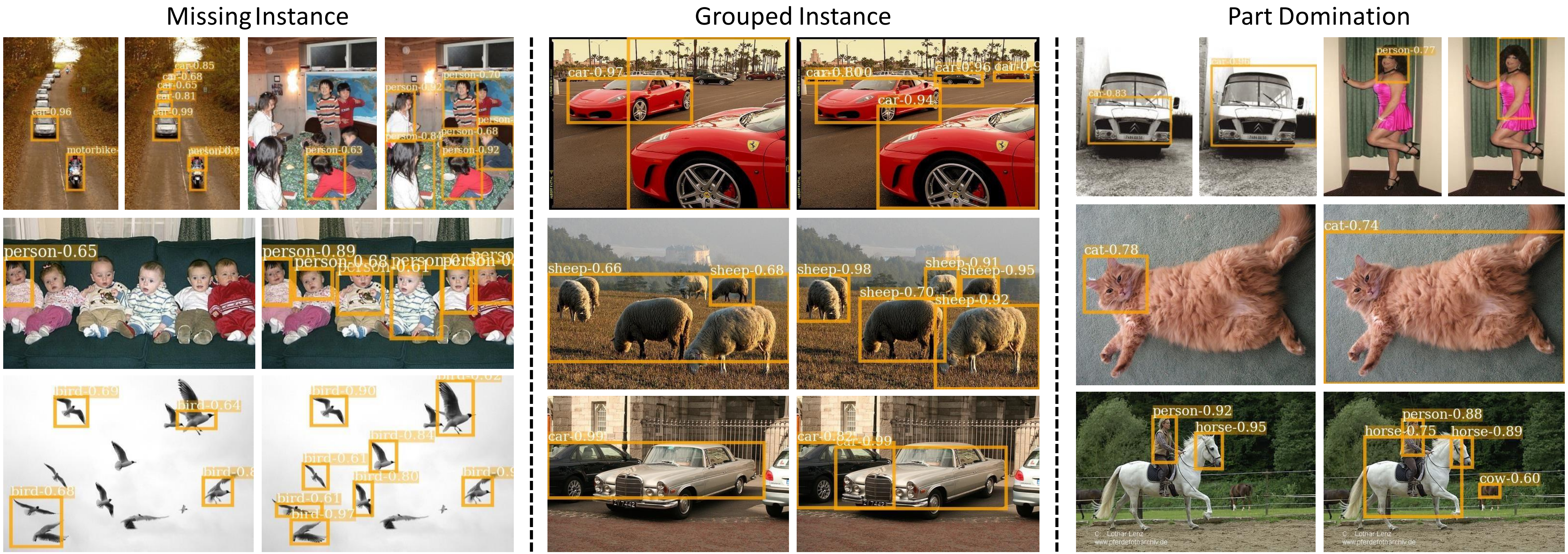
return $\hat{R}(c)$

防止伪标签的检测框在最重要的部位生成，而不是在整个物体上生成，这里采用了concrete DropBlock思路

Discriminative parts such as head are zeroed out。

Ren et al., Instance-aware, Context-focused, and Memory-efficient Weakly Supervised Object Detection, CVPR20

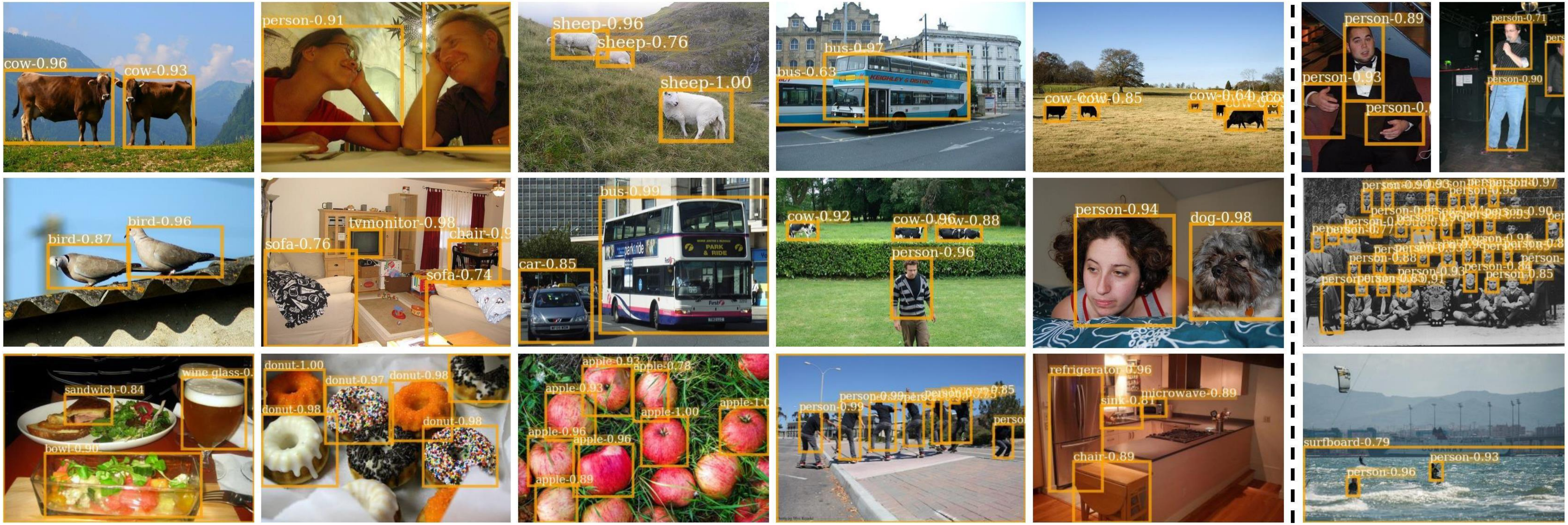
Qualitative Results



Comparison between our final WSOD model (right in pairs) and OICR (left in pairs)

Ren et al., Instance-aware, Context-focused, and Memory-efficient Weakly Supervised Object Detection, CVPR20

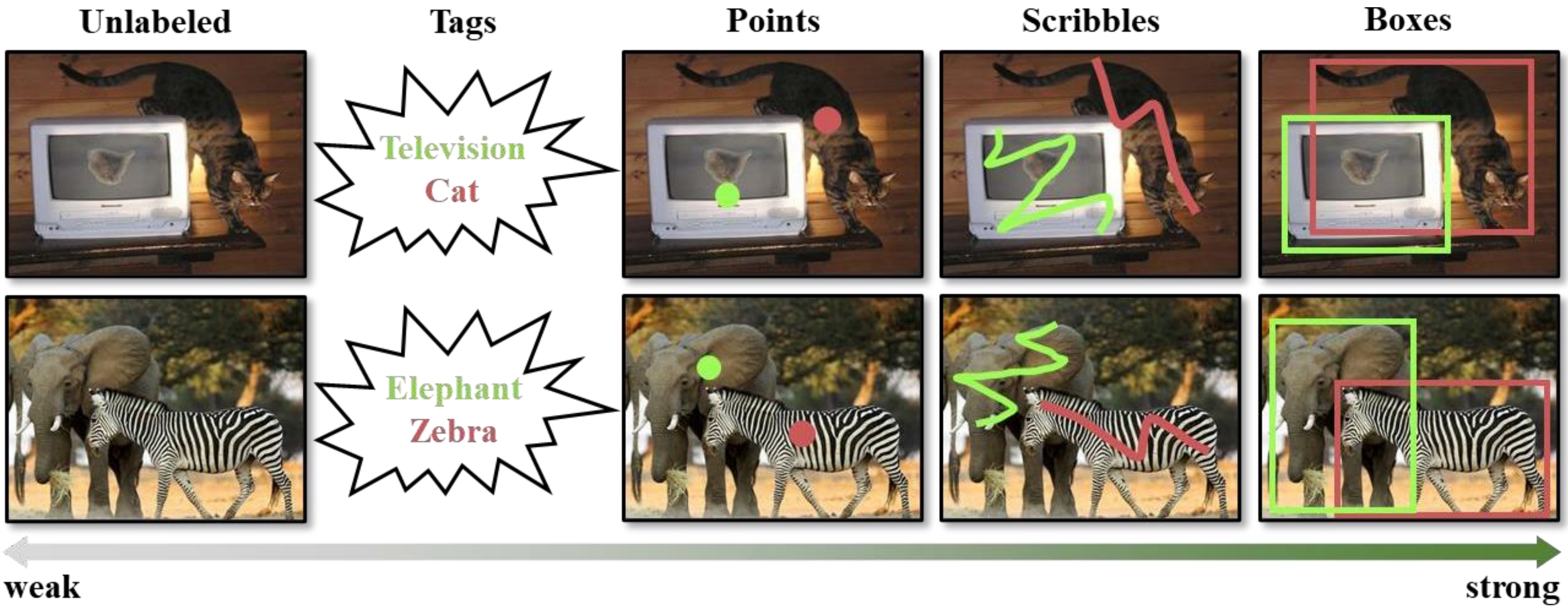
Qualitative Results



Additional visualizations on different datasets (top: VOC 2007, middle: VOC 2012, bottom: COCO) and examples of failure cases

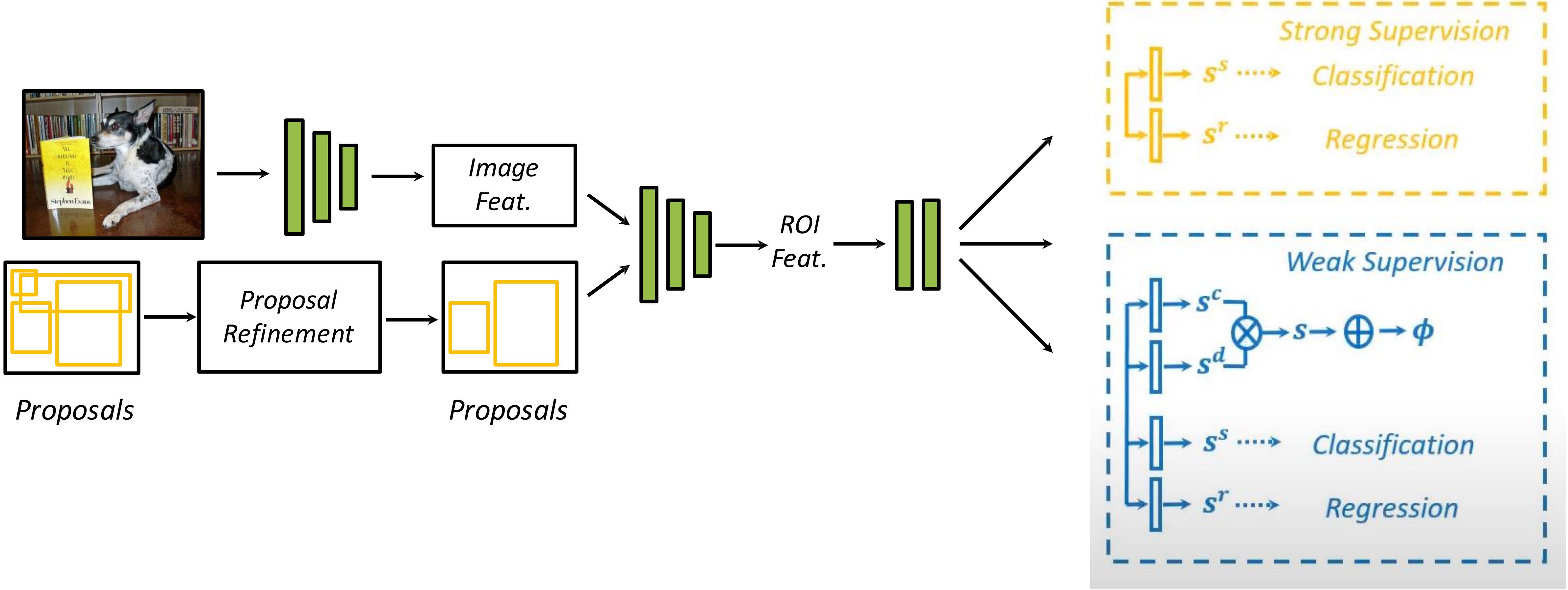
Ren et al., Instance-aware, Context-focused, and Memory-efficient Weakly Supervised Object Detection, CVPR20

Omni-Supervised Object Detection



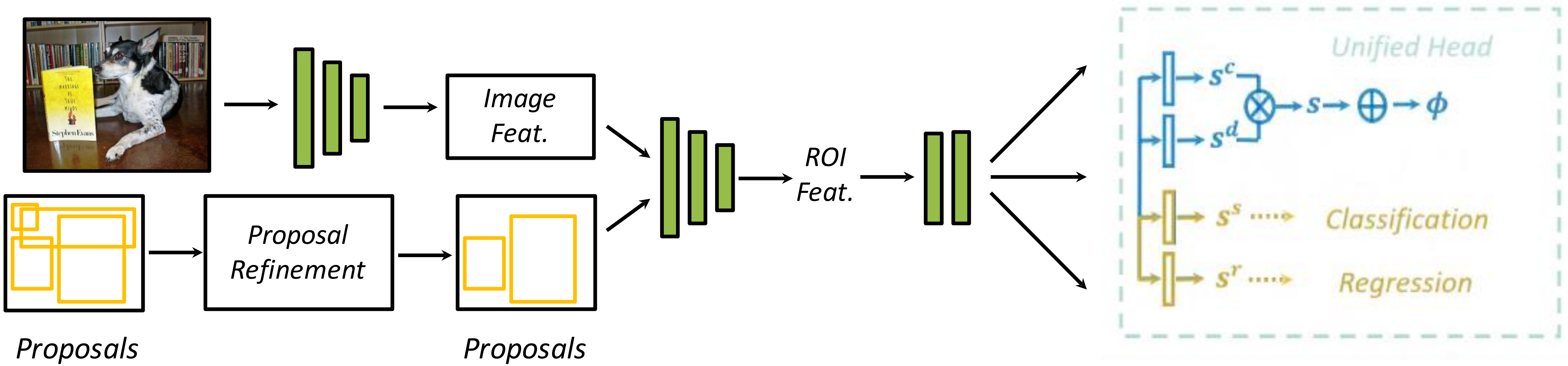
Ren et al., UFO^2: A Unified Framework towards Omni-supervised Object Detection, ECCV20

A Unified Omni-Supervised Framework



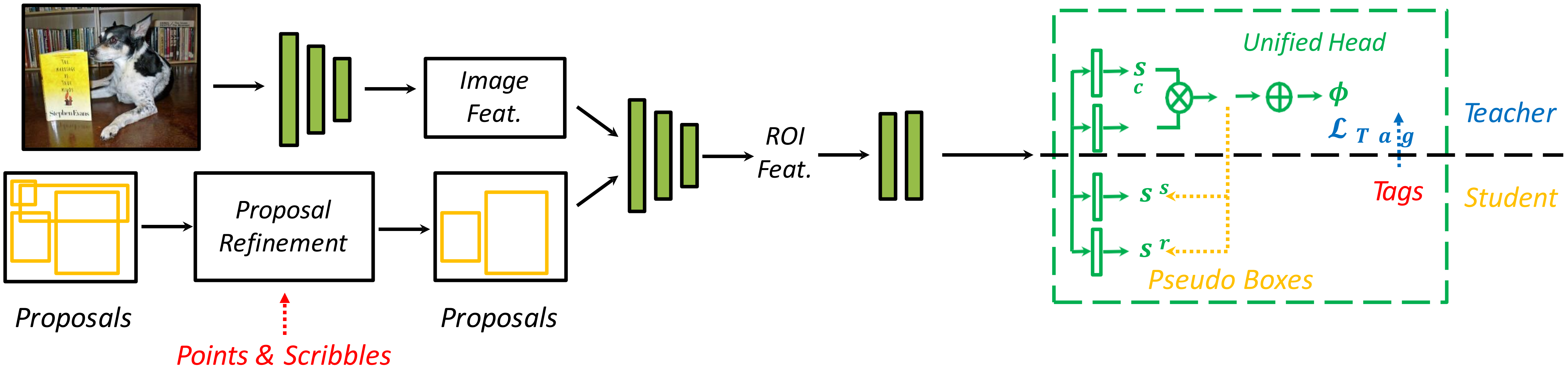
Ren et al., UFO^2: A Unified Framework towards Omni-supervised Object Detection, ECCV20

A Unified Omni-Supervised Framework



Ren et al., UFO^2: A Unified Framework towards Omni-supervised Object Detection, ECCV20

Partial Supervision (Tags, Points, Scribbles)



Teacher heads:

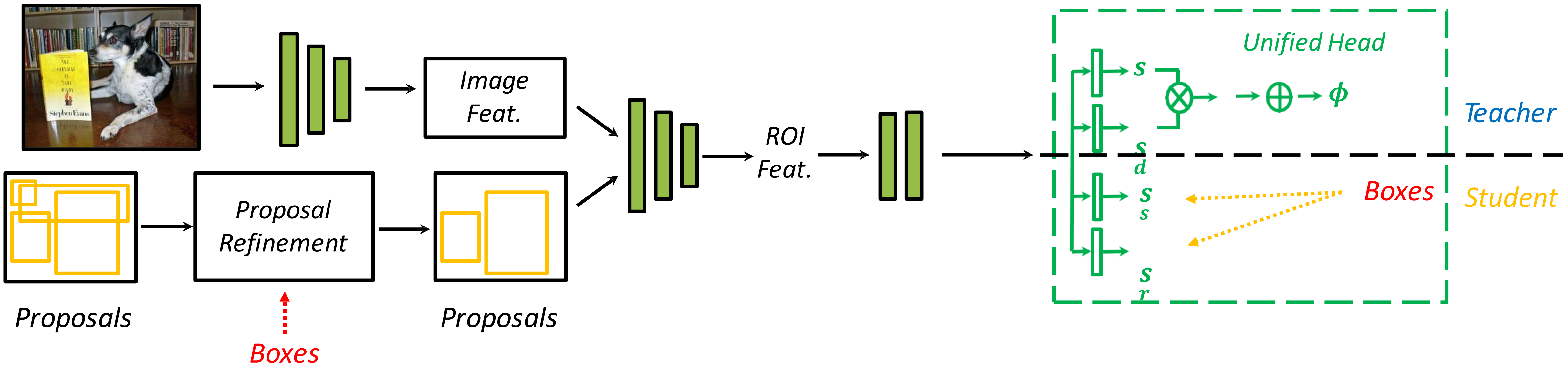
- No GT boxes available
- Image-level multi-label classification ($\mathcal{L}_{Tags}$)

Student heads:

- Generate pseudo-boxes online
- RoI classification
- RoI regression

Ren et al., UFO^2: A Unified Framework towards Omni-supervised Object Detection, ECCV20

Strong Supervision (Boxes)

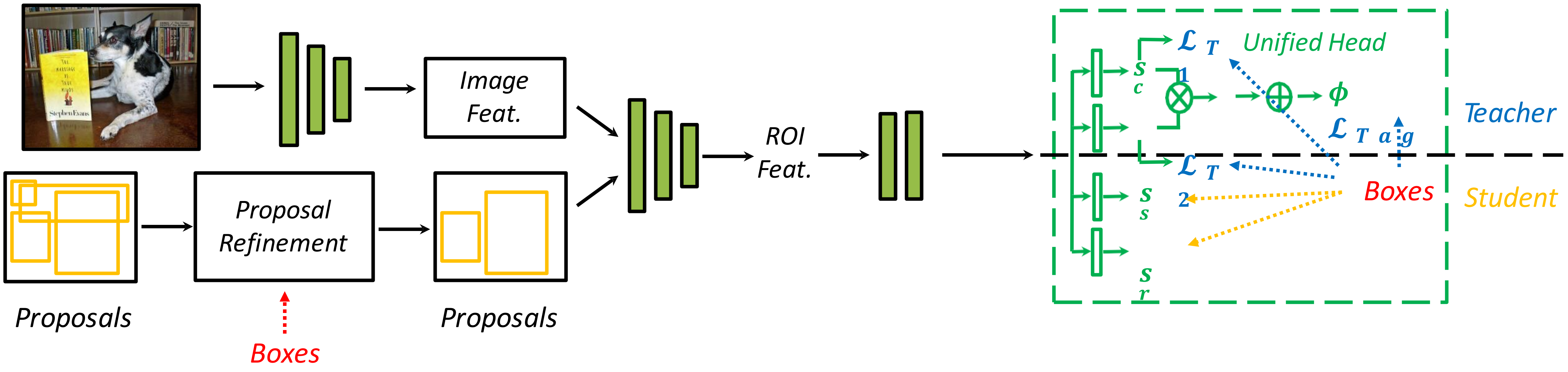


Naïve solution: directly supervise student heads using GT boxes

Issue: Weak Teacher & Strong Students

Ren et al., UFO^2: A Unified Framework towards Omni-supervised Object Detection, ECCV20

Strong Supervision (Boxes)

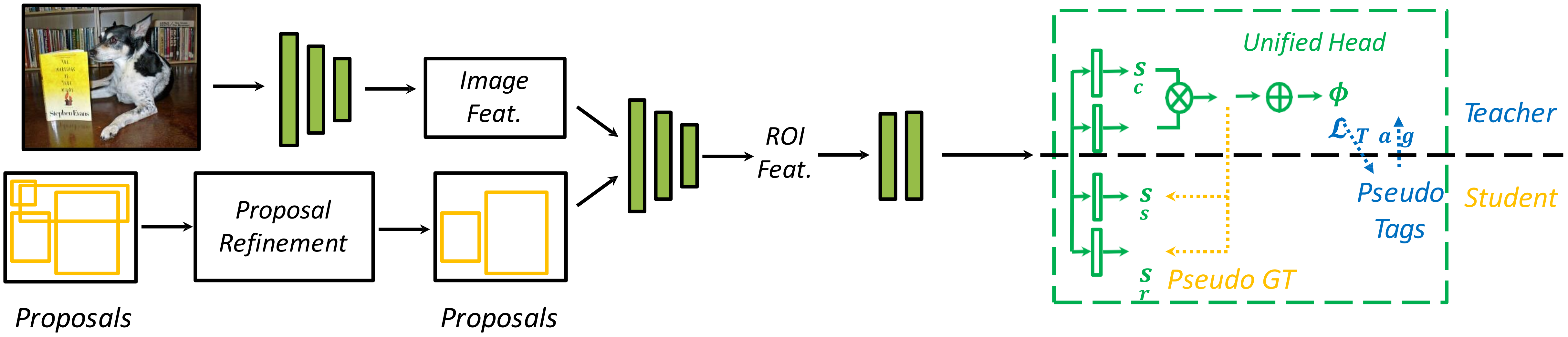


Make Teacher **stronger**:

- Image-level multi-label classification ($\mathcal{L}_{Tags}$)
- RoI classification ($\mathcal{L}_{T1}$)
- RoI objectness regularization ($\mathcal{L}_{T2}$)

Ren et al., UFO^2: A Unified Framework towards Omni-supervised Object Detection, ECCV20

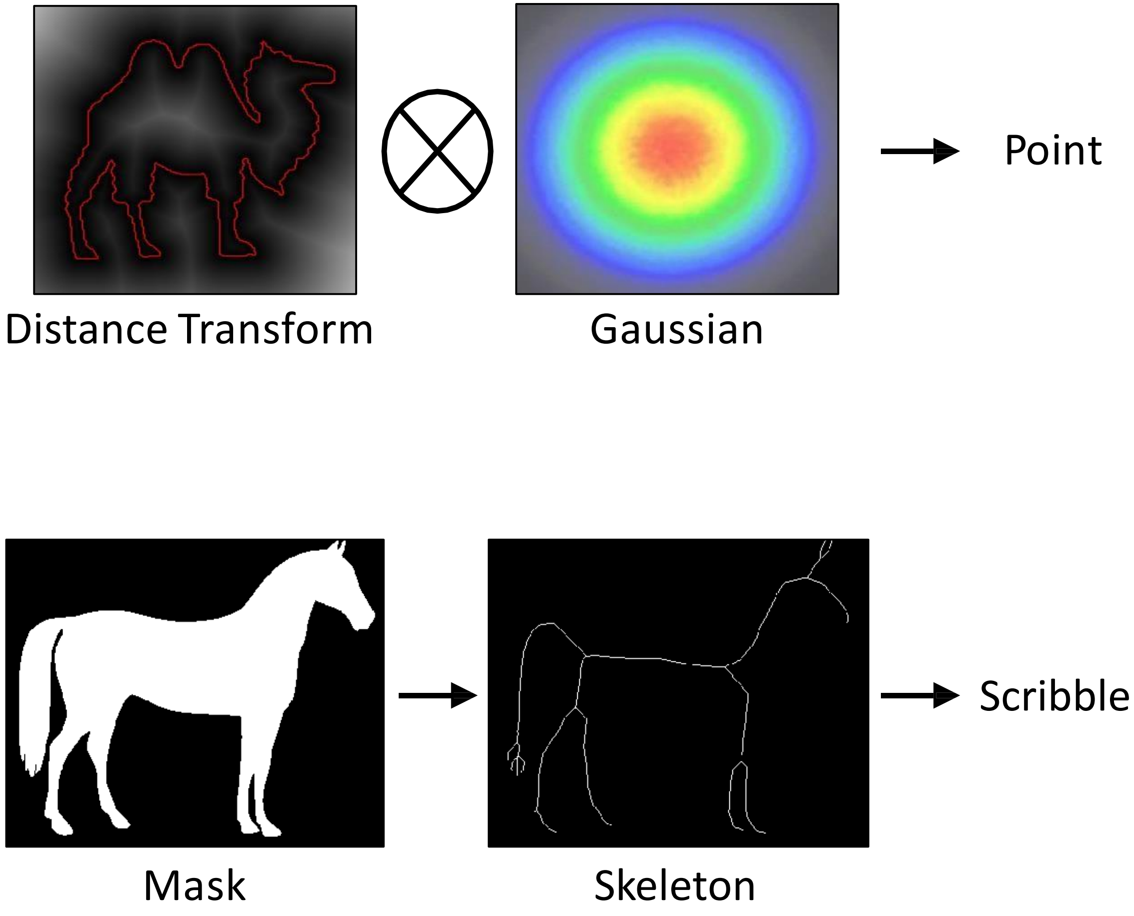
Unlabeled Data

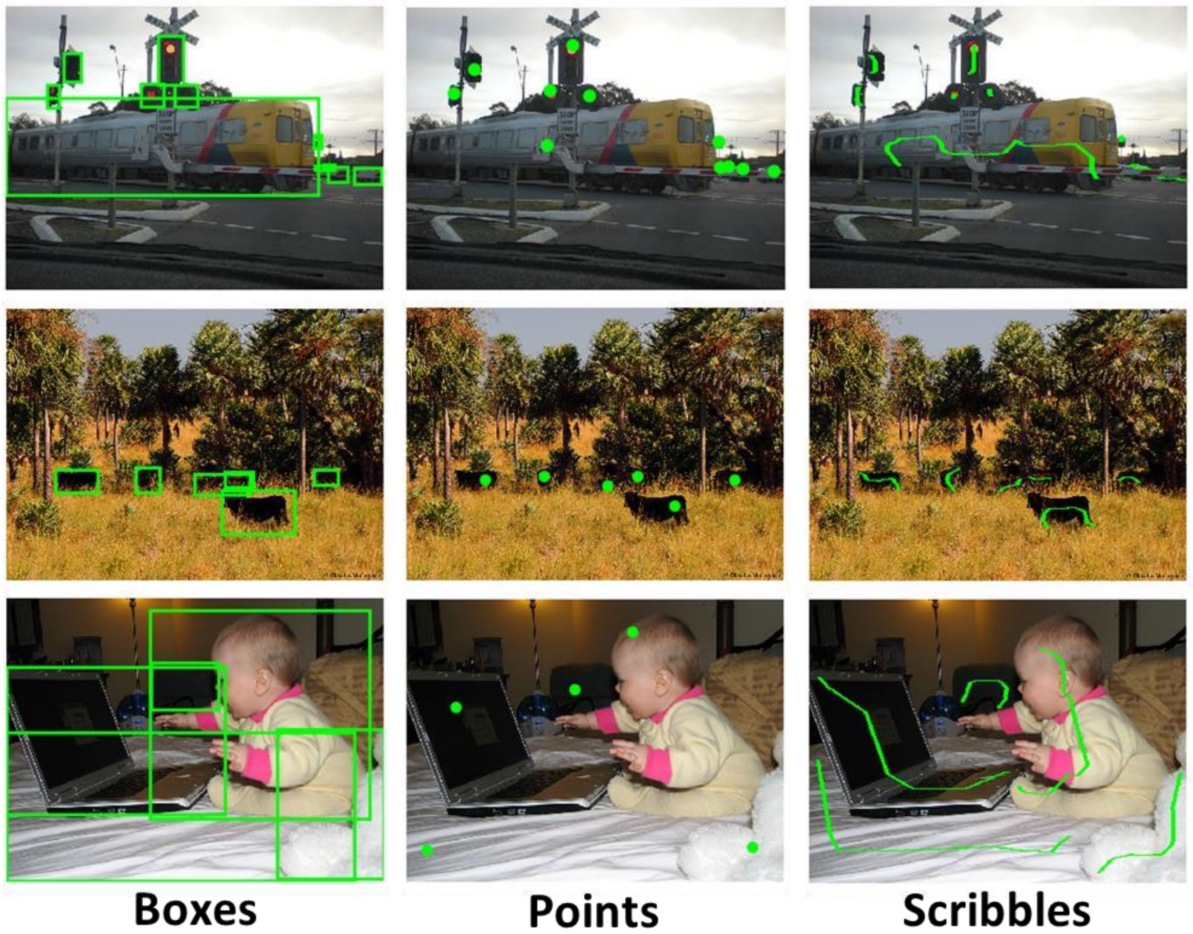


- Generate pseudo-tags by taking confident classes from teacher's predication
- Supervise teacher with pseudo-tags
- Generate pseudo-boxes to supervise student

Ren et al., UFO^2: A Unified Framework towards Omni-supervised Object Detection, ECCV20

Partially Labeled MS-COCO





Ren et al., UFO^2: A Unified Framework towards Omni-supervised Object Detection, ECCV20

Quantitative Results and Ablation Studies

Methods	Test-scale	Label	AP	AP-50
PCL [52]	multi	tags	8.5	19.4
C-MIDN [14]	multi	tags	9.6	21.4
WSOD2 [60]	multi	tags	10.8	22.7
Ours	multi	tags	**11.4**	**24.3**
Ours	single	tags	**10.8**	**23.1**
Ours	single	points	**12.4**	**27.0**
Ours	single	scribbles	**13.7**	**29.8**
Fast-RCNN [16]	single	boxes	18.9	38.6
Faster-RCNN [43]	single	boxes	21.2	41.5
Ours	single	boxes	**25.7**	**46.3**

Single model performance on COCO-80 (VGG-16)

Train	Methods	backbone	Labels	AP	Extra	Labels	AP	Δ
COCO-35	ours	VGG-16	tags	4.9	COCO-80	-	5.3	**8.2%**
COCO-115	ours	VGG-16	tags	12.9	Un-120	-	13.6	**5.4%**
COCO-35	ours	ResNet-50	tags	9.8	COCO-80	-	10.5	**7.1%**
COCO-35	ours	ResNet-50	boxes	29.1	COCO-80	tags	29.4	**1.0%**
COCO-35	ours	ResNet-50	boxes	29.1	COCO-80	points	30.1	**5.5%**
COCO-35	ours	ResNet-50	boxes	29.1	COCO-80	scribbles	30.9	**6.2%**
COCO-115	ours	ResNet-50	boxes	32.7	Un-120	-	33.9	**3.7%**

Model improvement with extra data and/or labels

Loss	$\mathcal{L}_C + \mathcal{L}_R$	$+\mathcal{L}_{T1}$	$+\mathcal{L}_{T2}$	$+\mathcal{L}_{T1} + \mathcal{L}_{T2}$	$+\mathcal{L}_{T1} + \mathcal{L}_{T2} + \mathcal{L}_{\text{Tags}}$
AP	22.6	24.8	25.1	25.5	25.7
AP-50	42.4	44.1	45.0	46.0	46.3

Strong teacher helps model

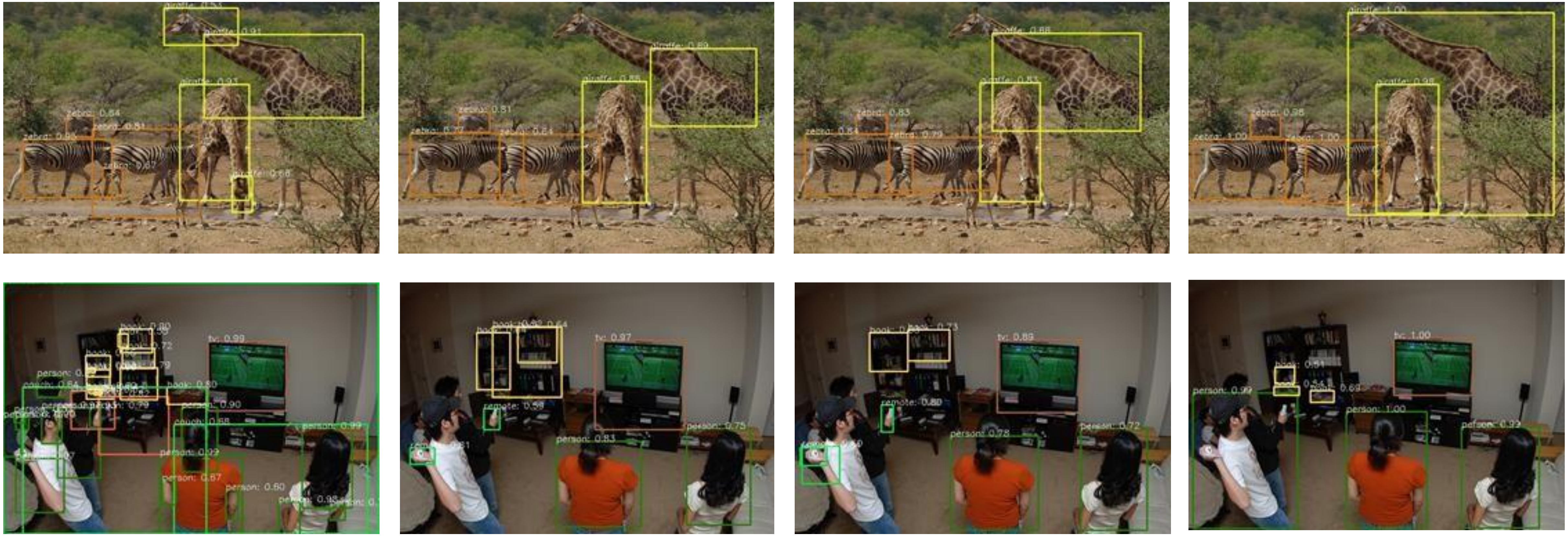
Methods	Tags	Ref_P	Ref_S	$\text{Ref}_P + \text{Con}_P$	$\text{Ref}_S + \text{Con}_S$
AP	10.8	11.1	11.6	12.4	13.7
AP-50	23.1	24.2	25.1	27.0	29.8

Localization constraints (Con_P/Con_S for points/scribbles) and proposal refinement (Ref_P/Ref_S) helps model

Ren et al., UFO^2: A Unified Framework towards Omni-supervised Object Detection, ECCV20

Qualitative Results

Tags	Points	Scribbles	Boxes



Ren et al., UFO^2: A Unified Framework towards Omni-supervised Object Detection, ECCV20

Budget-Aware Omni-Supervised Detection

Approx. per-img budget on COCO:

- Tags: 80s, points: 88.7s, scribbles: 160.4s, boxes: 346s

Given a fixed annotation budget (time), the most common strategy:

- STRONG: annotate using only boxes

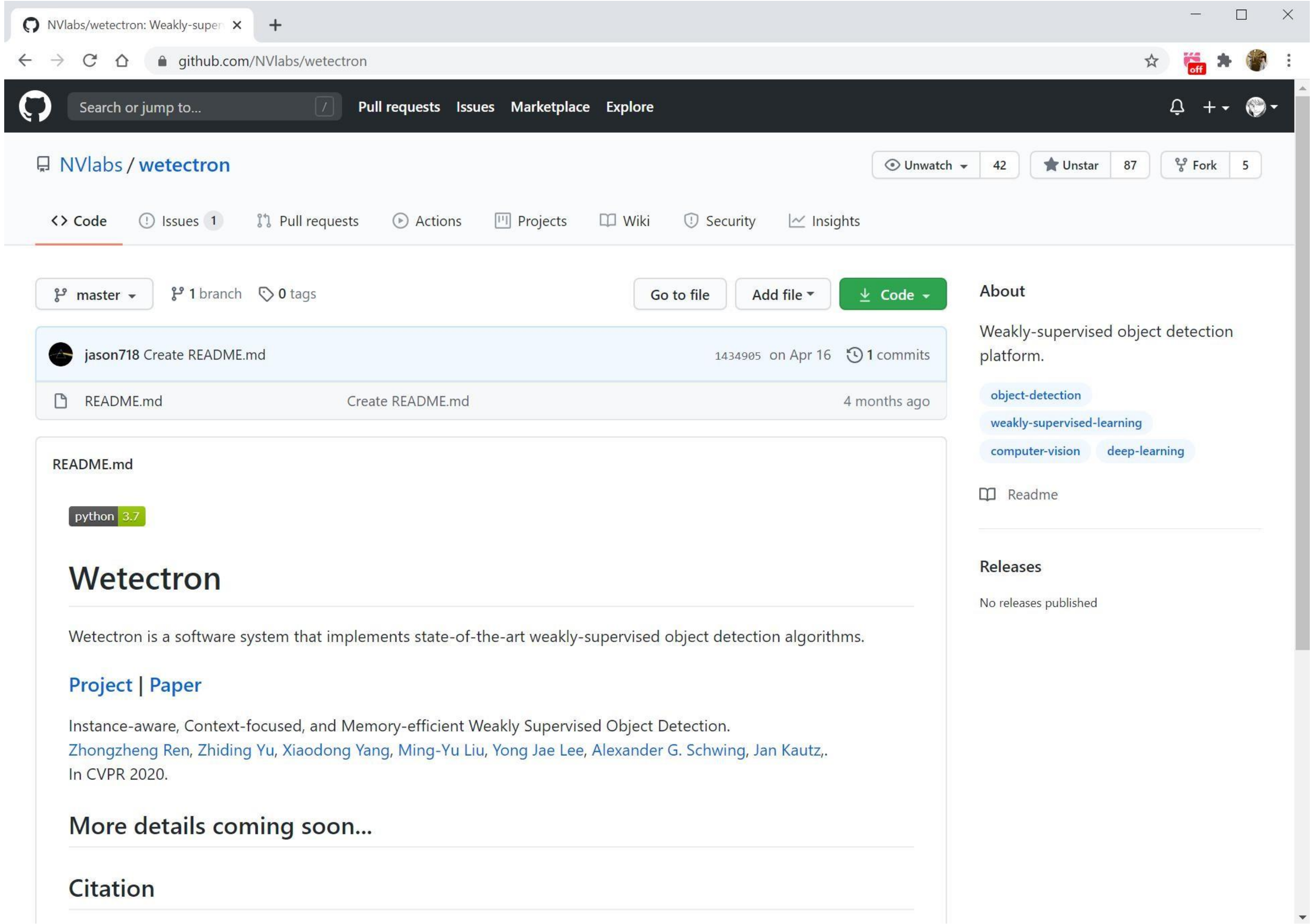
UFO^2 shows the possibility of a more efficient new policy:

- N%B: use N% budget for boxes and (1-N%) for points

Policy	Labels	# Labeled Images	AP
MOST	T	10000	3.00 ± 0.57
STRONG	**B + U**	**2312 + 7688**	**13.97 ± 0.98**
EQUAL	T + P + S + B + U	2500 + 2255 + 1250 + 5417	5.87 ± 0.70
80%B	**P + B + U**	**1804 + 1850 + 6346**	**14.11 ± 1.01**

Ren et al., UFO^2: A Unified Framework towards Omni-supervised Object Detection, ECCV20

Wetectron: A Partially Supervised Learning Platform



Code Repository

https://github.com/NVlabs/wetectron

Conclusions and Future Work

Conclusions

- Eliminating uncertainty is the key
- The diversity of supervisions create challenges but also brings abundant implicit constraints

Future Works

- Multi-modal/multi-task platform vs. single algorithms
- Post-deployment continuous interactive learning
- Inspiration from human baby learning

Thank You!