CutLER: Cut and Learn for Unsupervised Object Detection and Instance Segmentation

XuDong Wang[1,2], Rohit Girdhar[1], Stella X. Yu[2,3], Ishan Misra[1]
[1]FAIR - Meta AI; [2]UC Berkeley; [3]University of Michigan

CVPR 2023
Poster Session and ID: TUE-AM-297

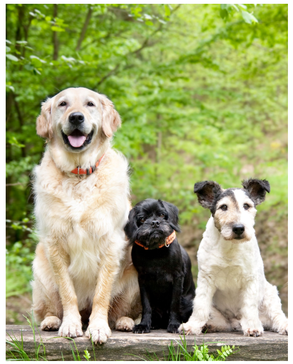
Berkeley UNIVERSITY OF CALIFORNIA Meta AI UNIVERSITY OF MICHIGAN

Goal

- Learn unsupervised object detection and instance segmentation model on ImageNet-1K without using any human annotations.

- Unsupervised representation learning for fully/semi-supervised object detection and instance segmentation tasks.

Our Pipeline: Cut-and-Learn (CutLER)



unlabeled data

ViT

MaskCut

pseudo masks

Detector

L_{drop}

self-training

Cut

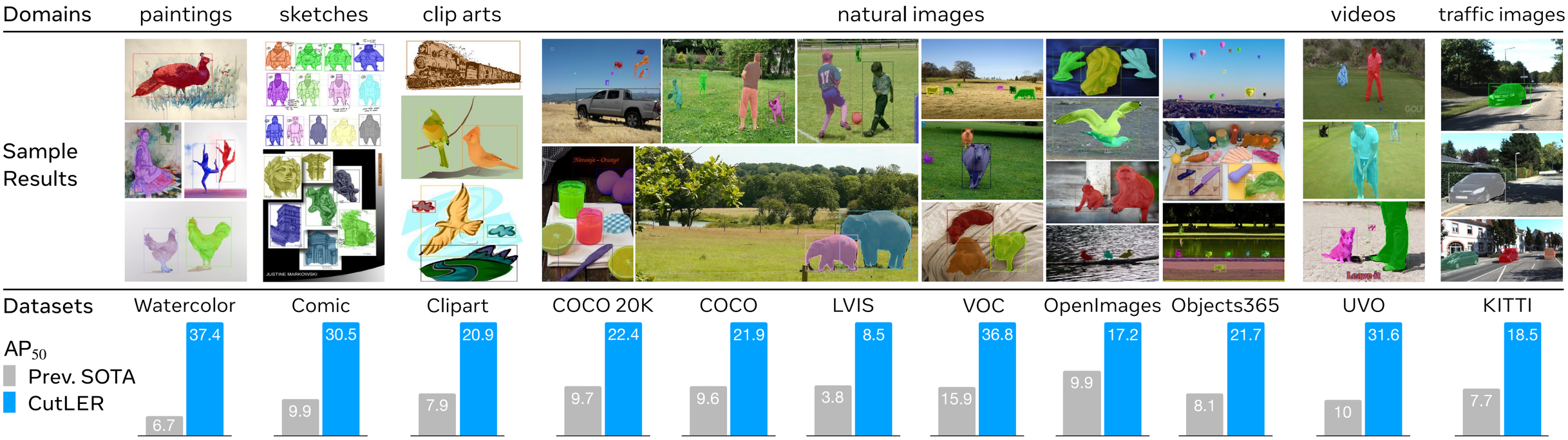
Learn

DropLoss: $\mathcal{L}_{\text{drop}}(r_i) = \mathbb{1}(\text{IoU}_i^{\max} > \tau^{\text{IoU}})\mathcal{L}_{\text{vanilla}}(r_i)$

Evaluate CutLER on 11 Different Datasets

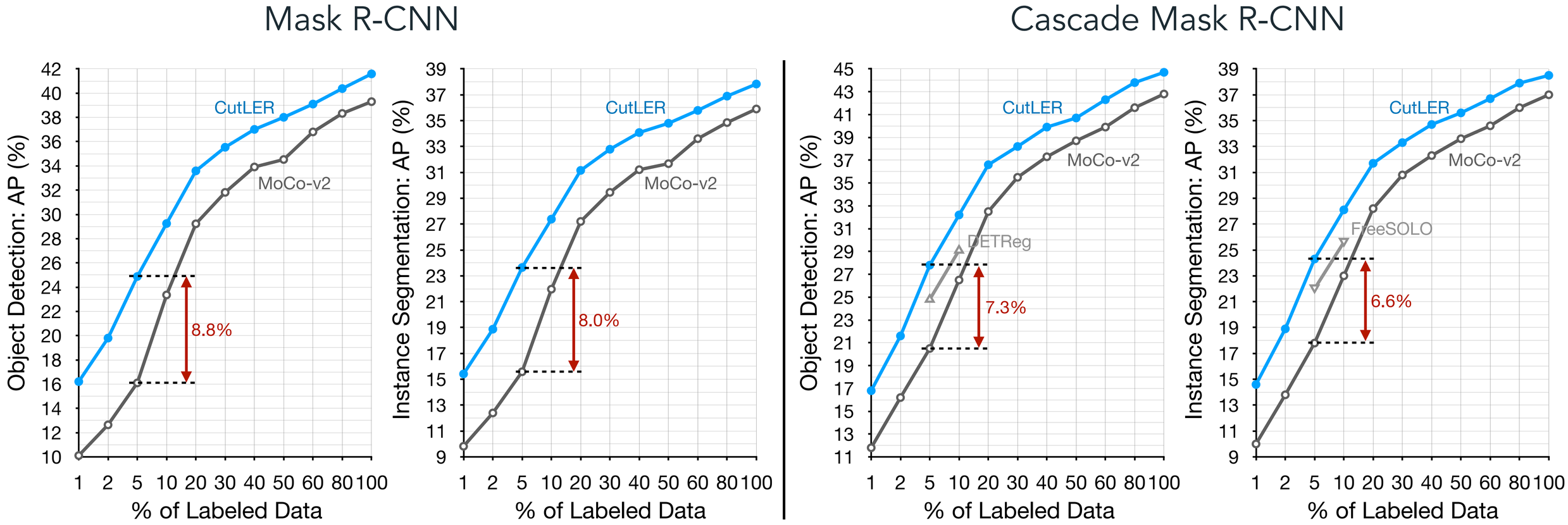
2.7 times higher precision

2.6 times higher recall



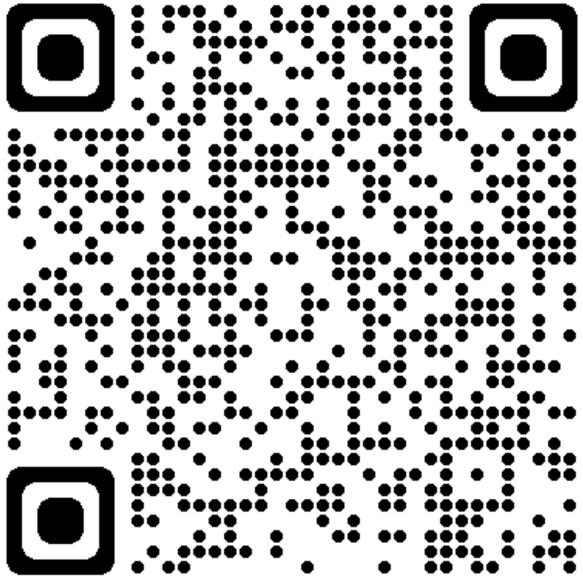
CutLER Is a Better Pretrained Model for Detection and Segmentation

Detectors are initialized with CutLER / baselines pre-trained on ImageNet-1K.



Our code is available!

https://github.com/
facebookresearch/CutLER



CutLER: Cut and Learn for Unsupervised Object Detection and Instance Segmentation

XuDong Wang[1,2], Rohit Girdhar[1], Stella X. Yu[2,3], Ishan Misra[1]
[1]FAIR - Meta AI; [2]UC Berkeley; [3]University of Michigan

CVPR 2023
Poster Session and ID: TUE-AM-297

Berkeley UNIVERSITY OF CALIFORNIA Meta AI UNIVERSITY OF MICHIGAN

Settings

Training time:

- unlabeled data WITHOUT ANY human annotation (pixel-/instance-/image-level annotations)

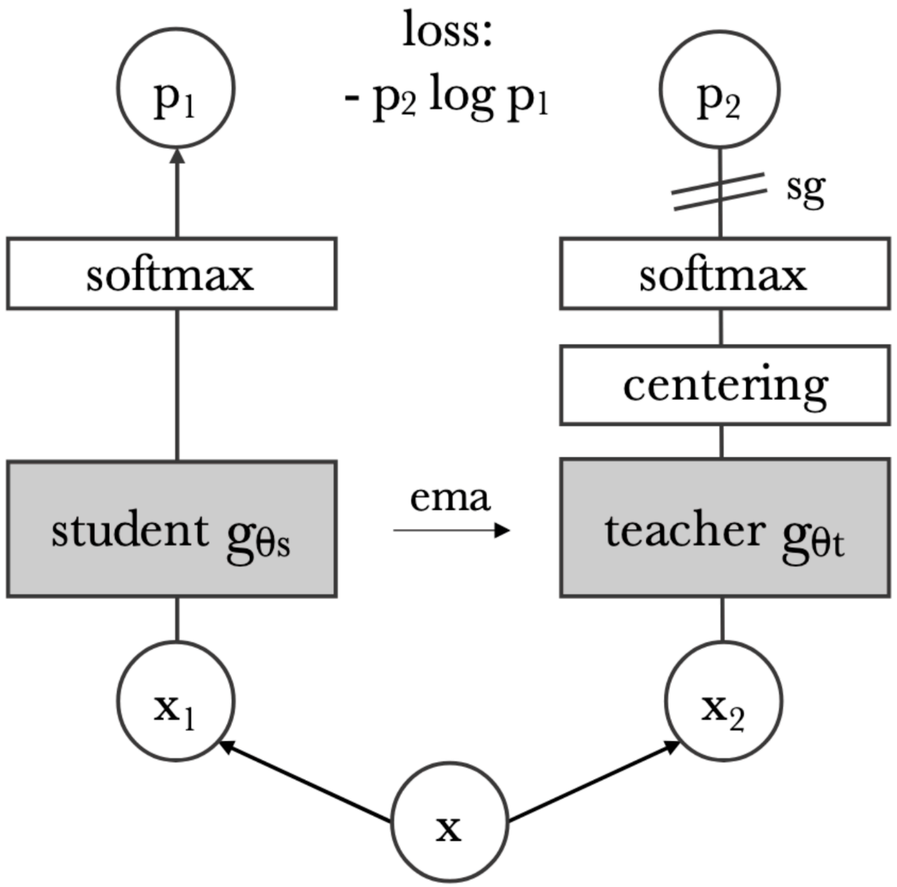
Inference time:

- provide bounding boxes and segmentation masks WITHOUT using labels

Key insight:

Simple probing and training mechanisms can amplify the innate localization ability of self-supervised models

DINO: Emerging Properties in SSL ViTs



Self Supervised Learning with Knowledge Distillation (KD)

Paradigm: train a student network to match the output of a teacher network

$$\min_{\theta_s} H(P_t(x), P_s(x))$$

Mathilde Caron, et al. Emerging properties in self-supervised vision transformers. ICCV 2021

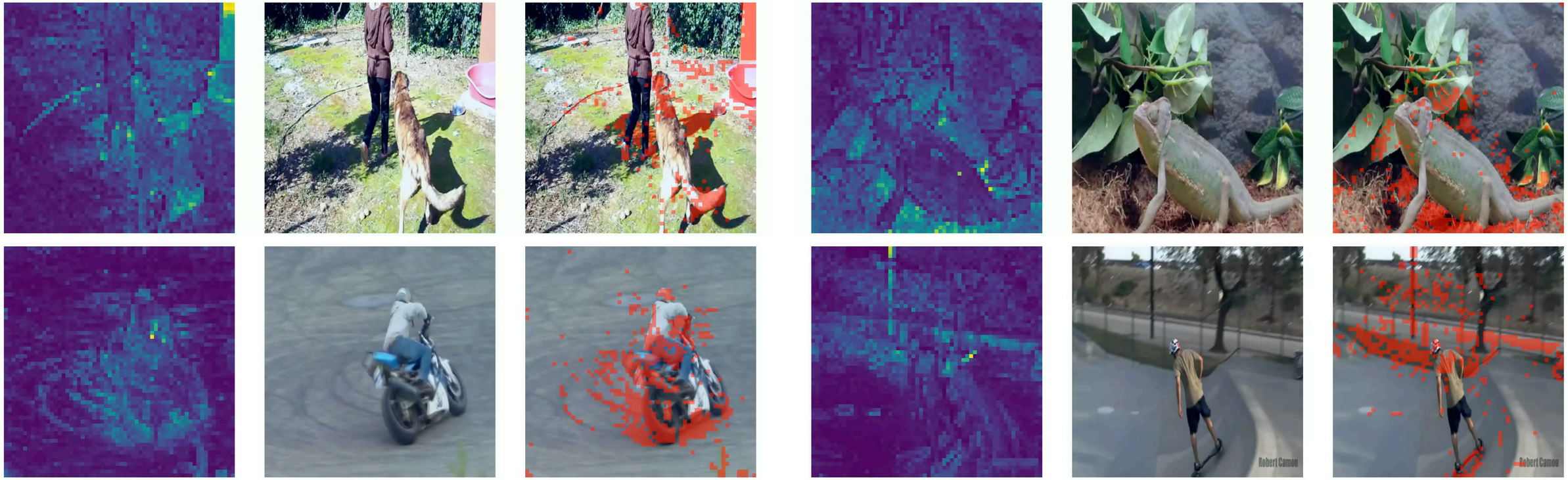
Emerging Property: discovering the semantic layout of scenes

Probing the self-attention map (self-attention for [CLS] token query)

Mathilde Caron, et al. Emerging properties in self-supervised vision transformers. ICCV 2021

Downsides of DINO's Results

- unable to distinguish different instances.
- unsatisfactory results when working in complex environments.
- only produce a single mask per image.



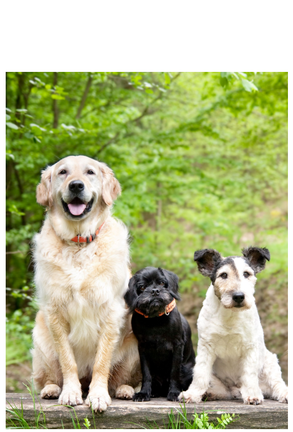
Mathilde Caron, et al. Emerging properties in self-supervised vision transformers. ICCV 2021

Berkeley UNIVERSITY OF CALIFORNIA Meta AI UNIVERSITY OF MICHIGAN

Desired Properties

- ✓ Unsupervised learning without any human annotation
- ✓ Multi-object detection and instance segmentation
- ✓ Zero-shot detector that can be applied to various domains.
- ✓ Compatible with various detection architectures
- ✓ Improved pre-trained model for fully-supervised and label-efficient learning

Our Pipeline: Cut-and-Learn (CutLER)



unlabeled data

ViT

MaskCut

pseudo masks

Detector

L_{drop}

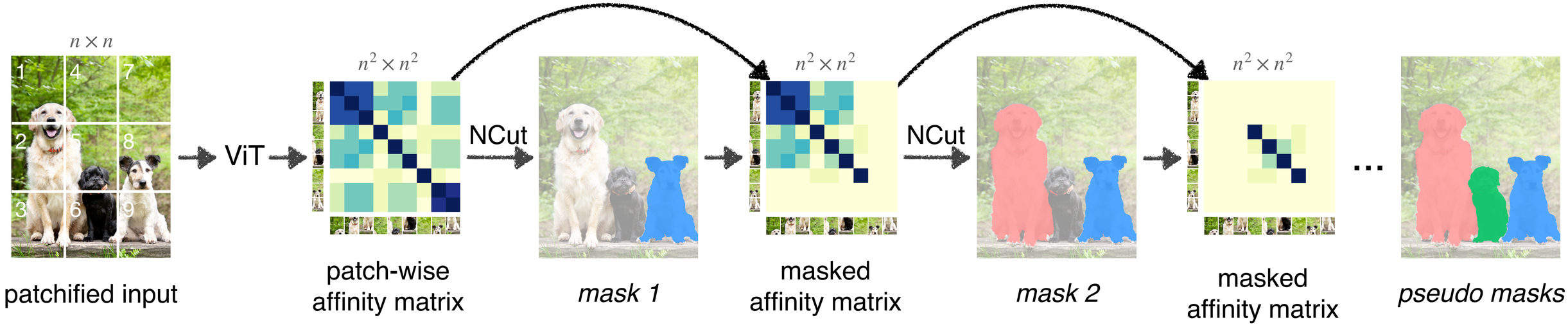
self-training

Cut

Learn

DropLoss: $\mathcal{L}_{\text{drop}}(r_i) = \mathbb{1}(\text{IoU}_i^{\max} > \tau^{\text{IoU}})\mathcal{L}_{\text{vanilla}}(r_i)$

Masked Normalized Cuts (MaskCut)

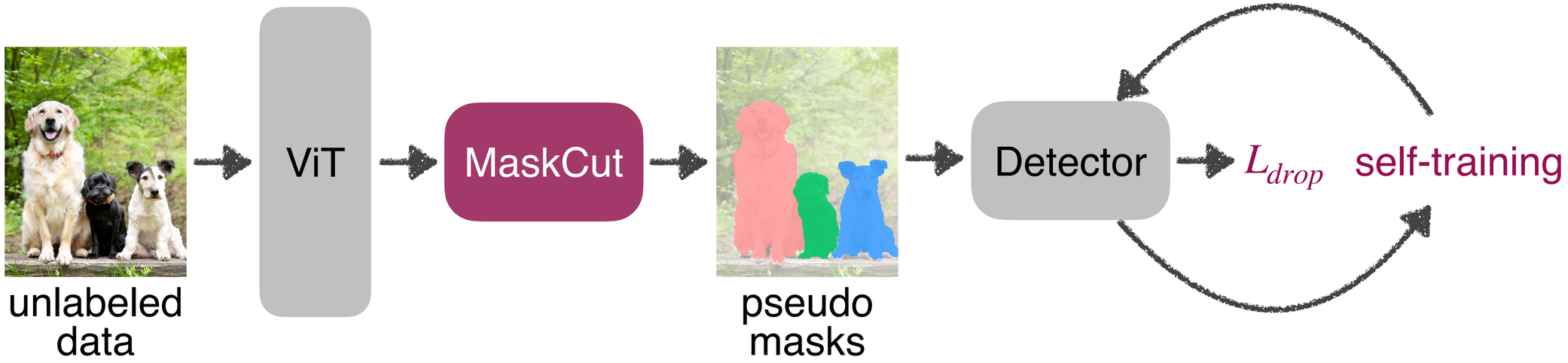


Mathilde Caron, et al. Emerging properties in self-supervised vision transformers. ICCV 2021

Berkeley UNIVERSITY OF CALIFORNIA Meta AI UNIVERSITY OF MICHIGAN

MaskCut Demos (w/ DINO)

Cut-and-Learn (CutLER)

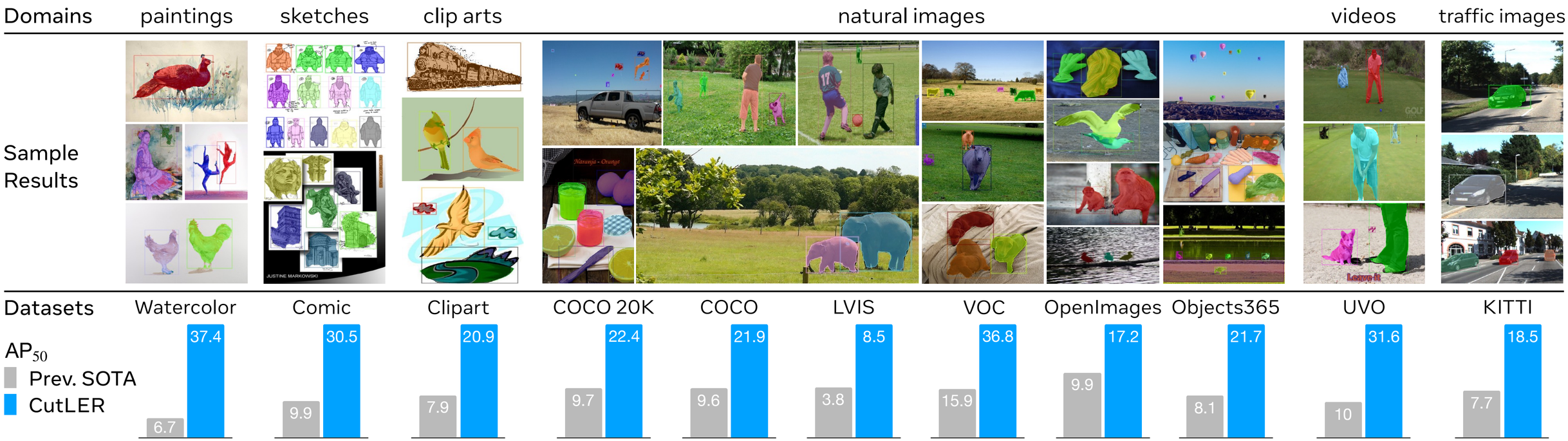


Train CutLER Solely on Unlabeled ImageNet-1K

Berkeley UNIVERSITY OF CALIFORNIA Meta AI UNIVERSITY OF MICHIGAN

Train CutLER Solely on Unlabeled ImageNet-1K

Evaluate CutLER on 11 Different Datasets



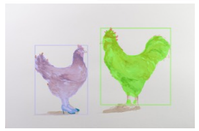
Berkeley UNIVERSITY OF CALIFORNIA Meta AI UNIVERSITY OF MICHIGAN

Train CutLER Solely on Unlabeled ImageNet-1K

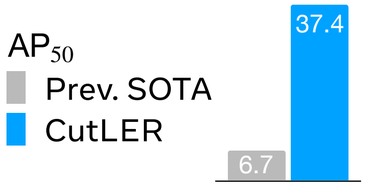
Evaluate CutLER on 11 Different Datasets - paintings

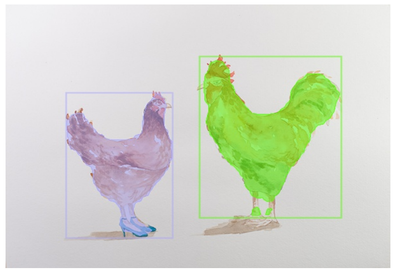
Domains paintings

Sample Results



Datasets Watercolor





Berkeley UNIVERSITY OF CALIFORNIA Meta AI UNIVERSITY OF MICHIGAN

Train CutLER Solely on Unlabeled ImageNet-1K

Evaluate CutLER on 11 Different Datasets - sketches

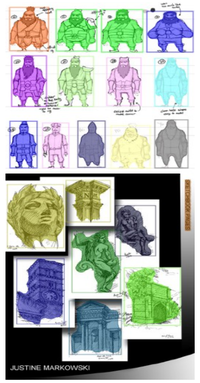
Domains

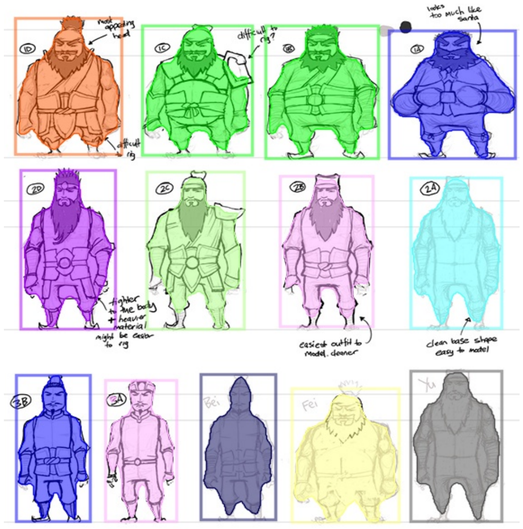
Sample Results

Datasets

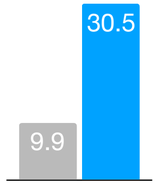
AP_{50}
- Prev. SOT
- CutLER

sketches





Comic



Berkeley UNIVERSITY OF CALIFORNIA Meta AI UNIVERSITY OF MICHIGAN

Train CutLER Solely on Unlabeled ImageNet-1K

Evaluate CutLER on 11 Different Datasets - clip arts

Domains

Sample Results

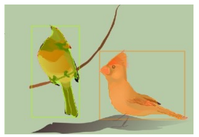
Datasets

AP_{50}

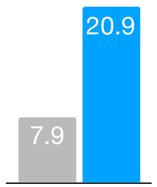
Prev. SOT

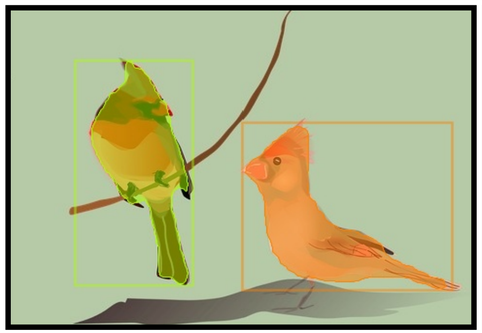
CutLER

clip arts



Clipart





Berkeley UNIVERSITY OF CALIFORNIA

Meta AI

UNIVERSITY OF MICHIGAN

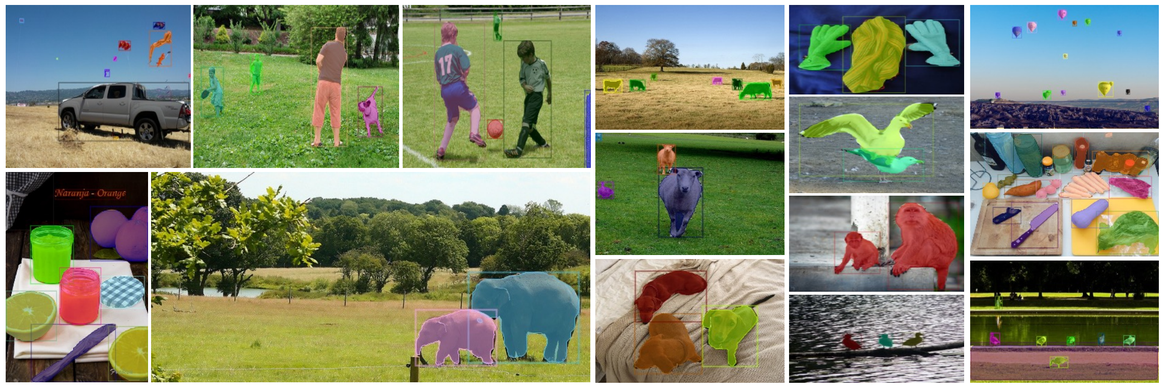
Train CutLER Solely on Unlabeled ImageNet-1K

Evaluate CutLER on 11 Different Datasets - natural images

Domains

natural images

Sample Results



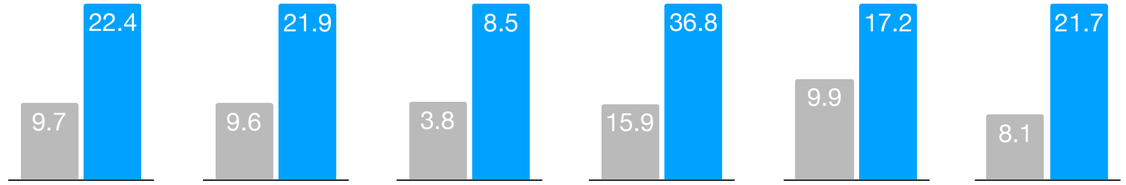
Datasets

COCO 20K COCO LVIS VOC OpenImages Objects365

AP_{50}

Prev. SOT

CutLER



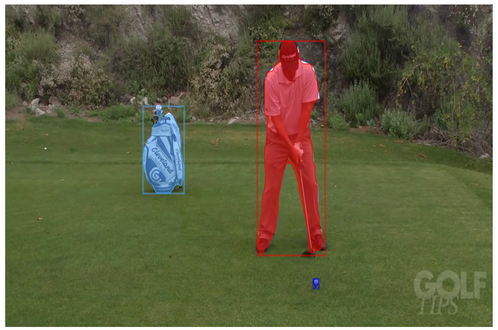
Berkeley UNIVERSITY OF CALIFORNIA Meta AI UNIVERSITY OF MICHIGAN

Train CutLER Solely on Unlabeled ImageNet-1K

Evaluate CutLER on 11 Different Datasets - videos

Domains

Sample Results



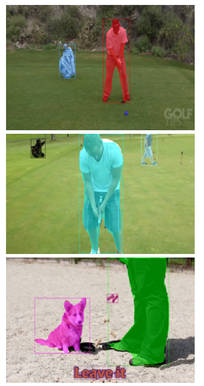
Datasets

AP_{50}

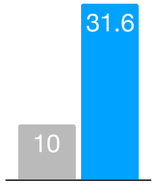
Prev. SOT

CutLER

videos



UVO



Berkeley UNIVERSITY OF CALIFORNIA

Meta AI

UNIVERSITY OF MICHIGAN

Train CutLER Solely on Unlabeled ImageNet-1K

Evaluate CutLER on 11 Different Datasets - traffic images

Domains

traffic images

Sample Results

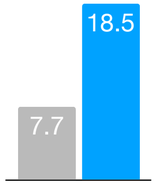
Datasets

KITTI

AP_{50}

Prev. SOT

CutLER



Berkeley UNIVERSITY OF CALIFORNIA

Meta AI

UNIVERSITY OF MICHIGAN

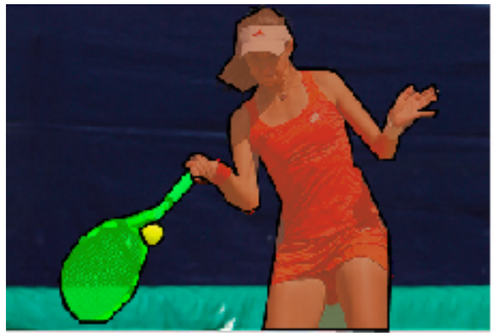
CutLER Can Detect and Segment Small Objects

Quantitative Results on 11 Datasets

2.7 times higher precision

2.6 times higher recall

Datasets →	Avg.		COCO		COCO20K		VOC		LVIS		UVO		Clipart		Comic		Watercolor		KITTI		Objects365		OpenImages	
Metrics →	AP_{50}	AR	AP_{50}	AR	AP_{50}	AR	AP_{50}	AR	AP_{50}	AR	AP_{50}	AR	AP_{50}	AR	AP_{50}	AR	AP_{50}	AR	AP_{50}	AR	AP_{50}	AR	AP_{50}	AR
Prev. SOTA [47]	9.0	13.4	9.6	12.6	9.7	12.6	15.9	21.3	3.8	6.4	10.0	14.2	7.9	15.1	9.9	16.3	6.7	16.2	7.7	7.1	8.1	10.2	9.9	14.9
CutLER	24.3	35.5	21.9	32.7	22.4	33.1	36.9	44.3	8.4	21.8	31.7	42.8	21.1	41.3	30.4	38.6	37.5	44.6	18.4	27.5	21.6	34.2	17.3	29.6
vs. prev. SOTA	**+15.3**	**+22.1**	**+12.3**	**+20.1**	**+12.7**	**+20.5**	**+21.0**	**+23.0**	**+4.6**	**+15.4**	**+21.7**	**+28.6**	**+13.2**	**+26.2**	**+20.5**	**+22.3**	**+30.8**	**+28.4**	**+10.7**	**+20.4**	**+13.5**	**+24.0**	**+7.4**	**+14.7**



COCO: 80 classes

LVIS: 1203 classes

vs. Previous SOTA:

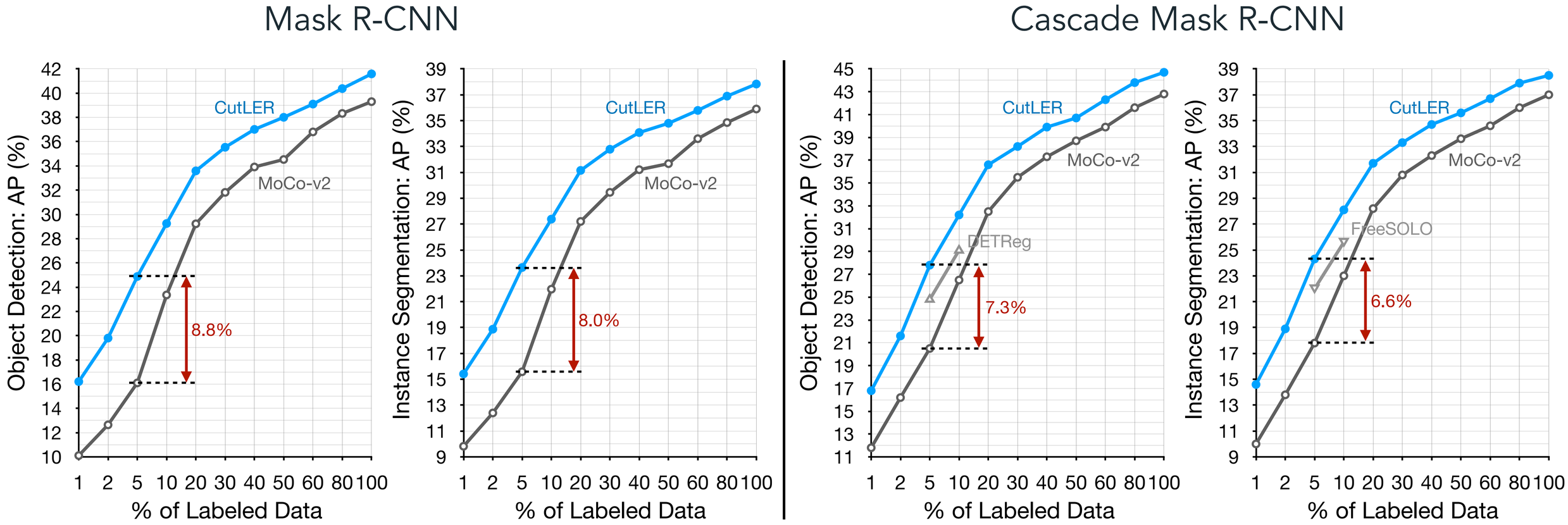
- ✓ better discriminate instances;
- ✓ discover more objects;
- ✓ produce precise segmentation masks for small objects

vs. Human Annotations

✓ locate novel instances that are overlooked by human annotators

CutLER Can Replace MoCo/FreeSOLO as a Pretrained Model

Detectors are initialized with CutLER / baselines pre-trained on ImageNet-1K.



Berkeley UNIVERSITY OF CALIFORNIA Meta AI UNIVERSITY OF MICHIGAN

Our code is available!

https://github.com/
facebookresearch/CutLER