Vision Transformers are Good Mask Auto-labelers

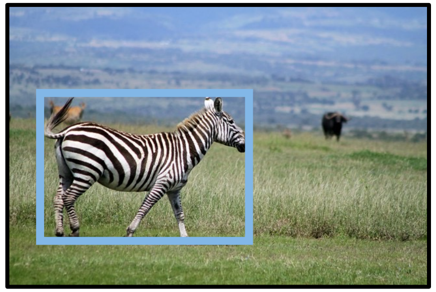
Shiyi Lan, Xitong Yang, Zhiding Yu, Zuxuan Wu,
Jose M. Alvarez, Anima Anandkumar

JUNE 18-22, 2023
CVPR
VANCOUVER, CANADA

Background

- Instance Segmentation Annotations are expensive
  - For COCO Object Detection Dataset, **78%** annotation time is spent on segmentation

Image Classification



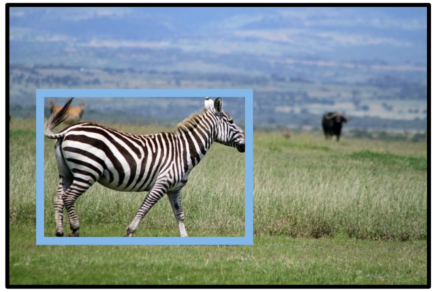
Object Detection

Instance Segmentation

Background

- Instance Segmentation Annotations are expensive
  - For COCO Object Detection Dataset, **78%** annotation time is spent on segmentation
- Auto-labeling Instance Segmentation with no grounding is **very hard**

Image Classification



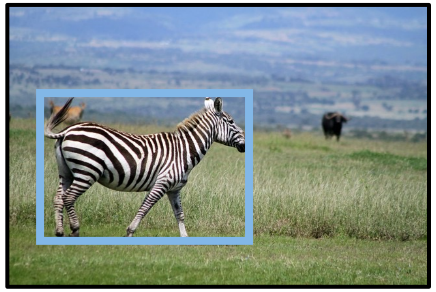
Object Detection

Instance Segmentation

Background

- Instance Segmentation Annotations are expensive
  - For COCO Object Detection Dataset, **78%** annotation time is spent on segmentation
- Auto-labeling Instance Segmentation with no grounding is **very hard**
- Given Ground-truth bounding boxes, mask auto-labeling becomes **easier**

Image Classification

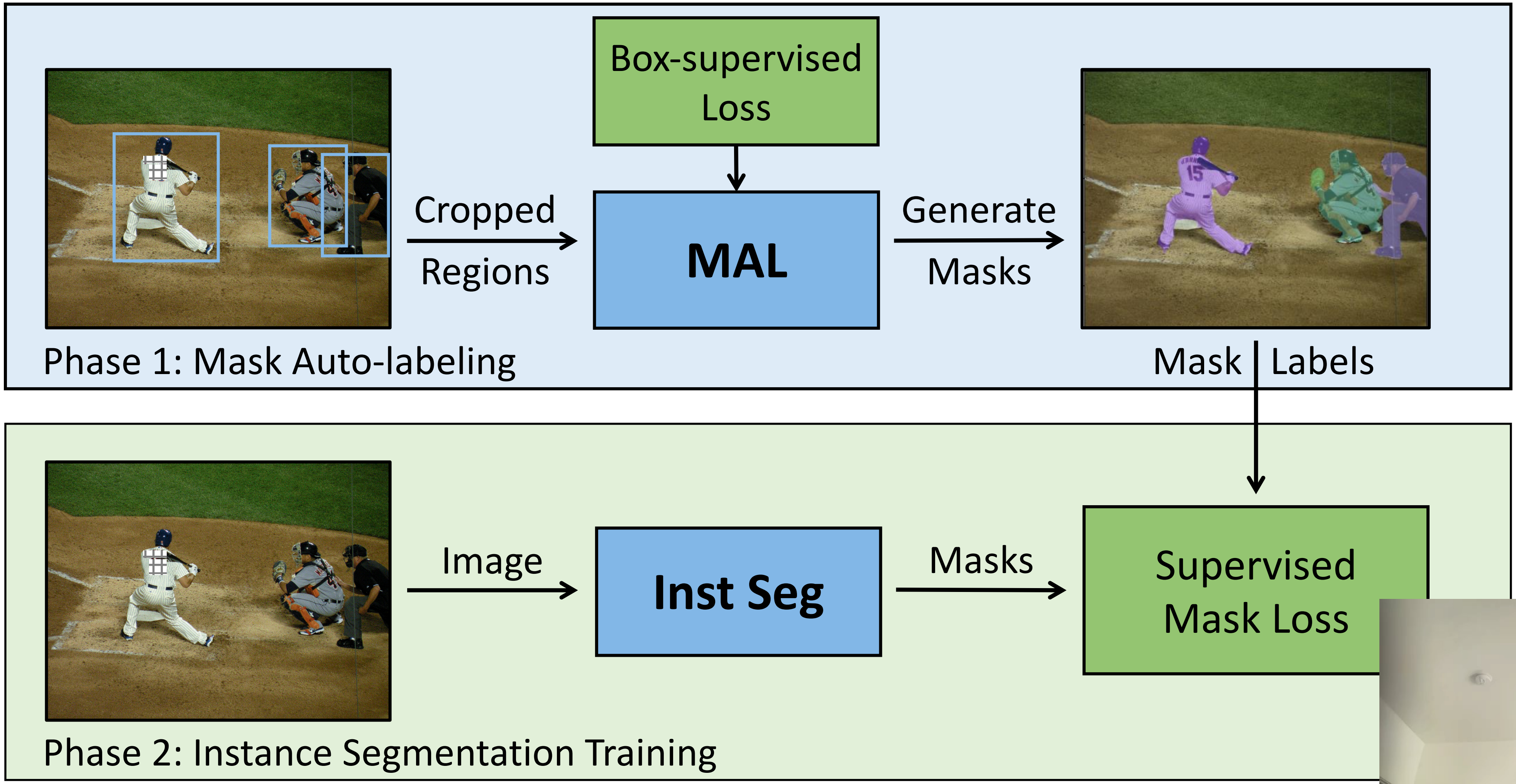


Object Detection

Instance Segmentation

Our Framework

Our novel mask auto-labeling network: **MAL** (Mask Auto-labeler)



Quantitative Results

Method	Labeler Backbone	InstSeg Backbone	InstSeg Model	Sup	(%)Mask AP_{val}	(%)Mask AP_{test}	(%)$Ret._{val}$	(%)$Ret._{test}$
Mask R-CNN* [24]	-	ResNet-101	Mask R-CNN	Mask	38.6	38.8	-	-
Mask R-CNN* [24]	-	ResNeXt-101	Mask R-CNN	Mask	39.5	39.9	-	-
CondInst [33]	-	ResNet-101	CondInst	Mask	38.6	39.1	-	-
SOLOv2 [31]	-	ResNet-50	SOLOv2	Mask	37.5	38.4	-	-
SOLOv2 [31]	-	ResNet-101-DCN	SOLOv2	Mask	41.7	41.8	-	-
SOLOv2 [31]	-	ResNeXt-101-DCN	SOLOv2	Mask	42.4	42.7	-	-
ConvNeXt [44]	-	ConvNeXt-Small [44]	Cascade R-CNN	Mask	44.8	45.5	-	-
ConvNeXt [44]	-	ConvNeXt-Base [44]	Cascade R-CNN	Mask	45.4	46.1	-	-
Mask2Former [41]	-	Swin-Small	Mask2Former	Mask	**46.1**	**47.0**	-	-
BBTP† [4]	-	ResNet-101	Mask R-CNN	Box	-	21.1	-	59.1
BoxInst [5]	-	ResNet-101	CondInst	Box	33.0	33.2	85.5	84.9
BoxLevelSet [6]	-	ResNet-101-DCN	SOLOv2	Box	35.0	35.4	83.9	83.5
DiscoBox [7]	-	ResNet-50	SOLOv2	Box	30.7	32.0	81.9	83.3
DiscoBox [7]	-	ResNet-101-DCN	SOLOv2	Box	35.3	35.8	84.7	85.9
DiscoBox [7]	-	ResNeXt-101-DCN	SOLOv2	Box	37.3	37.9	88.0	88.8
BoxTeacher [8]	-	Swin-Base	CondInst	Box	-	40.0	-	-
Mask Auto-Labeler	ViT-MAE-Base [13]	ResNet-50	SOLOv2	Box	35.0	35.7	93.3	93.0
Mask Auto-Labeler	ViT-MAE-Base [13]	ResNet-101-DCN	SOLOv2	Box	38.2	38.7	91.6	92.6
Mask Auto-Labeler	ViT-MAE-Base [13]	ResNeXt-101-DCN	SOLOv2	Box	38.9	39.1	91.7	[illegible]
Mask Auto-Labeler	ViT-MAE-Base [13]	ConvNeXt-Small [44]	Cascade R-CNN	Box	42.3	43.0	94.4	[illegible]
Mask Auto-Labeler	ViT-MAE-Base [13]	ConvNeXt-Base [44]	Cascade R-CNN	Box	42.9	43.3	**94.5**	[illegible]
Mask Auto-Labeler	ViT-MAE-Base [13]	Swin-Small [12]	Mask2Former [41]	Box	**43.3**	**44.1**	93.9	[illegible]

Table 1. Main results on COCO. Ret means the retention rate of $\frac{\text{box-supervised mask AP}}{\text{supervised mask AP}}$. MAL with SOLOv2/ResNeXt-10
DiscoBox with SOLOv2/ResNeXt-101 by 1.6% on val2017 and 1.3% on test-dev. Our best model (Mask2former/Swin-S
43.3% AP on val and 44.1% AP on test-dev.

Motivation

- Vision Transformers are very good at segmentation
  - Self-emerging segmentation, e.g. DINO [1], FAN [2]
  - High-performance segmentor, e.g. SegFormer [3], Mask2Former [4]

[1] Caron, Mathilde, et al. "Emerging properties in self-supervised vision transformers." Proceedings of the IEEE/CVF international conference on computer vision. 2021.
[2] Zhou, Daquan, et al. "Understanding the robustness in vision transformers." International Conference on Machine Learning. PMLR, 2022.
[3] Xie, Enze, et al. "SegFormer: Simple and efficient design for semantic segmentation with transformers." Advances in Neural Information Processing Systems 34 (2021): 12077-120

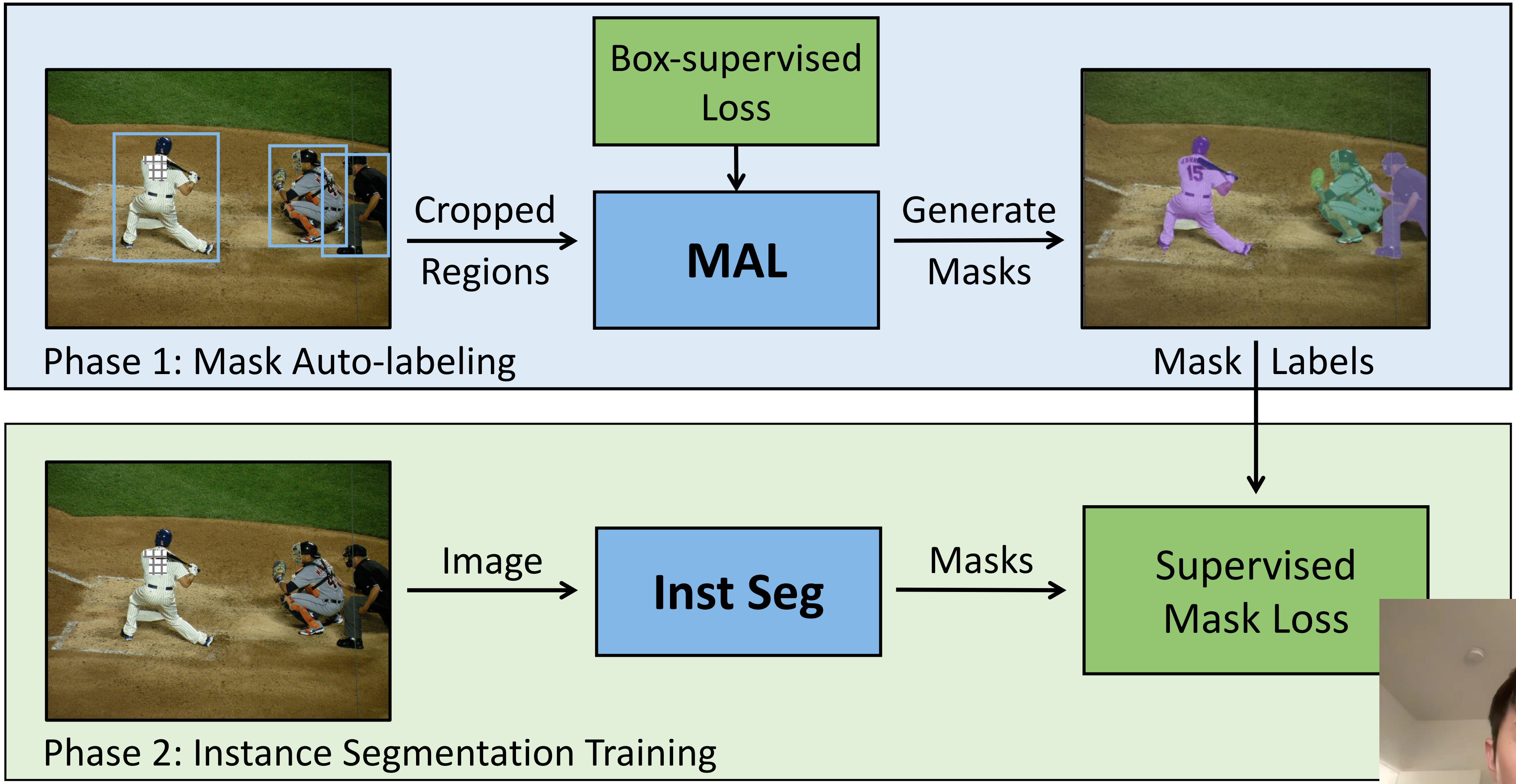
Motivation

- Vision Transformers are very good at segmentation
  - Self-emerging segmentation, e.g. DINO [1], FAN [2]
  - High-performance segmentor, e.g. SegFormer [3], Mask2Former [4]
- No need to cover detection in mask auto-labeling
  - Ground-truth bounding boxes are always better than predicted boxes
  - Model can focus on sole task

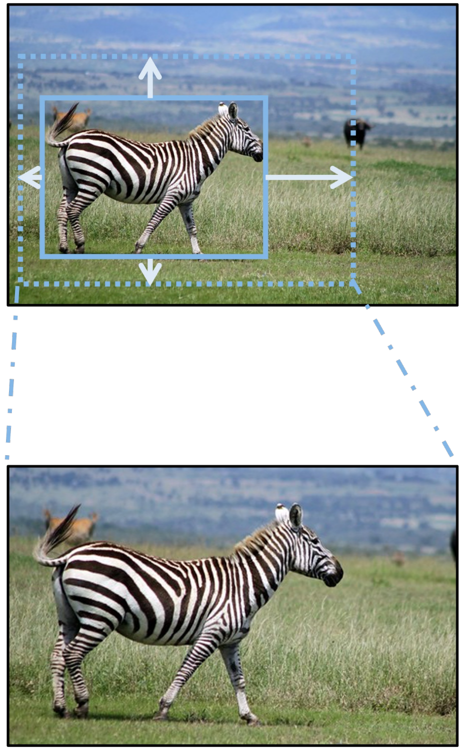
[1] Caron, Mathilde, et al. "Emerging properties in self-supervised vision transformers." Proceedings of the IEEE/CVF international conference on computer vision. 2021.
[2] Zhou, Daquan, et al. "Understanding the robustness in vision transformers." International Conference on Machine Learning. PMLR, 2022.
[3] Xie, Enze, et al. "SegFormer: Simple and efficient design for semantic segmentation with transformers." Advances in Neural Information Processing Systems 34 (2021): 12077-120

Our Framework

Our novel mask auto-labeling network: **MAL** (Mask Auto-labeler)

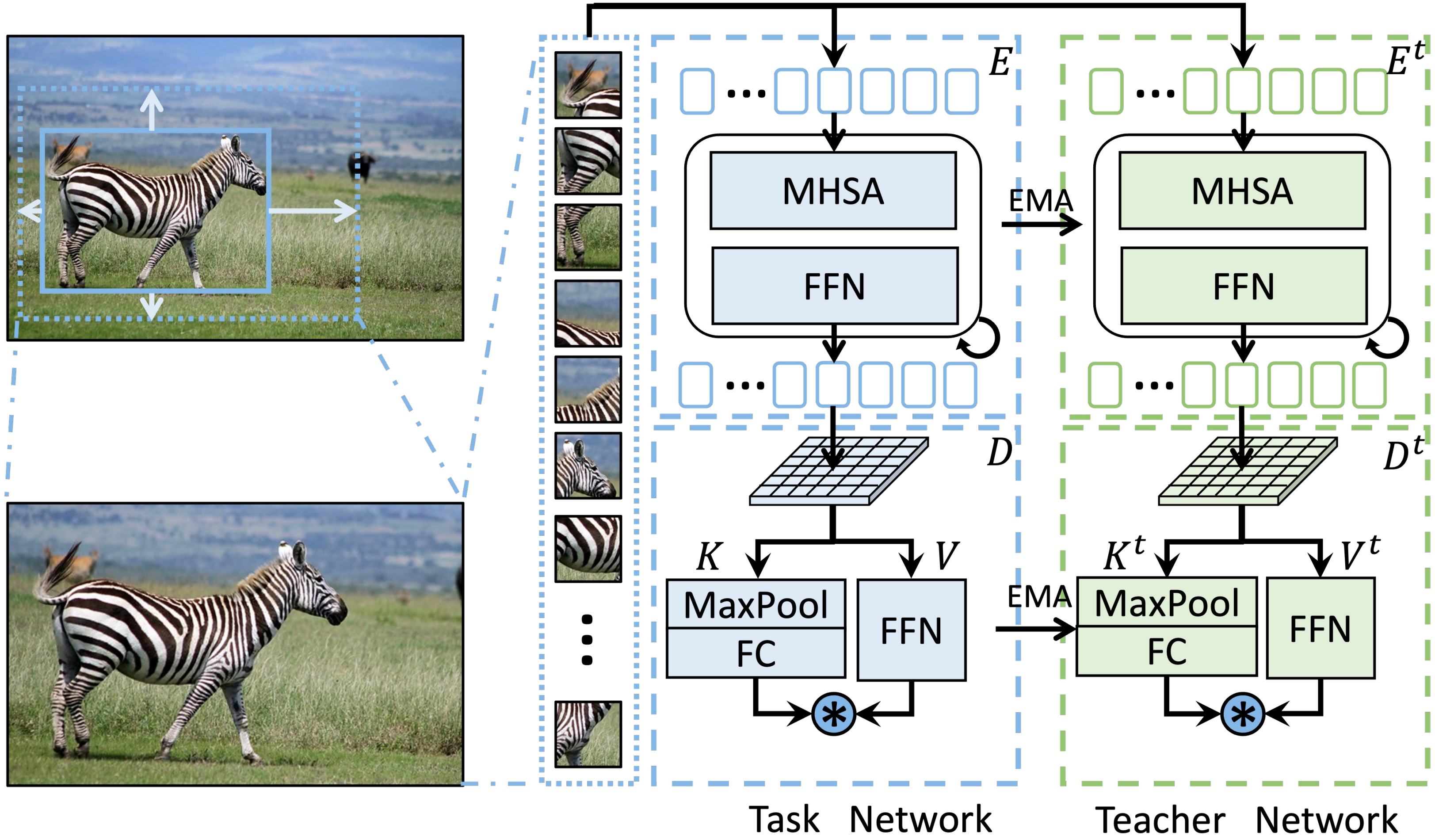


MAL Pipeline

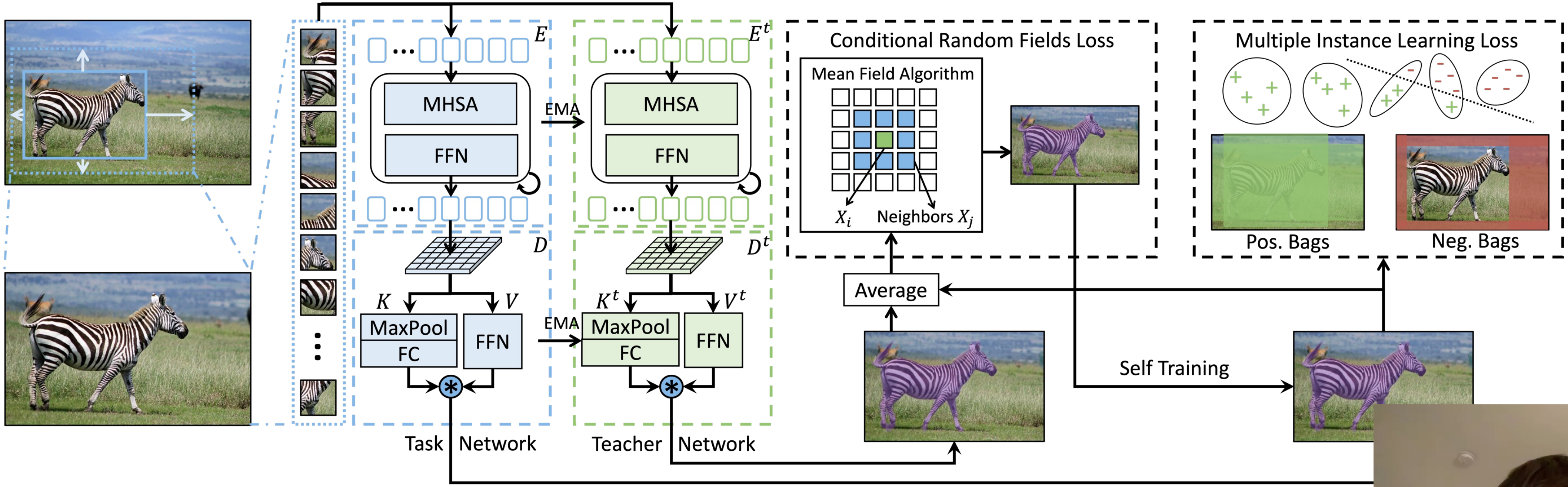


Randomly expand the GT box
and
crop the images

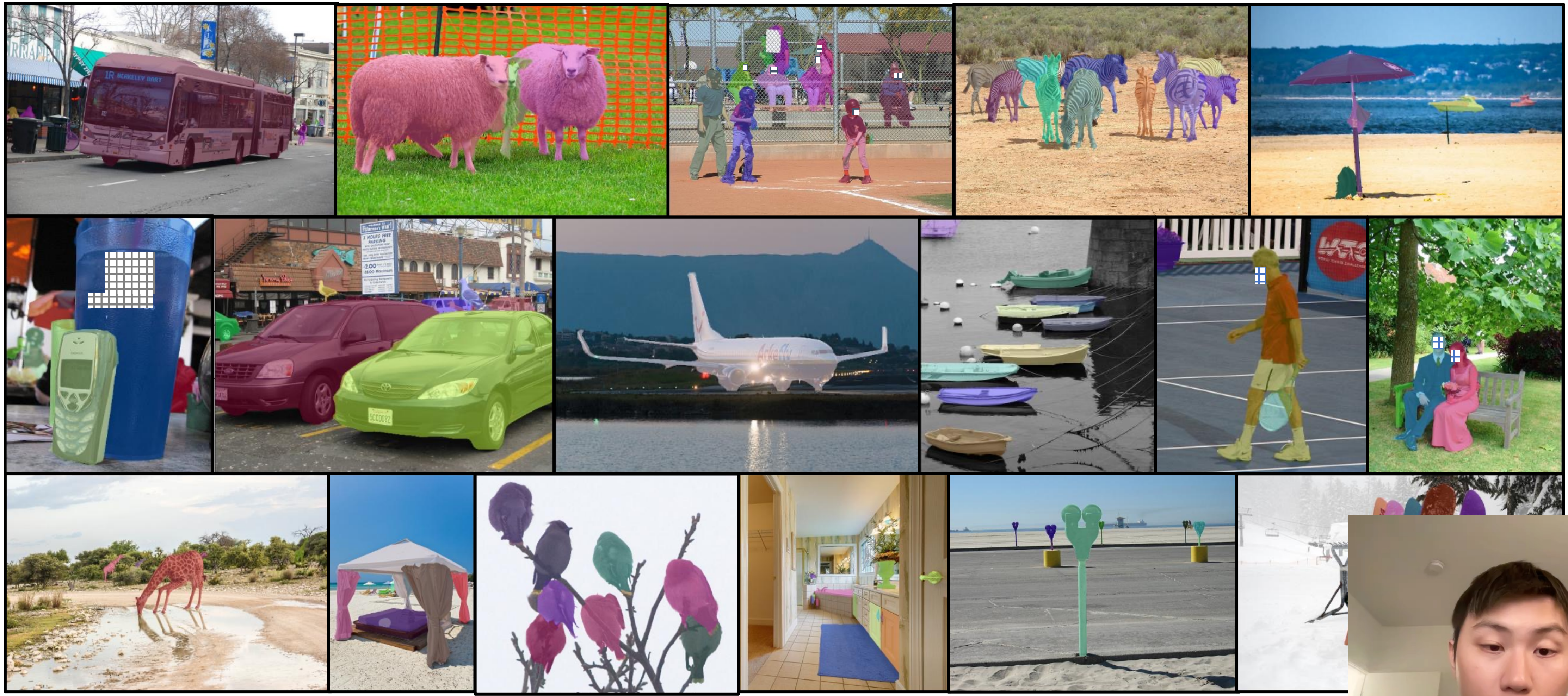
MAL Pipeline



MAL Pipeline



Qualitative Results



Quantitative Results

Method	Labeler Backbone	InstSeg Backbone	InstSeg Model	Sup	(%)Mask AP_{val}	(%)Mask AP_{test}	(%)$Ret._{val}$	(%)$Ret._{test}$
Mask R-CNN* [24]	-	ResNet-101	Mask R-CNN	Mask	38.6	38.8	-	-
Mask R-CNN* [24]	-	ResNeXt-101	Mask R-CNN	Mask	39.5	39.9	-	-
CondInst [33]	-	ResNet-101	CondInst	Mask	38.6	39.1	-	-
SOLOv2 [31]	-	ResNet-50	SOLOv2	Mask	37.5	38.4	-	-
SOLOv2 [31]	-	ResNet-101-DCN	SOLOv2	Mask	41.7	41.8	-	-
SOLOv2 [31]	-	ResNeXt-101-DCN	SOLOv2	Mask	42.4	42.7	-	-
ConvNeXt [44]	-	ConvNeXt-Small [44]	Cascade R-CNN	Mask	44.8	45.5	-	-
ConvNeXt [44]	-	ConvNeXt-Base [44]	Cascade R-CNN	Mask	45.4	46.1	-	-
Mask2Former [41]	-	Swin-Small	Mask2Former	Mask	**46.1**	**47.0**	-	-
BBTP† [4]	-	ResNet-101	Mask R-CNN	Box	-	21.1	-	59.1
BoxInst [5]	-	ResNet-101	CondInst	Box	33.0	33.2	85.5	84.9
BoxLevelSet [6]	-	ResNet-101-DCN	SOLOv2	Box	35.0	35.4	83.9	83.5
DiscoBox [7]	-	ResNet-50	SOLOv2	Box	30.7	32.0	81.9	83.3
DiscoBox [7]	-	ResNet-101-DCN	SOLOv2	Box	35.3	35.8	84.7	85.9
DiscoBox [7]	-	ResNeXt-101-DCN	SOLOv2	Box	37.3	37.9	88.0	88.8
BoxTeacher [8]	-	Swin-Base	CondInst	Box	-	40.0	-	-
Mask Auto-Labeler	ViT-MAE-Base [13]	ResNet-50	SOLOv2	Box	35.0	35.7	93.3	93.0
Mask Auto-Labeler	ViT-MAE-Base [13]	ResNet-101-DCN	SOLOv2	Box	38.2	38.7	91.6	92.6
Mask Auto-Labeler	ViT-MAE-Base [13]	ResNeXt-101-DCN	SOLOv2	Box	38.9	39.1	91.7	91.6
Mask Auto-Labeler	ViT-MAE-Base [13]	ConvNeXt-Small [44]	Cascade R-CNN	Box	42.3	43.0	94.4	**94.5**
Mask Auto-Labeler	ViT-MAE-Base [13]	ConvNeXt-Base [44]	Cascade R-CNN	Box	42.9	43.3	**94.5**	93.9
Mask Auto-Labeler	ViT-MAE-Base [13]	Swin-Small [12]	Mask2Former [41]	Box	**43.3**	**44.1**	93.9	93.8

Table 1. Main results on COCO. Ret means the retention rate of $\frac{\text{box-supervised mask AP}}{\text{supervised mask AP}}$. MAL with SOLOv2/ResNeXt-101 outperforms DiscoBox with SOLOv2/ResNeXt-101 by 1.6% on val2017 and 1.3% on test-dev. Our best model (Mask2former/Swin-Small) achieves 43.3% AP on val and 44.1% AP on test-dev.

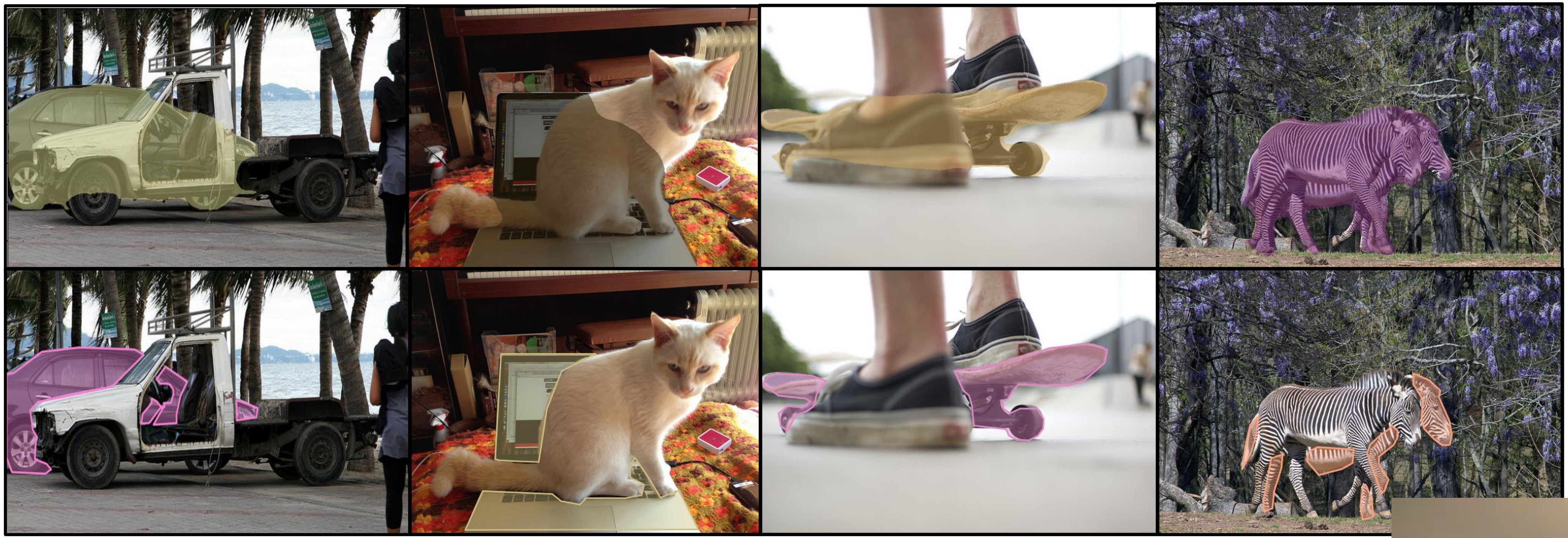
Quantitative Results

Method	Autolabeler Backbone	InstSeg Backbone	InstSeg Model	Training Data	Sup	(%)Mask AP_{val}	(%)Ret._{val}
Mask R-CNN [24]	-	ResNet-50-DCN	Mask R-CNN [24]	-	Mask	21.7	-
Mask R-CNN [24]	-	ResNet-101-DCN	Mask R-CNN [24]	-	Mask	23.6	-
Mask R-CNN [24]	-	ResNeXt-101-32x4d-FPN	Mask R-CNN [24]	-	Mask	25.5	-
Mask R-CNN [24]	-	ResNeXt-101-64x4d-FPN	Mask R-CNN [24]	-	Mask	25.8	-
Mask Auto-Labeler	ViT-MAE-Base [13]	ResNet-50-DCN	Mask R-CNN [24]	LVIS v1	Box	20.7	95.4
Mask Auto-Labeler	ViT-MAE-Base [13]	ResNet-101-DCN	Mask R-CNN [24]	LVIS v1	Box	23.0	**97.4**
Mask Auto-Labeler	ViT-MAE-Base [13]	ResNeXt-101-32x4d-FPN	Mask R-CNN [24]	LVIS v1	Box	23.7	92.9
Mask Auto-Labeler	ViT-MAE-Base [13]	ResNeXt-101-64x4d-FPN	Mask R-CNN [24]	LVIS v1	Box	**24.5**	95.0
Mask Auto-Labeler	ViT-MAE-Base [13]	ResNeXt-101-32x4d-FPN	Mask R-CNN [24]	COCO	Box	23.3	91.8
Mask Auto-Labeler	ViT-MAE-Base [13]	ResNeXt-101-64x4d-FPN	Mask R-CNN [24]	COCO	Box	**24.2**	**93.8**

Table 2. Main results on LVIS v1. Training data means the dataset we use for training MAL. We also finetune it on COCO and then generate pseudo-labels of LVIS v1. Compared with trained on LVIS v1 directly, MAL finetuned on COCO only caused around 0.35% mAP drop on the final results, which indicates the great potential of the open-set ability of MAL. Ret means the retention rate of $\frac{\text{box-supervised mask AP}}{\text{supervised mask AP}}$.

MAL Labels (top) v.s. Human Labels (bottom)

MAL Labels v.s. Human Labels



Improve Detection

InstSeg Backbone	Dataset	Mask Labels	(%)AP	(%)AP_{50}	(%)AP_{75}	(%)AP_S	(%)AP_M	(%)AP_L
ResNet-50-DCN [59]	LVIS v1	None	22.0	36.4	22.9	16.8	29.1	33.4
ResNet-50-DCN [59]	LVIS v1	GT mask	22.5	36.9	23.8	16.8	29.7	35.0
ResNet-50-DCN [59]	LVIS v1	MAL mask	22.6	37.2	23.8	17.3	29.8	34.6
ResNet-101-DCN [59]	LVIS v1	None	24.4	39.5	26.1	17.9	32.2	36.7
ResNet-101-DCN [59]	LVIS v1	GT mask	24.6	39.7	26.1	18.3	32.1	38.3
ResNet-101-DCN [59]	LVIS v1	MAL mask	25.1	40.0	26.7	18.4	32.5	37.8
ResNeXt-101-32x4d-FPN [53, 59]	LVIS v1	None	25.5	41.0	27.1	18.8	33.7	38.0
ResNeXt-101-32x4d-FPN [53, 59]	LVIS v1	GT mask	26.7	42.1	28.6	19.7	34.7	39.4
ResNeXt-101-32x4d-FPN [53, 59]	LVIS v1	MAL mask	26.3	41.5	28.3	19.5	34.5	39.6
ResNeXt-101-64x4d-FPN [53, 59]	LVIS v1	None	26.6	42.0	28.3	19.8	34.7	39.9
ResNeXt-101-64x4d-FPN [53, 59]	LVIS v1	GT mask	27.2	42.8	29.2	20.2	35.7	41.0
ResNeXt-101-64x4d-FPN [53, 59]	LVIS v1	MAL mask	27.2	42.7	29.1	19.8	35.9	40.7
ConvNeXt-Small [44]	COCO	None	51.5	70.6	56.1	34.8	55.2	66.9
ConvNeXt-Small [44]	COCO	GT mask	51.8	70.6	56.3	34.5	55.9	66.6
ConvNeXt-Small [44]	COCO	MAL mask	51.7	70.5	56.2	35.2	55.7	66.8

Table 7. Results of detection by adding different mask supervision. The models are evaluated on COCO val2017 and LVIS v1 mask supervision using ground-truth masks or mask pseudo-labels, we can get around 1% improvement on different AP metr v1. On COCO val2017, the detection performance also benefits from mask pseudo-labels. Although the improvement is less th the improvement is consistent over different random seeds.

Thanks!