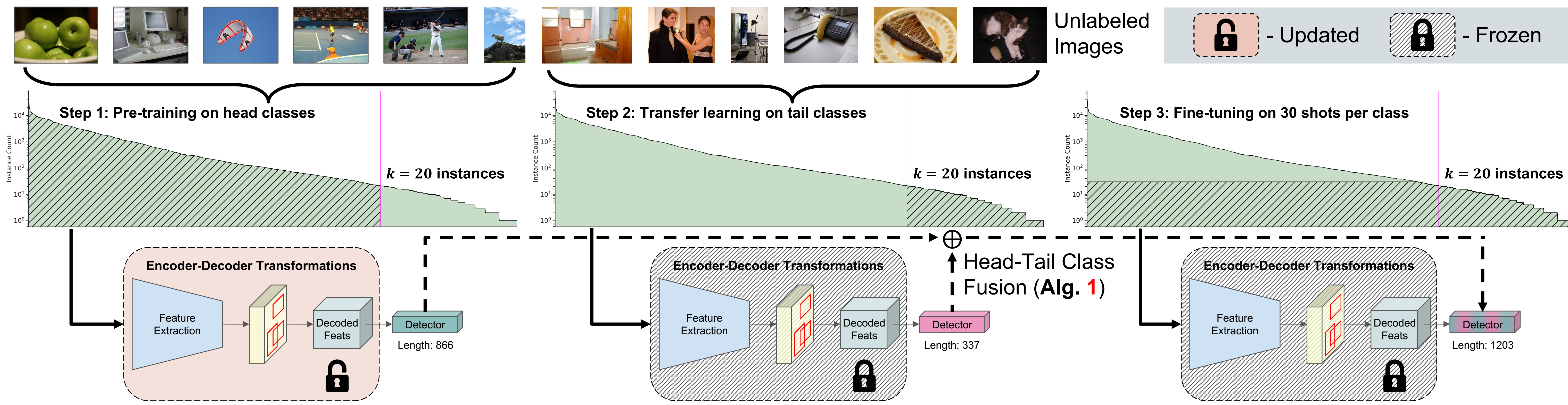


SimLTD: Simple Supervised and Semi-Supervised Long-Tailed
Object Detection

Phi Vu Tran

The SimLTD Framework: Improved Head-to-Tail Model Transfer for Long-Tailed Detection

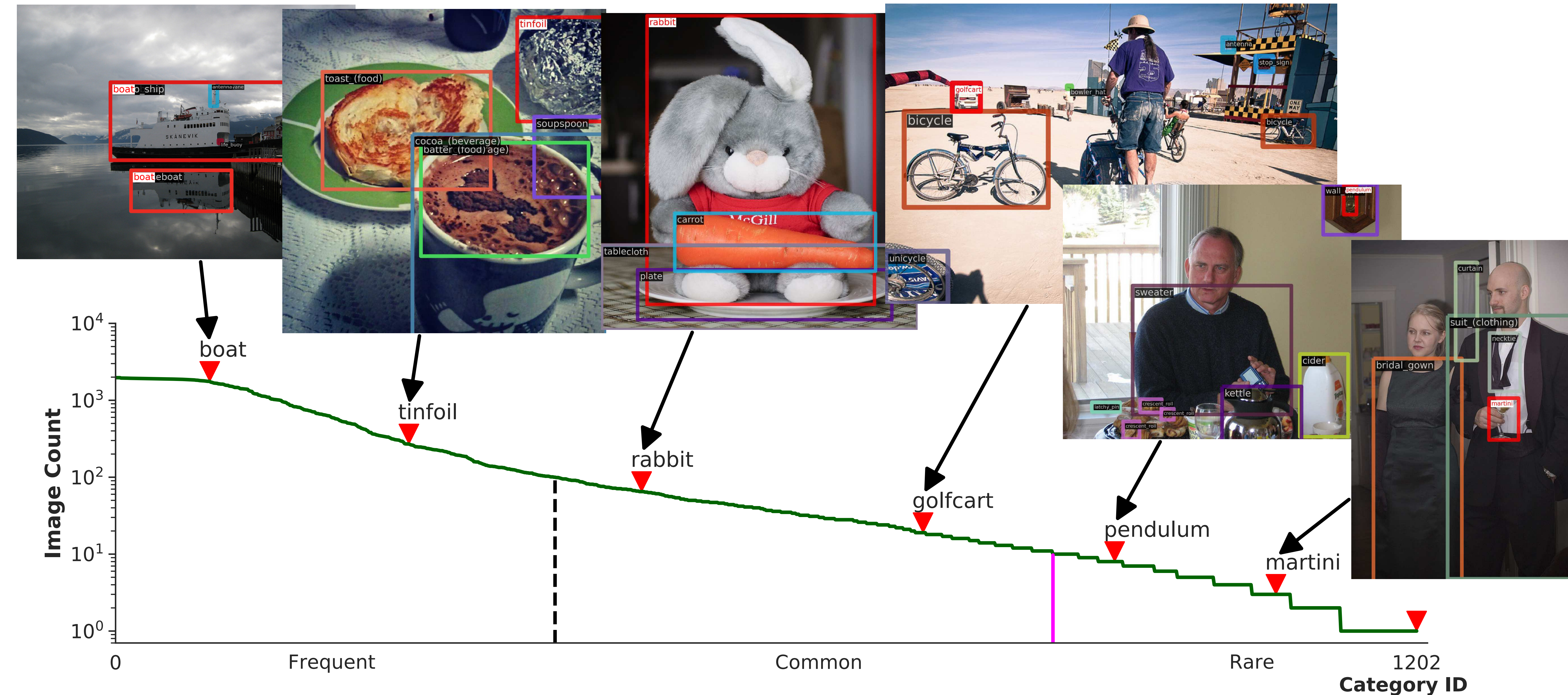


Overview of SimLTD: We partition the long-tailed distribution into two disjoint sets—*frequent + common head* categories and *rare tail* classes—and perform supervised and semi-supervised training in three easy steps:

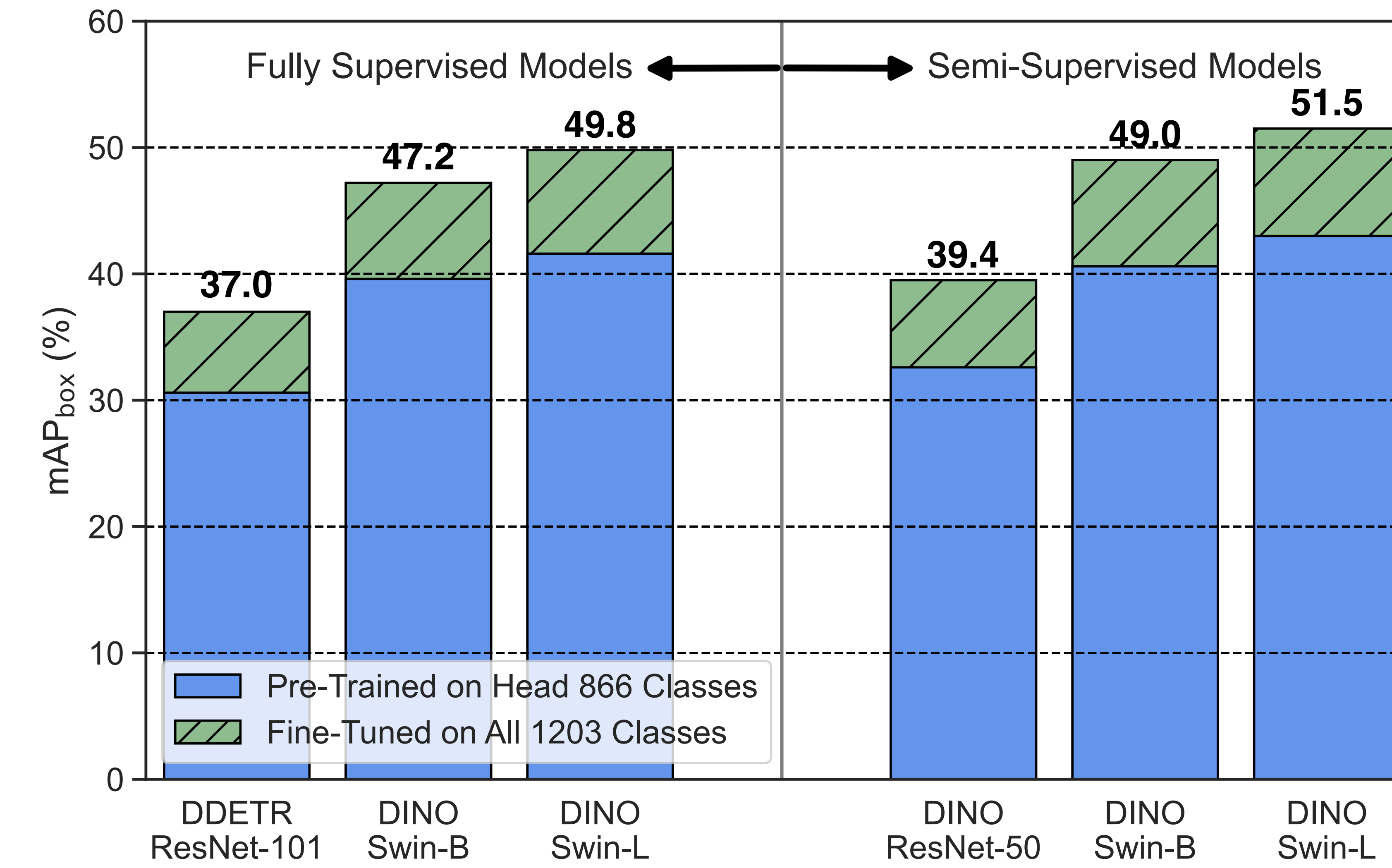
- 1 Pre-training on head classes (hatched region) with optional unlabeled images;
- 2 Transfer learning on tail classes (hatched region) with optional unlabeled images;
- 3 Fine-tuning on a balanced sample of head and tail classes for exemplar replay.

Robust Long-Tailed Detection Without Additional Image-Level Supervision

- ✗ Existing models require extra ImageNet labels, which are infeasible for bespoke applications outside of LVIS.
- ✓ Our SimLTD leverages unlabeled images, which are easy to collect without the burden of human annotations.

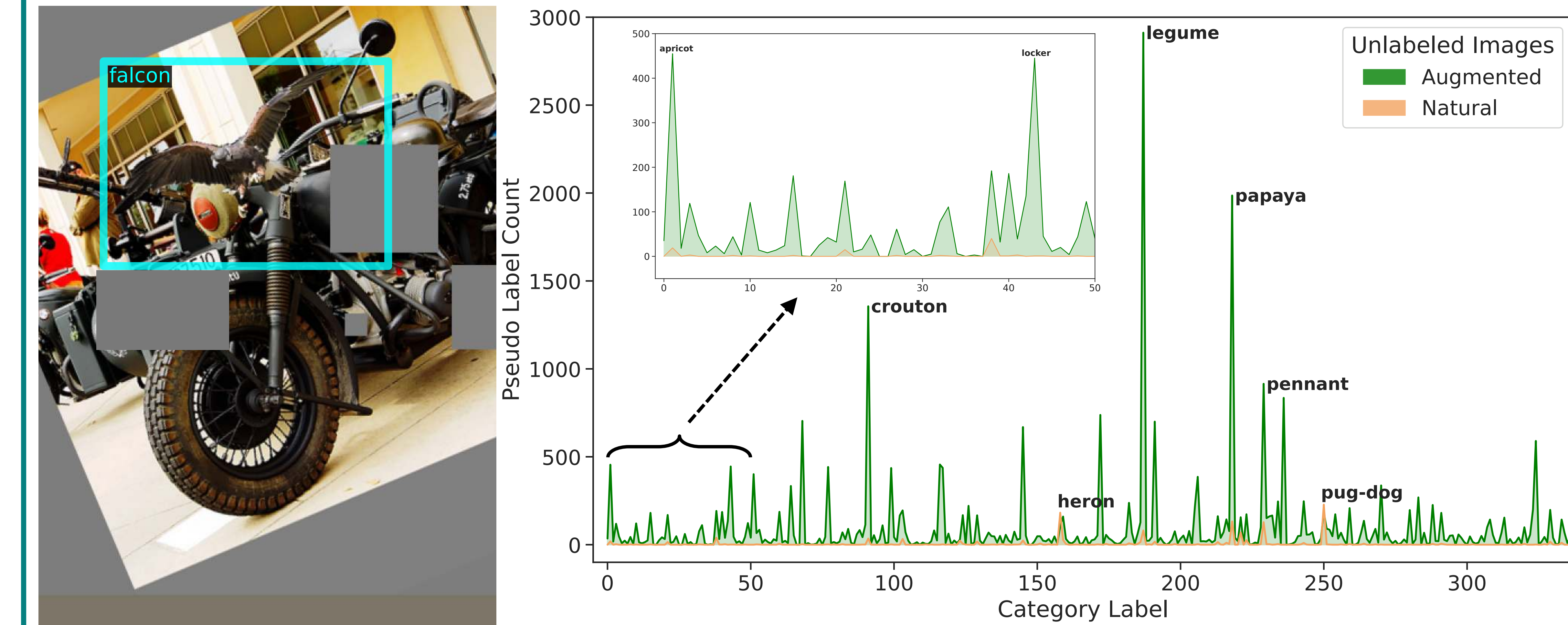


What Makes for Effective Long-Tailed Detection?

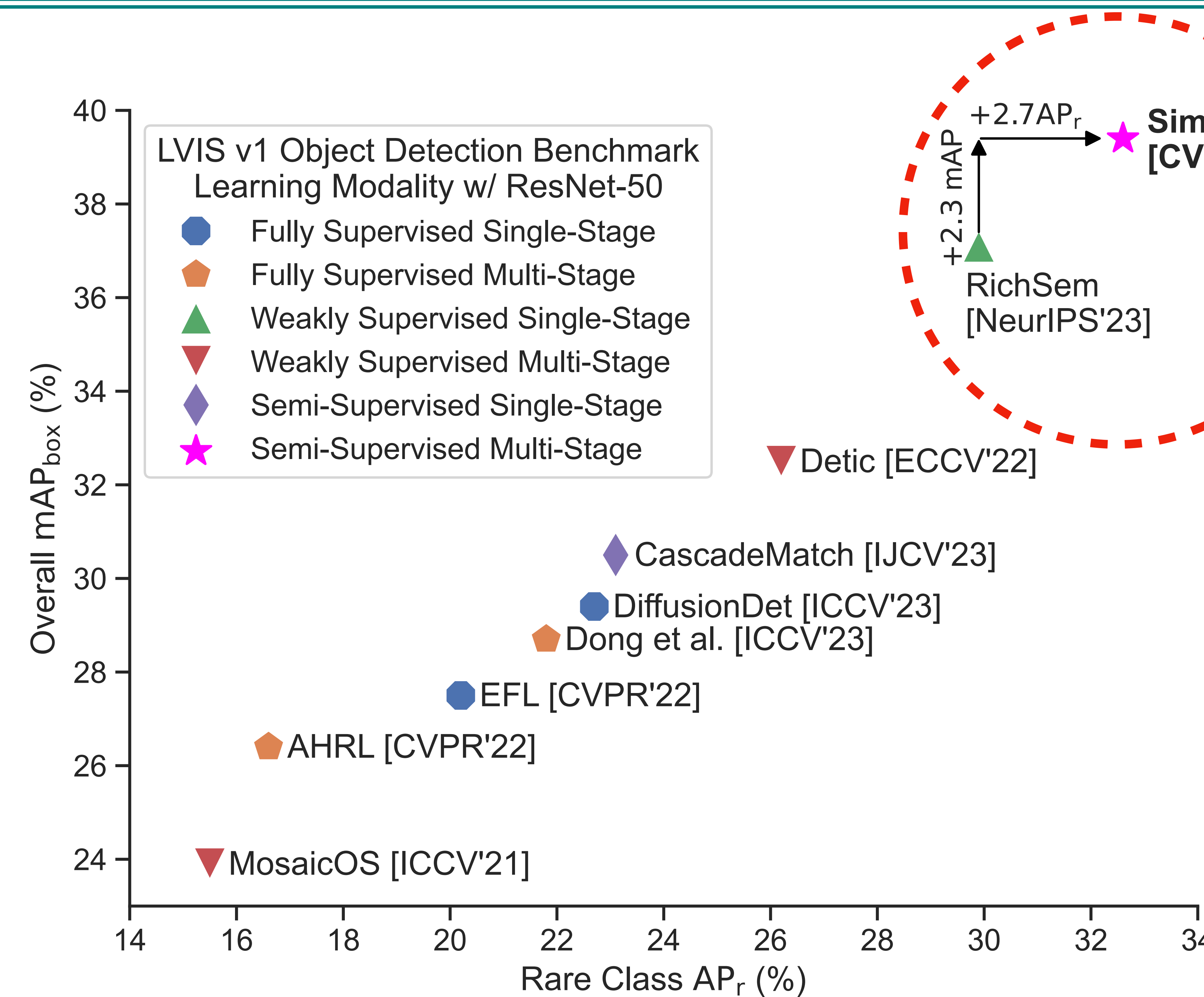


Our contribution: we find that stronger pre-trained head models (solid blue bars) are more powerful long-tailed detectors (hatched green bars) across both supervised and semi-supervised settings.

Learning to Detect Genuinely Rare Objects by Pasting Onto Unlabeled Images



Advancing Long-Tailed Detection on the LVIS v1 Benchmark — Compelling Performance and Scalability Across Backbones and Detectors



Method	Extra External Data	Base Detector	Backbone	GPU Hrs	mAP _{box}	AP _r	AP _c	AP _f
Seesaw Loss [CVPR'21] [36]		Faster R-CNN	R101-FPN	—	27.8	18.7	27.0	32.8
NorCal [NeurIPS'21] [28]		Mask R-CNN	R101-FPN	—	28.1	20.8	26.5	30.9
AHRL [CVPR'22] [20]		Mask R-CNN	R101-FPN	—	28.7	19.3	27.6	31.4
EFL [CVPR'22] [21]		RetinaNet	R101-FPN	—	29.2	23.5	27.4	33.8
SimLTD Supervised		Faster R-CNN	R101-FPN	83	31.7	24.3	32.1	34.6
SimLTD Supervised		Deformable DETR	R101	234	37.0	31.9	36.6	39.6
Dong et al. [ICCV'23] [12]		Deformable DETR	R50	—	28.7	21.8	28.4	32.0
Detic [ECCV'22] [49]		Deformable DETR	R50	—	31.7	21.4	30.7	37.5
RichSem [NeurIPS'23] [27]		DINO	R50	—	35.1	26.0	32.6	41.8
SimLTD Supervised		Faster R-CNN	R50-FPN	81	29.0	20.9	29.7	31.9
SimLTD Supervised		Deformable DETR	R50	215	35.0	32.0	34.0	37.5
RichSem [NeurIPS'23] [27]		DINO	Swin-T	—	38.8	30.8	36.4	45.0
SimLTD Supervised		DINO	Swin-T	310	41.1	33.6	40.1	45.4
DiffusionDet [ICCV'23] [6]		4 @ 300	Swin-B	—	42.0	34.8	40.9	46.4
RichSem [NeurIPS'23] [27]		DINO	Swin-B	—	46.4	38.5	45.1	51.3
SimLTD Supervised		DINO	Swin-B	414	47.2	42.7	46.7	49.9
RichSem [NeurIPS'23] [27]		DINO	Swin-L	—	49.7	42.8	49.2	53.4
SimLTD Supervised		DINO	Swin-L	460	49.8	45.4	50.4	52.4
MosaicOS [ICCV'21] [45]	ImageNet-1K Labels	Faster R-CNN	R50-FPN	—	23.9	15.5	22.4	29.3
RichSem [NeurIPS'23] [27]	ImageNet-21K Labels	Faster R-CNN + CLIP	R50-FPN	—	30.6	27.6	29.7	32.9
SimLTD Softer Teacher [35]	COCO-unlabeled2017	Faster R-CNN	R50-FPN	392	30.3	23.3	30.3	33.3
SimLTD MixTeacher [25]	COCO-unlabeled2017	Faster R-CNN	R50-FPN	434	31.8	23.4	32.1	35.1
Detic [ECCV'22] [49]	ImageNet-21K Labels	Deformable DETR + CLIP	R50	—	32.5	26.2	31.3	36.6
RichSem [NeurIPS'23] [27]	ImageNet-21K Labels	DINO + CLIP	R50	—	37.1	29.9	35.6	42.0
SimLTD MixPL [7]	Objects365-unlabeled	DINO	R50	447	39.1	31.5	38.5	43.1
SimLTD MixPL [7]	COCO-unlabeled2017	DINO	R50	446	39.4	32.6	38.5	43.6
RichSem [NeurIPS'23] [27]	ImageNet-21K Labels	DINO + CLIP	Swin-T	—	41.6	37.3	39.7	45.5
SimLTD MixPL [7]	COCO-unlabeled2017	DINO	Swin-T	482	42.5	35.7	42.4	45.6
RichSem [NeurIPS'23] [27]	ImageNet-21K Labels	DINO + CLIP	Swin-B	—	48.2	46.5	46.5	51.0
SimLTD MixPL [7]	COCO-unlabeled2017	DINO	Swin-B	794	49.0	43.4	49.0	51.5
RichSem [NeurIPS'23] [27]	ImageNet-21K Labels	DINO + CLIP	Swin-L	—	52.0	50.2	51.5	53.3
SimLTD MixPL [7]	COCO-unlabeled2017	DINO	Swin-L	1168	51.5	45.0	52.4	53.3