



Scaling Novel Object Detection with Weakly Supervised Detection Transformers

Tyler LaBonte^{1,2}, Yale Song¹, Xin Wang¹, Vibhav Vineet¹, Neel Joshi¹

¹Microsoft Research, ²Georgia Tech



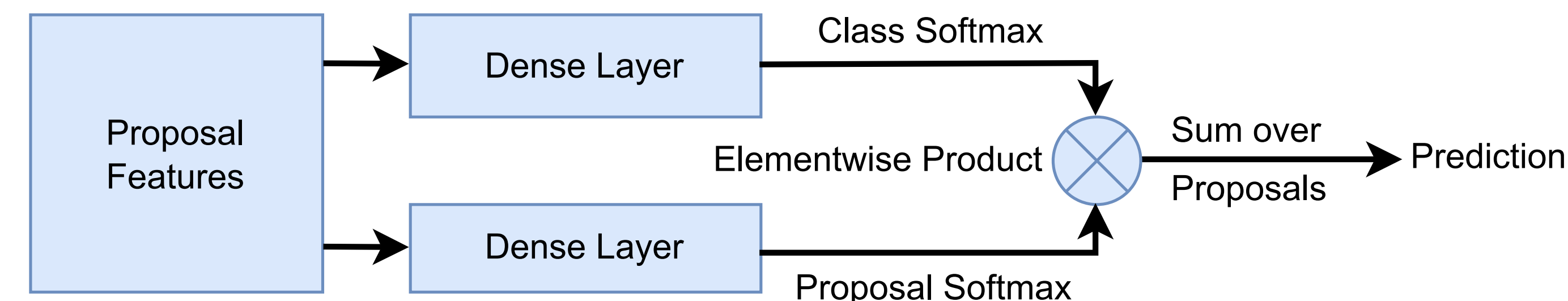
Effortless Detection of Novel Objects

- Object detection annotation is expensive and difficult
- **Weakly supervised object detection** (WSOD) uses only class labels
- Pretrain on fully-annotated source dataset, then transfer knowledge to target dataset of novel objects with image-level class labels
- **Goal: effortless detection of novel objects without labeling**



Previous Work

- Primary WSOD paradigm is **multiple instance learning** (MIL)
 - Aggregates class features over dataset to localize objects
 - 2016: Bilen and Vedaldi introduced deep learning framework for MIL [1]



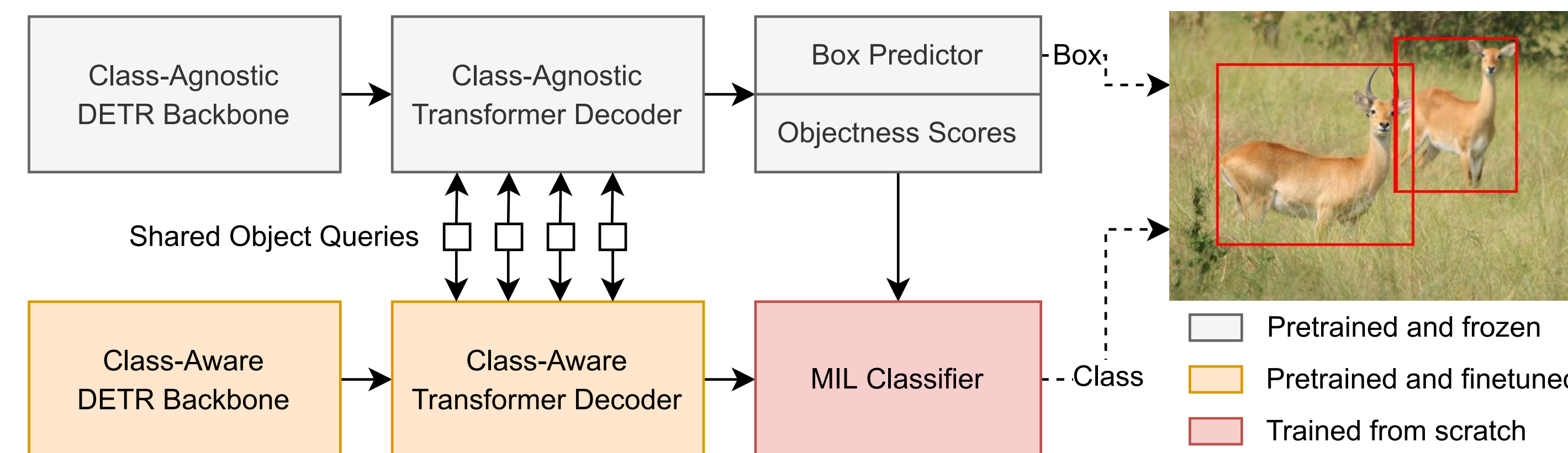
- Current approaches are **not scalable**
 - Require multiple rounds of training and refinement [3, 5, 8]
 - Optimized for datasets with only 60 pretraining and 20 novel classes

Our Contributions

- We propose Weakly Supervised Detection Transformer which **scales to 1000s of novel classes with a single pretrain-finetune step**
- We introduce new large-scale experimental setups for WSOD and call for the community to move beyond toy datasets to complex settings
- We also identify and rectify a weakness of a standard regularization method and explore sparsity for proposal noise reduction

Weakly Supervised Detection Transformer

- Combines proposal generation of two-stage CNN model with scalability and simplicity of one-stage Transformer approach



Large-Scale Novel Object Detection Results

- We utilize FSOD dataset [4] with 800 pretraining and 200 novel classes (175K boxes) constructed from ILSVRC and Open Images
 - **Classes maximally separated** w.r.t. semantic hierarchy
- 4X the pretraining classes and 2X the novel classes than shown previously

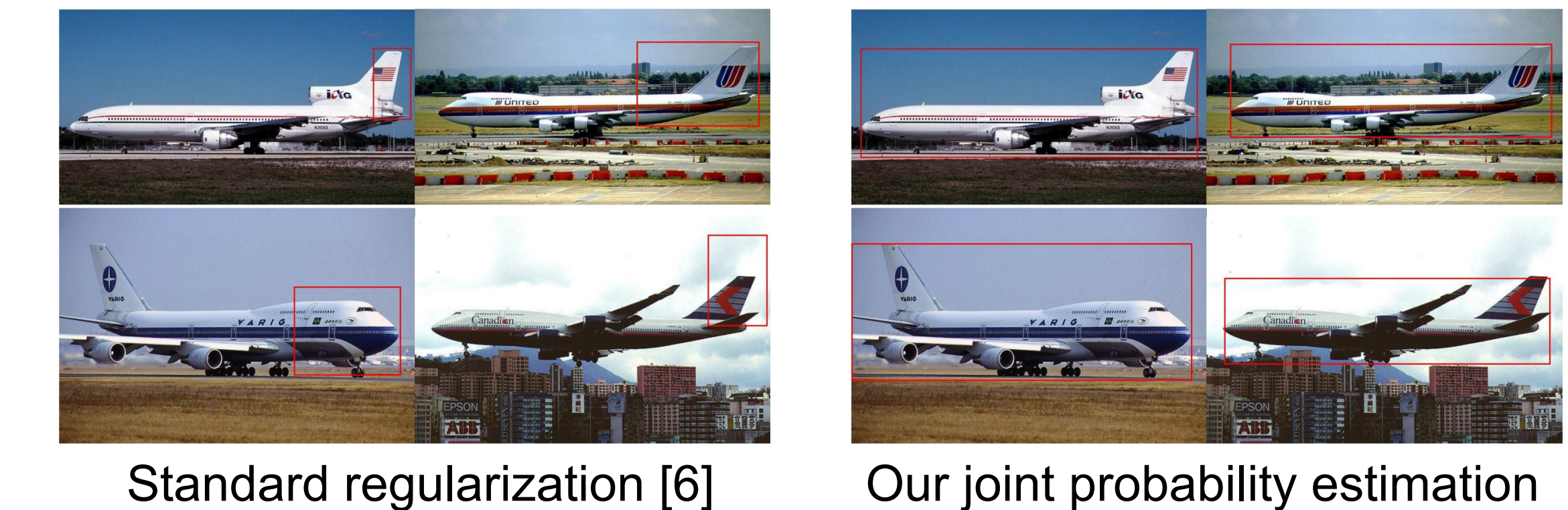
Method	mAP	AP50	mAR
Zhong <i>et al.</i> [8]	20.6	32.7	34.4
WS-DETR Base	13.9	20.0	60.1
WS-DETR Sparse	28.5	38.5	68.0
WS-DETR Joint	28.6	37.8	65.3
WS-DETR Full	28.6	38.2	67.4
Supervised DETR [2]	47.7	64.0	76.3

- We study iNaturalist dataset [7] for species detection (560K boxes)
 - 13 superclasses and **2,854 subclasses—5X that of Open Images**
 - Less “pure” of a novel classes setting than FSOD, but realistic

Method	13 Superclasses		2,854 Subclasses	
	mAP	AP50	mAP	AP50
Zhong <i>et al.</i> [8]	44.1	76.7	-	-
WS-DETR Base	0.2	0.4	1.7	3.7
WS-DETR Sparse	61.1	79.3	30.4	38.2
WS-DETR Joint	54.8	70.0	22.1	29.8
WS-DETR Full	60.7	78.7	35.4	43.5
Supervised DETR [2]	79.2	93.6	51.5	58.8

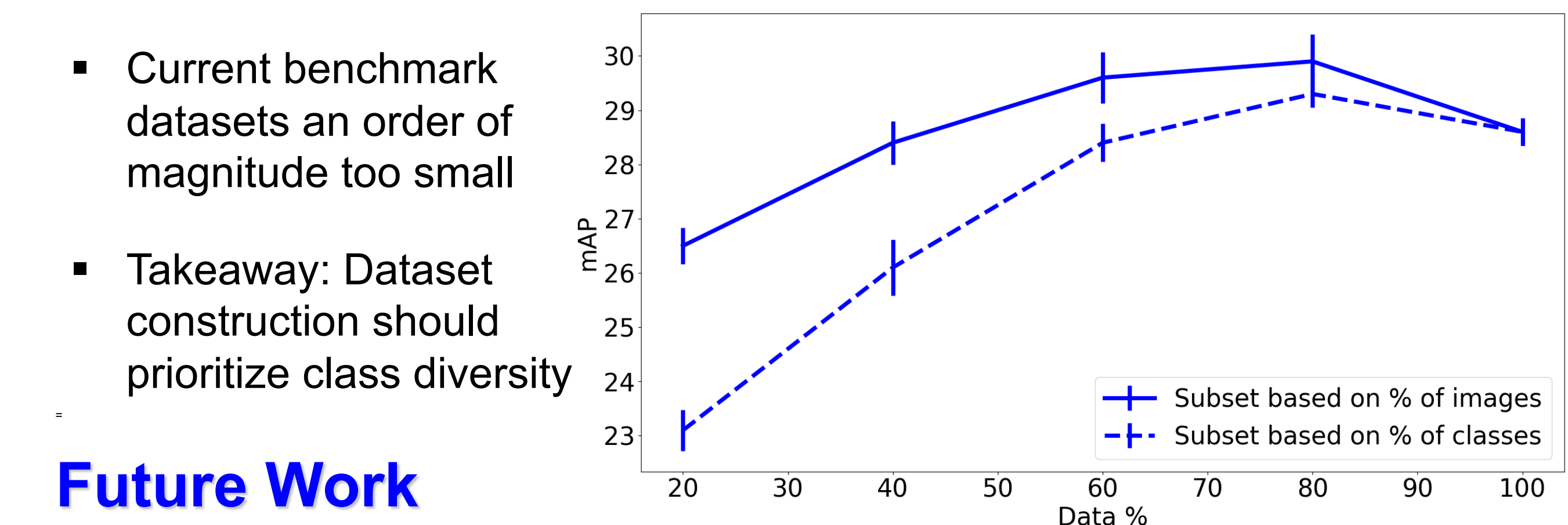
Joint Probability Estimation

- We show a standard regularization technique [8] **overfits to classification features** [5] and we introduce a solution
- Uses pretrained “objectness” for joint probability of object and class



Scaling Study

- Can’t pretrain two DETRs? No problem! Using the class-agnostic DETR in classification stream only drops performance by 1.6 mAP
- First large-scale rigorous study of WSOD scaling: **class quantity is more important than image quantity** [4, 6]



Future Work

- Use noisy Web searches and generated image captions as labels
- Use Transformer attention to refine box predictions [3]
- Integrate self-supervised detection-aware pretraining [9]

References

- [1] Bilen and Vedaldi. Weakly Supervised Deep Detection Networks. *CVPR*, 2016.
- [2] Carion *et al.* End-to-End Object Detection with Transformers. *ECCV*, 2020.
- [3] Huang *et al.* Comprehensive Attention Self-Distillation for WSOD. *NeurIPS*, 2020.
- [4] Fan *et al.* Few-Shot Object Detection with Attention-RPN and Multi-Relation Detector. *CVPR*, 2020.
- [5] Tang *et al.* Multiple Instance Detection Network with Online Instance Classifier Refinement. *CVPR*, 2017.
- [6] Uijlings *et al.* Revisiting Knowledge Transfer for Training Object Class Detectors. *CVPR*, 2018.
- [7] Van Horn *et al.* The iNaturalist Species Classification and Detection Dataset. *CVPR*, 2018.
- [8] Zhong *et al.* Boosting WSOD with Progressive Knowledge Transfer. *ECCV*, 2020.
- [9] Zhong *et al.* DAP: Detection-Aware Pre-training with Weak Supervision. *CVPR*, 2021.