
SOTA WORKS
Object Detection on COCO
Nov 2022

OBJECT DETECTION ON COCO (TOP1 – TOP4)

AILab/Tsinghua/NJU /CUHK/Sense Time	BAAI/HuazhongU /ZJU/BIT	Baidu/ANU/BHU/PKU	MSFT
InternImage-DCNv3-H	EVA	Group DETR v2	FocalNet-H (DINO)
65.4 mAP on COCO	64.7 mAP on COCO	64.5 mAP on COCO	64.4 mAP on COCO
Deformable CNN : Mask R - CNN base plus deformable kernal	Masked Visual Representation Learning at Scale	ViT-Huge + DINO + Group DETR training method	Hierarchical contextualization + Gated aggregation + Affine transformation
Paper tables with annotated results for InternImage: Exploring Large-Scale Vision Foundation Models with Deformable Convolutions | Papers With Code	EVA: Exploring the Limits of Masked Visual Representation Learning at Scale | Papers With Code	Group DETR v2: Strong Object Detector with Encoder-Decoder Pretraining | Papers With Code	Focal Modulation Networks | Papers With Code

OBJECT DETECTION ON COCO (TOP5 – TOP8)

Tsinghua/MSRT Asia	MSFT	HKUST/Tsinghua/IDEA	MSFT Asia
FD-SwinV2-G	BEiT-3	DINO	SwinV2-G
64.2 mAP on COCO	63.7 mAP on COCO	63.3 mAP on COCO	63.1 mAP on COCO
Swin Transformer + Feature Distillation + Masked image modeling	Multiway transformers + masked "language" modeling	DETR with Improved deNoising anchOr boxes	residual-post-norm + log-spaced position bias + SimMIM
Contrastive Learning Rivals Masked Image Modeling in Fine-tuning via Feature Distillation | Papers With Code	Image as a Foreign Language: BEiT Pretraining for All Vision and Vision-Language Tasks | Papers With Code	DINO: DETR with Improved DeNoising Anchor Boxes for End-to-End Object Detection | Papers With Code	Swin Transformer V2: Scaling Up Capacity and Resolution | Papers With Code

OBJECT DETECTION ON COCO (TOP9 – TOP12)

MSFT	UW/Meta/MSFT/UCLA	UCLA/MSFT/UW/UWM	HZUST/MSFT
Florence-CoSwin-H	GLIPv2	GLIP	Soft Teacher + Swin-L
62.4 mAP on COCO	62.4 mAP on COCO	61.5 mAP on COCO	61.3 mAP on COCO
data curation + model pretraining + task adaptations + training infrascturue	a grounded VL understanding model (Localization + VL understanding)	Swin-L, multi-scale	HTC++, multi-scale
Florence: A New Foundation Model for Computer Vision | Papers With Code	GLIPv2: Unifying Localization and Vision-Language Understanding | Papers With Code	Grounded Language-Image Pre-training | Papers With Code	End-to-End Semi-Supervised Object Detection with Soft Teacher | Papers With Code

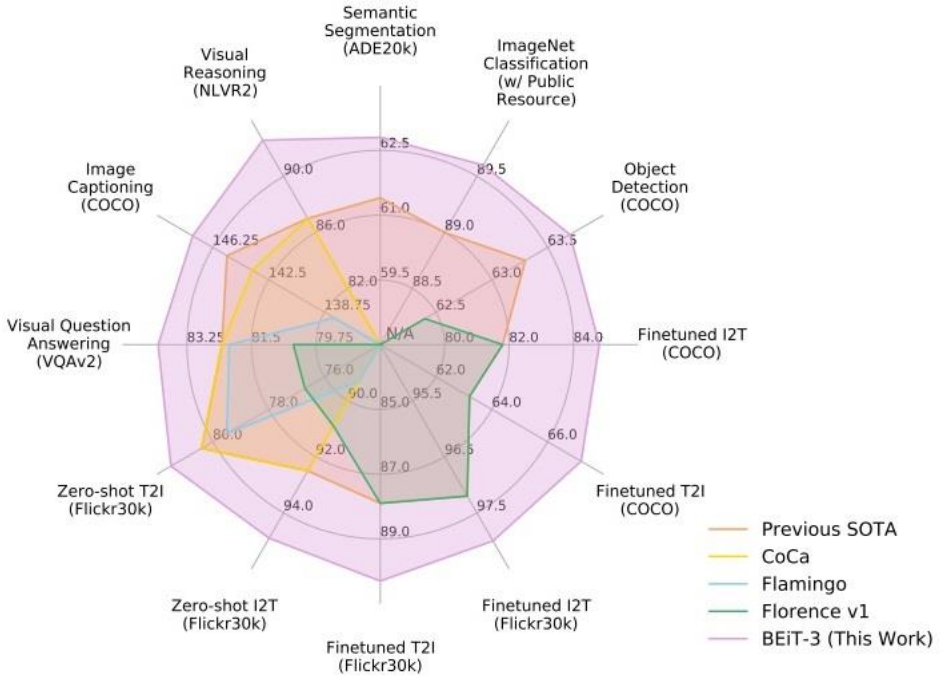
COMPARISON AND TAKEAWAYS

Backbone + Pre-training + Scaling-up

ViT huge
Swin Tran
Muti-way Tran

Multi-reception
Heuristic
Training strategy

Data-sets
Different
application tasks



THANK YOU