Supplementary Material

We start by providing the full implementation details of DETReg and include the complete PASCAL VOC results. We then follow with additional analysis of DETReg pretraining as well as class agnostic performance and visualization.

Implementation Details. Based on the ablations presented in Section 4.5, the default experiment settings are as follows. For region proposals, we compute Selective Search boxes online using the "fast" preset of the OpenCV implementation [4] and unless otherwise noted, we use the DETReg Top-K region selection variant (see Section 3.1) and set $K = 30$ proposals per-image. We initialize the ResNet50 backbone of DETReg with SwAV [6], which was pretrained with multi-crop views for 800 epochs on IN1K, and fix it throughout the pretraining stage. A similar SwAV encoder is used to encode region proposals, which are first cropped and resized to $128x128$. In the object embedding branch, f_{emb} and f_{box} are MLPs with 2 hidden layers of size 256 followed by a ReLU [44] nonlinearity. The output sizes of f_{emb} and f_{box} are 512 and 4. f_{cat} is implemented as a single fully-connected layer with 2 outputs. We run the pretraining experiments using a batch size of 24 per GPU on an NVIDIA DGX, V100 x8 GPUs machine, following the hyperparameter settings and image augmentations from existing works [5, 71]. Similarly, cropped regions are augmented before being fed to the encoder to obtain embeddings $\boldsymbol{z}_i$. When finetuning, we drop the f_{emb} branch, and set the size of the last fully-connected layer of f_{cat} to be the number of classes in the target dataset plus a background class.

Object Detection in Full Data Regimes

We reported DETReg results on the PASCAL VOC benchmark in Section 4.1. Here we include the full table, containing more past pretraining approaches using three different object detectors (see Table 8). We observe that using the Deformable-DETR detector, the supervised pretraining baseline is superior to past pretraining approaches and that DETReg pretraining improves over it by 4 points (AP).

Semi-supervised Learning

We reported DETReg results and comparisons to other pretraining approaches like [6, 62] when using limited amounts of data. In Table 9, we include comparisons to semi-supervised works [34, 42, 53, 65] that leverage both the labeled and unlabeled data in training via auxiliary losses.

DETReg Analysis

In Section 4.5 we analyzed DETReg, including the model ablations, class agnostic results, visualization and robustness. Here we further examine the pretrained DETReg model including the class agnostic results, and TopK selection policy.

Method	Detector	AP	AP_{50}	AP_{75}
Supervised	FRCN	56.1	82.6	62.7
InsDis [61]		55.2	80.9	61.2
Jigsaw [25]		48.9	75.1	52.9
NPID++ [43]		52.3	79.1	56.9
SimCLR [10]		51.5	79.4	55.6
PIRL [43]		54.0	80.7	59.7
BoWNet [22]		55.8	81.3	61.1
MoCo [28]		55.9	81.5	62.6
MoCo-v2 [13]		57.0	82.4	63.6
SwAV [6]		56.1	82.6	62.7
DenseCL [57]		58.7	82.8	65.2
DetCo [64]		58.2	82.7	65.0
ReSim [62]		59.2	82.9	65.9
Supervised	DETR	54.1	78.0	58.3
UP-DETR [16]		57.2	80.1	62.0
Supervised	DDETR	59.5	82.6	65.6
SwAV [6]		61.0	83.0	68.1
DETReg		**63.5**	**83.3**	**70.3**

Table 8. **Object detection finetuned on PASCAL VOC.** The model is finetuned on PASCAL VOC `trainval07+2012` and evaluated on `test07`. Models are based on Faster-RCNN [49] (FRCN), DETR [5], and Deformable DETR [71] (DDETR). Bold values indicate an improvement ≥ 0.3 AP.

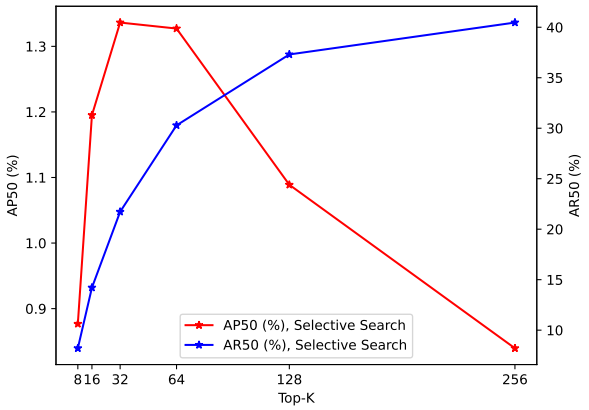

Figure 5. **Top-K proposals performance of Selective Search.** Using different values of K, w evaluate the class agnostic performance of Selective Search on MS COCO 2017 validation split.

Improved Encoder, improved DETReg. We test how DETReg performs when object embeddings are obtained with different image encoders. Specifically, we pretrain DETReg on IN100 using SwAV trained for 400 epochs compared to a superior variant trained for 800 epochs with multi-crops. We finetune on MS COCO with 1% data and observe the improved encoder achieves 1 AP improvement (27.7 vs 26.7).

DETReg TopK selection policy. Using Selective Search, we examine the class agnostic performance when using TopK

Method	Approach	Detector	COCO 1%	COCO 2%	COCO 5%	COCO 10%
CSD [34]	Auxiliary	FRCN	10.5 ± 0.1	13.9 ± 0.1	18.6 ± 0.1	22.5 ± 0.1
STAC [53]			14.0 ± 0.6	18.3 ± 0.3	24.4 ± 0.1	28.6 ± 0.2
U-T [42]			20.8 ± 0.1	24.3 ± 0.1	28.3 ± 0.1	31.5 ± 0.1
S-T [65]			$\mathbf{20.5 \pm 0.4}$	–	$\mathbf{30.7 \pm 0.1}$	$\mathbf{34.0 \pm 0.1}$
Supervised	Pretraining	DDETR	11.31 ± 0.3	15.22 ± 0.32	21.33 ± 0.2	26.34 ± 0.1
SwAV			11.79 ± 0.3	16.02 ± 0.4	22.81 ± 0.3	27.79 ± 0.2
ReSim			11.07 ± 0.4	15.26 ± 0.26	21.48 ± 0.1	26.56 ± 0.3
DETReg			$\mathbf{14.58 \pm 0.3}$	$\mathbf{18.69 \pm 0.2}$	$\mathbf{24.80 \pm 0.2}$	$\mathbf{29.12 \pm 0.2}$

Table 9. **Object detection using k% of the labeled data on COCO.** The models are trained on `train2017` using k% and then evaluated on `val2017`. Methods like [42] utilize auxiliary losses during the training stage using unlabeled data, whereas DETReg utilizes unlabeled data during the pretraining stage only.

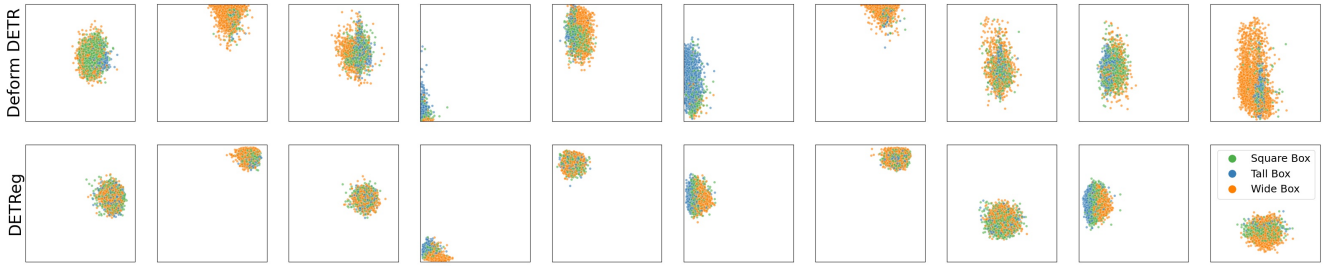

Figure 6. **DETReg slots specialize in specific areas in the image and uses a variety of box sizes much like Deformable DETR**. Each square corresponds to a DETR slot, and shows the location of its bounding box predictions. We compare 10 random slots of the supervised Deformable DETR **(top)** and unsupervised DETReg **(bottom)** decoder for the MS COCO 2017 val dataset. Each point shows the center coordinate of the predicted bounding box, where following a similar plot in [5], a green point represents a square bounding box, a orange point is a large horizontal bounding box, and a blue point is a large vertical bounding box. Deformable DETR has been trained on MS COCO 2017 data, while DETReg has only been trained on unlabeled ImageNet data. Similar DETReg and Deformable DETR slots were manually chosen for illustration.

policy. We report the precision and recall in Figure 5. In this paper, we have used $K = 30$ (see Figure 7), which emphasizes precision over recall. This might imply that DETReg performs well given high precision proposals.

DETReg Slots Visualization. We examine the learned object queries slots (see Figure 6) and observe they are similar to those in Deformable DETR, despite not using any human annotated data. Nevertheless, the Deformable DETR slots have greater variance with respect to locations and they tend to specialize more in particular boxes shapes.

Class Agnostic Object Detection. The quantitative results in Section 4.5 indicate that DETReg improves over Selective Search. The included qualitative examples of DETReg on MS COCO (see Figure 8) supports a similar conclusion, indicating that DETReg outperforms Selective Search but still much behind the ground truth labeled data.

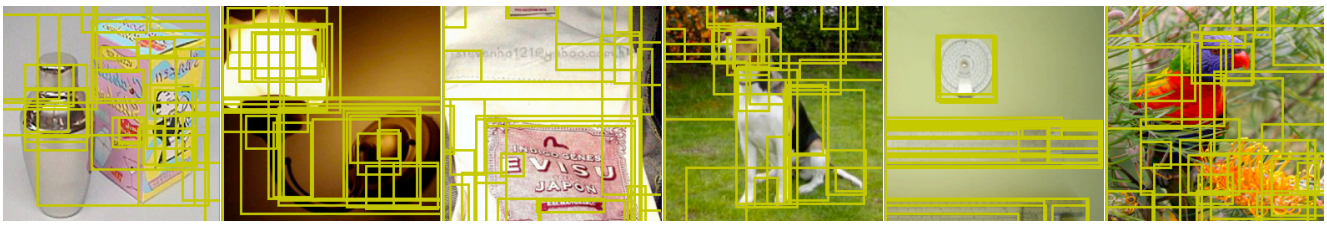

Figure 7. **TopK Selective Search proposals on ImageNet.** Using K=30, the proposals typically cover objects and parts-of-objects in the image.

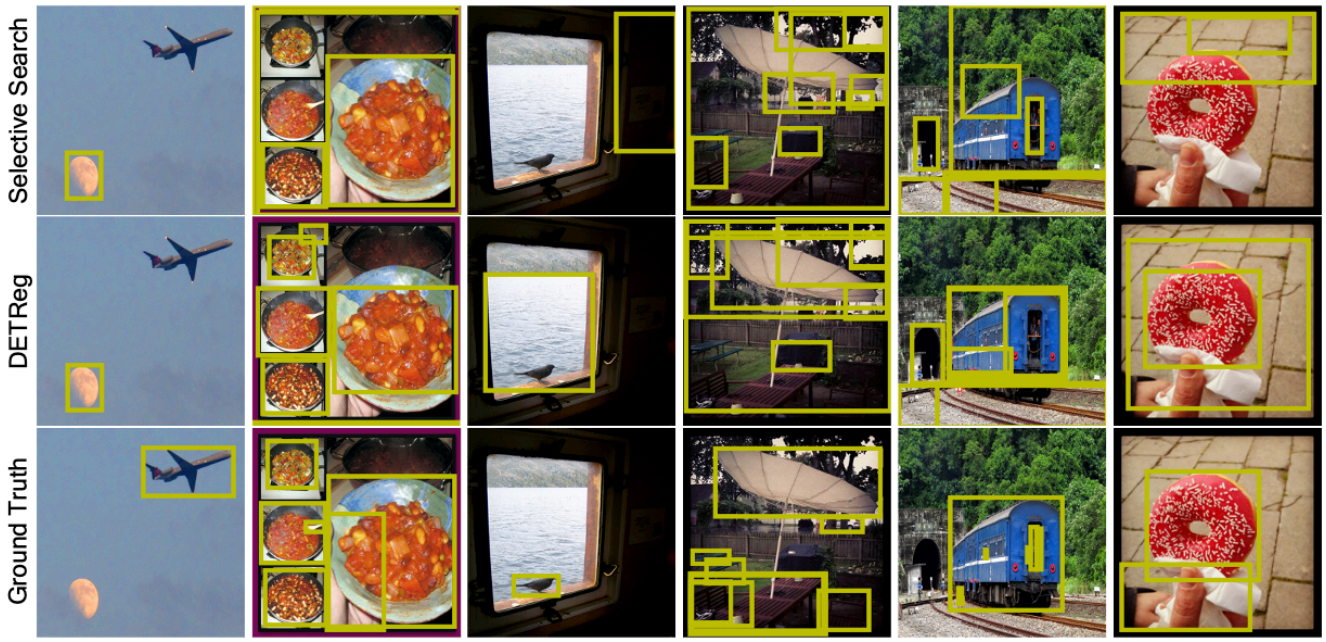

Figure 8. **Class Agnostic object detection visualization.** Examples predictions using Selective Search and DETReg on random MS COCO images. For every image annotated with M boxes, only the top M predictions are shown.