Supplementary Materials of Omni-DETR: Omni-Supervised Object Detection with Transformers

Pei Wang[2,⋆] Zhaowei Cai[1,†] Hao Yang[1] Gurumurthy Swaminathan[1]
Nuno Vasconcelos[2] Bernt Schiele[1] Stefano Soatto[1]
AWS AI Labs[1] UC San Diego[2]
{zhaoweic,haoyng,gurumurs,bschiel,soattos}@amazon.com {pew062,nuno}@ucsd.edu

A. Experimental Implementation Details

In this supplement, we show the details that are not presented in the paper due to the page limitation.

A.1. Annotation Cost Calculation in Table 11

In this section, we explain how the numbers in Table 11 of the main paper are calculated. It is non-trivial to compute the labeling time for each type of annotation because it depends on several factors like the annotation tools or platforms, the quality requirement of the annotations, the crowdsourcing protocol used, etc. In our work, we mainly follow [1, 3, 5, 7, 8, 11] for the calculation.

We denote the averaged number of categories per image as C_{avg}, the averaged number of instances per image as I_{avg}, and the overall number of categories for a dataset as C. We first list the statistics information for each dataset in Table A. Then we consider the labeling time calculation for each weak annotation.

TagsU According to [1,8], collecting image-level class labels takes $\sim$ 1 second per category per image. Thus, the expected annotation time is equal to C on COCO, VOC, Objects365. On Bees or CrowdHuman, TagsU is not considered since they only have one category.

PointsU According to [1], it takes 0.9 seconds on average to annotate one point. Thus, the time is $0.9 \times I_{avg}$.

PointsK We follow the computation of [8]. It takes 1 second to eliminate every non-existing class, and $C - C_{avg}$ seconds in total. [8] reports that annotators take a median of 2.4 seconds to click on the first instance of a class and 0.9 seconds for every additional instance. Thus the total labeling time is $(C - C_{avg}) + 2.4 \times C_{avg} + 0.9 \times (I_{avg} - C_{avg})$ on COCO, VOC, Objects365. On Bees or CrowdHuman, the time is equal to that of PointsU because since they only have one category.

TagsK It takes about 1 second to count a number [3]. Thus, on COCO, VOC, Objects365, the estimated time is $(C - C_{avg}) + 1.0 \times C_{avg} + 1.0 \times (I_{avg} - C_{avg}) = C + I_{avg} - C_{avg}$. Because this computation is derived from multi-class data domains [8], it is not applicable on Bees or CrowdHuman where they have only one category. For this reason, we simply estimate the TagsK cost of the single-class dataset as k times of the PointsK cost. Here k is the averaged proportion of TagsK cost over PointsK cost on three multi-class datasets (VOC, COCO and Objects365), i.e., $k = (21/22.9 + 84.2/88.7 + 375.8/381.7)/3 = 0.95$, where these numbers are from Table 11 of the main paper (the columns of TagsK and PointsK cost). Thus, the costs of TagsK on Bees and CrowdHuman are computed by $0.95 \times 6.4 = 6.1$, $0.95 \times 20.4 = 19.4$, respectively, where these numbers are from Table 11 of the main paper (the columns of TagsK and PointsK cost).

	COCO	VOC	Objects365	Bees	CrowdHuman
C	80	20	365	1	1
C_{avg}	3.5	1.4	5	1	1
I_{avg}	7.7	2.4	15.8	7.14	22.64

Table A. Dataset statistic. The information is provided by [2, 4, 5, 8, 9, 11].

BoxesEC [7] reports that it takes 7 seconds for one Extreme Clicking box, so the time is $7 \times I_{avg}$ seconds.

BoxesU Similarly, since annotating a high quality box needs 35 seconds [11], it takes $35 \times I_{avg}$ seconds per image.

Fully Following [8], the total time for full annotation is computed by $(C - C_{avg}) + 35 \times I_{avg}$ on COCO, VOC, Objects365. On Bees and CrowdHuman, the time is equal to that of BoxesU because there is no category labeling.

A.2. Datasets and Splitting Details in Section 5.5

In the paper, for each dataset, Figure 4 of the main paper shows two different mixture policies, and three different budgets for each mixture policy. Table B reports the detailed mixture percentages and other information. The training set used to be split into labeled and omni-labeled data is presented as follows for each dataset.

Bees [2] The total number of images is 3596. Since Bees does not split the dataset officially, we randomly sample

⋆ Work done during internship at Amazon. †Corresponding author.

Dataset	Omni-label	Fully (%)	None (%)	TagsK (%)	PointsU (%)	BoxesEC (%)	cost (hours)	mAP
Bees	Fully+TagsK+BoxesEC	5	0	80	0	15	25	39.9
		10	0	46	0	44	50	52.0
		20	0	34	0	46	75	57.7
	Fully+PointsU+BoxesEC	5	0	0	80	15	25	35.7
		10	0	0	46	44	50	51.5
		20	0	0	34	46	75	57.1
CrowdHuman	Fully+TagsK+BoxesEC	5	0	80	0	15	330	33.6
		10	0	46	0	44	660	35.4
		20	0	34	0	46	990	38.4
	Fully+PointsU+BoxesEC	5	0	0	80	15	330	30.4
		10	0	0	46	44	660	35.0
		20	0	0	34	46	990	38.2
VOC	Fully+None+BoxesEC	8	80	0	0	12	63.1	41.0
		10	29	0	0	61	126.2	48.0
		20	19	0	0	61	189.3	51.2
	Fully+PointsU+BoxesEC	8	0	0	91	1	63.1	40.9
		10	0	0	33	57	126.2	47.1
		20	0	0	22	58	189.3	50.4
COCO	Fully+None+BoxesEC	8	79	0	0	13	1.1×10^3	33.0
		10	26	0	0	64	2.3×10^3	35.8
		20	16	0	0	64	3.4×10^3	37.8
	Fully+PointsU+BoxesEC	8	0	0	91	1	1.1×10^3	33.3
		10	0	0	30	60	2.3×10^3	36.0
		20	0	0	18	62	3.4×10^3	38.0
Objects365	Fully+None+BoxesEC	8	75	0	0	17	2.4×10^3	10.4
		10	7	0	0	83	4.8×10^3	13.0
		25	25	0	0	50	7.2×10^3	13.9
	Fully+PointsU+BoxesEC	8	0	0	86	6	2.4×10^3	10.5
		10	0	0	8	82	4.8×10^3	13.1
		25	0	0	34	41	7.2×10^3	13.9

Table B. The details of omni-supervision experiments in Section 5.5 of the main paper.

80% images as the training set, after removing the broken images. The model is evaluated on the rest 20% data.

CrowdHuman [10] The official training set of 15,000 images is split by different percentages for the omni-supervision experiments. The model is evaluated on the official validation set.

VOC [4] We combine VOC07 *trainval* set and VOC12 *trainval* set as the training set with 22136 images in total, which is used for the omni-supervision experiments. The model is evaluated on the VOC07 *test* set.

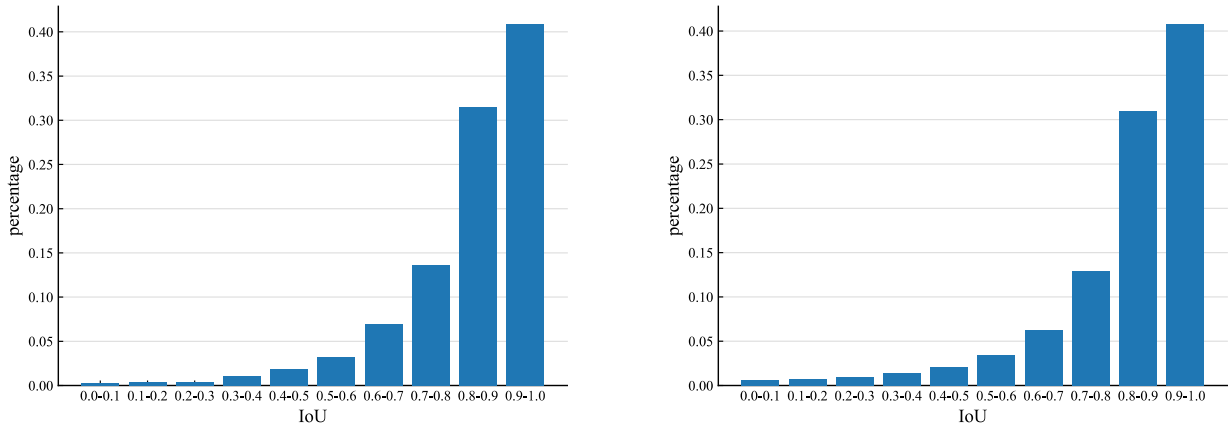
COCO [5] COCO *train2017* set of 118,287 images is used as the training set for splitting. The model is evaluated on the COCO *val2017* set.

Objects365 [9] To have faster experiments, 93,455 images are sampled from the Objects365 official training set as the training set for the omni-supervision experiments. In the process, since this dataset is long-tailed, we ensure that there is at least one image per category. Performance is evaluated on the official validation set.

The cost number (the second column from right) in Table B is computed by considering the mixture ratio, the dataset size and the cost per image in Table 11 of the main paper. For example, for the first row of Bees, the cost is $25 = (3596*0.05*249.9+3596*0.8*6.1+3596*0.15*50)/3600$.

A.3. The Simulation of Extreme Clicking Boxes

Because Extreme Clicking [7] does not release the annotations except for VOC [4], we simulate the boxes generated by Extreme Clicking for the other four datasets in our experiments. In detail, for each dataset, Gaussian noise is added to the ground truth bounding box coordinates, such that the distribution of mean Intersection over Union (mIoU) between the simulated boxes and ground truth boxes is close to the mIoU distribution between the given Extreme Clicking boxes and the ground truth boxes on VOC. The value of mIoU can be controlled by varying the covariance matrix of the Gaussian noise. Figure A shows the comparison of mIoU distribution of Extreme Clicking (left) and our simulation on COCO 10%Fully+90%BoxesEC setting (right) as an example. Their statistics comparison is: 1) Mean: 0.83 v.s. 0.82; Std: 0.15 v.s. 0.16; Second-order moment: 0.02 v.s. 0.02. These have shown our simulation is close to Extreme Clicking.



(a) Extreme Clicking on VOC. mean is 0.83; std is 0.15; second-order moment is 0.02.

(b) Simulated Extreme Clicking on COCO. mean is 0.82; std is 0.16; second-order moment is 0.02.

Figure A. The distribution of mIoU between the BoxesEC and ground truth.

A.4. Other Implementation and Training Details

The number of epoch for Burn-In stage depends on the size of the labeled data in our experiments. The total epoch number is chosen until the training saturated. They are shown in Table C. All models of Deformable DETR are trained with total batch size of 16. For other hyperparameters, we mainly follow the settings of Unbiased Teacher [6] and Deformable DETR [12]. For example, in (2) of the main paper, weight $\alpha = 2$, $\beta = 5$ to balance the classification loss ($\mathcal{L}^{cls}$) and regression loss ($\mathcal{L}^{\text{box}}$). $\mathcal{L}^{cls}$ is the focal loss with default hyperparameters, and $\mathcal{L}^{\text{box}} = 2\mathcal{L}_{\text{iou}} + 5\mathcal{L}_{\text{L1}}$ combines generalized IoU loss and L1 loss. EMA smoothing constant $k = 0.9996$ in (3) of the main paper. Weights $\lambda_{\text{iou}} = 2$, $\lambda_{L1} = 5$ in (9) for box matching. The object query number is $K = 300$ in Section 4 of the main paper.

Dataset	Fully (%)	Burn-In Epochs	Total Epochs
Bees	5	600	1000
	10	400	1000
	20	200	1000
	30	100	1000
CrowdHuman	5	400	800
	10	200	500
	20	100	500
	30	80	500
VOC	8	300	800
	10	200	500
	20	100	500
	30	80	200
COCO	1	400	800
	5	80	500
	8	60	300
	10	40	200
	20	20	150
	30	15	100
Objects365	8	130	300
	10	100	200
	20	50	200
	25	40	200
	30	30	200

Table C. The epoch setting details of omni-supervision training

References

[1] Amy Bearman, Olga Russakovsky, Vittorio Ferrari, and Li Fei-Fei. What's the point: Semantic segmentation with point supervision. In *ECCV*, pages 549–565. Springer, 2016.

[2] Bees. https://lila.science/datasets/boxes-on-bees-and-pollen.

[3] Counting. http://www.blog.republicofmath.com/how-long-does-it-take-to-count-to-one-trillion/.

[4] Mark Everingham, Luc Van Gool, Christopher KI Williams, John Winn, and Andrew Zisserman. The pascal visual object classes (voc) challenge. *IJCV*, 88(2):303–338, 2010.

[5] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *ECCV*, pages 740–755. Springer, 2014.

[6] Yen-Cheng Liu, Chih-Yao Ma, Zijian He, Chia-Wen Kuo, Kan Chen, Peizhao Zhang, Bichen Wu, Zsolt Kira, and Peter Vajda. Unbiased teacher for semi-supervised object detection. *arXiv preprint arXiv:2102.09480*, 2021.

[7] Dim P Papadopoulos, Jasper RR Uijlings, Frank Keller, and Vittorio Ferrari. Extreme clicking for efficient object annotation. In *ICCV*, pages 4930–4939, 2017.

[8] Zhongzheng Ren, Zhiding Yu, Xiaodong Yang, Ming-Yu Liu, Alexander G Schwing, and Jan Kautz. Ufo^2: A unified framework towards omni-supervised object detection. In

ECCV, pages 288–313. Springer, 2020.

[9] Shuai Shao, Zeming Li, Tianyuan Zhang, Chao Peng, Gang Yu, Xiangyu Zhang, Jing Li, and Jian Sun. Objects365: A large-scale, high-quality dataset for object detection. In *CVPR*, pages 8430–8439, 2019.

[10] Shuai Shao, Zijian Zhao, Boxun Li, Tete Xiao, Gang Yu, Xiangyu Zhang, and Jian Sun. Crowdhuman: A benchmark for detecting human in a crowd. *arXiv preprint arXiv:1805.00123*, 2018.

[11] Hao Su, Jia Deng, and Li Fei-Fei. Crowdsourcing annotations for visual object detection. In *AAAI workshop*, 2012.

[12] Xizhou Zhu, Weijie Su, Lewei Lu, Bin Li, Xiaogang Wang, and Jifeng Dai. Deformable detr: Deformable transformers for end-to-end object detection. *arXiv preprint arXiv:2010.04159*, 2020.