# Supplementary material - Not All Labels Are Equal: Rationalizing The Labeling Costs for Training Object Detection

Ismail Elezi[1]* Zhiding Yu[2] Anima Anandkumar[2,3] Laura Leal-Taixé[1] Jose M. Alvarez[2]
[1]TUM [2]NVIDIA [3]CALTECH

## Abstract

*In this supplementary material, we start by giving a discussion of the limitations and the broader impact of our paper (Section 1). We then continue by providing the exact numbers for all the graphs in the main paper (Section 2). We continue by providing more extended results in MS-COCO (Section 3) and devising a pseudo-label for class model (Section 4). We then compare our method with SSL methods that are boosted by active learning (Section 5), before we conclude by a discussion of engineering tricks that we considered in our work (Section 6).*

## 1. Limitations and broader impact

**Limitations.** Our method is task-specific and limited to a set of known categories. Furthermore, our method is less suitable for acquiring datasets for multi-task networks. Finally, we only experimented with a vanilla pseudo-labeling method, which might not reach the best possible results.

**Broader impact.** Our work introduces a unified framework for reducing the labeling costs needed to train object detection networks. It provides a way of using all the samples in the dataset, be they labeled or not, in an optimized way to reach high accuracy. Manually selecting and labeling frames takes a tremendous amount of time and labor, so by selecting the right data for annotation and training, our approach can positively impact by reducing storage and labeling costs on industries such as autonomous driving that require large amounts of labeled data. Our approach uses a single model with the minimum amount of training data to maximize performance from an ecological standpoint. Our results suggest that this is a practical approach to address inefficiencies in training data selection for real-world applications such as autonomous driving. As our approach requires fewer training resources, thus we also reduce the carbon footprint.

*Work performed while interning at NVIDIA.

## 2. Exact numbers for the experiments given in the main paper

In the paper, we provide plots for the main experiments due to the limited space. In Tables 1, 2a, 2b, 3, 4a, 4b, we summarize the exact numbers corresponding to Figures 3a, 3b, 3c, 4a, 4b and 4c of the main paper. We provide the mean and the standard deviation for each method and AL cycle. Each experiment has been run three times.

## 3. Extended results in MS-COCO

In the paper we presented aggregated results on the number of classes one acquisition function performs another, and in pseudo-label performance boost per class. We presented results by aggregating them over the five cycles of active learning. For completeness, in Figures 1 and 2, we provide results for each AL cycle in isolation. We see the same trend as in the paper.

## 4. Pseudo-labels for class

In this experiment we analyze different methods for obtaining pseudo-labels. Precisely, instead of obtaining pseudo-labels using the confidence score and a threshold $\tau$ independently of the class, we consider the $k\%$ most confident objects for each class and add them as pseudo-labels. In Table 5a we present the results where we pseudo-label the top $20\%$, top $30\%$, top $40\%$ most confident predictions for class and compare them with the results of our method described in the main paper. We see that the methods where we pseudo-label per class work well, but worse than our method. Thus, for both simplicity and perormance, we choose to use our class-agnostic method.

## 5. Comparison with semi-supervised learning methods

In this section, we compare our method with semi-supervised learning methods that are combined with active learning methods. These results are similar to those of Tab. 3 of the main paper, where we used different active learning acquisition functions (random, entropy, and inconsis-

| Cycle | Random | Entropy | Core-Set [1] | LLAL [2] | Ensemble [3] | MC-dropout [4] | CDAL-RL [5] | MI-AOD [6] | PM [7] | Ours |
|---|---|---|---|---|---|---|---|---|---|---|
| 0 | 60.82±0.2 | 61.23±0.8 | 62.36±0.5 | 60.95±0.4 | 60.82±0.2 | 60.82±0.2 | 61.45 ±0.2 | 62.20 ±0.2 | 61.30 ±0.5 | **63.25**±0.2 |
| 1 | 64.23±0.2 | 63.57±0.9 | 65.90±0.4 | 64.91±0.5 | 65.70±0.9 | 66.90±0.3 | 65.30 ±0.2 | 65.60 ±0.2 | 65.56 ±0.3 | **70.95** ±0.1 |
| 2 | 66.33±0.2 | 66.94±0.2 | 67.63±0.2 | 66.90±0.3 | 69.20±0.3 | 68.40±0.2 | 68.20 ±0.3 | 69.25 ±0.2 | 68.43 ±0.1 | **72.88**±0.1 |
| 3 | 67.51±0.2 | 68.70±0.2 | 68.88±0.5 | 69.05±0.5 | 71.50±0.2 | 70.80±0.4 | 70.30 ±0.2 | 70.35 ±0.2 | 70.77 ±0.1 | **73.55**±0.2 |
| 4 | 68.60±0.5 | 69.82±0.1 | 69.44±0.3 | 70.35±0.6 | 72.90±0.3 | 71.90±0.5 | 71.60 ±0.2 | 70.80 ±0.2 | 72.52 ±0.1 | **74.75**±0.2 |
| 5 | 69.27±0.2 | 70.18±0.3 | 70.16±0.1 | 71.49±0.7 | 74.29±0.2 | 73.81±0.0 | 72.20 ±0.2 | 72.00 ±0.2 | 73.52 ±0.5 | **75.60**±0.2 |

Table 1. **VOC07+12**. Comparison to state-of-the-art active learning methods. We initially use $2,000$ randomly sampled images and, in every other cycle, we label $1,000$ extra images. Our method outperforms all the other methods, including ensembles, by a large margin.

| Cycle | Random | SSL-cons. [8] | SSL-PL [9] | Ours |
|---|---|---|---|---|
| 0 | 60.82±0.2 | **63.25**±0.2 | **63.25**±0.2 | **63.25**±0.2 |
| 1 | 64.23±0.2 | 67.19 ±0.1 | 69.60 ±0.5 | **70.95** ±0.1 |
| 2 | 66.33±0.2 | 69.44 ±0.1 | 70.90 ±0.5 | **72.88**±0.1 |
| 3 | 67.51±0.2 | 71.13 ±0.1 | 71.80 ±0.1 | **73.55**±0.2 |
| 4 | 68.60±0.5 | 72.18 ±0.1 | 72.60 ±0.2 | **74.75**±0.2 |
| 5 | 69.27±0.2 | 73.10 ±0.1 | 73.30 ±0.2 | **75.60**±0.2 |

(a)

| Cycle | Random | Entropy | Inconsistency | Combined | Ours |
|---|---|---|---|---|---|
| 0 | **63.25**±0.2 | **63.25**±0.2 | **63.25**±0.2 | **63.25**±0.2 | **63.25**±0.2 |
| 1 | 67.19 ±0.1 | 67.24±0.1 | 67.39±0.9 | 68.40±0.3 | **70.95** ±0.1 |
| 2 | 69.44±0.1 | 70.05±0.1 | 70.42±0.6 | 70.84±0.8 | **72.88**±0.1 |
| 3 | 71.13±0.1 | 72.13±0.2 | 72.43±0.4 | 72.93±0.3 | **73.55**±0.2 |
| 4 | 72.18±0.1 | 73.48±0.6 | 72.80±1.0 | 73.66±0.2 | **74.75**±0.2 |
| 5 | 73.1±0.1 | 74.77±0.2 | 74.90±0.3 | 74.98±0.5 | **75.60**±0.2 |

(b)

Table 2. **VOC07+12**. a) Comparison to two semi-supervised learning methods. We initially use $2,000$ randomly sampled images and, in every other cycle, we label $1,000$ extra images. Our method outperforms both of them by a large margin. b) Ablation study on the effect of entropy, inconsistency, unified score, and our method in VOC07+12. We observe that doing active learning with either entropy or consistency outperforms the semi-supervised model, that the unified score performs better than either of the individual scores, and that our method reaches the best overall results.

| Cycle | Random | Entropy | Core-Set [1] | Ensemble [3] | MC-dropout [4] | PM [7] | Ours |
|---|---|---|---|---|---|---|---|
| 0 | 25.63±0.4 | 25.63±0.4 | 25.63±0.4 | 27.50±0.3 | 27.50±0.3 | **27.70** ±0.1 | 27.50±0.3 |
| 1 | 28.40±0.1 | 28.57±0.2 | 28.10±0.5 | 28.65±0.1 | 28.70±0.2 | 29.28 ±0.1 | **30.07**±0.4 |
| 2 | 29.40±0.2 | 29.47±0.1 | 29.57±0.1 | 29.75±0.2 | 29.42±0.2 | 30.51±0.1 | **31.63**±0.1 |
| 3 | 30.20±0.6 | 30.37±0.1 | 30.40±0.4 | 30.43±0.1 | 30.24±0.2 | 31.20 ±0.1 | **32.10**±0.1 |
| 4 | 31.03±0.1 | 31.17±0.2 | 31.17±0.3 | 31.20±0.2 | 31.03±0.1 | 31.86 ±0.1 | **32.43**±0.1 |
| 5 | 31.47±0.3 | 31.50±0.2 | 31.87±0.2 | 31.75±0.1 | 31.22±0.1 | 32.27 ±0.1 | **32.80**±0.0 |

Table 3. **MS-COCO**. Comparison to state-of-the-art methods. In this case, we initially use $5,000$ randomly sampled images, and, in every active learning cycle, we label $1,000$ extra images. Our method outperforms all the other methods, including ensembles, by a large margin.

| Cycle | Random | SSL-cons. [8] | SSL-PL [9] | Ours |
|---|---|---|---|---|
| 0 | 25.63±0.4 | **27.50**±0.3 | **27.50**±0.3 | **27.50**±0.3 |
| 1 | 28.40±0.1 | 28.52 ±0.1 | 28.70 ±0.2 | **30.07**±0.4 |
| 2 | 29.40±0.2 | 29.20 ±0.1 | 29.42 ±0.2 | **31.63**±0.1 |
| 3 | 30.20±0.6 | 30.20 ±0.1 | 30.24 ±0.2 | **32.10**±0.1 |
| 4 | 31.03±0.1 | 31.03 ±0.1 | 31.03 ±0.1 | **32.43**±0.1 |
| 5 | 31.47±0.3 | 31.47 ±0.1 | 31.22±0.1 | **32.80**±0.0 |

(a)

| Cycle | Random | Entropy | Inconsistency | Combined | Ours |
|---|---|---|---|---|---|
| 0 | **27.50**±0.3 | **27.50**±0.3 | **27.50**±0.3 | **27.50**±0.3 | **27.50**±0.3 |
| 1 | 28.53±0.1 | 28.97±0.3 | 28.90±0.3 | 29.17±0.2 | **30.07**±0.4 |
| 2 | 29.20±0.1 | 29.77±0.1 | 29.90±0.7 | 30.17±0.2 | **31.63** ±0.1 |
| 3 | 29.87±0.1 | 29.93±0.4 | 30.85±0.4 | 30.90±0.2 | **32.10** ±0.1 |
| 4 | 31.03±0.1 | 31.10±0.1 | 31.55±0.1 | 31.55±0.1 | **32.43** ±0.1 |
| 5 | 31.10±0.1 | 31.40±0.1 | 31.65±0.3 | 31.50±0.1 | **32.80** ±0.0 |

(b)

Table 4. **MS-COCO**. a) Comparison to two semi-supervised learning methods. We initially use $5,000$ randomly sampled images and, in every other cycle, we label $1,000$ extra images. Our method outperforms both of them by a large margin. b) Ablation study on the effect of entropy, inconsistency, unified score, and our method in MS-COCO. We observe that doing active learning with either entropy or consistency outperforms the semi-supervised model, that the unified score performs better than either of the individual scores, and that our method reaches the best overall results.

tency) on top of the consistency-based SSL [8]. Here, we add a diversity method (core set [1]) and we also experiment with the pseudo-labeling SSL method. We show the results in Tab. 6. As we show, both entropy and inconsistency when combined with either form of semi-supervision improve over the baseline. However, our method still significantly outperforms these naive combinations. The core-set method [1] harms the training in both cases.

| Cycle | Top 20% | Top 30% | Top 40% | Ours |
|---|---|---|---|---|
| 0 | 63.25±0.3 | 63.25±0.3 | 63.25±0.3 | **63.25**±0.3 |
| 1 | 69.42±0.1 | 69.63±0.1 | 69.67±0.3 | **70.95**±0.1 |
| 2 | 71.84±0.3 | 71.84±0.3 | 72.14±0.2 | **72.88**±0.1 |
| 3 | 73.34±0.3 | 73.47±0.1 | 73.56±0.2 | **73.55**±0.1 |
| 4 | 74.53±0.1 | 74.53±0.2 | 74.32±0.2 | **74.75**±0.1 |
| 5 | 75.13±0.1 | 75.39±0.2 | 75.19±0.2 | **75.60**±0.1 |

(a)

| Cycle | Random | Balanced quarter | Unified |
|---|---|---|---|
| 0 | 53.66±0.2 | **63.37**±0.2 | 63.25±0.2 |
| 1 | 67.39±0.4 | 68.12±0.2 | **68.40**±0.3 |
| 2 | 69.90±0.5 | 70.12±0.6 | **70.84**±0.8 |
| 3 | 71.38±1.2 | 71.92±0.3 | **72.93**±0.3 |
| 4 | 73.53±0.4 | 73.48±0.3 | **73.66**±0.2 |
| 5 | 74.30±0.4 | 73.90±0.4 | **74.98**±0.5 |

(b)

Table 5. **VOC07+12**. a) The results of adding top $k\%$ most confident pseudo-labels for class, compared to the results of our method. *Top* $20\%$, *Top* $30\%$, *Top* $40\%$ represent the methods where we choose to pseudo-label the most confident $20\%$, $30\%$ and $40\%$ pseudo-labels per class. *Ours* represent our method where we pseudo-label all the objects for which the network's confidence is greater than $0.99$. b) Accuracy as a function of label/unlabeled sampling strategy. *Random* refers to random sampling from the entire dataset, *Balanced quarter* refers to having a quarter of labeled samples; *Unified* refers to half of the samples being labeled. Our balanced strategy outperforms the other two strategies. Note that in order to check only the effect of balancing, we do not add pseudo-labels during the training.

| Method/Cycle | 1 | 2 | 3 | 4 | 5 |
|---|---|---|---|---|---|
| CSD [8] | 67.19 | 69.44 | 71.13 | 72.18 | 73.10 |
| CSD [8] + incons. | 67.39 | 70.42 | 72.43 | 72.80 | 74.90 |
| CSD [8] + entropy | 67.24 | 70.05 | 72.13 | 73.48 | 74.77 |
| CSD [8] + coreSet | 62.93 | 65.35 | 67.15 | 69.32 | 70.62 |
| PL [9] | 69.60 | 70.90 | 71.80 | 72.60 | 73.30 |
| PL [9] + incons. | 68.32 | 69.93 | 72.13 | 73.42 | 73.74 |
| PL [9] + entropy | 66.87 | 71.48 | 72.30 | 74.06 | 74.62 |
| PL [9] + coreSet | 66.35 | 67.56 | 70.59 | 72.06 | 72.93 |
| Ours | **70.95** | **72.88** | **73.55** | **74.75** | **75.60** |

Table 6. **VOC07+12**. Comparison to other semi-supervised active learning methods. We use the consistency-based method [8] and pseudo-labeling method [9] as semi-supervised learning method, and use random sampling, entropy, inconsistency and core set as active learning method. We show that our method significantly outperforms the other approaches.

# 6. Engineering tricks to consider

## 6.1. Non-maximum suppression

We found the effect of non-maximum suppression (NMS) to be very important in all AL methods. Without applying NMS, active learning methods did not work better than a random sampling method. We hypothesize that this happens because if we do not apply NMS, the number of detected boxes is in the hundreds, so by sheer chance, some of them might have high acquisition scores. Considering that in a real-world scenario these boxes would be *killed* by NMS, we conclude that these boxes should not be used to compute an acquisition score. Thus, for every image, we apply NMS before proceeding with the computation of the acquisition score.

## 6.2. Balanced mini-batches

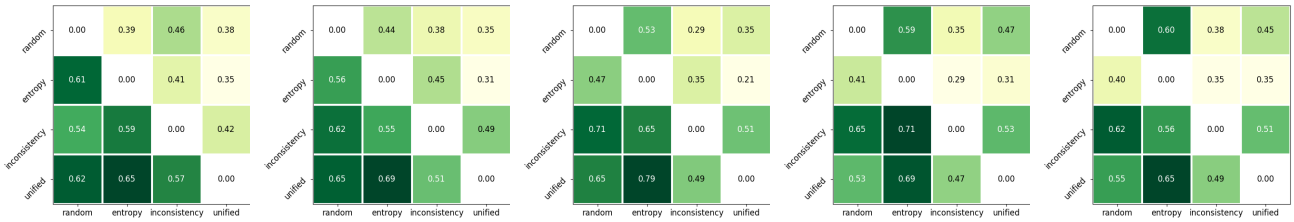
In the main paper, for every experiment, we force that half of the samples in a mini-batch are labeled. In this experiment, we evaluate the effect of varying the number of labeled samples in a mini-batch. In particular, we compare our results to having only half of the samples labeled, and a random approach. In order to be able to quantify the effect of balancing, we do all the experiments without adding pseudo-labels. We present the results in Table 5b. We observe that our strategy of balancing the mini-batches so they contain an equal number of labeled and unlabeled samples, performs best by up to $1pp$ in all AL cycles except the zeroth one, when it gets outperformed by $0.12pp$ by the strategy where only a quarter of samples contain labels. We also observe that the strategy where we do only random sampling consistently reaches the worst results. In fact, in the zeroth AL cycle it gets outperformed by the balanced strategies by almost $10pp$. This can be explained by the fact that the number of labeled samples $(2,000)$ is much lower than the number of unlabeled samples $(14,651)$, so in a mini-batch of size $32$, in average, only $3.86$ samples have labels. In some mini-batches, the number of labeled samples is $0$, and thus the loss function becomes completely self-supervised. We observe that when the number of labeled samples increases (by labeling other images during AL stage), the overall performance increases, but it still lags behind the the balanced strategies.

## 6.3. Balancing the losses

Similar to balancing the mini-batches, we experiment with balancing the weight of the consistency loss. We show the experiments in Tab. 7, and these results are complementary to those of the balanced mini-batches. Similar to the results with balanced mini-batches, we see that we reach the top performance where we do not use any weight (weight=1) for the consistency loss.

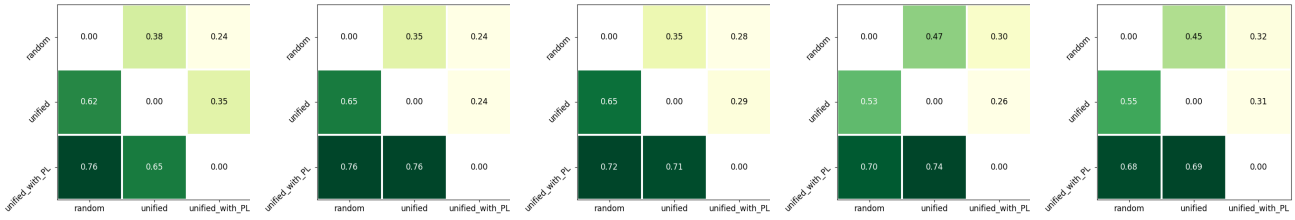
# Acknowledgments

This research was partially funded by the Humboldt Foundation through the Sofja Kovalevskaja Award and a Humboldt Research Fellowship. The authors thank Jiwoong Choi and Aljosa Osep for useful discussion.




(1) (2) (3) (4) (5)

Figure 1. **MS-COCO.** The percentage of classes where one acquisition function outperforms another. Numbers 1-5 represent the active learning cycle. Example: taking the entry "unified" in the y-axis, and "entropy" in the x-axis in (1), we get the value 0.65 which means that "unified" acquisition function outperforms the "entropy" acquisition function in 65% of classes during the first active learning cycle.




(1) (2) (3) (4) (5)

Figure 2. **MS-COCO.** The percentage of classes where our unified acquisition function outperforms random with and without pseudo-labels. Numbers 1-5 represent the active learning cycle. Example: taking the entry "unified_with_PL" in the y-axis, and "random" in the x-axis in (1), we get the value 0.76 which means that our method outperforms the random acquisition function in 76% of classes during the first active learning cycle.

| Method/Cycle | 1 | 2 | 3 | 4 | 5 |
|---|---|---|---|---|---|
| weight=0.25 | 69.75 | 65.55 | 73.39 | 73.89 | 74.18 |
| weight=0.50 | 69.62 | 65.90 | 73.06 | 74.37 | 74.48 |
| **Ours (weight=1)** | **70.95** | **72.88** | **73.55** | **74.75** | **75.60** |
| weight=2.00 | 69.92 | 65.74 | 73.31 | 74.12 | 74.22 |
| weight=4.00 | 70.09 | 63.58 | 72.07 | 73.78 | 74.06 |

Table 7. **VOC07+12.** Accuracy as a function of the weight of the unsupervised loss.

## References


[1] Ozan Sener and Silvio Savarese. Active learning for convolutional neural networks: A core-set approach. In *ICLR*, 2018. 2

[2] Donggeun Yoo and In So Kweon. Learning loss for active learning. In *CVPR*, 2019. 2

[3] William H. Beluch, Tim Genewein, Andreas Nürnberger, and Jan M. Köhler. The power of ensembles for active learning in image classification. In *CVPR*, 2018. 2

[4] Yarin Gal, Riashat Islam, and Zoubin Ghahramani. Deep bayesian active learning with image data. In *ICML*, 2017. 2

[5] Sharat Agarwal, Himanshu Arora, Saket Anand, and Chetan Arora. Contextual diversity for active learning. In *ECCV*, 2020. 2

[6] Tianning Yuan, Fang Wan, Mengying Fu, Jianzhuang Liu, Songcen Xu, Xiangyang Ji, and Qixiang Ye. Multiple instance active learning for object detection. In *CVPR*, 2021. 2

[7] Jiwoong Choi, Ismail Elezi, Hyuk-Jae Lee, Clément Farabet, and Jose M. Alvarez. Active learning for deep object detection via probabilistic modeling. In *ICCV*, 2021. 2

[8] Jisoo Jeong, Seungeui Lee, Jeesoo Kim, and Nojun Kwak. Consistency-based semi-supervised learning for object detection. In *NeurIPS*, 2019. 2, 3

[9] Dong-Hyun Lee. Pseudo-label: The simple and efficient semi-supervised learning method for deep neural networks. *ICMLW*, 2013. 2, 3