Training Ensembles with Inliers and Outliers for Semi-supervised Active Learning -Supplementary-

Vladan Stojnić Zakaria Laskar Giorgos Tolias
Visual Recognition Group, Faculty of Electrical Engineering, Czech Technical University in Prague
stojnvla,laskazak,toliageo@fel.cvut.cz

In the supplementary, we provide the following information and material.

- Additional comparison to the recent MQNet [PSB+22].
- The impact of the weights used for pseudo-labels.
- Additional experiments for a smaller initial labeled set.
- Average performance and standard deviation across seeds for all evaluated methods.
- The full list of implementation details of the experiments in the main paper.
- The full algorithm in the form of pseudo-code.
- The details regarding the inlier/outlier class splits for CIFAR100 and TinyImages.
- Training time comparison with other methods.
- Analysis of the relation between inlier rate and performance.
- Results with an ensemble classifier during inference.
- Results on an additional dataset split on ImageNet.
- The impact of pseudo-label accuracy on the effectiveness of semi-supervision.

I. Results on the setup from MQNet

We evaluate our method on the experimental setup of MQNet [PSB+22] and present the results in Table I. We perform this experiment due to the following differences: (1) to use the same inlier/outlier class splits and the same initial labels set and unlabeled set, (2) to perform experiments without SSL pre-training as in their work and (3) to perform the network training with the same hyper-parameters as in their work. To be sure for a direct comparison, we implement our method in their own implementation framework. Results confirm the same observations as in our own setup; our method outperforms MQNet even without semi-supervision.

II. Impact of pseudo-label weights

We evaluate the impact of weights $w_t(x)$ used for pseudo-labels by setting them all to 1.0. Results are presented in Figure I, which shows that weights provide a benefit if an ensemble is not used, while results with and without weights are comparable in the case ensemble is used. This is because ensembles improve pseudo-label accuracy, so assigning high weights for them is safe.

In Figure II, we show the evolution of pseudo-label weights over active learning rounds. It is observed that over active learning rounds, correct pseudo-labels are getting higher weights meaning that the classifier is becoming certain about those predictions. In contrast, incorrect pseudo-labels mostly have weights in the lower middle of the range.

III. Experiment with a smaller labeled set

In Figure III, we present additional results on the ImageNet dataset for the case when the initial labeled set L_0 contains 5 examples per class and the budget is set to 100. Results show that our approach outperforms all other recent state-of-the-art competitors and baseline methods by a large margin. The variant without semi-supervision is either second best or close to second best over all cases.

Method	Dataset		
	CIFAR10	CIFAR100	ImageNet
MQNet	89.51	52.82	54.11
Ours w/o semi	91.63	54.23	58.00
Ours	92.95	56.20	62.30

Table I. Results after 10 acquisition rounds on the setup from the MQNet paper. Outlier ratio is 0.6. Reported values for MQNet are taken from [PSB+22].

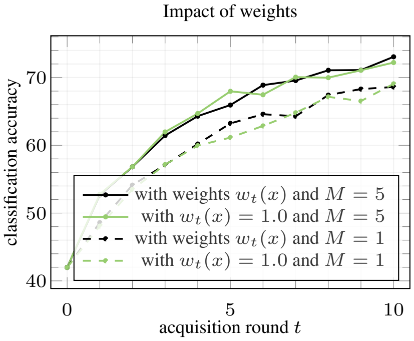

Figure I. Comparison of classification accuracy for our approach with semi-supervision with and without weights for pseudo-labels in Equation 3 of the main paper. Results are presented on ImageNet dataset with a 0.5 outlier ratio.

IV. Detailed results

We present detailed results, including standard deviation, for the main experiments from the paper. These results are presented in Table II and Table III.

V. Implementation details

For the backbone of all experiments, we use ResNet18 [HZRS16]. For CIFAR100 and TinyImageNet experiments, we use the variant commonly used for CIFAR experiments. It is standard practice to use this variant [KPKC21, KBKI21] which uses a kernel of size 3 and stride 1 instead of 7 and 2, respectively, in the first convolutional layer[1]. For ImageNet experiments, we use the standard version with a kernel size of 7 and stride 2 in the first convolutional layer. SSL pre-training is performed for 700 epochs using a batch of size 32, 64, and 100 for CIFAR100, TinyImageNet, and ImageNet, respectively, initial learning rate equal to 1e-1 with cosine annealing and SGD optimizer. The result is used as initialization for classifier training, which is performed for 10 epochs using a batch of size 32, learning rate equal to 5e-4, and Adam optimizer for the training on the labeled set. In the experiments, this setup is fixed for all methods we compare with.

For the semi-supervised training, we continue training from the point where training on the labeled set stopped. We do this for 3 epochs, where we consider one full pass through the unlabeled set as the epoch. We use a batch size of 512, where half of the batch comes from the unlabeled set and the other half comes from the labeled set. All members of the ensemble use the same architecture and are initialized by the same weights coming from the self-supervised pre-training, which is performed only once. They are also trained using the same optimization hyper-parameters. Ensemble members differ only by the seed used in the implementation for randomization, which affects training data shuffling and the randomization of the augmentations per image. During training, we use random horizontal flipping as the augmentation on CIFAR100 and TinyImageNet, while on ImageNet, we first perform random resized cropping and then random horizontal flipping. Pseudo-code of our method is presented in Algorithm 1.

We run CoreSet, BADGE, CCAL, SIMILAR, and MQNet using the provided implementations [2], after integrating them into our implementation framework. We implement LfOSA by ourselves.

Algorithm 1 Overview of the approach.

```
procedure AL(labeled set $L_0$, unlabeled set $U_0$, do-semi, do-filtering)
    $f_{\text{init}} \leftarrow$ SSL on $L_0 \cup U_0$    ▷ self-supervised pre-training
    for $t \in [0, \dots, T]$ do    ▷ active learning rounds
        for $i \in [1, \dots, M]$ do    ▷ supervised training, $M$ models
            $f_{t_i} \leftarrow \arg\min_f \mathcal{L}(L_t; f)$    ▷ start from $f_{\text{init}}$, train $\mathcal{L}$
        end for
        if do-semi is true and $t \neq 0$ then    ▷ semi-supervision
            for $x \in U_0$ do $\hat{y}_t(x) \leftarrow \arg\max_j F_t(x)_j$    ▷ pseudo-label
            for $x \in U_0$ do $w_t(x) \leftarrow 1 - \frac{H(F_t(x))}{\log(K+1)}$    ▷ weights
            for $i \in [1, \dots, M]$ do    ▷ semi-supervised training, $M$ models
                $f^{'}_{t_i} \leftarrow \arg\min_f \mathcal{L}_{\text{semi}}(L_t, U_t; f)$    ▷ train longer with $\mathcal{L}_{\text{semi}}$
            end for
        end if
        for $x \in U_t$ do    ▷ loop to estimate acquisition score
            if $t = 0$ then
                $a_t(x) \sim \mathcal{U}_{[0,1]}$    ▷ random chance
            else
                $\tilde{a}_t(x) \leftarrow 1 - \frac{\left|\left\{i \colon \hat{y}'_{t_i}(x) = \hat{y}'_t(x)\right\}\right|}{M}$    ▷ VR score
                if do-filtering is true then
                    $a_t(x) \leftarrow \tilde{a}_t(x) \mathbb{1}_{\hat{y}_t(x) \neq C_o}$    ▷ filtering
                else
                    $a_t(x) \leftarrow \tilde{a}_t(x)$    ▷ no filtering
                end if
            end if
        end for
        $A_t \leftarrow \text{top}_{\text{B}}\{a_t(x) : x \in U_t\}$    ▷ example selection based on largest score
        annotate($A_t$)    ▷ annotators assign labels
        $L_{t+1} \leftarrow L_t \cup A_t$    ▷ update the labeled set
        $U_{t+1} \leftarrow U_t \setminus A_t$    ▷ update the unlabeled set
    end for
end procedure
```

VI. Benchmark details

The original CIFAR100 consists of 100 categories. We use 20 of them as inlier classes, and the rest are used to form the outlier class. The former correspond to large omnivores and herbivores, medium-sized mammals, and small mammals. This particular way of splitting classes is performed in prior work, but without publicly sharing the list of images per split [DZC+21]. Therefore, we adopt the same

[1] This architecture is used for SLL by CCAL, but not for the classifier, even though we found it to be beneficial.

[2] https://github.com/RUC-DWBI-ML/CCAL
https://github.com/decile-team/distil
https://github.com/kaist-dmlab/MQNet

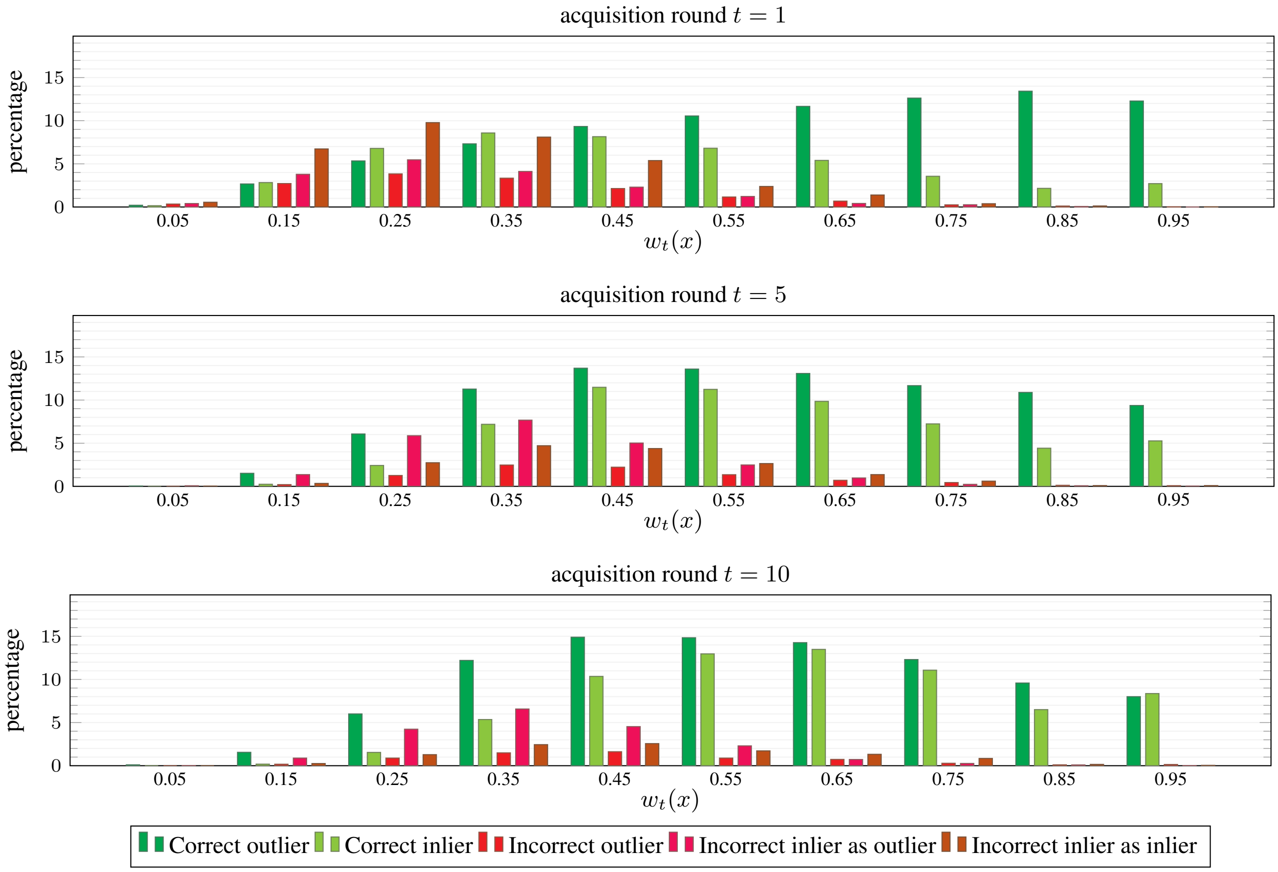

Figure II. Distribution of weights $w_t(x)$ for different types of pseudo-labels. *Correct inlier/outlier*: example pseudo-labeled correctly. *Incorrect outlier*: outlier example wrongly pseudo-labeled as an inlier (as any of the inlier classes). *Incorrect inlier as outlier*: inlier example incorrectly pseudo-labeled as an outlier. *Incorrect inlier as inlier*: inlier example wrongly pseudo-labeled into the wrong inlier class. Y-axis shows the percentage of outlier/inlier examples from each type, *i.e. Correct outlier* and *Incorrect outlier* sum to 100, and *Correct inlier*, *Incorrect inlier as outlier* and *Incorrect inlier as inlier* also sum to 100.

class splits and define our own image splits, which we will publicly share. The test set is formed by examples coming from the original test split and contains all images of inlier classes. It consists of 2000 images in total.

The original TinyImageNet consists of 200 categories. We use 25 categories corresponding to land animals as inlier classes, and the rest are used to form the outlier class. The test set is formed by examples from the original validation split and contains all images of inlier classes. It consists of 1250 images in total.

We provide the inlier/outlier class splits for CIFAR100, TinyImageNet, and ImageNet datasets. While CIFAR100 splits are obtained from CCAL [DZC+21], TinyImageNet, and ImageNet splits are created from scratch for our work. The ids for classes used as inliers are listed below, while the ids of outlier classes will be released with the code.

1. CIFAR100: 3, 42, 43, 88, 97, 15, 19, 21, 32, 39, 35, 63, 64, 66, 75, 37, 50, 65, 74, 80
2. TinyImageNet: 29, 54, 114, 159, 171, 197, 94, 174, 192, 28, 1, 11, 5, 24, 83, 128, 82, 108, 118, 98, 180, 62, 163, 111, 78
3. ImageNet: n02085620, n02086240, n02086910, n02087046, n02089867, n02089973, n02090622, n02091831, n02093428, n02099849, n02100583, n02104029, n02105505, n02106550, n02107142, n02108089, n02109047, n02113799, n02113978, n02114855, n02116738, n02119022, n02123045, n02138441, n02326432

VII. Training time comparison

We present training time comparison of different active learning methods in Figure IV. The timings include all steps required for each method between two acquisition rounds. The proposed method has increased training time mainly due to the use of ensembles and semi-supervised learning.

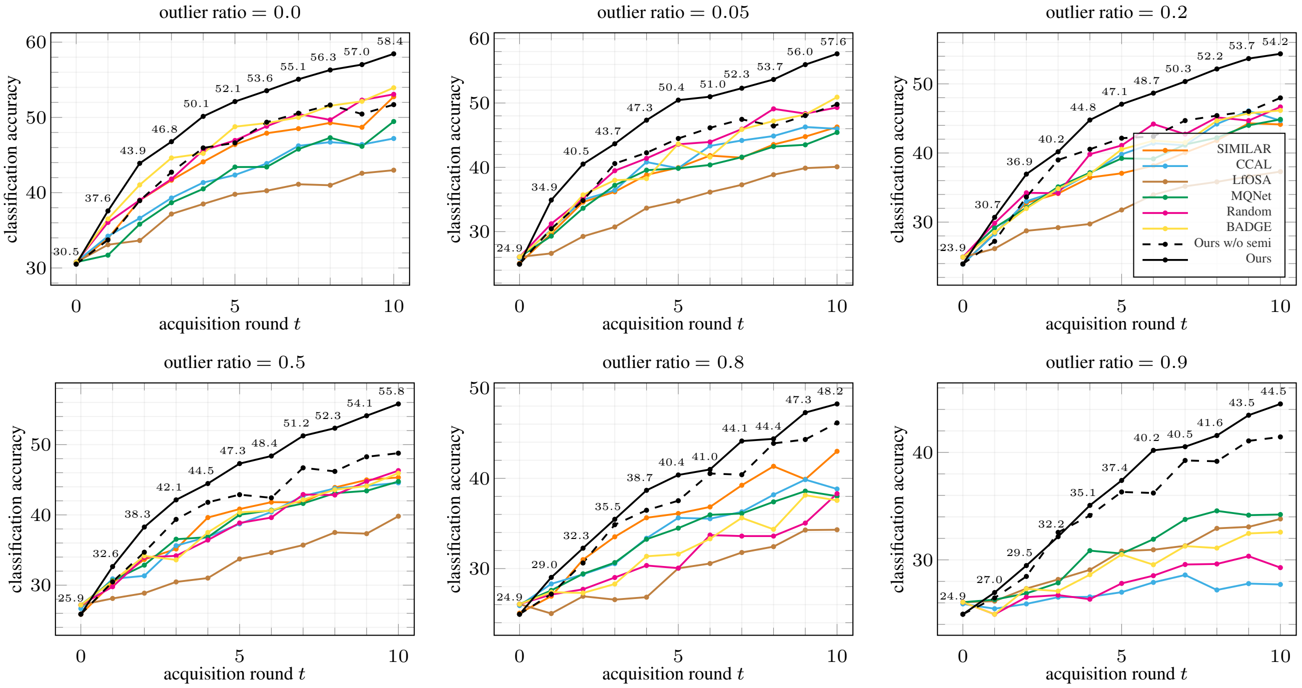

Figure III. Comparison of classification accuracy over multiple active learning rounds for varying outlier ratios on ImageNet when initial labeled set L_0 contains 5 examples (in contrast to 20 in the main paper) per inlier class and budget is equal to 100. SIMILAR is excluded for 0.9 outlier ratio since we were not able to run it even on a machine with 800GB of RAM.

However, it is still faster than the current SoA method, SIMILAR, which performs computationally heavy optimization for selection. The times of CCAL are higher because we use the official implementation, which is not optimized for efficiency. Note that a 3-network ensemble reduces training time without a significant drop in performance and that recent developments train a single network and save its checkpoints to form the ensemble [LA16, WWL+21], which can speed up the proposed method. Since ensembles and semi-supervised learning improve any other method, as shown in the main paper, this time increase comes with performance benefits for every method.

We report in detail the timings of the different steps for our method for round $t = 5$. Supervised training with labeled examples takes 172s. Estimating pseudo-labels and their weights takes 107s. Semi-supervised training with all examples takes 591s. Estimation of the acquisition scores for all unlabeled examples takes 110s. The lower times of our method at $t = 0$ are due to training only one network and not an ensemble, not performing semi-supervised training, and using random sampling at this round.

VIII. Analysis of inlier rate and performance

In Figures 3 and 4 of the main paper we observe that LfOSA achieves to acquire with the highest percentage of inliers, but its accuracy is quite low. We further analyze this. In Figure V, we present histograms of pairwise similarities of selected examples for one active learning round for LfOSA and the proposed method. We compare the overall distribution of similarities with the similarities within the same class. We see that examples of the same class selected by LfOSA are much more similar to each other than the examples selected by our method; the two distributions are roughly the same for our method, while for LfOSA this is not the case. This evidence makes us conclude that the high inlier rate does not lead to high performance due to the low variability of the selected examples.

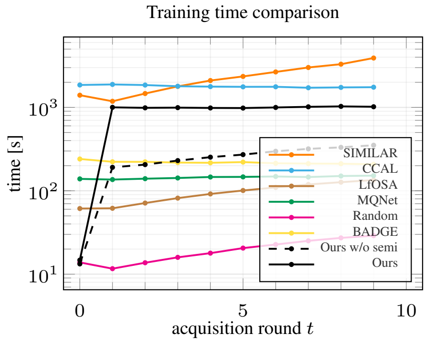

Figure IV. Training time comparison of different active learning methods. Results are presented on ImageNet with a 0.5 outlier ratio. Timings are measured on the NVIDIA A100 GPU and with the use of 6 cores of AMD EPYC 7543.

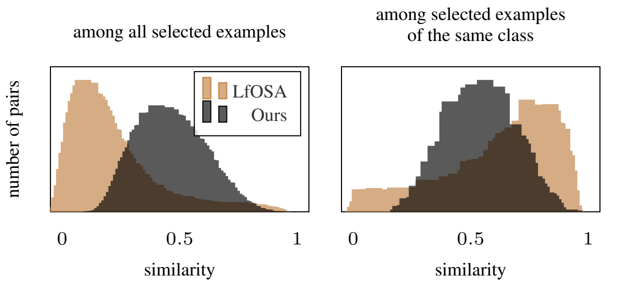

Figure V. Histograms of pairwise similarities between the selected examples for one acquisition round.

IX. Ensembles for inference

In all our experiments, we use ensembles to improve the accuracy of pseudo-labels and the selection, while during inference we use only one network from the ensemble to reduce the inference complexity and to have a fair comparison with prior methods. However, ensembles can also improve the final classification accuracy. We perform a single experiment for that and present results in Figure VI. The use of the ensemble noticeably improves the classification accuracy over all active learning rounds.

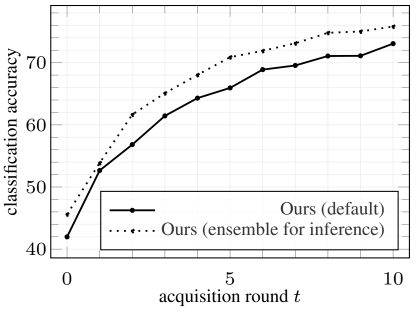

Figure VI. Comparison of classification accuracy for our method with and without using the ensemble network for inference. Results are presented on ImageNet with a 0.5 outlier ratio.

X. Additional dataset split on ImageNet

To test whether the proposed method is favored by the split of inlier classes, we derive another split of inlier/outlier classes on ImageNet. This split, as inlier classes contains 25 classes that are randomly sampled from all the carnivore animal classes, while the outlier class contains images coming from 700 different classes. [3] We present the results on this split of the data in Figure VII. We observe the same behavior as in the original split. Our method outperforms the competitors, while our method without the semi-supervision and SIMILAR are competing for the second best.

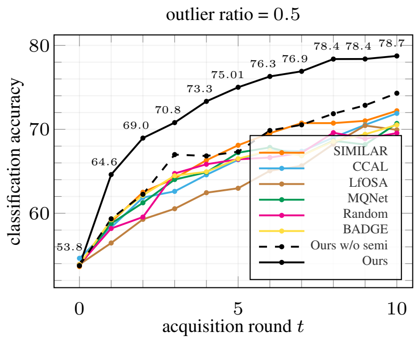

Figure VII. Comparison of classification accuracy over multiple active learning rounds on the additional inlier/outlier split of ImageNet when the initial labeled set L_0 contains 20 examples per inlier class and the budget size is equal to 500.

XI. Impact of pseudo-label accuracy on semi-supervision

Our method uses semi-supervision to improve the classification accuracy, and semi-supervision is dependent on the quality of pseudo-labels. In all our experiments it appears that the achieved level of pseudo-label accuracy is high enough to provide improvements. To study the case of lower pseudo-label accuracy, we perform an analysis without including the SSL initialization. In this case, the networks are randomly initialized.

In Figure VIII, we see that although pseudo-label accuracy is lower, our method with semi-supervision still significantly outperforms the version without semi-supervision. This shows us that even in the case of lower quality of pseudo-labels our method is still able to perform well.

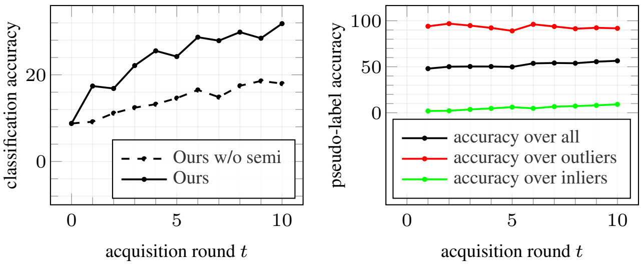

Figure VIII. Comparison of classification accuracy (left) over active learning rounds for our method with and without semi-supervision when networks are randomly initialized (SSL is skipped). The corresponding pseudo-label accuracy is shown on the right for all examples or for only inliers or outliers separately. Results on ImageNet with a 0.5 outlier ratio.

[3] The full list of the inlier/outlier classes and images used in the initial labeled and unlabeled sets will be released with the code.

References

[DZC+21] Pan Du, Suyun Zhao, Hui Chen, Shuwen Chai, Hong Chen, and Cuiping Li. Contrastive coding for active learning under class distribution mismatch. In *ICCV*, 2021. 2, 3

[HZRS16] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, 2016. 2

[KBKI21] Suraj Kothawade, Nathan Beck, Krishnateja Killamsetty, and Rishabh Iyer. Similar: Submodular information measures based active learning in realistic scenarios. In *NeurIPS*, 2021. 2

[KPKC21] Kwanyoung Kim, Dongwon Park, Kwang In Kim, and Se Young Chun. Task-aware variational adversarial active learning. In *CVPR*, 2021. 2

[LA16] Samuli Laine and Timo Aila. Temporal ensembling for semi-supervised learning. *ICLR*, 2016. 4

[PSB+22] Dongmin Park, Yooju Shin, Jihwan Bang, Youngjun Lee, Hwanjun Song, and Jae-Gil Lee. Meta-query-net: Resolving purity-informativeness dilemma in open-set active learning. In *NeurIPS*, 2022. 1

[WWL+21] Feng Wang, Guoyizhe Wei, Qiao Liu, Jinxiang Ou, Xian Wei, and Hairong Lv. Boost neural networks by checkpoints. In *NeurIPS*, 2021. 4

Method	acquisition round										
	0	1	2	3	4	5	6	7	8	9	10
Ours	45.76	55.47	60.53	63.15	65.22	66.05	68.69	69.46	70.48	71.97	72.22
	(±3.27)	(±1.98)	(±2.04)	(±1.48)	(±0.64)	(±0.95)	(±0.90)	(±0.81)	(±1.32)	(±0.53)	(±1.11)
Ours w/o semi	45.76	52.03	55.62	57.97	59.34	61.92	62.34	64.26	66.18	64.59	67.55
	(±3.27)	(±1.87)	(±1.10)	(±2.26)	(±1.25)	(±2.19)	(±1.41)	(±1.21)	(±0.83)	(±1.42)	(±1.16)
CCAL	45.79	50.67	53.70	55.41	56.99	58.83	60.27	61.12	62.67	64.53	65.90
	(±2.72)	(±1.11)	(±2.00)	(±2.17)	(±1.94)	(±1.19)	(±1.40)	(±1.53)	(±1.64)	(±0.26)	(±0.71)
LfOSA	44.96	48.29	51.84	54.02	55.28	55.92	58.11	59.52	60.03	62.37	63.06
	(±2.39)	(±2.12)	(±2.21)	(±0.74)	(±0.97)	(±2.04)	(±1.83)	(±1.40)	(±1.23)	(±1.60)	(±1.52)
MQNet	44.96	50.54	54.19	55.97	59.41	60.67	61.87	62.72	64.05	65.89	65.38
	(±2.39)	(±1.60)	(±1.43)	(±2.07)	(±1.45)	(±1.80)	(±2.09)	(±1.78)	(±1.32)	(±1.33)	(±0.91)
SIMILAR	45.42	51.17	55.74	57.23	60.64	61.06	62.58	62.53	64.61	65.95	65.87
	(±3.46)	(±2.99)	(±0.76)	(±1.76)	(±1.14)	(±0.75)	(±1.48)	(±1.57)	(±0.58)	(±1.36)	(±0.70)
Random	44.96	52.53	55.63	58.83	61.89	61.41	62.64	63.31	63.82	65.41	65.52
	(±2.39)	(±0.92)	(±1.43)	(±1.35)	(±1.42)	(±1.66)	(±1.17)	(±0.94)	(±1.53)	(±1.96)	(±1.03)
BADGE	44.96	52.05	56.13	58.51	60.19	61.55	63.17	64.18	64.99	67.02	68.59
	(±2.39)	(±2.64)	(±1.23)	(±1.92)	(±1.56)	(±1.16)	(±1.14)	(±1.71)	(±1.65)	(±0.78)	(±0.70)

(a) Results for 0.0 outlier ratio.

Method	acquisition round										
	0	1	2	3	4	5	6	7	8	9	10
Ours	41.12	54.85	60.03	62.38	65.71	66.91	67.42	69.87	72.02	72.10	73.22
	(±2.96)	(±2.61)	(±1.05)	(±0.84)	(±0.60)	(±0.66)	(±1.05)	(±0.74)	(±1.07)	(±1.26)	(±0.78)
Ours w/o semi	41.12	47.41	54.29	56.38	58.37	59.44	62.86	61.17	63.74	64.35	64.66
	(±2.96)	(±2.90)	(±2.10)	(±1.53)	(±2.33)	(±1.87)	(±2.16)	(±2.91)	(±2.01)	(±2.06)	(±3.09)
CCAL	41.71	45.89	49.82	54.51	55.41	57.49	59.28	61.41	63.12	62.43	64.51
	(±2.88)	(±2.97)	(±1.09)	(±1.25)	(±2.18)	(±1.38)	(±1.92)	(±0.91)	(±1.23)	(±1.79)	(±0.98)
LfOSA	41.26	46.61	48.74	50.74	52.78	54.19	56.66	58.77	60.10	61.71	62.74
	(±2.13)	(±3.65)	(±2.94)	(±1.02)	(±1.79)	(±2.47)	(±2.84)	(±1.79)	(±1.60)	(±1.18)	(±1.51)
MQNet	41.26	46.66	51.39	56.30	57.42	59.36	61.74	62.77	62.22	63.87	66.37
	(±2.13)	(±2.95)	(±3.71)	(±0.77)	(±2.36)	(±2.90)	(±0.63)	(±2.14)	(±2.24)	(±1.73)	(±1.77)
SIMILAR	41.41	47.68	50.75	54.78	58.00	58.96	60.59	61.81	63.79	63.95	64.98
	(±2.87)	(±3.22)	(±3.20)	(±1.16)	(±0.95)	(±1.79)	(±1.06)	(±2.15)	(±1.29)	(±1.21)	(±1.31)
Random	41.26	49.98	52.21	55.66	58.27	61.30	61.62	64.05	61.74	65.55	64.24
	(±2.13)	(±2.68)	(±1.67)	(±3.07)	(±3.66)	(±1.18)	(±2.26)	(±1.76)	(±2.33)	(±1.04)	(±1.99)
BADGE	41.26	48.72	51.78	55.89	58.03	60.93	61.54	63.84	64.75	65.52	66.45
	(±2.13)	(±2.89)	(±1.38)	(±1.84)	(±1.96)	(±2.70)	(±2.40)	(±1.65)	(±0.85)	(±0.92)	(±2.01)

(b) Results for 0.05 outlier ratio.

Method	acquisition round										
	0	1	2	3	4	5	6	7	8	9	10
Ours	40.66	52.66	57.47	61.65	63.90	66.13	67.28	69.57	70.64	71.73	71.36
	(±2.89)	(±1.79)	(±2.00)	(±0.74)	(±1.54)	(±1.55)	(±0.86)	(±1.39)	(±0.90)	(±1.37)	(±0.79)
Ours w/o semi	40.66	46.48	52.38	53.38	57.50	58.85	61.12	63.01	63.52	63.74	65.20
	(±2.89)	(±2.23)	(±2.21)	(±2.30)	(±1.82)	(±2.16)	(±1.22)	(±0.93)	(±1.88)	(±1.15)	(±1.59)
CCAL	40.05	46.61	52.88	54.08	56.59	56.90	59.07	60.29	61.15	63.26	61.84
	(±2.58)	(±1.45)	(±1.22)	(±0.49)	(±1.47)	(±2.42)	(±0.96)	(±2.14)	(±1.91)	(±1.43)	(±1.86)
LfOSA	40.94	43.74	48.35	50.13	50.83	53.71	56.94	57.20	59.22	61.01	62.62
	(±2.90)	(±3.19)	(±2.02)	(±1.43)	(±1.88)	(±1.92)	(±0.83)	(±2.63)	(±1.63)	(±2.08)	(±1.07)
MQNet	40.94	45.94	51.14	53.09	54.40	57.15	56.59	57.74	60.11	61.79	62.37
	(±2.90)	(±1.64)	(±0.48)	(±2.32)	(±1.80)	(±2.29)	(±1.57)	(±1.91)	(±3.91)	(±1.85)	(±2.03)
SIMILAR	40.91	47.38	51.95	53.90	55.34	57.97	61.41	62.16	60.96	61.87	63.58
	(±2.95)	(±1.88)	(±2.91)	(±2.23)	(±0.74)	(±1.51)	(±1.30)	(±1.40)	(±2.87)	(±1.72)	(±2.44)
Random	40.94	46.24	52.06	52.38	56.51	58.26	57.90	58.90	63.28	62.18	64.26
	(±2.90)	(±0.85)	(±2.42)	(±1.95)	(±3.02)	(±1.47)	(±2.63)	(±1.81)	(±0.62)	(±1.89)	(±1.47)
BADGE	40.94	46.69	51.15	52.99	56.66	58.78	58.96	61.02	61.81	63.36	63.81
	(±2.90)	(±2.63)	(±1.70)	(±1.99)	(±0.96)	(±1.48)	(±2.40)	(±3.09)	(±2.53)	(±2.31)	(±1.85)

(c) Results for 0.2 outlier ratio.

Table II. Mean and standard deviation for different methods on ImageNet dataset.

Method	acquisition round										
	0	1	2	3	4	5	6	7	8	9	10
Ours	41.98	52.66	56.82	61.44	64.30	65.94	68.88	69.55	71.07	71.10	73.06
	(±3.60)	(±1.68)	(±0.97)	(±2.37)	(±1.47)	(±1.27)	(±0.73)	(±1.34)	(±0.49)	(±0.63)	(±1.28)
Ours w/o semi	41.98	47.42	53.12	56.26	58.51	60.90	61.30	63.25	63.86	64.37	66.16
	(±3.60)	(±2.83)	(±2.07)	(±1.00)	(±3.63)	(±2.09)	(±1.13)	(±1.87)	(±3.89)	(±0.86)	(±2.33)
CCAL	43.79	47.76	50.45	51.70	54.14	56.80	59.79	59.55	61.09	60.22	63.12
	(±1.77)	(±2.08)	(±2.70)	(±1.84)	(±1.95)	(±2.44)	(±0.95)	(±2.17)	(±1.83)	(±2.18)	(±1.09)
LfOSA	41.76	45.87	47.97	51.94	51.87	55.36	56.02	56.05	59.42	60.38	61.94
	(±2.62)	(±3.44)	(±1.92)	(±1.31)	(±1.37)	(±2.60)	(±2.32)	(±2.85)	(±1.61)	(±0.97)	(±1.52)
MQNet	41.76	47.87	50.13	53.87	54.67	55.42	56.93	60.37	59.42	62.46	62.10
	(±2.62)	(±1.45)	(±3.47)	(±3.13)	(±2.08)	(±1.86)	(±2.67)	(±1.14)	(±1.56)	(±0.69)	(±2.18)
SIMILAR	41.84	47.65	52.91	54.93	57.20	58.56	59.09	61.47	63.33	63.90	65.14
	(±3.62)	(±0.99)	(±1.66)	(±2.26)	(±1.11)	(±2.22)	(±1.37)	(±1.27)	(±1.63)	(±0.55)	(±1.21)
Random	41.76	46.91	49.23	52.26	55.17	56.93	58.05	58.35	60.42	61.28	61.33
	(±2.62)	(±2.36)	(±1.66)	(±1.94)	(±1.30)	(±1.73)	(±1.80)	(±1.37)	(±0.69)	(±1.81)	(±2.28)
BADGE	41.76	47.60	49.76	51.55	54.50	56.45	57.82	56.80	59.94	61.78	61.89
	(±2.62)	(±1.56)	(±2.04)	(±4.35)	(±3.41)	(±2.25)	(±1.23)	(±3.02)	(±1.32)	(±2.09)	(±1.54)

(a) Results for 0.5 outlier ratio.

Method	acquisition round										
	0	1	2	3	4	5	6	7	8	9	10
Ours	41.12	48.48	52.77	56.69	61.15	62.96	63.57	66.43	67.3	67.68	68.98
	(± 2.96)	(±1.02)	(±1.86)	(±0.51)	(±1.57)	(±1.15)	(±1.08)	(±1.27)	(±1.92)	(±1.81)	(±1.28)
Ours w/o semi	41.12	43.1	48.7	49.92	54.98	56.58	56.83	59.95	61.58	61.57	63.02
	(± 2.96)	(±1.56)	(±1.35)	(±2.30)	(±1.81)	(±3.48)	(±2.70)	(±0.89)	(±1.32)	(±2.38)	(±1.30)
CCAL	41.71	45.57	47.01	49.10	48.66	50.90	51.47	53.38	55.07	54.66	53.26
	(±2.88)	(±1.60)	(±1.62)	(±1.50)	(±1.78)	(±1.42)	(±1.57)	(±1.05)	(±2.63)	(±1.17)	(±2.71)
LfOSA	41.26	44.13	47.39	48.75	51.26	52.22	54.51	54.59	56.94	58.03	60.05
	(±2.13)	(±2.63)	(±2.78)	(±1.86)	(±1.10)	(±1.56)	(±1.32)	(±2.02)	(±2.60)	(±3.09)	(±1.66)
MQNet	41.26	44.19	46.34	50.08	50.24	49.79	50.93	52.29	53.09	53.65	54.46
	(±2.13)	(±1.40)	(±1.28)	(±2.27)	(±1.74)	(±1.78)	(±2.12)	(±1.79)	(±1.47)	(±1.93)	(±1.43)
SIMILAR	41.41	46.69	48.85	50.11	54.13	55.22	58.94	55.92	59.34	59.47	61.84
	(±2.87)	(±2.25)	(±2.68)	(±1.30)	(±1.38)	(±0.94)	(±0.94)	(±3.43)	(±1.65)	(±1.68)	(±1.59)
Random	41.26	43.50	45.06	46.34	47.50	48.74	50.93	51.62	52.26	52.99	53.30
	(±2.13)	(±3.14)	(±2.04)	(±1.33)	(±2.06)	(±2.30)	(±2.79)	(±1.44)	(±1.04)	(±4.22)	(±1.52)
BADGE	41.26	44.72	45.89	47.10	46.54	49.15	51.65	49.49	52.66	53.14	55.09
	(±2.13)	(±1.31)	(±1.82)	(±1.85)	(±2.55)	(±1.62)	(±0.66)	(±3.42)	(±2.20)	(±1.93)	(±1.59)

(b) Results for 0.8 outlier ratio.

Method	acquisition round										
	0	1	2	3	4	5	6	7	8	9	10
Ours	41.12	48.00	50.00	53.15	56.21	58.35	61.30	62.00	64.80	64.78	65.22
	(±2.96)	(±0.85)	(±2.46)	(±1.79)	(±1.09)	(±2.26)	(±1.10)	(±0.25)	(±0.79)	(±0.72)	(±1.06)
Ours w/o semi	41.12	42.27	44.90	49.94	51.82	54.66	54.34	57.41	58.86	59.26	62.14
	(±2.96)	(±2.12)	(±1.84)	(±0.94)	(±2.08)	(±2.69)	(±2.88)	(±1.25)	(±2.09)	(±2.61)	(±0.84)
CCAL	41.71	43.38	45.23	45.82	47.52	47.66	47.65	47.78	49.23	51.38	50.45
	(±2.88)	(±1.99)	(±1.32)	(±2.30)	(±2.40)	(±2.00)	(±1.96)	(±1.38)	(±2.52)	(±1.41)	(±2.37)
LfOSA	41.26	44.19	45.20	48.88	49.73	51.12	51.90	52.72	55.52	56.16	57.90
	(±2.13)	(±2.18)	(±1.59)	(±1.51)	(±1.67)	(±1.28)	(±1.66)	(±1.77)	(±1.21)	(±1.10)	(±1.24)
MQNet	41.26	42.03	45.81	45.50	47.10	48.83	47.14	48.72	47.95	50.29	51.17
	(±2.13)	(±2.56)	(±2.56)	(±2.28)	(±1.52)	(±1.40)	(±1.13)	(±1.34)	(±4.34)	(±1.79)	(±2.36)
Random	41.26	42.51	43.30	43.60	45.86	45.70	46.83	47.62	48.88	49.81	51.70
	(±2.13)	(±1.89)	(±3.84)	(±2.21)	(±1.87)	(±2.38)	(±1.55)	(±0.98)	(±3.09)	(±1.87)	(±1.92)
BADGE	41.26	43.68	43.94	45.44	45.54	46.85	47.92	47.60	47.18	46.86	49.07
	(±2.13)	(±2.86)	(±2.45)	(±0.96)	(±0.45)	(±2.82)	(±1.32)	(±0.98)	(±2.56)	(±3.91)	(±1.59)

(c) Results for 0.9 outlier ratio.

Table III. Mean and standard deviation for different methods on ImageNet dataset.