Active Learning for Deep Object Detection via Probabilistic Modeling -Supplementary Material-

Jiwoong Choi[1,3], Ismail Elezi[2,3], Hyuk-Jae Lee[1], Clement Farabet[3], and Jose M. Alvarez[3]
[1]Seoul National University, [2]Technical University of Munich, [3]NVIDIA
{jwchoi, hjlee}@capp.snu.ac.kr, ismail.elezi@tum.de, {cfarabet, josea}@nvidia.com

A. Parameter Sensitivity

A.1. Accuracy as a function of K

In the main paper, we presented experiments using K=4 as the number of components in the mixture model. In Tab. 1 we analyze the sensitivity of our results with respect to the number of components in the GMM. Specifically, we provide numbers for K=1, 2, 4, and 8. As in the main paper, we repeat the experiment three times and provide the average mAP and standard deviation for the normal (IoU>0.5) and the strict metric (IoU>0.75). We also provide the number of parameters and the forward time for each of these instantiations. As shown, the accuracy of K=1 is much lower than that of K=4, especially for IoU>0.75. Moreover, for K=1, epistemic uncertainty cannot be estimated (see Eq.1 in the main paper). The accuracy remains stable for other configurations with minor variations in mAP. However, as the number of parameters is proportional to K, there are significant variations in terms of the number of parameters and forward time. Given these results, we selected K=4 as a good trade-off between accuracy (normal and strict metric) and computing cost. In practice, the larger K, the more difficult to train the GMM due to fluctuation. This would be the reason for a drop in accuracy when K=8.

K	mAP (%)		# of parameters (M)	Forward time (sec)
# of mixture	IoU>0.5	IoU>0.75		
1	69.89±0.23	45.18±0.24	29.6	0.021
2	70.29±0.29	45.98±0.38	37.6	0.025
4	70.19±0.36	46.11±0.38	52.3	0.031
8	70.01±0.29	45.69±0.28	81.8	0.051

Table 1: **VOC07:** mAP and computing cost as a function of the number of components in the mixture model. Model parameters in millions (M) and forward time in seconds (sec).

A.2. Accuracy as a function of input image resolution

In order to check for the robustness of our method with respect to the image size, here we compare the performance of the network trained using higher resolution images (512×512). The experiment is analogous to the experiment we showed in Tab.1a in the main paper. We compare the results of SSD [3], with the results of our method. As we can see in Tab. 2, as expected, increasing the resolution of the input image yields a significant improvement in mAP score for all the methods. For high-resolution input images, our method outperforms SSD in the normal metric (IoU>0.5) by 0.51 percent points (*pp*), and shows significant improvement when evaluated in the strict metric (IoU>0.75), with an improvement of 2.49 *pp*. That is, our method is notably better in those scenarios where we need a higher intersection between the predicted bounding box and the ground truth.

Method	SSD 512 (512×512)		SSD 300 (300×300)	
	IoU>0.5	IoU>0.75	IoU>0.5	IoU>0.75
SSD [3]	73.22±0.35	45.74±0.70	69.29±0.51	43.36±1.24
$Ours_{gmm}$	73.50±0.12	48.23±0.53	70.19±0.36	46.11±0.38
$Ours_{eff}$	73.73±0.16	48.12±0.33	70.45±0.06	46.18±0.26

Table 2: **VOC07:** mAP (in %) as a function of the resolution of the input image.

A.3. Accuracy as a function of budget number in active learning

In the main paper, we used a budget of 1k following the setup of [6] to enable direct comparison on VOC07+12. In order to check for the mAP with respect to a budget number in active learning, we further compare the mAP for cases of 9k, 3k, and 1k as the budget number. We summarize the results of this experiment in Tab. 3. As in the main paper, we report the performance using the average of mAP and standard deviation for three independent trials. As shown in the last active learning iteration, as expected, we can see that the smaller the budget number yields a higher accuracy improvement in active learning.

B. More visual examples selected by our approach

Fig. 1 shows more representative examples selected by our active learning approach. Each uncertainty value (bold

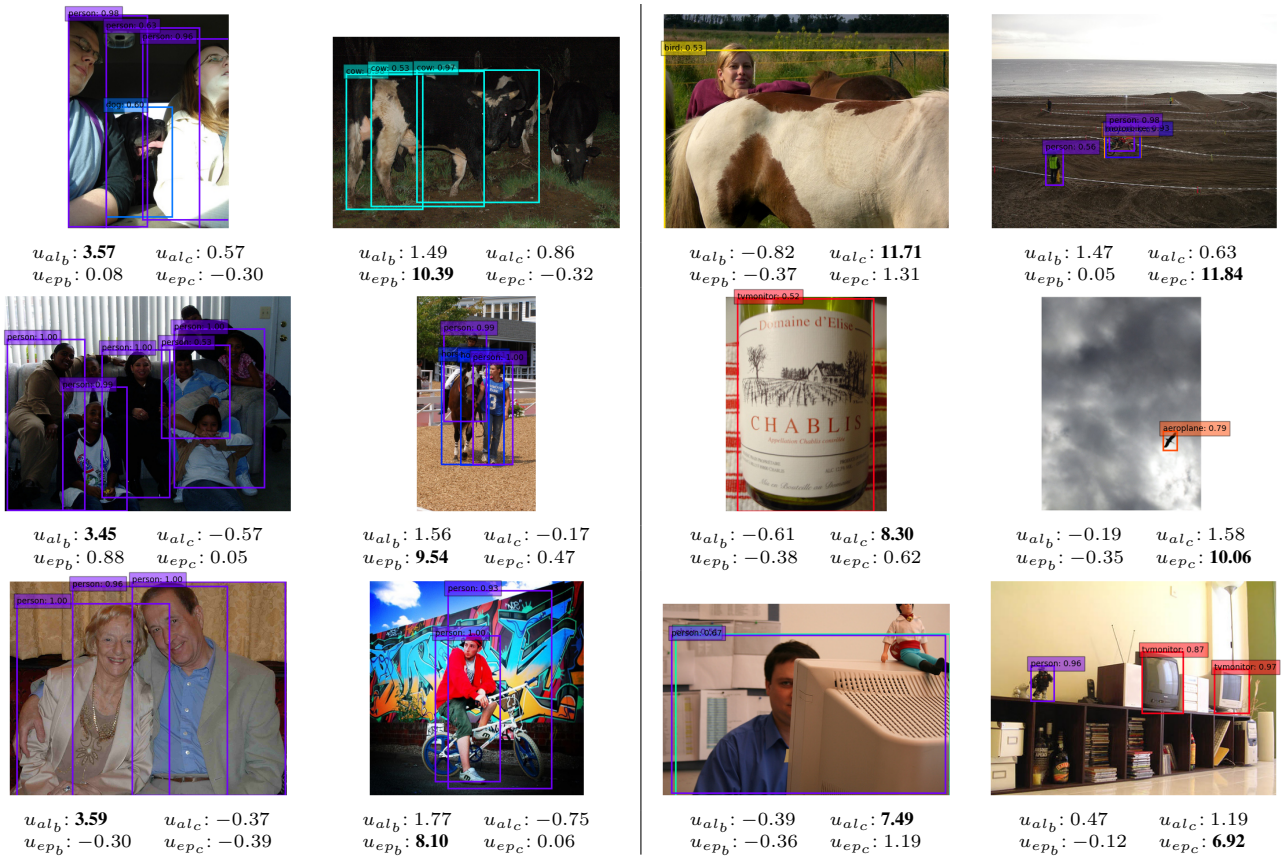

Figure 1: Examples of aleatoric and epistemic uncertainties for inaccurate detections. Best viewed in screen.

# of images	Budget number		
	9k	3k	1k
1k	0.5254±0.0017	0.5254±0.0017	0.5254±0.0017
2k	-	-	0.6130±0.0040
3k	-	-	0.6556±0.0051
4k	-	0.6797±0.0011	0.6843±0.0043
5k	-	-	0.7077±0.0019
6k	-	-	0.7252±0.0027
7k	-	0.7335±0.0049	0.7352±0.0025
8k	-	-	0.7453±0.0025
9k	-	-	0.7509±0.0014
10k	0.7493±0.0043	0.7550±0.0012	0.7598±0.0021

Table 3: **VOC07+12:** mAP as a function of the budget number in active learning.

numbers in Fig. 1) provides a different insight into some particular failure. From left to right and top to bottom: One of the several bounding boxes detected as person is false positive; One of the several bounding boxes detected as cow is false positive; A horse is misclassified as a bird; A motorbike is misclassified as bicycle; One of the several bounding boxes detected as person is false positive; One of the several bounding boxes detected as horse is false positive; A bottle is misclassified as a TV/monitor; A bird is misclassified as an aeroplane; One of the several bounding boxes detected as person is false positive; One of the several bounding boxes detected as person is false positive; A person is misclassified as a chair; A toy (not in the PASCAL VOC dataset) is misclassified as a person.

C. Values in the plots of VOC07+12

In the main paper, we present plots for active learning results using VOC07+12 in Fig. 4 and Fig. 5. Tab. 4, Tab. 5, and Tab. 6 summarize the actual numbers used to create the plots. As mentioned in the paper, in Tab. 4, numbers corresponding to Random [3] , Entropy [4], Core-set [5], and LLAL [6] are taken from [6]. For MC-dropout [1], to further verify the influence in the number of forward passes, we include two instances: using 25 (the one included in the main paper) and 50 forward passes. As we can see in Tab. 5, the variation in accuracy for these two approaches is negligible while the compute needed is significantly larger for the one using 50 forward passes.

D. Discussion of the classification loss

In addition to Eq. 5 and Eq. 9 in the main paper (called *Type-1* loss), we can train the proposed object detection net-

# of labeled images	Random [3]	Entropy [4]	Core-set [5]	LLAL [6]	$Ours_{gmm}$	$Ours_{eff}$
1k	0.5262±0.0062	0.5262±0.0062	0.5262±0.0062	0.5238±0.0028	0.5254±0.0017	0.5254±0.0017
2k	0.6082±0.0019	0.6123±0.0081	0.6236±0.0052	0.6095±0.0042	0.6130±0.0040	0.6121±0.0050
3k	0.6423±0.0022	0.6357±0.0091	0.6590±0.0043	0.6491±0.0047	0.6556±0.0051	0.6657±0.0027
4k	0.6633±0.0018	0.6694±0.0021	0.6763±0.0021	0.6690±0.0028	0.6843±0.0043	0.6849±0.0014
5k	0.6751±0.0017	0.6870±0.0015	0.6888±0.0048	0.6905±0.0045	0.7077±0.0019	0.7073±0.0012
6k	0.6860±0.0050	0.6982±0.0011	0.6944±0.0032	0.7035±0.0055	0.7252±0.0027	0.7185±0.0016
7k	0.6927±0.0016	0.7018±0.0027	0.7016±0.0013	0.7149±0.0066	0.7352±0.0025	0.7318±0.0045
8k	0.7010±0.0017	0.7112±0.0012	0.7083±0.0012	0.7213±0.0060	0.7453±0.0025	0.7429±0.0044
9k	0.7044±0.0047	0.7166±0.0031	0.7115±0.0016	0.7273±0.0030	0.7509±0.0014	0.7483±0.0028
10k	0.7117±0.0016	0.7222±0.0024	0.7171±0.0025	0.7338±0.0028	0.7598±0.0021	0.7584±0.0026

Table 4: **VOC07+12:** Comparison to published work using a single model for scoring.

# of labeled images	MC-dropout [1] (50 fwd)	MC-dropout [1] (25 fwd)	Ensemble [2]	$Ours_{gmm}$	$Ours_{eff}$
1k	0.5235 ± 0.0004	0.5235±0.0004	0.5254±0.0017	0.5254±0.0017	0.5254±0.0017
2k	0.6059 ± 0.0026	0.6059±0.0028	0.6020±0.0093	0.6130±0.0040	0.6121±0.0050
3k	0.6660 ± 0.0023	0.6690±0.0030	0.6570±0.0099	0.6556±0.0051	0.6657±0.0027
4k	0.6890 ± 0.0018	0.6840±0.0019	0.6920±0.0034	0.6843±0.0043	0.6849±0.0014
5k	0.7060 ± 0.0045	0.7080±0.0041	0.7150±0.0018	0.7077±0.0019	0.7073±0.0012
6k	0.7200 ± 0.0012	0.7190±0.0050	0.7290±0.0027	0.7252±0.0027	0.7185±0.0016
7k	0.7367 ± 0.0015	0.7381±0.0003	0.7429±0.0004	0.7352±0.0025	0.7318±0.0045
8k	0.7468 ± 0.0027	0.7475±0.0056	0.7491±0.0041	0.7453±0.0025	0.7429±0.0044
9k	0.7549 ± 0.0013	0.7558±0.0023	0.7589±0.0025	0.7509±0.0014	0.7483±0.0028
10k	0.7567 ± 0.0048	0.7601±0.0018	0.7590±0.0032	0.7598±0.0021	0.7584±0.0026

Table 5: **VOC07+12:** Accuracy comparison to MC-dropout and ensemble. For MC-dropout, we include two instances: using 25 forward passes and using 50 forward passes.

	SSD [3]	Ensemble [2]	MC-dropout [1]	$Ours_{gmm}$	$Ours_{eff}$
# of parameters (M)	26.29	78.87	26.29	52.35	41.12
Forward time (sec)	0.023	0.069	0.412	0.031	0.029

Table 6: **VOC07+12:** Model parameters in millions (M) and forward time in seconds (sec) using a resolution of 300 × 300 for the input image and $K = 4$.

work with the following classification loss:

$$L_{cl}^{Pos}(y,c) = -\sum_{i\in Pos}^{N} y_G^i \, log \sum_{k=1}^{K} \pi^{ik} Softmax(\hat{c}_p^{ik})$$
$$L_{cl}^{Neg}(y,c) = -\sum_{i\in Neg}^{M\times N} y_0^i \, log \sum_{k=1}^{K} \pi^{ik} Softmax(\hat{c}_p^{ik}), \quad (1)$$

where y_G and y_0 are one-hot vectors having 1 in ground-truth class G and background class 0, respectively. The remaining parameters are the same as Eq. 5 in the main paper. The parameters of the loss for $Ours_{eff}$ are the same as Eq. 1 except for the class probability $\hat{\mu}_p^{ik}$. For *Type-1* loss, the weights tend to be concentrated in one of the mixture distributions in training. For Eq. 1 (called *Type-2* loss), however, this trend tends to be alleviated. Tab. 7 and Tab. 8 show the results of active learning of two classification losses on VOC07 and MS-COCO, respectively. As shown, there is no significant difference in the accuracy of the two loss functions on VOC07, but there is a large difference on MS-COCO. Although the *Type-2* loss mitigates the weight bias compared to the *Type-1* loss, the accuracy improvement is not sufficient for larger dataset with more classes. For this reason, in the main paper, we show the experimental results based on the *Type-1* loss, but a study on a classification loss design that can improve the overall active learning performance while resolving the weight bias is needed in the future.

Model	Cls. loss	mAP in % (# images)		
		1st (2k)	2nd (3k)	3rd (4k)
$Ours_{gmm}$	Type-1	62.43±0.10	67.32±0.12	69.43±0.11
	Type-2	62.64±0.21	67.20±0.22	69.40±0.14
$Ours_{eff}$	Type-1	62.91±0.16	67.61±0.17	69.66±0.17
	Type-2	62.23±0.25	67.30±0.23	69.57±0.16

Table 7: Ablation study of classification losses on **VOC07**.

Model	Cls. loss	mAP in % (# images)		
		1st (5k)	2nd (6k)	3rd (7k)
$Ours_{gmm}$	Type-1	27.70±0.08	29.28±0.05	30.51±0.12
	Type-2	27.38±0.16	28.69±0.22	29.55±0.14
$Ours_{eff}$	Type-1	27.33±0.04	29.06±0.08	30.02±0.05
	Type-2	27.46±0.13	28.32±0.32	29.21±0.14

Table 8: Ablation study of classification losses on **MS-COCO**.

References

[1] Di Feng, Xiao Wei, Lars Rosenbaum, Atsuto Maki, and Klaus Dietmayer. Deep active learning for efficient training of a lidar 3d object detector. In *IEEE Intelligent Vehicles Symposium (IV)*, 2019.

[2] Elmar Haussmann, Michele Fenzi, Kashyap Chitta, Jan Ivanecky, Hanson Xu, Donna Roy, Akshita Mittel, Nicolas Koumchatzky, Clement Farabet, and Jose M Alvarez. Scalable active learning for object detection. In *IEEE Intelligent Vehicles Symposium (IV)*, 2020.

[3] Wei Liu, Dragomir Anguelov, Dumitru Erhan, Christian Szegedy, Scott Reed, Cheng-Yang Fu, and Alexander C Berg. Ssd: Single shot multibox detector. In *European Conference on Computer Vision (ECCV)*, 2016.

[4] Soumya Roy, Asim Unmesh, and Vinay P. Namboodiri. Deep active learning for object detection. In *British Machine Vision Conference (BMVC)*, 2018.

[5] Ozan Sener and Silvio Savarese. Active learning for convolutional neural networks: A core-set approach. In *International Conference on Learning Representations (ICLR)*, 2018.

[6] Donggeun Yoo and In So Kweon. Learning loss for active learning. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019.