Localization-Aware Active Learning for Object Detection: Supplementary Materials

Chieh-Chi Kao[1], Teng-Yok Lee[2], Pradeep Sen[1], and Ming-Yu Liu[2]

[1] University of California, Santa Barbara, Santa Barbara, CA 93106, USA
[2] Mitsubishi Electric Research Laboratories, Cambridge, MA 02139, USA

This document includes the data and analysis of the proposed methods that are not covered in the main paper due to the spcae limitation. We first define the abbreviation for all methods as following:

Abbreviation	Method
R	Random
C	Classification
LS	Localization Stability
LS+C	Localization Stability and Classification
LT/C	Localization Tightness and Classification
LT/C(GT)	Localization Tightness and Classification with Ground Truth
3in1	Localization Stability, Localization Tightness, and Classification

This abbreviation is used in all the text, figures, and tables in this document.

1 Design of Localization-Tightness Metric

Given the measurement of localization tightness, we need to design a metric to utilize it for active learning. The most intuitive way is to use the localization tightness alone to decide the score for each box. However, in our experiments it does not help for selecting samples to annotate. We further analyze it by showing the images selected by different methods as shown in Fig. 1. When using only localization tightness as the cue to calculate the score of each detected box for active learning, it tends to find images (Fig. 1, first row) that have tiny objects (e.g., airplane, bird), which are not chosen that often by other methods (Fig. 1, second row). However, these classes are easier ones that the detector already does well so that the overall performance of using localization tightness alone is worse than other metrics.

Based on the observations in Sec. 3.2 in the main paper, we would like to find images contain boxes that have disagreement in classification and localization results. When designing a metric using localization tightness, there are two important qeustions: "How to define the score for an image with detected boxes?" and "How to define the score for a detected box?" For the first question, two methods have been tested: using the lowest score of all boxes (min(.)), and using a weighted sum of all boxes (wsum(.)), where the weight is P_{max} of each box. For the second question, different metrics have been tested as following, where P ($P_{max}(B)$ in the main paper) is the highest probability out of

Fig. 1: **First row:** Example images selected for annotation by the method using information from localization only to evaluate the score of each box. **Second row:** Example images selected for annotation by the method using classification uncertainty only (C).

K categories of box B, and T ($T(B)$ in the main paper) is the localization tightness of box B. For a set of unlabeled images, the following methods choose images with lower scores to annotate in active learning.

min(|T+P-1|) This metric is the one (**LT/C**) we used in the main paper. It selects images with boxes that have disagreement between classification and localization results. It also picks images contain boxes that are not very certain in both classification and localization results.

min(-|P-T|) Different from LT/C, this metric only selects images with boxes that have disagreement between classification and localization results. It does not select boxes that are not very certain in both two outputs.

wsum(|T+P-1|) This method uses the same metric as LT/C to evaluate the score of each box. However, instead of using the highest score out of all boxes as the score of an imgae, it uses a weighted sum across all boxes.

wsum(T) This method uses only the information from localization outputs when deciding the score of each box. Images with boxes that have low localization tightness will be chosen by this method.

Fig. 2 shows the mean average precision (mAP) curves of different metrics using localization tightness, and the experimental setup is the same as mentioned in Sec. 4.2 in the main paper. The proposed LT/C outperforms the rest metrics clearly at the first half of the experiment. Among the second half, LT/C is still the best among all metrics, but the gap between LT/C and the others becomes smaller.

The difference between LT+C and max(|P-T|) is selecting images with boxes that are both uncertain in classification and localization outputs. We hypothesize that images with uncertainty in both outputs are more informative, which make LT/C better than

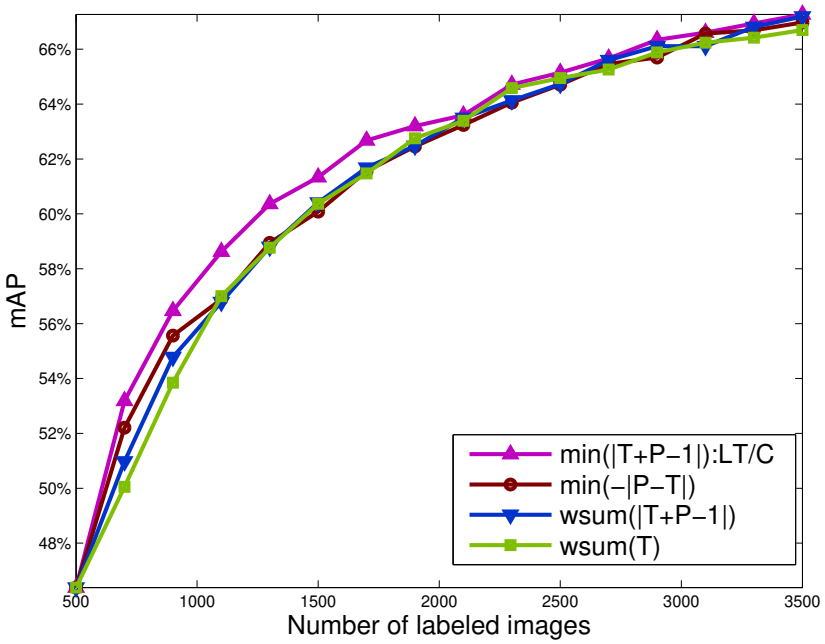


Fig. 2: Mean average precision curve of different metrics of localization tightness on PASCAL 2007 detection dataset. Each point in the plot is an average of 5 trials.

max(|P-T|). Also, given the same metric for calculating the score of a detected box, LT/C and wsum(|T+P-1|) use different strategy to define the score of an image. The overlapping ratio of images sampled by these two methods is only 17.9% (an average over 5 trials), which implies that how to define the score of an image greatly affects the sampling process.

2 Discussion of Extreme Cases

There could be extreme cases that the proposed methods may not be helpful. For instance, if perfect candidate windows are available (LT/C), or feature extractors are resilient to Gaussian noise (LS+C).

If we have very precise candidate windows, which means that we need only the classification part and it is not a detection problem anymore. While this might be possible for few special object classes (e.g. human faces), to our knowledge, there is no perfect region proposal algorithms that can work for all type of objects. As shown in our experiments, even state-of-the-art object detectors can still incorrectly localize objects. Furthermore, when perfect candidates are available, the localization tightness will always be 1, and our LT/C degenerates to classification uncertainty method (C), which can still work for active learning.

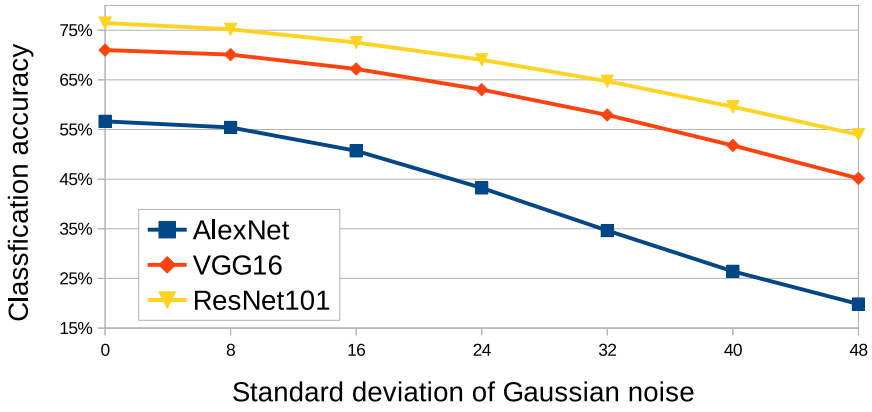


Fig. 3: Top-1 classification accuracy of different neural network models when input images are corrupted by Gaussian noise on PASCAL 2012 validation dataset.

Also, we have tested the resiliency to Gaussian noise of state-of-the-art feature extractors (AlexNet, VGG16, ResNet101). Classification task on the validation set of ImageNet (ILSVRC2012) is used as the testbed. The results demonstrate that none of these state-of-the-art feature extractors is resilient to noise. Moreover, if the feature extractor is robust to noise, the localization stability will always be 1, and our LS+C degenerates to classification uncertainty method (C), which can still work for active learning.

We have tested the resiliency to Gaussian noise of state-of-the-art feature extractors (AlexNet, VGG16, ResNet101). Classification task on the validation set of ImageNet (ILSVRC2012) is used as the testbed. Pre-trained models are used as the classifier and input images are corrupted by Gaussian noise of different levels. Fig. 3 shows the top-1 classification accuracy under different standard deviation of Gaussian noise. With the largest standard deviation, the accuracy can drop 23-37%. It demonstrates that none of these state-of-the-art feature extractors is resilient to noise. Goodfellow et. al [1] also hypothesized that NNs with non-linear modules (e.g., sigmoid) mainly work in linear region, could be vulnerable to local perturbation such as Gaussian noise.

3 Full Experimental Results

In this section, the full results from the experiments of active learning methods on the PASCAL and MS COCO datasets are presented. These results are not covered in the main paper due to the easiness of reading and space constraint.

Results of Using Localization Stability Only: As an ablation experiment, results for the method using localization stability only (LS) are added into the plot of mAP curves and the table of classwise APs. Table 1 and Table 2 show the average precision for each method after 3 rounds of active learning on the PSACAL 2012 validation and PASCAL

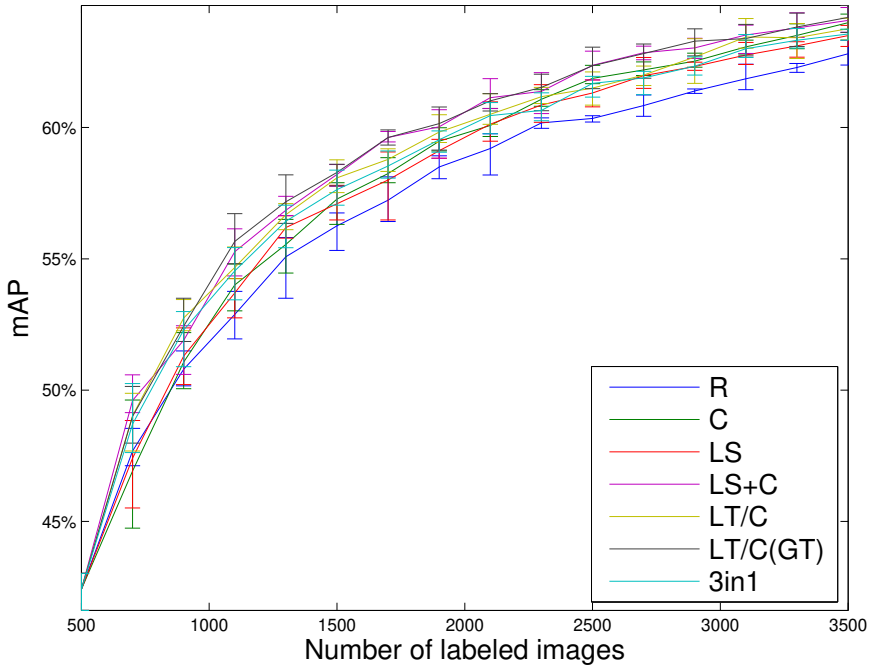


Fig. 4: Mean average precision curve of different active learning methods on the PASCAL 2012 detection dataset. Each point in the plot is an average of 5 trials. The error bars represent the minimum and maximum values out of 5 trials at each point. This is a full version (LS and 3in1 added) of Fig. 5a in the main paper.

2007 testing set. Fig. 4 and Fig. 5 show the mAP curves of each active learning method on the PASCAL 2012 and 2007 datasets. Each point in the plot is an average of 5 trials. Also, error bars that represent the minimum and maximum values out of 5 trials are added at each point to show the distribution of 5 trials. Fig. 6a and Fig. 6b show the relative saving in labeled images of each active learning method on the PASCAL 2012 and 2007 datasets. As shown in Fig. 6a and Fig. 6b, LS outperforms the random sampling for the most cases. Also, combining the localization stability with the classification uncertainty (LS+C) works better than using either only the localization stability (LS) or classification uncertainty (C).

Results of Using 3 Cues: In order to see that if the localization-uncertainty measurements have complementary information, we further combine all cues for selecting informative images. As images with high classification uncertainty, low localization stability, and low localization tightness should be selected for annotation, the score of the i-th image (I_i) image is defined as follows: $U_C(I_i) - \lambda_{ls} S_I(I_i) - \lambda_{lt} T_I(I_i)$ where λ_{ls} and λ_{lt} are set to 1 across all the experiments in this paper.

method	aero	bike	bird	boat*	bottle*	bus	car	cat	chair*	cow	table*
R	71.1	61.5	54.7	28.4	32.0	68.1	57.9	75.4	25.8	44.2	36.4
C	70.7	62.9	54.7	25.5	30.8	66.1	56.2	**78.1**	26.4	**54.5**	36.7
LS	**75.1**	61.3	**57.6**	**34.7**	35.1	65.1	58.2	75.4	29.3	43.9	38.5
LS+C	73.9	63.7	56.9	29.6	**35.2**	66.5	58.5	77.9	31.3	50.8	40.7
LT/C	69.8	**64.6**	54.6	29.5	33.8	**70.3**	**59.7**	75.5	29.5	46.3	**41.8**
3in1	72.9	63.8	52.7	29.5	33.6	66.4	57.2	76.0	**31.5**	48.5	41.6

method	dog	horse	mbike	persn	plant*	sheep	sofa	train	tv	mAP
R	73.0	61.9	67.3	68.1	21.6	51.9	41.0	**65.5**	51.7	52.9
C	**76.9**	**68.3**	67.7	67.4	22.5	57.7	40.8	63.6	52.5	54.0
LS	70.7	57.5	66.1	68.5	23.0	56.1	40.3	64.2	53.6	53.7
LS+C	73.8	65.4	66.9	68.4	**24.8**	**58.0**	**44.9**	64.2	53.9	**55.3**
LT/C	73.0	62.5	**69.0**	**70.8**	23.2	56.5	42.8	64.3	55.9	54.7
3in1	72.2	62.6	67.6	68.8	24.5	57.6	43.6	63.0	**57.1**	54.5

Table 1: Average precision for each method on the PASCAL 2012 validation set after 3 rounds of active learning (the number of labeled images in the training set is 1,100). This is a full version (LS and 3in1 added) of Table 1 in the main paper. All the experimental settings are the same with Table 1 in the main paper.

method	aero	bike	bird	boat*	bottle*	bus	car	cat	chair*	cow	table*
R	61.6	67.2	54.1	40.0	33.6	64.5	73.0	73.9	34.5	60.8	52.2
C	56.9	68.0	54.9	36.8	34.4	68.1	71.7	**75.5**	34.0	**68.6**	51.0
LS	**64.4**	63.9	**56.3**	**45.1**	38.0	65.5	73.7	71.2	38.6	62.7	57.0
LS+C	61.5	64.4	55.8	40.2	38.7	66.3	73.8	74.7	39.6	68.0	56.3
LT/C	57.6	**69.7**	52.9	41.1	38.4	**69.7**	74.4	71.8	36.4	61.2	58.1
3in1	57.6	65.1	53.3	37.1	**39.0**	68.0	**74.6**	73.9	**39.8**	64.9	**58.5**

method	dog	horse	mbike	persn	plant*	sheep	sofa	train	tv	mAP
R	69.3	**74.7**	66.6	67.1	25.9	52.1	54.2	66.1	54.9	57.3
C	71.4	74.7	65.2	65.9	24.9	60.0	53.9	63.0	57.4	57.8
LS	67.6	69.0	64.6	67.1	**29.6**	56.2	57.3	**68.6**	53.6	58.5
LS+C	**71.5**	73.8	67.2	66.7	27.7	**61.3**	57.0	65.6	57.4	**59.4**
LT/C	69.5	74.3	66.2	**67.8**	28.0	55.5	56.3	65.5	58.2	58.6
3in1	70.4	73.7	**67.3**	67.3	27.4	59.9	**58.0**	65.1	**59.2**	59.0

Table 2: Average precision for each method on the PASCAL 2007 testing set after 3 rounds of active learning (the number of labeled images in the training set is 1,100). This is a full version (LS and 3in1 added) of Table 2 in the main paper. All the experimental settings are the same with Table 2 in the main paper.

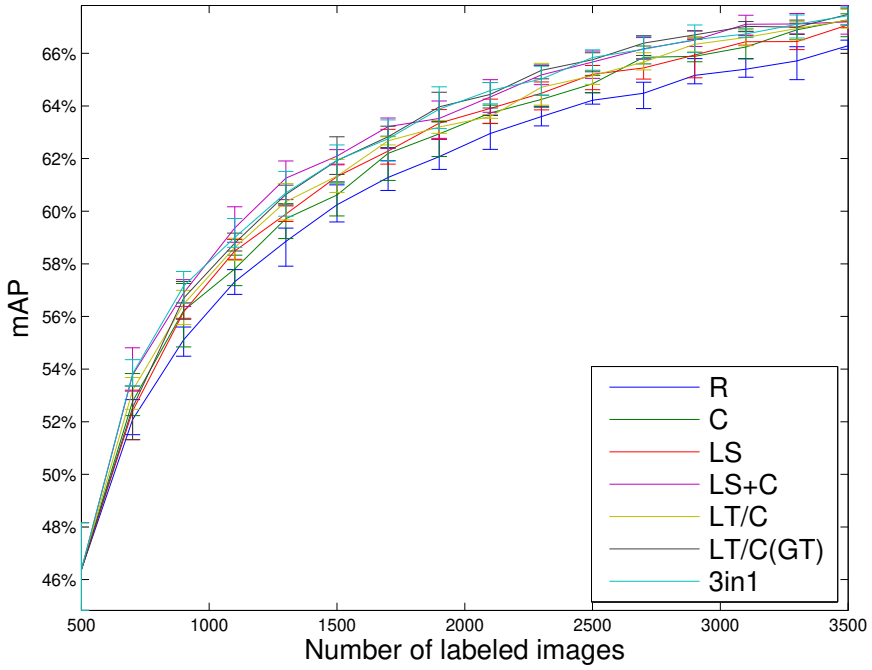
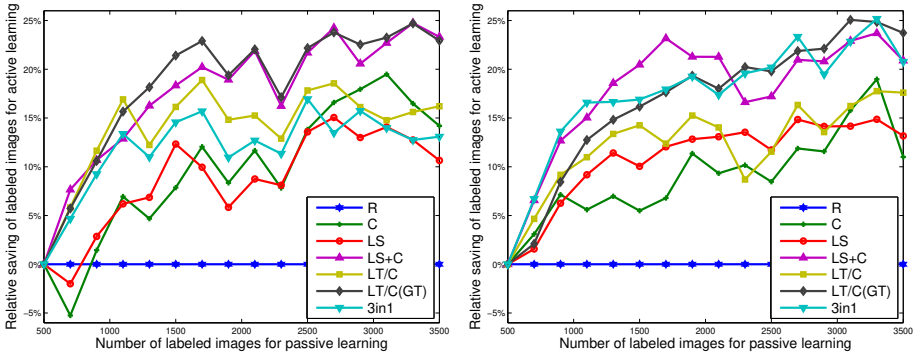


Fig. 5: Mean average precision curve of different active learning methods on the PASCAL 2007 detection dataset. Each point in the plot is an average of 5 trials. The error bars represent the minimum and maximum values out of 5 trials at each point. This is a full version (LS and 3in1 added) of Fig. 7a in the main paper.

On PASCAL 2012, combining all cues together does not work better than either LS+C or LT/C (Fig. 6a). On PASCAL 2007, 3in1 is compatible with LS+C, and better than LT/C (Fig. 6b). It seems that localization-uncertainty measurements do not have complementary information. We further analyze the overlapping ratio between images chosen by different active learning methods in Table 3 and Table 4. When we compare the overlapping ratio between 3in1 and three other metrics (C, LS, LT/C), both C and LS have an overlapping ratio around 30%, but LT/C has only about 10%. This implies that among the three cues, LT/C provides the least information in 3in1 method. We notice that the images chosen by 3in1 method are highly overlapped with LS+C (over 60%), but 3in1 does not outperform LS+C. Our hypothesis is that the images (about one third of total images) chosen differently by 3in1 and LS+C make this difference in performance.

mAP Plots with Error Bars: In the original mAP plots of the FRCNN on the MS COCO dataset (Fig. 8a in the main paper) and the SSD on the PASCAL 2007 dataset (Fig. 9a in the main paper), only the average of multiple trials is plotted. Here we add the error bars that represent the minimum and maximum values of multiple trials to the plot. This



(a) PASCAL 2012 (b) PASCAL 2007

Fig. 6: Relative saving of labeled images for different active learning methods on the (a) PASCAL 2012 validation dataset and (b) PASCAL 2007 testing set. (a) and (b) are full versions (LS and 3in1 added) of Fig. 5b and Fig. 7b in the main paper.

Table 3: Overlapping ratio between 200 images chosen by different active learning methods on the PASCAL 2012 dataset after the first round of active learning. Each number shown in the table is an average over 5 trials.

Method	R	C	LS	LS+C	LT/C
C	3.5%				
LS	4.0%	2.7%			
LS+C	4.4%	34.7%	34.6%		
LT/C	5.0%	5.9%	2.4%	5.2%	
3in1	4.6%	30.4%	25.7%	62.4%	8.8%

shows the distribution of the result from different trials. Fig. 7 and Fig. 8 show the mAP curves of the FRCNN on the MS COCO dataset and the SSD on the PASCAL 2007 dataset. Three methods (R, C, and LS+C) are tested in these two experiments.

4 Visualization of The Selection Process

The most popular metric used for measuring the performance of an object detector is mAP. We also use this metric to evaluate the performance of different active learning methods. If one active learning method selects more informative images to label and add them into the training set, the detector trained on this set will have a higher mAP. Besides this final numerical result, we are curious about what images are chosen in the selection process by different active learning methods, and how these chosen images are related to the average precision.

In order to visualize the selection process, we first visualize the PASCAL 2012 training set [2] by using t-Distributed Stochastic Neighbor Embedding (t-SNE) [3].

Table 4: Overlapping ratio between 200 images chosen by different active learning methods on the PASCAL 2007 dataset after the first round of active learning. Each number shown in the table is an average over 5 trials.

Method	R	C	LS	LS+C	LT/C
C	4.1%				
LS	4.2%	3.5%			
LS+C	4.3%	34.0%	39.7%		
LT/C	5.6%	5.9%	4.5%	5.7%	
3in1	3.9%	30.5%	32.0%	65.3%	12.0%

After knowing the distibution of the PASCAL 2012 training set, we further visualize the chosen images in the selection process by different active learning methods.

Visualization of the PASCAL 2012 Dataset: We first visualize the PASCAL 2012 training set (5,717 images) by using t-SNE with VGG16 model [4]. t-SNE is a technique for dimensionality reduction that is tailored for visualizing high-dimensional datasets. Features extracted from the conv5_3 layer are used as the high-dimensional vector for each image in the PASCAL 2012 training set. The visualization of the PASCAL 2012 training set by embedding each image to a point on the 2D plane is shown in Fig. 9. Each data point in Fig. 9 represents one image in the dataset. Images with objects from only one class are represented by markers other than dots. Note that there might be objects belong to different classes shown in one image. Red dots ($>$1cls) are used for representing those images. For each class, there is a certain region that images locate at. For example, images of aeroplanes (orange plus signs) are located at the top-right part, and images of cats (green squares) are located at the bottom-center part.

For those images have objects from muliple classes, we cannot tell what classes are included in each of them from Fig. 9. Therefore, another visualization is shown in Fig. 13 by considering whether one image has objects from a certain class or not. For example, each orange plus sign in Fig. 13a represents an image which has at least one aeroplane in it, and each black dot represents an image that has no aeroplane in it. Given Fig. 9 and Fig. 13, we now have a better understanding about the distribution of the dataset, and the relationship between different classes. For example, in the left part of the scatter plot in Fig. 9, we notice that there are many images that have objects belong to multiple classes (red dots). From Fig. 13, we know that these images may contain people, chairs, tables, sofas, bottles, plants, and TVs. Actually, these images are regular scenes in a living room, just like the 4 images shown in Fig. 9. With these information, we can further analyze the selection process of different active learning methods.

Visualization of Different Active Learning Methods: We would like to visualize the selection process of different active learning methods. The experimental settings are the same with Sec. 4.1 in the main paper. For the analysis and visualization in this section, we only use one trial instead of using the average of 5 trials for the easiness of reading. The baseline FRCNN detector [5] is trained on a training set of 500 labeled

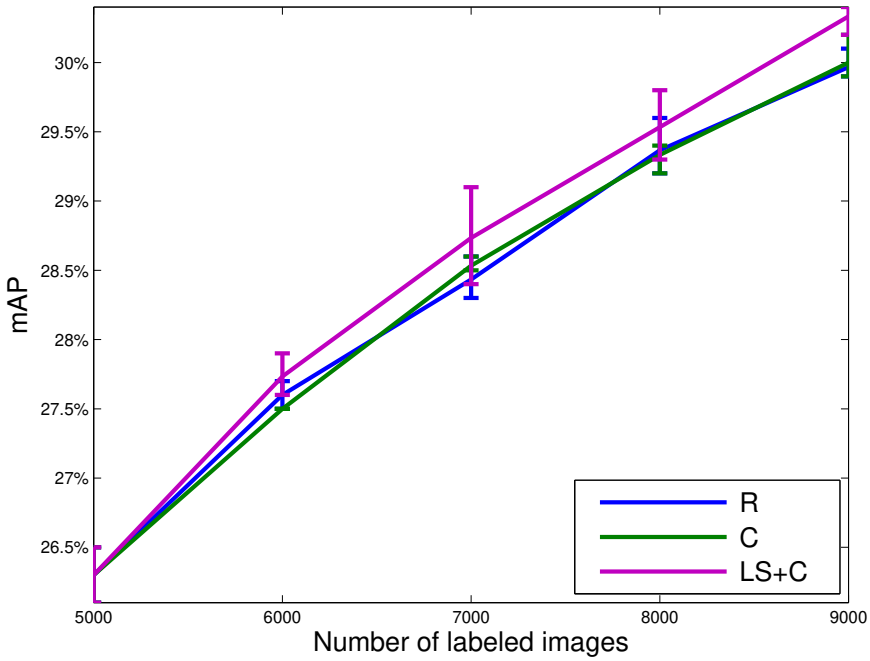


Fig. 7: Mean average precision curve of different active learning methods on the MS COCO validation set. Each point in the plot is an average of 3 trials. The error bars represent the minimum and maximum values out of 3 trials at each point. This is a full version of Fig. 9a in the main paper.

images, and then each active learning algorithm is executed for 3 rounds. In each round, we select 200 images, add these images to the existing training set. After 3 rounds, each method has selected 600 images for annotation, and a set with 1,100 labeled images is used to train the detector.

Table 5 shows the average precision for each method on the PSACAL 2012 validation set after 3 rounds of active learning. As defined in the main paper, catergories with AP lower than 40% in passive learning (R) are defined as difficult categories. These difficult classes are marked by an asterisk in Table 5. We further analyze the selection result of different methods by a visualization as shown in Fig. 12. There are total 5,217 images (500 images in the initial training set of this trial are not included) in each graph. 600 images selected for annotation by each active learning method are represented by green asterisks, and the rest 4,617 images that have not been chosen are represented by black dots.

We have two major observations from the visualzation results on the PASCAL 2012 dataset. First, the random sampling (R) method selects images for annotation across all categories, no matter it is a difficult class or an easy class. Compared to the other methods, lots of images of cats and cars are selected by R (blue rectangles in Fig. 12a

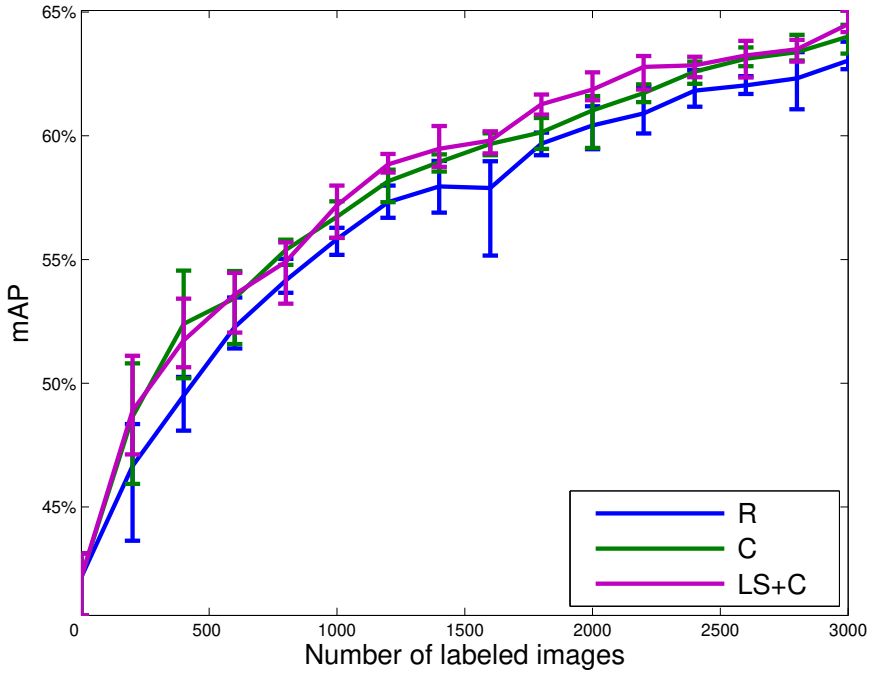


Fig. 8: Mean average precision curve of different active learning methods with SSD on the PASCAL 2007 testing set. Each point in the plot is an average of 5 trials. The error bars represent the minimum and maximum values out of 5 trials at each point. This is a full version of Fig. 10a in the main paper.

and Fig. 14a). However, these classes are relatively easy so the room for improvements is not that large. Also, the selected images are not informative so that even many images are selected in these classes, there is no large improvement over the other methods.

Second, as mentioned in Sec. 4.1 in the main paper, the proposed method LS+C outperforms the baseline method C especially in the difficult categories. There is a $10\times$ difference between difficult and non-difficult categories in the improvement of LS+C over C as shown in Fig. 6a in the main paper. These 5 difficult categories are: boat, bottle, chair, table, and plant. Fig. 13 shows that all difficult categories but boat locate at the left part of the 2D plane. These categories also are the ones show in scenes of a living room (Fig. 9), as mentioned in the previous section. By visual inspection, the red rectangles in Fig. 12c and Fig. 12b show that the proposed LS+C tends to select more images for annotation in these difficult classes than the baseline method C. Quantitative results are shown in Fig. 10. The proposed LS+C selects images that contain objects belong to difficult classes much more than the baseline method C. By selecting more images for annoation, the proposed LS+C gets more improvement in these difficult classes. In contrast, for easy classes (catergories with AP higher than 70% in passive

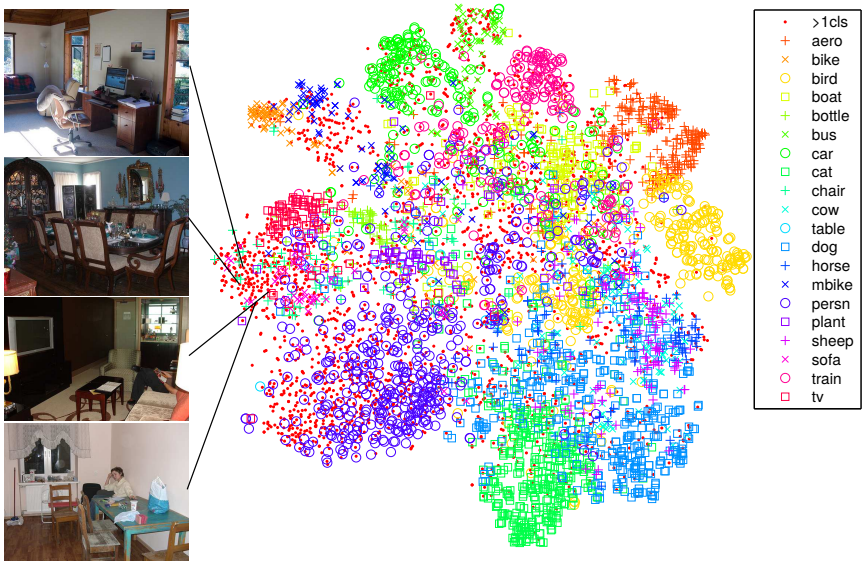


Fig. 9: t-SNE embeddings of images on the PASCAL 2012 training set. VGG16 is used for generating high-dimensional vectors of images that used for the embedding. Each data point in the scatter plot is an image. "$>$1cls" represents an image that has objects belong to different classes. Images marked by only one class means that all the objects in the image belong to the same class. Images on the left are examples contain objects belong to difficult classes. As defined in Table 5 ,the difficult classes are boat, bottle, chair, table, and plant.

learning) like cat and dog, the baseline method C selects more images than the proposed LS+C as shown in Fig. 11. These observations indicate that C focuses on non-difficult categories to get an overall improvement in mAP, but does not perform well in difficult categories.

References

1. Goodfellow, I.J., Shlens, J., Szegedy, C.: Explaining and harnessing adversarial examples. In: ICLR. (2015)
2. Everingham, M., Van Gool, L., Williams, C.K.I., Winn, J., Zisserman, A.: The PASCAL Visual Object Classes (VOC) challenge. International Journal of Computer Vision (IJCV) **88** (2010) 303–338
3. Maaten, L.v.d., Hinton, G.: Visualizing data using t-sne. Journal of Machine Learning Research **9** (2008) 2579–2605
4. Simonyan, K., Zisserman, A.: Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:1409.1556 (2014)

method	aero	bike	bird	boat*	bottle*	bus	car	cat	chair*	cow	table*
R	68.3	61.5	54.2	27.8	30.4	68.2	58.2	76.3	28.4	44.8	31.1
C	72.8	66.6	50.8	28.5	34.8	64.3	54.3	77.5	27.2	**53.2**	36.3
LS+C	68.1	**68.0**	52.0	**34.2**	34.9	**70.0**	**59.9**	74.4	30.3	44.2	**42.1**
LT/C	**74.8**	64.8	**60.1**	28.7	**36.4**	63.9	58.1	**79.7**	**31.0**	51.1	38.1

method	dog	horse	mbike	persn	plant*	sheep	sofa	train	tv	mAP
R	73.7	64.1	67.9	66.7	21.9	52.4	41.7	**64.8**	55.5	52.9
C	**79.0**	**70.4**	66.5	69.0	21.9	59.6	38.8	60.6	54.5	54.3
LS+C	73.6	63.3	**69.7**	**71.7**	**28.5**	**60.2**	40.6	64.4	**59.0**	55.5
LT/C	72.9	66.0	66.9	67.2	23.7	56.4	**50.4**	64.3	54.6	**55.5**

Table 5: Average precision for each method on the PASCAL 2012 validation set after 3 rounds of active learning (the number of labeled images in the training set is 1,100). Each number shown in the table is the result of one trial (different from Table 1 in the main paper which shows the average over 5 trials) and displayed in percentage. Numbers in bold are the best results per column, and underlined numbers are the second best results. Catergories with AP lower than 40% in passive learning (R) are defined as difficult categories and marked by an asterisk.

5. Ren, S., He, K., Girshick, R., Sun, J.: Faster R-CNN: Towards real-time object detection with region proposal networks. In: Advances in Neural Information Processing Systems (NIPS). (2015)

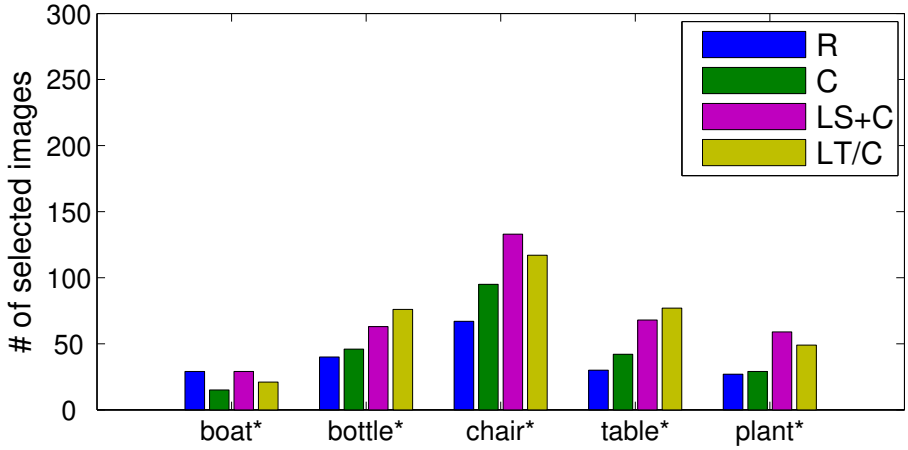


Fig. 10: The number of selected images that contain objects belong to difficult classes by different active learning methods.

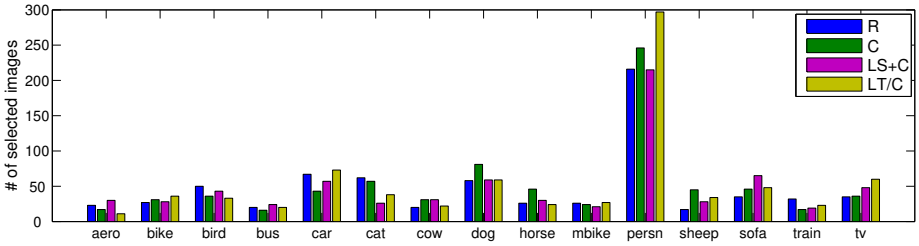
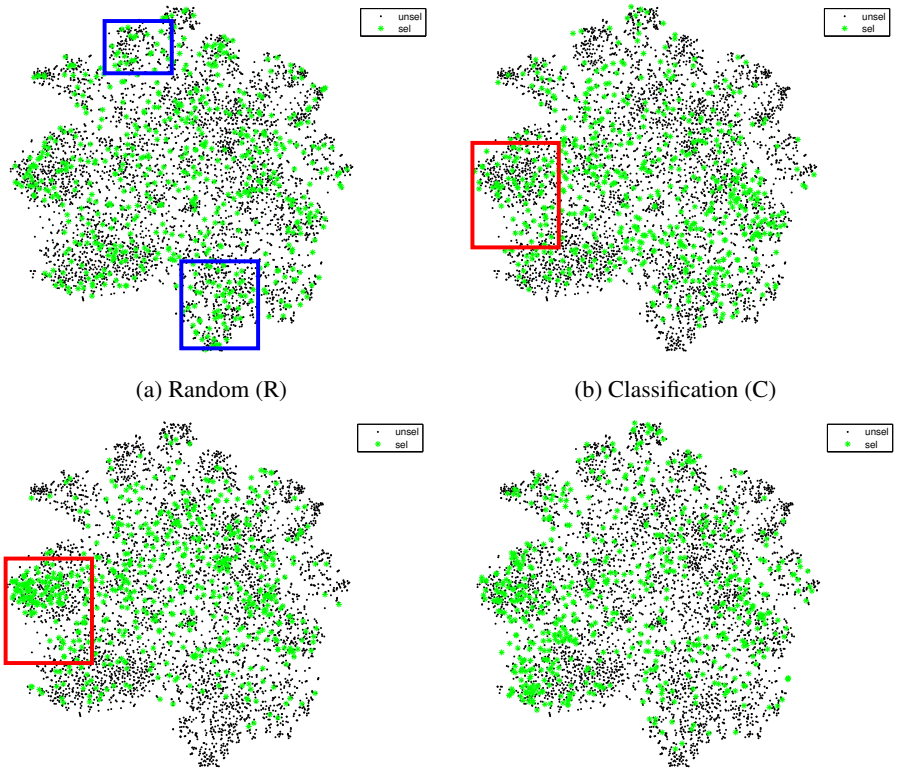


Fig. 11: The number of selected images that contain objects belong to non-difficult classes by different active learning methods.



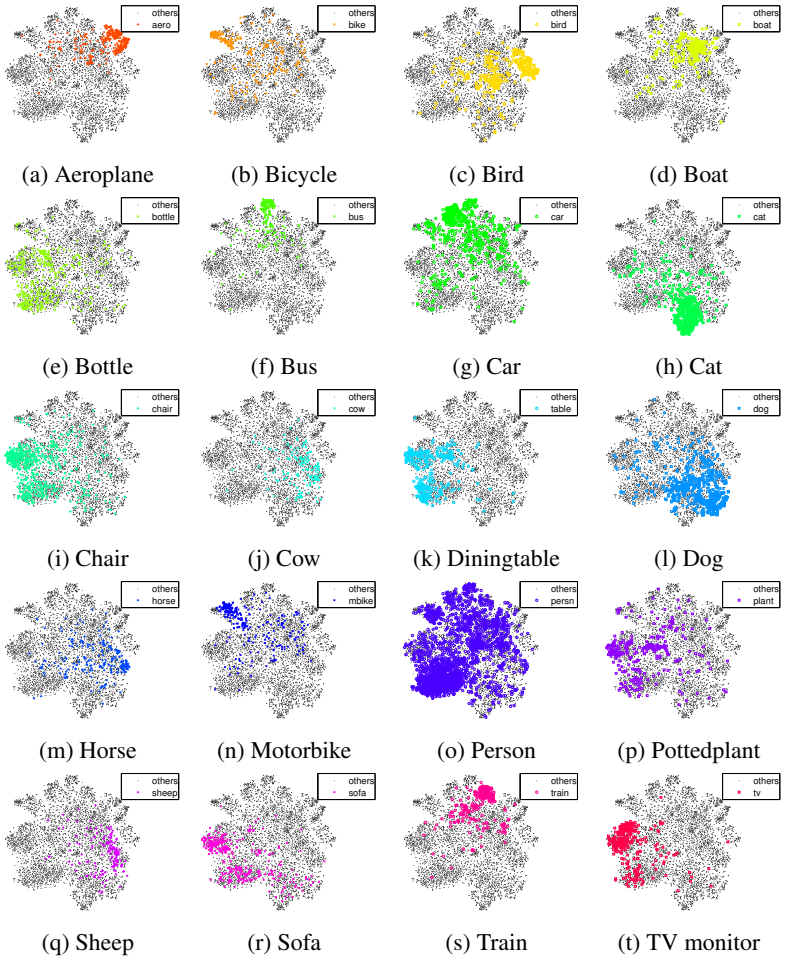
(a) Random (R)

(b) Classification (C)

(c) Localization stability + classification (LS+C)

(d) Localization tightness + classification (LT/C)

Fig. 12: The visualization of selection results by different active learning methods. Green asterisks (sel) are the images selected for annotation by each active learning method, and black dots (unsel) are the images that have not been selected. A detailed version of this graph with class-wise information is shown in Fig. 14.



(a) Aeroplane (b) Bicycle (c) Bird (d) Boat

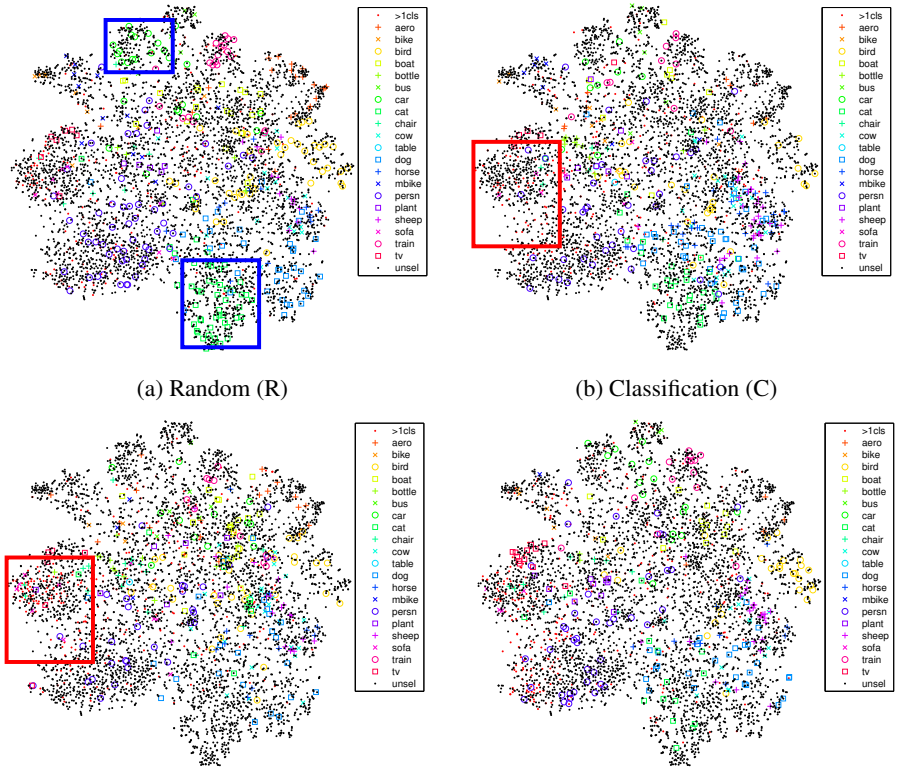
(e) Bottle (f) Bus (g) Car (h) Cat

(i) Chair (j) Cow (k) Diningtable (l) Dog

(m) Horse (n) Motorbike (o) Person (p) Pottedplant

(q) Sheep (r) Sofa (s) Train (t) TV monitor

Fig. 13: t-SNE embeddings of images for each category on the PASCAL 2012 training set. Different from Fig. 9, each colored point in the graphs represents an image that includes at least one object belongs to the target class. For example, each orange plus sign in (a) represents an image which has at least one aeroplane in it.



(a) Random (R)

(b) Classification (C)

(c) Localization stability + classification (LS+C)

(d) Localization tightness + classification (LT/C)

Fig. 14: The visualization of selection results by different active learning methods. Different from Fig. 12, each colored marker not only represents a selected image, but also indicates the class that objects contained in the image belong to.