Localization-Aware Active Learning for Object Detection

Kao, C.-C.; Lee, T.-Y.; Sen, P.; Liu, M.-Y.

Abstract

Active learning is a class of algorithms that iteratively searches for the most informative samples to include in a training dataset has been shown to be effective at annotating data for image classification. However, the use of active learning for object detection is still largely unexplored as determining informativeness of an object-location hypothesis is more difficult. In this paper, we address this issue and present two metrics for measuring the informativeness of an object hypothesis, which allow us to leverage active learning to reduce the amount of annotated data needed to achieve a target object detection performance. Our first metric measures "localization tightness" of an object hypothesis, which is based on the overlapping ratio between the region proposal and the final prediction. Our second metric measures "localization stability" of an object hypothesis, which is based on the variation of predicted object locations when input images are corrupted by noise. Our experimental results show that by augmenting a conventional active-learning algorithm designed for classification with the proposed metrics, the amount of labeled training data required can be reduced up to 25%. Moreover, on PASCAL 2007 and 2012 datasets our localization-stability method has an average relative improvement of 96.5% and 81.9% over the baseline method using classification only.

arXiv

Localization-Aware Active Learning for Object Detection

Chieh-Chi Kao
University of California, Santa Barbara
chiehchi.kao@gmail.com

Teng-Yok Lee
Mitsubishi Electric Research Laboratories
tlee@merl.com

Pradeep Sen
University of California, Santa Barbara
psen@ece.ucsb.edu

Ming-Yu Liu
Mitsubishi Electric Research Laboratories
seanmingyuliu@gmail.com

Abstract

Active learning—a class of algorithms that iteratively searches for the most informative samples to include in a training dataset—has been shown to be effective at annotating data for image classification. However, the use of active learning for object detection is still largely unexplored as determining informativeness of an object-location hypothesis is more difficult. In this paper, we address this issue and present two metrics for measuring the informativeness of an object hypothesis, which allow us to leverage active learning to reduce the amount of annotated data needed to achieve a target object detection performance. Our first metric measures "localization tightness" of an object hypothesis, which is based on the overlapping ratio between the region proposal and the final prediction. Our second metric measures "localization stability" of an object hypothesis, which is based on the variation of predicted object locations when input images are corrupted by noise. Our experimental results show that by augmenting a conventional active-learning algorithm designed for classification with the proposed metrics, the amount of labeled training data required can be reduced up to 25%. Moreover, on PASCAL 2007 and 2012 datasets our localization-stability method has an average relative improvement of 96.5% and 81.9% over the baseline method using classification only.

1. Introduction

Prior works have shown that with a large amount of annotated data, convolutional neural networks (CNNs) can be trained to achieve a super-human performance for various visual recognition tasks. As tremendous efforts are dedicated into the discovery of effective network architectures and training methods for further advancing the performance, we argue it is also important to investigate into effective approaches for data annotation as data annotation is essential but expensive.

Data annotation is especially expensive for the object-detection task. Compared to annotating image class, which can be done via a multiple-choice question, annotating object location requires a human annotator to specify a bounding box for an object. Simply dragging a tight bounding box to enclose an object can cost 10-times more time than answering a multiple-choice question [27, 18]. Consequently, a higher pay rate has to be paid to a human labeler for annotating images for an object detection task. In addition to the cost, it is more difficult to monitor and control the annotation quality.

Active learning [23] is a machine learning procedure that is useful in reducing the amount of annotated data required to achieve a target performance. It has been applied to various computer-vision problems including object classification [11, 5], image segmentation [14, 3], and activity recognition [7, 8]. Active learning starts by training a baseline model with a small, labeled dataset, and then applying the baseline model to the unlabeled data. For each unlabeled sample, it estimates whether this sample contains critical information that has not been learned by the baseline model. Once the samples that bring the most critical information are identified and labeled by human annotators, they can be added to the initial training dataset to train a new model, which is expected to perform better. Compared to *passive learning*, which randomly selects samples from the unlabeled dataset to be labeled, active learning can achieve the same accuracies with fewer but more informative labeled samples.

Multiple metrics for measuring how informative a sample is have been proposed for the classification task, including maximum uncertainty, expected model change, density weighted, and so on [23]. The concept behind several of them is to evaluate how uncertain the current model is for an unlabeled sample. If the model could not assign a high probability to a class for a sample, then it implies the model

is uncertain about the class of the sample. In other words, the class of the sample would be very informative to the model. This sample would require human to clarify.

Since an object-detection problem can be considered as an object-classification problem once the object is located, existing active learning approaches for object detection [1, 25] mainly measure the information in the classification part. Nevertheless, in addition to classification, the accuracy of an object detector also relies on its localization ability. Because of the importance of localization, in this paper we present an active learning algorithm tailored for object detection, which considers the localization of detected objects. Given a baseline object detector which detects bounding boxes of objects, our algorithm evaluates the uncertainty of both the classification and localization.

Our algorithm is based on two quantitative metrics of the localization uncertainty.

1. *Localization Tightness (LT)*: The first metric is based on how tight the detected bounding boxes can enclose true objects. The tighter the bounding box, the more certain the localization. While it sounds impossible to compute the localization tightness for non-annotated images because the true object locations are unknown, for object detectors that follow the propose-then-classify pipeline [6, 21], we estimate the localization tightness of a bounding box based on its changes from the intermediate proposal (a box contains any kind of foreground objects) to the final class-specific bounding box.

2. *Localization Stability (LS)*: The second metric is based on whether the detected bounding boxes are sensitive to changes in the input image. To evaluate the localization stability, our algorithm adds different amounts of Gaussian noise to pixel values of the image, and measures how the detected regions vary with respect to the noise. This one can be applied to all kinds of object detectors, especially those that do not have an explicit proposal stage [20, 17].

The contributions of this paper are two-fold:

1. We present different metrics to quantitatively evaluate the localization uncertainty of an object detector. Our metrics consider different aspects of object detection in spite that the ground truth of object locations is unknown, making our metrics suited for active learning.

2. We demonstrate that to apply active learning for object detection, both the localization and the classification of a detector should be considered when sampling informative images. Our experiments on benchmark datasets show that considering both the localization and classification uncertainty outperforms the existing

active-learning algorithm works on the classification only and passive learning.

2. Related Works

We now review active learning approaches used for image classification. For more detail of active learning, Settles's survey [23] provides a comprehensive review. In this paper, we use the maximum uncertainty method in the classification as the baseline method for comparison. The uncertainty based method is used for CAPTCHA recognition [26], image classification [10], and automated and manual video annotation [13]. It also has been applied to different learning models including decision trees [15], SVMs [28], and Gaussian processes [12]. We choose uncertainty-based method since it is efficient to compute.

Active learning is also applied for object detection tasks in various specific applications, such as satellite images [1] and vehicle images [25]. Vijayanarasimhan *et al.* [30] propose an approach to actively crawl images from the web to train part-based linear SVM detector. Note that these methods only consider information from the classifier, while our methods aim to consider the localization part as well.

Current state-of-the-art object detectors are based on deep-learning. They can be classified into two categories. Given an input image, the first category explicitly generates region proposals, following by feature extraction, category classification, and fine-tuning of the proposal geometry [6, 21]. The other category directly outputs the object location and class without the intermediate proposal stage, such as YOLO [20] and SSD [17]. This inspires us to consider localization stability, which can be applied to both categories.

Besides active learning, there are other research directions to reduce the cost for annotation. Temporal coherence of the video frames are used to reduce the annotation effort for training detectors [19]. Domain adaptation [9] is used to transfer the knowledge from an image classifier to an object detector without the annotation of bounding boxes. Papadopoulos *et al.* [18] suggest to simplify the annotation process from drawing a bounding box to simply answering a Yes/No question whether a bounding box tightly encloses an object. Russakovsky *et al.* [22] integrate multiple inputs from both computer vision and humans to label objects.

3. Active Learning for Object Detection

The goal of our algorithm is to train an object detector that takes an image as input and outputs a set of rectangular bounding boxes. Each bounding box has the location and the scale of its shape, and a probability mass function of all classes. To train such an object detector, the training and validation images of the detector are annotated with an bounding box per object and its category. Such an anno-
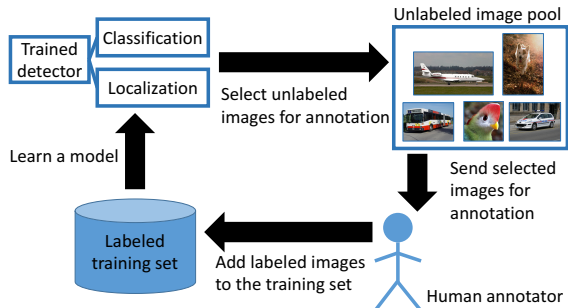
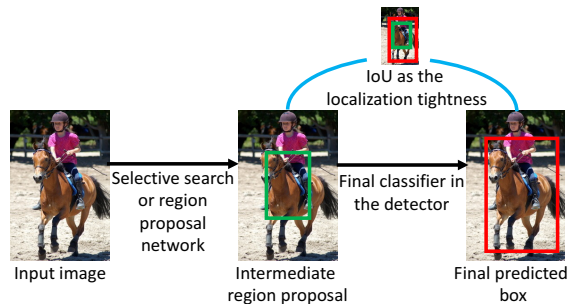Figure 1: A round of active learning for object detection.

Figure 2: The process of calculating the tightness of each predicted box. Given an intermediate region proposal, the detector refines it to a final predicted box. The IoU calculated by the final predicted box and its corresponding region proposal is defined as the localization tightness of that box.

tation is commonly seen in public datasets including PAS-CAL VOC [4] and MS COCO [16].

We first review the basic active learning framework for object detection in Sec. 3.1. It also reviews the measurement of classification uncertainty, which is the major measurement for object detection in previous active learning algorithms for object detection [23, 1, 25]. Based on this framework, we extend the uncertainty measurement to also consider the localization result of a detector, as described in Sec. 3.2 and 3.3.

3.1. Active Learning with Classification Uncertainty

Fig. 1 overviews our active learning algorithm. Our algorithm starts with a small training set of annotated images to train a baseline object detector. In order to improve the detector by training with more images, we continue to collect images to annotate. Other than annotating all newly collected images, based on different characteristics of the current detector, we select a subset of them for human annotators to label. Once being annotated, these selected images are added to the training set to train a new detector. The entire process continues to collect more images, select a subset with respect to the new detector, annotate the selected ones with humans, re-train the detector and so on. Hereafter we call such a cycle of data collection, selection, annotation, and training as a *round*.

A key component of active learning is the selection of images. Our selection is based on the uncertainty of both the classification and localization. The classification uncertainty of a bounding box is the same as the existing active learning approaches [23, 1, 25]. Given a bounding box B, its classification uncertainty $U_B(B)$ is defined as $U_B(B) = 1 - P_{max}(B)$ where $P_{max}(B)$ is highest probability out of all classes for this box. If the probability on a single class is close to 1.0, meaning that the probabilities for other classes are low, the detector is highly certain about its class. To the contrast, when multiple classes have similar probabilities, each probability will be low because the sum

of probabilities of all classes must be one.

Based on the classification uncertainty per box, given the i-th image to evaluate, say I_i, its classification uncertainty is denoted as $U_C(I_i)$, which is calculated by the maximum uncertainty out of all detected boxes within.

3.2. Localization Tightness

Our first metric of the localization uncertainty is based on the *Localization Tightness* (LT) of a bounding box. The localization tightness measures how tight a predicted bounding box can enclose true foreground objects. Ideally, if the ground-truth locations of the foreground objects are known, the tightness can be simply computed as the IoU (Intersection over Union) between the predicted bounding box and the ground truth. Given two boxes B^1 and B^2, their IoU is defined as: $IoU(B^1, B^2) = \frac{B^1 \cap B^2}{B^1 \cup B^2}$.

Because the ground truth is unknown for an image without annotation, an estimate for the localization tightness is needed. Here we design an estimate for object detectors that involves the adjustment from intermediate region proposals to the final bounding boxes. Region proposals are the bounding boxes that might contain any foreground objects, which can be obtained via the selective search [29] or a region proposal network [21]. Besides classifying the region proposals into specific classes, the final stage of these object detectors can even adjust the location and scale of region proposals based on the classified object classes. Fig. 2 illustrates the typical pipeline of these detectors where the region proposal (green) in the middle is adjusted to the red box in the right.

As the region proposal is trained to predict the location of foreground objects, the refinement process in the final stage is actually related to how well the region proposal predicts. If the region proposal locates the foreground object perfectly, there is no need to refine it. Based on

(a)   (b)

Figure 3: Images preferred by LT/C. Top rows show two figures are two cases that will be selected by LT/C, which are images with certain category but loose bounding box (a) or images with tight bounding box but uncertain about the category (b).

this observation, we use the IoU value between the region proposal and the refined bounding box to estimate the localization tightness between an adjusted bounding box and the unknown ground truth. The estimated tightness T of j-th predicted box B_0^j can be formulated as following: $T(B_0^j) = IoU(B_0^j, R_0^j)$, where R_0^j is the corresponding region proposal fed into the final classifier that generates B_0^j.

Once the tightness of all predicted boxes are estimated, we can extend the selection process to consider not only the classification uncertainty but also the tightness. Namely, we want to select images with inconsistency between the classification and the localization, as following:

- Given a predicted box that is absolutely certain about its classification result ($P_{max} = 1$), but it cannot tightly enclose a true object ($T = 0$). An example is shown in Figure 3 (a).

- Reversely, if the predicted box can tightly enclose a true object ($T = 1$) but the classification result is uncertain (low P_{max}). An example is shown in Figure 3 (b).

The score of a box is denoted as J, which is computed per Equ. 1. Both conditions above can get value close to zero.

$$J(B_0^j) = |T(B_0^j) + P_{max}(B_0^j) - 1| \quad (1)$$

As each image can have multiple predicted boxes, we calculate the score per image as: $T_I(I_i) = min_j J(B_0^j)$. Unlabeled images with low score will be selected to annotate in active learning. Since both the localization tightness and classification outputs are used in this metric, later we use LT/C to denotes methods with this score.

3.3. Localization Stability

The concept behind the localization stability is that, if the current model is stable to noise, meaning that the detection result does not dramatically change even if the input
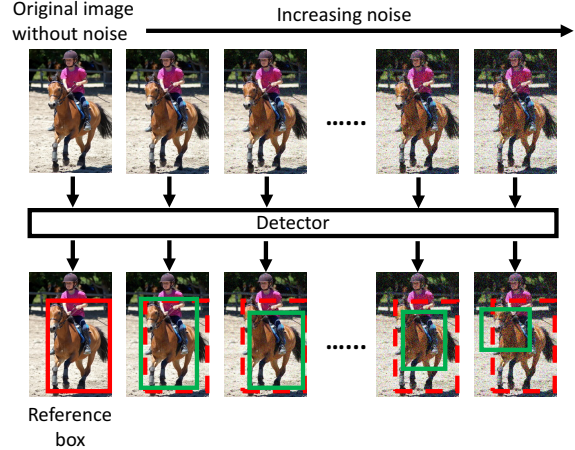

Figure 4: The process of calculating the localization stability of each predicted box. Given one input image, a reference box (red) is predicted by the detector. The change in predicted boxes (green) from noisy images is measured by the IoU of predicted boxes (green) and the corrsponding reference box (dashed red).

unlabeled image is corrupted by noise, the current model already understands this unlabeled image well so there is no need to annotate this unlabeled image. In other words, we would like to select images that have large variation in the localization prediction of bounding boxes when the noise is added into the image.

Fig. 4 overviews the idea to calculate the localization stability of an unlabeled image. We first detect bounding boxes in the original image with the current model. These bounding boxes when noise is absent are called reference boxes. The j-th reference box is denoted as B_0^j. For each noise level n, a noise is added to each pixel of the image. We use Gaussian noise where the standard deviation is proportional to the level n; namely, the pixel value can be changed more for higher level. After detecting boxes in the image with noise level n, for each reference box (the red box in Fig. 4), we find a corresponding box (green) in the noisy image to calculate how the reference box varies. The corresponding box is denoted as $C_n(B_0^j)$, which has the highest IoU value among all bounding boxes that overlap B_0^j.

Once all the corresponding boxes from different noise levels are detected, we can tell that the model is stable to noise on this reference box if the box does not significantly change across the noise levels. Therefore, the localization stability of each reference box B_0^j can be defined as the average of IoU between the reference box and corresponding boxes across all noise levels. Given N noise levels, it is calculated per Equ. 2:

$$S_B(B_0^j) = \frac{\sum_{n=1}^N IoU(B_0^j, C_n(B_0^j))}{N}, \qquad (2)$$

With the localization stability of all reference boxes, the localization stability of this unlabeled image, says I_i, is defined based on their weighted sum per Equ. 3 where M is the number of reference boxes. The weight of each reference box is its highest class probability in order to prefer boxes with high probability as foreground objects but high uncertainty to their locations.

$$S_I(I_i) = \frac{\sum_{j=1}^M P_{max}(B_0^j) S_B(B_0^j)}{\sum_{j=1}^M P_{max}(B_0^j)}. \qquad (3)$$

4. Experimental Results

Reference Methods: Since no prior work does active learning for deep learning based object detectors, we designate two informative baselines that show the impact of proposed methods.

- **Random (R):** Randomly choose samples from the unlabeled set, label them, and put them into labeled training set.

- **Classification only (C):** Select images only based on the classification uncertainty U_c in Sec. 3.1.

Our algorithm with two different metrics for the localization uncertainty are tested. First, the localization stability (Section 3.3) is combined with the classification information (**LS+C**). As images with high classification uncertainty and low localization stability should be selected for annotation, the score of the i-th image (I_i) image is defined as follows: $U_C(I_i) - \lambda S_I(I_i)$,where λ is the weight to combine both, which is set to 1 across all the experiments in this paper. Second, the localization tightness of predicted boxes is combined with the classification information (**LT/C**) as defined in Section 3.2.

We also test three variants of our algorithm. One uses the localization stability only (**LS**). Another is the localization tightness of predicted boxes combined with the classification information but using the localization tightness calculated from ground-truth boxes (**LT/C(GT)**) instead of the estimate used in LT/C. The other is combining all 3 cues together (**3in1**).

For the easiness of reading, data for LS and 3in1 are shown in the supplementary result. Our supplementary result also includes the mAP curves with error bars that indicate the minimum and maximum average precision (AP) out of multiple trials of all methods. Furthermore, experiments with different designs of LT/C are included in the supplementary result.

Datasets: We validated our algorithm on three datasets (PASCAL 2012, PASCAL 2007, MS COCO [4, 16]). For each dataset, we started from a small subset of the training set to train the baseline model, and selected from the remained training images for active learning. Since objects in training images from these datasets have been annotated with bounding boxes, our experiments used these bounding boxes as annotation without asking human annotators.

Detectors: The object detector for all datasets is the Faster-RCNN (FRCNN) [21], which contains the intermediate stage to generate region proposals. We also tested our algorithm with the Single Shot multibox Detector (SSD) [17] on the PASCAL 2007 dataset. Because the SSD does not contain a region proposal stage, the tests for localization tightness were skipped. Both FRCNN and SSD used VGG16 [24] as the pre-trained network in the experiments shown in this paper.

4.1. FRCNN on PASCAL 2012

Experimental Setup: We evaluate all the methods with the FRCNN model [21] using the RoI warping layer [2] on the PASCAL 2012 object-detection dataset [4] that consists of of 20 classes. Its training set (5,717 images) is used to mimic a pool of unlabeled images, and the validation set (5,823 images) is used for testing. Input images are resized to have 600 pixels on the shortest side for all FRCNN models in this paper.

The numbers shown in following sections on PASCAL datasets are averages over 5 trails for each method. All trials start from the same baseline object detectors, which are trained with 500 images selected from the unlabeled image pool. After then, each active learning algorithm is executed in 15 rounds. In each round, we select 200 images, add these images to the existing training set, and train a new model. Each model is trained with 20 epoches.

Our experiments used Gaussian noise as the noise source for the localization stability. We set the number of noise level N to 6. The standard deviations of these levels are {8, 16, 24, 32, 40, 48} where the pixels range from [0, 255].

Results: Fig. 5a and Fig. 5b show the mAP curve and the relative saving of labeled images, respectively, for different active learning methods. We have three major observations from the results on the PASCAL 2012 dataset. First, LT/C(GT) outperforms all other methods in most of the cases as shown in Fig. 5b. This is not surprising since LT/C(GT) is based on the ground-truth annotations. In the region that achieves the same performance as passive learning with a dataset of 500 to 1,100 labeled images, the performance of the proposed LT/C is similar to LT/C(GT), which represents the full potential of LT/C. This implies that LT/C using the estimate of tightness of predicted boxes (Section 3.2) can achieve results close to its upper bound.

Second, in most of the cases, active learning approaches
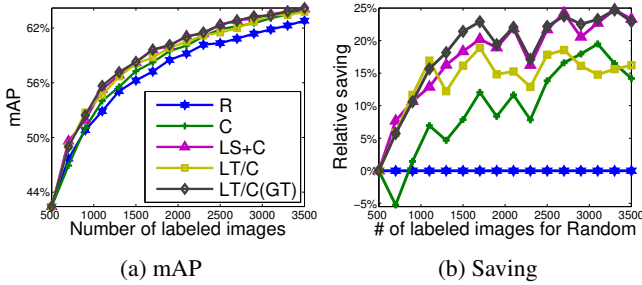
(a) mAP   (b) Saving

Figure 5: (a) Mean average precision curve of different active learning methods on PASCAL 2012 detection dataset. Each point in the plot is an average of 5 trials. (b) Relative saving of labeled images for different methods.

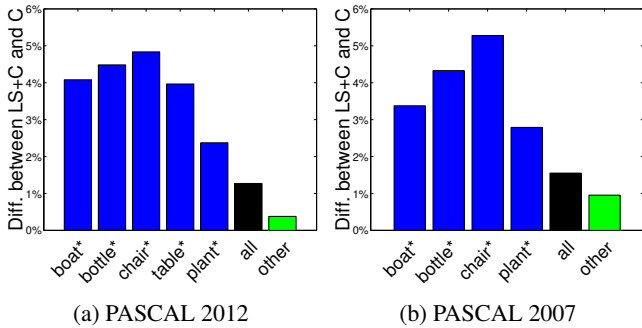(a) PASCAL 2012   (b) PASCAL 2007

Figure 6: The difference in difficult classes (blue bars) between the proposed method (LS+C) and the baseline method (C) in average precision on (a) PASCAL 2012 dataset (b) PASCAL 2007 dataset. Black and green bars are the average improvements of LS+C over C for all classes and non-difficult classes.

work better than random sampling. The localization stability with the classfication uncertainty (LS+C) has the best performance among all methods other than LT/C(GT). In terms of average saving, LS+C and LT/C have 96.5% and 36.3% relative improvement over the baseline method C.

Last, we also note that the proposed LS+C method has more improvements in the difficult categories. We further analyze the performance of each method by inspecting the AP per category. Table 1 shows the average precision for each method on the PASCAL 2012 validation set after 3 rounds of active learning, meaning that every model is trained on a dataset with 1,100 labeled images. For categories with AP lower than 40% in passive learning (R), we treat them as difficult categories, which have a asterisk next to their name. For these difficult categories (blue bars) in Fig. 6a, we notice that the improvement of LS+C over C is large. For those 5 difficult categories the average improvement of LS+C over C is 3.95%, while the average improvement is only 0.38% (the green bar in Fig. 6a) for the rest
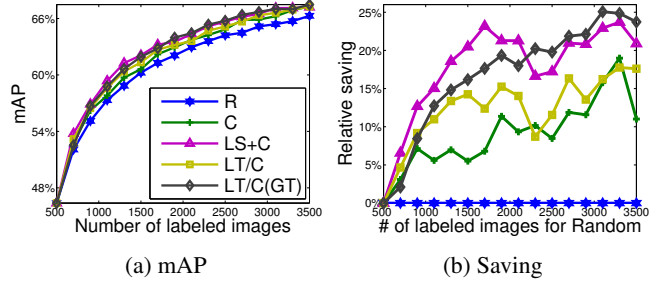

(a) mAP   (b) Saving

Figure 7: (a) Mean average precision curve of different active learning methods on PASCAL 2007 detection dataset. Each point in the plot is an average of 5 trials. (b) Relative saving of labeled images for different methods.

15 non-difficult categories. This $10\times$ difference shows that adding the localization information into active learning for object detection can greatly help the learning for difficult categories. It is also noteworthy that for those 5 difficult categories, the baseline method C performs slightly worse than random sampling by 0.50% in average. It indicates that C focuses on non-difficult categories to get an overall improvement in mAP.

4.2. FRCNN on PASCAL 2007

Experimental Setup: We evaluate all the methods with the FRCNN model [21] using the RoI warping layer [2] on the PASCAL VOC 2007 object-detection dataset [4] that consists of 20 classes. Both training and validation sets (total 5,011 images) are used as the unlabeled image pool, and the test set (4,952 images) is used for testing. All the experimental settings are the same as the experiments on the PASCAL 2012 dataset as mentioned Section 4.1.

Results: Fig. 7a and Fig. 7b show the mAP curve and relative saving of labeled images for different active learning methods. In terms of average saving, LS+C and LT/C have 81.9% and 45.2% relative improvement over the baseline method C. Table 2 shows the AP for each method on the PASCAL 2007 test set after 3 rounds of active learning. The proposed LS+C and LT/C are better than the baseline classification-only method (C) in terms of mAP.

It is interesting to see that LS+C method has the same behavior as shown in the experiments on the PASCAL 2012 dataset. Namely, LS+C also outperforms the baseline model C on difficult categories. As the setting in experiments on the PASCAL 2012 dataset, categories with AP lower than 40% in passive learning (R) are considered as difficult categories. For those 4 difficult categories, the average improvement in AP of LS+C over C is 3.94%, while the average improvement is only 0.95% (the green bar in Fig. 6b) for the other 16 categories.

method	aero	bike	bird	boat*	bottle*	bus	car	cat	chair*	cow	table*	dog	horse	mbike	persn	plant*	sheep	sofa	train	tv	mAP
R	<u>71.1</u>	61.5	<u>54.7</u>	28.4	32.0	68.1	57.9	75.4	25.8	44.2	36.4	73.0	61.9	67.3	68.1	21.6	51.9	41.0	**65.5**	51.7	52.9
C	70.7	62.9	54.7	25.5	30.8	66.1	56.2	**78.1**	26.4	**54.5**	36.7	**76.9**	**68.3**	<u>67.7</u>	67.4	22.5	<u>57.7</u>	40.8	63.6	52.5	54.0
LS+C	**73.9**	<u>63.7</u>	**56.9**	**29.6**	**35.2**	66.5	<u>58.5</u>	<u>77.9</u>	**31.3**	<u>50.8</u>	<u>40.7</u>	<u>73.8</u>	<u>65.4</u>	66.9	<u>68.4</u>	**24.8**	**58.0**	**44.9**	64.2	<u>53.9</u>	**55.3**
LT/C	69.8	**64.6**	54.6	<u>29.5</u>	<u>33.8</u>	**70.3**	**59.7**	75.5	<u>29.5</u>	46.3	**41.8**	73.0	62.5	**69.0**	**70.8**	<u>23.2</u>	56.5	<u>42.8</u>	<u>64.3</u>	**55.9**	<u>54.7</u>

Table 1: Average precision for each method on PASCAL 2012 validation set after 3 rounds of active learning (number of labeled images in the training set is 1,100). Each number shown in the table is an average of 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 asterisk.

method	aero	bike	bird	boat*	bottle*	bus	car	cat	chair*	cow	table	dog	horse	mbike	persn	plant*	sheep	sofa	train	tv	mAP
R	**61.6**	67.2	54.1	40.0	33.6	64.5	73.0	73.9	34.5	60.8	52.2	69.3	**74.7**	<u>66.6</u>	<u>67.1</u>	25.9	52.1	54.2	**66.1**	54.9	57.3
C	56.9	<u>68.0</u>	<u>54.9</u>	36.8	34.4	<u>68.1</u>	71.7	**75.5**	34.0	**68.6**	51.0	<u>71.4</u>	<u>74.7</u>	65.2	65.9	24.9	**60.0**	53.9	63.0	<u>57.4</u>	57.8
LS+C	<u>61.5</u>	64.4	**55.8**	<u>40.2</u>	**38.7**	66.3	<u>73.8</u>	74.7	**39.6**	<u>68.0</u>	<u>56.3</u>	**71.5**	73.8	**67.2**	66.7	<u>27.7</u>	<u>61.3</u>	**57.0**	<u>65.6</u>	57.4	**59.4**
LT/C	57.6	**69.7**	52.9	**41.1**	<u>38.4</u>	**69.7**	**74.4**	71.8	<u>36.4</u>	61.2	**58.1**	69.5	74.3	66.2	**67.8**	**28.0**	55.5	<u>56.3</u>	65.5	**58.2**	<u>58.6</u>

Table 2: Average precision for each method on PASCAL 2007 test set after 3 rounds of active learning (number of labeled images in the training set is 1,100). The other experimental settings are the same as shown in Table 1.
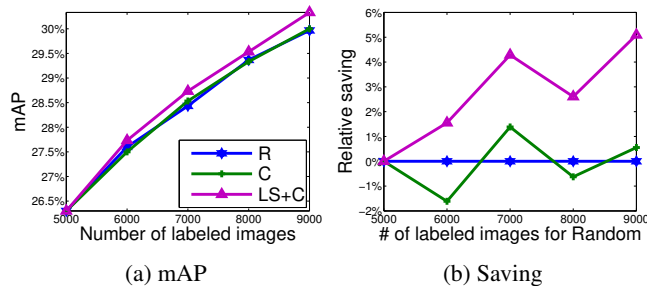

(a) mAP

(b) Saving

Figure 8: (a) Mean average precision curve (@IoU=0.5) of different active learning methods on MS COCO detection dataset. (b) Relative saving of labeled images for different methods. Each point in the plots is an average of 3 trials.
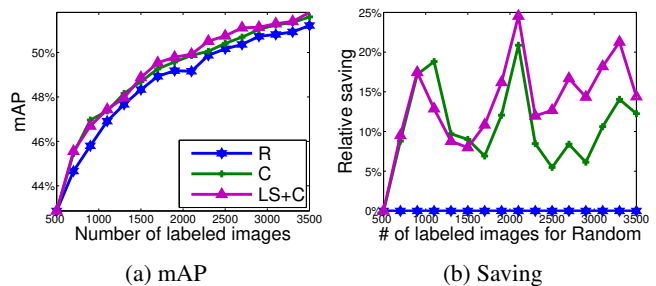

(a) mAP

(b) Saving

Figure 9: (a) Mean average precision curve of SSD with different active learning methods on PASCAL 2007 detection dataset. (b) Relative saving of labeled images for different methods. Each point in the plots is an average of 5 trials.

4.3. FRCNN on MS COCO

Experimental Setup: For the MS COCO object-detection dataset [16], we evaluate three methods: passive learning (R), the baseline method using classification only (C), and the proposed LS+C. Our experiments still use the FRCNN model [21] with the RoI warping layer [2]. Compared to the PASCAL datasets, the MS COCO has more categories (80) and more images (80k for training and 40k for validation). Our experiments use the training set as the unlabeled image pool, and the validation set for testing.

The numbers shown in this section are averages over 3 trails for each method. All trials start from the same baseline object detectors, which are trained with 5,000 images selected from the unlabeled image pool. After then, each active learning algorithm is executed in 4 rounds. In each round, we select 1,000 images, add these images to the existing training set, and train a new model. Each model is trained with 12 epochs.

Results: Fig. 8a and Fig. 8b show the mAP curve and the relative saving of labeled images for the testing meth-

ods. Fig. 8a shows that classification-only method (C) does not have improvement over passive learning (R), which is not similar to the observations for the PASCAL 2012 in Section 4.1 and the PASCAL 2007 in Section 4.2. By incorporating the localization information, LS+C method can achieve 5% relative saving in the amount of annotation compared with passive learning, as shown in Fig. 8b.

4.4. SSD on PASCAL 2007

Experimental Setup: Here we test our algorithm on a different object detector: the single shot multibox detector (SSD) [17]. The SSD is a model without an intermediate region-proposal stage, which is not suitable for the localization-tightness based methods. We test the SSD on the PASCAL 2007 dataset where the training and validation sets (total 5,011 images) are used as the unlabeled image pool, and the test set (4,952 images) is used for testing. Input images are reiszed to 300×300.

Similar to the experimental settings in Section 4.1 and 4.2, the numbers shown in this section are averages over 5 trails.All trials start from the same baseline object detectors

which are trained with 500 images selected from the unlabeled image pool. After then, each active learning algorithm is executed in 15 rounds. A difference from previous experiments is that each model is trained with 40,000 iterations, not a fixed number of epochs. In our experiments, the SSD takes more iterations to converge. Consequently, when the number of labeled images in the training set is small, a fixed number of epochs means training with fewer number of iterations and the SSD cannot converge.

Results: Fig. 9a and Fig. 9b show the mAP curve and the relative saving of labeled images for the testing methods. Fig. 9a shows that both active learning method (C and LS+C) have improvements over passive learning (R). Fig. 9b shows that in order to achieve the same performance of passive learning with a training set consists of 2,300 to 3,500 labeled images, the proposed method (LS+C) can reduce the amount of image for annotation (12 - 22%) more than the baseline active learning method (C) (6 - 15%). In terms of average saving, LS+C is 29.0% better than the baseline method C.

5. Discussion

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.

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. Please refer to the supplemental material for more details.

Estimate of Localization Tightness: Our experiment shows that if the ground truth of bounding box is known, localization tightness can achieve best accuracies,

but the benefit degrades when using the estimated tightness instead. To analyze the impact of the estimate, after we trained the FRCNN-based object detector with 500 images of PASCAL2012 training set, we collected the ground-truth-based tightness and the estimated values of all detected boxes in the 5,215 test images.

Here shows a scatter plot where the coordinates of each point represents the two scores of a detected box. As this scatter plot shows an upper-triangular distribution, it implies that our estimate is most accurate when the proposals can tightly match the final detection boxes. Otherwise, it could be very different from the ground-truth value. This could partially explain why using the estimated cannot achieve the same performance as the ground-truth-based tightness.
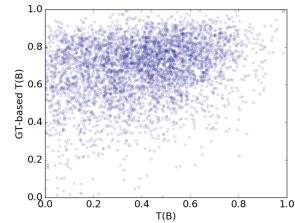
Computation Speed: Regarding the speed of our approach, as all testing object detector are CNN-based, the main speed bottleneck lies in the forwarding propagation. In our experiment with FRCNN-based detectors, for instance, forwarding propagation used 137 milliseconds per image, which is 82.5% of the total time when considering only classification uncertainty. The calculation of T_I has similar speed as U_C. The calculation of localization stability S_I needs to run the detector multiple times, and thus is slower than calculating other metrics.

Nevertheless, as these metrics are fully automatic to calculate, using our approach to reduce the number of images to annotate is still cost efficient. Considering that drawing a box from scratch can take 20 seconds in average [27], and checking whether a box tightly encloses an object can take 2 seconds [18], the extra overhead to check images with our metrics is small, especially that we can reduce 20 - 25% of images to annotate.

Acknowledgments

References

[1] A. Bietti. Active learning for object detection on satellite images. Technical report, California Institute of Technology, Jan 2012. 2, 3

[2] J. Dai, K. He, and J. Sun. Instance-aware semantic segmentation via multi-task network cascades. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2016. 5, 6, 7

[3] S. Dutt Jain and K. Grauman. Active image segmentation propagation. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2016. 1

[4] M. Everingham, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman. The PASCAL Visual Object Classes (VOC) challenge. *International Journal of Computer Vision (IJCV)*, 88(2):303–338, 2010. 3, 5, 6

[5] A. Freytag, E. Rodner, and J. Denzler. Selecting influential examples: Active learning with expected model output changes. In *European Conference on Computer Vision (ECCV)*. Springer, 2014. 1

[6] R. Girshick. Fast R-CNN. In *International Conference on Computer Vision (ICCV)*, 2015. 2

[7] M. Hasan and A. K. Roy-Chowdhury. Continuous learning of human activity models using deep nets. In *European Conference on Computer Vision (ECCV)*. Springer, 2014. 1

[8] M. Hasan and A. K. Roy-Chowdhury. Context aware active learning of activity recognition models. In *International Conference on Computer Vision (ICCV)*, 2015. 1

[9] J. Hoffman, S. Guadarrama, E. Tzeng, R. Hu, J. Donahue, R. Girshick, T. Darrell, and K. Saenko. LSDA: Large scale detection through adaptation. In *Advances in Neural Information Processing Systems (NIPS)*, 2014. 2

[10] R. Islam. Active learning for high dimensional inputs using bayesian convolutional neural networks. Master's thesis, Department of Engineering, University of Cambridge, 8 2016. 2

[11] A. Kapoor, K. Grauman, R. Urtasun, and T. Darrell. Active learning with gaussian processes for object categorization. In *International Conference on Computer Vision (ICCV)*. IEEE, 2007. 1

[12] A. Kapoor, K. Grauman, R. Urtasun, and T. Darrell. Gaussian processes for object categorization. *International Journal of Computer Vision (IJCV)*, 88(2):169–188, 2010. 2

[13] V. Karasev, A. Ravichandran, and S. Soatto. Active frame, location, and detector selection for automated and manual video annotation. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2014. 2

[14] K. Konyushkova, R. Sznitman, and P. Fua. Introducing geometry in active learning for image segmentation. In *The IEEE International Conference on Computer Vision (ICCV)*, December 2015. 1

[15] D. D. Lewis and J. Catlett. Heterogeneous uncertainty sampling for supervised learning. In *International Conference on Machine Learning (ICML)*. Morgan Kaufmann, 1994. 2

[16] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick. Microsoft coco: Common objects in context. In *European Conference on Computer Vision (ECCV)*, 2014. 3, 5, 7

[17] W. Liu, D. Anguelov, D. Erhan, C. Szegedy, S. Reed, C.-Y. Fu, and A. C. Berg. Ssd: Single shot multibox detector. In *European Conference on Computer Vision (ECCV)*, 2016. 2, 5, 7

[18] D. P. Papadopoulos, J. R. R. Uijlings, F. Keller, and V. Ferrari. We don't need no bounding-boxes: Training object class detectors using only human verification. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2016. 1, 2, 8

[19] A. Prest, C. Leistner, J. Civera, C. Schmid, and V. Ferrari. Learning object class detectors from weakly annotated video. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, 2012. 2

[20] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi. You only look once: Unified, real-time object detection. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2016. 2

[21] S. Ren, K. He, R. Girshick, and J. Sun. Faster R-CNN: Towards real-time object detection with region proposal networks. In *Advances in Neural Information Processing Systems (NIPS)*, 2015. 2, 3, 5, 6, 7

[22] O. Russakovsky, L. J. Li, and L. Fei-Fei. Best of both worlds: Human-machine collaboration for object annotation. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2015. 2

[23] B. Settles. Active learning literature survey. *University of Wisconsin, Madison*, 52(55-66):11, 2010. 1, 2, 3

[24] K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014. 5

[25] S. Sivaraman and M. M. Trivedi. Active learning for on-road vehicle detection: A comparative study. *Mach. Vision Appl.*, 25(3):599–611, Apr. 2014. 2, 3

[26] F. Stark, C. Hazirbas, R. Triebel, and D. Cremers. Captcha recognition with active deep learning. In *GCPR Workshop on New Challenges in Neural Computation*, Aachen, Germany, 2015. 2

[27] H. Su, J. Deng, and L. Fei-Fei. Crowdsourcing annotations for visual object detection. In *Workshops at the Twenty-Sixth AAAI Conference on Artificial Intelligence*, 2012. 1, 8

[28] S. Tong and D. Koller. Support vector machine active learning with applications to text classification. *J. Mach. Learn. Res.*, 2:45–66, Mar. 2002. 2

[29] J. R. R. Uijlings, K. E. A. van de Sande, T. Gevers, and A. W. M. Smeulders. Selective search for object recognition. *International Journal of Computer Vision (IJCV)*, 104(2):154–171, 2013. 3

[30] S. Vijayanarasimhan and K. Grauman. Large-scale live active learning: Training object detectors with crawled data and crowds. *International Journal of Computer Vision (IJCV)*, 108(1-2):97–114, 2014. 2

Localization-Aware Active Learning for Object Detection

Supplementary Materials

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/highest score of all boxes

Figure 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).

(min(.)/max(.)), 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 is the highest probability out of K categories of box B, and T is the localization tightness of box B:

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.

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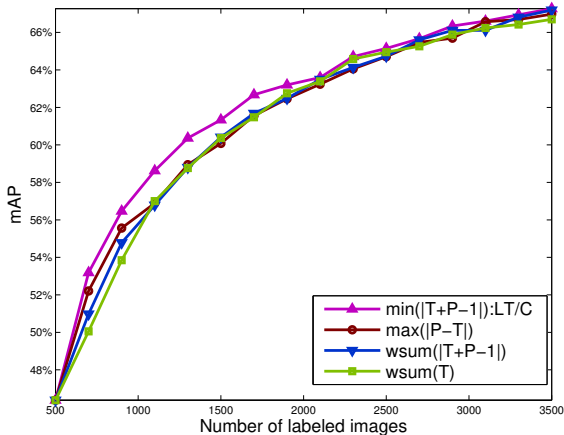
Figure 2: Mean average precision curve of different metrics of localization tightness on PASCAL 2007 detection dataset. Each point in the plot is an average of 5 trials.

For a set of unlabeled images, min(|T+P-1|), wsum(|T+P-1|), and wsum(T) methods choose images with lower scores to annotate in active learning; max(|P-T|) selects images with higher scores for annotation. 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 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

As mentioned in Sec. 5 in the main paper, 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).

We had discussed the case of perfect candidate windows in the main paper. In the following, we discuss more about the case of feature extractors are resilient to
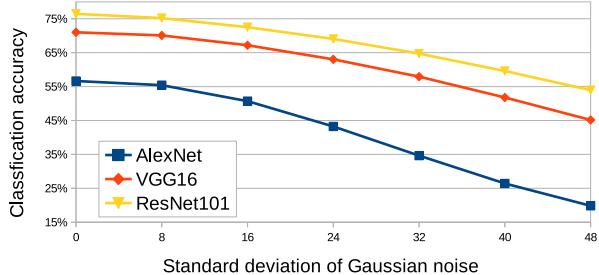

Figure 3: Top-1 classification accuracy of different neural network models when input images are corrupted by Gaussian noise on PASCAL 2012 validation dataset.

noise. 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 [2] 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 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
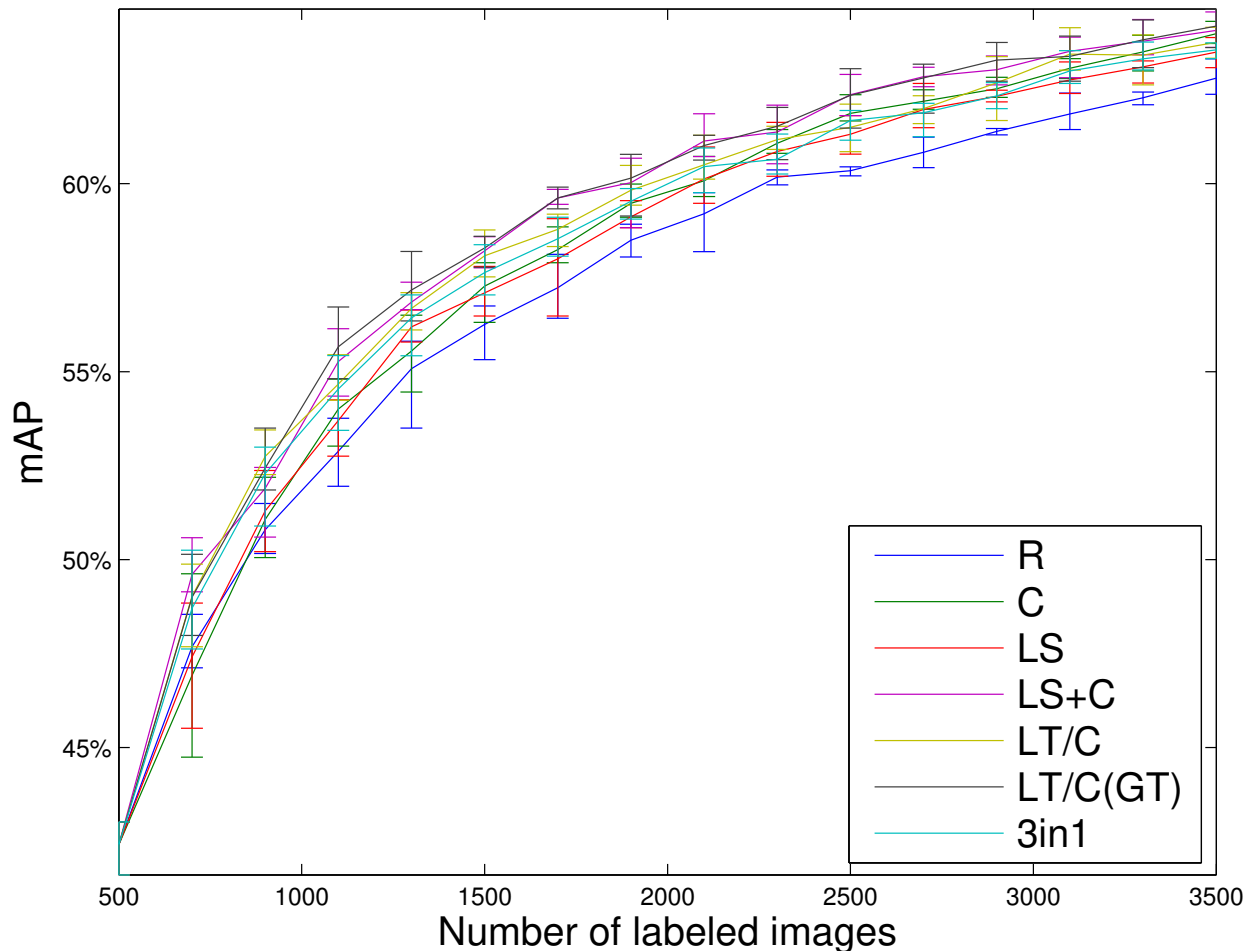
Figure 4: Mean average precision curve of different active learning methods on the PASCAL 2012 detection dataset. Each point in the plot is an average of 5 trials. The error bars represent the minimum and maximum values out of 5 trials at each point. This is a full version (LS and 3in1 added) of Fig. 5a in the main paper.

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.

On PASCAL 2012, combining all cues together does not work better than either LS+C or LT/C (Fig. 6a). On PAS-

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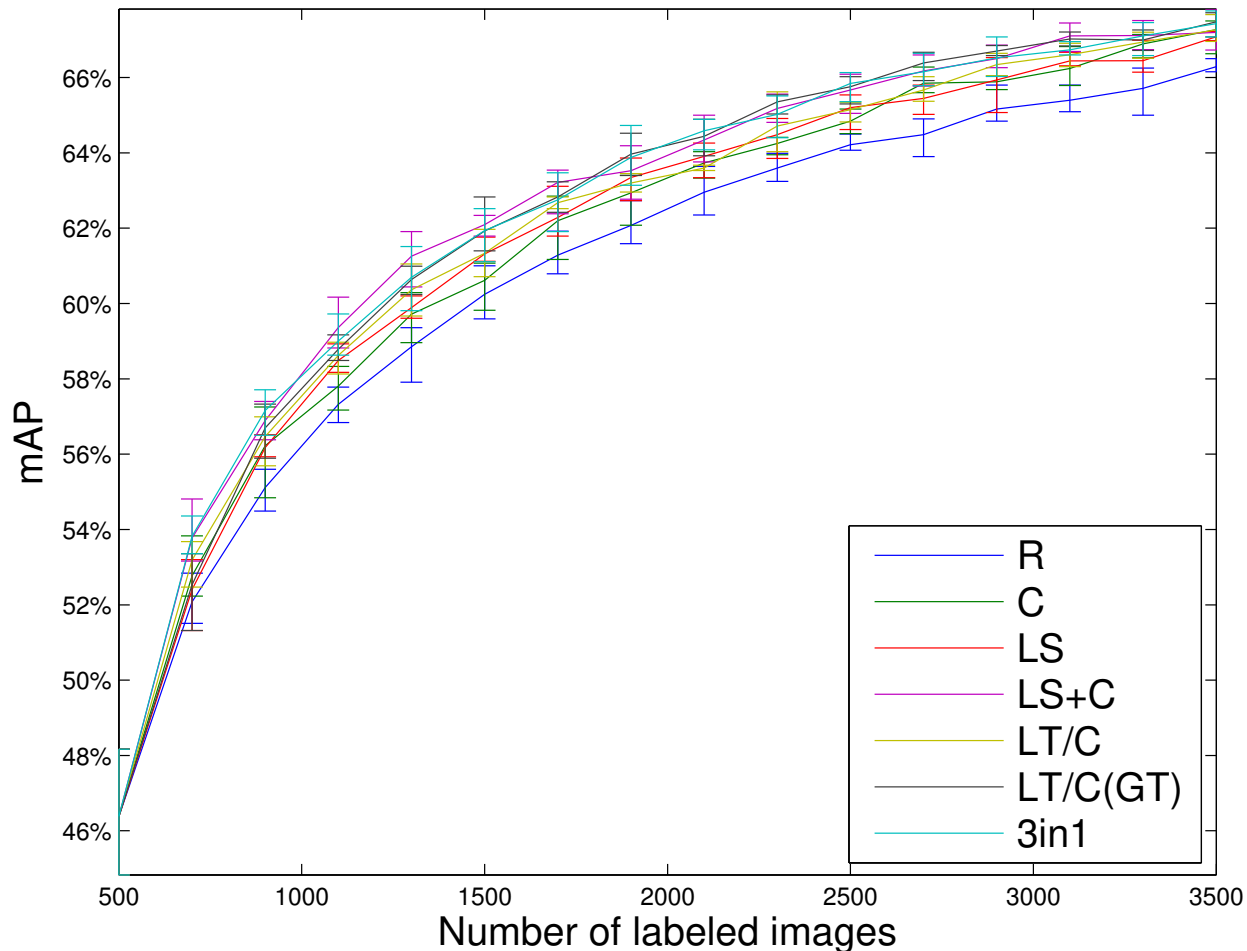
Figure 5: Mean average precision curve of different active learning methods on the PASCAL 2007 detection dataset. Each point in the plot is an average of 5 trials. The error bars represent the minimum and maximum values out of 5 trials at each point. This is a full version (LS and 3in1 added) of Fig.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

method	aero	bike	bird	boat*	bottle*	bus	car	cat	chair*	cow	table*	dog	horse	mbike	persn	plant*	sheep	sofa	train	tv	mAP
R	71.1	61.5	54.7	28.4	32.0	<u>68.1</u>	57.9	75.4	25.8	44.2	36.4	73.0	61.9	67.3	68.1	21.6	51.9	41.0	**65.5**	51.7	52.9
C	70.7	62.9	54.7	25.5	30.8	66.1	56.2	**78.1**	26.4	**54.5**	36.7	**76.9**	**68.3**	<u>67.7</u>	67.4	22.5	<u>57.7</u>	40.8	63.6	52.5	54.0
LS	**75.1**	61.3	**57.6**	**34.7**	<u>35.1</u>	65.1	58.2	75.4	29.3	43.9	38.5	70.7	57.5	66.1	68.5	23.0	56.1	40.3	64.2	53.6	53.7
LS+C	<u>73.9</u>	63.7	<u>56.9</u>	<u>29.6</u>	**35.2**	66.5	<u>58.5</u>	<u>77.9</u>	**31.3**	<u>50.8</u>	40.7	<u>73.8</u>	<u>65.4</u>	66.9	68.4	**24.8**	**58.0**	**44.9**	64.2	53.9	**55.3**
LT/C	69.8	**64.6**	54.6	29.5	33.8	**70.3**	**59.7**	75.5	29.5	46.3	**41.8**	73.0	62.5	**69.0**	**70.8**	23.2	56.5	42.8	<u>64.3</u>	**55.9**	<u>54.7</u>
3in1	72.9	<u>63.8</u>	52.7	29.5	33.6	66.4	57.2	76.0	**31.5**	48.5	<u>41.6</u>	72.2	62.6	67.6	<u>68.8</u>	<u>24.5</u>	57.6	<u>43.6</u>	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.

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 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.

method	aero	bike	bird	boat*	bottle*	bus	car	cat	chair*	cow	table*	dog	horse	mbike	persn	plant*	sheep	sofa	train	tv	mAP
R	61.6	67.2	54.1	40.0	33.6	64.5	73.0	73.9	34.5	60.8	52.2	69.3	**74.7**	66.6	67.1	25.9	52.1	54.2	66.1	54.9	57.3
C	56.9	68.0	54.9	36.8	34.4	68.1	71.7	**75.5**	34.0	68.6	51.0	71.4	74.7	65.2	65.9	24.9	60.0	53.9	63.0	57.4	57.8
LS	**64.4**	63.9	**56.3**	**45.1**	38.0	65.5	73.7	71.2	38.6	62.7	57.0	67.6	69.0	64.6	67.1	**29.6**	56.2	57.3	**68.6**	53.6	58.5
LS+C	61.5	64.4	55.8	40.2	38.7	66.3	73.8	74.7	39.6	68.0	56.3	**71.5**	73.8	67.2	66.7	27.7	**61.3**	57.0	65.6	57.4	**59.4**
LT/C	57.6	**69.7**	52.9	41.1	38.4	69.7	74.4	71.8	36.4	61.2	58.1	69.5	74.3	66.2	**67.8**	28.0	55.5	56.3	65.5	58.2	58.6
3in1	57.6	65.1	53.3	37.1	**39.0**	68.0	**74.6**	73.9	**39.8**	64.9	**58.5**	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.

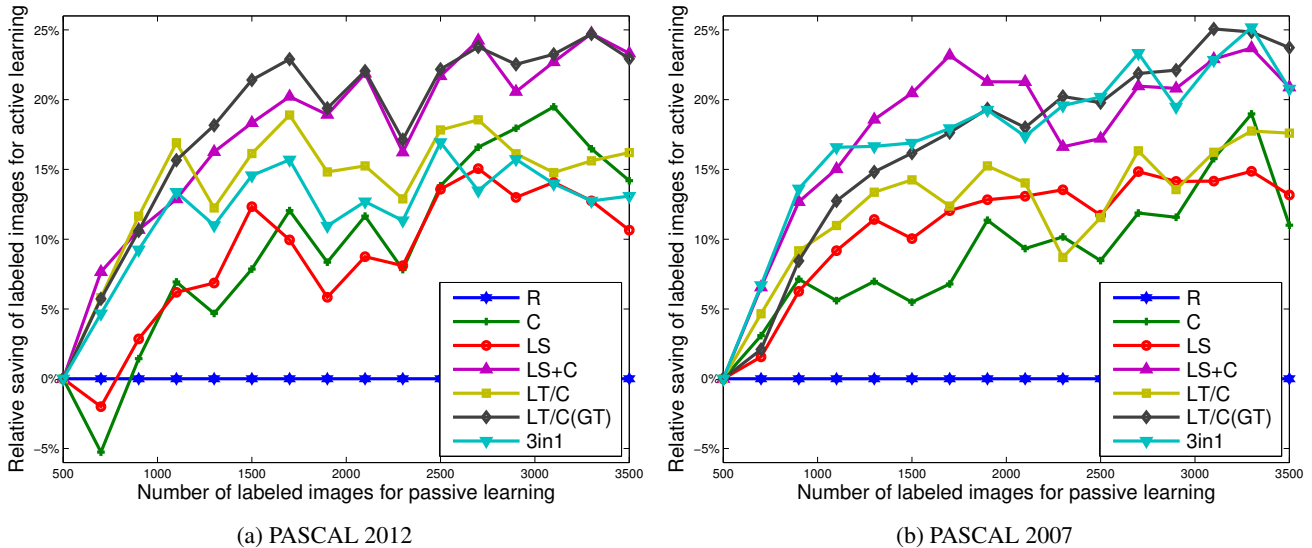(a) PASCAL 2012

(b) PASCAL 2007

Figure 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.

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

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

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.

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 [1] by using t-Distributed Stochastic Neighbor Embedding (t-SNE) [3]. 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 [5]. t-SNE is a technique
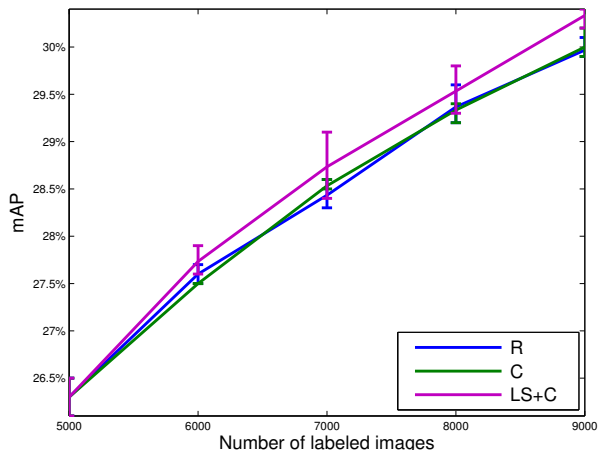
Figure 7: Mean average precision curve of different active learning methods on the MS COCO validation set. Each point in the plot is an average of 3 trials. The error bars represent the minimum and maximum values out of 3 trials at each point. This is a full version of Fig. 9a in the main paper.
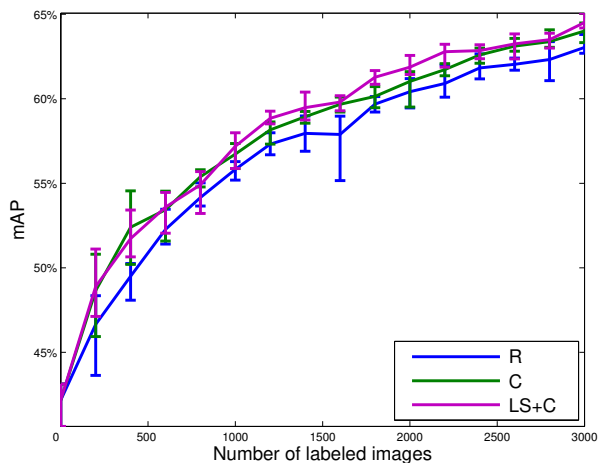

Figure 8: Mean average precision curve of different active learning methods with SSD on the PASCAL 2007 testing set. Each point in the plot is an average of 5 trials. The error bars represent the minimum and maximum values out of 5 trials at each point. This is a full version of Fig. 10a in the main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

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 [4] is trained on a training set of 500 labeled 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.

Figure 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.

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 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× 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
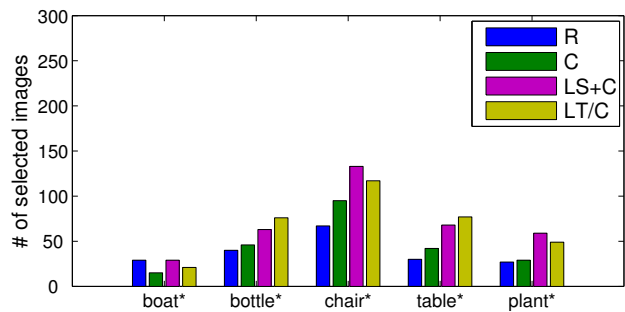

Figure 10: The number of selected images that contain objects belong to difficult classes by different active learning methods.

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 an-

method	aero	bike	bird	boat*	bottle*	bus	car	cat	chair*	cow	table*	dog	horse	mbike	persn	plant*	sheep	sofa	train	tv	mAP
R	68.3	61.5	<u>54.2</u>	27.8	30.4	<u>68.2</u>	<u>58.2</u>	76.3	28.4	44.8	31.1	<u>73.7</u>	64.1	<u>67.9</u>	66.7	21.9	52.4	<u>41.7</u>	**64.8**	55.5	52.9
C	<u>72.8</u>	<u>66.6</u>	50.8	28.5	34.8	64.3	54.3	<u>77.5</u>	27.2	**53.2**	36.3	**79.0**	**70.4**	66.5	<u>69.0</u>	21.9	<u>59.6</u>	38.8	60.6	54.5	54.3
LS+C	68.1	**68.0**	52.0	**34.2**	<u>34.9</u>	**70.0**	**59.9**	74.4	<u>30.3</u>	44.2	**42.1**	73.6	63.3	**69.7**	**71.7**	**28.5**	**60.2**	40.6	<u>64.4</u>	**59.0**	<u>55.5</u>
LT/C	**74.8**	64.8	**60.1**	<u>28.7</u>	**36.4**	63.9	58.1	**79.7**	**31.0**	<u>51.1</u>	<u>38.1</u>	72.9	<u>66.0</u>	66.9	67.2	<u>23.7</u>	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.
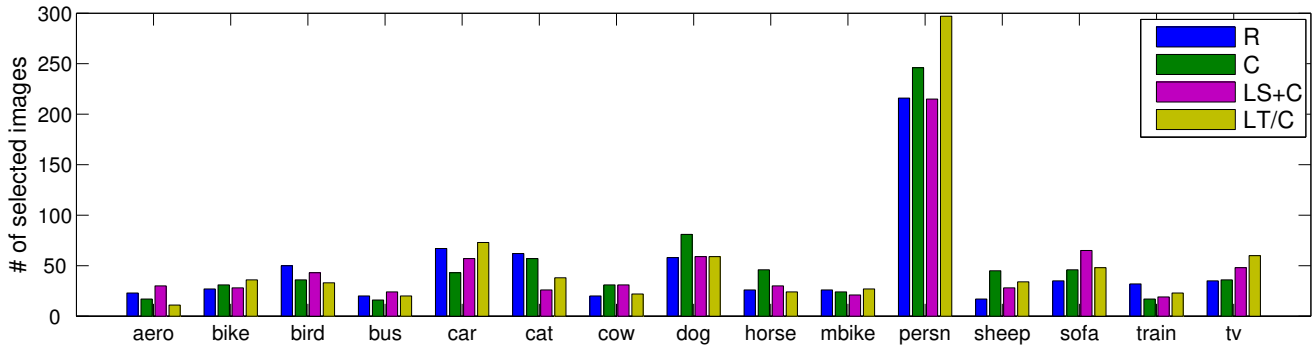

Figure 11: The number of selected images that contain objects belong to non-difficult classes by different active learning methods.

notation 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 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] M. Everingham, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman. The PASCAL Visual Object Classes (VOC) challenge. *International Journal of Computer Vision (IJCV)*, 88(2):303–338, 2010. 6

[2] I. J. Goodfellow, J. Shlens, and C. Szegedy. Explaining and harnessing adversarial examples. In *ICLR*, 2015. 2

[3] L. v. d. Maaten and G. Hinton. Visualizing data using t-sne. *Journal of Machine Learning Research*, 9(Nov):2579–2605, 2008. 6

[4] S. Ren, K. He, R. Girshick, and J. Sun. Faster R-CNN: Towards real-time object detection with region proposal networks. In *Advances in Neural Information Processing Systems (NIPS)*, 2015. 6

[5] K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014. 6

(a) Random (R)

(b) Classification (C)

(c) Localization stability + classification (LS+C)
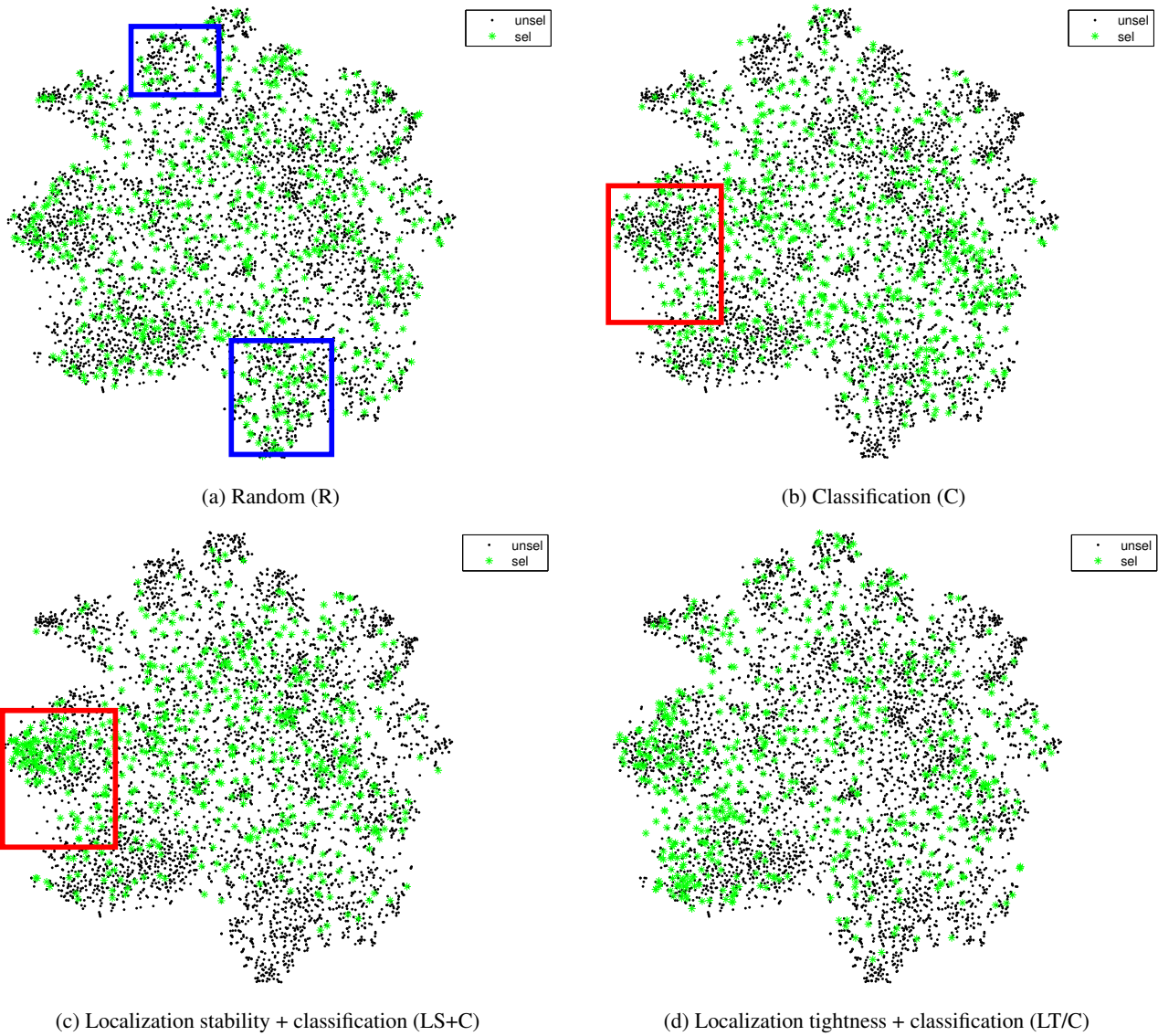
(d) Localization tightness + classification (LT/C)

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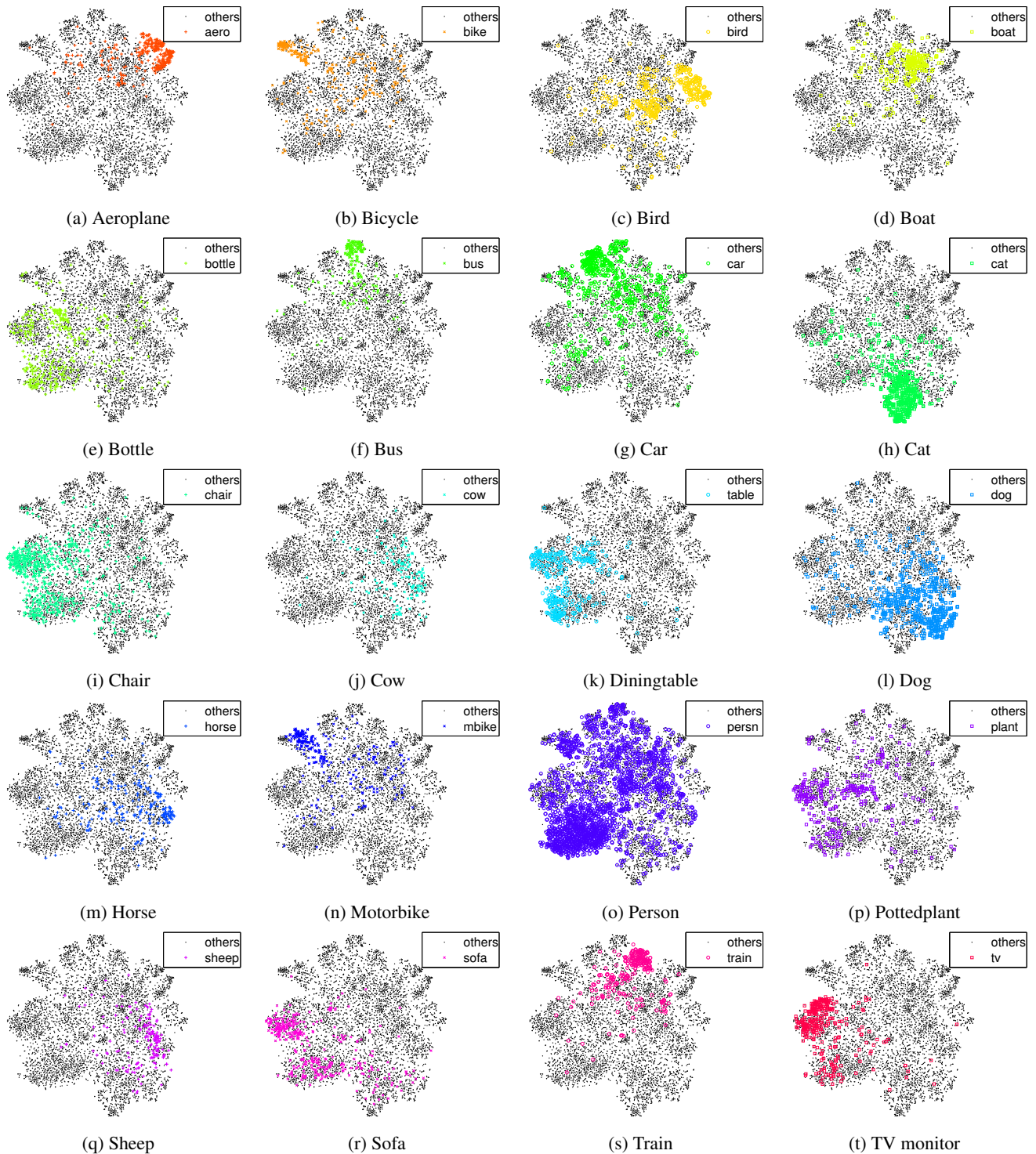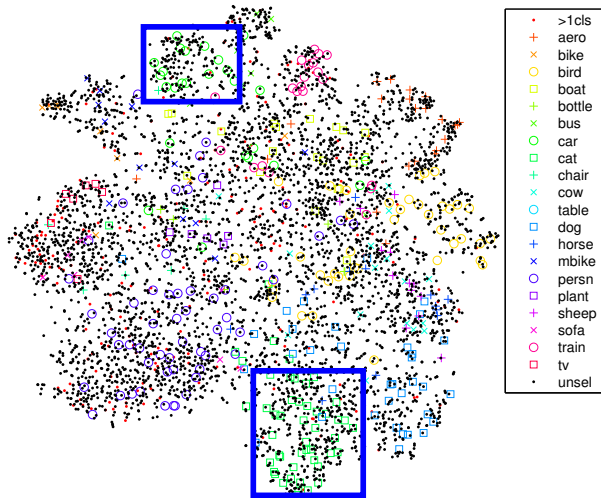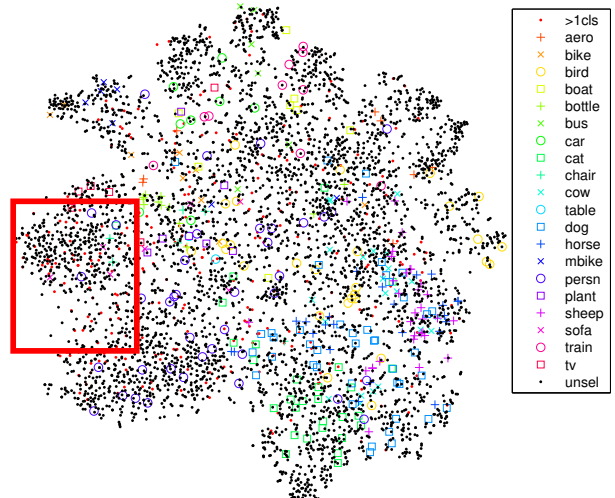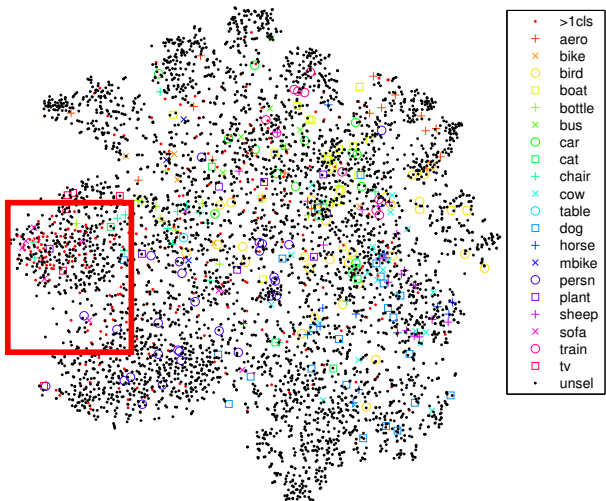
Figure 13: t-SNE embeddings of images for each category on the PASCAL 2012 training set. Different from Fig. 9, each colored point in the graphs represents an image that includes at least one object belongs to the target class. For example, each orange plus sign in (a) represents an image which has at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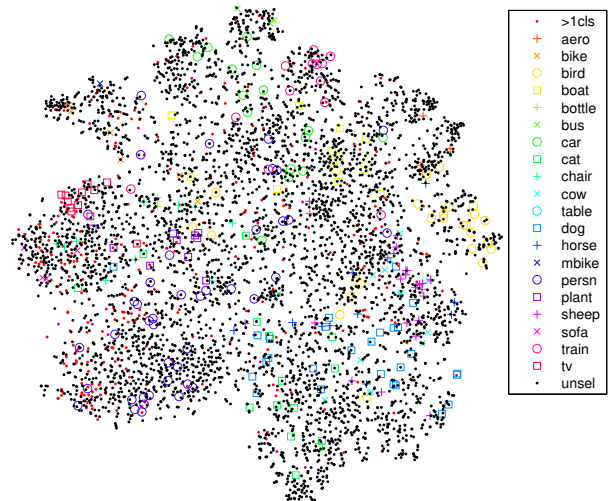
(a) Random (R)

(b) Classification (C)

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

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

Figure 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.