Active Object Detection with Epistemic Uncertainty and Hierarchical Information Aggregation

Younghyun Park, Soyeong Kim, Wonjeong Choi, Dong-Jun Han, Jaekyun Moon
Korea Advanced Institute of Science and Technology (KAIST)
{dnffkf369, best004, dnjswjd5457, djhan93}@kaist.ac.kr, jmoon@kaist.edu

Abstract

Despite the huge success of object detection, the training process still requires an immense amount of labeled data. Active learning has been proposed as a practical solution, but existing works on active object detection do not utilize the concept of epistemic uncertainty, which is an important metric for capturing the usefulness of the sample. Previous works also pay little attention to the relation between bounding boxes when computing the informativeness of an image. In this paper, we propose a new active object detection strategy that improves these two shortcomings of existing methods. We specifically consider a Bayesian framework and propose a new module termed model evidence head (MEH), to take advantage of epistemic uncertainty in object detection. We also propose hierarchical uncertainty aggregation (HUA), which realigns all bounding boxes into multiple levels and aggregates uncertainties in a bottom-up order, to compute the informativeness of an image. Experimental results show that our method outperforms existing state-of-the-art methods by a considerable margin.

1. Introduction

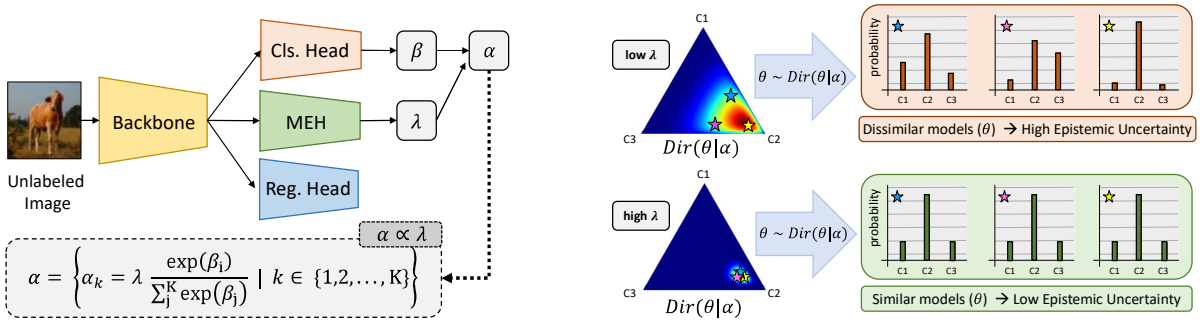
Computer vision tasks such as semantic segmentation [3, 13, 16] and object detection [10, 12, 15] typically require a large labeled dataset to train the model. Labeling all data samples in complex vision tasks requires intensive labor of human experts. Active learning, which gradually labels a set of samples based on the informativeness (e.g., uncertainty), is a promising solution for this problem.

Although active learning has been extensively studied for classification, only a few past works focus on active object detection [6, 8, 19, 21, 22] despite its practical importance. Furthermore, existing works on active object detection have two limitations. First, when computing the informativeness of an image, most previous works only use the aleatoric uncertainty, not taking the *epistemic uncertainty* into account. Epistemic uncertainty, also known as knowledge uncertainty, captures the lack of knowledge of a model (caused by a lack of data) and can be reduced when we have large amounts of data. Aleatoric uncertainty, on the other hand, captures the noise inherent in the observed data and is irreducible. As stated in [7, 9, 14], epistemic uncertainty can reflect the usefulness of samples better than aleatoric uncertainty. Secondly, previous works on active object detection generally ignore the relation between objects when computing the informativeness of an image: informativeness is often defined as the maximum or mean of the uncertainty values of all bounding boxes capturing the image. This can be a problem because a cluttered image with many objects belonging to various categories can be enforced to have a similar uncertainty value relative to just a simple image with only a few objects belonging to a single category.

Contributions. In this paper, we propose a new active learning method tailored to object detection which can solve the above two problems. First, we propose to utilize epistemic uncertainty in object detection to select samples in low density region. To this end, we adopt a Bayesian framework which employs Dirichlet-Categorical distribution and design a new module termed model evidence head (MEH), which solely predicts the model evidence independently of the class confidence. Secondly, we propose hierarchical uncertainty aggregation (HUA), a new method for computing the informativeness of an image. HUA reorganizes all bounding boxes into several levels and aggregates uncertainties corresponding to each level in a bottom-up manner. The proposed method beats SoTA works with RetinaNet and SSD as base models on PASCAL VOC and MS-COCO.

2. Proposed Method

Active learning has multiple cycles. At each cycle, the network selects the most informative data from the unlabeled data pool. Human oracles then label the selected data and update labeled/unlabeled data pools. In the following, we describe our method for computing the informativeness of an image in object detection. Subsection 2.1 describes how we estimate epistemic uncertainty of a bounding box. Based on the results in Subsection 2.1, we describe how we compute the informativeness of an image in Subsection 2.2.



(a) Forward propagation of an unlabeled image

(b) Calculation of epistemic uncertainty using model sampling

Figure 1. An overview of uncertainty computation of a bounding box in an unlabeled image. (a) First, an image goes through the network and classification head produces class confidences $\beta = \{\beta_i\}_{i=1}^K$ while model evidence head (MEH) produces model evidence λ. β and λ are used to compute a parameter set α of Dirichlet distribution $Dir(\theta|\alpha)$. (b) Now based on α, parameters θ of categorical distribution $Cat(\theta)$ are sampled from $Dir(\theta|\alpha)$. Epistemic uncertainty is then computed. Note that a larger λ indicates a larger α, making $Dir(\theta|\alpha)$ sharper; sharp $Dir(\theta|\alpha)$ produces similar $Cat(\theta)$, decreasing the epistemic uncertainty.

2.1. Evidential Learning for Epistemic Uncertainty

Fig. 1 shows the process for computing epistemic uncertainty. Under Dirichlet-Categorical Bayesian framework, we propose to predict class confidences $\beta = \{\beta_k\}_{k=1}^K$ and model evidence λ, to obtain concentration parameter $\alpha = \{\alpha_k\}_{k=1}^K$ and construct prior Dirichlet distribution $Dir(\theta|\alpha)$ for each bounding box. Epistemic uncertainty is then computed using model ensembles $\theta \sim Dir(\theta|\alpha)$ as

$$\underbrace{\mathcal{I}[x,\theta]}_{\text{Epistemic Unc.}} = \underbrace{\mathcal{H}\left[\mathbb{E}_{p(\theta|\alpha)}[p(x|\theta)]\right]}_{\text{Total Unc.}} - \underbrace{\mathbb{E}_{p(\theta|\alpha)}\left[\mathcal{H}[p(x|\theta)]\right]}_{\text{Aleatoric Unc.}} \quad (1)$$

where $\mathcal{H}$ denotes Shannon entropy, and θ parameteriezes categorical likelihood $Cat(\theta)$. $p(x|\theta)$ and $p(\theta|\alpha)$ are probability functions of categorical and Dirichlet distributions.

Evidential object detector. Typical object detectors, where classification head predicts parameters θ of categorical distribution, cannot compute epistemic uncertainty. Inspired by evidential deep learning (EDL) [2, 18, 23], we adopt a classification head that predicts high-order Dirichlet distribution $Dir(\theta|\alpha)$, which is a conjugate prior of lower-order categorical likelihood $Cat(\theta)$. To fit our evidential model to data, we first compute the marginal likelihood $p(x|\alpha) = \int p(x|\theta)p(\theta|\alpha)d\theta$, which can be written in a closed-form thanks to the Dirichlet-Categorical conjugacy:

$$p(x = k|\alpha) = \frac{\alpha_k}{\sum_c \alpha_c}. \quad (2)$$

The network is then optimized to minimize negative log marginal likelihood $L_{cls} = -\sum_k y_k \log(p(x = k|\alpha))$, where y is an one-hot label vector. At inference, the expected probability for the k-th category is computed as $\hat{p}_k = \alpha_k/S$, where $S = \sum \alpha_k$ is the Dirichlet strength.

To obtain the concentration parameter α, previous works [18, 23] apply ReLU as $\alpha'_k = \text{ReLU}(\beta_k) + 1$. However, the absence of exponential terms in ReLU tends to make model overly underconfident, resulting in much lower mAP performance. For example, in the case of 80-way classification, β_k should be 7820 to achieve $\hat{p}_k = 0.99$ even when $\beta_i = 0$ for $i \neq k$. Instead, we apply softmax as $\alpha_k = \frac{\exp(\beta_k)}{\sum_c \exp(\beta_c)}$ to produce sufficiently confident prediction.

Model evidence head (MEH). Although softmax enables confident prediction, it enforces the Dirichlet strength S to be 1. This makes the entropy of Dirichlet unchangeable and deprives the model of the ability to predict model evidence. To tackle this issue, we introduce a model evidence head (MEH) which solely predicts model evidence λ. When λ is obtained from MEH and class confidences $\{\beta_k\}_{k=1}^K$ are obtained from classification head, we propose to re-scale the concentration parameter α as

$$\alpha_k = \lambda \frac{\exp(\beta_k)}{\sum_c \exp(\beta_c)}. \quad (3)$$

Here, λ re-scales S and makes the distribution $Dir(\theta|\alpha)$ either concentrated or flat. Fig. 1a describes the role of MEH and Fig. 1b illustrates the effect of λ. We further validate the effects of softmax and λ in Section 3.

When training MEH, we start from the intuition that the model would be uncertain when it predicts a high loss from its prediction. We interpret the inverse of the output of MEH, $\frac{1}{\lambda}$, as a predictive value for target loss: when the target loss $l_{s,i}$ and loss prediction $\hat{l}_{s,i} = \frac{1}{\lambda}_{s,i}$ are given for bounding box i at scale s, the loss for image I is defined as

$$L_{MEH}(I) = \sum_s \sum_i (\hat{l}_{s,i} - l_{s,i})^2. \quad (4)$$

Note that λ is utilized only for re-scaling the Dirichlet strength; λ has no effect on $p(x|\theta)$ or $\hat{p}_k$, since λ will naturally vanish when division α_k/S occurs. Hence, the MEH network Φ_{MEH} and remaining network $\Phi \setminus \Phi_{MEH}$ can be updated in a disjoint manner. We in turn optimize Φ_{MEH} and $\Phi \setminus \Phi_{MEH}$, thus training of Φ_{MEH} never affects the performance of the primary object detector; this resolves the instability issue of previous EDL approaches.

(a) Bounding boxes are matched to the nearest object

(b) Uncertainties are aggregated to get a single score

Figure 2. An overview of the proposed hierarchical uncertainty aggregation. (a) Bounding boxes are grouped or ignored based on IoU with the near objects. (b) Bounding boxes in an image are hierarchically realigned based on object, scale, category to which the boxes belong. Uncertainty of bounding boxes in the same level are aggregated and then passed to the higher level. For the above example, a set of functions (sum, max, sum) is chosen for each information level (object, scale, class).

Applications. Evidential learning can be applied not only to softmax-based object detectors like SSD [12] but also to sigmoid-based detectors like RetinaNet [10]: given sigmoid-based prediction $p \in [0, 1]$, focal loss can be defined as $FL(p_t) = -(1-p_t)^{\gamma}\log(p_t)$, $p_t = p$ if $y = 1$, $p_t = 1 - p$ otherwise. To enable evidential learning, we replace sigmoid-based binary prediction p by Dirichlet-Categorical multiclass prediction $p(x|\alpha)$ in equation (2).

2.2. Hierarchical Uncertainty Aggregation (HUA)

Based on the epistemic uncertainty of a bounding box, in this subsection, we propose a new method for computing the total informativeness score (or uncertainty) of an image. It is evident that informativeness of an image cannot be fully captured by a single bounding box. However, recent active object detection methods [4, 21, 22] typically compute informativeness of an image as the mean/maximum of all bounding boxes. Instead, we propose to realign bounding boxes into multiple levels and aggregate uncertainties in a bottom-up order, as described in Fig. 2.

Filtering bounding boxes. Single-stage object detectors such as RetinaNet [10] and SSD [12] generate bounding boxes at every scale, pixel and anchors. Since most boxes correspond to background, we first filter out background boxes whose $\max_k(\hat{p}_k)$ is lower than threshold γ_{score}.

Realigning bounding boxes. Besides uncertainty score, each bounding box contains much more information: object, scale, category it belongs to. Based on these information, we propose to realign the boxes into a hierarchical structure. First, boxes are matched to the nearest object depending on the IoU score. For example in Fig. 2a, the blue box is matched to "object", but the green box is ignored since IoU is lower than the threshold γ_{IoU}. Secondly, boxes are further grouped based on the scale to which they belong. Lastly, boxes are divided based on the category $(\text{argmax}_k\, \hat{p}_k)$ to which they belong.

Uncertainty aggregation. Once the bounding boxes are fully realigned, individual uncertainty scores are unified in a hierarchical order. As shown in Fig. 2b, uncertainties of the boxes in a lower level (e.g., class) are aggregated into a single value through a predefined aggregation function; the aggregated value is then passed to an upper level (e.g, scale). This aggregation repeats from the "class level" to "object level". Note that different types of aggregation functions can be adopted at different levels of information. We specifically propose to adopt sum operation when aggregating the uncertainties at the object level, since this reflects the number of objects in total informativeness of the image.

2.3. Selection of Informative Image

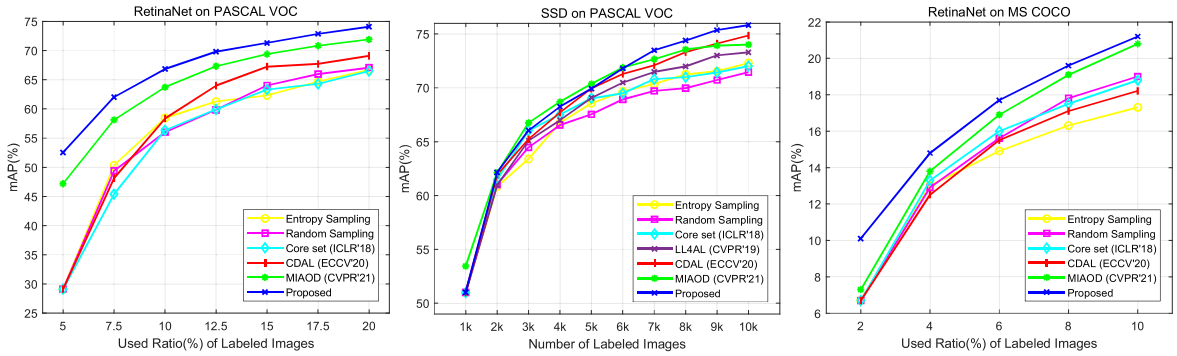
Now the epistemic uncertainties of all unlabeled images can be computed at each active learning cycle. While previous works typically select the top-k uncertain images, we propose to select filtered-out images as well. We stress that these images are also valuable since the machine was incapable of sensing any objects due to a lack of knowledge. We empirically found that composing 15% of selections with filtered images considerably increases the performance.

3. Experiments

Experimental details. For fair comparisons with previous works [1,20,22], we adopt RetinaNet [10] and SSD [12] which use ResNet50 and VGG16 as backbones. The structure of MEH is the same as the regression head. As for hyperparameter setting, we follow the settings of [1, 20, 22].

Dataset. Our work is validated on PASCAL VOC [5] and MS-COCO [11]. In the first active learning cycle for PASCAL VOC, 5% from 16,551 samples are randomly selected and 2.5% of the remaining set is additionally labeled until it reaches 20%. As for MS-COCO with 117,267 samples, labeled sets increased from 2% to 10% in steps of 2%.

Baselines. We compare our work with state-of-the-art works of MI-AOD [22], CDAL [1], LL4AL [20], Coreset [17]. Also, as basic baselines, *entropy sampling*, *ran-*



(a) RetinaNet on PASCAL VOC (b) SSD on PASCAL VOC (c) RetinaNet on MS-COCO

Figure 3. mAP score at the end of each active learning cycle.

dom sampling with vanilla RetinaNet/SSD are considered. 5 independent networks are trained with different seeds and the averaged performance is reported. For our scheme, we adopted (sum, max, sum) operations for each information level (object, scale, class) when applying HUA.

Comparison with state-of-the-arts. Fig. 3 shows mAP scores under various settings. Our proposed method exhibits a clear edge over other baselines giving the best performance. In Figs. 3a and 3c, the proposed method shows superiority in early cycles, proving that the proposed evidential focal loss is effective when labeled data is extremely limited. In Fig. 3b, SSD lags a bit at early cycles due to the inability of using evidential focal loss. However, our proposed methods (MEH and HUA) gradually increase the performance by selecting informative images.

Effects of MEH and HUA. Table 1 shows ablation studies on the proposed methods. While *ReLU* computes α with ReLU as in [18, 23], *Soft* uses softmax. All methods except *HUA* compute uncertainty of an image as the mean of all boxes, as in [4, 22]. At every cycle, proposed methods (Soft, MEH, HUA) increase the performance by a large margin.

Method	Ratio (%) of Labeled Samples						
	5.0	7.5	10.0	12.5	15.0	17.5	20.0
Random	29.13	50.35	58.45	61.27	62.31	64.67	66.72
Entropy	29.13	49.41	56.02	59.83	64.03	65.96	67.08
ReLU	22.46	27.83	31.39	33.18	35.61	37.03	38.95
ReLU+MEH	22.46	29.10	32.71	34.95	37.16	38.47	39.84
Soft	52.53	58.26	61.13	65.05	66.44	68.41	69.55
Soft+MEH	52.53	59.68	64.78	67.80	68.28	70.85	71.84
Soft+MEH+HUA	**52.53**	**62.02**	**66.84**	**69.82**	**71.31**	**72.86**	**74.08**

Table 1. mAP (%) of RetinaNet on PASCAL VOC.

Effect of model evidence λ. Fig. 4 illustrates the effect of λ. It can be seen that in a region where λ is high (warm area in 2^{nd} column), uncertainty becomes much smaller (compare 3^{rd} and 4^{th} columns).

Examples of easy and hard samples. For intuitive understanding, we display examples of easy and hard samples in Fig. 5. Easy examples tend to have just one unoccluded object. But, hard examples tend to have numerous objects, which are unclear or heavily occluded.

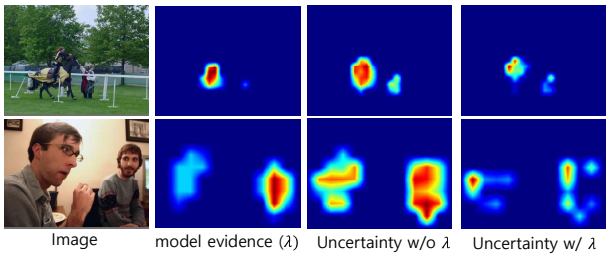


Figure 4. Effect of model evidence (λ) when calculating epistemic uncertainty. Warm color indicates high value. In the area with high model evidence, epistemic uncertainty decreases.

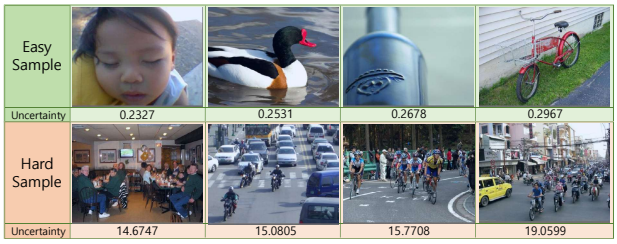


Figure 5. Examples of easy and hard samples. Images and their informativeness score are displayed together.

4. Conclusion

We proposed a new active learning method for object detection. With Bayesian framework and model evidence head, our scheme estimates the epistemic uncertainty of a bounding box. Also, our hierarchical aggregation strategy provides a new guideline for computing informativeness of an image. Our scheme presents an up-and-coming direction for active object detection, where estimating epistemic uncertainty accurately yet quickly is of crucial importance.

5. Acknowledgement

This work was conducted by Center for Applied Research in Artificial Intelligence (CARAI) grant funded by DAPA and ADD (UD190031RD).

References

[1] Sharat Agarwal, Himanshu Arora, Saket Anand, and Chetan Arora. Contextual diversity for active learning. In *European Conference on Computer Vision*, pages 137–153. Springer, 2020. 3

[2] Alexander Amini, Wilko Schwarting, Ava Soleimany, and Daniela Rus. Deep evidential regression. *Advances in Neural Information Processing Systems*, 33:14927–14937, 2020. 2

[3] Liang-Chieh Chen, Yukun Zhu, George Papandreou, Florian Schroff, and Hartwig Adam. Encoder-decoder with atrous separable convolution for semantic image segmentation. In *Proceedings of the European conference on computer vision (ECCV)*, pages 801–818, 2018. 1

[4] Jiwoong Choi, Ismail Elezi, Hyuk-Jae Lee, Clement Farabet, and Jose M Alvarez. Active learning for deep object detection via probabilistic modeling. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 10264–10273, 2021. 3, 4

[5] Mark Everingham, Luc Van Gool, Christopher KI Williams, John Winn, and Andrew Zisserman. The pascal visual object classes (voc) challenge. *International journal of computer vision*, 88(2):303–338, 2010. 3

[6] Di Feng, Xiao Wei, Lars Rosenbaum, Atsuto Maki, and Klaus Dietmayer. Deep active learning for efficient training of a lidar 3d object detector. In *2019 IEEE Intelligent Vehicles Symposium (IV)*, pages 667–674. IEEE, 2019. 1

[7] Danijar Hafner, Dustin Tran, Timothy Lillicrap, Alex Irpan, and James Davidson. Reliable uncertainty estimates in deep neural networks using noise contrastive priors. 2018. 1

[8] Elmar Haussmann, Michele Fenzi, Kashyap Chitta, Jan Ivanecky, Hanson Xu, Donna Roy, Akshita Mittel, Nicolas Koumchatzky, Clement Farabet, and Jose M Alvarez. Scalable active learning for object detection. In *2020 IEEE intelligent vehicles symposium (iv)*, pages 1430–1435. IEEE, 2020. 1

[9] Eyke Hüllermeier and Willem Waegeman. Aleatoric and epistemic uncertainty in machine learning: An introduction to concepts and methods. *Machine Learning*, 110(3):457–506, 2021. 1

[10] Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. Focal loss for dense object detection. In *Proceedings of the IEEE international conference on computer vision*, pages 2980–2988, 2017. 1, 3

[11] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *European conference on computer vision*, pages 740–755. Springer, 2014. 3

[12] Wei Liu, Dragomir Anguelov, Dumitru Erhan, Christian Szegedy, Scott Reed, Cheng-Yang Fu, and Alexander C Berg. Ssd: Single shot multibox detector. In *European conference on computer vision*, pages 21–37. Springer, 2016. 1, 3

[13] Jonathan Long, Evan Shelhamer, and Trevor Darrell. Fully convolutional networks for semantic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3431–3440, 2015. 1

[14] Vu-Linh Nguyen, Mohammad Hossein Shaker, and Eyke Hüllermeier. How to measure uncertainty in uncertainty sampling for active learning. *Machine Learning*, 111(1):89–122, 2022. 1

[15] Joseph Redmon, Santosh Divvala, Ross Girshick, and Ali Farhadi. You only look once: Unified, real-time object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 779–788, 2016. 1

[16] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical image computing and computer-assisted intervention*, pages 234–241. Springer, 2015. 1

[17] Ozan Sener and Silvio Savarese. Active learning for convolutional neural networks: A core-set approach. *arXiv preprint arXiv:1708.00489*, 2017. 3

[18] Murat Sensoy, Lance Kaplan, and Melih Kandemir. Evidential deep learning to quantify classification uncertainty. *Advances in Neural Information Processing Systems*, 31, 2018. 2, 4

[19] Jong-Chyi Su, Yi-Hsuan Tsai, Kihyuk Sohn, Buyu Liu, Subhransu Maji, and Manmohan Chandraker. Active adversarial domain adaptation. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 739–748, 2020. 1

[20] Donggeun Yoo and In So Kweon. Learning loss for active learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 93–102, 2019. 3

[21] Weiping Yu, Sijie Zhu, Taojiannan Yang, Chen Chen, and Mengyuan Liu. Consistency-based active learning for object detection. *arXiv preprint arXiv:2103.10374*, 2021. 1, 3

[22] Tianning Yuan, Fang Wan, Mengying Fu, Jianzhuang Liu, Songcen Xu, Xiangyang Ji, and Qixiang Ye. Multiple instance active learning for object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5330–5339, 2021. 1, 3, 4

[23] Xujiang Zhao, Feng Chen, Shu Hu, and Jin-Hee Cho. Uncertainty aware semi-supervised learning on graph data. *Advances in Neural Information Processing Systems*, 33:12827–12836, 2020. 2, 4