TSBP: Improving Object Detection in Histology Images via Test-time Self-guided Bounding-box Propagation

Tingting Yang[1], Liang Xiao[1 ✉], and Yizhe Zhang[1 ✉]

[1]School of Computer Science and Engineering, Nanjing University of Science and Technology, China
yizhe.zhang.cs@gmail.com

Abstract. A global threshold (e.g., 0.5) is often applied to determine which bounding boxes should be included in the final results for an object detection task. A higher threshold reduces false positives but may result in missing a significant portion of true positives. A lower threshold can increase detection recall but may also result in more false positives. Because of this, using a preset global threshold (e.g., 0.5) applied to all the bounding box candidates may lead to suboptimal solutions. In this paper, we propose a Test-time Self-guided Bounding-box Propagation (TSBP) method, leveraging Earth Mover's Distance (EMD) to enhance object detection in histology images. TSBP utilizes bounding boxes with high confidence to influence those with low confidence, leveraging visual similarities between them. This propagation mechanism enables bounding boxes to be selected in a controllable, explainable, and robust manner, which surpasses the effectiveness of using simple thresholds and uncertainty calibration methods. Importantly, TSBP does not necessitate additional labeled samples for model training or parameter estimation, unlike calibration methods. We conduct experiments on gland detection and cell detection tasks in histology images. The results show that our proposed TSBP significantly improves detection outcomes when working in conjunction with state-of-the-art deep learning-based detection networks. Compared to other methods such as uncertainty calibration, TSBP yields more robust and accurate object detection predictions while using no additional labeled samples. The code is available at https://github.com/jwhgdeu/TSBP.

Keywords: Object Detection · Histology Images · Self-guided Detection Refinement · Test-time Method

1 Introduction

Object detection methods enable automated localization and classification of key objects in images, and play an important role in various areas such as medical image analysis and disease diagnosis. In recent years, with the rapid development of deep learning, object detection has made significant improvements in the
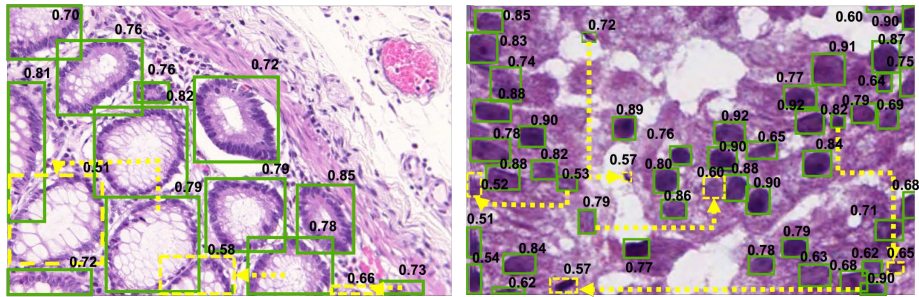
Fig. 1. Shown are examples of detected bounding boxes (b-boxes) with varying confidence scores in two test images. The yellow arrows highlight how higher-confident b-boxes influence lower confident b-boxes in TSBP based on visual similarities.

field of biomedical research, driving innovation in medical image diagnostics and biomedical studies [3,4,13].

A confidence score of each detected bounding-box is often generated by a detection model, which reflects the level of certainty that the detection is correct [5]. Fig. 1 showcases the predicted results of object detection models on two histology images. Higher confidence thresholds (e.g., 0.90) can effectively reduce false positives but lead to false negatives, while lower thresholds (e.g., 0.6) can increase the recall rate at the expense of more false positives. During the deployment (test-time) of an object detection model, using confidence thresholds to determine which bounding-boxes to include is critical but fewer studies were conducted in this aspect [19]. A common approach involves testing and evaluating different thresholds on a validation set, followed by selecting the optimal threshold as the deployment confidence threshold [21]. However, this approach has two notable limitations. First, object detection tasks involve diverse objects with varying categories, sizes, and shapes. Consequently, using a single threshold may not effectively cater to all the objects. Second, estimating a threshold using data inevitably incurs the data shift problem [14]. That is, when the test set exhibits data shift with respect to the training and validation sets, the estimated threshold can be ineffective for test samples.

Predictions with higher confidence scores are statistically more likely to be correct than those with lower confidence scores. In this paper, we aim to utilize bounding-box predictions with high confidence to rectify the bounding-box predictions with lower confidence. More specifically, we propose a new method called Test-time Self-guided Bounding-box Propagation (TSBP). TSBP propagates information of bounding-boxes with high confidence to influence bounding-boxes with lower confidence, with the goal of adjusting the class labels associated to bounding-boxes with low prediction confidence. Our main contributions can be summarized in the following three aspects.

- We explore a practical test-time task aimed at improving object detection performance in histology images. We illustrate the efficacy of a test-time optimization procedure in refining predictions by leveraging visual similarities

among test samples. Our findings underscore the potential benefits of this approach in enhancing detection accuracy.

– To tackle this challenge, we introduce a novel and adaptable method termed Test-time Self-guided Bounding-box Propagation (TSBP). TSBP employs an iterative matching and selection process, leveraging high-confidence bounding-boxes to refine low-confidence ones by exploiting their visual similarities. Unlike conventional confidence calibration techniques, TSBP operates without such requirements for label adjustment.

– Experiments, conducted on gland detection and cell detection tasks in histology images, show that our TSBP method can substantially enhance detection outcomes. Compared to known confidence calibration methods, TSBP leads more robust and accurate object detection predictions.

2   Related Work

Our TSBP method is related to confidence calibration, as TSBP can be viewed as a method for dynamically adjusting (or calibrating) the confidence score for each prediction (bounding-box) of a model. The goal of confidence calibration is to align the predicted confidences of a model with their actual accuracy [7]. Studies have shown that deep neural networks (DNNs) tend to make overconfident predictions [16], leading to inaccurate confidence estimation. Post-processing calibration, which requires parameter estimation on a validation set, is the most widely-used approach for confidence calibration, and such methods include histogram binning [22], Bayesian binning [15], Platt scaling [17], and Beta calibration [10]. In [12], calibration results were improved by considering both confidence scores and bounding-box information of object detection models. In [6], histogram binning calibration was modified under the condition of bounding-box sizes.

In traditional classification tasks, the maximum class probability (MCP) is commonly used as the confidence score [9]. This approach assigns the highest softmax output as the confidence score, which means that even incorrect predictions can result in high confidence values. Considering that MCP can lead to the overlap of confidence scores between correct and incorrect predictions, a confidence module called ConfidNet [2] was introduced into the original classification model to output confidence predictions that are closer to the true class probability (TCP). To suppress low-quality prediction boxes, CPSS-FAT [20] multiplies a consistent location quality estimation (CLQE) score by a class score as the confidence score.

3   Methodology

Fig. 2 provides a high-level overview of the proposed method. We utilize a trained object detection model (e.g., DiffusionDet [1]) on a set of test images, denoted as $X = x_1, x_2, \ldots, x_n$. For each image $x_i \in \mathbb{R}^{3 \times H \times W}$, we obtain a collection of bounding-boxes $\text{box}_{i,k} = u_{i,k}, v_{i,k}, w_{i,k}, h_{i,k}, c_{i,k}, s_{i,k}, k = 1, 2, \ldots, z_i$. As the
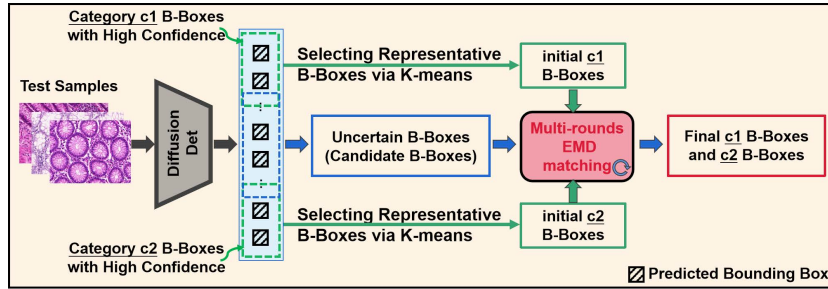
Fig. 2. Overview of Test-time Self-guided Bounding-box Propagation (TSBP).

number of generated bounding-boxes may vary for each image during inference, we use z_i to represent the count of bounding-boxes produced for image x_i. Here, $(u_{i,k}, v_{i,k}, w_{i,k}, h_{i,k})$ indicates the positional attributes of the corresponding bounding-box: $(u_{i,k}, v_{i,k})$ denotes the top-left corner coordinates, while $(w_{i,k}, h_{i,k})$ represent its width and height, respectively. $c_{i,k}$ denotes the category label of the bounding-box, while $s_{i,k}$ represents its confidence score, ranging from 0 to 1. A high value of $s_{i,k}$ indicates the model's strong confidence in the category label assignment, whereas a low value suggests being uncertain. Below, we describe our proposed method for refining the label assignment of the bounding-boxes, utilizing the higher-confidence bounding-boxes to influence those with lower confidence.

3.1 Initialization

In the subsequent matching process, we need to compare the feature similarity between bounding-boxes. Therefore, we introduce an additional feature information $\text{feat}_{i,k}$ for each bounding-box $\text{box}_{i,k}$. $\text{feat}_{i,k}$ is obtained by cropping the image patch inside the bounding-box $\text{box}_{i,k}$ at the corresponding position on the original image x_i, and then applying a feature extractor, e.g., ResNet-50 [8], to the cropped image patch. We denote feature extractor as π, and then the feature information $\text{feat}_{i,k}$ is obtained by $\text{feat}_{i,k} = \pi(x_i[v_{i,k} : (v_{i,k} + h_{i,k}), u_{i,k} : (u_{i,k}+w_{i,k})])$. TSBP takes all the bounding-boxes as its input. The task of TSBP is to reassign class categories to those bounding-boxes with low confidence scores.

For simplicity, suppose there are two class categories, denoted as $c1$ and $c2$. We set confidence thresholds, s^{c1} and s^{c2}. For bounding-boxes $\text{box}_{i,k}$ with $c_{i,k} = c1$ and confidence scores $s_{i,k} > s^{c1}$, we consider them as high-confident (HF) $c1$ bounding-boxes. Similarly, for bounding-boxes $\text{box}_{i,k}$ with $c_{i,k} = c2$ and confidence scores $s_{i,k} > s^{c2}$, we consider them as high-confident (HF) $c2$ bounding-boxes. The found HF $c1$ and $c2$ bounding-boxes are set as the initial confirmed $c1$ and $c2$ bounding-boxes. We then select a set of representative bounding-boxes from HF $c1$ and $c2$ bounding-boxes for the following Multi-Round EMD Matching step. To achieve more efficient and effective matching performance, we use a classic K-means algorithm (the number K of clusters

is set as 25 by default). All the bounding-boxes not added to the confirmed bounding-box set are considered as candidate bounding-boxes.

To ensure the reliability of the subsequent multi-round EMD matching and bounding-box propagation, we proceed to compute distance constraints for each class category. For every HF $c1$ bounding-box, we compute the shortest Euclidean distance between this bounding-box and the other HF $c1$ bounding-boxes. This process is repeated for all HF $c1$ bounding-boxes, and then the average of the found shortest distances is computed to establish the distance constraint for class category $c1$, denoted as D^{c1}_{max}. Similarly, we execute the same procedure for HF $c2$ bounding-boxes to obtain D^{c2}_{max}. Both D^{c1}_{max} and D^{c2}_{max} will be applied in the initial stage of the multi-round EMD matching.

3.2   Multi-Round EMD Matching and Bounding-Box Propagation

The multi-round bounding-box propagation comprises two stages. In the first stage, propagation is carried out with stricter constraints when admitting newly confirmed bounding-boxes (via utilizing the above obtained D^{c1}_{max} and D^{c2}_{max}). This is done to ensure that the bounding-boxes added during this early stage are more likely to have correct label assignments. In the second stage, these constraints are relaxed to allow propagation to occur across a wider range of candidate bounding-boxes.

In each round of propagation, suppose there are a candidate bounding-boxes and m confirmed bounding-boxes. Among these confirmed bounding-boxes, t are of class category $c1$ and the rest, $m - t$, are of class category $c2$. The input to Earth Mover's Distance (EMD) consists of two signatures: $P = p_1, p_2, \ldots, p_a$ and $Q = q_1, q_2, \ldots, q_t, q_{t+1}, \ldots, q_m$, where P represents the candidate bounding-boxes and Q represents the confirmed bounding-boxes. Specifically, Q comprises t confirmed $c1$ boxes and $m - t$ confirmed $c2$ boxes.

The distance between p_i and q_j is denoted as $D_{i,j}$, reflecting the cost of matching a candidate bounding-box p_i with a confirmed box q_j. $D_{i,j}$ is computed using the Euclidean distance between the features of p_i and q_j (features obtained in Section 3.1). The matching flow between p_i and q_j is denoted by $f_{i,j}$, with a potential value of either 0 or 1. The matching optimization goal is:

$$\min_f \sum_{1 \leq i \leq a} \sum_{1 \leq j \leq m} f_{i,j} D_{i,j}, \tag{1}$$

subject to

$$\sum_{1 \leq i \leq a} f_{i,j} \leq 1, \ 1 \leq j \leq m, \tag{2}$$

$$\sum_{1 \leq j \leq m} f_{i,j} \leq 1, \ 1 \leq i \leq a, \tag{3}$$

$$\sum_{1 \leq i \leq a} \sum_{1 \leq j \leq m} f_{i,j} = \min(a, m), \tag{4}$$

$$f_{i,j} \in \{0,1\}, 1 \leq i \leq a, 1 \leq j \leq m. \tag{5}$$

Equation (2) indicates that a confirmed bounding-box ($c1$ or $c2$) can be matched with at most one candidate bounding-box. Equation (3) indicates that a candidate bounding-box can be matched with at most one of the confirmed bounding-boxes. When the matching flow $f_{i,j}$ is 1, it means candidate bounding-box p_i has been successfully matched with confirmed bounding-box q_j.

In the **first stage** of propagation, for each confirmed $c1$ bounding-box associated with q_j, where $1 \leq j \leq t$, we identify p_{i^*} such that $f_{i^*,j} = 1$. If $D_{i^*,j}$ is smaller than the distance constraint D_{max}^{c1} (obtained in Section 3.1), we include the corresponding candidate bounding-box p_{i^*} in the confirmed $c1$ bounding-box set. The same procedure is followed for $c2$ bounding-boxes associated with q_j, where $t < j \leq m$. The confirmed box set Q is then updated to a new state with the newly added $c1$ and $c2$ bounding-boxes, while the remaining candidate bounding-boxes proceed to the next round of EMD matching (with updated P).

The **second stage** of propagation starts when no new bounding-boxes were added to the confirmed bounding-box sets for both $c1$ and $c2$ categories in the last round of propagation. Consequently, the D_{max}^{c1} and D_{max}^{c2} constraints are relaxed to allow more bounding-boxes to be added to the confirmed sets, **that is,** any matched pairs found in the EMD matching are used for adding the candidate bounding-boxes to the confirmed bounding-box sets. Specifically, for each confirmed $c1$ bounding-box associated with q_j, where $1 \leq j \leq t$, we find p_{i^*} such that $f_{i^*,j} = 1$, and then add the corresponding candidate bounding-box p_{i^*} to the confirmed $c1$ bounding-box set. The same procedure is applied for $c2$ bounding-boxes associated with q_j, where $t < j \leq m$.

The termination condition for multiple rounds of bounding-box propagation occurs when the number of candidate bounding-boxes in P becomes 0. At this point, we report all the bounding boxes in the confirmed bounding-box sets for both $c1$ and $c2$ class categories to form the final detection results.

4   Experiments

4.1   Datasets

To validate the effectiveness of our proposed method on histology images, we conduct experiments on two datasets: GlaS [18] and MoNuSeg [11]. The GlaS dataset consists of 165 images from 16 colon adenocarcinoma tissue slides. Among them, 85 images are used as the training set, 60 images are used as test set A, and 20 images are used as test set B. The MoNuSeg dataset contains different cell nuclei from multiple organs. It includes 30 images (around 22,000 cells) for training and 14 images (around 7000 cells) for testing. The codes for all the experiments will be made publicly available.

4.2   Main Results

We use the bounding-box predictions from a well-trained DiffusionDet [1] with a confidence threshold set at 0.50 as the baseline. The inputs for the proposed TSBP method are obtained from the same DiffusionDet model with all of its output bounding-boxes (without applying the threshold). We compare TSBP with

Table 1. F-scores (%) of the object detection on the GlaS and MoNuSeg datasets.

Dataset		Baseline	LC [17]	BC [10]	HB [22]	TSBP
Labeled Samples Used		0%	20%	20%	20%	0%
GlaS	testA	85.94	89.37	87.76	89.46	**90.39**
	testB	69.18	72.63	66.67	68.35	**73.45**
MoNuSeg		83.23	81.66	82.85	83.35	**83.60**

GT   LC   BC   HB   TSBP (ours)
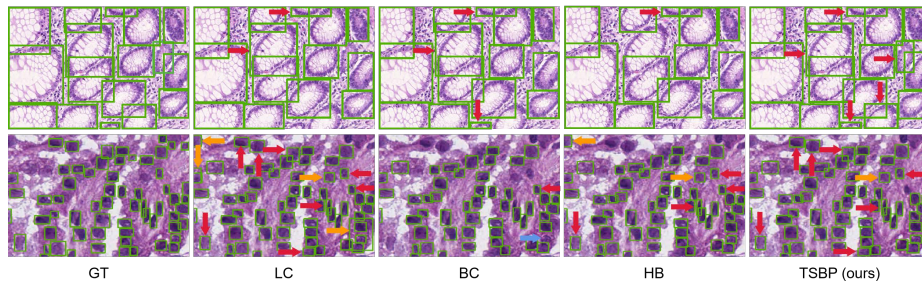
Fig. 3. Top: Gland detection. Bottom: Cell nucleus detection. The red arrows indicate additional true positives compared to the baseline results. The orange arrows indicate additional false positives compared to the baseline results. The blue arrow indicates missed detections compared to the baseline results.

a range of confidence calibration methods, Logic Calibration (LC) [17,12], Beta Calibration (BC) [10,12], and Histogram Binning (HB) [22,12][1]. Since calibration methods require additional labeled data for parameter estimations, we allocate 20% of the samples in the test sets for parameter estimations for calibration methods. We then perform model evaluation (including our method) using the remaining 80% of samples in the test sets. For confidence calibration methods, after calibration, we use a threshold set at 0.50 to decide which bounding-boxes to include in their final detection outputs.

Table 1 summarizes the quantitative results of these methods, it can be observed that TSBP achieves the highest F-scores on both the tested datasets. Note that the calibration methods require **additional labeled data** for parameter estimation, while TSBP does not use such additional labeled samples. Fig. 3 shows some visual results of TSBP and the competing methods. Compared to the baseline method, TSBP further increases the number of true positives. In the first row of the visual results, our method has the highest number of additional true positives, and TSBP finds all the remaining missed detection boxes compared to the other results. In the second row of the prediction images, the LC method has the highest number of additional true positives (LC adds eight true positives, while TSBP adds seven true positives), but, the LC method introduces more false positives (LC adds four false positives, while TSBP adds one false positives). In comparison, our method achieves a better balance between missed detections and false positives.

---

[1] https://github.com/EFS-OpenSource/calibration-framework

Table 2. F-scores (%) with varying s^{c1} used in TSBP. For comparison, other methods use s^{c1} as the threshold for selecting bounding-boxes.

		threshold= 0.70			$s^{c1}= 0.70$	threshold= 0.60			$s^{c1} = 0.60$
		Baseline	LC [17]	HB [22]	TSBP	Baseline	LC [17]	HB [22]	TSBP
Labeled Samples Used		0	20%	20%	0	0	20%	20%	0
GlaS	testA	69.45	82.34	86.97	**88.42**	79.30	86.78	86.97	**89.29**
	testB	51.13	68.35	65.36	**72.73**	59.31	70.66	68.35	**72.73**
MoNuSeg		72.62	79.97	80.59	**80.83**	80.59	83.20	83.35	**83.76**

Table 3. Resulting F-score (%) with varying K in the K-means step of TSBP.

GlaS	testA	90.21 ($K = 15$)	90.11 ($K = 20$)	90.39 ($K = 25$)	89.75 ($K = 30$)
	testB	73.45 ($K = 5$)	73.14 ($K = 10$)	72.94 ($K = 20$)	70.93 ($K = 25$)
MoNuSeg		83.52 ($K = 15$)	83.56 ($K = 20$)	83.60 ($K = 25$)	83.71 ($K = 30$)

Table 4. Error rates (%) of detection in different stages of TSBP.

GlaS		MoNuSeg	
26.79 (first satge)	37.25 (second stage)	50.66 (first stage)	53.44 (second stage)

4.3   Ablation and Additional Studies

On Initial Bounding-Box Selection. The parameter s^{c1} and s^{c2} determine the inclusion criteria for bounding-boxes in the initial confirmed bounding-box set. Here, we examine the impact of selecting the initial confidence threshold s^{c1} on the detection performance. As shown in Table 2, TSBP yields better detection results with a lower initial threshold (0.60). When a higher threshold (0.70) is applied, the initial confirmed bounding-boxes may fail to encompass the diversity of object appearances in the test samples, resulting in sub-optimal performance of bounding-box propagation. Notably, compared to other methods across varying confidence thresholds, TSBP consistently achieves better results. K **in the K-means step of TSBP.** In Table 3, we show that different values of K lead to performance fluctuations in a small range on the GlaS (testA) and MoNuSeg datasets. Due to the smaller size of the testB in GlaS dataset, smaller value of K gives better results.
Error rates in different stages of TSBP. Table 4 demonstrates that the error rate in the first stage is lower than that in the second stage, consistent with the method design, where early propagated samples are more likely to be assigned correct labels.

5   Conclusion

The abundance of visual information in histology images can be leveraged during test time to improve the performance of key object detection. This paper introduced a novel perspective and approach to enhance object detection in histology images in test time. Specifically, we designed a versatile matching-based algorithm (TSBP) that leverages detections with high model confidence to influence detections with low model confidence, aiming to rectify potential errors and

enhance the overall accuracy of the detection results. Experiments on two object detection tasks in histology images showed the effectiveness of our proposed method. Compared to confidence calibration methods, our method requires no additional labeled data for model training and parameter estimation. Future work will focus on developing a more automatic and robust procedure for forming the initial condition of TSBP. Further studies of utilizing TSBP for test-time model training can be considered an interesting future research target.

Disclosure of Interests. The authors have no competing interests to decalre.

References

1. Chen, S., Sun, P., Song, Y., Luo, P.: Diffusiondet: Diffusion model for object detection. In: ICCV. pp. 19773–19786 (2023)
2. Corbière, C., Thome, N., Bar-Hen, A., Cord, M., Pérez, P.: Addressing failure prediction by learning model confidence. Advances in Neural Information Processing Systems **32** (2019)
3. Davri, A., Birbas, E., Kanavos, T., Ntritsos, G., Giannakeas, N., Tzallas, A.T., Batistatou, A.: Deep learning on histopathological images for colorectal cancer diagnosis: A systematic review. Diagnostics **12**(4), 837 (2022)
4. Echle, A., Rindtorff, N.T., Brinker, T.J., Luedde, T., Pearson, A.T., Kather, J.N.: Deep learning in cancer pathology: A new generation of clinical biomarkers. British Journal of Cancer **124**(4), 686–696 (2021)
5. Erhan, D., Szegedy, C., Toshev, A., Anguelov, D.: Scalable object detection using deep neural networks. In: CVPR. pp. 2155–2162. IEEE Computer Society (2014)
6. Gilg, J., Teepe, T., Herzog, F., Rigoll, G.: The box size confidence bias harms your object detector. In: WACV. pp. 1471–1480 (2023)
7. Guo, C., Pleiss, G., Sun, Y., Weinberger, K.Q.: On calibration of modern neural networks. In: International Conference on Machine Learning. pp. 1321–1330. PMLR (2017)
8. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: CVPR. pp. 770–778 (2016)
9. Hendrycks, D., Gimpel, K.: A baseline for detecting misclassified and out-of-distribution examples in neural networks. In: ICLR (2017)
10. Kull, M., Silva Filho, T., Flach, P.: Beta calibration: A well-founded and easily implemented improvement on logistic calibration for binary classifiers. In: Artificial Intelligence and Statistics. pp. 623–631. PMLR (2017)
11. Kumar, N., Verma, R., Sharma, S., Bhargava, S., Vahadane, A., Sethi, A.: A dataset and a technique for generalized nuclear segmentation for computational pathology. IEEE Transactions on Medical Imaging **36**(7), 1550–1560 (2017)
12. Kuppers, F., Kronenberger, J., Shantia, A., Haselhoff, A.: Multivariate confidence calibration for object detection. In: CVPR Workshops. pp. 326–327 (2020)

13. Van der Laak, J., Litjens, G., Ciompi, F.: Deep learning in histopathology: The path to the clinic. Nature Medicine **27**(5), 775–784 (2021)
14. Munir, M.A., Khan, M.H., Sarfraz, M.S., Ali, M.: Towards improving calibration in object detection under domain shift. In: NeurIPS (2022)
15. Naeini, M.P., Cooper, G., Hauskrecht, M.: Obtaining well calibrated probabilities using Bayesian binning. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 29 (2015)
16. Nguyen, A., Yosinski, J., Clune, J.: Deep neural networks are easily fooled: High confidence predictions for unrecognizable images. In: CVPR. pp. 427–436 (2015)
17. Platt, J., et al.: Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods. Advances in large margin classifiers **10**(3), 61–74 (1999)
18. Sirinukunwattana, K., Pluim, J.P., Chen, H., Qi, X., Heng, P.A., Guo, Y.B., Wang, L.Y., Matuszewski, B.J., Bruni, E., Sanchez, U., et al.: Gland segmentation in colon histology images: The GlaS challenge contest. Medical Image Analysis **35**, 489–502 (2017)
19. Wenkel, S., Alhazmi, K., Liiv, T., Alrshoud, S., Simon, M.: Confidence score: The forgotten dimension of object detection performance evaluation. Sensors **21**(13), 4350 (2021)
20. Yang, X., Wu, J., He, L., Ma, S., Hou, Z., Sun, W.: CPSS-FAT: A consistent positive sample selection for object detection with full adaptive threshold. Pattern Recognition **141**, 109627 (2023)
21. Yue, X., Li, H., Shimizu, M., Kawamura, S., Meng, L.: Deep learning-based real-time object detection for empty-dish recycling robot. In: 2022 13th Asian Control Conference (ASCC). pp. 2177–2182. IEEE (2022)
22. Zadrozny, B., Elkan, C.: Obtaining calibrated probability estimates from decision trees and naive bayesian classifiers. In: ICML. vol. 1, pp. 609–616 (2001)