Towards Accurate Tumor Budding Detection: A Benchmark Dataset and A Detection Approach Based on Implicit Annotation Standardization and Positive-Negative Feature Coupling

Rui-Qing Sun[1*], Zeng Fan[1*], Boyang Dai[1], Yiyan Su[2], Qun Hao[1,3†], Chuyang Ye[1†], and Shaohui Zhang[1†]

[1] Beijing Institute of Technology, Beijing 100081, China
[2] Shanxi Medical University, Taiyuan 030001, China
[3] Changchun University of Science and Technology, Changchun 130022, China
{qhao,chuyang.ye,zhangshaohui}@bit.edu.cn

Abstract. The detection of tumor budding on histopathological images provides vital information for treatment planning and prognosis prediction. As manual identification of tumor budding is labor-intensive, automated tumor budding detection is desired. However, unlike other tumor cell detection tasks, tumor budding involves clusters of multiple tumor cells, which is more likely to be confused with other clusters of cells with similar appearances. It becomes challenging for existing cell detection methods to discriminate tumor budding from other cells. Additionally, the lack of public datasets for tumor budding detection hinders further development of accurate tumor budding detection methods. To address these challenges, to the best of our knowledge, we introduce the first publicly available benchmark dataset for tumor budding detection. The dataset consists of 410 images with H&E staining and the corresponding bounding box annotations of 3,968 cases of tumor budding made by experts. Moreover, based on this dataset, we propose a designated approach *Tumor Budding Detection Network* (TBDNet) for tumor budding detection with improved detection performance. On top of standard objection detection backbones, we develop two major components in TBDNet, *Iteratively Distilled Annotation Relocation* (IDAR) and *Rotational Feature Decoupling And Recoupling* (RFDAR). First, as different experts have different standards for budding boundaries in the annotation, the detection model may receive inconsistent knowledge during model training. Therefore, we introduce the IDAR module that implicitly standardizes the annotations. IDAR relocates the annotations via iterative model distillation so that the relocated annotations are consistent for training the detection model. Second, to reduce the interference from cells with similar features, i.e., negative samples, to tumor budding, i.e., positive samples, we develop the RFDAR module. RFDAR enhances feature extraction via positive-negative feature coupling regularized by prior feature

* These authors contributed equally.
† Corresponding authors

distributions, so that it is better capable of distinguishing tumor budding. The results on the benchmark show that our approach outperforms state-of-the-art detection methods by a noticeable margin. All code and data are available at https://github.com/J-F-AN/TumorBuddingDetection .

Keywords: tumor budding detection · computational pathology · benchmark dataset.

1 Introduction

Tumor budding refers to the phenomenon where tumor cells detach from the main tumor mass and infiltrate surrounding tissues in small clusters [6]. It has been identified as a key indicator of tumor invasiveness and metastatic potential. The accurate quantification of tumor budding can inform effective treatment, which prolongs patient survival and improves treatment outcome [17]. Traditional quantitative analysis of tumor budding relies on manual detection by pathologists in pathological images, a labor-intensive process. Therefore, there is a pressing need for automated tumor budding detection. Although tumor budding detection has not been extensively studied before, it can be formulated as an object detection problem. With the development of object detection technologies based on *deep learning* (DL) [12, 13, 20], DL detection models, including generic ones [10, 11, 18, 19, 23] or those designed for tumor cell detection [1, 2, 16], can be applied to tumor budding detection. For example, Bokhorst et al. [2] use a general DL detection model to automatically detect tumor budding in colorectal cancer. Their subsequent work [1] trains the network using a semi-supervised approach, which further improves detection accuracy under data-limited conditions. Piansaddhayanaon et al. [16] propose a module for tumor cell detection aimed at enhancing the performance of existing two-stage detection frameworks, and it can be used for budding detection as well. However, tumor budding detection is generally more challenging than other tumor cell detection tasks, as budding typically presents as clusters of tumor cells, which are prone to being mistaken for other cells or cell clusters with similar morphological characteristics. Therefore, direct application of existing detection methods to tumor budding only achieves suboptimal performance. Moreover, there is currently no publicly available dataset for tumor budding detection, which further hinders the development of accurate budding detection methods.

To address these challenges, we introduce the first publicly available benchmark for tumor budding detection, and it is named *Tumor Budding Detection Dataset* (TBDD). TBDD contains 410 images with H&E staining and 3,968 annotated cases of tumor budding in the form of bounding boxes. Based on this benchmark dataset, we also propose *Tumor Budding Detection Network* (TBDNet), which is a designated approach to tumor budding detection with improved detection performance. There are two major contributions, *Iteratively Distilled Annotation Relocation* (IDAR) and *Rotational Feature Decoupling And Recoupling* (RFDAR) in TBDNet. First, due to the irregular shape of tumor budding, experts tend to have different standards when determining budding boundaries

for annotation. Such a discrepancy can lead to inconsistent knowledge learning during model training, which adversely affects the detection performance. To tackle this issue, we introduce IDAR that implicitly standardizes the bounding box annotations. IDAR uses a teacher-student framework to relocate the boxes, and with iterative model distillation the final student model produces consistently relocated annotations for subsequent model training. Second, to better distinguish between tumor budding, i.e., positive samples, and non-budding targets, i.e., negative samples, with similar features, RFDAR enhances the feature extraction by decoupling and recoupling the features of positive and negative samples, under the regularization of prior feature distributions. Such positive-negative feature coupling encourages the model to focus on the discriminative features for distinguishing tumor budding and non-budding targets. Qualitative and quantitative evaluation results on the benchmark dataset TBDD show that TBDNet outperforms existing detection methods.

2 Method

The contribution of this work is summarized in Fig. 1, which gives an overview of the data curation for TBDD and the proposed TBDNet. Their detailed descriptions are given below.

2.1 Dataset Curation for Tumor Budding Detection

TBDD was collected for colorectal cancer patients, as colorectal cancer is a leading cause of cancer-related morbidity and mortality [22]. Histopathology images were acquired for 410 patients over two years on a KFBIO KF-PRO-400-HI scanner. The samples were processed, sectioned, and stained with H&E; then they were examined with 40x magnification to capture clear, high-resolution whole slide images. For each patient, only an 800×800 patch with tumor budding that is clinically informative was cropped from the whole slide image and annotated. Three experienced pathologists annotated tumor budding areas on the 800×800 images with bounding boxes, each handling a disjoint subset of the images, and all annotations were cross-reviewed according to the criteria of the International Tumor Budding Consensus Conference (ITBCC) to ensure consistency. The released dataset contains both the cropped images and their annotations.

2.2 Iteratively Distilled Annotation Relocation

As tumor budding typically has variable shapes depending on the composition of tumor cells, different experts can determine the sizes and locations of annotation bounding boxes with different standards [9]. This annotation inconsistency hinders effective training of detection models [21]. To address this issue, we propose the IDAR method, which gradually relocates the annotation through iterative knowledge distillation [7] and produces consistent standardized bounding boxes better suited for model learning.

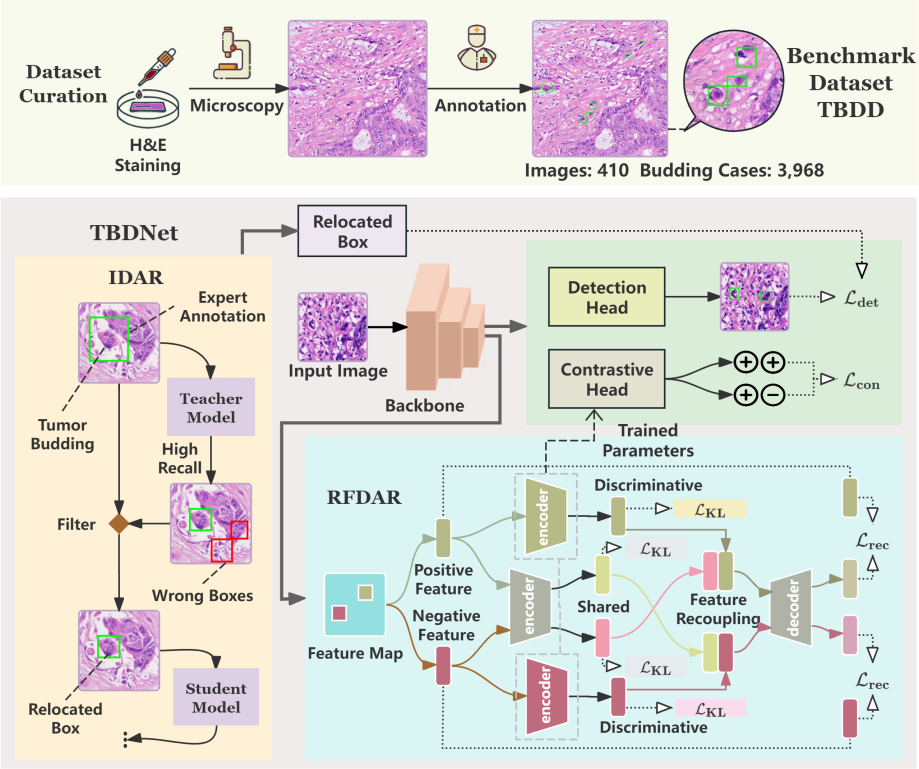

Fig. 1. An overview of the proposed work. First, we introduce a benchmark dataset TBDD for tumor budding detection. Then, we develop TBDNet for tumor budding detection based on TBDD. TBDNet comprises two major components, IDAR and RFDAR. IDAR implicitly standardizes the annotation to avoid inconsistent knowledge learning, whereas RFDAR uses positive-negative feature coupling to enhance the extraction of discriminative features for distinguishing tumor budding from other targets.

In IDAR, we first train a teacher model with the manually annotated data. The bounding boxes generated by the teacher model are then used as soft labels to train the student model. In the next iteration, the student model serves as the new teacher model to train a new student model. With the iterative distillation, the size and location of the bounding boxes stabilize, which implicitly achieves standardized annotation that both conforms to the original annotation and becomes consistent across samples. Notably, during each iteration, when generating soft-label bounding boxes, to ensure all annotated budding cases are preserved, the teacher model predicts with a low confidence threshold of 0.05. As false positives may be produced with the low threshold, the teacher outputs are then filtered based on the manual annotations with an *Intersection over Union* (IoU) threshold of 0.2.

2.3 Rotational Feature Decoupling and Recoupling

Then, we train the detection model with the standardized annotation. The tumor budding cell clusters can be visually similar to other cell clusters, where their cytoplasm has similar appearances and their nucleus appearances are different. Such subtle differences can be difficult to learn for existing detection frameworks. Therefore, we design RFDAR to address the issue.

Our method is based on a standard object detection backbone. Here, we choose YOLOv5 [10] due to its ease of deployment and high performance, but other backbones are also applicable. RFDAR uses a contrastive learning strategy by adding a contrastive head to YOLOv5 [10] (see Fig. 1), which encourages discrimination between the positive samples of tumor budding and the negative samples of similar non-tumor-budding cell clusters. However, the features extracted by the backbone network contain both discriminative ones and those that are shared between tumor budding and other cell clusters. This is possibly because the nucleus components are different and the cytoplasm components are similar between positive and negative samples, and naive training of the detection model cannot optimally preserve the discriminative features. Therefore, in conventional contrastive learning, the separation of discriminative features of positive and negative samples is hindered by the shared features.

To address this problem, RFDAR seeks to decompose the features into discriminative ones and shared ones, and the decomposition is learned by coupling the features of positive and negative samples. Then, only the discriminative features are used in contrastive learning to better distinguish positive and negative samples. Specifically, during model training the backbone features are first categorized as features of positive samples and negative samples based on the annotation. As shown in Fig. 1, both positive and negative features are fed into an encoder of discriminative features and an encoder of shared features for feature decoupling. The shared features are then rotated between the positive and negative samples, i.e., the positive shared features are recoupled with the negative discriminative features and the negative shared features are recoupled with the positive discriminative features. Since the shared features are similar between positive and negative samples, the original features should be reconstructed from the recoupled features.

$$\mathcal{L}_{\text{rec}} = \frac{1}{2}\left(\|\mathbf{x}_{\text{orig}} - \mathbf{x}^{(1)}_{\text{recouple}}\|_2^2 + \|\mathbf{x}_{\text{orig}} - \mathbf{x}^{(2)}_{\text{recouple}}\|_2^2\right), \tag{1}$$

Therefore, by minimizing a reconstruction loss $\mathcal{L}_{\text{rec}}$ measured by the mean squared error of the positive and negative features, RFDAR learns how to decompose the raw features into discriminative and shared ones.

To avoid degeneration and improve the stability of the decoupling, we introduce additional regularization for training RFDAR based on the prior feature distributions. Specifically, from the raw features, we fit the distributions of positive and negative features with a *Gaussian mixture model* (GMM) [4]. As the raw features are high-dimensional, before model fitting we apply *principal component analysis* (PCA) [8] to the features for dimensionality reduction. Then,

the first principal component that best reflects the major variation in the samples is used for fitting the GMM. We assume that the GMM comprises three components, the discriminative features of positive samples, the discriminative features of negative samples, and the shared features of the positive and negative samples. With the three distributions, the *Kullback–Leibler* (KL) divergence loss $\mathcal{L}_{\mathrm{KL}}$ is computed as regularization for the discriminative feature of positive samples, the discriminative feature of negative samples, and the shared feature with respect to their corresponding prior distributions (see Fig. 1).

The encoders of discriminative and shared features are trained by minimizing the sum of $\mathcal{L}_{\mathrm{rec}}$ and $\mathcal{L}_{\mathrm{KL}}$ with the other network parameters fixed.

Finally, with the trained RFDAR module (frozen), we perform the training of the detection model aided by contrastive learning. Specifically, the standard detection loss $\mathcal{L}_{\mathrm{det}}$ of YOLOv5 [10] is combined with the following contrastive loss motivated by [3] and [25]:

$$\mathcal{L}_{\mathrm{con}} = \frac{1}{N}\sum_{i=1}^{N}\left[(1-y_i)\cdot(d_i)^2 + y_i\cdot(\max(0, m-d_i))^2\right], \tag{2}$$

where y_i denotes the label disagreement ($y_i = 0$ for two samples of the budding class and $y_i = 1$ for samples of different classes) of the i-th sample pair, d_i represents the Euclidean distance between the sample pair, N is the total number of sample pairs, and m is a predefined hyperparameter representing the maximum valid distance between inter-class pairs. Note that the features are normalized by their L2 norm before computing d_i, and we set $m = 1.2$ based on experiments on a validation set [5, 24]. Then, the total loss is

$$\mathcal{L}_{\mathrm{total}} = \mathcal{L}_{\mathrm{det}} + \lambda\cdot\mathcal{L}_{\mathrm{con}}, \tag{3}$$

where λ is a balancing hyperparameter set to 1.5 based on the validation set.

2.4 Implementation Details

Both teacher and student models in IDAR use the YOLOv5 detector as the backbone, with the iteration count empirically fixed to two based on experimental tuning. Each encoder in RFDAR comprises two fully connected layers. It maps the input to a 256-dimensional space with ReLU activation [15], followed by two parallel linear layers that generate the mean and log variance with a dimension of 128. The decoder in RFDAR also comprises two fully connected layers, where the first layer has a channel width of 256 with ReLU activation and the second one projects the output of the first layer back to the original feature space. The contrastive head uses the encoder trained in RFDAR for computing the contrastive loss. All these auxiliary modules are employed solely during training, and at inference time the model collapses to the vanilla YOLOv5 detector, incurring no additional parameters or latency.

Model training is performed with the default settings of YOLOv5 [10] when applicable. The images are cropped into 384×384 patches with an overlap of 128 pixels in each dimension as model input. The predictions on the input patches are merged for test images.

Table 1. Detection performance of different methods on TBDD. The best results are highlighted in bold.

Category	Method	F1 (%) ↑	P (%) ↑	R (%) ↑
CNN-based (two-stage)	Faster R-CNN (2016) [18]	59.11	60.61	57.69
	RetinaNet (2017) [19]	58.06	55.75	60.58
	ReCasNet (2023) [16]	53.26	56.98	49.99
CNN-based (one-stage)	YOLO11 (2024) [11]	57.07	58.97	55.29
	MambaYOLO (2024) [23]	56.46	58.80	54.30
Transformer-based	SwinT (2021) [14]	40.48	45.11	36.71
	Deformable DETR (2021) [26]	45.37	41.67	49.79
Ours	**TBDNet**	**63.29**	**63.59**	**62.98**

3 Experiments

3.1 Experimental Setup and Detection Performance

The 410 images in TBDD were split into a training set of 276 images, a validation set of 94 images, and a test set of 40 images. We compared TBDNet with several representative object detection methods, including CNN-based two-stage detectors Faster R-CNN [18], RetinaNet [19], and ReCasNet [16], CNN-based one-stage detectors YOLO11 [11] and MambaYOLO [23], and Transformer-based approaches Deformable DETR [26] and SwinT [14]. All these methods applied their default settings and used the same experimental settings as TBDNet for fair comparison.

The detection performance was quantitatively evaluated with the F1-score (F1), precision (P), and recall (R). The results are presented in Table 1. TBDNet achieves better performance than all competing methods. The CNN-based two-stage detectors tend to perform better among the competing methods. Compared with the second best method Faster R-CNN, TBDNet improves the F1-score by 4.18%. Note that our method is based on YOLOv5, and it outperforms (by over 6%) more recent versions of YOLO detectors, YOLO11 and MambaYOLO, whereas the baseline YOLOv5 has similar performance to the two competitors (see the ablation study later in Table 2).

Visualization examples of the detection results are shown in Fig. 2. Compared with the competing methods, TBDNet produces more true positive and fewer false positive budding detection results. This observation agrees with the better precision and recall values of TBDNet in Table 1.

3.2 Ablation Study

To further validate the effectiveness of each module in our method, we performed an ablation study. The results are shown in Table 2. First, compared to the YOLOv5 baseline model, one iteration of IDAR improves the F1-score, and it is further increased after the second iteration. This confirms the benefit of IDAR.

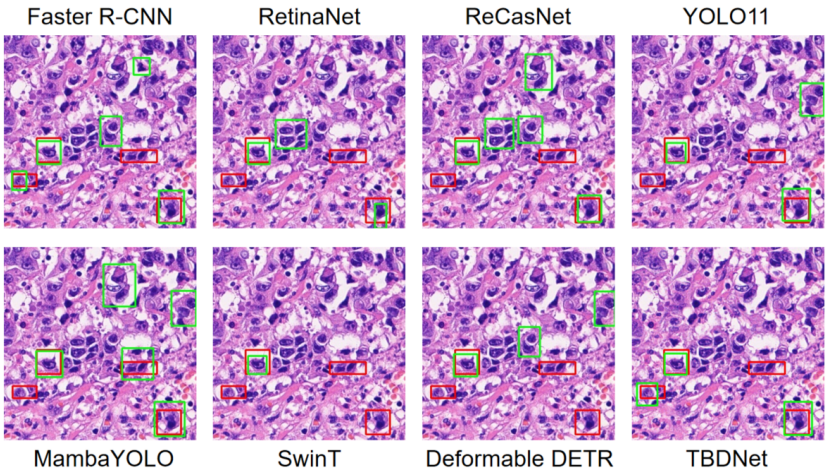

Fig. 2. Visualization examples of detection results. The green box is the output of the model, and the red box is the annotation.

Table 2. Ablation study with different configurations. The baseline uses YOLOv5. Iter represents the number of iterations in IDAR. C-Head represents the use of the contrastive head. The best results are highlighted in bold.

Baseline	Iter	C-Head	$\mathcal{L}_{\text{rec}}$	$\mathcal{L}_{\text{KL}}$	F1 (%) ↑	P (%) ↑	R (%) ↑
✓	-	-	-	-	57.42	59.18	55.77
✓	1	-	-	-	58.63	57.67	59.62
✓	2	-	-	-	60.77	60.48	61.06
✓	2	✓	-	-	61.23	62.94	59.62
✓	2	✓	✓	-	59.36	56.52	62.50
✓	2	✓	-	✓	57.00	57.28	56.73
✓	2	✓	✓	✓	**63.29**	**63.59**	**62.98**

Second, the contrastive head itself improves the F1-score after the application of IDAR. When the full RFDAR is also applied (with both $\mathcal{L}_{\text{rec}}$ and $\mathcal{L}_{\text{KL}}$), the best performance is achieved, which shows the benefit of RFDAR. Note that if only $\mathcal{L}_{\text{rec}}$ or $\mathcal{L}_{\text{KL}}$ is used in RFDAR, the results are worse than the use of the contrastive head, because $\mathcal{L}_{\text{rec}}$ mainly increases recall by encouraging the detection of positive cases, while $\mathcal{L}_{\text{KL}}$ tightens the posterior distribution to improve precision. Using either loss alone therefore fails to balance sensitivity and specificity.

4 Conclusion

We have curated and released a tumor budding detection dataset TBDD and proposed TBDNet for better performance of tumor budding detection. TBDD comprises 410 H&E images with 3,968 cases of annotated tumor budding. In TBDNet, we have made two major contributions IDAR and RFDAR. IDAR implicitly standardizes the annotation by relocating the annotation bounding boxes via iterative model distillation. This avoids knowledge inconsistency during model training due to the different standards of different experts for budding boundaries. RFDAR better distinguishes tumor budding from other similar cells or cell clusters by extracting more discriminative features via positive-negative feature coupling. Experimental results show that TBDNet outperforms existing methods for the challenging tumor budding detection task by a noticeable margin.

Acknowledgments. This work was supported by the National Natural Science Foundation of China (62275020).

Disclosure of Interests. The authors have no competing interests to declare that are relevant to the content of this article.

References

1. Bokhorst, J.M., Nagtegaal, I.D., Zlobec, I., Dawson, H., Sheahan, K., Simmer, F., Kirsch, R., Vieth, M., Lugli, A., van der Laak, J., et al.: Semi-supervised learning to automate tumor bud detection in cytokeratin-stained whole-slide images of colorectal cancer. Cancers **15**(7), 2079 (2023)
2. Bokhorst, J.M., Rijstenberg, L., Goudkade, D., Nagtegaal, I., van der Laak, J., Ciompi, F.: Automatic detection of tumor budding in colorectal carcinoma with deep learning. In: Computational Pathology and Ophthalmic Medical Image Analysis: First International Workshop, COMPAY 2018, and 5th International Workshop, OMIA 2018, Held in Conjunction with MICCAI 2018, Granada, Spain, September 16-20, 2018, Proceedings 5. pp. 130–138. Springer (2018)
3. Chopra, S., Hadsell, R., LeCun, Y.: Learning a similarity metric discriminatively, with application to face verification. In: 2005 IEEE computer society conference on computer vision and pattern recognition (CVPR'05). vol. 1, pp. 539–546. IEEE (2005)
4. Dempster, A.P., Laird, N.M., Rubin, D.B.: Maximum likelihood from incomplete data via the EM algorithm. Journal of the royal statistical society: series B (methodological) **39**(1), 1–22 (1977)
5. Hadsell, R., Chopra, S., LeCun, Y.: Dimensionality reduction by learning an invariant mapping. In: 2006 IEEE computer society conference on computer vision and pattern recognition (CVPR'06). vol. 2, pp. 1735–1742. IEEE (2006)
6. Hase, K., Shatney, C., Johnson, D., Trollope, M., Vierra, M.: Prognostic value of tumor "budding" in patients with colorectal cancer. Diseases of the colon & rectum **36**(7), 627–635 (1993)

7. Hinton, G.: Distilling the knowledge in a neural network. arXiv preprint arXiv:1503.02531 (2015)
8. Hotelling, H.: Analysis of a complex of statistical variables into principal components. Journal of educational psychology **24**(6), 417 (1933)
9. Jiao, L., Zhang, F., Liu, F., Yang, S., Li, L., Feng, Z., Qu, R.: A survey of deep learning-based object detection. IEEE access **7**, 128837–128868 (2019)
10. Jocher, G.: Ultralytics YOLOv5 (2020). `https://doi.org/10.5281/zenodo.3908559`, `https://github.com/ultralytics/yolov5`
11. Jocher, G., Qiu, J.: Ultralytics YOLO11 (2024), `https://github.com/ultralytics/ultralytics`
12. Krizhevsky, A., Sutskever, I., Hinton, G.E.: Imagenet classification with deep convolutional neural networks. Advances in neural information processing systems **25** (2012)
13. LeCun, Y., Bengio, Y., Hinton, G.: Deep learning. nature **521**(7553), 436–444 (2015)
14. Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y., Zhang, Z., Lin, S., Guo, B.: Swin Transformer: Hierarchical vision transformer using shifted windows. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 10012–10022 (2021)
15. Nair, V., Hinton, G.E.: Rectified linear units improve restricted boltzmann machines. In: Proceedings of the 27th international conference on machine learning (ICML-10). pp. 807–814 (2010)
16. Piansaddhayanaon, C., Santisukwongchote, S., Shuangshoti, S., Tao, Q., Sriswasdi, S., Chuangsuwanich, E.: ReCasNet: Improving consistency within the two-stage mitosis detection framework. Artificial Intelligence in Medicine **135**, 102462 (2023)
17. Prall, F., Nizze, H., Barten, M.: Tumour budding as prognostic factor in stage I/II colorectal carcinoma. Histopathology **47**(1), 17–24 (2005)
18. Ren, S., He, K., Girshick, R., Sun, J.: Faster R-CNN: Towards real-time object detection with region proposal networks. IEEE transactions on pattern analysis and machine intelligence **39**(6), 1137–1149 (2016)
19. Ross, T.Y., Dollár, G.: Focal loss for dense object detection. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 2980–2988 (2017)
20. Rumelhart, D.E., Hinton, G.E., Williams, R.J.: Learning representations by back-propagating errors. nature **323**(6088), 533–536 (1986)
21. Song, H., Kim, M., Park, D., Shin, Y., Lee, J.G.: Learning from noisy labels with deep neural networks: A survey. IEEE transactions on neural networks and learning systems **34**(11), 8135–8153 (2022)
22. Sung, H., Ferlay, J., Siegel, R.L., Laversanne, M., Soerjomataram, I., Jemal, A., Bray, F.: Global cancer statistics 2020: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. CA: a cancer journal for clinicians **71**(3), 209–249 (2021)
23. Wang, Z., Li, C., Xu, H., Zhu, X.: Mamba YOLO: SSMs-Based YOLO for object detection. arXiv preprint arXiv:2406.05835 (2024)
24. Wu, Z., Xiong, Y., Yu, S.X., Lin, D.: Unsupervised feature learning via non-parametric instance discrimination. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 3733–3742 (2018)
25. Zhou, H., Liu, Y., Liu, Z., Luo, P., Wang, X.: Talking face generation by adversarially disentangled audio-visual representation. In: Proceedings of the AAAI conference on artificial intelligence. vol. 33, pp. 9299–9306 (2019)

26. Zhu, X., Su, W., Lu, L., Li, B., Wang, X., Dai, J.: Deformable DETR: Deformable transformers for end-to-end object detection. In: International Conference on Learning Representations (2020)