Gall Bladder Cancer Detection from US Images with only Image Level Labels

Soumen Basu[1(✉)], Ashish Papanai[1], Mayank Gupta[1], Pankaj Gupta[1,2], and Chetan Arora[1]

[1] Indian Institute of Technology, Delhi, India
soumen.basu@cse.iitd.ac.in
[2] Postgraduate Institute of Medical Education and Research, Chandigarh, India

Abstract. Automated detection of Gallbladder Cancer (GBC) from Ultrasound (US) images is an important problem, which has drawn increased interest from researchers. However, most of these works use difficult-to-acquire information such as bounding box annotations or additional US videos. In this paper, we focus on GBC detection using only image-level labels. Such annotation is usually available based on the diagnostic report of a patient, and do not require additional annotation effort from the physicians. However, our analysis reveals that it is difficult to train a standard image classification model for GBC detection. This is due to the low inter-class variance (a malignant region usually occupies only a small portion of a US image), high intra-class variance (due to the US sensor capturing a 2D slice of a 3D object leading to large viewpoint variations), and low training data availability. We posit that even when we have only the image level label, still formulating the problem as object detection (with bounding box output) helps a deep neural network (DNN) model focus on the relevant region of interest. Since no bounding box annotations is available for training, we pose the problem as weakly supervised object detection (WSOD). Motivated by the recent success of transformer models in object detection, we train one such model, DETR, using multi-instance-learning (MIL) with self-supervised instance selection to suit the WSOD task. Our proposed method demonstrates an improvement of AP and detection sensitivity over the SOTA transformer-based and CNN-based WSOD methods. Project page is at https://gbc-iitd.github.io/wsod-gbc.

Keywords: Weakly Supervised Object Detection · Ultrasound · Gallbladder Cancer

1 Introduction

GBC is a deadly disease that is difficult to detect at an early stage [12,15]. Early diagnosis can significantly improve the survival rate [14]. Non-ionizing radiation,

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/978-3-031-43907-0_20.

H. Greenspan et al. (Eds.): MICCAI 2023, LNCS 14220, pp. 206–215, 2023.
https://doi.org/10.1007/978-3-031-43907-0_20

low cost, and accessibility make US a popular non-invasive diagnostic modality for patients with suspected gall bladder (GB) afflictions. However, identifying signs of GBC from routine US imaging is challenging for radiologists [11]. In recent years, automated GBC detection from US images has drawn increased interest [3,5] due to its potential for improving diagnosis and treatment outcomes. Many of these works formulate the problem as an object detection, since training a image classification model for GBC detection seems challenging due to the reasons outlined in the abstract (also see Fig. 1).

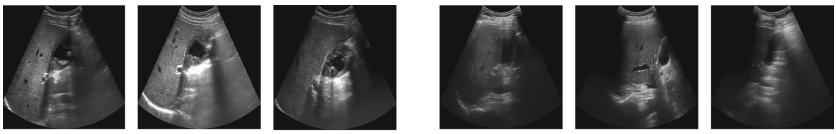

Fig. 1. (a) Low inter-class variability. The first two GBs show benign wall thickening, and the third one shows malignant thickening. However, the appearance of the GB in all three images is very similar. (b) High intra-class variability. All three images have been scanned from the same patient, but due to the sensor's scanning plane, the appearances change drastically.

Recently, GBCNet [3], a CNN-based model, achieved SOTA performance on classifying malignant GB from US images. GBCNet uses a two-stage pipeline consisting of object detection followed by classification, and requires bounding box annotations for GB as well as malignant regions for training. Such bounding box annotations surrounding the pathological regions are time-consuming and require an expert radiologist for annotation. This makes it expensive and non-viable for curating large datasets for training large DNN models. In another recent work, [5] has exploited additional unlabeled video data for learning good representations for downstream GBC classification and obtained performance similar to [3] using a ResNet50 [13] classifier. The reliance of both SOTA techniques on additional annotations or data, limits their applicability. On the other hand, the image-level malignancy label is usually available at a low cost, as it can be obtained readily from the diagnostic report of a patient without additional effort from clinicians.

Instead of training a classification pipeline, we propose to solve an object detection problem, which involves predicting a bounding box for the malignancy. The motivation is that, running a classifier on a focused attention/ proposal region in an object detection pipeline would help tackle the low inter-class and high intra-class variations. However, since we only have image-level labels available, we formulate the problem as a Weakly Supervised Object Detection (WSOD) problem. As transformers are increasingly outshining CNNs due to their ability to aggregate focused cues from a large area [6,9], we choose to use transformers in our model. However, in our initial experiments SOTA WSOD methods for transformers failed miserably. These methods primarily rely on training a classification pipeline and later generating activation heatmaps using attention and

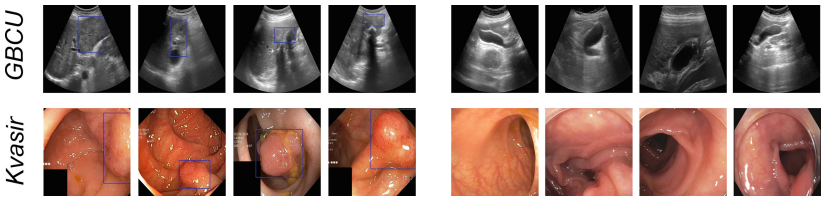

Fig. 2. Samples from the GBCU [3] and Kvasir-SEG [17] datasets. Four images from each of the disease and non-disease classes are shown on the left and right, respectively. Disease locations are shown by drawing bounding boxes.

drawing a bounding box circumscribing the heatmaps [2,10] to show localization. However, for GBC detection, this line of work is not helpful as we discussed earlier.

Inspired by the success of the Multiple Instance Learning (MIL) paradigm for weakly supervised training on medical imaging tasks [20,22], we train a detection transformer, DETR, using the MIL paradigm for weakly supervised malignant region detection. In this, one generates region proposals for images, and then considers the images as bags and region proposals as instances to solve the instance classification (object detection) under the MIL constraints [8]. At inference, we use the predicted instance labels to predict the bag labels. Our experiments validate the utility of this approach in circumventing the challenges in US images and detecting GBC accurately from US images using only image-level labels.

Contributions: The key contributions of this work are:

- We design a novel DETR variant based on MIL with self-supervised instance learning towards the weakly supervised disease detection and localization task in medical images. Although MIL and self-supervised instance learning has been used for CNNs [24], such a pipeline has not been used for transformer-based detection models.
- We formulate the GBC classification problem as a weakly supervised object detection problem to mitigate the effect of low inter-class and large intra-class variances, and solve the difficult GBC detection problem on US images without using the costly and difficult to obtain additional annotation (bounding box) or video data.
- Our method provides a strong baseline for weakly supervised GBC detection and localization in US images, which has not been tackled earlier. Further, to assess the generality of our method, we apply our method to Polyp detection from Colonoscopy images.

2 Datasets

Gallbladder Cancer Detection in Ultrasound Images: We use the public GBC US dataset [3] consisting of 1255 image samples from 218 patients. The dataset contains 990 non-malignant (171 patients) and 265 malignant (47 patients) GB images (see Fig. 2 for some sample images). The dataset contains image labels as well as bounding box annotations showing the malignant regions. Note that, we use only the image labels for training. We report results on 5-fold cross-validation. We did the cross-validation splits at the patient level, and all images of any patient appeared either in the train or validation split.

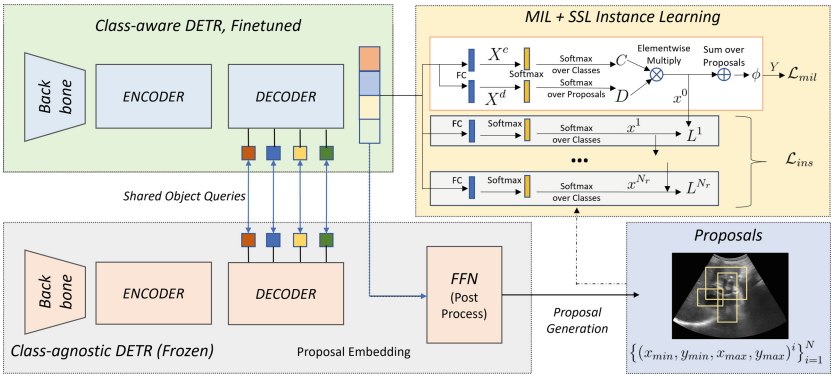

Fig. 3. Overview of the proposed Weakly Supervised DETR architecture. The location information in the object queries learned by the class-agnostic DETR ensures generation of high-quality proposals. The MIL framework uses the proposal embeddings generated at the class-aware branch.

Polyp Detection in Colonoscopy Images: We use the publicly available Kvasir-SEG [17] dataset consisting of 1000 white light colonoscopy images showing polyps (see Fig. 2). Since Kvasir-SEG does not contain any control images, we add 600 non-polyp images randomly sampled from the PolypGen [1] dataset. Since the patient information is not available with the data, we use random stratified splitting for 5-fold cross-validation.

3 Our Method

Revisiting DETR: The DETR [6] architectures utilize a ResNet [13] backbone to extract 2D convolutional features, which are then flattened and added with a positional encoding, and fed to the self-attention-based transformer encoder. The decoder uses cross-attention between learned object queries containing positional embedding, and encoder output to produce output embedding containing the class and localization information. The number of object queries, and the decoder

output embeddings is set to 100 in DETR. Subsequently, a feed-forward network generates predictions for object bounding boxes with their corresponding labels and confidence scores.

Proposed Architecture: Fig. 3 gives an overview of our method. We use a COCO pre-trained class-agnostic DETR as proposal generator. The learned object queries contain the embedded positional information of the proposal. Class-agnostic indicates that all object categories are considered as a single object class, as we are only interested in the object proposals. We then finetune a regular, class-aware DETR for the WSOD task. This class-aware DETR is initialized with the checkpoint of the class-agnostic DETR. The learned object queries from the class-agnostic DETR is frozen and shared with the WSOD DETR during finetuning to ensure that the class-aware DETR attends similar locations of the object proposals. The class-agnostic DETR branch is frozen during the finetuning phase. We finally use the MIL-based instance classification with the self-supervised instance learning over the finetuning branch. For GBC classification, if the model generates bounding boxes for the input image, then we predict the image to be malignant, since the only object present in the data is the cancer.

MIL Setup: The decoder of the fine-tuning DETR generates R d-dimensional output embeddings. Each embedding corresponds to a proposal generated by the class-agnostic DETR. We pass these embeddings as input to two branches with FC layers to obtain the matrices $X^c \in \mathbb{R}^{R \times N_c}$ and $X^r \in \mathbb{R}^{R \times N_c}$, where R is the number of object queries (same as proposals) and N_c is the number of object (disease) categories. Let $\sigma(\cdot)$ denote the softmax operation. We then generate the class-wise and detection-wise softmax matrices $C \in \mathbb{R}^{R \times N_c}$ and $D \in \mathbb{R}^{R \times N_c}$, where $C_{ij} = \sigma((X^c)_j^T)i$ and $D_{ij} = \sigma(X_i^r)j$, and X_i denotes the i-th row of X. C provides classification probabilities of each proposal, and D provides the relative score of the proposals corresponding to each class. The two matrices are element-wise multiplied and summed over the proposal dimension to generate the image-level classification predictions, $\phi \in \mathbb{R}^{N_c}$:

$$\phi_j = \sum_{i=1}^{R} C_{ij} \cdot D_{ij} \tag{1}$$

Notice, $\phi_j \in (0, 1)$ since C_{ij} and D_{ij} are normalized. Finally, the negative log-likelihood loss between the predicted labels, and image labels $y \in \mathbb{R}^{N_c}$ is computed as the MIL loss:

$$\mathcal{L}_{\text{mil}} = -\sum_{i=1}^{N_c} [y_i \log \phi_i + (1 - y_i) \log (1 - \phi_i)] \tag{2}$$

The MIL classifier further suffers from overfitting to the distinctive classification features due to the mismatch of classification and detection probabilities [24]. To tackle this, we further use a self-supervised module to improve the instances.

Self-supervised Instance Learning: Inspired by [24], we design a instance learning module with N_r blocks in a self-supervised framework to refine the instance scores with instance-level supervision. Each block consists of an `FC` layer. A class-wise softmax is used to generate instance scores $x^n \in \mathbb{R}^{R\times(N_c+1)}$ at n-th block. N_c+1 includes the background/ no-finding class. Instance supervision of each layer (n) is obtained from the scores of the previous layer $(x^{(n-1)})$. The instance supervision for the first layer is obtained from the `MIL` head. Suppose $\hat{y}^n \in \mathbb{R}^{R\times(N_c+1)}$ is the pseudo-labels of the instances. An instance (p_j) is labelled 1 if it overlaps with the highest-scoring instance by a chosen threshold.

Table 1. Weakly supervised disease detection performance comparison of our method and SOTA baselines in GBC and Polyps. We report Average Precision at IoU 0.25 (AP_{25}).

Method	**GBC** AP_{25}	**Polyp** AP_{25}
TS-CAM [10] (ICCV 2021)	0.024 ± 0.008	0.058 ± 0.015
SCM [2] (ECCV 2022)	0.013 ± 0.001	0.082 ± 0.036
OD-WSCL [21] (ECCV 2022)	0.482 ± 0.067	0.239 ± 0.032
WS-DETR [19] (WACV 2023)	0.520 ± 0.088	0.246 ± 0.023
Point-Beyond-Class [18] (MICCAI 2022)	0.531 ± 0.070	0.283 ± 0.022
Ours	0.628 ± 0.080	0.363 ± 0.052

Table 2. Ablation study. Performance of `MIL`-framework variants on `DETR`. We compare the AP and detection sensitivity.

Design	**GBC**		**Polyp**	
	AP_{25}	**Sens.**	AP_{25}	**Sens.**
MIL + DETR	0.520 ± 0.088	0.833 ± 0.034	0.246 ± 0.023	0.882 ± 0.034
MIL + SSL + DETR (Ours)	0.628 ± 0.080	0.861 ± 0.089	0.363 ± 0.052	0.932 ± 0.022

Otherwise, the instance is labeled 0 as defined in Eq. 3:

$$m_j^n = \underset{i}{\operatorname{argmax}}\, x_{ij}^{(n-1)} \;; \qquad \hat{y}_{ij}^n = \begin{cases} 1, & IoU(p_j, p_{m_j^n}) \geq \tau \\ 0, & \text{otherwise} \end{cases} \tag{3}$$

The loss over the instances is given by Eq. 4:

$$\mathcal{L}_{ins} = -\frac{1}{N_r}\sum_{n=1}^{N_r}\frac{1}{R}\sum_{i=1}^{R}\sum_{j=1}^{N_c+1} w_i^n \hat{y}_{ij}^n \log x_{ij}^n \tag{4}$$

Here x_{ij}^n denotes the score of i-th instance for j-th class at layer n. Following [24], the loss weight $w_i^n = x_{i\,m_j^n}^{(n-1)}$ is applied to stabilize the loss. Assuming λ to be a scaling value, the overall loss function is given in Eq. 5:

$$\mathcal{L} = \mathcal{L}_{mil} + \lambda \mathcal{L}_{ins} \tag{5}$$

4 Experiments and Results

Experimental Setup: We use a machine with Intel Xeon Gold 5218@2.30GHz processor and 8 Nvidia Tesla V100 GPUs for our experiments. The model is trained using SGD with LR 0.001 (for MIL head), weight decay 10^{-6}, and momentum 0.9 for 100 epochs with batch size 32. The LR at backbone and transformer are 0.003, and 0.0003, respectively. We use a cosine annealing of the LR.

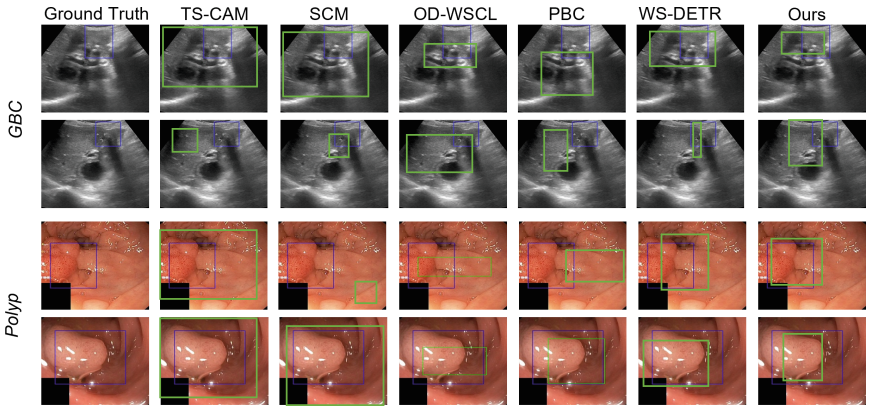

Fig. 4. Qualitative analysis of the predicted bounding boxes. Ground truths are in blue, and predictions are in green. We compare with SOTA WSOD techniques and our proposed method. Our method predicts much tighter bounding boxes that cover the clinically significant disease regions. (Color figure online)

Comparison with SOTA: Table 1 shows the bounding box localization results of the WSOD task. Our method surpasses all latest SOTA WSOD techniques by 9 points, and establishes itself as a strong WSOD baseline for GBC localization in US images. Our method also achieves 7-point higher AP score for polyp detection. We present visualizations of the predicted bounding boxes in Fig. 4 which shows that the localization by our method is more precise and clinically relevant as compared to the baselines.

Generality of the Method: We assess the generality of our method by applying it to polyp detection on colonoscopy images. The applicability of our method on two different tasks - (1) GBC detection from US and (2) Polyp detection from Colonoscopy, indicates the generality of the method across modalities.

Ablation Study: We show the detection sensitivity to the self-supervised instance learning module in Table 2 for two variants, (1) vanilla MIL head on DETR, and (2) MIL with self-supervised instance learning on DETR. Table 2 shows the Average Precision and detection sensitivity for both diseases. The results establish the benefit of using the self-supervised instance learning. Other ablations related to the hyper-parameter sensitivity is given in Supplementary Fig. S1.

Classification Performance: We compare our model with the standard CNN-based and Transformer-based classifiers, SOTA WSOD-based classifiers, and SOTA classifiers using additional data or annotations (Table 3). Our method beats the SOTA weakly supervised techniques and achieves 1.2% higher sensitivity for GBC detection. The current SOTA GBC detection models require additional bound-

Table 3. Performance comparison of our method and other SOTA methods in GBC classification. We report accuracy, specificity, and sensitivity.

Type	Method	Acc.	Spec.	Sens.
CNN Classifier	ResNet50 [13]	0.867 ± 0.031	0.926 ± 0.069	0.672 ± 0.147
	InceptionV3 [23]	0.869 ± 0.039	0.913 ± 0.032	0.708 ± 0.078
Transformer Classifier	ViT [9]	0.803 ± 0.078	0.901 ± 0.050	0.860 ± 0.068
	DEIT [25]	0.829 ± 0.030	0.900 ± 0.040	0.875 ± 0.063
	PVTv2 [26]	0.824 ± 0.033	0.887 ± 0.057	0.894 ± 0.076
	RadFormer [4]	0.921 ± 0.062	0.961 ± 0.049	0.923 ± 0.062
Additional Data/ Annotation	USCL [7]	0.889 ± 0.047	0.895 ± 0.054	0.869 ± 0.097
	US-UCL [5]	0.920 ± 0.034	0.926 ± 0.043	0.900 ± 0.046
	GBCNet [3]	0.921 ± 0.029	0.967 ± 0.023	0.919 ± 0.063
	Point-Beyond-Class [18]	0.929 ± 0.013	0.983 ± 0.042	0.731 ± 0.077
SOTA WSOD	TS-CAM [10]	0.862 ± 0.049	0.879 ± 0.049	0.751 ± 0.045
	SCM [2]	0.795 ± 0.101	0.783 ± 0.130	0.849 ± 0.072
	OD-WSCL [21]	0.815 ± 0.144	0.805 ± 0.129	0.847 ± 0.214
	WS-DETR [19]	0.839 ± 0.042	0.843 ± 0.028	0.833 ± 0.034
WSOD	Ours	0.834 ± 0.057	0.817 ± 0.061	0.861 ± 0.089

Table 4. Comparison with SOTA WSOD baselines in classifying Polyps from Colonoscopy images.

Method	Acc.	Spec.	Sens.
TS-CAM [10]	0.704 ± 0.017	0.394 ± 0.042	0.891 ± 0.054
SCM [2]	0.751 ± 0.026	0.523 ± 0.014	0.523 ± 0.016
OD-WSCL [21]	0.805 ± 0.056	0.609 ± 0.076	0.923 ± 0.034
WS-DETR [19]	0.857 ± 0.071	0.812 ± 0.088	0.882 ± 0.034
Point-Beyond-Class [18]	0.953 ± 0.007	0.993 ± 0.004	0.924 ± 0.011
Ours	0.878 ± 0.067	0.785 ± 0.102	0.932 ± 0.022

ing box annotation [3] or, US videos [5,7]. However, even without these additional annotations/ data, our method reaches 86.1% detection sensitivity. The results for polyp classification are reported in Table 4. Although our method has a slightly lower specificity, the sensitivity surpasses the baselines reported in literature [16], and the SOTA WSOD based baselines.

5 Conclusion

GBC is a difficult-to-detect disease that benefits greatly from early diagnosis. While automated GBC detection from US images has gained increasing interest from researchers, training a standard image classification model for this task is challenging due to the low inter-class variance and high intra-class variability of malignant regions. Current SOTA models for GBC detection require costly bounding box annotation of the pathological regions, or additional US video data, which limit their applicability. We proposed to formulate GBC detection as a weakly supervised object detection/ localization problem using a DETR with self-supervised instance learning in a MIL framework. Our experiments show that the approach achieves competitive performance without requiring additional annotation or data. We hope that our technique will simplify the model training at the hospitals with easily available data locally, enhancing the applicability and impact of automated GBC detection.

References

1. Ali, S., et al.: A multi-centre polyp detection and segmentation dataset for generalisability assessment. Scientific Data **10**(1), 75 (2023)
2. Bai, H., Zhang, R., Wang, J., Wan, X.: Weakly supervised object localization via transformer with implicit spatial calibration. In: ECCV. pp. 612–628. Springer (2022). https://doi.org/10.1007/978-3-031-20077-9_36
3. Basu, S., Gupta, M., Rana, P., Gupta, P., Arora, C.: Surpassing the human accuracy: Detecting gallbladder cancer from USG images with curriculum learning. In: CVPR, pp. 20886–20896 (2022)
4. Basu, S., Gupta, M., Rana, P., Gupta, P., Arora, C.: Radformer: transformers with global-local attention for interpretable and accurate gallbladder cancer detection. Med. Image Anal. **83**, 102676 (2023)
5. Basu, S., Singla, S., Gupta, M., Rana, P., Gupta, P., Arora, C.: Unsupervised contrastive learning of image representations from ultrasound videos with hard negative mining. In: MICCAI, pp. 423–433. Springer (2022). https://doi.org/10.1007/978-3-031-16440-8_41
6. Carion, N., Massa, F., Synnaeve, G., Usunier, N., Kirillov, A., Zagoruyko, S.: End-to-end object detection with transformers. In: Vedaldi, A., Bischof, H., Brox, T., Frahm, J.-M. (eds.) ECCV 2020. LNCS, vol. 12346, pp. 213–229. Springer, Cham (2020). https://doi.org/10.1007/978-3-030-58452-8_13
7. Chen, Y., et al.: USCL: pretraining deep ultrasound image diagnosis model through video contrastive representation learning. In: de Bruijne, M., et al. (eds.) MICCAI 2021. LNCS, vol. 12908, pp. 627–637. Springer, Cham (2021). https://doi.org/10.1007/978-3-030-87237-3_60

8. Dietterich, T.G., Lathrop, R.H., Lozano-Pérez, T.: Solving the multiple instance problem with axis-parallel rectangles. Artif. Intell. **89**(1–2), 31–71 (1997)
9. Dosovitskiy, A., et al.: An image is worth 16x16 words: Transformers for image recognition at scale. arXiv preprint arXiv:2010.11929 (2020)
10. Gao, W., et al.: Ts-cam: Token semantic coupled attention map for weakly supervised object localization. In: ICCV, pp. 2886–2895 (2021)
11. Gupta, P.: Imaging-based algorithmic approach to gallbladder wall thickening. World J. Gastroenterol. **26**(40), 6163 (2020)
12. Gupta, P., et al.: Locally advanced gallbladder cancer: a review of the criteria and role of imaging. Abdominal Radiol. **46**(3), 998–1007 (2021)
13. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: CVPR, pp. 770–778 (2016)
14. Hong, E.K., et al.: Surgical outcome and prognostic factors in patients with gallbladder carcinoma. Ann. Hepato-Biliary-Pancreat. Surg. **18**(4), 129–137 (2014)
15. Howlader, N., et al.: Seer cancer statistics review, 1975–2014, national cancer institute, pp. 1–12. Bethesda, MD pp (2017)
16. Jha, D., et al.: Real-time polyp detection, localization and segmentation in colonoscopy using deep learning. IEEE Access **9**, 40496–40510 (2021)
17. Jha, D., et al.: Kvasir-SEG: a segmented polyp dataset. In: Ro, Y.M., et al. (eds.) MMM 2020. LNCS, vol. 11962, pp. 451–462. Springer, Cham (2020). https://doi.org/10.1007/978-3-030-37734-2_37
18. Ji, H., et al.: Point beyond class: A benchmark for weakly semi-supervised abnormality localization in chest x-rays. In: MICCAI. pp. 249–260. Springer (2022). https://doi.org/10.1007/978-3-031-16437-8_24
19. LaBonte, T., Song, Y., Wang, X., Vineet, V., Joshi, N.: Scaling novel object detection with weakly supervised detection transformers. In: WACV, pp. 85–96 (2023)
20. Qian, Z., et al.: Transformer based multiple instance learning for weakly supervised histopathology image segmentation. In: MICCAI, pp. 160–170. Springer Nature Switzerland Cham (2022). https://doi.org/10.1007/978-3-031-16434-7_16
21. Seo, J., Bae, W., Sutherland, D.J., Noh, J., Kim, D.: Object discovery via contrastive learning for weakly supervised object detection. In: ECCV, pp. 312–329. Springer (2022). https://doi.org/10.1007/978-3-031-19821-2_18
22. Shao, Z., Bian, H., Chen, Y., Wang, Y., Zhang, J., Ji, X., et al.: Transmil: transformer based correlated multiple instance learning for whole slide image classification. NeurIPS **34**, 2136–2147 (2021)
23. Szegedy, C., Vanhoucke, V., Ioffe, S., Shlens, J., Wojna, Z.: Rethinking the inception architecture for computer vision. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 2818–2826 (2016)
24. Tang, P., Wang, X., Bai, X., Liu, W.: Multiple instance detection network with online instance classifier refinement. In: CVPR, pp. 2843–2851 (2017)
25. Touvron, H., Cord, M., Douze, M., Massa, F., Sablayrolles, A., Jégou, H.: Training data-efficient image transformers & distillation through attention. In: ICML, pp. 10347–10357. PMLR (2021)
26. Wang, W., et al.: Pvtv 2: Improved baselines with pyramid vision transformer (2021)