HHS Public Access

Author manuscript

Phys Med Biol. Author manuscript; available in PMC 2024 August 07.

Published in final edited form as:

Phys Med Biol. ; 68(16): . doi:10.1088/1361-6560/ace499.

Tumor detection under cystoscopy with transformer-augmented deep learning algorithm

Xiao Jia[1,2,5], **Eugene Shkolyar**[3,4,5], **Mark A Laurie**[2,3], **Okyaz Eminaga**[3,4], **Joseph C Liao**[3,4,*], **Lei Xing**[2,*]

[1]School of Control Science and Engineering, Shandong University, Jinan, People's Republic of China

[2]Department of Radiation Oncology, Stanford University, Stanford, CA, United States of America

[3]Department of Urology, Stanford University, Stanford, CA, United States of America

[4]VA Palo Alto Health Care System, Palo Alto, CA, United States of America

[5]Equal contribution.

Abstract

Objective.—Accurate tumor detection is critical in cystoscopy to improve bladder cancer resection and decrease recurrence. Advanced deep learning algorithms hold the potential to improve the performance of standard white-light cystoscopy (WLC) in a noninvasive and cost-effective fashion. The purpose of this work is to develop a cost-effective, transformer-augmented deep learning algorithm for accurate detection of bladder tumors in WLC and to assess its performance on archived patient data.

Approach.—'CystoNet-T', a deep learning-based bladder tumor detector, was developed with a transformer-augmented pyramidal CNN architecture to improve automated tumor detection of WLC. CystoNet-T incorporated the self-attention mechanism by attaching transformer encoder modules to the pyramidal layers of the feature pyramid network (FPN), and obtained multi-scale activation maps with global features aggregation. Features resulting from context augmentation served as the input to a region-based detector to produce tumor detection predictions. The training set was constructed by 510 WLC frames that were obtained from cystoscopy video sequences acquired from 54 patients. The test set was constructed based on 101 images obtained from WLC sequences of 13 patients.

Main results.—CystoNet-T was evaluated on the test set with 96.4 F1 and 91.4 AP (Average Precision). This result improved the benchmark of Faster R-CNN and YOLO by 7.3 points in F1 and 3.8 points in AP. The improvement is attributed to the strong ability of global attention of CystoNet-T and better feature learning of the pyramids architecture throughout the training. The model was found to be particularly effective in highlighting the foreground information for precise localization of the true positives while favorably avoiding false alarms

*Authors to whom any correspondence should be addressed. jliao@stanford.edu and lei@stanford.edu.

Significance.—We have developed a deep learning algorithm that accurately detects bladder tumors in WLC. Transformer-augmented AI framework promises to aid in clinical decision-making for improved bladder cancer diagnosis and therapeutic guidance.

Keywords

AI-assisted diagnosis; cystoscopy; tumor detection; deep learning; transformer

1. Introduction

Bladder cancer (BCa) is the sixth most common cancer in the US, with an estimated 81,180 new cases in 2022 (Siegel et al 2022). White light cystoscopy (WLC) is the standard endoscopic tool to evaluate the inner surface of the bladder for cancer screening or surveillance. If a tumor or an indeterminate lesion were identified, a WLC-enabled endoscopic surgical procedure called transurethral resection of bladder tumor (TURBT) is performed to remove the tumor and/or biopsy the lesion to establish the pathological diagnosis. About a million cystoscopies are performed annually in the US for screening or surveillance of BCa (O'Sullivan et al 2022). Owing to the high prevalence of disease and high risk of recurrence and progression, adequate tumor detection in cystoscopy is critical for diagnosis, risk stratification, and guiding complete resection. However, not all cancerous areas are readily visible using WLC. Up to 40% of BCa lesions are undetected during initial WLC (Burger et al 2013, Oude Elferink and Witjes 2014).

Blue light cystoscopy (BLC), also known as fluorescence cystoscopy or photodynamic diagnosis, is an adjunct to WLC that enhances detection of BCa through selective tumor uptake of the imaging agent hexaminolevulinate (Daneshmand et al 2014). Despite reported benefit in improved tumor detection, adoption of BLC remains limited due to the increased cost of specialized cystoscopic equipment and imaging agent, as well as clinical workflow demand to instill the imaging agent into the bladder 1 hour in advance of BLC. Moreover, a recent randomized controlled trial of BLC versus WLC failed to demonstrate a benefit to recurrence-free survival at 3 years, but did confer some additional cost (Heer et al 2022). Cost-effective, noninvasive, and easily adoptable adjunct imaging technologies are needed to address the diagnostic shortcomings of both WLC and BLC.

Deep learning has yielded breakthroughs not only in natural image analysis, but also in biomedical applications such as artificial intelligence (AI)-assisted diagnosis. Several efforts have been dedicated to developing deep learning-based approaches that can automatically identify bladder lesions to enhance medical decision-making in WLC (Eminaga et al 2018, Shkolyar et al 2019a, 2019b, Chang et al 2020, Ikeda et al 2020, Yang et al 2021). Previous work by us (Shkolyar et al 2019a, 2019b, Chang et al 2020) has developed a Faster R-CNN (Ren et al 2015)/YOLO (Redmon and Farhadi 2018)-based model, CystoNet, for automated cystoscopic detection of bladder tumor using deep learning.

The evolution of computational technologies has revealed the depth and scale of feature representations are of importance to deep learning-based model performance (He et al 2016, Lin et al 2017). Multi-scale feature extraction with deeper network architectures may improve performance, by learning higher-level representations and richer semantics

of tumors on both input image and feature maps. Furthermore, attention mechanisms can potentially improve tumor detection by augmenting feature maps for model learning. Several attempts have been made to apply attention in conjunction with convolutional neural networks (CNNs) with improvement in accuracy (Hu et al 2018, Shen et al 2021).

Transformers were introduced by Vaswani et al (2017) as a new attention-based building block for natural language processing, and have recently gained popularity in computer vision. Transformers utilize non-local self-attention mechanisms, which can explicitly model interactions of all pixels on feature maps (Carion et al 2020). We hypothesize that augmenting feature learning by combining a CNN detector with transformer-based self-attention can improve bladder tumor detection under cystoscopy.

Herein we propose 'CystoNet-T', a transformer-augmented deep learning-based model for automated tumor detection during standard WLC. CystoNet-T builds upon our prior work (Shkolyar et al 2019a, 2019b, Chang et al 2020) that aims to improve WLC tumor detection by utilizing a transformer-augmented pyramidal CNN architecture. Feature learning is enhanced by aggregating global context information via transformer-based self-attention as well as integrating low- and high-resolution features via feature pyramid networks. This represents the first study to investigate the potential of transformers and attention mechanisms in cystoscopy.

CystoNet-T was trained and evaluated on an annotated development dataset from patients undergoing clinic flexible cystoscopy and TURBT. We show that this new framework enables accurate tumor detection across a range of sizes, morphologies, and locations.

2. Materials and methods

2.1. CystoNet-T for bladder tumor detection

An overview of the CystoNet-T structure is shown in figure 1. CystoNet-T is composed of three main components: a ResNet50 backbone to extract compact feature representations, a transformer encoder with a feature pyramidal architecture to augment feature learning, and a region-based detector that determines the final detection prediction.

Backbone.—The backbone of ResNet50 (He et al 2016) consists of one convolution level C1 and four residual levels {R2, R3, R4, R5}. The output width and height in {C1, R2, R3, R4, R5} have a spatial scale of {1/2, 1/4, 1/8, 1/16, 1/32} of the input image $\mathbf{x} \in \mathbb{R}^{3 \times H_0 \times W_0}$ (with 3 color channels). Feature activation maps $\mathbf{f}_i \in \mathbb{R}^{C_i \times H_i \times W_i}$ are generated by residual level i and connected to the pyramid components. The standard architecture of ResNet50 with the standard values $C_5 = 2048$ and $H_5, W_5 = \frac{H_0}{32}, \frac{W_0}{32}$ was followed.

Transformer encoder.—Transformer encoder expects an input sequence to learn the global context information through self-attention. Hence, we flatten the high-level feature activation $\mathbf{f}_5$ by first reducing its channel dimension (from $C_5 = 2048$ to a smaller dimension $C_d = 256$) with a 1×1 convolution, then collapsing the spatial dimensions to one to produce a $C_d \times H_5 W_5$ feature map. The resulting feature maps are supplemented with positional

encoding to preserve positional information, and then utilized as the input to the transformer encoder. We applied the transformer with 6 encoder layers, each with a standard architecture (Carion et al 2020, Dosovitskiy et al 2020). Figure 1(b) illustrates a transformer encoder layer that consists of a multi-head self-attention module, layer normalization, and a feed forward network (FFN). The FFN is composed of two layers of 1×1 convolutions with ReLU activations. The output of the transformer is reshaped from $C_d \times H_5 W_5$ to $C_d \times H_5 \times W_5$ and added to the input of FPN. We note that while the encoder input and encoder output have the same dimensionality, they may not have a one-to-one correspondence in terms of the information they represent. The encoder output typically consists of encoded representations of the input sequence, which are often aggregated or summarized representations capturing the important features of the input.

Transformer-augmented FPN.—Feature pyramid networks (FPNs) can construct feature pyramids from a single-scale input using a top-down pathway and lateral connections (Lin et al 2017). FPN architecture allows feature integration across resolutions and semantic levels, making it particularly suitable for the detection of various tumor sizes and morphologies. We built four pyramid levels referring to the number of residual modules (figure 1 (a)). Here feature maps in the FPN top-down pathway are augmented by the transformer encoder with non-local features aggregation and attention weights. The resulting feature set of the transformer-augmented FPN is called {P2a, P3a, P4a, P5a} and serves as the pyramid component to produce multi-scale predictions. Furthermore, a potential issue related to the data interface of a transformer model is the possibility of information loss during the data conversion process at the encoder input. Given the computational demands of transformers, our approach in designing CystoNet-T involves operating the encoder input at a higher semantic level (R5) derived from ResNet. This may lead to the loss of certain fine-grained details, such as very small or subtle visual patterns, in the encoder input representation. To mitigate information loss and retain as much relevant information as possible, we have incorporated lateral connections from the FPN into CystoNet-T. These connections establish links to the R2–R4 stages of ResNet, allowing CystoNet-T to access both high-level and low-level information throughout the network. By incorporating these connections, we ensure the preservation and propagation of important information that might otherwise be lost or diluted during the forward pass through the network.

Region-based detector.—After pyramidal features are extracted and augmented via global attention, they are fed into a region-based detector to produce detection predictions. The resulting bounding boxes give an indication of the lesion location, which can provide auxiliary information for diagnosis and guide TURBT. The predictor head follows the Faster R-CNN approach described in (Ren et al 2015). CystoNet-T is developed by attaching the region proposal network (RPN) to all levels of transformer-augmented FPN, i.e. {P2a, P3a, P4a, P5a}, to generate multi-scale proposal regions. We use region of interest (RoI) pooling to reshape the proposals into fixed-size feature maps, which serve as the input to a sequence of fully connected (fc) layers. The final detection output is generated in parallel by a softmax classifier and a bounding box regressor.

2.2. Sub-networks

Positional encoding.—Network activations are associated with the spatial positions of image features. Since the transformer expects a sequence of vectors as input, we use two-dimensional positional embedding to retain positional information (Vaswani et al 2017). The positional encoding ensures that the model can differentiate between different positions within the input sequence, compensating for the lack of inherent spatial awareness in the transformer architecture. Here, we use two sets of fixed absolute encoding, each for one of the feature axes, i.e. x-embedding, and y-embedding, each with size $C_d / 2$. The final positional encoding with size C_d is generated by concatenating the x and y embedding. Specifically, for a feature pixel with position (x, y), its positional encoding is defined as $[PE(x): PE(y)]$, where [:] denotes concatenation and function PE is defined by:

$$\begin{aligned} PE(x, 2i) &= \sin(x \,/\, 10000^{x \,/\, C_d}) \\ PE(x, 2i+1) &= \cos(x \,/\, 10000^{x \,/\, C_d}), \end{aligned} \tag{1}$$

where C_d denotes the dimensionality of the encoder input and output. In our model, we use $C_d = 256$. $i \in [0, 1, \ldots, C_d / 2)$ is the dimension for each embedding of the spatial coordinates x and y.

Region proposal network.—The region proposal network (RPN) was introduced by Ren et al (2015) to generate a set of RoIs before task-specific detection. Anchor was defined in RPN as a set of reference boxes with multiple scales and aspect ratios to cover objects of various shapes and sizes, which is of utility in bladder tumor detection. In its original design, anchors were assigned on the last backbone layer R5, however we designed the anchors with scales of $\{32^2, 64^2, 128^2, 256^2\}$, each with multiple aspect ratios of {1: 1, 1: 2, 2: 1} for {P2a, P3a, P4a, P5a}, respectively. As such, anchors are assigned to each of the transformer-augmented FPN components, allowing the model to densely cover all locations in all scale levels of the input. We also apply an RoI pooling layer as max-pooling, which performs downsampling of arbitrarily sized features at the proposal regions and produces feature maps with a small, fixed spatial size. Here 7 × 7 is used by default.

2.3. Dataset

With institutional review board approval, WLC images were collected from patients undergoing clinic flexible cystoscopy and TURBT in the operating room.

The dataset was constructed from 611 WLC images containing histologically confirmed papillary urothelial carcinoma, where tumors were annotated and outlined by expert urologists. The dataset was built to cover as many varieties of bladder tumors as possible, such as having differences in morphology, size, location, and illumination. To implement the detection algorithm, a ground truth setting consisting of a tumor bounding box was generated based on the outline annotation to fit the contour of a tumor. We illustrate the variation of the WLC dataset in figure 2 in terms of relative tumor size, which calculates the size ratio between the tumor location box (bounding box sufficiently fitting the tumor area) and the corresponding WLC frame. The values scattered along the x-axis indicate large variations in tumor scales for detection.

The training set contained 510 labeled frames obtained from WLC sequences of 54 patients. The test set contained 101 labeled frames acquired from a separate set of 13 patients. Some frames contain more than one tumor and thus increase the complexity of detection. We imposed the constraint that a patient's records cannot be divided across different sets to facilitate the evaluation of the effectiveness and generalizability of the proposed method. This constraint ensures that the model's performance can be assessed on unseen data and prevents it from becoming overly specialized to the training dataset.

2.4. Implementation details

CystoNet-T was implemented using PyTorch (Paszke et al 2019) on a single *NVIDIA GeForce GTX TITAN Xp.* The backbone and encoder were pretrained on COCO dataset (Carion et al 2020), and the remaining layer weights were randomly initialized. For the region-based detector head, the Intersection-over-Union (IoU) threshold of nonmaximum suppression (NMS) was set to 0.7 following (Ren et al 2015), which helps remove RoIs that overlap with others that have higher scores. The anchor was assigned to negative if it had an IoU lower than 0.3. Dropout was employed during the training phase of CystoNet-T, encouraging the model to learn more robust and generalized representations.

CystoNet-T was trained end-to-end to minimize multi-task loss function $L(p, p^*, t, t^*)$ which measures the matching of the predicted and annotated lesions:

$$L(p, p^*, t, t^*) = L_{\text{cls}}(p, p^*) + p^* L_{\text{reg}}(t, t^*), \tag{2}$$

where p is the predicted probability distribution of tumor (positive) and background (negative), t is a vector showing the four coordinates of the predicted bounding box. p^* and t^* represent the ground-truth label and bounding-box regression target, respectively. $p^* = 1$ if the target is an area presenting a tumor, and $p^* = 0$ otherwise. $L_{\text{cls}}(p, p^*) = -\log(p_{p^*})$ for true class p^* is a log loss for positive-negative classification. We used the smooth L_1 loss defined in (Girshick 2015) for bounding box regression:

$$\text{smooth}_{L_1}(x) = \begin{cases} 0.5x^2 & \text{if} \quad |x| < 1 \\ |x| - 0.5 & \text{otherwise,} \end{cases} \tag{3}$$

and $L_{\text{reg}}(t, t^*) = \text{smooth}_{L_1}(t - t^*)$ is activated for tumor targets.

We used scale augmentation, resizing the input images such that the shorter side was at least 640 and at most 800 pixels while the longer side was at most 960. We applied simple transformation techniques, such as random rotations and flips, provided by the PyTorch library to artificially increase the size and diversity of the training dataset and enhance the model's generalization ability. We trained the model with an initial learning rate of 10^{-4}. The training schedule was set to 300 epochs with a decay factor of 0.1 after 200 epochs.

2.5. Evaluation metrics

We evaluated CystoNet-T performance as the accurate identification of lesion location in a frame containing tumor. We followed the measurements described in (Bernal et al 2017).

IoU metric evaluates the degree of overlap between the ground truth (GT) and prediction (PR), and it is calculated as an area of intersection divided by the area of union between the ground truth and predicted box:

$$\mathrm{IoU} = \frac{|\ \mathrm{PR} \cap \mathrm{GT}\ |}{|\ \mathrm{PR} \cup \mathrm{GT}\ |}, \tag{4}$$

where ∩ represents the set intersection and ∪ represents set union. For IoU threshold at λ, true positive (TP) is a detection for which IoU(GT, PR) ⩾ λ and false positive (FP) is a detection for which IoU(GT, PR) < λ. False negative (FN) is a ground-truth tumor missed together with GT for which IoU(GT, PR) < λ. We use λ = 0.5 in our case to evaluate detection accuracy. For multiple predictions generated on the input, only one TP is considered per tumor. The detection with the highest score is considered a TP and others are considered to be FP. Recall (R) is the proportion of TPs among all GTs and precision (P) is the proportion of TPs among all detections produced by the model. F1 is the harmonic average of R and P that calculates the balance between precision and recall:

$$\begin{aligned} \mathrm{R} &= \frac{\mathrm{TP}}{\mathrm{TP+FN}} \\ \mathrm{P} &= \frac{\mathrm{TP}}{\mathrm{TP+FP}} \\ \mathrm{F1} &= \frac{2 \times \mathrm{P} \times \mathrm{R}}{\mathrm{P} + \mathrm{R}} \end{aligned} \tag{5}$$

We also report the average precision evaluated at IoU threshold λ = 0.5 simply as AP to provide an overall evaluation, calculating the area under the precision-recall curve.

3. Results

In figure 3, we display the self-attention of the transformer with a reference point (red) on the test input. Attention mechanisms in the encoder allow the model to focus on the informative and relevant regions when predicting tumors. Furthermore, self-attention enables global reasoning over whole image context, and thus can model relations and interactions between features of different elements.

With global attention performed, we show activation maps of the transformer-augmented FPN in figure 4. Feature representations at the low-resolution level of FPN (P5 in figure 4) were augmented by aggregating non-local context of self-attention and activations across resolutions and semantic levels. Predictions were made on pyramidal levels of P5a–P2a and together contributed to the final performance of precise tumor detection.

Representative bladder tumor detection images are shown in figure 5. Predictions were made by CystoNet-T on the test set. The results cover scenes with tumors of various sizes (from large to small), morphologies, and locations (see figure 5(A)-(J)). We observed accurate predictions (blue and orange) closely associated to the ground truth (green) across a range of lesion cases, including challenging examples of clustered multifocal tumors (figure 5(K)-(M)), complex image backgrounds (figure 5(N)), and obscure tumor regions of low contrast with the background (figure 5(O)). We believe that precise localization was enabled

by leveraging both attention maps from global computations and feature representations from pyramidal levels.

A full comparison of detection performance of CystoNet-T and the latest detectors for bladder tumor detection is presented in table 1. Faster R-CNN and YOLO in the comparison represent the adapted versions of Faster R-CNN (Ren et al 2015) and YOLO (Redmon and Farhadi 2018) that were optimized for cystoscopic detection of bladder tumors in our previous work (Shkolyar et al 2019a, 2019b, Chang et al 2020). They were also the most popular and competitive detectors on detection tasks. We attempted to further optimize Faster R-CNN by upgrading the original backbone of VGG to ResNet50 to align it with competitors. FPN (Lin et al 2017) was also optimized for bladder tumor detection (by Yoo et al (Yoo et al 2022)) and served as a baseline method of the ResNet50 + FPN backbone for comparison with the proposed CystoNet-T. Models were trained and tested with the same data setting to ensure a fair comparison. CystoNet-T demonstrated significantly better performance on bladder tumor detection, achieving 96.4 F1 and 91.4 AP on the test set. This improves the well-established benchmark detectors of YOLO and Faster R-CNN by 8.1 points in R and 5.7 points in P, resulting in a 7.3 F1 and 3.8 AP improvement in detection performance. The gains of CystoNet-T arise from the effective exploitation of multi-scale information and global scene reasoning via transformer-augmented FPN. This development leads to better feature learning throughout training and highlights the foreground information for precise localization of the true positives while favorably avoiding false alarms We further evaluated the importance of self-attention mechanism by quantitatively comparing CystoNet-T with the baseline FPN. Without global self-attention of transformers, performance drops by 3.2 points in F1 and 1 point in AP, we thus conclude that non-local computations allowed by encoder attention are important for achieving accurate detection. Additionally, the reported results provide insights into the model's performance on the unseen data from the test set, indicating its robustness against overfitting.

4. Discussion

This study aimed to develop an AI-assisted approach to improve the accuracy of automated bladder tumor detection, with the eventual goal of seamless integration with WLC in clinical settings. Compared to natural image processing, bladder tumor detection presents greater challenges due to the low contrast between lesions and the bladder wall, as well as considerable intraclass variation in tumor morphologies and sizes. Recent advances in deep learning-based image processing have shown the potential to optimize detection performance through enhanced feature learning capabilities and sophisticated network architectures. The attention mechanism provides a powerful tool for capturing and emphasizing the important features for tumor identification, leading to more accurate detection outcomes. Furthermore, the self-attention mechanism employed by the transformer is particularly useful for lesion detection tasks. It enables the model to capture both local and global information, allowing it to focus on fine-grained details within specific lesion regions while considering the broader context and global relationships between different regions (please refer to figure 4 for an illustration of feature augmentation through self-attention). This selective attention improves the model's ability to discriminate between lesions

and normal tissues and enhances its capability to detect lesions with various variations. Additionally, since tumors vary in size, incorporating features from different resolutions can further enhance the performance of tumor detection. These factors served as strong motivation for the development of a specialized framework for cystoscopic detection of bladder tumors. Our framework incorporates multi-scale attention-augmented lesion feature extraction, allowing us to effectively address the challenges associated with accurate tumor detection in cystoscopy.

The majority of existing methods have focused on image-level classification, generating text predictions per image input. Early work on classifying cystoscopy images based on hand-engineered color features has achieved good sensitivity for tumor identification, but with a false positive rate of 50% (Gosnell et al 2018). Convolutional neural networks (CNNs) were introduced in the work of Eminaga et al (2018) for WLC image classification, where training and validation were performed on a curated WLC image atlas. The work of (Ikeda et al 2020, Yang et al 2021) evaluated typical CNN architectures for their performance in distinguishing images of bladder tumors, and they demonstrated accuracy gains from increased network depth.

Compared to image-level classification, the output of region-level detection algorithms can indicate the lesion location on the input image which is of clinical utility. In our previous work (Shkolyar et al 2019a, 2019b, Chang et al 2020) we developed a deep learning model, CystoNet, for cystoscopic detection of bladder tumors, with a CNN detector based on benchmark methods of Faster R-CNN (Ren et al 2015) and YOLO (Redmon and Farhadi 2018). We recognize the importance of region-level detection algorithms in two regards. One, the bounding box detection of tumors enables explainable and reliable AI predictions, and two, an indication of the lesion location can provide auxiliary information for diagnosis and guide urologists in performing targeted tumor resection.

The deep learning-based algorithm presented in this paper, CystoNet-T, provides an effective solution to the challenge of accurate tumor detection in WLC images. CystoNet-T is an upgraded version of CystoNet that enables more precise tumor localization by transformer-augmented pyramidal feature learning. To the best of our knowledge, this is the first attempt to explore the potential of global self-attention of transformers in cystoscopic tumor detection. We also evaluated multiple competitive baselines and demonstrated the superior performance of CystoNet-T over these alternatives.

To show that CystoNet-T achieves better performance for tumor detection, figure 6 plots the precision-recall curve of CystoNet-T and the leading deep learning-based methods (YOLO, Faster R-CNN, FPN) for bladder tumor detection. Models follow the same configuration described in table 1, area under the precision-recall curve referring to the AP values. CystoNet-T showed the best performance in terms of area under the precision-recall curve with the highest 91.4 AP, against the three other methods. Figure 7 shows the results of CystoNet-T (figure 7(A*)-(D*)) and the benchmark detector of Faster R-CNN (figure 7(A)-(D)) on the test set. Faster R-CNN generated false alarms (IoU(GT, PR) < 0.5) in cases of a large and obscure tumor (figure 7(A)) and a small tumor located at the very top corner of the image (figure 7(B)). Faster R-CNN produced no activation in figure 7(C) with a tumor with

irregular morphology present in the frame. CystoNet-T had no errors in the same examples compared to Faster R-CNN (see figure 7(A*)-(C*)). CystoNet-T enables precise localization of predicting boxes around bladder tumor regions, across a range of sizes, morphologies, and locations. transformer-augmented FPN is the key element that significantly contributes to the performance gain, which improves network feature learning by aggregating information of global self-attention and feature pyramids.

CystoNet-T runs at 4 FPS (frame per second) on a single GPU of *NVIDIA GeForce GTX TITAN Xp,* with inference time of ~0.231 s for each WLC image input, similar to FPN with ~0.227 s inference time. As a reference, benchmark methods of Faster R-CNN and YOLO run at 5 and 6 FPS, respectively. CystoNet-T increases runtime due to the extra costs of feature pyramids and transformer self-attention, but with clear gains in detection accuracy.

There are a few limitations to our work. In figure 7(D) and (D*) we show a failure case with multifocal tumors. Both Faster R-CNN and CystoNet-T missed one of the two tumors located on the bottom edge, failing to generate a positive alarm on the lesion region. CystoNet-T is likely to further improve with information aggregation from temporal dimension and training data with more morphologies. To make the model applicable in real-time, more efficient pre-processing and inference schemes are needed to reduce frame redundancy during WLC screening. Despite these limitations, this study represents a significant step forward in accurate tumor detection, with performance superior over current state-of-the-art models.

In the future, sequential data of cystoscopy videos will be included for model development and validation, extending the model from still-frame analysis. In addition, training and performance evaluation will need to be done with an extended variety of tumor morphologies, such as flat lesions. Conducting further studies with larger and more diverse datasets would also be valuable for validating the generalization capability of the proposed CystoNet-T model and assessing potential overfitting. Furthermore, we recognize the significance of achieving high segmentation accuracy in the context of the TURBT procedure. Mask R-CNN (He et al 2017) is a widely used framework for instance detection and segmentation. It employs ResNet-FPN as the backbone for feature extraction. The proposed CystoNet-T, with its transformer-augmented FPN backbone, can be viewed as an evolution of the FPN detector branch of Mask R-CNN. By incorporating transformer-based self-attention to each pyramid layer, it enables global feature aggregation. A more comprehensive study capable of generating real-time and accurate tumor outlines can be developed in the future by using the detection results of the CystoNet-T as valuable prompts for lesion segmentation. We note that CystoNet-T is quite general and easily generalized to various other endoscopic imaging modalities without significant modifications. We believe the findings of this study will facilitate AI-assisted applications in real-world cystoscopy workflow with increased accuracy of cancer detection and aid in the prompt diagnosis of bladder cancer using WLC.

5. Conclusion

We have developed CystoNet-T, a transformer-augmented deep-learning algorithm, for accurate bladder tumor detection in WLC. This method greatly improves our previous work, CystoNet. Our experiments show that CystoNet-T achieves excellent results across a range of tumor sizes, morphologies, and locations. Hence, it holds promise in aiding bladder cancer screening and improving the diagnostic decision-making of WLC.

Acknowledgments

The authors gratefully acknowledge research support from NIH R01 CA260426 (JCL and LX), Department of Veterans Affairs BLR&D I01 BX005598 (JCL), Natural Science Foundation of Shandong Province for Distinguished Young Scholars (Overseas) 2023HWYQ-027 (X J), and the Urology Care Foundation (E S).

Data availability statement

The data cannot be made publicly available upon publication because they contain sensitive personal information. The data that support the findings of this study are available upon reasonable request from the authors.

References

Bernal J et al. 2017 Comparative validation of polyp detection methods in video colonoscopy: Results from the MICCAI 2015 endoscopic vision challenge IEEE Trans. Med. Imaging 36 1231–49 [PubMed: 28182555]

Burger M et al. 2013 Photodynamic diagnosis of non-muscle-invasive bladder cancer with hexaminolevulinate cystoscopy: A meta-analysis of detection and recurrence based on raw data Eur. Urol 64 846–54 [PubMed: 23602406]

Carion N, Massa F, Synnaeve G, Usunier N, Kirillov A and Zagoruyko S 2020 End-to-end object detection with transformers European Conference on Computer Vision pp 213–29

Chang T, Shkolyar E, Jia X, Lee T, Mach K, Xing L and Liao J 2020 V12–01 real-time augmented bladder tumor detection with deep learning J. Urol 203 e1110–1110

Daneshmand S et al. 2014 Hexaminolevulinate blue-light cystoscopy in non-muscle-invasive bladder cancer: Review of the clinical evidence and consensus statement on appropriate use in the USA Nat. Rev. Urol 11 589–96 [PubMed: 25245244]

Dosovitskiy A et al. 2020 An image is worth 16x16 words: transformers for image recognition at scale arXiv:2010.11929

Eminaga O, Eminaga N, Semjonow A and Breil B 2018 Diagnostic classification of cystoscopic images using deep convolutional neural networks, JCO Clinical Cancer Inf. 2 1–8

Girshick R 2015 Fast r-cnn Proc. of the IEEE Int. Conf. on Computer Vision (10.1109/iccv.2015.169)

Gosnell ME, Polikarpov DM, Goldys EM, Zvyagin AV and Gillatt DA 2018 Computer-assisted cystoscopy diagnosis of bladder cancer Urologic Oncol.:Semin. Original Invest 36 8.e9–8.e15

He K, Gkioxari G, Dollár P and Girshick R 2017 Mask r-cnn Proce. of the IEEE Int. Conf. on Computer Vision (10.1109/iccv.2017.322)

He K, Zhang X, Ren S and Sun J 2016 Deep residual learning for image recognition Proc. of The IEEE Conf. on Computer Vision and Pattern Recognition (10.1109/cvpr.2016.90)

Heer R et al. 2022 A randomized trial of photodynamic surgery in non-muscle-invasive bladder cancer New Engl. J. Med., Evidence 1 10

Hu H, Gu J, Zhang Z, Dai J and Wei Y 2018 Relation networks for object detection Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition (10.1109/cvpr.2018.00378)

Ikeda A, Nosato H, Kochi Y, Kojima T, Kawai K, Sakanashi H, Murakawa M and Nishiyama H 2020 Support system of cystoscopic diagnosis for bladder cancer based on artificial intelligence J. Endourol 34 352–8 [PubMed: 31808367]

Lin T-Y, Dollár P, Girshick R, He K, Hariharan B and Belongie S 2017 Feature pyramid networks for object detection Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition (10.1109/cvpr.2017.106)

O'Sullivan S, Janssen M, Holzinger A, Nevejans N, Eminaga O, Meyer C and Miernik A 2022 Explainable artificial intelligence (xai): closing the gap between image analysis and navigation in complex invasive diagnostic procedures World J. Urol 40 1125–34 [PubMed: 35084542]

Oude Elferink P and Witjes JA 2014 Blue-light cystoscopy in the evaluation of non-muscle-invasive bladder cancer Ther. Adv. Urol 6 25–33 [PubMed: 24489606]

Paszke A et al. 2019 Pytorch: an imperative style, high-performance deep learning library Adv. Neural Inf. Process. Syst 32 8026–37

Redmon J and Farhadi A 2018 Yolov3: An incremental improvement arXiv:1804.02767

Ren S, He K, Girshick R and Sun J 2015 Faster r-cnn: Towards real-time object detection with region proposal networks Adv. Neural Inf. Process. Syst 28 1137–49

Shen Y, Jia X and Meng MQ-H 2021 Hrenet: a hard region enhancement network for polyp segmentation Int. Conf. on Medical Image Computing and Computer-Assisted Intervention (10.1007/978-3-030-87193-2_53)

Shkolyar E, Jia X, Chang TC, Trivedi D, Mach KE, Meng MQ-H, Xing L and Liao JC 2019a Augmented bladder tumor detection using deep learning Eur.n Urol 76 714–8

Shkolyar E, Jia X, Xing L and Liao J 2019b LBA-20 Automated cystoscopic detection of bladder cancer using deep-learning J. Urol 201 e1000–1

Siegel RL, Miller KD, Fuchs HE and Jemal A 2022 Cancer statistics CA: A Cancer J. Clinicians 72 7–33

Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez AN, Kaiser Ł and Polosukhin I 2017 Attention is all you need Adv. Neural Inf. Process. Syst 30 6000–10 arXiv:1706.03762

Yang R, Du Y, Weng X, Chen Z, Wang S and Liu X 2021 Automatic recognition of bladder tumours using deep learning technology and its clinical application Int. J. Med. Robot. Comput. Assist. Surg 17 e2194

Yoo JW, Koo KC, Chung BH, Baek SY, Lee SJ, Park KH and Lee KS 2022 Deep learning diagnostics for bladder tumor identification and grade prediction using RGB method Sci. Rep 12 17699 [PubMed: 36271252]

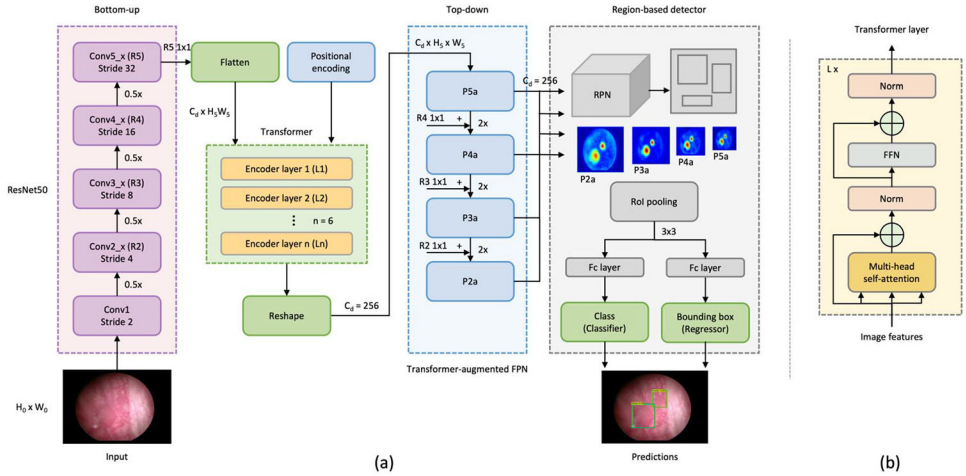

Figure 1.

CystoNet-T framework for bladder tumor detection. (a) CystoNet-T architecture. CystoNet-T uses the backbone of ResNet50 and FPN to generate high-level and multi-scale feature representations. Feature maps of the last residual block are flattened and supplemented with positional encoding before passing into the transformer encoder. {P2a, P3a, P4a, P5a} are feature pyramid components augmented by the transformer. A region-based detector is attached to all pyramid levels as the predictor head to produce the detection result, i.e. class and bounding box. (b) Illustration of a building block for the transformer encoder layer with multi-head self-attention. FPN = feature pyramid network; RoI = region of interest; RPN = region proposal network; Fc layer = fully connect layer; FFN = feed forward network.

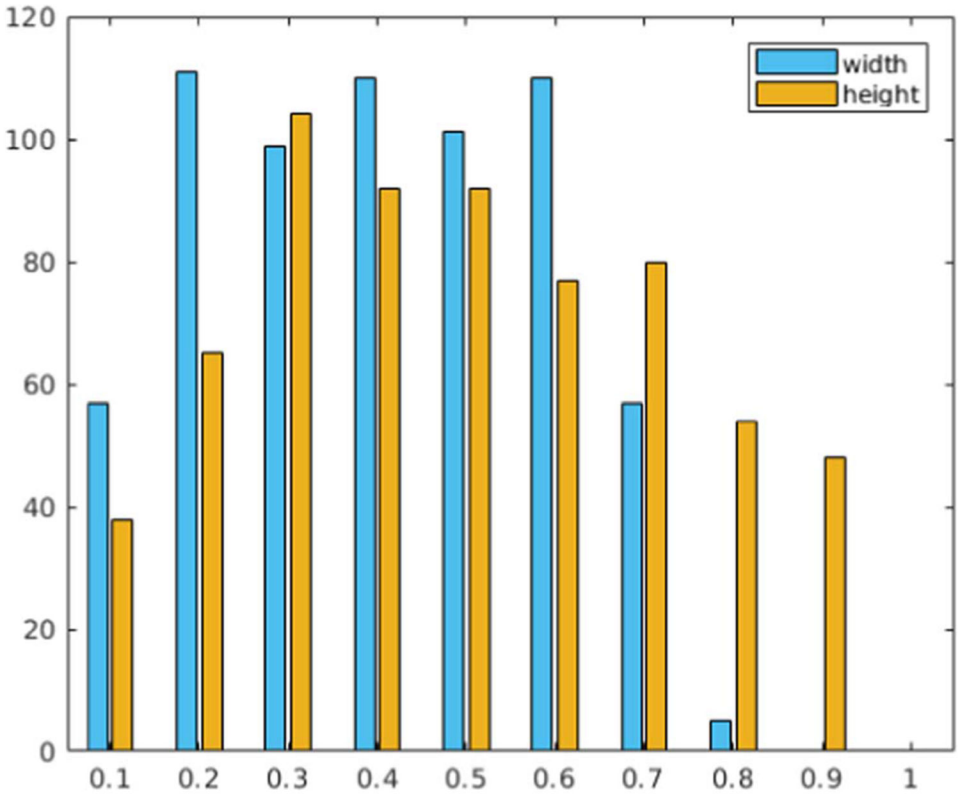

Figure 2.

Distribution of the relative tumor size in the WLC dataset. The values of *x*-axis represent the size ratio between the tumor location box and the corresponding WLC image.

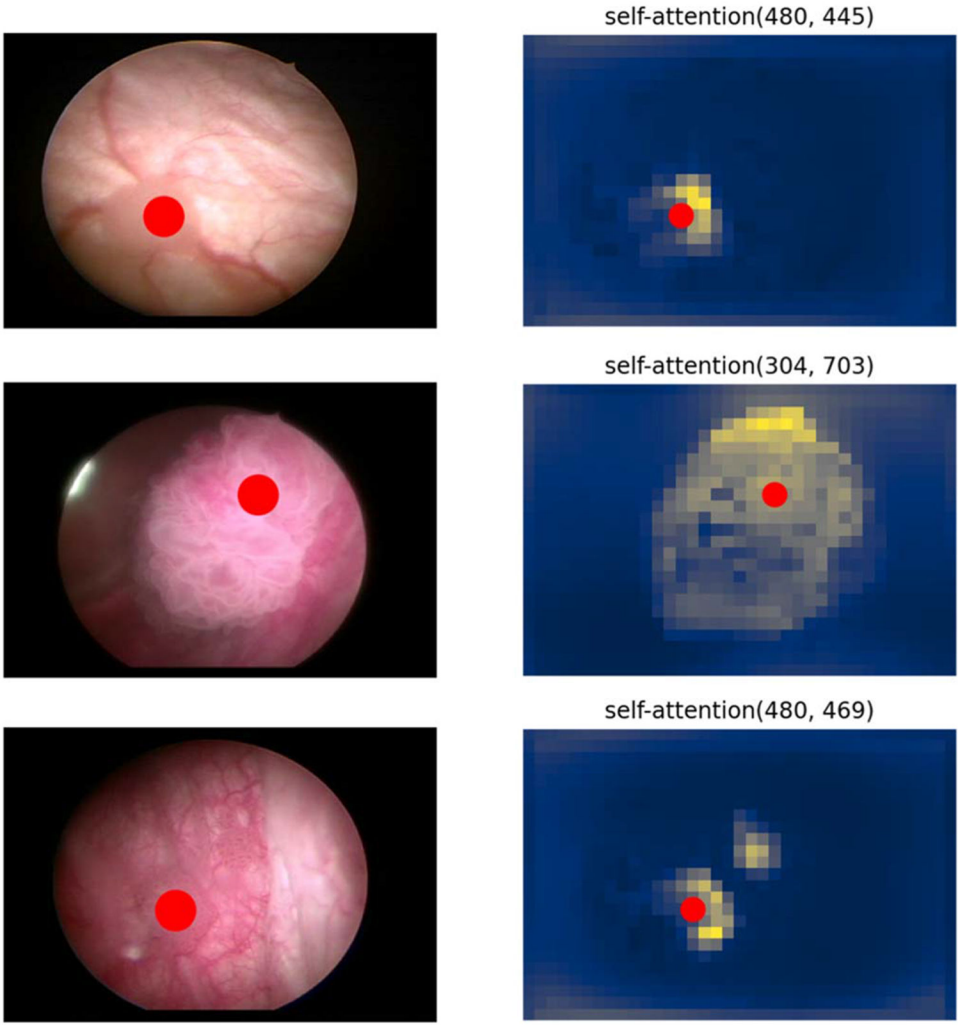

Figure 3.

Attention maps of the last encoder layer of transformer. Red dots represent the reference points of self-attention. From top to bottom shows predictions on a test set image with a small tumor region (top); a large tumor region (middle); multifocal tumor regions (bottom).

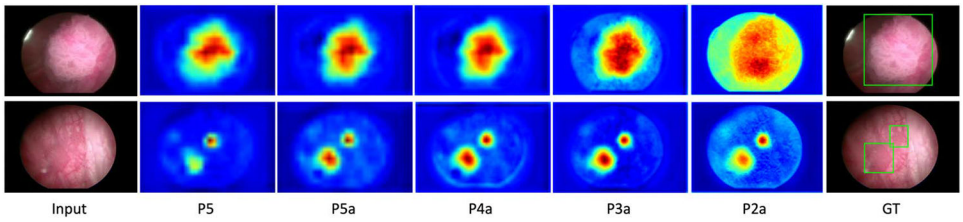

Figure 4.

Feature activation maps of the pyramid levels of transformer-augmented FPN. From left to right: input of a test set image; low-resolution activation maps of FPN w/o transformer augmentation; activation maps of transformer-augmented sets {P5a, P4a, P3a, P2a} at different resolution levels (low to high); ground truth boxes (green). Predictions are augmented by aggregating information from pyramidal feature levels and non-local self-attention.

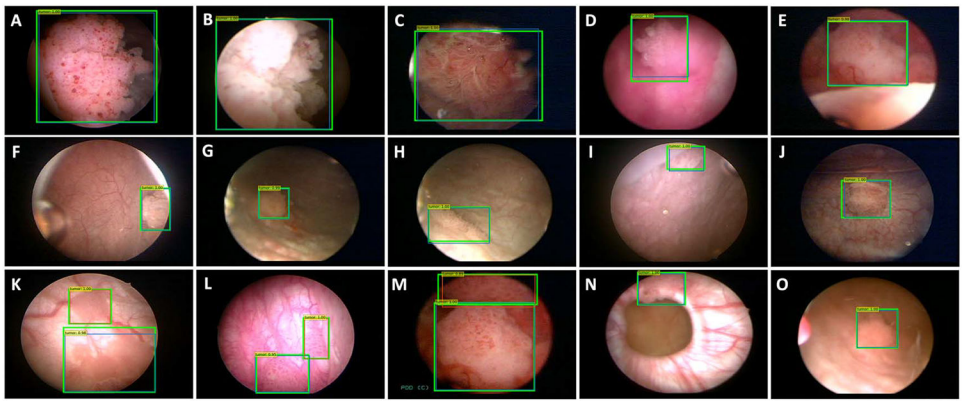

Figure 5.

Representative examples of bladder tumor detection using CystoNet-T. Ground truth is in green and the predicted detection is in blue and orange. Each output box is shown with a yellow class label of tumor and an associated prediction score in [0, 1]. CystoNet-T enables precise detection for tumors of various sizes, morphologies, and locations (A–J), and some challenging scenes with multifocal tumor presence (K-M), complex background (N), or obscure lesions with limited background contrast (O).

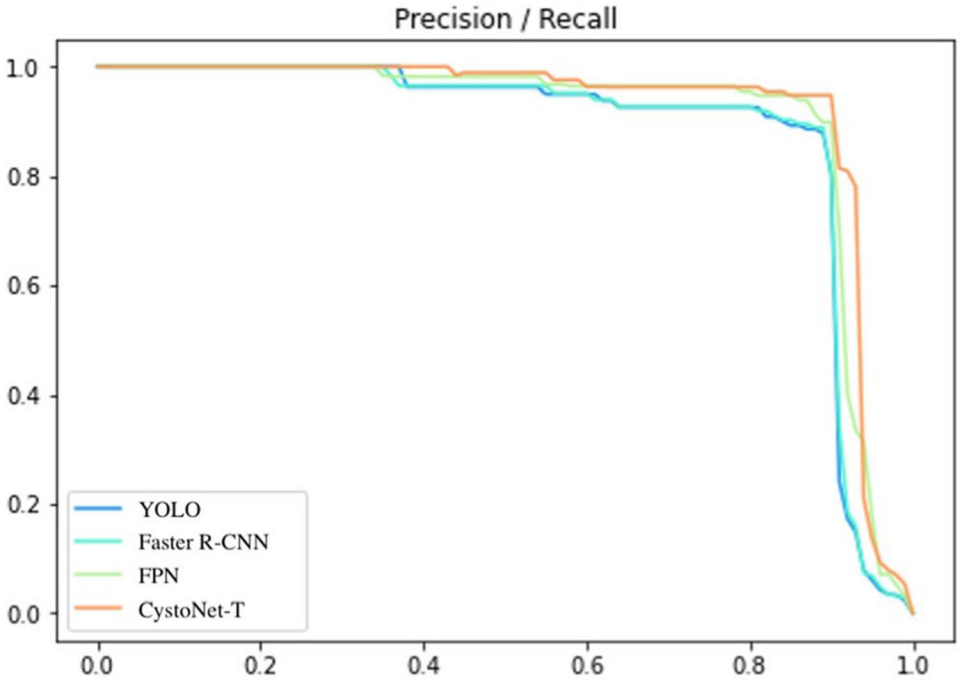

Figure 6.

Precision-recall curve of CystoNet-T versus competitive CNN-based methods (YOLO, Faster R-CNN, FPN) in detecting bladder tumors. All settings are identical to the same methods in table 1. Area under the precision-recall curve is calculated as AP in table 1. CystoNet-T shows the best performance with the largest area under the precision-recall curve of 91.4 AP.

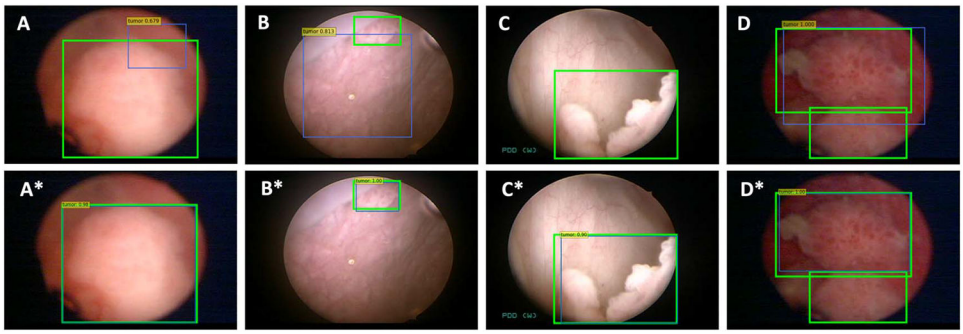

Figure 7.

Comparison of prediction results on the test set. Benchmark detector of Faster R-CNN (top) and CystoNet-T (bottom). Green boxes represent ground truth. Blue boxes with a yellow class label and prediction score attached on top indicate predictions generated by detectors. Faster R-CNN shows poor results in detection of a large, obscure tumor (A) and a small tumor located at the very top corner of the image (B), and shows no activation in the inference of tumor with irregular morphology (C). CystoNet-T can generate accurate detections very close to ground truth (A*–C*). Faster R-CNN and CystoNet-T exhibit errors in D and D*, respectively, where both failed to accurately detect one of two tumor regions presenting on the bottom edge.

Table 1.

Comparison of bladder tumor detection with benchmark detectors of Faster R-CNN and YOLO on the test set. FPN is the baseline of our method. The comparison items, namely Faster R-CNN, YOLO, and FPN, are adapted versions of the original Faster R-CNN (Ren et al 2015), YOLO (Redmon and Farhadi 2018), and FPN (Lin et al 2017), respectively, that have been optimized for cystoscopic detection of bladder tumors (Shkolyar et al 2019a, 2019b, Chang et al 2020, Yoo et al 2022). Methods were trained and tested with the same data settings for a fair comparison. 9 frames on the test set have two tumors present. The sum of TP and FN for each model is 111, which is equal to the total number of distinct tumors in the test set. Boldface indicates the best performance. CystoNet-T shows a significant improvement of 3.8 AP over the benchmark. TP = true positive; FP = false positive; FN = false negative; R = recall; P = precision; AP = average precision evaluated at IoU threshold $\lambda = 0.5$.

Method	Backbone	TP	FP	FN	R [%]	P [%]	F1 [%]	AP [%]
YOLO	Darknet	98	11	13	88.3	89.9	89.1	87.5
Faster R-CNN	ResNet50	99	13	12	89.2	88.4	88.8	87.6
FPN	ResNet50 + FPN	103	7	8	92.8	93.6	93.2	90.4
CystoNet-T	ResNet50 + transformer-augmented FPN	**108**	**5**	**3**	**97.3**	**95.6**	**96.4**	**91.4**