TEmory: A Temporal-Memory Approach to Weakly Supervised Colonic Polyp Frame Detection

Yufei Liu[1], Jianzhe Gao[1], Zhiming Luo[1,2(✉)], and Shaozi Li[1,2]

[1] Department of Artificial Intelligence, Xiamen University, Xiamen, Fujian, China
[2] Fujian Key Laboratory of Big Data Application and Intellectualization for Tea Industry, Wuyi University, China
zhiming.luo@xmu.edu.cn

Abstract. In recent years, weakly supervised colonic polyp frame detection based on colonoscopy, widely recognized as the 'gold standard' for colorectal polyps screening, has attracted significant attention. However, in colonoscopy videos, a minority of normal frames exhibit a distribution that differs from the majority, potentially affecting the assessment of frame abnormality. To address these challenges, we propose TEmory, a model featuring a **T**emporal **E**ncoder and Me**mory** Unit, designed for weakly supervised colonic polyp frame detection with a comprehensive understanding of normal frame characteristics. Specifically, the Temporal Encoder leverages the contextual information of adjacent frames within video segments, enhancing the encoding's expressive power. Additionally, the Memory Unit adeptly captures and retains the essential traits of both normal tissues and polyps with heightened precision and exhaustiveness, fortifying the model's robustness against the nuances of minority normal structures. Experimental outcomes on one of the most extensive and challenging colonoscopy video datasets indicate TEmory's state-of-the-art performance, showcasing a 1.48% improvement in average precision (AP) over recent advanced techniques. The code of this project is at https://github.com/Liu-Yufei/TEMory.

Keywords: Polyp Detection; Weakly Supervised Video Anomaly Detection; Weakly Supervised Learning; Temporal Encoder; Memory bank.

1 Introduction

According to the World Health Organization (WHO), colorectal cancer is the second most frequent cause of cancer [16]. The primary cause of these fatalities is the malignant transformation of colorectal polyps [14]. It is early detection of these polyps that is crucial for enhancing patient prognoses. Colonoscopy, a gold standard in the diagnosis of polyps, is pivotal in detecting precancerosis [7,9]. Recent advancements in deep learning have precipitated an influx of artificial

(a) Intestinal exsufflation (b) Intestinal insufflation (c) Appendiceal orifice (d) Ileocecal valve (e) Ileocecal orifice

Fig. 1. (a) A colonoscopic photo of exsufflation state. As a result of the reduced air pressure, the intestinal tract exhibits pronounced folds. **(b) Simultaneous insufflation state of the same intestinal segment of (a).** Most of colonoscopic video frames are as shown in (b). Differences between (a) and (b) may lead models to classify exsufflation state as abnormal and overlook polyps within exsufflation frames. **(c)-(e) Specialized structures in colonoscopy.** Appearances of specialized structures are considerably similar to polyps, presenting difficulty in distinguishing them from polyps.

intelligence applications in the analysis of medical imagery. Subject to manual endoscopic screening' high misdiagnosis rate [1,5,13,17], weakly supervised colonic polyp frame detection has emerged as a focal point of interest and active research within the medical imaging and computer vision communities.

This task aims to enable the model to achieve frame-level detection outcomes through video-level training data. Before 2022, researchers typically applied anomaly detection techniques from natural videos [11,19,18,15,4,12] to colonoscopy videos. Tian et al. [13] reformulate the task of polyp frame detection in videos into a weakly supervised video anomaly frame detection problem by constructing the Contrastive Transformer-based Multiple Instance Learning (CTMIL) model. Gao et al. [5] propose the Temporal Prototype Network (TPNet), enhancing weakly supervised polyp frame detection with a temporal encoder and a prototype-based memory bank.

As illustrated in Fig. 1, despite the successes of these methodologies in abnormal frame detection in colonoscopic videos, they fall short in comprehending the intestinal constriction induced by air exsufflation performed by colonoscopy doctors and certain distinctive normal structures within the intestinal tract. Clinically, physicians assess the presence of polyps in the current exsufflation frame by referencing an insufflation one at the identical location in the frame sequence. These specialized structures may vary in different individuals, but their relative positioning remains consistent. In summary, both exsufflation state and specialized structures make the assessment of adjacent frames particularly important.

Improving polyp detection in colonoscopy videos requires two key strategies: exploiting the inter-frame relationships and temporal characteristics to facilitate the robust extraction of information and intensifying the focus on learning paradigms that pertain to the understanding of normal structures.

Drawing from the analysis, we propose TEmory, a novel weakly supervised polyp frame detection model. Specifically, the Temporal Encoder is designed based on Bidirectional Long Short Time Memory (Bi-LSTM) [10] and Global and Local Multi-Head Self-Attention (GL-MHSA) [20], exploiting temporal informa-

tion and augmenting the descriptive power of embeddings, enabling the precise capture of information and delineation of associations between adjacent frames. Additionally, the Memory Unit, constituted by multiple parallel self-attention memory banks, not only elevates the dimensionality of the input feature space but also comprehensively stores a variety of features pertaining to specialized structures, thereby enhancing the capability for feature retrieval.

To summarize, the main contributions of this study are as follows:

- We propose TEmory, a novel weakly supervised colonic polyp frame detection model.
- We design a Temporal Encoder to fully leverage temporal information, thereby enhancing the expressiveness of feature encoding.
- We design multiple parallel self-attention memory modules to constitute the Memory Unit, enabling deeper interaction between input features and memory modules for enhanced feature retrieval and preservation.
- Extensive experiments demonstrate that our TEmory achieves new state-of-the-art performance on one of the largest and most challenging colonoscopy video datasets.

2 Method

As shown in Fig. 2, the video segment $v \in \mathbb{R}^{T\times H\times W}$ is fed into a pre-trained I3D model [3] to extract frame features. These features then pass through a Temporal Encoder and a Memory Unit, producing temporally and memory-enhanced features. The enhanced features are concatenated and processed by a linear layer to generate an anomaly score for each frame. During training, the top k anomaly scores are averaged and supervised using video-level ground truth labels. During testing, the anomaly scores for each frame are compared with a preset threshold to detect anomalies. anomalies. anomalies. anomalies.

2.1 Overall Architecture

The model takes video segment $v \in \mathbb{R}^{T\times H\times W}$ as input, which is then fed into a pretrained I3D model. Where T is the number of video frames and $W \times H$ denotes the resolution of each frame.

$$X_{I3D} = \text{I3D}\left(v\right), \tag{1}$$

where $X_{I3D} \in \mathbb{R}^{T\times D_{I3D}}$ and D_{I3D} is the dimensional length. The extracted features are processed by the Temporal Encoder to obtain temporally-enhanced features $X_{TE} \in \mathbb{R}^{T\times D}$:

$$X_{TE} = \text{Temporal Encoder}\left(X_{I3D}\right), \tag{2}$$

utilized to capture the temporal and relational dependencies within video segments, obtaining high-quality embeddings.

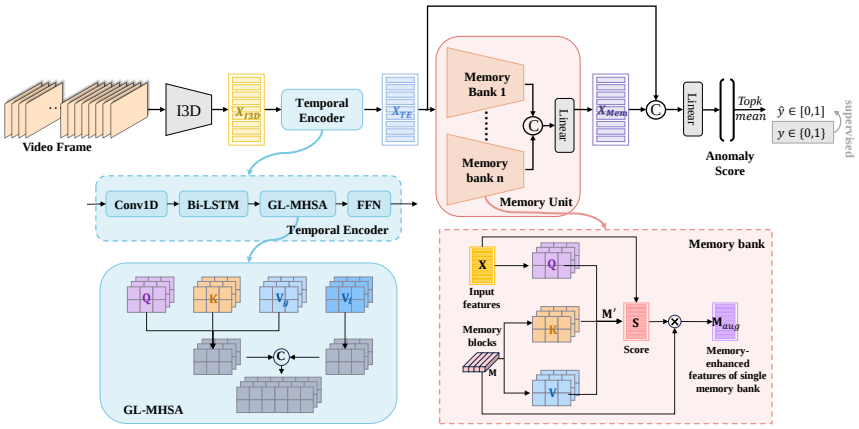


Fig. 2. The framework of our TEmemory model consists of two parts: Temporal Encoder and Memory Unit.

Then, X_{TE} is processed through multiple parallel memory banks within the Memory Unit, which use attention mechanisms to update and integrate individual features, forming memory-enhanced features $X_{Mem} \in \mathbb{R}^{T \times D}$:

$$X_{Mem} = \text{Memory Unit}\left(X_{TE}\right). \tag{3}$$

The memory-enhanced features are concatenated with the temporally-enhanced features and mapped through a Linear layer to obtain the anomaly scores:

$$y_T = \text{Linear}\left(\text{Concat}\left(X_{TE}, X_{Mem}\right)\right), \tag{4}$$

where $y_T \in \mathbb{R}^T$. Since the training utilizes video-level annotations, where $y \in \{0, 1\}$, to supervise the training with ground truth labels, we takes the mean of the top K largest anomaly scores as the output $\hat{y} \in [0, 1]$:

$$\hat{y} = \text{Mean}\left(\text{topK}\left(y_T\right)\right). \tag{5}$$

2.2 Temporal Encoder

To leverage temporal information and enhance feature encoding for capturing inter-frame associations, this paper employs a Temporal Encoder for colonoscopy video segments. After preprocessing, extracted features are fed into the Temporal Encoder, then passed into a Bi-LSTM network to learn temporal information. Using a GL-MHA mechanism, we capture each frame's relative importance. Finally, a feedforward network layer derives the temporally enhanced features.

Specifically, the input features $X_{I3D} \in \mathbb{R}^{T \times D_{I3D}}$ are subjected to temporal modeling using a Bi-LSTM as follows:

$$X_{LSTM} = \text{Bi-LSTM}\left(\text{Conv1D}\left(X_{I3D}\right)\right), \tag{6}$$

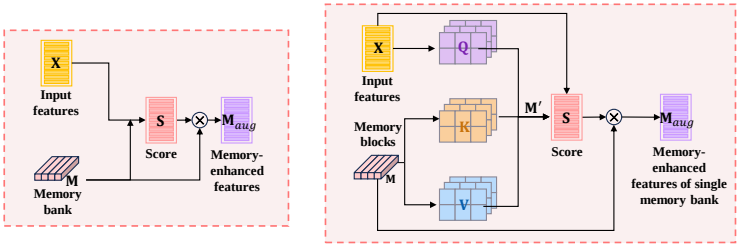


Fig. 3. Classical vs. multi-head attention updated memory bank.

where Conv1D denotes a one-dimensional convolutional layer designed to extract local features and capture spatial information within each video segment.

Then, a linear transformation is applied to X_{LSTM} to derive the matrices for queries (Q), keys (K), global values (V_g), and temporal mask values (V_l):

$$\begin{aligned} Q &= X_{LSTM} \times W^{Q}, \\ K &= X_{LSTM} \times W^{K}, \\ V_g &= X_{LSTM} \times W^{V_g}, \\ V_l &= X_{LSTM} \times W^{V_l}. \end{aligned} \tag{7}$$

Global modeling is then conducted using self-attention:

$$X_{att} = \text{Concat}\left(\text{Softmax}\left(\frac{QK^T}{\sqrt{d}}\right) V_g; \text{Softmax}\left(T_m\right) V_l\right) + X_{LSTM}, \tag{8}$$

where d represents the dimensionality of the vectors that make up Q and K, and T_m denotes the temporal mask, calculated as follows:

$$T_m = -\frac{|i-j|}{e^{\tau}}, \tag{9}$$

where τ is the sensitivity factor that balances the local distance. Finally, we further processes X_{att} using a FFN:

$$X_{TE} = \text{FFN}\left(X_{att}\right) + X_{att}, \tag{10}$$

where $X_{TE} \in \mathbb{R}^{T \times D}$, FFN denotes a linear layer following the Gaussian Error Linear Unit (GELU) [6] activation function.

2.3 Memory Unit

The exsufflation state and specialized structures exhibit varied manifestations. A single memory module cannot fully store the critical video features, as classical memory modules lack the dimensionality for deep interaction with input features, hindering effective retrieval. Therefore, we propose the Memory Unit, composed of multiple parallel memory modules updated via an attention mechanism. We

have improved the original update mechanisms of each memory module by integrating a multi-head attention mechanism, as shown in Fig. 3. Specifically, after $X_{TE} \in \mathbb{R}^{T\times D}$ passes through multiple linear layers, it transforms into the tensor required for multi-head attention; the memory module weights $\mathbf{M}$ are processed similarly:

$$\begin{aligned} \mathbf{Q} &= X_{TE} \times W^{Q}, \\ \mathbf{K} &= \mathbf{M} \times W^{K}, \\ \mathbf{V} &= \mathbf{M} \times W^{V}. \end{aligned} \tag{11}$$

The computed $\mathbf{Q}$, $\mathbf{K}$, $\mathbf{V}$ are modeled using the multi-head attention mechanism to obtain $\mathbf{M}' \in \mathbb{R}^{T\times D}$:

$$\mathbf{M}' = \text{Multi-head attention}\left(\mathbf{Q}, \mathbf{K}, \mathbf{V}\right). \tag{12}$$

Subsequently, the transpose of the memory unit weights $\mathbf{M}'$ is multiplied, normalized, and passed through a Sigmoid activation to yield the query scores $\mathbf{S} \in \mathbb{R}^{T\times T}$:

$$\mathbf{S} = \text{Sigmoid}\left(\frac{\mathbf{X}_{TE}\mathbf{M}'^{T}}{\sqrt{D}}\right), \tag{13}$$

The query scores $\mathbf{S}$ are then multiplied by the memory module weights $\mathbf{M}$ to obtain the memory-enhanced features $\mathbf{M}_{aug} \in \mathbb{R}^{T\times D}$:

$$\mathbf{M}_{aug} = \mathbf{S}\mathbf{M}, \tag{14}$$

where $\mathbf{M}_{aug} \in \mathbb{R}^{T\times D}$ is the output of the feature $\mathbf{X}_{TE}$ after entering a single memory unit.

In addition to the improvements mentioned above, as shown in Fig. 2, due to the potential limitation of a single memory module in storing features, a single memory module is expanded into multiple memory modules. Assuming the output of the i-th memory module is $\mathbf{M}_{aug_i} \in \mathbb{R}^{T\times D}$, $i \in \{1, 2, \ldots, n\}$, after all $\mathbf{M}_{aug_i}$ are computed in parallel, they are concatenated into $\mathbb{M}_{aug} \in \mathbb{R}^{T\times D\times N_{banks}}$. Finally, a linear layer with $W \in \mathbb{R}^{N_{banks}\times 1}$ (where N_{banks} is the number of memory modules) is applied to produce the final output $\mathbf{X}_{Mem} \in \mathbb{R}^{T\times D}$.

2.4 Loss Function

Given the final video scores $\hat{y} \in [0, 1]$, we employ a binary cross-entropy loss function for supervised training:

$$Loss = -\sum_{i=1}^{B}\left(y_i \log\left((\widehat{y_i}\right) + \left(1 - y_i\right)\log\left(1 - \left(\widehat{y_i}\right)\right)\right)\right), \tag{15}$$

where $y \in \{0, 1\}$ represents the labels, and B denotes the size of the batch.

Table 1. Compared with advanced methods

Method	Year	Feature	AP(%)↑	AUC(%)↑
DeelMIL [11]	2018	I3D(RGB)	68.53	89.41
GCN-Ano [19]	2019	I3D(RGB)	75.39	92.13
CLAWS [18]	2020	I3D(RGB)	80.42	95.62
AR-Net [15]	2020	I3D(RGB)	71.58	88.59
MIST [4]	2021	I3D(RGB)	72.85	94.53
RTFM [12]	2021	I3D(RGB)	77.96	96.30
CTMIL [13]	2022	I3D(RGB)	86.63	98.41
TPNet [5]	2023	I3D(RGB)	90.74	98.97
ours	**2024**	**I3D(RGB)**	**92.22**	**99.43**

3 Experiment

3.1 Experiment Settings

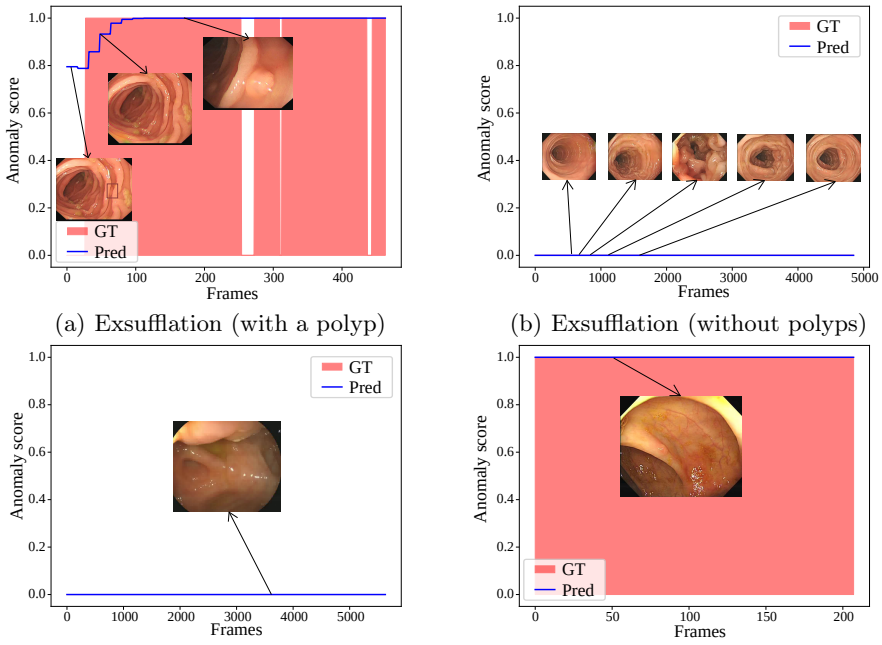
Dataset We use a combined dataset of HyperKvasir [2] and LDPolypVideo [8], totaling over one million frames with varied polyps. In alignment with the methodology of [5] and [13], the training set includes 61 normal and 102 polyp videos, while the test set has 30 normal and 60 polyp videos. Training annotations are at the video-level frames; testing uses frame-level frames.

Evaluation Metrics Consistent with the approach outlined in [5] and [13], this study employs the frame-level Area Under the Receiver Operating Characteristic curve (AUC) and AP as the metrics for evaluation.

Implement Details To ensure the fairness of the experiments, this paper adheres to the experimental setup proposed by Gao *et al.* [5] and Tian *et al.* [13], conducting training on the PyTorch platform using an NVIDIA 2080 Ti GPU. Each video is segmented into 32 video clips, and a pre-trained I3D model [3] is utilized to extract 2048-dimensional features from the mixed5c layer. The Adam optimization algorithm is employed, with a training batch size set to 16, over 2500 epochs, and a learning rate of $1e-4$, without the application of any data augmentation techniques. Additionally, the number of memory blocks within each memory module of the Memory Unit is set to 256, and the dimensionality D in X_{TE} and X_{Mem} is configured to 512.

3.2 Comparison with Advanced Methods

Quantitive Evaluation To provide a comprehensive evaluation of the performance of the model proposed in this paper, this chapter compares it with the most recent state-of-the-art weakly supervised video anomaly frame detection models [11,19,18,15,4,12], as well as the latest techniques designed specifically



(a) Exsufflation (with a polyp)

(b) Exsufflation (without polyps)

(c) Specialized structures (without polyps)

(d) Small-sized polyp morphology

Fig. 4. Anomaly score visualization: 'GT' denotes ground truth, while 'Pred' signifies the prediction conclusion.

for polyp frame detection, CTMIL [13] and TPNet [5]. In this experiment, our model maintains consistent parameter settings with the aforementioned models and conducts a fair comparison using the same features extracted from a pre-trained I3D model. As depicted in Table 1, the AP of our model is 92.22%, which represents an improvement of approximately 1.5% over TPNet. The AUC is 99.43%, marking an enhancement of about 0.46% over TPNet and nearing the 100% threshold. Surpassing the state-of-the-art methods in both AP and AUC metrics, our model substantiates the effectiveness of integrating associations between different frames and mapping data into a higher-dimensional space for storage, retrieval, and updating.

Qualitative Evaluation In Fig. 4(a) and 4(b), analyze the large intestine under exsufflation, with and without polyps, showing the model's ability to differentiate normal and abnormal frames without false positives or negatives. Fig. 4(c) illustrating the model's ability to accurately differentiate between specialized structures and polyps. Fig. 4(d) shows the model's response when polyps are of a smaller size, indicating that the proposed model has a strong capability to identify even minor polyps.

The experiments confirm the model's efficacy in distinguishing between normal and polyp structures and in reducing exsufflation-related interference in colonoscopies. The model adeptly learns specialized characteristics, enhancing its robustness in the exsufflated intestinal environment.

Table 2. Ablation experiment

Multiple memory banks	GL-MHSA	MHA-updated memory bank	Bi-LSTM	AP(%)↑	AUC(%)↑
				86.20	98.58
✓				90.29	98.88
✓	✓			91.77	99.31
✓	✓	✓		91.17	99.10
✓	✓		✓	91.60	99.22
✓	✓	✓	✓	**92.22**	**99.43**

3.3 Ablation Experiment

Ablation studies elucidate the contributions of various components in this chapter's methodology. Key insights from Table 2 include:

(1) Multiple memory modules significantly boost precision. This suggests that utilizing multiple memory modules allows for more effective storage and retrieval of information, thereby improving the model's ability to distinguish between colorectal polyps and normal tissue.

(2) GL-MHSA led to an improvement in performance, which proves the importance of acquiring and integrating global and local temporal information.

(3) Combining Bi-LSTM with multi-head attention outperforms their individual use, capturing and storing temporal patterns more effectively through enhanced guidance and broader retrieval.

(4) Compared to the baseline model, the inclusion of multiple memory modules, GL-MHSA, multi-head attention for updating memory modules, and Bi-LSTM are indispensable. This demonstrates that after the Temporal Encoder captures extensive information, the memory unit stores essential features and enables more precise and robust colorectal polyp frame detection.

3.4 Parameter Experiment

Table 3. Parameter Experiment

Memory banks	AP(%)↑	AUC(%)↑
1	91.82	99.19
2	**92.22**	**99.43**
4	91.23	99.13
8	91.28	99.14

We investigate the impact of the number of memory modules in the Memory Unit on model performance. The number of memory blocks ranged from 1 to

8, and the results are summarized in Table 3. The data indicates that with two memory modules, both AP and AUC reach their highest performance levels.

4 Conclusion

In this work, we introduce TEmory, a novel weakly supervised polyp frame detection method. Our approach employs a Temporal Encoder to capture global and interactive relationships within colonoscopy videos, thereby improving robustness by addressing inter-frame relationships and mitigating exsufflation event interference. Additionally, we propose a Memory Unit composed of multiple parallel memory modules, each updated via a multi-head attention mechanism. This Memory Unit effectively stores and retrieves salient information, enabling deeper interaction between memory modules and input features to select more representative features. This enhances the model's ability to distinguish polyps from normal structures. Extensive experiments demonstrate that our method achieves state-of-the-art performance in weakly supervised polyp frame detection.

Acknowledgments. This work is supported by the National Natural Science Foundation of China (No. 62276221, No. 62376232); the Open Project Program of Fujian Key Laboratory of Big Data Application and Intellectualization for Tea Industry, Wuyi University (No. FKLBDAITI202203).

References

1. Ahn, S.B., Han, D.S., Bae, J.H., Byun, T.J., Kim, J.P., Eun, C.S.: The miss rate for colorectal adenoma determined by quality-adjusted, back-to-back colonoscopies. Gut and liver **6**(1), 64 (2012)
2. Borgli, H., Thambawita, V., Smedsrud, P.H., Hicks, S., Jha, D., Eskeland, S.L., Randel, K.R., Pogorelov, K., Lux, M., Nguyen, D.T.D., et al.: Hyperkvasir, a comprehensive multi-class image and video dataset for gastrointestinal endoscopy. Scientific data **7**(1), 283 (2020)
3. Carreira, J., Zisserman, A.: Quo vadis, action recognition? a new model and the kinetics dataset. In: proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 6299–6308 (2017)
4. Feng, J.C., Hong, F.T., Zheng, W.S.: Mist: Multiple instance self-training framework for video anomaly detection. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 14009–14018 (2021)
5. Gao, J., Luo, Z., Tian, C., Li, S.: Tpnet: Enhancing weakly supervised polyp frame detection with temporal encoder and prototype-based memory bank. In: Chinese Conference on Pattern Recognition and Computer Vision (PRCV). pp. 470–481. Springer (2023)
6. Hendrycks, D., Gimpel, K.: Gaussian error linear units (gelus). arXiv preprint arXiv:1606.08415 (2016)

7. Itoh, H., Misawa, M., Mori, Y., Kudo, S.E., Oda, M., Mori, K.: Positive-gradient-weighted object activation mapping: visual explanation of object detector towards precise colorectal-polyp localisation. International Journal of Computer Assisted Radiology and Surgery **17**(11), 2051–2063 (2022)
8. Ma, Y., Chen, X., Cheng, K., Li, Y., Sun, B.: Ldpolypvideo benchmark: a large-scale colonoscopy video dataset of diverse polyps. In: Medical Image Computing and Computer Assisted Intervention–MICCAI 2021: 24th International Conference, Strasbourg, France, September 27–October 1, 2021, Proceedings, Part V 24. pp. 387–396. Springer (2021)
9. Macrae, F.A., Bendell, J., Tanabe, K., Savarese, D., Grover, S.: Clinical presentation, diagnosis, and staging of colorectal cancer. UpToDate Retrieved from (https://www. uptodate. com/contents/clinical-presentation-diagnosis-and-stagingof-colorectal-cancer). Accessed on **2**, 2016 (2016)
10. Shi, X., Chen, Z., Wang, H., Yeung, D.Y., Wong, W.K., Woo, W.c.: Convolutional lstm network: A machine learning approach for precipitation nowcasting. Advances in neural information processing systems **28** (2015)
11. Sultani, W., Chen, C., Shah, M.: Real-world anomaly detection in surveillance videos. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 6479–6488 (2018)
12. Tian, Y., Pang, G., Chen, Y., Singh, R., Verjans, J.W., Carneiro, G.: Weakly-supervised video anomaly detection with robust temporal feature magnitude learning. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 4975–4986 (2021)
13. Tian, Y., Pang, G., Liu, F., Liu, Y., Wang, C., Chen, Y., Verjans, J., Carneiro, G.: Contrastive transformer-based multiple instance learning for weakly supervised polyp frame detection. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 88–98. Springer (2022)
14. Tresca, A.: The stages of colon and rectal cancer. The New York times (2010)
15. Wan, B., Fang, Y., Xia, X., Mei, J.: Weakly supervised video anomaly detection via center-guided discriminative learning. In: 2020 IEEE international conference on multimedia and expo (ICME). pp. 1–6. IEEE (2020)
16. World Health Organization: Colorectal cancer fact sheet (2023-07-11), `https://www.who.int/zh/news-room/fact-sheets/detail/colorectal-cancer`
17. Xu, J., Zhao, R., Yu, Y., Zhang, Q., Bian, X., Wang, J., Ge, Z., Qian, D.: Real-time automatic polyp detection in colonoscopy using feature enhancement module and spatiotemporal similarity correlation unit. Biomedical Signal Processing and Control **66**, 102503 (2021)
18. Zaheer, M.Z., Mahmood, A., Astrid, M., Lee, S.I.: Claws: Clustering assisted weakly supervised learning with normalcy suppression for anomalous event detection. In: Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXII 16. pp. 358–376. Springer (2020)
19. Zhong, J.X., Li, N., Kong, W., Liu, S., Li, T.H., Li, G.: Graph convolutional label noise cleaner: Train a plug-and-play action classifier for anomaly detection. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 1237–1246 (2019)
20. Zhou, H., Yu, J., Yang, W.: Dual memory units with uncertainty regulation for weakly supervised video anomaly detection. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 37, pp. 3769–3777 (2023)