

YONA: You Only Need One Adjacent Reference-frame for Accurate and Fast Video Polyp Detection

Yuncheng Jiang^{1,2,4,*}, Zixun Zhang^{1,2,4} ^{*}, Ruimao Zhang³, Guanbin Li⁵,
Shuguang Cui^{1,2}, and Zhen Li^{1,2,4} [✉]

¹ SSE, The Chinese University of Hong Kong, Shenzhen

² FNii, The Chinese University of Hong Kong, Shenzhen

³ SDS, The Chinese University of Hong Kong, Shenzhen

⁴ Shenzhen Research Insitute of Big Data

⁵ School of Computer Science and Engineering, Sun Yat-sen University
yunchengjiang@link.cuhk.edu.cn, lizhen@cuhk.edu.cn

Abstract. Accurate polyp detection is essential for assisting clinical rectal cancer diagnoses. Colonoscopy videos contain richer information than still images, making them a valuable resource for deep learning methods. However, unlike common fixed-camera video, the camera-moving scene in colonoscopy videos can cause rapid video jitters, leading to unstable training for existing video detection models. In this paper, we propose the **YONA (You Only Need one Adjacent Reference-frame)** method, an efficient end-to-end training framework for video polyp detection. YONA fully exploits the information of one previous adjacent frame and conducts polyp detection on the current frame without multi-frame collaborations. Specifically, for the foreground, YONA adaptively aligns the current frame’s channel activation patterns with its adjacent reference frames according to their foreground similarity. For the background, YONA conducts background dynamic alignment guided by inter-frame difference to eliminate the invalid features produced by drastic spatial jitters. Moreover, YONA applies cross-frame contrastive learning during training, leveraging the ground truth bounding box to improve the model’s perception of polyp and background. Quantitative and qualitative experiments on three public challenging benchmarks demonstrate that our proposed YONA outperforms previous state-of-the-art competitors by a large margin in both accuracy and speed.

Keywords: Video Polyp Detection · Colonoscopy · Feature Alignment · Contrastive Learning.

1 Introduction

Colonoscopy plays a crucial role in identifying and removing early polyps and reducing mortality rates associated with rectal cancer. Over the past few years,

^{*} Equal contribution

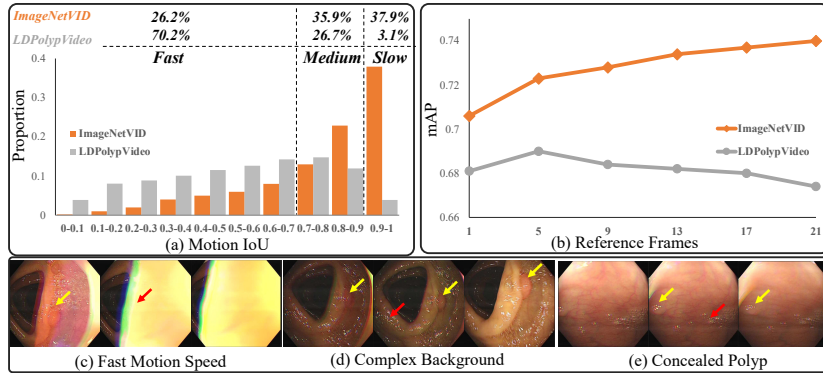


Fig. 1. (a) The histogram of the motion IoUs distribution on two datasets. Lower motion IoU denotes a faster target moving speed. The proportion of slow, medium and fast-moving targets is listed at the top of the figure. (b) The performance of FGFA [26] using multiple reference frames increases on ImageNetVID while decreasing on LDPolypVideo. (c) The typical challenges in colonoscopy videos. Yellow arrows point to the polyp, and red arrows point to distraction that causes false detection.

the research community has devoted great effort to understanding colonoscopy videos using either optical flow [23,22] or temporal information aggregation [12,16,19,5] between multiple frames.

However, those works are mainly designed based on the experience of previous natural video object detection studies, ignoring the inherent uniqueness of the colonoscopy motion patterns. Thus, we rethink the video polyp detection task and conclude three core challenges in colonoscopy videos. **1) Fast motion speed.** In Fig. 1(a), we show the target motion speed [26]⁶ on ImageNetVID [14] (natural) and LDPolypVideo [9] (colonoscopy) dataset. The motion speed in ImageNetVID evenly distributes in three intervals. In contrast, most targets in LDPolypVideo fall in the fast speed zone, leading to a large variance in the adjacent foreground features, like motion blur or occlusion, as shown in Fig. 1(c). Thus we conjecture that collaborating too many frames for polyp video detection will increase the misalignment between adjacent frames and leads to poor detection performance. Fig. 1(b) shows the performance of FGFA [26] on two datasets with increasing reference frames. The different trends of the two lines confirm our hypothesis. **2) Complex background** Different from the common camera-fixed videos, the camera-moving of colonoscopy video will introduce large disturbances between adjacent frames (e.g., specular reflection, bubbles, water, etc.), as shown in Fig. 1(d). Those abnormalities disrupt the integrity of background structures and thus affect the effect of multi-frame fusion. **3) Concealed polyps** As shown in Fig. 1(e), we noticed that some polyps could be seen as concealed objects in the colonoscopy video since such polyps have a very similar

⁶ averaged intersection-over-union scores of target in the nearby frames (± 10 frames)

appearance to the intestine wall. The model will be confused by such frames in inference and result in high false-positive or false-negative predictions.

To address the above issues, we propose the **YONA** framework, which fully exploits the reference frame information and only needs one adjacent reference frame for accurate video polyp detection. Specifically, we propose the Foreground Temporal Alignment (FTA) module to explicitly align the foreground channel activation patterns between adjacent features according to their foreground similarity. In addition, we design the Background Dynamic Alignment (BDA) module after FTA that further learns the inter-frame background spatial dynamics to better eliminate the influence of motion speed and increase the training robustness. Finally, parallel to FTA and BDA, we introduce the Cross-frame Box-assisted Contrastive Learning (CBCL) that fully utilizes the box annotations to enlarge polyp and background discrimination in embedding space.

In summary, our contributions are in three-folds: (1) To the best of our knowledge, we are the first to investigate the obstacles to the development of existing video polyp detectors and conclude that two-frame collaboration is enough for video polyp detection. (2) We propose the YONA, a novel framework for video polyp detection. It composes the foreground and background alignment modules to align the features under the fast-moving condition. It further introduces the cross-frame contrastive learning module to enhance the model’s discrimination ability of polyps and intestine walls. (3) Extensive experiments demonstrate that our YONA achieves new state-of-the-art performance on three large-scale public video polyp detection datasets.

2 Method

The whole pipeline is shown in Fig. 2. We leverage the CenterNet [25] as the base detector. Given a clip of a colonoscopy video, we take the current frame as anchor I^a and its adjacent previous frame as reference I^r . The binary maps M^a, M^r are generated using the bounding box of anchor and reference, where the foreground pixels are assigned with 1 while the background with 0. At each step, YONA first extracts multi-scale features from I^a, I^r using the backbone. Then, multi-scale features are fused and up-sampled to the resolution of the first stage as the intermediate features F^a, F^r . Then, we conduct foreground temporal alignment (Fig. 2(a)) on intermediate features to align their channel activation pattern. Next, the enhanced anchor feature \tilde{F} is further refined by the background dynamic alignment module (Fig. 2(b)) to mitigate the rapid dynamic changes in the spatial field. The BDA’s output F^* is used to compute the detection loss. Meanwhile, the intermediate features and binary maps are used to calculate the contrastive loss during training to improve the model’s perception of polyp and background (Fig. 2(c)).

Overall, the whole network is optimized with the combination loss function in an end-to-end manner. The final loss is composed of the same detection loss with CenterNet and our proposed contrastive loss, formulated as $\mathcal{L} = \mathcal{L}_{\text{detection}} + \lambda_{\text{contrast}} \mathcal{L}_{\text{contrast}}$.

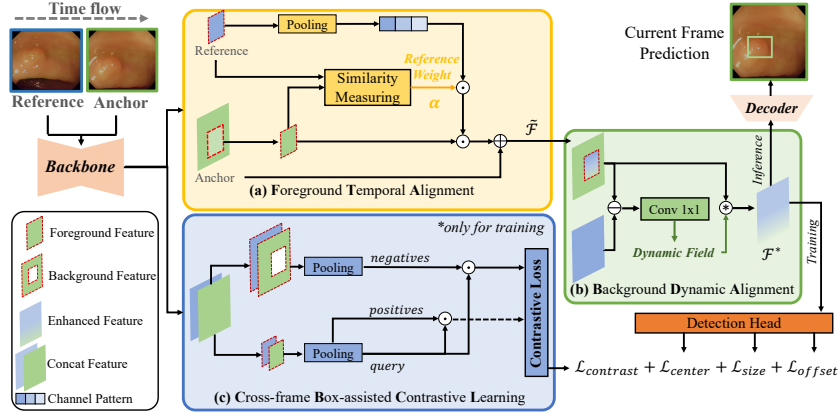


Fig. 2. Illustration of our proposed video polyp detection framework, YONA. It first aligns the foreground channel patterns between the anchor and reference frame in (a). Then it extracts polyp context guided by dynamic field in (b). Meanwhile, YONA enhances the discrimination ability via contrastive learning in (c) during training. The final output of (b) is used to predict the bounding box of the current frame.

2.1 Foreground Temporal Alignment

Since the camera moves at a high speed, the changes in the frame are very drastic for both foreground and background targets. As a result, multi-frame (reference > 3) fusion may easily incorporate more noise features into the aggregation features. On the other hand, the occluded or distorted foreground context may also influence the quality of aggregation. Thus we propose to conduct temporal alignment between adjacent features by leveraging the foreground context of only **one** adjacent reference frame. It is designed to align the certain channel's activation pattern of anchor feature to its preceding reference feature. Specifically, given the intermediate features F^a, F^r and reference binary map M^r , we first pooling F^r to 1D channel pattern f^r by the binary map on the spatial dimension ($\mathbb{R}^{N \times C \times H \times W} \rightarrow \mathbb{R}^{N \times C \times 1}$) and normalize it to $[0, 1]$:

$$f^r = \text{norm}[\text{Pooling}(F^r)]$$

$$\text{Pooling}(F^r) = \text{sum}_{HW} [F^r(x, y)] / \text{sum}[M^r(x, y)] \quad \text{if } M^r(x, y) = 1 \quad (1)$$

Then, the foreground temporal alignment is implemented by channel attention mechanism, where the attention maps are computed by weighted dot-product. We obtain the enhanced anchor feature by adding the attention maps with the original anchor feature through skip connection to keep the gradient flow.

$$\tilde{F} = [\alpha f^r \odot F^a(x, y)] \oplus F^a \quad \text{if } M^r(x, y) = 1 \quad (2)$$

where α is the adaptive weight by similarity measuring.

At the training stage, the ground truth boxes of the reference frame are used to generate the binary map M^r . During the inference stage, we conduct FTA

only if the validated bounding box of the reference frame exists, where "validated" denotes the confidence scores of detected boxes are greater than 0.6. Otherwise, we will skip this process and feed the original inputs to the next module.

Adaptive Re-weighting by Similarity Measuring As discussed above, due to video jitters, adjacent frames may change rapidly at the temporal level, and directly fusing the reference feature will introduce noisy information and misguide the training. Thus we designed an adaptive re-weighting method by measuring the feature similarity, where the weight indicates the importance of the reference feature to the anchor feature. Specifically, if the foreground feature of the reference is close to the anchor, it is assigned a larger weight at all channels. Otherwise, a smaller weight is assigned. For efficiency, we use the cosine similarity metric [8] to measure the similarity, where f^a is the 1D channel pattern of F^a computed with Eq. 1:

$$\alpha = \exp\left(\frac{f^r \cdot f^a}{\|f^r\| \|f^a\|}\right) \quad (3)$$

2.2 Background Dynamic Alignment

The traditional convolutional-based object detector can detect objects well when the background is stable. However, once it receives obvious interference, such as light or shadow, the background changes may cause the degradation of spatial correlation and lead to many false-positive predictions. Motivated by the inter-frame difference method [20], we first mine the dynamic field of adjacent background contents, then consult to deformable convolution [3] to learn the inherent geometric transformations according to the intensity of the dynamic field. In practice, given the enhanced anchor feature $\tilde{\mathcal{F}}$ from FTA and reference feature F^r , the inter-frame difference is defined as the element-wise subtraction of enhanced anchor and reference feature. Then a 1×1 convolution is applied on the difference to generate dynamic field \mathcal{D} , which encodes all spatial dynamic changes between adjacent frames.

$$\mathcal{D} = \text{Conv}_{1 \times 1}(\tilde{\mathcal{F}} - F^r) \quad (4)$$

Finally, a 3×3 deformable convolution embeds the spatial dynamic changes of \mathcal{D} on the enhanced anchor feature $\tilde{\mathcal{F}}$.

$$\mathcal{F}^* = \text{DeConv}_{3 \times 3}(\tilde{\mathcal{F}}, \mathcal{D}) \quad (5)$$

where \mathcal{D} works as the deformable offset and \mathcal{F}^* is the final aligned anchor feature. Then the enhanced anchor feature is fed into three detection heads composed of a 3×3 Conv and a 1×1 Conv to produce center, size, and offset features for detection loss:

$$\mathcal{L}_{\text{detection}} = \mathcal{L}_{\text{focal}}^{\text{center}} + \lambda_{\text{size}} \mathcal{L}_{\text{L1}}^{\text{size}} + \lambda_{\text{offset}} \mathcal{L}_{\text{L1}}^{\text{offset}} \quad (6)$$

where $\mathcal{L}_{\text{focal}}$ is focal loss and \mathcal{L}_{L1} is L1 loss.

2.3 Cross-frame Box-assisted Contrastive Learning

Typically, in colonoscopy videos, some concealed polyps appear very similar to the intestine wall in color and texture. Thus, an advanced training strategy is required to distinguish such homogeneity. Inspired by recent studies on supervised contrastive learning [18], we select the foreground and background region on both two frames guided by ground truth boxes to conduct contrastive learning. In practice, Given a batch of intermediate feature maps $F^a, F^r \in \mathbb{R}^{N \times T \times C \times H \times W}$ and corresponding binary maps $M^a, M^r \in \mathbb{R}^{N \times T \times H \times W}$, we first concatenate the anchor and reference at the batch-wise level as $\hat{F} \in \mathbb{R}^{NT \times C \times H \times W}$ and $\hat{M} \in \mathbb{R}^{NT \times H \times W}$ to exploit the cross-frame information. Then we extract the foreground and background channel patterns of cross-frame feature \hat{F} using the Eq. 1 base on $\hat{M}(x, y) = 1$ and $\hat{M}(x, y) = 0$, respectively. After that, for each foreground channel pattern, which is the "query", we randomly select another different foreground feature as the "positive", while all the background features in the same batch are taken as the "negatives". Finally, we calculate the one-step contrastive loss by InfoNCE [18]:

$$\mathcal{L}_j^{\text{NCE}} = -\log \frac{\exp(q_j \cdot i^+ / \tau)}{\exp(q_j \cdot i^+ / \tau) + \sum_{i^- \in \mathcal{N}_j} \exp(q_j \cdot i^- / \tau)} \quad (7)$$

where $q_j \in \mathbb{R}^C, j = 0, \dots, NT$ is the query feature, $i^+ \in \mathbb{R}^C$ and $i^- \in \mathbb{R}^{NT \times C}$ are positives and negatives. \mathcal{N}_j denote embedding collections of the negatives. We repeat this process until every foreground channel pattern is selected and sum all steps as the final contrastive loss:

$$\mathcal{L}_{\text{contrast}} = \frac{1}{NT} \sum_{j=1}^{NT} \mathcal{L}_j^{\text{NCE}} \quad (8)$$

3 Experiments

We evaluate the proposed method on three public video polyp detection benchmarks: SUN Colonoscopy Video Database [10,7] (train set: 19,544 frames, test set: 12,522 frames), LDPolypVideo [9] (train set: 20,942 frames, test set: 12,933 frames), and CVC-VideoClinicDB [1] (train set: 7995 frames, test set: 2030 frames). For the fairness of the experiments, we keep the same dataset settings for YONA and all other methods.

We use ResNet-50 [6] as our backbone and CenterNet [25] as our base detector. Following the same setting in CenterNet, we set $\lambda_{\text{size}} = 0.1$ and $\lambda_{\text{off}} = 1$. We set $\lambda_{\text{contrast}} = 0.3$ by ablation study. Detailed results are listed in the supplement. We randomly crop and resize the images to 512×512 and normalize them using ImageNet settings. Random rotation and flip with probability $p = 0.5$ are used for data augmentation. We set the batch size $N = 32$. Our model is trained using the Adam optimizer with a weight decay of 5×10^{-4} for 64 epochs. The initial learning rate is set to 10^{-4} and gradually decays to 10^{-5} with cosine annealing. All models are trained with PyTorch [11] framework. The training setting of other competitors follows the best settings given in their paper.

Table 1. Performance comparison with other image/video-based detection models. P, R, and F1 denote the precision, recall, and F1-score. †: results from the original paper with the same data division. The best score is marked as **red**, while the second best score is marked as **blue**.

Methods	SUN Database			LDPolypVideo			CVC-VideoClinic			FPS
	P	R	F1	P	R	F1	P	R	F1	
Faster-RCNN [13]	77.2	69.6	73.2	68.8	46.7	55.6	84.6	98.2	90.9	44.7
FCOS [17]	75.7	64.1	69.4	65.1	46.0	53.9	92.1	74.1	82.1	42.0
CenterNet [25]	74.6	65.4	69.7	70.6	43.8	54.0	92.0	80.5	85.9	51.5
Sparse-RCNN [15]	75.5	73.7	74.6	71.6	47.9	57.4	85.1	96.4	90.4	40.0
DINO [21]	81.5	72.3	76.6	68.3	51.1	58.4	93.1	89.3	91.2	23.0
FGFA [26]	78.9	70.4	74.4	68.8	48.9	57.2	94.5	89.2	91.7	1.8
OptCNN† [23]	-	-	-	-	-	-	84.6	97.3	90.5	-
AIDPT† [22]	-	-	-	-	-	-	90.6	84.5	87.5	-
MEGA [2]	80.4	71.6	75.7	69.2	50.1	58.1	91.6	87.7	89.6	8.1
TransVOD [24]	79.3	69.6	74.1	69.2	49.2	57.5	92.1	91.4	91.7	8.4
STFT [19]	81.5	72.4	76.7	72.1	50.4	59.3	91.9	92.0	92.0	12.5
Ours-YONA	83.3	74.9	78.9	75.4	53.1	62.3	92.8	93.8	93.3	46.3

3.1 Quantitative and Qualitative Comparison

Quantitative Comparison The comparison results are shown in Tab. 1. Following the standard of [1], the Precision, Recall, and F1-scores are used for evaluation. Firstly, compared with the CenterNet baseline, our YONA with three novel designs significantly improved the F1 score by 9.2%, 8.3%, and 7.4% on three benchmarks, demonstrating the effectiveness of the model design. Besides, YONA achieves the best trade-off between accuracy and speed compared with all other image-based SOTAs across all datasets. Second, for video-based competitors, previous video object detectors with multiple frame collaborations lack the ability for accurate detection on challenging datasets. Specifically, YONA surpasses the second-best STFT [19] by 2.2%, 3.0%, and 1.3% on F1 score on three datasets and 33.8 on FPS. All the results confirm the superiority of our proposed framework for accurate and fast video polyp detection.

Qualitative Comparison Fig. 3 visualizes the qualitative results of YONA with other competitors [25,19]. Thanks to this one-adjacent-frame framework, our YONA can not only prevent the false positive caused by part occlusion (1st and 2nd clips) but also capture useful information under severe image quality (2nd clip). Moreover, our YONA shows robust performance even for challenging scenarios like concealed polyps (3rd clip).

3.2 Ablation Study

We investigated the effectiveness of each component in YONA on the SUN database, as shown in Tab. 2. It can be observed that all the modules are necessary for precise detection compared with the baseline results. Due to the large

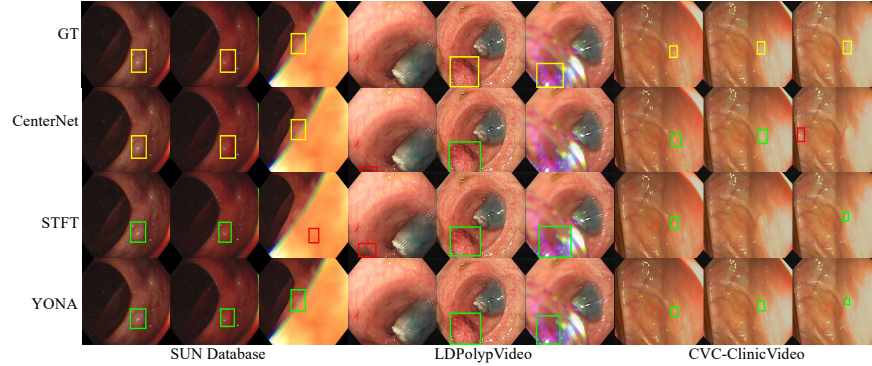


Fig. 3. Qualitative results of polyp detection on some video clips. The yellow, green, and red denote the ground truth, true positive, and false positive, respectively

Table 2. Ablation studies of YONA under different settings. Ada means the adaptive re-weighting by similarity measuring; CW denotes the channel-wise attention [4]; CA denotes the channel-aware attention [19].

FTA	CW [4]	CA [19]	Ada	BDA	CBCL	Precision	Recall	F1	FPS
						74.6	65.4	69.7	51.5
✓						74.0	63.9	68.6 _{↓1.1}	49.7
✓			✓			80.9	70.1	75.1 _{↑5.4}	48.5
	✓		✓			78.0	65.2	71.1 _{↑1.4}	48.3
		✓	✓			80.4	68.4	73.9 _{↑4.2}	45.2
✓			✓	✓		82.0	72.2	76.8 _{↑7.1}	46.3
✓			✓	✓	✓	83.3	74.9	78.9_{↑9.2}	46.3

variance of colonoscopy image content, the F1 score slightly decreases if directly adding FTA without the adaptive re-weighting strategy. Adding the adaptive weight greatly improves the F1 score by 5.4. Moreover, we use other two mainstream channel attention mechanisms to replace our proposed FTA for comparison. Compared with them, our FTA with adaptive weighting achieves the largest gain over the baseline and higher FPS. Overall, by combining all the proposed methods, our model can achieve new state-of-the-art performance.

4 Conclusion

Video polyp detection is a currently challenging task due to the fast-moving property of colonoscopy video. In this paper, We proposed the YONA framework that requires only one adjacent reference frame for accurate and fast video polyp detection. To address the problem of fast-moving polyps, we introduced the foreground temporal alignment module, which explicitly aligns the channel patterns of two frames according to their foreground similarity. For the complex

background content, we designed the background dynamic alignment module to mitigate the large variances by exploiting the inter-frame difference. Meanwhile, we employed a cross-frame box-assisted contrastive learning module to enhance the polyp and background discrimination based on box annotations. Extensive experiment results confirmed the effectiveness of our method, demonstrating the potential for practical use in real clinical applications.

Acknowledgements

This work was supported in part by Shenzhen General Program No. JCYJ20220530143600001, by the Basic Research Project No. HZQB-KCZY2021067 of Hetao Shenzhen HK S&T Cooperation Zone, by Shenzhen-Hong Kong Joint Funding No. SGDX20211123112401002, NSFC with Grant No. 62293482, by Shenzhen Outstanding Talents Training Fund, by Guangdong Research Project No. 2017ZT07X152 and No. 2019CX01X104, by the Guangdong Provincial Key Laboratory of Future Networks of Intelligence (Grant No. 2022B1212010001), by the Guangdong Provincial Key Laboratory of Big Data Computing, The Chinese University of Hong Kong, Shenzhen, by the NSFC 61931024&81922046, by the Shenzhen Key Laboratory of Big Data and Artificial Intelligence (Grant No. ZDSYS201707251409055), and the Key Area R&D Program of Guangdong Province with grant No. 2018B030338001, by zelixir biotechnology company Fund, by Tencent Open Fund.

References

1. Bernal, J.J., Histace, A., Masana, M., Angermann, Q., Sánchez-Montes, C., Rodríguez, C., Hammami, M., Garcia-Rodriguez, A., Córdova, H., Romain, O., et al.: Polyp detection benchmark in colonoscopy videos using gtcreator: A novel fully configurable tool for easy and fast annotation of image databases. In: Proceedings of 32nd CARS conference (2018)
2. Chen, Y., Cao, Y., Hu, H., Wang, L.: Memory enhanced global-local aggregation for video object detection. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 10337–10346 (2020)
3. Dai, J., Qi, H., Xiong, Y., Li, Y., Zhang, G., Hu, H., Wei, Y.: Deformable convolutional networks. In: Proceedings of the IEEE international conference on computer vision. pp. 764–773 (2017)
4. Fu, J., Liu, J., Tian, H., Li, Y., Bao, Y., Fang, Z., Lu, H.: Dual attention network for scene segmentation. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 3146–3154 (2019)
5. González-Bueno Puyal, J., Brandao, P., Ahmad, O.F., Bhatia, K.K., Toth, D., Kader, R., Lovat, L., Mountney, P., Stoyanov, D.: Polyp detection on video colonoscopy using a hybrid 2d/3d cnn. *Medical Image Analysis* **82**, 102625 (Nov 2022)
6. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 770–778 (2016)

7. Itoh, H., Misawa, M., Mori, Y., Oda, M., Kudo, S.E., Mori, K.: Sun colonoscopy video database (2020), <http://amed8k.sundatabase.org/>
8. Luo, C., Zhan, J., Xue, X., Wang, L., Ren, R., Yang, Q.: Cosine normalization: Using cosine similarity instead of dot product in neural networks. In: Artificial Neural Networks and Machine Learning–ICANN 2018: 27th International Conference on Artificial Neural Networks, Rhodes, Greece, October 4–7, 2018, Proceedings, Part I 27. pp. 382–391. Springer (2018)
9. 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)
10. Misawa, M., Kudo, S.e., Mori, Y., Hotta, K., Ohtsuka, K., Matsuda, T., Saito, S., Kudo, T., Baba, T., Ishida, F., et al.: Development of a computer-aided detection system for colonoscopy and a publicly accessible large colonoscopy video database (with video). *Gastrointestinal endoscopy* **93**(4), 960–967 (2021)
11. Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., Lin, Z., Gimelshein, N., Antiga, L., Desmaison, A., Kopf, A., Yang, E.Z., DeVito, Z., Raison, M., Tejani, A., Chilamkurthy, S., Steiner, B., Fang, L., Bai, J., Chintala, S.: Pytorch: An imperative style, high-performance deep learning library. *Neural Information Processing Systems* (2019)
12. Qadir, H.A., Balasingham, I., Solhusvik, J., Bergsland, J., Aabakken, L., Shin, Y.: Improving automatic polyp detection using cnn by exploiting temporal dependency in colonoscopy video. *IEEE journal of biomedical and health informatics* **24**(1), 180–193 (2019)
13. Ren, S., He, K., Girshick, R., Sun, J.: Faster r-cnn: Towards real-time object detection with region proposal networks. *Advances in neural information processing systems* **28** (2015)
14. Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z., Karpathy, A., Khosla, A., Bernstein, M., Berg, A.C., Fei-Fei, L.: ImageNet Large Scale Visual Recognition Challenge. *International Journal of Computer Vision (IJCV)* **115**(3), 211–252 (2015)
15. Sun, P., Zhang, R., Jiang, Y., Kong, T., Xu, C., Zhan, W., Tomizuka, M., Li, L., Yuan, Z., Wang, C., et al.: Sparse r-cnn: End-to-end object detection with learnable proposals. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 14454–14463 (2021)
16. Tajbakhsh, N., Gurudu, S.R., Liang, J.: Automatic polyp detection in colonoscopy videos using an ensemble of convolutional neural networks. In: 2015 IEEE 12th International Symposium on Biomedical Imaging (ISBI). pp. 79–83. IEEE (2015)
17. Tian, Z., Shen, C., Chen, H., He, T.: Fcos: Fully convolutional one-stage object detection. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 9627–9636 (2019)
18. Wang, W., Zhou, T., Yu, F., Dai, J., Konukoglu, E., Van Gool, L.: Exploring cross-image pixel contrast for semantic segmentation. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 7303–7313 (2021)
19. Wu, L., Hu, Z., Ji, Y., Luo, P., Zhang, S.: Multi-frame collaboration for effective endoscopic video polyp detection via spatial-temporal feature transformation. 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. 302–312. Springer (2021)

20. Zhan, C., Duan, X., Xu, S., Song, Z., Luo, M.: An improved moving object detection algorithm based on frame difference and edge detection. In: Fourth International Conference on Image and Graphics (ICIG 2007). pp. 519–523 (2007)
21. Zhang, H., Li, F., Liu, S., Zhang, L., Su, H., Zhu, J., Ni, L., Shum, H.Y.: Dino: Detr with improved denoising anchor boxes for end-to-end object detection. In: The Eleventh International Conference on Learning Representations (2022)
22. Zhang, Z., Shang, H., Zheng, H., Wang, X., Wang, J., Sun, Z., Huang, J., Yao, J.: Asynchronous in parallel detection and tracking (aipdt): Real-time robust polyp detection. In: Medical Image Computing and Computer Assisted Intervention–MICCAI 2020: 23rd International Conference, Lima, Peru, October 4–8, 2020, Proceedings, Part III 23. pp. 722–731. Springer (2020)
23. Zheng, H., Chen, H., Huang, J., Li, X., Han, X., Yao, J.: Polyp tracking in video colonoscopy using optical flow with an on-the-fly trained cnn. In: 2019 IEEE 16th International Symposium on Biomedical Imaging (ISBI 2019). pp. 79–82. IEEE (2019)
24. Zhou, Q., Li, X., He, L., Yang, Y., Cheng, G., Tong, Y., Ma, L., Tao, D.: Transvod: end-to-end video object detection with spatial-temporal transformers. *IEEE Transactions on Pattern Analysis and Machine Intelligence* (2022)
25. Zhou, X., Wang, D., Krähenbühl, P.: Objects as points. *arXiv preprint arXiv:1904.07850* (2019)
26. Zhu, X., Wang, Y., Dai, J., Yuan, L., Wei, Y.: Flow-guided feature aggregation for video object detection. In: Proceedings of the IEEE international conference on computer vision. pp. 408–417 (2017)

YONA: You Only Need One Adjacent Reference-frame for Accurate and Fast Video Polyp Detection Supplementary Material

Yuncheng Jiang^{1,2,4,*}, Zixun Zhang^{1,2,4*}, Ruimao Zhang³, Guanbin Li⁵,
Shuguang Cui^{1,2}, and Zhen Li^{1,2,4}✉

¹ SSE, The Chinese University of Hong Kong, Shenzhen

² FNii, The Chinese University of Hong Kong, Shenzhen

³ SDS, The Chinese University of Hong Kong, Shenzhen

⁴ Shenzhen Research Institute of Big Data

⁵ School of Computer Science and Engineering, Sun Yat-sen University
lizhen@cuhk.edu.cn

Table 1. Impact of $\lambda_{\text{contrast}}$ on F1-score.

$\lambda_{\text{contrast}}$	SUN Database	LDPolypVideo	CVC-VideoClinic
0.1	78	61.7	92.6
0.2	78.5	62.3	93.4
0.3	78.9	62.3	93.4
0.4	78.6	62.1	93.3
0.5	78.2	62	93.1
0.6	78	61.8	93
0.7	78.1	61.5	92.6
0.8	78.3	61.2	92.4
0.9	77.9	61.6	92.6
1	77.7	61.7	92.8

Table 2. Impact of the reference frame number on accuracy (F1-score) and inference speed (FPS).

Dataset	# Reference Frames						
	0	1*	2	4	6	8	10
SUN Database	69.7	78.9	78.1	76.4	75.6	73.7	73.0
LDPolypVideo	54.0	62.3	61.1	59.7	56.6	55.1	54.5
CVC-VideoClinic	85.9	93.3	92.6	91.2	89.8	87.5	86.4
FPS	51.5	46.3	33.6	26.5	18.9	13.4	10.8

* Equal contribution

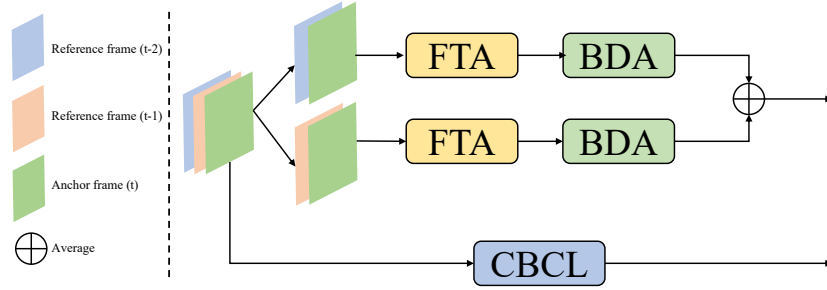


Fig. 1. Illustration of YONA framework with multiple reference frames. For simplicity, here we take two reference frames as an example. It’s noted that the design of multi-frame architecture is not the main concern and contribution of our work. Thus, we just adopt the naive average fusion strategy.

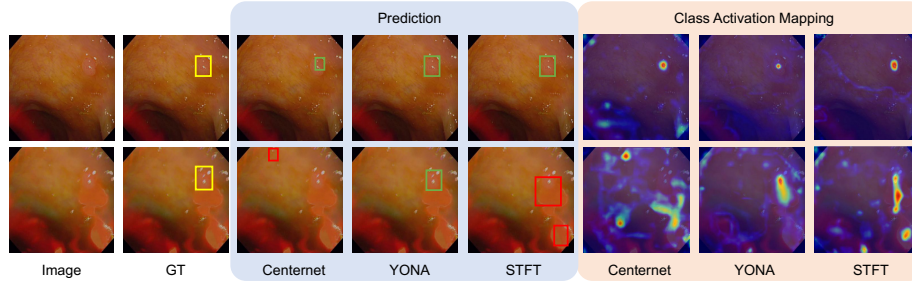


Fig. 2. Illustration of attention regions and predictions of different methods on two adjacent frames. Image-level detectors lack the ability to extract useful context information on blurred images, leading to false positive results. Thanks to the proposed foreground and background alignments, our YONA can fuse the context information from the adjacent frame and obtain valid predictions.