



This MICCAI paper is the Open Access version, provided by the MICCAI Society. It is identical to the accepted version, except for the format and this watermark; the final published version is available on SpringerLink.

# PolypSegTrack: Unified Foundation Model for Colonoscopy Video Analysis

Anwesa Choudhuri, Zhongpai Gao, Meng Zheng, Benjamin Planche, Terrence Chen, Ziyang Wu

United Imaging Intelligence, Boston, MA, USA  
`first.last@uii-ai.com`

**Abstract.** Early detection, accurate segmentation, classification and tracking of polyps during colonoscopy are critical for preventing colorectal cancer. Many existing deep-learning-based methods for analyzing colonoscopic videos either require task-specific fine-tuning, lack tracking capabilities, or rely on domain-specific pre-training. In this paper, we introduce *PolypSegTrack*, a novel foundation model that jointly addresses polyp detection, segmentation, classification and unsupervised tracking in colonoscopic videos. Our approach leverages a novel conditional mask loss, enabling flexible training across datasets with either pixel-level segmentation masks or bounding box annotations, allowing us to bypass task-specific fine-tuning. Our unsupervised tracking module reliably associates polyp instances across frames using object queries, without relying on any heuristics. We leverage a robust vision foundation model backbone that is pre-trained unsupervisedly on natural images, thereby removing the need for domain-specific pre-training. Extensive experiments on multiple polyp benchmarks demonstrate that our method significantly outperforms existing state-of-the-art approaches in detection, segmentation, classification, and tracking.

**Keywords:** Polyp Detection · Polyp Segmentation · Polyp Tracking.

## 1 Introduction

Early diagnosis of polyps in the gastrointestinal (GI) tract through colonoscopy is vital for preventing colorectal cancer. Automating the detection, segmentation, classification, and tracking of polyps in colonoscopic videos can greatly enhance the speed, accuracy, and consistency of polyp diagnosis. In recent years, deep learning methods [42, 10, 36, 11] has achieved remarkable progress in medical image analysis tasks such as segmentation and object detection. In particular, foundation models [34, 32]—pre-trained on large-scale, diverse datasets and then fine-tuned for specific tasks—have emerged as a promising direction for robust visual representation learning. Recent studies [28] have shown that self-supervised learning in these models can yield general-purpose features that transfer well to downstream tasks. Several works [12, 40] have focused on developing foundation models for colonoscopic video analysis, specifically for polyp detection and segmentation.

The aforementioned methods suffer from some drawbacks. The foundation models need to be fine-tuned separately on the different downstream tasks like detection and segmentation. However, treating these tasks independently neglects the inherent synergies between these tasks and limits the amount of data available for fine-tuning. Some recent works in computer vision [21, 41] have incorporated these synergies, but such synergies for colonoscopic video analysis is still lacking. Furthermore, the lack of large-scale video datasets with temporally dense segmentations has limited the development of effective tracking models for polyps, even though polyp tracking can be valuable for clinicians to generate exam/operation reports with accurate numbers and consistent labels of polyps. Additionally, the pre-training phase of the colonoscopic foundation models [40, 12] relies on domain-specific data, i.e., colonoscopic videos, which can be expensive to collect.

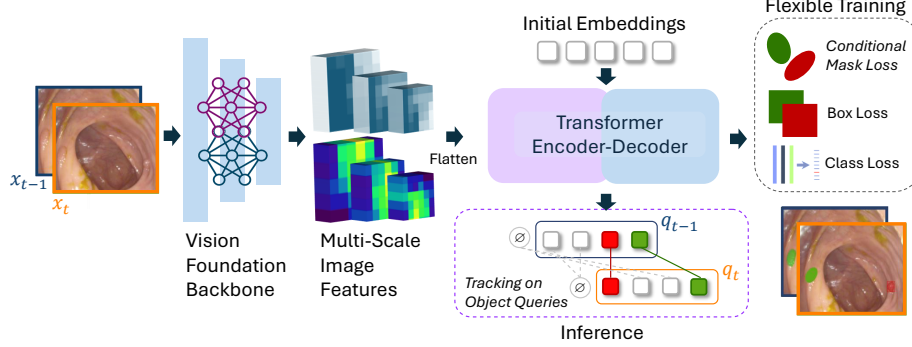
In this paper, we propose **PolypSegTrack**, a novel foundation model for polyp detection, segmentation, and unsupervised tracking in colonoscopic videos. Our multi-task learning framework prevents over-optimization on single tasks and improves generalization by exploiting different task commonalities, which is key to develop a robust foundation model. Our novel conditional mask loss allows us to exploit the interdependencies between detection and segmentation tasks, and train our model in a flexible manner adapting to different annotation types. We develop an unsupervised and non-heuristic tracking approach that uses object queries to assign track identities to polyps. Our model is initially pre-trained on natural images in an unsupervised manner, which reduces the reliance on large-scale, expensive domain-specific colonoscopic data.

We evaluate our model on a wide range of tasks: joint detection and segmentation on the ETIS, CVC-ColonDB, CVC-300, Kvasir-SEG and the CVC-Clinic-DB datasets; detection and classification on the KUMC dataset; and joint detection and tracking on a subset of the REAL-Colon dataset. In all the aforementioned tasks, our model achieves the state-of-the-art results.

## 2 Method

Fig. 1 shows the overview of our approach. Given a video consisting of  $T$  frames, the goal is to generate bounding boxes, segmentation masks, class probabilities, and track identities for every polyp in each video frame. Our model generates predictions in 2 stages: In stage 1, it produces bounding boxes, segmentation masks, and class probabilities of objects in each frame; and in stage 2, the objects are matched between every two consecutive frames to perform tracking.

Sec. 2.1 describes stage 1 of our approach, i.e., joint detection, segmentation and classification in each frame. In Sec. 2.2, we describe our novel conditional mask loss used to fine-tune our model jointly on training data containing segmentation masks, as well as on training data where only bounding box annotations are available. In Sec. 2.3, we describe stage 2 of our two-stage process, i.e., our non-heuristic and unsupervised tracking method during inference.



**Fig. 1.** Overview of the proposed approach. Our proposed conditional mask loss (Sec. 2.2) allows flexible training. Our unsupervised and non-heuristic tracking on object queries (Sec. 2.3) allows effective association of polyps across video frames.

## 2.1 Joint Detection, Segmentation, and Classification

To perform joint detection, segmentation and classification (stage 1 of our 2 stage process), our model consists of the following main components: a) a vision foundation model backbone to extract meaningful image features from every video frame, b) a transformer encoder and decoder to produce abstract object proposals or object queries, and c) prediction heads to produce the bounding boxes, segmentation masks and class probabilities using the object queries for each object.

**Vision foundation model backbone.** Given a video frame  $x_t \in \mathbb{R}^{H \times W}$ , the vision foundation model backbone extracts meaningful multi-scale image features from the frame (as shown in Fig. 1). Here  $H$  and  $W$  refer to the height and width of the video frame. We use a pre-trained DINOv2 [28] backbone for the feature extraction, followed by four multi-layer perceptrons (MLPs) to generate features at multiple scales ( $1/32^{\text{th}}$ ,  $1/16^{\text{th}}$ ,  $1/8^{\text{th}}$  and  $1/4^{\text{th}}$ ). DINOv2 [28] is pre-trained on natural images using self-supervised learning and has shown to generate general-purpose features that transfer well to downstream tasks. Note that the lack of supervision during pre-training is important to capture visual features which translates well to other domains, like colonoscopic videos.

**Transformer encoder-decoder.** The transformer encoder-decoder accepts a set of trainable initial embeddings  $e \in \mathbb{R}^{N \times C}$  and the flattened image features as inputs to produce  $N$  object queries  $q_t \in \mathbb{R}^{N \times C}$  for video frame  $x_t$  (as shown in Fig. 1). The object queries are abstract representations for all objects in the current frame.  $N$  refers to the maximum number of objects to be discovered in the current frame and  $C$  refers to the number of channels. We use MaskDINO (DETR with Improved Denoising Anchor Boxes) [21] as our choice of encoder-decoder. MaskDINO is trained on natural-image datasets where both segmentation masks and bounding boxes are available, so the encoder-decoder captures the synergies between the detection and segmentation tasks. However, note

that MaskDINO is not extensively trained on data where segmentation masks are lacking, which is a common case for our setting, i.e., for colonoscopic videos. **Prediction Heads.** There are three prediction heads: box head, classification head, and mask head, which are 3 layered, 1 layered and 3 layered fully connected networks respectively (omitted in Fig. 1 for clarity). The box head and the classification head accept the object queries and produce the bounding boxes and the class probabilities of the corresponding objects. The mask head accepts the object queries and generates intermediate queries, which are vectors that are the same size as the object queries. These intermediate queries are then multiplied with the image features and thresholded to produce the segmentation masks for individual objects. This design is following [11]. Note that the object queries are expressive and contain enough information about the respective objects that they represent such that when they are passed through the respective heads, they are able to produce the desired bounding boxes, class probabilities, and segmentation masks for the objects they represent.

## 2.2 Training with Conditional Mask Loss

During training, our model can be conditioned to learn from either segmentation mask and bounding box annotations, or only from bounding box annotations if the segmentation masks aren't available in the dataset. This flexibility allows us to train jointly from a wide range of datasets containing either types of annotations. Datasets with segmentation-based annotations offer fine-grained pixel-level localization for the model, but large-scale segmentation datasets are lacking for colonoscopic videos. Datasets with bounding box-based annotations, on the other hand, are available more commonly, even though, they offer only course-grained localization in the form of 4 points. To leverage learning from both kinds of datasets, we design a conditional mask loss which is activated only when segmentation annotations are available.

Specifically, the conditional mask loss  $\mathcal{L}_{\text{cond-mask}}$  is the combination of dice loss  $L_{\text{dice}}$  and mask-based cross entropy loss  $L_{\text{mask}}$  whenever segmentation annotations are present. This loss is 0 when only bounding box annotations are present. Formally, let  $s_{\text{gt}}^i$  refer to the segmentation annotations for a given ground truth object  $i$ . If the segmentation annotations aren't present,  $s_{\text{gt}}^i = \emptyset$ . Let  $K$  represent the number of ground truth objects in the current image. Following DETR [8], we first match the ground truth objects with the predicted object queries (some object queries remain unmatched, since  $N > K$ ). Let  $\sigma_i$  represent a match between an object query with the ground truth object  $i$ . Then the conditional mask loss  $\mathcal{L}_{\text{cond-mask}}$  for the given image is defined as follows.

$$\mathcal{L}_{\text{cond-mask}} = \sum_{i=1}^K \mathbb{I}_{\{s_{\text{gt}}^i \neq \emptyset\}} [L_{\text{mask}}(\sigma_i) + L_{\text{dice}}(\sigma_i)] \quad (1)$$

Our overall training objective is to minimize the a total multi-task loss function  $\mathcal{L}$ .

$$\mathcal{L} = \alpha_c \mathcal{L}_{\text{cls}} + \alpha_b \mathcal{L}_{\text{bbox}} + \alpha_m \mathcal{L}_{\text{cond-mask}} \quad (2)$$

Here,  $\mathcal{L}_{\text{cls}}$  refers to the cross entropy loss for class prediction and  $\mathcal{L}_{\text{bbox}}$  is a combination of  $L1$  loss and the generalized IoU loss calculated for the matched predicted objects in the current image and the ground truth following DETR [8].  $\alpha_c$ ,  $\alpha_b$  and  $\alpha_m$  are scaling factors to balance the loss terms. They are set to 1, 1 and 10 in all our experiments. The conditional mask loss is scaled  $10\times$  higher than the bounding box loss to address the training data imbalance, where data points with only bounding boxes are  $3.5\times$  more numerous than those with both bounding boxes and segmentation masks.

### 2.3 Unsupervised Tracking in the Space of Object Queries

In stage 2 of our two-stage approach, we temporally associate object instances between frames for tracking during inference (shown in Fig. 1). In prior works on tracking [6], this step often involves heuristics like computing mask overlap, which may not generalize well in case of large camera motion or occlusions, both of which are common for colonoscopic videos. Heuristic IoU-based tracking calculates overlap between segmentation masks or bounding boxes of objects across frames in the pixel space. Polyps with high overlap in consecutive frames share the same identity. To avoid heuristic post-processing, we match the object queries in the query space, following MinVIS [17], which shows that object queries from image-based models are consistent across frames, enabling lightweight tracking without any temporal training. Specifically, given two consecutive video frames  $x_{t-1}$  and  $x_t$ , we obtain the set of object queries  $q_{t-1}$  and  $q_t$ , as described in Sec. 2.1. We perform tracking by using the Hungarian matching algorithm on a cost matrix  $M \in \mathbb{R}^{N \times N}$ , where every element is  $M^{i,j} = S_{\text{cosine}}(q_{t-1}^i, q_t^j)$ . Here,  $S_{\text{cosine}}(q_{t-1}^i, q_t^j)$  represents the cosine similarity between the  $i^{\text{th}}$  element in  $q_{t-1}$  and  $j^{\text{th}}$  element in  $q_t$ . Since the appearance of objects change gradually in a video, the object queries representing the same object only change slightly in consecutive frames, leading to a high similarity between these queries.

This approach of per-frame matching in the query-space is less affected by occlusions, as compared to directly matching masks or bounding boxes in the pixel space, because the object queries are not directly tied to the spacial positions of objects in each frame. Further, we do not need heuristics to handle the birth and death of object instances in this framework. Since the number of queries ( $N$ ) is set to a high limit, it is larger than the actual number of instances ( $K$ ). So, there are queries that produce empty objects (represented by  $\emptyset$  in Fig. 1). The death of an object instance happens when its query is matched to such an empty query for more than five frames. If an object query is matched to an empty query for less than five frames, the query is carried forward and concatenated to the next frame’s object queries. The birth of an instance is correctly handled if the matched query embeddings have been null before the actual birth of the object instance. Since the matching process does not need training, tracking can be performed in an unsupervised manner. The unsupervised tracking is particularly useful because of the lack of availability of densely annotated open-source colonoscopic videos with polyps.

**Table 1.** Joint detection and segmentation performance on seen datasets.

Type	Method	Venue	Kvasir-SEG				CVC-ClinicDB			
			Dice	IoU	Pre.	Rec.	Dice	IoU	Pre.	Rec.
Det	PraNet [15]	MICCAI'20	89.1	82.9	-	-	89.4	83.5	-	-
	UACANet [20]	ACM MM'21	91.4	86.1	-	-	93.6	88.9	-	-
	SSFormer-L [39]	MICCAI'22	92.2	87.1	-	-	90.7	85.6	-	-
	Polyp-PVT [14]	CAAI'23	92.2	86.9	-	-	93.4	88.4	-	-
	PVT-CAS [33]	WACV'23	92.2	87.2	-	-	93.6	88.9	-	-
Seg	Def. DETR [42]	ICLR'21	-	-	90.2	76.0	-	-	95.5	94.1
	DAB-DETR [24]	CVPR'22	-	-	90.7	80.2	-	-	94.0	92.6
	DINO [9]	ICLR'23	-	-	90.2	76.0	-	-	95.5	92.7
Joint	QueryNet [10]	MICCAI'24	93.3	88.3	91.7	82.6	94.2	89.4	97.0	97.0
	<b>Ours</b>		<b>94.7</b>	<b>91.0</b>	<b>98.0</b>	<b>97.0</b>	<b>95.6</b>	<b>91.8</b>	<b>98.4</b>	<b>98.9</b>

**Table 2.** Joint detection and segmentation performance on unseen datasets.

Type	Method	ETIS				CVC-ColonDB				CVC-300			
		Dice	IoU	Pre.	Re.	Dice	IoU	Pre.	Re.	Dice	IoU	Pre.	Re.
Det	PraNet [15]	66.5	58.1	-	-	74.7	66.1	-	-	87.5	79.7	-	-
	UACANet [20]	77.0	69.0	-	-	75.9	68.7	-	-	91.3	85.1	-	-
	SSFormer-L [39]	80.1	72.8	-	-	81.3	73.5	-	-	90.3	83.8	-	-
	Polyp-PVT [14]	78.1	69.7	-	-	81.3	72.9	-	-	89.8	82.8	-	-
	PVT-CAS [33]	78.6	70.8	-	-	81.6	73.5	-	-	89.2	82.3	-	-
Seg	Def. DETR [42]	-	-	72.6	70.2	-	-	79.9	82.6	-	-	90.5	91.8
	DAB-DETR [24]	-	-	73.6	71.2	-	-	77.5	78.2	-	-	88.5	90.0
	DINO [9]	-	-	71.3	68.3	-	-	77.5	78.2	-	-	91.7	91.7
Joint	QueryNet [10]	81.9	74.0	74.9	77.4	82.8	75.9	83.5	85.3	92.0	86.0	91.8	93.3
	<b>Ours</b>	<b>91.4</b>	<b>85.3</b>	<b>94.2</b>	<b>93.8</b>	<b>83.3</b>	<b>76.3</b>	<b>88.8</b>	<b>91.7</b>	<b>93.2</b>	<b>87.8</b>	<b>98.6</b>	<b>97.4</b>

### 3 Experiments

#### 3.1 Evaluation Datasets and Metrics

**Datasets.** We evaluate the detection and segmentation performance of our model on five popular polyp datasets as benchmarks: CVC-ClinicDB [2], Kvasir-SEG [18], CVC-ColonDB [38], ETIS [3] and CVC-300 [37]. We follow the same setting as PraNet [15], that is, only 900 images from the Kvasir-SEG dataset and 550 images from the CVC-ClinicDB dataset are used for training, and the remaining images are used to test the learning ability of our method. The other 3 datasets are completely unseen during training and are used to test the generalizability of our method. We also evaluate the detection and classification performance of polyps on the KUMC [22] validation dataset. To evaluate the unsupervised tracking consistency of our method, we use 1000 consecutive frames from three videos from the REAL-Colon dataset [5] without any fine-tuning. Note that there are currently no openly available datasets, to the best of our knowledge, to evaluate joint detection, segmentation, classification and tracking together, hence we evaluate on the aforementioned tasks to cover all the tasks.

**Metrics.** For joint detection and segmentation, we use the precision and recall metrics to evaluate the detection performance and the dice and IoU scores to measure the segmentation accuracy to be consistent with prior works [10, 15]. For detection and classification on KUMC [22] dataset, we report the F1 score

**Table 3.** Results on the KUMC [22] dataset.

Method	F1 Score
YOLOv4 [7]	57.2
FasterRCNN [35]	57.7
RetinaNet [23]	59.0
SSD [25]	66.5
TimeSformer [4]	75.8
CORP [16]	78.2
FAME [13]	76.9
ProViCo [30]	78.6
VCL [31]	78.1
ST-Adapter [29]	74.9
Endo-FM [40]	84.1
<b>Ours</b>	<b>90.9 <math>\pm</math> 0.5</b>

**Table 4.** Tracking results on a subset of the REAL-colon [5] dataset.

Method	DetA	AssA	HOTA	MOTA	IDF1
IoU	57.7	28.2	39.5	33.6	37.0
<b>Ours</b>	57.7	<b>49.9</b>	<b>53.2</b>	<b>34.3</b>	<b>52.7</b>

**Table 5.** Ablation on the ETIS dataset with Resnet-50, Swin-Large and DINOv2 backbones.

	Dice	IoU	Pre.	Re.
<b>Ours</b> (R50)	82.5	76.4	83.9	87.5
<b>Ours</b> (SwinL)	89.9	83.3	88.8	91.7
<b>Ours</b> (DINOv2)	<b>91.4</b>	<b>85.3</b>	<b>94.2</b>	<b>93.8</b>

following prior work [40]. To evaluate tracking, we report the object tracking metrics of DetA (detection accuracy) [26], AssA (association accuracy) [26], HOTA [26], MOTA (multi-object tracking accuracy) and IDF1 following prior works on multi-object tracking [27].

### 3.2 Quantitative Results

**Performance on joint detection and segmentation.** Tab. 1 shows the performance of recent models on the held-out validation images of the Kvasir-SEG and the CVC-ClonicDB datasets. Tab. 2 shows the performance of these models on the unseen CVC-ColonDB, ETIS and CVC-300 datasets. We observe that our model outperforms other methods for all the datasets, sometimes by a large margin (as seen for the ETIS dataset).

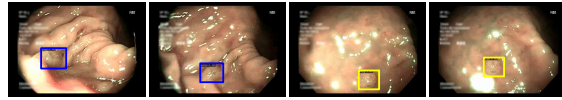
**Performance on polyp detection and classification.** We evaluate different methods on the KUMC dataset in Tab. 3. Our method outperforms the next best method, EndoFM [40], which is a also foundation model, significantly.

**Tracking performance.** To the best of our knowledge, we haven’t seen any methods performing tracking on polyps. To analyze the tracking performance of our method, we use a subset of the REAL-Colon dataset. Tab. 4 shows the comparison of our method with heuristic-based IoU matching for tracking (row 1). Note that, the detection results are generated using the same model and hence the detection accuracy (DetA) is exactly the same for both these methods.

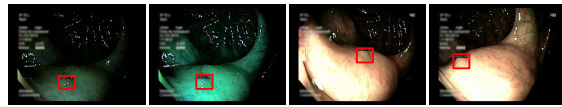
**Effect of different backbones.** Tab. 5 shows the performance of our model with different backbones, Resnet-50, Swin-L, and DINOv2. We see that DINOv2, being a general-purpose foundation model, outperforms the other models.

### 3.3 Qualitative Results

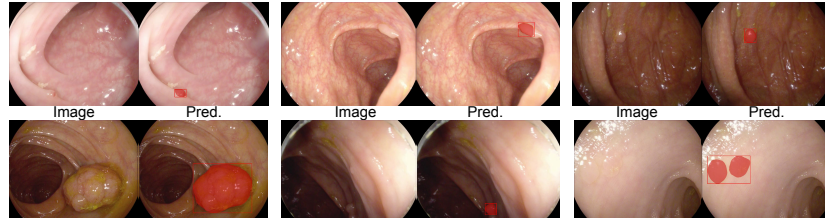
Fig. 2 shows two examples of joint detection, classification, and tracking on 2 videos of the KUMC dataset (top and bottom row), along with a comprehensive report generated for both videos. In the first example, we see 2 polyps (purple and yellow bounding boxes). Our model correctly identifies them as 2 different

				Total video frames: 145				
ID	Type	Fr. Ct.	1st Fr.	Last Fr.				
0	AD (100%)	56	0	56				
3	AD (92%)	63	66	130				

				Total video frames: 240				
ID	Type	Fr. Ct.	1st Fr.	Last Fr.				
0	HP (51%)	198	1	239				

**Fig. 2.** Detection, classification, and tracking on 2 videos (top and bottom row) of the KUMC dataset, along with comprehensive reports generated for each video. The reports summarize the polyps IDs, the polyp type (AD: cancerous or HP: benign) with prediction confidences, frame count (Fr. Ct.) of the polyps, their frame of first appearance (1st. Fr.) and their frame of last appearance (Last Fr.).



**Fig. 3.** Detection and segmentation results on a few images from the ETIS dataset. Our model is able to discover hard to see polyps in diverse scenes. Prediction results are shown in red.

polyps. We see a similar example in the bottom row, where our model is robust to different lighting conditions (green light and white light). Fig. 3 shows three examples from the ETIS dataset where we perform joint detection and segmentation on small and hard-to-see polyps. Additional results are shown in the supplementary video.

### 3.4 Training Data

For joint detection and segmentation (Tab. 1 and Tab. 2), our model is trained on 900 and 550 images of Kvasir-SEG and CVC-ClinicDB respectively, along with PolypDB [19], PolypGen [1], and KUMC [22] training datasets, to align our backbone and encoder-decoder to colonoscopic videos. This one-step fine-tuning removes the need to first pre-train our model on colonoscopic videos like in prior works with colonoscopic foundation models [40, 12]. Note that KUMC [22] dataset only has bounding-box-based annotations (with 28k training images), whereas PolypDB, PolypGen have available segmentation masks (with 6k frames combined). The KUMC dataset has polyp classification information (i.e., whether a polyp is AD: cancerous or HP: non-cancerous). For this experiment, both classes are just treated as polyps. For detection and classification (Tab. 3), our model is directly fine-tuned on the KUMC [22] training dataset. For the unsupervised tracking experiment (Tab. 4), we directly use the trained model from



joint detection and segmentation (Tab. 1 and Tab. 2), and test it on the REAL-Colon data without fine-tuning. For Fig. 2, we directly use the tracking module on top of the model trained on the detection and classification task (Tab. 3).

## 4 Conclusion

In this paper, we introduced PolypSegTrack, a novel foundation model that jointly addresses polyp detection, segmentation, classification and unsupervised tracking in colonoscopic videos. Our novel conditional mask loss enables flexible training and our unsupervised and non-heuristic tracking approach reliably tracks polyp instances across video frames. Extensive experiments on multiple polyp benchmarks demonstrate that our method significantly outperforms existing state-of-the-art approaches in polyp detection, segmentation, classification and tracking.

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

## References

1. Ali, S., Jha, D., Ghatwary, N., Realdon, S., Cannizzaro, R., Salem, O.E., Lamarque, D., Daul, C., Riegler, M.A., Anonsen, K.V., et al.: A multi-centre polyp detection and segmentation dataset for generalisability assessment. *Scientific Data* (2023)
2. Bernal, J., Sánchez, F.J., Fernández-Esparrach, G., Gil, D., Rodríguez, C., Vilarinho, F.: Wm-dova maps for accurate polyp highlighting in colonoscopy: Validation vs. saliency maps from physicians. *CMIG* (2015)
3. Bernal, J., Tajkbaksh, N., Sanchez, F.J., Matuszewski, B.J., Chen, H., Yu, L., Angermann, Q., Romain, O., Rustad, B., Balasingham, I., et al.: Comparative validation of polyp detection methods in video colonoscopy: results from the miccai 2015 endoscopic vision challenge. *IEEE Trans Med Imaging* (2017)
4. Bertasius, G., Wang, H., Torresani, L.: Is space-time attention all you need for video understanding? In: *ICML* (2021)
5. Biffi, C., Antonelli, G., Bernhofer, S., Hassan, C., Hirata, D., Iwatate, M., Maieron, A., Salvagnini, P., Cherubini, A.: Real-colon: A dataset for developing real-world ai applications in colonoscopy. *Scientific Data* (2024)
6. Bochinski, E., Eiselein, V., Sikora, T.: High-speed tracking-by-detection without using image information. In: *AVSS*. IEEE (2017)
7. Bochkovskiy, A., Wang, C.Y., Liao, H.Y.M.: Yolov4: Optimal speed and accuracy of object detection. *arXiv:2004.10934* (2020)
8. Carion, N., Massa, F., Synnaeve, G., Usunier, N., Kirillov, A., Zagoruyko, S.: End-to-end object detection with transformers. In: *ECCV*. Springer (2020)
9. Caron, M., Touvron, H., Misra, I., Jégou, H., Mairal, J., Bojanowski, P., Joulin, A.: Emerging properties in self-supervised vision transformers. In: *ICCV* (2021)
10. Chai, J., Luo, Z., Gao, J., Dai, L., Lai, Y., Li, S.: Querynet: A unified framework for accurate polyp segmentation and detection. In: *MICCAI* (2024)
11. Cheng, B., Misra, I., Schwing, A.G., Kirillov, A., Girdhar, R.: Masked-attention mask transformer for universal image segmentation. In: *CVPR* (2022)

12. Dermeyer, P., Kalra, A., Schwartz, M.: Endodino: A foundation model for gi endoscopy. arXiv:2501.05488 (2025)
13. Ding, S., Li, M., Yang, T., Qian, R., Xu, H., Chen, Q., Wang, J., Xiong, H.: Motion-aware contrastive video representation learning via foreground-background merging. In: CVPR (2022)
14. Dong, B., Wang, W., Fan, D.P., Li, J., Fu, H., Shao, L.: Polyp-pvt: Polyp segmentation with pyramid vision transformers. arXiv:2108.06932 (2021)
15. Fan, D.P., Ji, G.P., Zhou, T., Chen, G., Fu, H., Shen, J., Shao, L.: Pranel: Parallel reverse attention network for polyp segmentation. In: MICCAI. Springer (2020)
16. Hu, K., Shao, J., Liu, Y., Raj, B., Savvides, M., Shen, Z.: Contrast and order representations for video self-supervised learning. In: ICCV (2021)
17. Huang, D.A., Yu, Z., Anandkumar, A.: Minvis: A minimal video instance segmentation framework without video-based training. NeurIPS (2022)
18. Jha, D., Smedsrud, P.H., Riegler, M.A., Halvorsen, P., De Lange, T., Johansen, D., Johansen, H.D.: Kvasir-seg: A segmented polyp dataset. In: MMM (2020)
19. Jha, D., Tomar, N.K., Sharma, V., Trinh, Q.H., Biswas, K., Pan, H., Jha, R.K., Durak, G., Hann, A., Varkey, J., et al.: Polypdb: A curated multi-center dataset for development of ai algorithms in colonoscopy. arXiv:2409.00045 (2024)
20. Kim, T., Lee, H., Kim, D.: Ucanet: Uncertainty augmented context attention for polyp segmentation. In: MM (2021)
21. Li, F., Zhang, H., Xu, H., Liu, S., Zhang, L., Ni, L.M., Shum, H.Y.: Mask dino: Towards a unified transformer-based framework for object detection and segmentation. In: CVPR (2023)
22. Li, K., Fathan, M.I., Patel, K., Zhang, T., Zhong, C., Bansal, A., Rastogi, A., Wang, J.S., Wang, G.: Colonoscopy polyp detection and classification: Dataset creation and comparative evaluations. Plos one (2021)
23. Lin, T.Y., Goyal, P., Girshick, R., He, K., Dollár, P.: Focal loss for dense object detection. In: ICCV (2017)
24. Liu, S., Li, F., Zhang, H., Yang, X., Qi, X., Su, H., Zhu, J., Zhang, L.: Dab-detr: Dynamic anchor boxes are better queries for detr. arXiv:2201.12329 (2022)
25. Liu, W., Anguelov, D., Erhan, D., Szegedy, C., Reed, S., Fu, C.Y., Berg, A.C.: Ssd: Single shot multibox detector. In: ECCV. Springer (2016)
26. Luiten, J., Osep, A., Dendorfer, P., Torr, P., Geiger, A., Leal-Taixé, L., Leibe, B.: Hota: A higher order metric for evaluating multi-object tracking. IJCV (2021)
27. Meinhardt, T., Kirillov, A., Leal-Taixé, L., Feichtenhofer, C.: Trackformer: Multi-object tracking with transformers. In: CVPR (2022)
28. Oquab, M., Darcet, T., Moutakanni, T., Vo, H., Szafraniec, M., Khalidov, V., Fernandez, P., Haziza, D., Massa, F., El-Nouby, A., et al.: Dinov2: Learning robust visual features without supervision. arXiv:2304.07193 (2023)
29. Pan, J., Lin, Z., Zhu, X., Shao, J., ST-Adapter, H.L.: Parameter-efficient image-to-video transfer learning for action recognition. Preprint at <https://arxiv.org/abs/2206.13559> (2022)
30. Park, J., Lee, J., Kim, I.J., Sohn, K.: Probabilistic representations for video contrastive learning. In: CVPR (2022)
31. Qian, R., Ding, S., Liu, X., Lin, D.: Static and dynamic concepts for self-supervised video representation learning. In: ECCV. Springer (2022)
32. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al.: Learning transferable visual models from natural language supervision. In: ICML. PmLR (2021)
33. Rahman, M.M., Marculescu, R.: Medical image segmentation via cascaded attention decoding. In: WACV (2023)

34. Ravi, N., Gabeur, V., Hu, Y.T., Hu, R., Ryali, C., Ma, T., Khedr, H., Rädle, R., Rolland, C., Gustafson, L., et al.: Sam 2: Segment anything in images and videos. arXiv:2408.00714 (2024)
35. Ren, S., He, K., Girshick, R., Sun, J.: Faster r-cnn: Towards real-time object detection with region proposal networks. TPAMI (2016)
36. Ronneberger, O., Fischer, P., Brox, T.: U-net: Convolutional networks for biomedical image segmentation. In: MICCAI. Springer (2015)
37. Tajbakhsh, N., Gurudu, S.R., Liang, J.: Automated polyp detection in colonoscopy videos using shape and context information. IEEE Trans Med Imaging (2015)
38. Vázquez, D., Bernal, J., Sánchez, F.J., Fernández-Esparrach, G., López, A.M., Romero, A., Drozdal, M., Courville, A.: A benchmark for endoluminal scene segmentation of colonoscopy images. Journal of healthcare engineering (2017)
39. Wang, J., Huang, Q., Tang, F., Meng, J., Su, J., Song, S.: Stepwise feature fusion: Local guides global. In: MICCAI. Springer (2022)
40. Wang, Z., Liu, C., Zhang, S., Dou, Q.: Foundation model for endoscopy video analysis via large-scale self-supervised pre-train. In: MICCAI. Springer (2023)
41. Zhang, H., Li, F., Zou, X., Liu, S., Li, C., Yang, J., Zhang, L.: A simple framework for open-vocabulary segmentation and detection. In: ICCV (2023)
42. Zhu, X., Su, W., Lu, L., Li, B., Wang, X., Dai, J.: Deformable detr: Deformable transformers for end-to-end object detection. arXiv:2010.04159 (2020)