Supplementary materials

1 Effect of loss function

The performance of the model depends not only on the architecture but also on the loss function during training. As with most state-of-the-art models, we use weighted BCE-Dice loss function to train our DALA. The impact of the weight factor λ on the model performance is shown in Table S1 and can be observed that $\mathcal{L}_{BCE}$ with $\lambda = 0.6$ has better segmentation.

Table S1: Effect of the hybrid loss function with different λ values.

λ	SUN-SEG-Easy(Unseen)		SUN-SEG-Hard(Unseen)	
	$S_\alpha \uparrow$	Dice↑	$S_\alpha \uparrow$	Dice↑
1	0.829	0.755	0.799	0.725
0.8	0.834	0.762	0.807	0.730
0.6	**0.837**	**0.768**	**0.808**	0.733
0.5	0.830	0.759	0.805	**0.736**
0.4	0.822	0.750	0.798	0.729
0.2	0.813	0.743	0.791	0.721
0	0.795	0.726	0.768	0.709

2 Comparison on the HD95 score

Hausdorff distance(95^{th} per centile, HD95) is a boundary-based metric that measures the distance between boundaries. The HD95 of predicted map $\boldsymbol{P}$ and the ground-truth $\boldsymbol{G}$ is computed as:

$$HD95(\boldsymbol{P}, \boldsymbol{G}) = max\left\{h(\boldsymbol{P}_b, \boldsymbol{G}_b), h(\boldsymbol{G}_b, \boldsymbol{P}_b)\right\} \tag{S1}$$

where $\boldsymbol{P}_b$ and $\boldsymbol{G}_b$ denote the predicted boundary points and the ground-truth boundary points in the $\boldsymbol{P}$ and $\boldsymbol{G}$. $h(\boldsymbol{P}_b, \boldsymbol{G}_b) = \hat{max}_{a \in \boldsymbol{P}_b}\left\{min_{b \in \boldsymbol{G}_b} \parallel a - b \parallel\right\}$ denotes the one-way hausdorff distance from $\boldsymbol{P}_b$ to $\boldsymbol{G}_b$, and $\hat{max}\{\cdot\}$ refers to the calculation of the 95^{th} percentile of the distances. We show the compared results in Table S2, where the best score is highlighted in bold.

Table S2: Quantitative comparison on HD95 for two unseen sub-datasets.

	Method	SUN-SEG-Easy(Unseen) $HD95 \downarrow$	SUN-SEG-Hard(Unseen) $HD95 \downarrow$
Image	UNet [1]	13.84	14.60
	UNet++ [2]	13.31	13,14
	ACSNet [3]	10.92	10.94
	PraNet [4]	12.66	12.90
	ColonSegNet [5]	11.55	11.29
	MedNeXt [6]	13.61	13.28
	MSRF-Net [7]	10.96	11.16
	TransNetR [8]	11.04	10.89
Video	PCSA [9]	12.90	12.69
	2/3D [10]	11.12	10.99
	FSNet [11]	12.02	12.02
	PNSNet [12]	11.75	11.57
	PNS+ [13]	10.66	**10.59**
	SSTAN [14]	10.53	11.09
	FLA-Net [15]	12.51	12.86
	DALA(ours)	**10.38**	10.72

3 Performance on the ASU-Mayo dataset

ASU-Mayo [16] contains 36,458 continuous frames from 38 videos, while it only provides 3,856 pixel-level labels for 10 positive videos. We only adopt the positive part and split the videos into 60% for training, 20% for validation, and 20% for testing. We compare our model to several competitive segmentation models, including the image-based models, ACSNet [3], PraNet [4], ColonSegNet [5], MRSF-Net [7], TransNetR [8], and the video-based models, 2/3D [10], FSNet [11], PNS+ [13] and SSTAN [14]. The quantitative results from Table S3 show that our approach surpasses other SOTA methods on the S_α score and *Dice* score.

Table S3: Quantitative comparison on the ASU-Mayo dataset.

	Method	$S_\alpha \uparrow$	$Dice \uparrow$
Image	ACSNet [3]	0.809	0.740
	PraNet [4]	0.792	0.733
	ColonSegNet [5]	0.669	0.709
	MSRF-Net [7]	0.843	0.772
	TransNetR [8]	0.855	0.776
Video	2/3D [10]	0.840	0.761
	FSNet [11]	0.784	0.725
	PNSNet [12]	0.871	0.780
	PNS+ [13]	0.880	0.792
	SSTAN [14]	0.875	0.800
	DALA(ours)	**0.896**	**0.808**

4 More details on limitations and challenges

VPS is a challenging task in medical imaging and its overall accuracy is not high enough. This section discusses in detail some common issues within visualization attributes. According to PNS+ [13], these visualization attributes are classified into ten categories, e.g., intestinal contents in a colonoscopy video may disturb the results of detection or segmentation. Classification criteria of visualization attributes are detailed in Table S4. Attribute-based comparisons with several cutting-edge image-based or video-based models on the S_α score are presented in Table S5 and Table S6. Figure S1 shows the qualitative results of the existing cutting-edge model and our model struggling to produce accurate segmentation on every visual attribute.

In terms of S_α score, Table S5 and Table S6 show that our DALA consistently outperforms the other competitors on three attributes (i.e., SI, IB, and FM). More specifically, since colon polyps usually have fuzzy boundaries, most methods

Table S4: List of ten types of visual attributes and their descriptions.

Attribute	Description
SI	***Surgical Instruments***. The endoscopic surgical procedures involve the positioning of instruments, such as snares,forceps, knives, and electrodes.
IB	***Indefinable Boundaries.*** The foreground and background areas around the object have similar color.
HO	***Heterogeneous Object.*** Object regions have distinct colors.
GH	***Ghosting.*** Object has anomaly RGB-colored boundary due to fast moving or insufficient refresh rate.
FM	***Fast-motion.*** The average per-frame object motion in a clip, computed as the Euclidean distance of polyp centroids betweenconsecutive frames, is larger than 20 pixels.
SO	***Small Object.*** The average ratio between the object size and the image area in a clip is smaller than 0.05.
LO	***Large Object.*** The average ratio between the object size and the image area in a clip is larger than 0.15.
OC	***Occlusion.*** Polyp object becomes partially or fully occluded.
OV	***Out-of-view.*** Polyp object is partially clipped by the image boundaries.
SV	***Scale-variation.*** The average area ratio among any pair of bounding boxes enclosing the target object in a clip is smaller than 0.5.

Table S5: Attribute-based performance on SUN-SEG-Easy(Unseen) in terms of S_α score.

	Method	SI	IB	HO	GH	FM	SO	LO	OC	OV	SV
Image	UNet [1]	0.675	0.548	0.768	0.715	0.633	0.593	0.648	0.670	0.640	0.620
	UNet++ [2]	0.701	0.542	0.782	0.739	0.647	0.591	0.678	0.683	0.665	0.617
	ACSNet [3]	0.789	0.612	0.896	0.820	0.704	0.663	0.787	0.770	0.759	0.705
	TransNetR [8]	0.805	0.659	0.873	**0.859**	0.691	0.674	0.779	**0.789**	0.770	0.729
Video	2/3D [10]	0.809	0.625	**0.899**	0.835	0.728	0.667	**0.820**	0.783	0.778	0.719
	PNSNet [12]	0.789	0.592	0.871	0.820	0.723	0.619	0.768	0.749	0.751	0.705
	PNS+ [13]	0.819	0.667	0.883	0.844	0.738	0.690	0.796	0.782	**0.798**	0.734
	SSTAN [14]	0.800	0.590	0.874	0.825	0.709	0.688	0.804	0.776	0.763	0.730
	DALA(ours)	**0.829**	**0.694**	0.870	0.841	**0.759**	**0.699**	0.813	0.780	0.792	**0.749**

Table S6: Attribute-based performance on SUN-SEG-Hard(Unseen) in terms of S_α score.

	Method	SI	IB	HO	GH	FM	SO	LO	OC	OV	SV
Image	UNet [1]	0.618	0.619	0.663	0.676	0.713	0.689	0.633	0.658	0.659	0.658
	UNet++ [2]	0.654	0.604	0.665	0.696	0.714	0.681	0.660	0.676	0.677	0.678
	ACSNet [3]	0.770	0.681	0.828	0.795	0.817	0.738	0.810	**0.828**	0.806	0.759
	TransNetR [8]	0.786	0.671	0.808	0.803	0.724	0.740	0.719	0.719	0.743	0.710
Video	2/3D [10]	0.768	0.662	**0.865**	0.784	0.797	0.737	**0.853**	0.827	0.808	0.765
	PNSNet [12]	0.746	0.631	0.803	0.780	0.778	0.743	0.805	0.790	0.794	0.758
	PNS+ [13]	0.770	0.703	0.817	0.801	0.823	**0.793**	0.792	0.808	0.807	**0.795**
	SSTAN [14]	0.759	0.683	0.791	**0.811**	0.807	0.758	0.820	0.779	0.815	0.765
	DALA(ours)	**0.778**	**0.714**	0.856	0.809	**0.834**	0.788	0.839	0.824	**0.829**	0.768

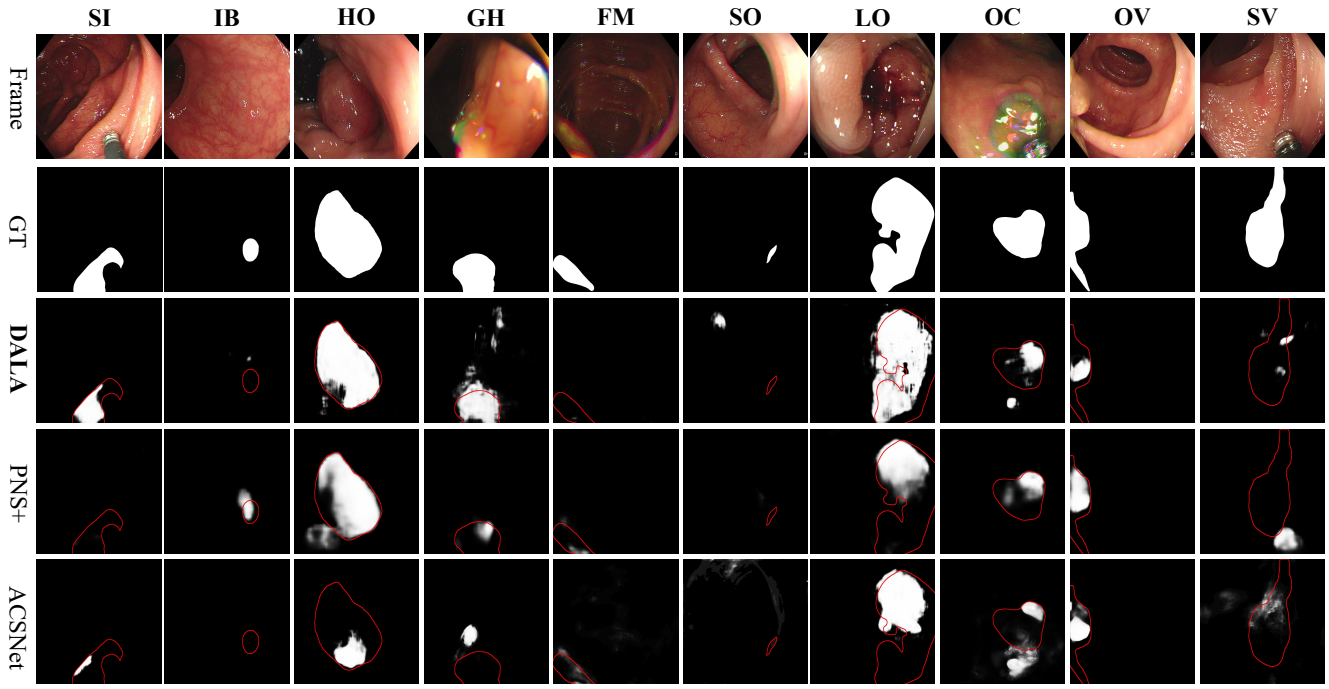

Figure S1: Qualitative results are taken from ten visual attributes.

cannot handle the VPS task with IB attribute. In contrast, DALA obtained the best score (S_α = 0.694/0.714) on this challenging IB attribute of SUN-SEG-Easy (Unseen)/-Hard (Unseen). In addition, DALA also achieves the best result on FM attribute which suggests that DALA has better stabilization and segmentation in colonoscopy videos with fast

camera moves. On the whole, the IB and SO attributes show lower scores, suggesting that these two attributes are the most challenging issues in colonoscopy. On the contrary, the HO and LO attributes consistently maintain higher scores than the other attributes, making polyps easier to segment. This is in line with our basic assumption that visual features are more pronounced in these cases.

We can observe from Figure S1 that existing cutting-edge models (i.e., ACSNet and PNS+) and our model (DALA) still lack sufficient robustness in particular cases of most attributes. Over- and under-segmentation on surgical instruments (1^{st} column, SI) and distinct colors (4^{th} column, GH) show that these models do not perceive accurate polyp-related representations and fail to learn semantics. Besides, as for the HO (3^{rd} column) and LO (7^{th} column) attributes, ACSNet and PNS+ fail to capture the whole polyp due to significant appearance changes. DALA also has difficulty segmenting complete boundaries. Furthermore, the misidentification of the SV attribute (last column) is due to the lack of diversity of polyp shapes in the training set. The above issues motivate us to seek better learning paradigms to improve the robustness of our models.

We can also observe difficulty in being localized when the object is too small (6^{th} column, SO) or similar in color to the surrounding tissue (2^{nd} column, IB). Finally, the lack of temporo-spatial modeling will lead to the false prediction in the FM, OV, and OC attributes. Considering how to model the time-space relationship can reduce performance degradation due to occlusion or out-of-bounds, while increasing temporal stability under camera moves at a high speed.

5 Evaluation for polyp detection

Model details. We explore the performance of DALA on polyp detection tasks to improve clinical applicability. The core parts of DALA (pre-trained ConvNeXt, Multi-Scale Frame Alignment, and Local Attention) are retained while we replace the segmentation head (Decoder) with the detection head of CenterNet [17] (as shown in Figure S2). The detection head is composed of a 3×3 Conv and a 1×1 Conv to produce center, size, and offset features for detection loss:

$$\mathcal{L}_{detection} = \mathcal{L}_{focal}^{center} + \lambda_{size}\mathcal{L}_{L1}^{size} + \lambda_{offset}\mathcal{L}_{L1}^{offset} \tag{S2}$$

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

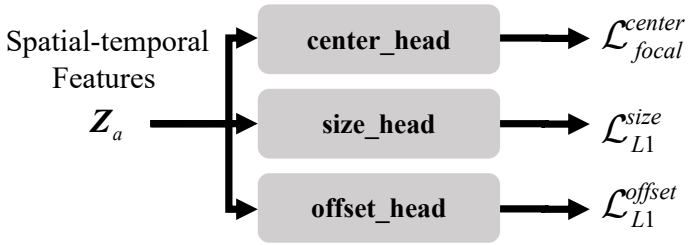

Figure S2: Detection head for polyp detection.

Datasets. We evaluate DALA on two public video polyp detection benchmarks: SUN-SEG (train set: 19,544 frames, test set: 12,351 frames of SUN-SEG-Hard(Unseen) and 8,640 frames of SUN-SEG-Hard(Unseen)) and CVC-VideoClinicDB [18] (train set: 7995 frames, test set: 2030 frames). For the fairness of the experiments, we keep the same dataset settings for DALA and all other methods.

Implementation details. Following the same setting in CenterNet, we set $\lambda_{size} = 0.1$ and $\lambda_{offset} = 1$. We set the batchsize N = 16. Our model is trained using the Adam optimizer with a weight decay of 5×10^{-4} for 100 epochs. The initial learning rate is set to 10^{-3} and gradually decays to 10^{-4} with cosine annealing. All models are trained with PyTorch framework. The training setting of other competitors follows the best settings given in their paper. We choose precision, recall, and mAP as evaluation metrics.

The comparison results are shown in Table S7. Firstly, compared with the CenterNet baseline, our DALA with three novel designs significantly improved the mAP score by 2.1%, 12.8%, and 10.1% on three benchmarks, demonstrating the effectiveness of the model design. Second, for video-based competitors, previous video object detectors with multiple frame collaborations lack the ability for accurate detection on challenging datasets. Specifically, DALA surpasses the second-best YONA [27] by 1.1%, 2.4%, and 0.6% on mAP score on three benchmarks and 6.3 on FPS. All the results confirm the superiority of our DALA for accurate and fast video polyp detection.

Table S7: Performance comparison with other image/video-based models for polyp detection.

	Method	SUN-SEG-Easy(Unseen)			SUN-SEG-Hard(Unseen)			CVC-VideoclinicDB			FPS
		P	R	mAP	P	R	mAP	P	R	mAP	
Image	Faster-RCNN [19]	0.755	0.831	0.850	0.740	0.692	0.691	0.833	**0.978**	0.916	46.8
	FCOS [20]	0.761	0.825	0.843	0.698	0.653	0.666	0.915	0.754	0.821	43.9
	CenterNet [17]	0.790	0.824	0.859	0.706	0.642	0.658	0.920	0.783	0.851	53.3
	DINO [21]	0.808	0.849	0.872	0.793	0.704	0.719	0.925	0.880	0.925	34.6
	YOLOv9-S [22]	0.821	0.816	0.837	0.776	0.691	0.714	0.905	0.860	0.897	**83.1**
Video	FGFA [23]	0.801	0.823	0.836	0.721	0.683	0.709	**0.945**	0.896	0.929	3.2
	MEGA [24]	0.822	0.809	0.831	0.740	0.709	0.723	0.906	0.871	0.900	10.2
	TransVOD [25]	0.847	0.836	0.857	0.761	0.729	0.735	0.919	0.900	0.911	10.4
	STFT [26]	0.865	0.844	0.872	0.840	0.738	0.759	0.919	0.925	0.939	15.6
	YONA [27]	**0.878**	0.840	0.869	0.852	0.749	0.772	0.931	0.926	0.946	49.7
	DALA(ours)	0.867	**0.852**	**0.880**	**0.864**	**0.761**	**0.786**	0.926	0944	**0.952**	56.0

6 Validation on Clinical Metrics

The core goal of polyp screening is to minimize underdiagnosis (high sensitivity) and ensure diagnostic accuracy, which we complement with clinically relevant metrics: specificity (Spec), positive predictive value (PPV) and negative predictive value (NPV). The results of the clinical metrics are shown in TableS8.

Table S8: Comparison of clinical metrics on two unseen sub-datasets.

	Method	SUN-SEG-Easy(Unseen)				SUN-SEG-Hard(Unseen)			
		Sen↑	Spec↑	PPV↑	NPV↑	Sen↑	Spec↑	PPV↑	NPV↑
Image	UNet	0.420	0.811	0.713	0.702	0.429	0.879	0.742	0.733
	UNet++	0.457	0.796	0.702	0.696	0.467	0.863	0.751	0.748
	ACSNet	0.601	0.925	0.904	0.890	0.618	0.913	0.875	0.869
	PraNet	0.524	0.876	0.869	0.833	0.512	0.879	0.786	0.777
	ColonSegNet	0.594	0.918	0.893	0.879	0.599	0.889	0.861	0.834
	MedNeXt	0.506	0.863	0.843	0.814	0.508	0.883	0.775	0.761
	MSRF-Net	0.600	0.928	0.902	0.888	0.605	0.896	0.858	0.849
	TransNetR	0.652	0.950	0.933	0.919	0.655	0.936	0.899	0.853
Video	PCSA	0.398	0.786	0.690	0.681	0.415	0.855	0.730	0.719
	2/3D	0.603	0.933	0.908	0.895	0.607	0.909	0.868	0.853
	FSNet	0.493	0.852	0.829	0.793	0.491	0.881	0.769	0.757
	PNSNet	0.574	0.905	0.884	0.878	0.579	0.899	0.839	0.826
	PNS+	0.630	0.963	0.944	0.927	0.623	0.929	0.917	0.871
	SSTAN	0.662	0.956	0.930	0.911	**0.676**	0.939	0.891	0.860
	FLA-Net	0.506	0.855	0.845	0.806	0.522	0.890	0.861	0.850
	DALA(ours)	**0.721**	**0.997**	**0.990**	**0.979**	0.669	**0.949**	**0.893**	**0.885**

References

[1] O. Ronneberger, P. Fischer, and T. Brox, "U-Net: Convolutional networks for biomedical image segmentation," in *Proc. Int. Conf. Med. Image Comput. Comput.-Assist. Intervent.*, 2015, pp. 234–241.

[2] Z. Zhou, M. M. R. Siddiquee, N. Tajbakhsh, and J. Liang, "UNet++: Redesigning skip connections to exploit multiscale features in image segmentation," *IEEE Trans. Med. Imag.*, vol. 39, no. 6, pp. 1856–1867, Jun. 2020.

[3] R. Zhang et al., "Adaptive context selection for polyp segmentation," in *Proc. Int. Conf. Med. Image Comput. Comput.-Assist. Intervent.*, Cham, Switzerland: Springer, 2020, pp. 253–262.

[4] D.-P. Fan et al., "PraNet: Parallel reverse attention network for polyp segmentation," in *Proc. Int. Conf. Med. Image Comput. Comput.-Assist. Intervent.* Cham, Switzerland: Springer, 2020, pp. 263–273.

[5] D. Jha et al., " Real-Time Polyp Detection, Localization and Segmentation in Colonoscopy Using Deep Learning," *IEEE Access.*, vol. 9, pp. 40496-40510, 2023.

[6] S. Roy et al., "Mednext: transformer-driven scaling of convnets for medical image segmentation," in *Proc. Int. Conf. Med. Image Comput. Comput.-Assist. Intervent.* Cham, Switzerland: Springer, 2023, pp. 405–415.

[7] A. Srivastava et al., "MSRF-Net: A multi-scale residual fusion network for biomedical image segmentation," *IEEE J. Biomed. Health Informat.*, vol. 26, no. 5, pp. 2252–2263, May 2022.

[8] D. Jha et al., "Transnetr: transformer-based residual network for polyp segmentation with multi-center out-of-distribution testing," *Proc. Int. Conf. Medical Imaging Deep Learn.*, 2024, pp. 1372–1384.

[9] Y. Gu et al., "Pyramid constrained self-attention network for fast video salient object detection," in *Proc. AAAI Conf. Artif. Intell.*, Vol. 34. No. 07. 2020.

[10] J. G.-B. Puyal et al., "Endoscopic polyp segmentation using a hybrid 2d/3d cnn," in *Proc. Int. Conf. Med. Image Comput. Comput.-Assist. Intervent.* Springer, 2020, pp. 295–305.

[11] G. -P. Ji, K. Fu, Z. Wu, D. -P. Fan, J. Shen and L. Shao, "Full-Duplex Strategy for Video Object Segmentation," in *IEEE Int. Conf. Comput. Vis. (ICCV)*, 2021, pp. 4902-4913.

[12] G.-P. Ji et al., "Progressively normalized self-attention network for video polyp segmentation," in *Proc. Int. Conf. Med. Image Comput. Comput.-Assist. Intervent.* Cham, Switzerland: Springer, 2021, pp. 142–152.

[13] G.-P. Ji et al., "Video polyp segmentation: A deep learning perspective," *Mach. Intell. Research.*, vol. 19, no. 6, pp. 531–549, 2022.

[14] X. Zhao et al., "Semi-supervised spatial temporal attention network for video polyp segmentation," in *Proc. Int. Conf. Med. Image Comput. Comput.-Assist. Intervent.* Cham, Switzerland: Springer, 2022, pp. 456–466.

[15] J. Lin et al., "Shifting more attention to breast lesion segmentation in ultrasound videos," in *Proc. Int. Conf. Med. Image Comput. Comput.-Assist. Intervent.* Cham, Switzerland: Springer, 2023, pp. 497–507.

[16] N. Tajbakhsh, S. R. Gurudu, and J. Liang, "Automated polyp detection in colonoscopy videos using shape and context information," in *IEEE Trans. Med. Imaging* vol. 35, no. 2, pp. 630–644, 2016.

[17] K. Duan et al., "Centernet: Keypoint triplets for object detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2019, pp. 6569–6578.

[18] J. J. Bernal et al., "Polyp detection benchmark in colonoscopy videos using gtcreator: A novel fully configurable tool for easy and fast annotation of image databases," in *Proc. 32nd CARS conf.*, 2018.

[19] S. Ren et al., "Faster r-cnn: Towards real-time object detection with region proposal networks," in *Proc. Adv. Neural Inf. Process. Syst.*, 2015.

[20] Z. Tian et al., " Fcos: Fully convolutional one-stage object detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2019, pp. 9627–9636.

[21] H. Zhang et al., "Dino: Detr with improved denoising anchor boxes for end-to-end object detection," in *Proc. Int. Conf. Learn. Represent.*, 2022.

[22] C. Y. Wang et al., "Yolov9: Learning what you want to learn using programmable gradient information," 2024, arXiv:2402.13616.

[23] X. Zhu et al., "Flow-guided feature aggregation for video object detection," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2017, pp. 408–417.

[24] H. Zheng et al., "Polyp tracking in video colonoscopy using optical flow with an on-the-fly trained cnn," in *Proc. IEEE Int. Symp. Biomed. Imaging*, 2019, pp. 79–82.

[25] Q. Zhou et al., "Transvod: end-to-end video object detection with spatial-temporal transformers.," in *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 45, no. 6, pp. 7853-7869, 2022.

[26] L. Wu et al., "Multi-frame collaboration for effective endoscopic video polyp detection via spatial-temporal feature transformation," in *Proc. Int. Conf. Med. Image Comput. Comput.-Assist. Intervent.* Springer, 2021, pp. 302–312.

[27] Y. Jiang et al., "Yona: you only need one adjacent reference-frame for accurate and fast video polyp detection," in *Proc. Int. Conf. Med. Image Comput. Comput.-Assist. Intervent.* Springer, 2023, pp. 44–54.