Diff-VPS: Video Polyp Segmentation via a Multi-task Diffusion Network with Adversarial Temporal Reasoning

Yingling Lu[1], Yijun Yang[1], Zhaohu Xing[1],
Qiong Wang[2] (✉), and Lei Zhu[1,3] (✉)

[1] The Hong Kong University of Science and Technology (Guangzhou)
[2] Guangdong Provincial Key Laboratory of Computer Vision and Virtual Reality
Technology, Shenzhen Institutes of Advanced Technology, Chinese Academy of
Sciences
[3] The Hong Kong University of Science and Technology
`leizhu@ust.hk`

Abstract. Diffusion Probabilistic Models have recently attracted significant attention in the community of computer vision due to their outstanding performance. However, while a substantial amount of diffusion-based research has focused on generative tasks, no work introduces diffusion models to advance the results of polyp segmentation in videos, which is frequently challenged by polyps' high camouflage and redundant temporal cues. In this paper, we present a novel diffusion-based network for video polyp segmentation task, dubbed as *Diff-VPS*. We incorporate multi-task supervision into diffusion models to promote the discrimination of diffusion models on pixel-by-pixel segmentation. This integrates the contextual high-level information achieved by the joint classification and detection tasks. To explore the temporal dependency, Temporal Reasoning Module (TRM) is devised via reasoning and reconstructing the target frame from the previous frames. We further equip TRM with a generative adversarial self-supervised strategy to produce more realistic frames and thus capture better dynamic cues. Extensive experiments are conducted on SUN-SEG, and the results indicate that our proposed Diff-VPS significantly achieves state-of-the-art performance. Code is available at https://github.com/lydia-yllu/Diff-VPS.

Keywords: Diffusion model · Video polyp segmentation · Multi-task learning.

1   Introduction

Colorectal cancer (CRC) has become the second leading cause of cancer-related deaths worldwide, accounting for approximately 10% of all cancer cases. By 2040 the burden of colorectal cancer will increase to 3.2 million new cases per year (an increase of 63%) and 1.6 million deaths per year (an increase of 73%) [3,23]. Colonoscopy plays a crucial role in detecting polyps and serves as a screening

tool for CRC. However, there still exists a high missing rate in the diagnosis and treatment process. Recently, emerging methods have utilized deep learning techniques to automatically detect polyps in video, from popular convolutional neural networks to attention mechanism [19,17,8,10,28,18,1,29]. Nevertheless, video polyp segmentation (VPS) always poses the ill-posed bottleneck, *i.e.*, the *high camouflage* of polyps in the camera, more specifically, the intrinsic complexity of polyps (shape and color variability, fuzzy boundaries) and the external shooting conditions (low boundary contrast, specular reflection, artifacts). Also, the *temporal cues* in colonoscopy videos require to be finely explored, which refers to the effective utilization of long-term temporal characteristics within a video dataset. While parts of the above methods achieved significant progress in segmenting polyps in videos, there still exists great potential to be explored considering their failure cases and limited discriminative ability.

Diffusion models have shown impressive performance in many vision tasks, such as image deblurring [12], super-resolution [14,5], and anomaly detection [24,4]. The denoising diffusion probabilistic model (DDPM) [7] is a generative model based on a Markov chain. Simulating a diffusion process, DDPM captures the data distribution as it guides the input data towards a predefined target distribution, i.e. Gaussian distribution. By representing the diffusion process of image intensity values over successive iterations, diffusion models adeptly grasp the intrinsic structure and texture within images, facilitating the distinct segregation of regions of interest from the background [2]. While several pioneers [22,25,21,27] applied DDPM to medical image classification and segmentation, there is no diffusion-based work on medical video lesion segmentation.

Motivated by the nature of diffusion models [11], in this paper, **1) we present the first diffusion network, dubbed as Diff-VPS, for the VPS task.** Polyps can be diagnosed into different categories (hyperplastic polyp, invasive carcinoma, *etc.*) depending on the size and morphological features [10]. Inspired by this, **2) we design the Multi-task Diffusion Model where classification and detection tasks are performed simultaneously to improve the discrimination on the segmentation task.** Such supervision significantly advances the discriminative and generalizable ability of diffusion models by high-level semantic information. Furthermore, to capture the dynamic appearance and maintain the temporal coherence in segmenting polyps, **3) we devise a Temporal Reasoning Module by reconstructing the target frame from the previous temporal information and employing an adversarial self-supervised strategy.** Extensive experiments are conducted on the SUN-SEG dataset. The superior performance on seen and unseen videos validates the effectiveness and versatility of our Diff-VPS.

2   Method

Figure 1 shows the schematic illustration of our network for video polyp segmentation. Given the target frame I_i and its previous frames $\{I_{i-\delta}, .., I_{i-1}\}$, we first encode the pixel-level mask M_i^0 of I_i into the latent space, and then apply diffusion forward process to its latent feature z_i^0 by the randomly sampled
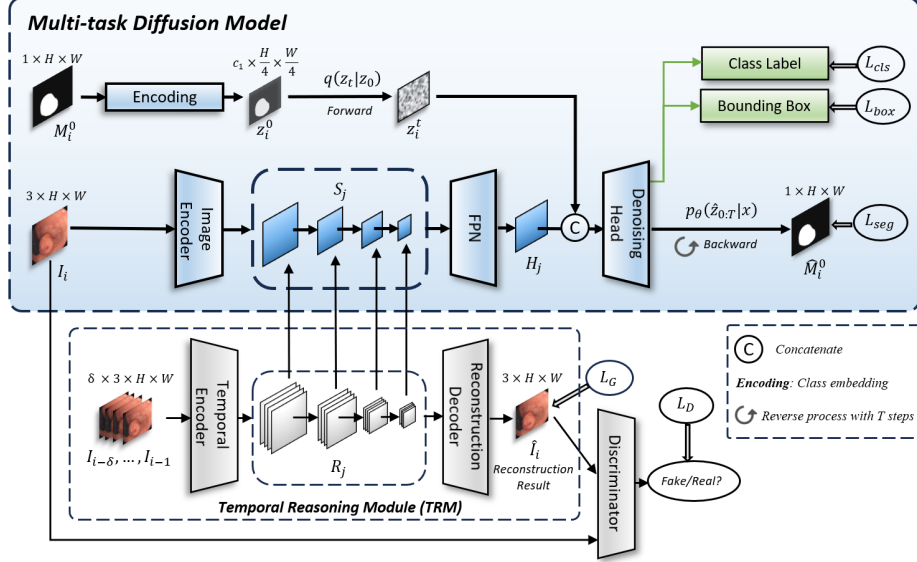
Fig. 1: **Overview of our Diff-VPS framework.** While the TRM extracts multi-scale temporal features $\mathcal{R}_j$ from the previous frames, the image encoder learns the spatial counterpart $\mathcal{S}_j$. The spatiotemporal prior $\mathcal{H}_j$ from $\mathcal{R}_j$ and $\mathcal{S}_j$ conditions the denoising process of our multi-task diffusion model.

timestep t to generate the corresponding noisy variable z_i^t. On the other hand, we feed the target frame I_i into the image encoder to extract its multi-scale spatial features. Simultaneously, we leverage the previous δ frames to abstract the multi-scale temporal features by reconstructing the target frame in our Temporal Reasoning Module. The temporal features are further integrated with the spatial feature scale-by-scale to construct the spatiotemporal prior. Finally, the denoising head predict the segmentation result $\hat{M}_i^0$ from the noisy variable z_i^t conditioned by the spatiotemporal prior.

2.1   Multi-task Diffusion Model (MDM)

Diffusion models. Our MDM is developed based on the conditional diffusion model [7,11], which belongs to the category of likelihood-based models inspired by non-equilibrium thermodynamics. The conditional diffusion model presupposes a forward noisy process where Gaussian noise is incrementally introduced to the data sample. This process is defined as:

$$q\left(\boldsymbol{z}_t \mid \boldsymbol{z}_0\right) = \mathcal{N}\left(\boldsymbol{z}_t; \sqrt{\bar{\alpha}_t}\boldsymbol{z}_0, \left(1 - \bar{\alpha}_t\right)\mathbf{I}\right), \tag{1}$$

where $\bar{\alpha}_t := \prod_{s=0}^{t} \alpha_s = \prod_{s=0}^{t}\left(1 - \beta_s\right)$ and β_s represents the noise scheduler. This process transforms the ground truth variable z_0 to a noisy variable z_t by the randomly sampled timestep $t \in [0, T)$, where $z_t \sim \mathcal{N}(0, \mathbf{I})$. When training,

the reverse process denoising head $f_\theta(z_t, \boldsymbol{x}, t)$ is trained to predict z_0 from z_t under the guidance of condition $\boldsymbol{x}$ by minimizing the objective function. When testing, the prediction $\hat{z}_0$ is reconstructed from the Gaussian noise z_T with the model f_θ conditioned by $\boldsymbol{x}$, in a markovian way ($\boldsymbol{z}_T \rightarrow \hat{\boldsymbol{z}}_{T-1} \rightarrow \cdots \rightarrow \hat{\boldsymbol{z}}_{T-\Delta} \rightarrow \cdots \rightarrow \hat{z}_0, 0 < \Delta < T$), which can be formulated as:

$$p_\theta\left(\hat{\boldsymbol{z}}_{0:T-1} \mid \boldsymbol{x}\right) = p\left(\boldsymbol{z}_T\right) p_\theta\left(\hat{\boldsymbol{z}}_{T-1} \mid \boldsymbol{z}_T, \boldsymbol{x}\right) \prod_{t=1}^{T-1} p_\theta\left(\hat{\boldsymbol{z}}_{t-1} \mid \hat{\boldsymbol{z}}_t, \boldsymbol{x}\right). \qquad (2)$$

Multi-task supervision. As shown in Fig. 1, in addition to the main task segmentation, we introduce polyp classification and detection as auxiliary tasks into the conditional diffusion model. Classification and detection tasks effectively exploit high-level semantic information from objects in frames, thus providing contextual and discriminative information for segmentation tasks. Specifically, the transformer-based image encoder receives the raw target frame I_i as input to generate multi-scale spatial features $\mathcal{S}_j$, specifically with the size of $c_j \times \frac{H}{2^{j+1}} \times \frac{W}{2^{j+1}}, j \in \{1, 2, 3, 4\}$. Meanwhile, the transformer-based temporal encoder takes the previous frames $\{I_{i-\delta}, .., I_{i-1}\}$ as input to produce multi-scale temporal features $\mathcal{R}_j$. Subsequently, $\mathcal{S}_j$ and $\mathcal{R}_j$ are integrated and fed into a feature pyramid network to construct the latent spatiotemporal priors $\mathcal{H}_j$. To predict the segmentation mask from the noisy mask z_i^t, we introduce a CNN-based denoising head f_θ. The multi-level spatiotemporal features $\mathcal{H}_j$ are concatenated with the noisy variable z_i^t in the latent space to provide conditional guidance for denoising. Then they are passed into an MLP layer to unify the channel dimension. These unified features are up-sampled to the same resolution and added together. After that, an MLP layer is adopted to fuse these features. Finally, the fused feature goes through a 1×1 convolutional layer to predict the segmentation mask $\hat{M}_i^0$. The denoising head also outputs the classification score $\hat{Y}_{cls}$ and the bounding box $\hat{Y}_{box}$ for instance detection in the target frame I_i. The segmentation loss $\mathcal{L}_{seg}$ consisting of pixel-wise cross-entropy loss, mean square error and IoU loss is applied during training. And the two auxiliary tasks are optimized by cross-entropy loss. Thus, the overall objective function of MDM is:

$$\mathcal{L}_{MDM} = \lambda_{seg} \cdot \mathcal{L}_{seg} + \lambda_{cls} \cdot \mathcal{L}_{ce}(\hat{Y}_{cls}, Y_{cls}) + \lambda_{det} \cdot \mathcal{L}_{ce}(\hat{Y}_{box}, Y_{box}), \qquad (3)$$

where $\lambda_{seg}, \lambda_{cls}, \lambda_{det}$ are the balancing hyper-parameters, empirically set as 0.5, 0.05, 0.2, respectively. Y_{cls} is the category of polyps in I_i, *i.e.*, Low-grade adenoma, High-grade adenoma, Hyperplastic polyp, Traditional serrated adenoma, Sessile serrated lesion, Invasive carcinoma [10]; Y_{box} is (x_c, y_c, w, h) of a bounding box for the target lesion. The multi-task supervision strategy critically enhances the discriminative and generalizable ability of conditional diffusion models. More importantly, object-level annotations mitigate the class-imbalanced teaser in pixel-level segmentation of polyps due to the scarcity of lesions.

2.2   Temporal Reasoning Module (TRM)

To capture the dynamic appearance and motion cues in colonoscopy videos, we design a temporal reasoning module as displayed in Fig. 1. To be specific, we

formulate a self-supervised proxy task that reconstructs the target frame I_i from the previous frames $\{I_{i-\delta}, \cdots, I_{i-1}\}$. We feed the previous frames into the temporal encoder to extract their multi-scale features frame-by-frame. Then, they are averaged along the temporal axis and aggregated into multi-scale temporal features $\{\mathcal{R}_j \in \mathbb{R}^{c_j \times \frac{H}{2^{j+1}} \times \frac{W}{2^{j+1}}} | j \in [1, 4]\}$. The CNN-based reconstruction decoder consisting of MLP layers receives and integrates multi-scale temporal features to predict the next frame $\hat{I}_i$. By reasoning the unseen future, the temporal encoder explores the temporal dependency from video sequences and learns the spatiotemporal discriminative features to provide guidance for segmenting polyps with high camouflage.

To produce more realistic frames, we further introduce an adversarial self-supervision on the reconstructed results inspired by [26]. Our designed TRM is treated as the generator $\mathcal{G}$ while the Inception3 network pre-trained on the ImageNet is adopted as the discriminator $\mathcal{D}$ to distinguish the generated frame $\hat{I}_i$ from the real frame I_i. For the fixed parameter of $\mathcal{G}$, we minimize the adversarial loss $\mathcal{L}_{\mathcal{D}}$ to optimize the discriminator $\mathcal{D}$ as follows:

$$\mathcal{L}_{\mathcal{D}} = -\log(1 - D(\hat{I}_i)) - \log D(I_i). \tag{4}$$

Simultaneously, for the fixed parameters of $\mathcal{D}$, we expect the generator $\mathcal{G}$ can spoof $\mathcal{D}$ to generate the frame more like a real counterpart:

$$\mathcal{L}_{\mathcal{G}} = \mathcal{L}_{MSE}(\hat{I}_i, I_i) - \lambda_{adv} \cdot \log D(\hat{I}_i), \tag{5}$$

where $\mathcal{L}_{TRM} = \mathcal{L}_{\mathcal{G}}$ and λ_{adv} is set to 0.001 to balance the reconstruction term of mean square error and the adversarial term.

Total Loss. By combing the MDM loss and the TRM loss, we compute the total loss $\mathcal{L}_{total}$ of Diff-VPS as follows:

$$\mathcal{L}_{total} = \lambda_{MDM} \cdot \mathcal{L}_{MDM} + \lambda_{TRM} \cdot \mathcal{L}_{TRM}, \tag{6}$$

where the hyper-parameters are empirically set as $\lambda_{MDM} = 0.75, \lambda_{TRM} = 0.25$. More details of the training and inference are displayed in the *Supplementary*.

3 Experiment

3.1 Datasets and Experimental Setting

We evaluate our proposed Diff-VPS on the largest video polyp dataset SUN-SEG [10]. SUN-SEG has 113 colonoscopy videos, including 100 positive cases with 49,136 polyp frames, and 13 negative cases with 109,554 non-polyp frames. Following [10], we use the re-organized positive cases in SUN-SEG where 14,176 frames are used for training and 24,736 frames are used for testing. We conduct extensive experiments on two testing sub-datasets (i.e. Easy and Hard) based on data distribution and each sub-dataset is divided into seen and unseen parts, where seen represents the visible case that divides one case into two parts for training and testing (e.g. case 7_2 for training and case 7_1 for testing).

Table 1: Quantitative comparison of two testing datasets with seen colonoscopy scenarios. The best values are highlighted in bold. ↑ denotes the higher, the better.

Model	Publish	SUN-SEG-Easy				SUN-SEG-Hard			
		$S_\alpha \uparrow$	$E_\phi^{mn} \uparrow$	$F_\beta^w \uparrow$	Dice↑	$S_\alpha \uparrow$	$E_\phi^{mn} \uparrow$	$F_\beta^w \uparrow$	Dice↑
COSNet [16]	$TPAMI_{19}$	0.845	0.836	0.727	0.804	0.785	0.772	0.626	0.725
MAT [32]	TIP_{20}	0.879	0.861	0.731	0.833	0.840	0.821	0.652	0.776
PCSA [6]	$AAAI_{20}$	0.852	0.835	0.681	0.779	0.772	0.759	0.566	0.679
2/3D [17]	$MICCAI_{20}$	0.895	0.909	0.819	0.856	0.849	0.868	0.753	0.809
AMD [15]	$NeurIPS_{21}$	0.471	0.526	0.114	0.245	0.480	0.536	0.115	0.231
DCF [30]	$ICCV_{21}$	0.572	0.591	0.357	0.398	0.603	0.602	0.385	0.443
FSNet [9]	$ICCV_{21}$	0.890	0.895	0.818	0.873	0.848	0.859	0.755	0.828
PNSNet [8]	$MICCAI_{21}$	0.906	0.910	0.836	0.861	0.870	0.892	0.787	0.823
PNS+ [10]	MIR_{22}	0.917	0.924	0.848	0.888	0.887	0.929	0.806	0.855
Diff-VPS	$OURS_{24}$	**0.930**	**0.966**	**0.894**	**0.908**	**0.900**	**0.947**	**0.851**	**0.868**

Table 2: Quantitative comparison of two testing datasets with unseen colonoscopy scenarios. The best values are highlighted in bold. †denotes image segmentation method.

Model	Publish	SUN-SEG-Easy				SUN-SEG-Hard			
		$S_\alpha \uparrow$	$E_\phi^{mn} \uparrow$	$F_\beta^w \uparrow$	Dice↑	$S_\alpha \uparrow$	$E_\phi^{mn} \uparrow$	$F_\beta^w \uparrow$	Dice↑
COSNet [16]	$TPAMI_{19}$	0.654	0.600	0.431	0.596	0.670	0.627	0.443	0.606
PCSA [6]	$AAAI_{20}$	0.680	0.660	0.451	0.592	0.682	0.660	0.442	0.584
MAT [32]	TIP_{20}	0.770	0.737	0.575	0.710	0.785	0.755	0.578	0.712
2/3D [17]	$MICCAI_{20}$	0.786	0.777	0.652	0.722	0.786	0.775	0.634	0.706
ACSNet† [31]	$MICCAI_{20}$	0.782	0.779	0.642	0.782	0.783	0.787	0.636	0.708
SANet† [20]	$MICCAI_{21}$	0.720	0.745	0.566	0.649	0.706	0.743	0.526	0.598
PNSNet [8]	$MICCAI_{21}$	0.767	0.744	0.616	0.676	0.767	0.755	0.609	0.675
FSNet [9]	$ICCV_{21}$	0.725	0.695	0.551	0.702	0.724	0.694	0.541	0.699
UACANet† [13]	$ACMMM_{21}$	0.831	0.856	0.754	0.757	0.824	0.848	0.734	0.739
PNS+ [10]	MIR_{22}	0.806	0.798	0.676	0.756	0.797	0.793	0.653	0.737
AutoSAM [18]	$arXiv_{23}$	0.815	0.855	0.716	0.753	0.822	0.866	0.714	0.759
Diff-VPS	$OURS_{24}$	**0.828**	**0.883**	**0.748**	**0.767**	**0.823**	**0.886**	**0.733**	**0.764**

Implementation details. The proposed framework was implemented on the Pytorch platform and trained by two NVIDIA 3090 GPUs. We trained our model for 15 epochs with a batch size of 16. A video clip of 5 frames with a patch size of 224×224 was fed into the network. We adopted the Adam optimizer with the initial learning rate of 1×10^{-4}, which was decayed by a polynomial scheduler.

3.2   Comparison with State-of-the-Arts

Baselines. We compare our Diff-VPS with popular image- and video-level object/polyp segmentation methods on SUN-SEG-Easy and SUN-SEG-Hard, *i.e.*, COSNet [16], MAT [32], PCSA [6], 2/3D [17], AMD [15], DCF [30], FSNet [9], PNSNet [8], PNS+ [10], ACSNet [31], SANet [20], UACANet [13], and AutoSAM [18] in Tab. 1 and 2.

Metrics. For a comprehensive evaluation, we use four metrics following [10]: (1) S-measure (S_α), which evaluates region and object aware structural similarity; (2) Enhanced-alignment measure (E_ϕ^{mn}), which measures pixel level matching and image-level statistics; (3) Weighted F-measure (F_β^w), which amends the "Equal-importance flaw" in Dice and F_β. (4) Dice coefficient, which measures the similarity between two sets of data.
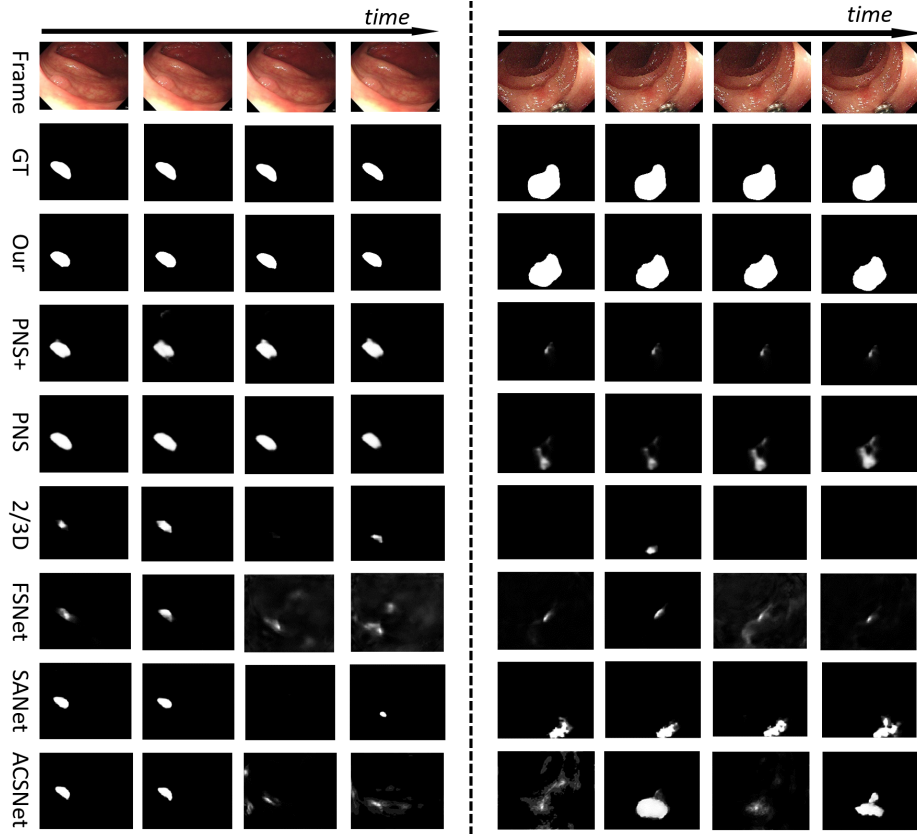
Fig. 2: Qualitative results on SUN-SEG. Left: a clip from easy-seen dataset case 75. Right: a clip from hard-unseen dataset case 36.

Quantitative Comparison. As shown in Tab. 1, our algorithm significantly outperforms the second-best method PNS+ on seen colonoscopy videos across four metrics by a remarkable margin (*e.g.*, for SUN-SEG-Hard, 1.3% in S_α, 1.8% in E_ϕ^{mn}, 4.6% in F_β^w, 1.2% in Dice). Also, in Tab. 2, our method is superior to the second-best method AutoSAM on unseen colonoscopy videos, especially for SUN-SEG-Easy (1.3% in S_α, 2.8% in E_ϕ^{mn}, 3.2% in F_β^w, 1.4% in Dice).

Qualitative Comparison. As displayed in Fig. 2, we showcase several polyp segmentation results of our Diff-VPS and the comparison methods on the SUN-SEG. Our method demonstrates robust performance in accurately locating and segmenting polyps across diverse and challenging scenarios. It has proficiency in managing polyps of varying dimensions, uniform areas, and textures. The versatility of our model allows it to navigate through intricate situations, providing reliable and precise segmentation results for polyps.

Table 3: Ablation studies on SUN-SEG. 'MDM' is multi-task diffusion model. 'TRM' is temporal reasoning module. 'ASS' is adversarial self-supervised strategy.

Models	MDM	TRM	ASS	Sub-dataset	Scenarios	S_α	E_ϕ^{mn}	F_β^w	Dice
#1				Easy	Seen	0.910	0.948	0.840	0.882
					Unseen	0.797	0.850	0.736	0.726
				Hard	Seen	0.870	0.935	0.841	0.837
					Unseen	0.802	0.859	0.721	0.717
#2	✓			Easy	Seen	0.920	0.965	0.899	0.896
					Unseen	0.808	0.863	0.748	0.740
				Hard	Seen	0.882	0.932	0.853	0.839
					Unseen	0.815	0.881	0.727	0.744
#3		✓		Easy	Seen	0.917	0.968	0.896	0.895
					Unseen	0.807	0.855	0.745	0.733
				Hard	Seen	0.872	0.932	0.832	0.833
					Unseen	0.807	0.870	0.726	0.725
#4		✓	✓	Easy	Seen	0.924	0.963	0.893	0.902
					Unseen	0.817	0.878	0.757	0.761
				Hard	Seen	0.874	0.937	0.844	0.847
					Unseen	0.823	0.883	0.744	0.750
Our method	✓	✓	✓	Easy	Seen	**0.930**	**0.966**	**0.894**	**0.908**
					Unseen	**0.828**	**0.883**	**0.748**	**0.767**
				Hard	Seen	**0.900**	**0.947**	**0.851**	**0.868**
					Unseen	**0.823**	**0.886**	**0.733**	**0.764**

3.3   Ablation Study

Extensive experiments are conducted on SUN-SEG in Tab. 3 to evaluate the effectiveness of our major components. To achieve this, we construct four baseline networks from our method. The first baseline (denoted as "#1") is to remove multi-task supervision and temporal reasoning module from our network. It means that "#1" equals the vanilla diffusion model [11]. Then, we introduce classification and detection supervisions into "#1" to construct "#2", and build "#3" by incorporating our TRM without adversarial self-supervision. Based on "#3", adversarial self-supervision is applied to construct "#4". Hence, "#4" equals removing the multi-task supervision from our full method.

Effectiveness of MDM. Observed from "#1" and "#2", the MDM strategy significantly improves the performance across all metrics. It indicates that category imbalance in pixel-level classification (i.e. segmentation) can be mitigated by applying detection and classification tasks. The categories of the detection and classification tasks are relatively balanced as they leverage object-level annotations for supervision. The multi-task supervision improves the generalizable ability of the model, which is also proved by incorporating MDM into #4 to construct our full method.

Effectiveness of TRM. As shown in Tab. 3, "#3" advances the performance of "#1" on seen and unseen videos, indicating that TRM plays a pivotal role in exploiting the temporal redundancy in colonoscopy videos. Compared to "#3", "#4" achieves a considerable improvement by adapting the model to the real data distribution and highlighting the perceptual quality of the samples. This demonstrates the effect of the generative adversarial self-supervision.

4   Conclusion

This paper presents the first diffusion-based multi-task framework for video polyp segmentation (VPS). The main idea is to improve the discriminative and generalizable ability of diffusion models by incorporating the multi-task supervision strategy. We effectively explore the temporal dependency by incorporating temporal reasoning module to mitigate the high camouflage of polyps in videos. Experiments demonstrate that our Diff-VPS is capable of state-of-the-art results on VPS benchmark dataset SUN-SEG, which can be a critical baseline for diffusion models on video object/lesion segmentation.

Disclosure of Interests The authors declare that they have no competing interests.

References

1. Ali, S., Espinel, Y., Jin, Y., Liu, P., Güttner, B., Zhang, X., Zhang, L., Dowrick, T., Clarkson, M.J., Xiao, S., et al.: An objective comparison of methods for augmented reality in laparoscopic liver resection by preoperative-to-intraoperative image fusion. arXiv preprint arXiv:2401.15753 (2024)
2. Amit, T., Shichrur, S., Shaharabany, T., Wolf, L.: Annotator consensus prediction for medical image segmentation with diffusion models. arXiv preprint arXiv:2306.09004 (2023)
3. Bernal, J., Sánchez, J., Vilarino, F.: Towards automatic polyp detection with a polyp appearance model. Pattern Recognition **45**(9), 3166–3182 (2012)
4. Chen, S., Sun, P., Song, Y., Luo, P.: Diffusiondet: Diffusion model for object detection. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 19830–19843 (2023)
5. Gao, S., Liu, X., Zeng, B., Xu, S., Li, Y., Luo, X., Liu, J., Zhen, X., Zhang, B.: Implicit diffusion models for continuous super-resolution. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 10021–10030 (2023)
6. Gu, Y., Wang, L., Wang, Z., Liu, Y., Cheng, M.M., Lu, S.P.: Pyramid constrained self-attention network for fast video salient object detection. In: Proceedings of the AAAI conference on artificial intelligence. vol. 34, pp. 10869–10876 (2020)
7. Ho, J., Jain, A., Abbeel, P.: Denoising diffusion probabilistic models. Advances in neural information processing systems **33**, 6840–6851 (2020)
8. Ji, G.P., Chou, Y.C., Fan, D.P., Chen, G., Fu, H., Jha, D., Shao, L.: Progressively normalized self-attention network for video polyp segmentation. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 142–152. Springer (2021)
9. Ji, G.P., Fu, K., Wu, Z., Fan, D.P., Shen, J., Shao, L.: Full-duplex strategy for video object segmentation. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 4922–4933 (2021)

10. Ji, G.P., Xiao, G., Chou, Y.C., Fan, D.P., Zhao, K., Chen, G., Van Gool, L.: Video polyp segmentation: A deep learning perspective. Machine Intelligence Research **19**(6), 531–549 (2022)
11. Ji, Y., Chen, Z., Xie, E., Hong, L., Liu, X., Liu, Z., Lu, T., Li, Z., Luo, P.: Ddp: Diffusion model for dense visual prediction. arXiv preprint arXiv:2303.17559 (2023)
12. Kawar, B., Elad, M., Ermon, S., Song, J.: Denoising diffusion restoration models. Advances in Neural Information Processing Systems **35**, 23593–23606 (2022)
13. Kim, T., Lee, H., Kim, D.: Uacanet: Uncertainty augmented context attention for polyp segmentation. In: Proceedings of the 29th ACM International Conference on Multimedia. pp. 2167–2175 (2021)
14. Li, H., Yang, Y., Chang, M., Chen, S., Feng, H., Xu, Z., Li, Q., Chen, Y.: Srdiff: Single image super-resolution with diffusion probabilistic models. Neurocomputing **479**, 47–59 (2022)
15. Liu, R., Wu, Z., Yu, S., Lin, S.: The emergence of objectness: Learning zero-shot segmentation from videos. Advances in Neural Information Processing Systems **34**, 13137–13152 (2021)
16. Lu, X., Wang, W., Ma, C., Shen, J., Shao, L., Porikli, F.: See more, know more: Unsupervised video object segmentation with co-attention siamese networks. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 3623–3632 (2019)
17. Puyal, J.G.B., Bhatia, K.K., Brandao, P., Ahmad, O.F., Toth, D., Kader, R., Lovat, L., Mountney, P., Stoyanov, D.: Endoscopic polyp segmentation using a hybrid 2d/3d cnn. In: Medical Image Computing and Computer Assisted Intervention–MICCAI 2020: 23rd International Conference, Lima, Peru, October 4–8, 2020, Proceedings, Part VI 23. pp. 295–305. Springer (2020)
18. Shaharabany, T., Dahan, A., Giryes, R., Wolf, L.: Autosam: Adapting sam to medical images by overloading the prompt encoder. arXiv preprint arXiv:2306.06370 (2023)
19. Tajbakhsh, N., Gurudu, S.R., Liang, J.: Automated polyp detection in colonoscopy videos using shape and context information. IEEE transactions on medical imaging **35**(2), 630–644 (2015)
20. Wei, J., Hu, Y., Zhang, R., Li, Z., Zhou, S.K., Cui, S.: Shallow attention network for polyp segmentation. In: MICCAI. pp. 699–708. Springer (2021)
21. Wolleb, J., Sandkühler, R., Bieder, F., Valmaggia, P., Cattin, P.C.: Diffusion models for implicit image segmentation ensembles. In: International Conference on Medical Imaging with Deep Learning. pp. 1336–1348. PMLR (2022)
22. Wu, J., Fu, R., Fang, H., Zhang, Y., Yang, Y., Xiong, H., Liu, H., Xu, Y.: Medsegdiff: Medical image segmentation with diffusion probabilistic model. In: Medical Imaging with Deep Learning. pp. 1623–1639. PMLR (2024)
23. Wu, Z., Lv, F., Chen, C., Hao, A., Li, S.: Colorectal polyp segmentation in the deep learning era: A comprehensive survey. arXiv preprint arXiv:2401.11734 (2024)
24. Wyatt, J., Leach, A., Schmon, S.M., Willcocks, C.G.: Anoddpm: Anomaly detection with denoising diffusion probabilistic models using simplex noise. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 650–656 (2022)
25. Xing, Z., Wan, L., Fu, H., Yang, G., Zhu, L.: Diff-unet: A diffusion embedded network for volumetric segmentation. arXiv preprint arXiv:2303.10326 (2023)
26. Xu, K., Wen, L., Li, G., Bo, L., Huang, Q.: Spatiotemporal cnn for video object segmentation. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 1379–1388 (2019)

27. Yang, Y., Fu, H., Aviles-Rivero, A.I., Schönlieb, C.B., Zhu, L.: Diffmic: Dual-guidance diffusion network for medical image classification. In: MICCAI. pp. 95–105. Springer (2023)
28. Yang, Y., Wang, S., Liu, L., Hickman, S., Gilbert, F.J., Schönlieb, C.B., Aviles-Rivero, A.I.: Mammodg: Generalisable deep learning breaks the limits of cross-domain multi-center breast cancer screening. arXiv preprint arXiv:2308.01057 (2023)
29. Yang, Y., Xing, Z., Zhu, L.: Vivim: a video vision mamba for medical video object segmentation. arXiv preprint arXiv:2401.14168 (2024)
30. Zhang, M., Liu, J., Wang, Y., Piao, Y., Yao, S., Ji, W., Li, J., Lu, H., Luo, Z.: Dynamic context-sensitive filtering network for video salient object detection. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 1553–1563 (2021)
31. Zhang, R., Li, G., Li, Z., Cui, S., Qian, D., Yu, Y.: Adaptive context selection for polyp segmentation. In: Medical Image Computing and Computer Assisted Intervention–MICCAI 2020: 23rd International Conference, Lima, Peru, October 4–8, 2020, Proceedings, Part VI 23. pp. 253–262. Springer (2020)
32. Zhou, T., Li, J., Wang, S., Tao, R., Shen, J.: Matnet: Motion-attentive transition network for zero-shot video object segmentation. IEEE Transactions on Image Processing **29**, 8326–8338 (2020)