# Highlighted Diffusion Model as Plug-In Priors for Polyp Segmentation

Yuhao Du, Yuncheng Jiang, Shuangyi Tan, *Student Member, IEEE*, Si-Qi Liu, Zhen Li, Guanbin Li, and Xiang Wan

***Abstract*—Automated polyp segmentation from colonoscopy images is crucial for colorectal cancer diagnosis. The accuracy of such segmentation, however, is challenged by two main factors. First, the variability in polyps' size, shape, and color, coupled with the scarcity of well-annotated data due to the need for specialized manual annotation, hampers the efficacy of existing deep learning methods. Second, concealed polyps often blend with adjacent intestinal tissues, leading to poor contrast that challenges segmentation models. Recently, diffusion models have been explored and adapted for polyp segmentation tasks. However, the significant domain gap between RGB-colonoscopy images and grayscale segmentation masks, along with the low efficiency of the diffusion generation process, hinders the practical implementation of these models. To mitigate these challenges, we introduce the Highlighted Diffusion Model Plus (HDM+), a two-stage polyp segmentation framework. This framework incorporates the Highlighted Diffusion Model (HDM) to provide explicit semantic guidance, thereby enhancing segmentation accuracy. In the initial stage, the HDM is trained using highlighted ground-truth data, which emphasizes polyp regions while suppressing the background in the images. This approach reduces the domain gap by focusing on the image itself rather than on the segmentation mask. In the subsequent second stage, we employ the highlighted features from the trained HDM's U-Net model as plug-in priors for polyp segmentation, rather than generating highlighted images, thereby increasing efficiency. Extensive experiments conducted on six polyp segmentation benchmarks demonstrate the effectiveness of our approach.**

***Index Terms*—Colonoscopy, diffusion models, polyp segmentation.**

## I. Introduction

Colorectal cancer (CRC) ranks as the third most frequently diagnosed cancer and the second leading cause of cancer-related deaths worldwide [1], [2]. The primary cause of CRC is associated with high-grade adenomatous polyps. Colonoscopy stands as the current gold standard for CRC screening and prevention due to its ability to reveal the location and characteristics of colorectal polyps. Research indicates that early colonoscopy has resulted in a 30% reduction in CRC incidence [3]. Therefore, timely detection and removal of these polyps by early screening is crucial for preventing CRC and reducing mortality rates. However, The bowel's intricate environment leads to a manual detection error rate exceeding 27% [4]. This emphasizes the urgent need for developing advanced computer-aided tools to support colonoscopists in colonoscopy procedures.

The rapid progress in artificial intelligence and computer vision has accelerated the development of such systems. In recent years, remarkable advancements have been made in designing efficient medical image segmentation algorithms to benefit clinicians in accurately detecting polyps. In particular, deep learning-based methods have shown promising performance in various scenarios, including assisting colonoscopy polyp detection and segmentation [5], [6], [7], [8], [9], [10], [11], [12]. Most segmentation models typically employ a UNet-based architecture, which consists of an encoder and a decoder constructed from multiple convolutional layers. More recently, there has been growing interest in utilizing Transformers [13] as fundamental building blocks to build attention-based segmentation models, which enables stronger capability for capturing global interactions [9].

Despite significant advancements in the field of automatic polyp segmentation research, there is still an observable gap between existing research and the development of clinically applicable computer-aided diagnosis systems. The main obstacles

Received 8 July 2024; revised 23 September 2024; accepted 16 October 2024. Date of publication 24 October 2024; date of current version 6 February 2025. This work was supported in part by the Shenzhen Science and Technology Program under Grant JCYJ20220818103001002, in part by the Guangdong Provincial Key Laboratory of Big Data Computing, in part by The Chinese University of Hong Kong, Shenzhen, in part by the Longgang District Special Funds for Science and Technology Innovation under Grant LGKCSDPT2023002, in part by the National Natural Science Foundation of China under Grant 62322608, and in part by the GuangDong Basic and Applied Basic Research Fund under Grant 2023A1515110706 and Grant 2024A1515010255. *(Yuhao Du and Yuncheng Jiang contributed equally to this work.) (Corresponding authors: Guanbin Li; Xiang Wan.)*

Yuhao Du, Yuncheng Jiang, Shuangyi Tan, and Xiang Wan are with the Shenzhen Research Institute of Big Data, The Chinese University of Hong Kong, Shenzhen 518172, China (e-mail: yuhaodu@link.cuhk.edu.cn; yunchengjiang@link.cuhk.edu.cn; shuangyitan@link.cuhk.edu.cn; wanxiang@sribd.cn).

Si-Qi Liu is with the Shenzhen Research Institute of Big Data, Shenzhen 518172, China (e-mail: siqiliu@sribd.cn).

Zhen Li is with the The Chinese University of Hong Kong, Shenzhen 518172, China (e-mail: lizhen@cuhk.edu.cn).

Guanbin Li is with the School of Computer Science and Engineering, Sun Yat-sen University, Guangzhou 510006, China, and also with Peng Cheng Laboratory, Shenzhen 518066, China (e-mail: liguanbin@mail.sysu.edu.cn).

The source code is available at https://github.com/DuYooho/HDM.

Digital Object Identifier 10.1109/JBHI.2024.3485767

2168-2194 © 2024 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission.
See https://www.ieee.org/publications/rights/index.html for more information.

Authorized licensed use limited to: The Chinese University of Hong Kong CUHK(Shenzhen). Downloaded on September 26,2025 at 07:38:58 UTC from IEEE Xplore. Restrictions apply.

are caused by two long-standing challenges. *1) Variety:* polyps often exhibit variations in shape, size, color, and texture, even when they belong to the same type. *2) Similarity:* the boundary between polyps and their surrounding tissues is usually unclear and blurred in colonoscopy imaging, lacking the necessary contrast for segmentation, especially in low-light conditions or unclean bowel preparation. These challenges introduce substantial uncertainty into the deep learning process, hampering the recognition capabilities of general segmentation models.

These challenges can be potentially mitigated by increasing the volume of well-annotated data and incorporating specific challenging samples for model training. However, the realm of medical data is often encumbered with privacy concerns, making it difficult to collect large-scale datasets with sufficient variety for network training. In addition, even when data is accessible, accurate manual annotations necessitate the involvement of experienced endoscopists for accurate annotating, which can be both costly and time-intensive. Recently, the Semantic Image Synthesis models exhibit pixel-level control ability over the image generation process and allow customization of the generated content [14], [15]. The Denoising Diffusion Probabilistic Models (DDPM) [16], [17] based methods have achieved promising performance in generating high-quality, realistic medical images [18], [19], [20]. Notably, Du et al. [21] have demonstrated that using binary masks as conditions to guide the diffusion models can generate realistic colonoscopy images with polyps in specified regions, where the generated images can improve the performance of downstream tasks like segmentation and detection. However, the issue of concealed polyps with unclear boundaries still remains and limits the effectiveness of existing segmentation models. In addition, Wu et al. [60] employed diffusion models to generate segmentation masks directly, aligning with the objective of polyp segmentation tasks. However, the significant domain gap between RGB-colonoscopy images and grayscale segmentation masks poses a challenge for achieving optimal performance as shown in the top part of Fig. 1. Furthermore, the low efficiency of the diffusion generation process hinders the practical application of the model in surgical settings.

In addressing the outlined limitations, we investigate the application of diffusion models to accentuate regions of interest and reduce the influence of non-essential background elements, thereby facilitating enhanced attention context for subsequent tasks. We introduce a novel diffusion-based segmentation framework, termed **H**ighlighted **D**iffusion **M**odel Plus (**HDM+**), designed for precise polyp segmentation. This framework integrates an innovative highlighted diffusion model (HDM) with a standard segmentation model. As illustrated in Fig. 1, the HDM+ framework adopts a two-stage training approach coupled with an end-to-end inference methodology. In the initial training phase, the highlighted diffusion model is refined to reduce the variance between the reconstructed images from the diffusion U-Net's outputs and the newly constructed ground-truth highlighted images, where these ground-truth images are generated by merging the original image with its binary mask. The second training stage diverges from the first by focusing on the direct generation of highlighted features, rather than constructing highlighted images using the pre-trained HDM. This strategy is shown to improve both the efficiency and effectiveness of training. The resulting highlighted features provide critical guidance for the further training of polyp segmentation models through an integrated prior fusion module. Specifically, we propose a weighting mechanism that employs normalized highlighted feature maps as weight matrices to enhance the original images, effectively diminishing noise in the highlighted features. Additionally, to address potential contextual misalignments caused by the direct addition of highlighted features to the original image, we introduce a cross-attention module. This module facilitates precise dot-wise attention coordination between the highlighted feature and the original images, ensuring alignment and thereby reducing irrelevant information, improving the model's recognition capabilities.

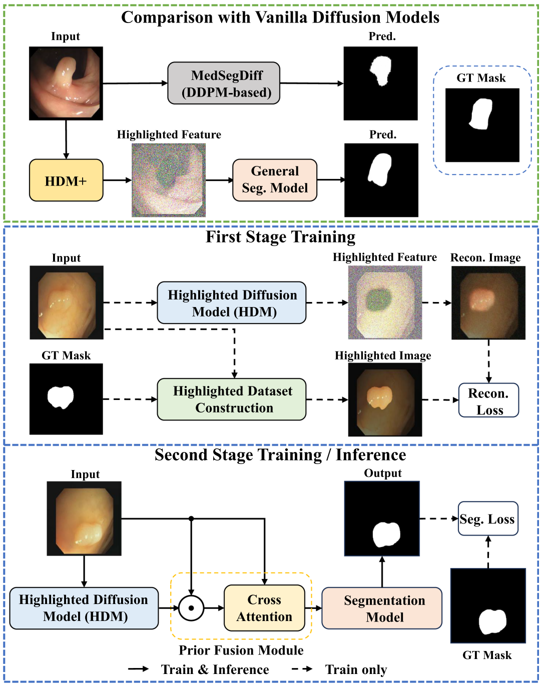

Fig. 1. Schematic overview of the proposed HDM+ pipeline for automatic polyp segmentation. The top section compares HDM+ with other diffusion-based polyp segmentation approaches. The middle section illustrates the workflow of the first training stage. The bottom section details the workflow of the second training stage and inference stage. Further details are provided in section III.

Our contributions are encapsulated as follows:

- *Highlighted Diffusion Model (HDM):* We propose a novel diffusion model based on DDPMs, specifically tailored for emphasizing potential polyp areas in colonoscopy images. Additionally, the HDM+ framework is introduced as an innovative segmentation guide. This represents the inaugural application of pre-trained diffusion models for explicit boundary guidance to augment segmentation performance.
- *Indirect Knowledge Transfer:* We adopt an approach that leverages highlighted features generated by the pre-trained HDM, instead of reconstructed images, as plug-in priors for directing the training of polyp segmentation models.

Authorized licensed use limited to: The Chinese University of Hong Kong CUHK(Shenzhen). Downloaded on September 26,2025 at 07:38:58 UTC from IEEE Xplore. Restrictions apply.

This method not only elevates segmentation accuracy but also streamlines the training process.

- *Prior Fusion Module:* A strategy is proposed for amalgamating original feature maps with the images within the prior fusion module to balance noise reduction and maintain background integrity. Furthermore, a cross-attention module is introduced to ensure effective alignment between highlighted features and original image features.
- *Extensive Experiments:* Extensive experiments are conducted across six polyp segmentation benchmarks, namely CVC-300, CVC-Clinicdb, Kvasir, CVC-Colondb, ETIS, and SUN-SEG. The results, both quantitative and qualitative, substantiate the enhanced segmentation performance achieved through our model across various pipelines.

## II. Related Work

### A. Automatic Polyp Segmentation

Automatic polyp segmentation in colonoscopy videos has been a widely studied topic in recent years. The primary techniques have gone through two periods of development. As in the early stage, polyp segmentation methods mainly rely on low-level image processing methods to extract hand-crafted features from color, shape, texture, and appearance to identify polyp from its surroundings [22], [23], [24], [25], [26]. For example, Tajbakhsh et al. [25] proposed an automated polyp detection method from colonoscopy videos, which fully utilizes context and shape to remove non-polyp structures and accurately locate polyps. Ameling et al. [27] adopted texture features, including grey-level-co-occurrence and local binary patterns, to achieve polyp segmentation. Further, Karkanis et al. [28] used the covariances of texture measurements to represent different polyp regions. However, the texture and shape of polyps highly differ in real-world applications, making the traditional methods suffer from unsatisfactory segmentation performance due to the limited expression ability of hand-crafted features.

Recently, the fully convolutional network (FCN) [29] has been widely applied for polyp segmentation tasks and has made significant progress. For instance, U-Net [30] is a famous structure for medical image segmentation, which consists of a downsampling path to capture context and an upsampling path to restore detailed information. Akbari et al. [31] proposed a polyp segmentation framework based on a fully connected CNN and adopted Otsu thresholding to extract the largest connected regions for segmenting polyp regions. Sun et al. [32] proposed an FCN-based polyp segmentation framework in which a dilated convolution is introduced to learn high-level semantic features without resolution reduction. Moreover, two variants of the U-Net architecture, including U-Net++ [33] and ResUNet++ [34], have been proposed and further improve original U-Net by dense connection, achieving promising segmentation performance. The methods mentioned above often focus on segmenting the entire polyp region, neglecting valuable boundary information. To address this, PsiNet [35] was introduced with three parallel decoders designed for contour extraction, mask prediction, and distance map estimation. SFA [36] implemented a boundary-sensitive loss to impose area boundary constraints, resulting in more precise predictions. SFA also explicitly applies an area-boundary constraint to supervise both polyp regions and boundaries. PraNet [7] proposed using reverse attention to locate polyp regions and implicitly refine object boundaries. SANet [8] introduced a color exchange operation to reduce color diversity in polyps and proposed a shallow attention module to select more useful shallow features. Polyp-PVT [9] was the first to apply transformer architecture for polyp segmentation, extracting long-term global contextual information. LDNet [37] incorporated a dynamic kernel generation and updating process to enhance the contrast between the polyp and background. Despite these advancements, the data-dependent nature of these complex models necessitates data expansion. To achieve greater efficiency in processing, EMTS-Net [38] presented an efficient framework for multitasking polyp segmentation and classification. However, it prioritizes efficiency at the expense of performance.

### B. Diffusion Models for Medical Image Segmentation

The diffusion models are a class of generative models that can be generally presented as a parameterized Markov chain, encompassing a forward diffusion process and a reverse denoising process [17], [39], [40], [41]. Each Markov step is modeled by a deep neural network that learns to inverse the diffusion process with a known Gaussian kernel. Ho et al. [17] first propose this idea by combining the diffusion probabilistic model with a score-based model that successfully achieves magnificent image generation. Afterward, a large number of researchers have devoted themselves to this area and developed more powerful methods [14], [42], [43], [44].

More recently, diffusion models have shown significant progress in assisting segmentation tasks in medical image analysis [18], [19], [45]. The denoising sampling process of the diffusion model has been utilized to generate an implicit ensemble of segmentation context that ultimately enhances the segmentation performance. Wu et al. [19] proposed the first general medical image segmentation framework based on DDPM by aggregating the information of images and predicted segmentation masks. Rahman et al. [18] leveraged the diffusion model to generate a distribution of segmentation masks that can present the conclusions of experts. Recently, the Semantic Image Synthesis models have emerged as a promising solution, which enables pixel-level control over the image generation process and allows customization of the generated content. For example, Wang et al. [14] were the first attempt to feed the semantic mask into the diffusion process as a condition to control the image generation. Guo et al. [45] used the predicted mask from the segmentation model as prior information to accelerate the diffusion process. Du et al. [21] have demonstrated that diffusion models possess the capability to synthesize high-quality data, which can be employed as additional training samples to enhance the performance of segmentation models. These common approaches involve either adapting the diffusion U-Net for downstream tasks directly or leveraging its generative power for large-scale image synthesis. However, challenges arise due to the domain gap

Authorized licensed use limited to: The Chinese University of Hong Kong CUHK(Shenzhen). Downloaded on September 26,2025 at 07:38:58 UTC from IEEE Xplore. Restrictions apply.

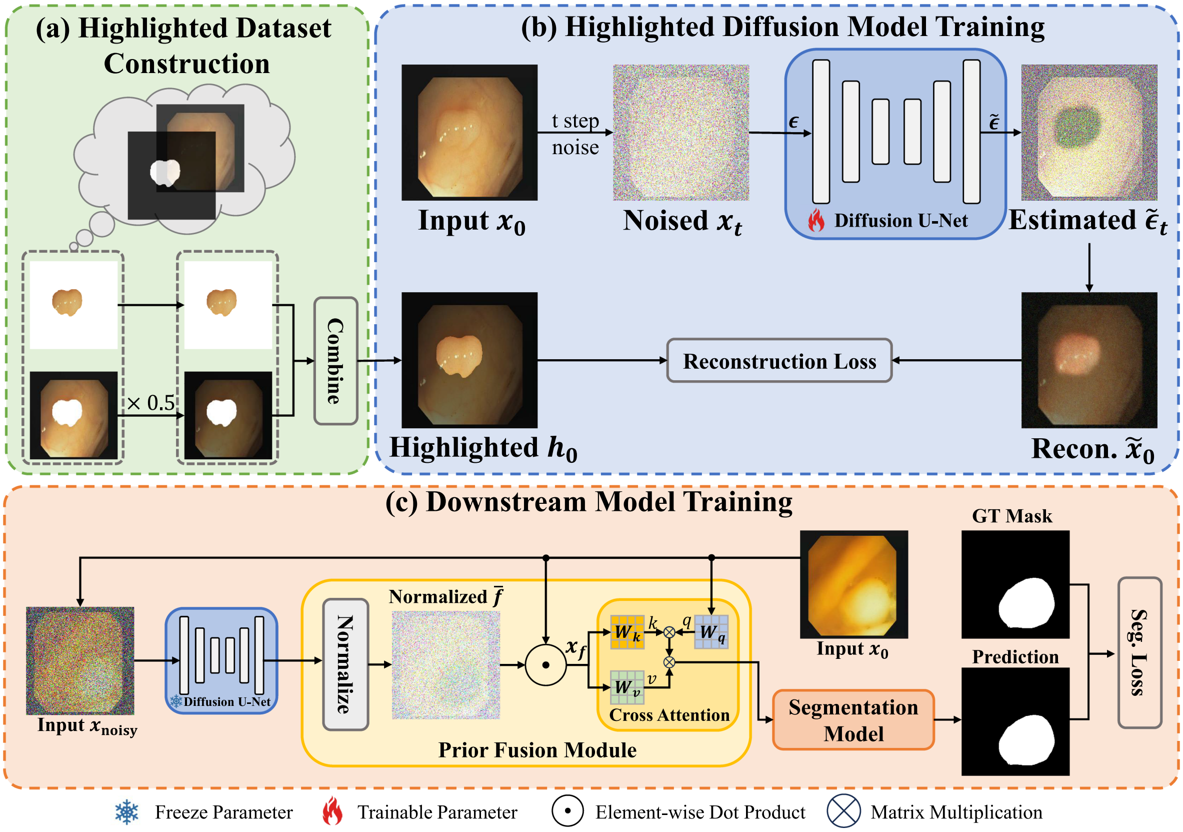

Fig. 2. Overview of the Highlighted Diffusion Model Plus (HDM+) framework, which adopts a two-stage training paradigm. In the first stage, (a) a specialized dataset is constructed, where polyp regions within the images are highlighted to enhance their features for better model training. (b) The diffusion U-Net model is trained to minimize the discrepancy between the reconstructed and original highlighted image. In stage two, (c) the pre-trained diffusion U-Net is utilized to generate highlight features for each corresponding input image. The images and features are fused in the prior fusion module, which further produces the prior context to train the downstream segmentation model.

between generative and discriminative tasks and the quality of data generated. In this work, we adopt a different approach, utilizing a specific diffusion model to assist in downstream tasks that bridge the domain gap and reduce bias from directly using generated images.

## III. Method

In this work, we introduce the Highlighted Diffusion Model (HDM) and a two-stage training framework, HDM+, specifically engineered for efficient and automated polyp segmentation, as depicted in Fig. 2. The HDM is adeptly designed to accentuate foreground elements (polyps) while attenuating background elements (intestinal walls) in sample generation. Once a pre-trained HDM is acquired, we freeze the weights and sample highlighted features from the original image. Then, these sampled features are integrated with the original images in the prior fusion module. This process yields prior contexts that are pivotal for guiding the polyp segmentation task. Subsequent sections will further explore the core of our methodology. Section III-A provides a concise introduction to Denoising Diffusion Probabilistic Models (DDPMs). The comprehensive training pipeline is detailed in Section III-B, focusing on the initial stage of diffusion model training, and in Section III-C, which addresses the second stage of segmentation model training.

### A. Preliminaries

The recent Denoising Diffusion Probabilistic Models (DDPMs) [16], [17] are classes of deep generative models. It has the forward and reverse process, where the forward process is a parameterized Markov Chain that gradually adds Gaussian noise to the original input. The forward process can be formulated as the joint distribution $q(\mathbf{x}_{1:T} \mid \mathbf{x}_0)$:

$$q\left(\mathbf{x}_{1:T} \mid \mathbf{x}_0\right) := \prod_{t=1}^{T} q\left(\mathbf{x}_t \mid \mathbf{x}_{t-1}\right),$$

$$q\left(\mathbf{x}_t \mid \mathbf{x}_{t-1}\right) := \mathcal{N}\left(\mathbf{x}_t; \sqrt{1-\beta_t}\mathbf{x}_{t-1}, \beta_t \mathbf{I}\right), \qquad (1)$$

where $q(\mathbf{x}_0)$ is the original data distribution with $\mathbf{x}_0 \sim q(\mathbf{x}_0)$, $\mathbf{x}_{1:T}$ are latents with the same dimension of $\mathbf{x}_0$ and $\beta_t$ is a variance schedule, which can be learned by parameterization [46] or held constant as hyper-parameters. A notable property of the forward process is that it admits sampling $\mathbf{x}_t$ at any timestep $t$:

$$q\left(\mathbf{x}_t \mid \mathbf{x}_0\right) := \mathcal{N}\left(\mathbf{x}_t; \sqrt{\bar{\alpha}_t}\mathbf{x}_0, (1-\bar{\alpha}_t))\mathbf{I}\right), \qquad (2)$$

where $\alpha_t := 1 - \beta_t$, and $\bar{\alpha}_t := \prod_{s=1}^{t} \alpha_x$ is the cumulative product of $\alpha_t$ values from $t = 1$ to the current time step $t$. In other words, $\mathbf{x}_t$ can be expressed using the closed form:

$$\mathbf{x}_t = \sqrt{\bar{\alpha}_t}\mathbf{x}_0 + \sqrt{1-\bar{\alpha}_t}\epsilon_t, \qquad (3)$$

Authorized licensed use limited to: The Chinese University of Hong Kong CUHK(Shenzhen). Downloaded on September 26,2025 at 07:38:58 UTC from IEEE Xplore. Restrictions apply.

where $\boldsymbol{\epsilon}_t \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$ are Gaussian noises at timestep $t$. While the reverse process, i.e., the diffusion process, is aiming to learn a model to reverse the forward process that reconstructs the original input data, which is defined as:

$$p_\theta\left(\mathbf{x}_{0:T}\right) := p\left(\mathbf{x}_T\right) \prod_{t=1}^{T} p_\theta\left(\mathbf{x}_{t-1} \mid \mathbf{x}_t\right),$$
$$p_\theta\left(\mathbf{x}_{t-1} \mid \mathbf{x}_t\right) := \mathcal{N}\left(\mathbf{x}_{t-1}; \boldsymbol{\mu}_\theta\left(\mathbf{x}_t, t\right), \sigma_t^2 \mathbf{I}\right), \tag{4}$$

where $p(\mathbf{x}_T)$ is the noised Gaussian transition from the forward process at timestep $T$. In this case, we only need to use deep-learning models to represent $\boldsymbol{\mu}_\theta$ with $\theta$ as the model parameters. According to the original paper [17], the loss function can be simplified as:

$$\mathcal{L}_{\text{simple}} := \mathbb{E}_{t, \mathbf{x}_t, \boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})}\left[\left\|\boldsymbol{\epsilon} - \boldsymbol{\epsilon}_\theta\left(\mathbf{x}_t, t\right)\right\|^2\right]. \tag{5}$$

Thus, instead of training the model $\boldsymbol{\mu}_\theta$ to predict $\tilde{\boldsymbol{\mu}}_t$, we can train the model $\boldsymbol{\epsilon}_\theta$ to predict $\tilde{\boldsymbol{\epsilon}}$, which is easier for parameterization and learning.

## B. Highlighted Diffusion Model

*1) Overview:* Conventional diffusion models typically utilize a simplified training objective, which revolves around minimizing the differences between the reconstructed noisy output generated by the diffusion U-Net model and the input Gaussian noise. However, our primary objective is to reconstruct images that exhibit a highlight on the polyp regions while simultaneously darkening the background regions. Consequently, the conventional loss function used in vanilla diffusion models proves to be unsuitable for our specific goal. In light of this, we propose an innovative variant of DDPMs to redefine the diffusion process.

In our approach, we steer the reconstruction process by comparing the reconstruction output initiated from the noisy output to the synthesized highlighted ground truths (detailed in section III-B-2). This strategy enables us to train a diffusion model capable of generating images in which the foreground regions are highlighted while the background elements are darkened. Through this alternative supervision method, we align our diffusion model with the specific goal of improving boundary visibility and emphasizing foreground subjects.

*2) Highlighted Dataset Construction:* In this part, we will describe the process of constructing the highlighted ground truths. Given an input image $\mathbf{x}_0 \in \mathbb{R}^{H \times W \times C}$ and the corresponding mask $\mathbf{c}_0 \in \mathbb{R}^{H \times W}$, the constructed highlighted image $\mathbf{h}_0 \in \mathbb{R}^{H \times W \times C}$ can be formulated as follows:

$$\begin{aligned} \mathbf{h}_0 &= f(\mathbf{x}_0), \\ f(\mathbf{x}) &:= \begin{cases} \mathbf{x}_{(h,w,c)} & , \mathbf{c}_{(h,w)} = 255 \\ \alpha \cdot \mathbf{x}_{(h,w,c)} & , \mathbf{c}_{(h,w)} = 0 \end{cases}, \end{aligned} \tag{6}$$

where $f(\cdot)$ represents the construction function controlling each pixel value of $\mathbf{h}_0$ at $(h, w, c)$, and $\alpha \in [0, 1]$ is a hyperparameter representing the scale factor.

**Algorithm 1:** One Training Iteration of HDM.

```
Input: $t \sim \text{Uniform}(\{1, \ldots, T\})$, $\mathbf{x}_0 \sim q(\mathbf{x}_0)$,
  $\boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$
Output: $\tilde{\mathbf{x}}_0$
  $\mathbf{x}_t = \sqrt{\bar{\alpha}_t}\mathbf{x}_0 + \sqrt{1 - \bar{\alpha}_t}\boldsymbol{\epsilon}$
  $\tilde{\boldsymbol{\epsilon}}_t = \boldsymbol{\epsilon}_\theta(\mathbf{x}_t, t)$
  $\tilde{\mathbf{x}}_0 = \frac{1}{\sqrt{\bar{\alpha}_t}}(\mathbf{x}_t - \sqrt{1 - \bar{\alpha}_t}\tilde{\boldsymbol{\epsilon}}_t)$
  $\mathbf{h}_0 = f(\mathbf{x}_0)$
  Take gradient descent step on $\nabla_\theta \|\mathbf{h}_0 - \tilde{\mathbf{x}}_0\|$
```

*3) Training Objective:* After generating the highlighted images, we proceed to compile a specialized training dataset consisting of triplet pairs. Each triplet is composed of the original input image, its corresponding segmentation mask, and the highlighted image. The Highlighted Diffusion Model (HDM) developed in this work is specifically oriented towards enhancing the polyp segmentation task. In line with this objective, during the testing phase, we deliberately exclude the use of the segmentation mask. Consequently, for HDM training, the focus is directed solely to the original and its highlighted counterpart as input. The HDM adheres to the conventional Denoising Diffusion Probabilistic Model (DDPM) protocol, characterized by unconditional training settings. Within this framework, during the training phase, the highlighted image is employed as the ground truth to guide the reconstruction process. In the sampling process, the original image alone is utilized to generate the highlight guidance, maintaining consistency with our established methodology. From (3), we observe that $\mathbf{x}_t$ can be directly obtained in a single step from the initial input $\mathbf{x}_0$. Similarly, the reverse process can be computed by estimating $\tilde{\mathbf{x}}_0$ from $\mathbf{x}_t$ using a closed-form expression:

$$\tilde{\mathbf{x}}_0 := g(\mathbf{x}_t, \tilde{\boldsymbol{\epsilon}}_t) = \frac{1}{\sqrt{\bar{\alpha}_t}}\left(\mathbf{x}_t - \sqrt{1 - \bar{\alpha}_t}\tilde{\boldsymbol{\epsilon}}_t\right). \tag{7}$$

This implies the feasibility of reconstructing an image, denoted as $\tilde{\mathbf{x}}_0$, from any given intermediate state $\mathbf{x}_t$ during the diffusion sequence. In our training framework, the fundamental objective is to reduce the disparity between the reconstructed image $\tilde{\mathbf{x}}_0$, which is generated from the model output $\tilde{\boldsymbol{\epsilon}}_t = \boldsymbol{\epsilon}_\theta(\mathbf{x}_t, t)$, and the designated highlighted image $\mathbf{h}_0$. To quantify this optimization process, we formulate the reconstruction loss function:

$$\mathcal{L}_{\text{recon.}} := \mathbb{E}_{t, \mathbf{x}_t, \mathbf{h}_0}\left[\left\|\mathbf{h}_0 - g(\mathbf{x}_t, \boldsymbol{\epsilon}_\theta\left(\mathbf{x}_t, t\right))\right\|\right], \tag{8}$$

where $\boldsymbol{\epsilon}(\cdot)$ represents the diffusion U-Net model with learnable parameters $\theta$, and $g(\cdot)$ refers to (7). This loss function guides the HDM to learn the intricacies of reconstructed images while preserving the prominence of foreground elements. The detailed procedure of one training iteration of HDM is shown in Algorithm 1.

*4) Highlighted Feature Extraction:* During the sampling process of diffusion models, the conventional strategy typically initiates with pure Gaussian noise that is independent of any input images. However, this is not suitable in the context of our framework. The primary goal of the HDM is to extract highlighted features that align with the distinct foreground and

Authorized licensed use limited to: The Chinese University of Hong Kong CUHK(Shenzhen). Downloaded on September 26,2025 at 07:38:58 UTC from IEEE Xplore. Restrictions apply.

background characteristics of the input images. Therefore, the conflict with the conventional strategy arises from initiating the sampling process with pure Gaussian noise, which fails to generate any useful highlighted regions due to its inability to acknowledge the inherent correlation between these samples and the images. An intuitive approach to address this issue is to employ the original images as an extra condition for the sampling process. This modification seems promising, as it seeks to align the generated features more closely with the underlying image content. However, the implementation introduces additional complexity to the models, increasing the computational demands. Moreover, the implicit conditions derived from the original images may not be accurately learned, leading to suboptimal feature alignment and introducing unwanted noise into the downstream model training.

Given these considerations, we present an alternative strategy to tackle this challenge. Our proposed approach maintains the established practice of unconditional training for the diffusion process. However, during the reverse sampling process, we depart from convention by initiating with the noised images instead of the pure Gaussian noise. As demonstrated in [47], this strategic alteration leverages the inherent information contained within the images, mitigating the misalignment issue while preserving the core benefits of unconditional training. Furthermore, it is worth noting that the U-Net architecture has garnered significant attention and recognition within the landscape of contemporary diffusion models. This architecture, well-suited for parameterizing the diffusion denoising process, has proven to be particularly effective. Prior studies [47] have demonstrated that the output from the final layer of the U-Net architecture encapsulates substantial semantic information. In light of these findings, our research centers on harnessing this information-rich output as a foundational substrate for shaping our feature representations.

Specifically, during the sampling process, we initiate by introducing noise to the images in accordance with a predetermined schedule, parameterized by $t = 100$. This noise injection process is mathematically defined as follows:

$$\mathbf{x}_{\text{noisy}} := \sqrt{\bar{\alpha}_t}\mathbf{x}_0 + \sqrt{1-\bar{\alpha}_t}\boldsymbol{\epsilon}_t. \tag{9}$$

Subsequently, we employ these "noised" images, denoted as $\mathbf{x}_{\text{noisy}}$, as the input to the U-Net model. The U-Net model processes these images and yields a set of features, denoted as $\mathbf{f} := \boldsymbol{\epsilon}_\theta(\mathbf{x}_{\text{noisy}}, t)$, which encapsulate relevant information regarding the image content and noise introduced during the sampling process.

### C. HDM+: Plugging-In Process for Polyp Segmentation

We further introduce the HDM+ segmentation framework that employs the highlighted features generated by the HDM as plug-in priors to assist in the downstream task of polyp segmentation, where the pre-trained HDM model is used to extract feature maps that exhibit distinct representations of foreground and background regions. These feature maps are then fed into polyp segmentation models as plug-in priors to enhance their performance.

Specifically, we designed a prior fusion module to effectively utilize the generated highlights feature maps. First, to maintain essential background information within these feature maps, we developed a specialized weighting mechanism. This mechanism involves normalizing the features, represented as $\bar{\mathbf{f}}$ within the [0,1] range. These normalized features serve as weighting factors for the original images $\mathbf{x}_0$. By applying these weights, we ensure that both the highlighted foreground semantic information and the crucial background information are preserved in the final feature maps. This approach allows for the full utilization of the contrast information inherent in the generated features, effectively maintaining the integrity of the background regions without introducing unexpected noise. The weighting mechanism can be represented as:

$$\mathbf{x}_f = \bar{\mathbf{f}} \circ \mathbf{x}_0, \tag{10}$$

where the symbol $\circ$ denotes element-wise multiplication, and $\mathbf{x}_f$ signifies the actual prior guidance fed into the segmentation models.

Subsequently, the highlighted features, serving as prior guidance, are fused with the original image to generate the input data for training the segmentation model. Initially, we explore an intuitive approach of concatenating the prior guidance to the original images through the channel dimension. However, we observed that there is a critical challenge associated with this strategy. As detailed in section IV-C, this straightforward approach can lead to misalignment between the highlighted features and the images, potentially diminishing the performance of downstream models. To overcome this, we eschew simple concatenation in favor of integrating a cross-attention module. This module is designed to mitigate the misalignment issues inherent in the fusion of feature maps and images. The cross-attention mechanism operates by initially transforming both inputs into linear representations using matrices $W_Q$, $W_K$, and $W_V$. The query matrix $Q$, derived from these transformations, is then fused with the key matrix $K$. Subsequently, this fused result undergoes further fusion with the value matrix $V$. This process results in a deep fusion of the two inputs, thereby eliminating the need for alignment and effectively addressing the misalignment concerns. The operation of the cross-attention module is mathematically formulated as follows:

$$\text{CA}(\mathbf{x}_1, \mathbf{x}_2) := \text{softmax}\left(\frac{W_Q\mathbf{x}_1 \otimes (W_K\mathbf{x}_2)^T}{\sqrt{d_k}}\right) \otimes W_V\mathbf{x}_2, \tag{11}$$

where $\text{CA}(\cdot)$ represents the cross-attention module, and $\otimes$ represents the matrix multiplication operation.

Above all, the overall training process of downstream tasks can be represented as:

$$\tilde{\mathbf{c}}_0 = \text{Seg}(\text{CA}(\mathbf{x}_0, \mathbf{x}_f)), \tag{12}$$

where $\tilde{\mathbf{c}}_0$ represents the segmentation prediction, and $\text{Seg}(\cdot)$ represents the segmentation model. The overall segmentation loss function can be formulated as:

$$\mathcal{L}_{\text{seg.}} = \mathcal{L}^w_{\text{IoU}}(\tilde{\mathbf{c}}_0, \mathbf{c}_0) + \mathcal{L}^w_{\text{BCE}}(\tilde{\mathbf{c}}_0, \mathbf{c}_0), \tag{13}$$

Authorized licensed use limited to: The Chinese University of Hong Kong CUHK(Shenzhen). Downloaded on September 26,2025 at 07:38:58 UTC from IEEE Xplore. Restrictions apply.

TABLE I
COMPARISONS OF DIFFERENT SETTINGS APPLIED ON FOUR POLYP SEGMENTATION BASELINES

| Methods | EndoScene | | ClinicDB | | Kvasir | | ColonDB | | ETIS | | **Overall** | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | mDice | mIoU | mDice | mIoU | mDice | mIoU | mDice | mIoU | mDice | mIoU | mDice | mIoU |
| MedSegDiff | 39.5 | 32.1 | 63.3 | 56.1 | 64.6 | 54.6 | 34.4 | 27.3 | 28.0 | 23.1 | 39.2 | 32.3 |
| PraNet | 87.1 | 79.7 | 89.9 | 84.9 | **89.8** | 84.0 | 70.9 | 64.0 | 62.8 | 56.7 | 74.0 | 67.5 |
| **HDM+PraNet** | **88.1** | **81.0** | **92.2** | **87.0** | 89.5 | **84.1** | **75.6** | **67.7** | **65.9** | **58.8** | **77.2** | **70.1** |
| FCBFormer | 88.0 | 81.3 | **89.7** | **84.2** | **91.9** | **86.7** | **78.4** | **70.0** | 74.9 | 66.4 | 80.8 | 73.2 |
| **HDM+FCBFormer** | **88.7** | **81.9** | 89.1 | 83.6 | 90.4 | 85.1 | 78.2 | 69.8 | **77.5** | **68.9** | **81.2** | **73.5** |
| Polyp-PVT | **90.0** | **83.3** | **93.7** | **88.9** | 91.7 | 86.4 | 80.8 | 72.7 | **78.7** | **70.6** | 83.3 | 76.0 |
| **HDM+Polyp-PVT** | 87.9 | 80.4 | 93.1 | 88.2 | **92.4** | **87.4** | **82.7** | **74.3** | 78.0 | 70.0 | **84.0** | **76.4** |

where $\mathbf{c}_0$ represents the ground truth, $\mathcal{L}^w_{\mathrm{IoU}}(\cdot)$ and $\mathcal{L}^w_{\mathrm{BCE}}(\cdot)$ are the weighted intersection over union (IoU) loss [48] and weighted binary cross entropy (BCE) loss [48].

## IV. Experiments

### A. Experimental Settings

*1) Dataset:* For the image-segmentation dataset, we performed experiments on five publicly available polyp segmentation datasets: EndoScene [49], CVC-ClincDB/CVC-612 [50], CVC-ColonDB [25], ETIS [51], and Kvasir [52]. In line with the standards set by PraNet [7] and Polyp-PVT [9], we used 1,450 images from the Kvasir and CVC-ClinicDB datasets for training. The test sets of these five datasets were combined to form our validation and test set for evaluation. For the video-segmentation dataset, we utilized the SUN-SEG [53] dataset. Specifically, we treated all frames in the dataset as individual images for input to the diffusion model. We used the entire training dataset for training and selected only the "unseen" dataset, comprising both easy and hard cases, for our validation and test sets.

*2) Compared Models:* We compare our *HDM+* with three SOTA medical image segmentation methods: PraNet [7], FCBFormer [54], and Polyp-PVT [9]. Also, we compare our *HDM+* with other diffusion-based models that generate masks directly, i.e., MedSegDiff [19]. For the video polyp segmentation, we chose the FCBFormer and Polyp-PVT models as the baseline methods.

*3) Metrics:* For image segmentation, we evaluated their performances using the widely-used mean Intersection over Union (mIoU) and mean Dice (mDice) metrics. For video polyp segmentation, we assessed their performance using various metrics, including structure measure ($S_\alpha$) [55], enhanced-alignment measure ($E_\phi$) [56], F-measure ($F_\beta$) [57], weighted F-measure ($F^\omega_\beta$) [58], mean Dice, mean IoU, and mean Specificity (mSpe). For hyperparameter settings, we followed the official settings of Polyp-PVT.

*4) Implementation Details:* All the training images were resized to a resolution of 352 × 352. Our experiments were conducted on a single NVIDIA A100 GPU using the AdamW optimizer and the LambdaLR scheduler. The training settings included a base learning rate of $1e-6$ and a batch size of 6. For the image-segmentation dataset, we trained the diffusion model for a maximum of 100 epochs. In the case of the video segmentation dataset, the model was trained for a maximum of 40 epochs.

### B. Quantitative Comparisons

Table I and Table II present a comparative analysis of our proposed approach against each corresponding baseline method. In the case of the image polyp segmentation, as depicted in Table I, MedSegDiff fails to predict much useful semantic information due to the large domain gap. However, our HDM combined with other segmentation models generally outperforms the corresponding baselines. Specifically, compared to PraNet, our method achieves a 3.2% and 2.6% improvement in mean Dice and mean IoU scores, respectively. In comparison to FCBFormer, our method yields a 0.4% and 0.3% increase. Against PolypPVT, our method achieves a 0.7% and 0.4% enhancement.

Moreover, for video polyp segmentation, as indicated in Table II, our approach exhibits superior performance across most of the metrics compared to the two baseline methods. Specifically, for FCBFormer, the model enhanced with our proposed highlighted features outperforms the baseline in all the metrics. For Polyp-PVT, our HDM+ outperforms the baseline method in terms of mean E-measure, mean F-measure, weighted F-measure, mean Dice, and mean Specificity while achieving comparable S-measure and mean IoU results. These results demonstrate that our model has a strong learning ability to effectively segment polyps.

### C. Ablation Study

In this section, we conducted ablation studies to explore the effectiveness of each component in our proposed prior fusion module. Results are presented in Table III and Table IV. In Table III, we compare the performances between three different feature processing methods. One directly uses normalized features computed by the HDM as the prior for the segmentation model, denoted as ($\bar{\mathbf{f}}$), another using $\bar{\mathbf{f}}$ added with the source image, denoted as ($\bar{\mathbf{f}} + \mathbf{x}_0$), while the other employs our proposed weighting mechanism, denoted as ($\mathbf{x}_f$). Our methods outperform the two base methods across the overall meanDice and meanIoU, demonstrating the effectiveness of the proposed highlighted feature processing process.

Authorized licensed use limited to: The Chinese University of Hong Kong CUHK(Shenzhen). Downloaded on September 26,2025 at 07:38:58 UTC from IEEE Xplore. Restrictions apply.

TABLE II
COMPARISONS OF DIFFERENT SETTINGS OF TWO POLYP SEGMENTATION BASELINES ON THE SUN-SEG DATASET

| Methods | $S_\alpha$ | $E_\phi$ | $F_\beta$ | $F_\beta^\omega$ | mDice | mIoU | mSpe |
|---|---|---|---|---|---|---|---|
| FCBFormer | 81.7 | 87.2 | 79.2 | 73.9 | 74.8 | 65.9 | 94.7 |
| **HDM+FCBFormer** | **82.7** | **88.3** | **79.8** | **74.6** | **75.9** | **66.8** | **95.4** |
| Polyp-PVT | **83.7** | 87.8 | 78.6 | 74.4 | 76.3 | **68.3** | 93.5 |
| **HDM+Polyp-PVT** | 83.2 | **88.6** | **80.4** | **75.3** | **76.5** | 67.9 | **94.7** |

TABLE III
ABLATION STUDY OF DIFFERENT FEATURE PROCESSING STRATEGIES

| Methods | PraNet | | | Polyp-PVT | | |
|---|---|---|---|---|---|---|
| | $\bar{\mathbf{f}}$ | $\bar{\mathbf{f}}+\mathbf{x}_0$ | $\mathbf{x}_f$ | $\bar{\mathbf{f}}$ | $\bar{\mathbf{f}}+\mathbf{x}_0$ | $\mathbf{x}_f$ |
| meanDice | 76.1 | 76.9 | **77.2** | 83.4 | 83.0 | **84.0** |
| meanIoU | 69.5 | 69.8 | **70.1** | 76.0 | 75.6 | **76.4** |

$\bar{\mathbf{f}}$ means using the normalized U-Net output directly as the guidance for the segmentation models, $\bar{\mathbf{f}}+\mathbf{x}_0$ means using the normalized U-Net output added with the source image $\mathbf{x}_0$, and $\mathbf{x}_f$ means using our proposed feature processing manner as described in Section III-B4.

TABLE IV
COMPARATIVE EVALUATION OF CONCATENATE (CONCAT) AND CROSS-ATTENTION (CROSS_ATTN) METHODS IN PRANET AND POLYP-PVT MODELS, USING MEAN-DICE AND MEAN-IOU METRICS ACROSS FIVE DISTINCT DATASETS TO ASSESS THE EFFECTIVENESS OF THESE METHODS IN VARIOUS CONTEXTS

| Methods | | PraNet | | Polyp-PVT | |
|---|---|---|---|---|---|
| | | concat | cross_attn | concat | cross_attn |
| EndoScene | mDice | 78.3 | **88.1** | 80.8 | **87.9** |
| | mIoU | 71.2 | **81.0** | 72.6 | **80.4** |
| ClinicDB | mDice | 86.1 | **92.2** | 88.8 | **93.1** |
| | mIoU | 80.7 | **87.0** | 83.0 | **88.2** |
| Kvasir | mDice | 83.1 | **89.5** | 83.3 | **92.4** |
| | mIoU | 75.6 | **84.1** | 75.5 | **87.4** |
| ColonDB | mDice | 65.6 | **75.6** | 70.6 | **82.7** |
| | mIoU | 58.3 | **67.7** | 62.2 | **74.3** |
| ETIS | mDice | 46.2 | **65.9** | 59.6 | **78.0** |
| | mIoU | 40.9 | **58.8** | 52.0 | **70.0** |
| **Overall** | mDice | 65.6 | **77.2** | 71.7 | **84.0** |
| | mIoU | 58.9 | **70.1** | 63.7 | **76.4** |

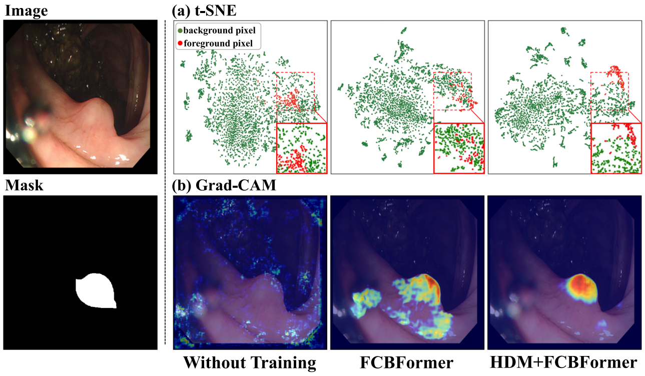

Fig. 3. Comparative visualization of t-SNE and Grad-CAM plots showing output features and layers in the untrained model, FCBFormer training, and enhanced HDM+FCBFormer training.

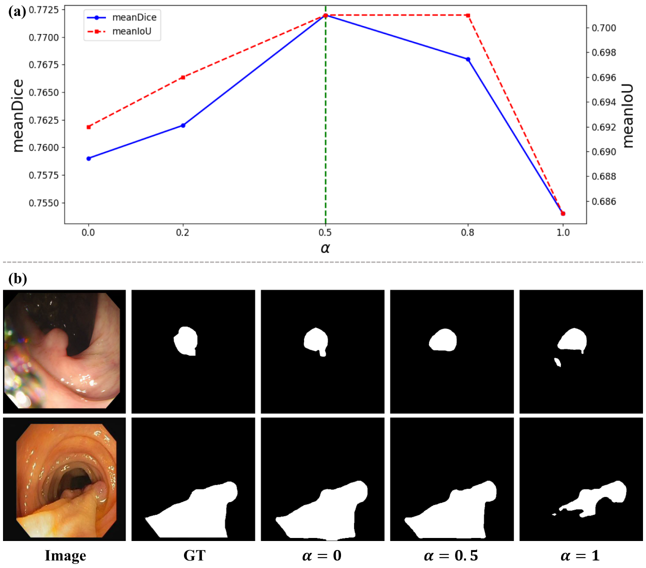

Fig. 4. Quantitative (a) and qualitative (b) comparisons of segmentation results using different scale factors ($\alpha$).

Similarly, we test the effectiveness of the cross-attention module by removing it from the prior fusion module and replacing it with a vanilla concatenate operation. As shown in Table IV, The cross-attention outperforms the concatenate in terms of all the metrics for both PraNet and Polyp-PVT models, which demonstrates the effectiveness of our proposed plug-in strategy based on cross-attention for feature fusion. The subpar performance of the concatenate method may be attributed to the misalignment of highlighted features and images, which could introduce noise into the dataset.

Furthermore, to validate the effectiveness of our proposed generative model in highlighting segmentation objects, we conducted an ablation study on the hyperparameter $\alpha$ in Equation (6), which controls the trade-off between the original image and the transformed image. We varied $\alpha$ across a range of values (0, 0.2, 0.5, 0.8, and 1.0), where $\alpha = 0$ represents the use of polyp regions only without any background information for guiding segmentation, and $\alpha = 1$ represents the use of the original image without transformation. The results, obtained using PraNet, are presented in Fig. 4(a), which plots the mean Dice score and mean IoU against different $\alpha$ values. Notably, our proposed setting of $\alpha = 0.5$ achieves the highest mean Dice score and mean IoU, demonstrating the superiority of our proposed generative model in highlighting segmentation objects and justifying our choice of $\alpha$. The optimal performance at $\alpha = 0.5$ suggests that our model strikes a balance between

Authorized licensed use limited to: The Chinese University of Hong Kong CUHK(Shenzhen). Downloaded on September 26,2025 at 07:38:58 UTC from IEEE Xplore. Restrictions apply.

preserving the original image information and highlighting the segmentation objects, leading to improved segmentation accuracy. Additionally, we provide visual examples of the segmentation results for different $\alpha$ values ($\alpha = 0$, $\alpha = 0.5$, and $\alpha = 1.0$) in Fig. 4(b). These visualizations illustrate that using $\alpha = 0.5$ yields better segmentation results, further supporting our findings.

### D. Qualitative Analyses

To demonstrate the main contribution of our HDM+ that the generated highlight feature can serve as an effective prior for downstream segmentation tasks, we offer a visual comparison of t-SNE and Grad-CAM [59], as shown in Fig. 3. The $1^{st}$ column shows the original colonoscopy image and its corresponding binary mask. The upper part of columns $2^{nd}$ to $4^{td}$ displays t-SNE plots for various settings, which is used to evaluate the model's ability to distinguish different classes in the embedding space. Our HDM+FCBFormer exhibits a clearer class cluster compared to FCBFormer baseline results, demonstrating a stronger ability to recognize concealed polyp features over the overwhelming background. Additionally, the lower part shows the Grad-CAM heatmaps that reveal the model's focus areas. The most accurate results illustrated that the proposed HDM provides effective priors to guide the segmentation's attention on critical polyp regions. The visualization comparison confirms our model's proficiency in maintaining the semantic integrity of the image while emphasizing diagnostically significant features.

Fig. 5 shows the highlighted features and the reconstructed images generated from the HDM. The third column visualizes the intermediate features generated by HDM, which are directly processed and utilized in the second stage for polyp segmentation, thereby bypassing the time-consuming reverse process inherent in diffusion models. Compared with the original image, the highlighted features preserve the semantic content while drawing emphasis on the polyp areas. To further demonstrate the capability of the pre-trained HDM in generating highlighted images, we utilize the highlighted features for the reverse process and obtain the samples displayed in the last column for verification. As can be seen, the samples indeed exhibit highlighted polyp regions. The figure also includes examples from the test set, underscoring our model's ability to generalize and consistently highlight polyp regions across different data sets.

Fig. 6 provides visual examples of polyp masks predicted by HDM+ and competing networks. Our method owns two advantages: 1) HDM+ demonstrates a stronger capability to adapt to diverse data conditions. Specifically, it maintains a stable recognition and segmentation ability of polyps across varying acquisition environments, e.g., low contrast, highlight distraction, water reflection, small objects, and rotation. 2) Thanks to the highlighted prior guidance, our model's segmentation results and the predicted edges are closer to the ground-truth labels. Such observations not only validate the efficacy of our highlighted features but also underscore the overall qualitative superiority of our proposed approach in enhancing the accuracy and reliability of polyp segmentation.

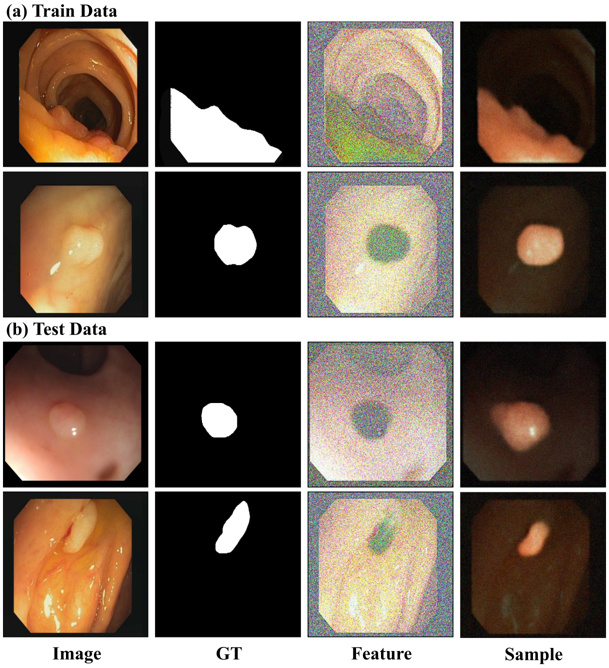

Fig. 5. Visualization of the highlighted features and reconstructed samples as produced by HDM, showcasing the method's ability to retain semantic information and emphasize polyp regions. The top panel (a) displays the train data, including the original images, ground truth masks, highlighted features, and reconstructed images. The bottom panel (b) depicts the test data with the same layout, indicating the model's robustness in generalizing the highlighting technique to unseen images. Pixel values in feature maps are rescaled to enhance visibility and samples are produced by the HDM beginning with a noised input image at timestep $t = 100$.

## V. Discussion

In recent years, the exploration of diffusion models in medical image segmentation has shown promising performance, particularly in the challenging domain of polyp segmentation from colonoscopy images. Our primary motivation is to enhance segmentation performance by addressing three key challenges: the significant variability in polyp size, shape, and color; the poor contrast between polyps and surrounding tissues; and the poor effectiveness and inefficiency of directly applying diffusion models as an end-to-end segmentation model. Previous works such as SANet [8] and ECC-PolypDet [12] have validated that small and hidden polyps significantly impact the network's ability to localize targets, leading to inaccurate segmentation boundaries. SANet incorporates a color exchange strategy to eliminate the interference caused by color variations, while ECC-PolypDet employs an additional contrastive learning module to enhance the network's ability to distinguish between polyps and surrounding tissues. However, these methods implicitly strengthen the fitting capability of neural networks, which may not be effective in all scenarios.

Our motivation is derived from the strong generative and semantic extraction capabilities of diffusion models, as identified in previous literature. The features output by trained diffusion models already contain coarse information about target locations and shapes. Initially, we attempted to directly obtain prediction

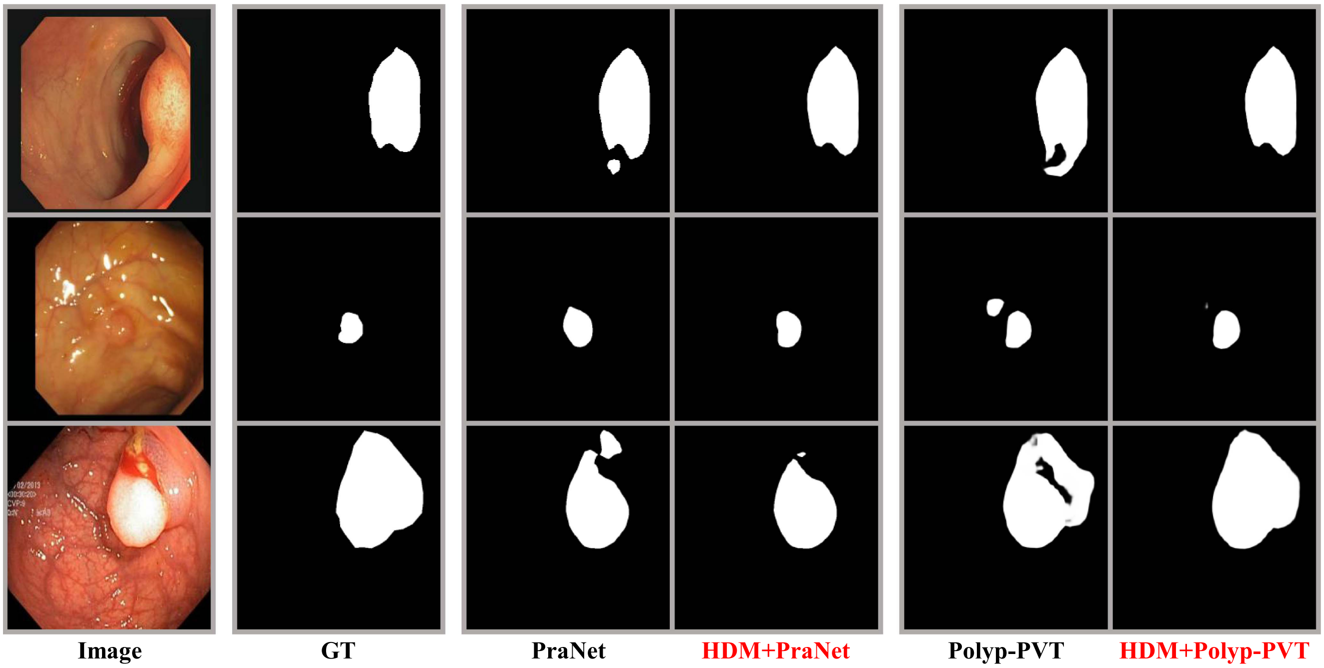

Fig. 6. Comparative visualization of segmentation results from test datasets, contrasting the performance of standard models against those augmented with our highlighted feature technique. The sequence of images showcases the original endoscopic images, the ground truth (GT) masks, and the outcomes from two distinct models, PraNet and Polyp-PVT, both with and without our enhancement ("HDM+").

results through a segmentation head using these features, but this approach yielded poor results, as confirmed by the results of MedSegDiff [19] shown in Table I. However, we further discovered that using this coarse information as guidance can significantly improve the performance of the original downstream segmentation model without any elaborate design. Our strategy can be seen as explicitly reinforcing the network by using the features from the diffusion model. Additionally, after training the HDM, we utilize only the diffusion U-Net output for the second stage of polyp segmentation, which can be easily integrated with any discriminative segmentation model in a plug-and-play manner. The computational costs are primarily reserved for HDM training, and once trained, the generation of highlighted features incurs negligible overhead compared to polyp segmentation. Extensive experiments validate that our proposed method can be effectively generalized to both CNN-based and Transformer-based models. Our work introduces a novel perspective and research direction for the subsequent design of polyp segmentation models.

Despite the promising performance of HDM+ across various benchmarks, certain limitations persist. One notable challenge arises still from polyps with extremely ambiguous boundaries, which remain difficult for the model to accurately segment. As illustrated in Fig. 7(a), the model struggled to distinguish the small target from the large background, leading to inaccurate segmentation. Additionally, the generation ability of HDM can affect the performance of downstream baseline models, where inaccurately highlighted features may introduce more noise and impact performance. As shown in Fig. 7(b), less accurate highlighted features generated by HDM can result in incorrect segmentation by the baseline models.

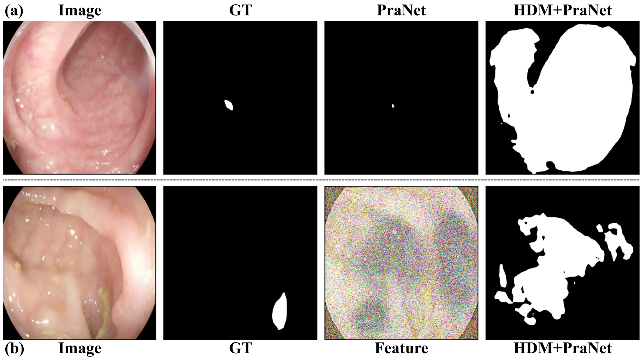

Fig. 7. Illustration of failure cases: (a) Inaccurate segmentation of image with unclear polyp boundary. (b) Failure case due to inaccurate HDM feature.

We identify several promising directions for future improvement. For the first challenge, while our primary contribution is the proposed simple and plug-and-play HDM module that can easily enhance the performance of baseline segmentation models, the final performance still heavily depends on the learning capability of the baseline model. Therefore, developing more fine-grained polyp segmentation baselines could help more effectively process polyps of varying sizes and shapes. For the second challenge, jointly training the polyp segmentation model and the HDM module may mitigate the issue of inaccurate highlighted features by using both diffusion and segmentation criteria to guide the models. Furthermore, as diffusion models continue to evolve, applying more robust and efficient diffusion models could capture more precise highlighted features, thereby further improving segmentation accuracy.

Authorized licensed use limited to: The Chinese University of Hong Kong CUHK(Shenzhen). Downloaded on September 26,2025 at 07:38:58 UTC from IEEE Xplore. Restrictions apply.

## VI. Conclusion

Precise polyp segmentation is pivotal in colorectal cancer diagnosis, yet challenges such as limited data availability and complex colonoscopic environments often hinder the performance of conventional discriminative models. This paper presents an innovative approach, employing a Highlighted Diffusion Model (HDM) to generate explicit references by distinctly highlighting the differences between polyps and the surrounding intestinal walls. These references serve as plug-in priors for polyp segmentation models. We introduce a sophisticated two-stage training and end-to-end inference framework named Highlighted Diffusion Model Plus (HDM+), specifically designed for accurate polyp segmentation. Specifically, to resolve issues of reference feature misalignment while maintaining the integrity of the original image data, the framework incorporates a cross-attention mechanism for effective feature integration. This is complemented by the utilization of processed features, which are instrumental in retaining essential image details, thereby enhancing the guidance provided to downstream models. Our extensive experiments across six renowned public polyp segmentation benchmarks validate the efficacy of our proposed network in addressing the challenges of polyp segmentation. We anticipate that this research will stimulate further exploration and innovation in the application of generative models to polyp segmentation and related fields.

## References

[1] M. Navarro, A. Nicolas, A. Ferrandez, and A. Lanas, "Colorectal cancer population screening programs worldwide in 2016: An update," *World J. Gastroenterol.*, vol. 23, no. 20, 2017, Art. no. 3632.

[2] P. Favoriti, G. Carbone, M. Greco, F. Pirozzi, R. E. M. Pirozzi, and F. Corcione, "Worldwide burden of colorectal cancer: A review," *Updates Surg.*, vol. 68, pp. 7–11, 2016.

[3] F. A. Haggar and R. P. Boushey, "Colorectal cancer epidemiology: Incidence, mortality, survival, and risk factors," *Clin. Colon Rectal Surg.*, vol. 22, no. 04, pp. 191–197, 2009.

[4] A. Leufkens, M. Van Oijen, F. Vleggaar, and P. Siersema, "Factors influencing the miss rate of polyps in a back-to-back colonoscopy study," *Endoscopy*, vol. 44, pp. 470–475, 2012.

[5] R. Zhang et al., "Automatic detection and classification of colorectal polyps by transferring low-level CNN features from nonmedical domain," *IEEE J. Biomed. Health Informat.*, vol. 21, no. 1, pp. 41–47, Jan. 2017.

[6] R. Zhang, G. Li, Z. Li, S. Cui, D. Qian, and Y. Yu, "Adaptive context selection for polyp segmentation," in *Proc. IEEE Int. Conf. Mach. Learn. Appl.*, 2020, pp. 253–262.

[7] D.-P. Fan et al., "PraNet: Parallel reverse attention network for polyp segmentation," in *Proc. IEEE Int. Conf. Mach. Learn. Appl.*, Springer, 2020, pp. 263–273.

[8] J. Wei, Y. Hu, R. Zhang, Z. Li, S. K. Zhou, and S. Cui, "Shallow attention network for polyp segmentation," in *Proc. IEEE Int. Conf. Mach. Learn. Appl.*, 2021, pp. 699–708.

[9] B. Dong, W. Wang, D.-P. Fan, J. Li, H. Fu, and L. Shao, "Polyp-PVT: Polyp segmentation with pyramid vision transformers," *CAAI Artif. Intell. Res.*, vol. 2, 2023, Art. no. 9150015.

[10] Y. Jiang, Z. Zhang, R. Zhang, G. Li, S. Cui, and Z. Li, "Yona: You only need one adjacent reference-frame for accurate and fast video polyp detection," in *Proc. Int. Conf. Med. Image Comput. Comput.-Assist. Intervention*, 2023, pp. 44–54.

[11] X. Xiong, S. Li, and G. Li, "Unpaired image-to-image translation based domain adaptation for polyp segmentation," in *Proc. IEEE Int. Symp. Biomed. Imag.*, 2023, pp. 1–5.

[12] Y. Jiang et al., "ECC-PolypDet: Enhanced CenterNet with contrastive learning for automatic polyp detection," *IEEE J. Biomed. Health Informat.*, vol. 28, no. 8, pp. 4785–4796, Aug. 2024.

[13] A. Vaswani et al., "Attention is all you need," in *Proc. Adv. Neural Inf. Process. Syst.*, 2017, pp. 6000–6010.

[14] W. Wang et al., "Semantic image synthesis via diffusion models," 2022, *arXiv:2207.00050*.

[15] Z. Dorjsembe, H.-K. Pao, S. Odonchimed, and F. Xiao, "Conditional diffusion models for semantic 3D brain MRI synthesis," *IEEE J. Biomed. Health Inform.*, vol. 28, no. 7, pp. 4084–4093, 2024.

[16] J. Sohl-Dickstein, E. Weiss, N. Maheswaranathan, and S. Ganguli, "Deep unsupervised learning using nonequilibrium thermodynamics," in *Proc. Int. Conf. Mach. Learn.*, 2015, pp. 2256–2265.

[17] J. Ho, A. Jain, and P. Abbeel, "Denoising diffusion probabilistic models," in *Proc. Adv. Neural Inf. Process. Syst.*, 2020, pp. 6840–6851.

[18] A. Rahman, J. M. J. Valanarasu, I. Hacihaliloglu, and V. M. Patel, "Ambiguous medical image segmentation using diffusion models," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2023, pp. 11536–11546.

[19] J. Wu, H. Fang, Y. Zhang, Y. Yang, and Y. Xu, "MedSegDiff: Medical image segmentation with diffusion probabilistic model," in *Proc. Med. Imag. Deep Learn.*, 2023.

[20] A. Bozorgpour, Y. Sadegheih, A. Kazerouni, R. Azad, and D. Merhof, "DermoSegDiff: A boundary-aware segmentation diffusion model for skin lesion delineation," in *Proc. Predictive Intell. Med.*, 2023, pp. 146–158.

[21] Y. Du et al., "ARSDM: Colonoscopy images synthesis with adaptive refinement semantic diffusion models," in *Proc. Int. Conf. Med. Image Comput. Comput.-Assist. Intervention*, 2023, pp. 339–349.

[22] C. Van Wijk, V. F. Van Ravesteijn, F. M. Vos, and L. J. Van Vliet, "Detection and segmentation of colonic polyps on implicit isosurfaces by second principal curvature flow," *IEEE Trans. Med. Imag.*, vol. 29, no. 3, pp. 688–698, Mar. 2010.

[23] S. Hwang and M. E. Celebi, "Polyp detection in wireless capsule endoscopy videos based on image segmentation and geometric feature," in *Proc. 2010 IEEE Int. Conf. Acoust. Speech Signal Process.*, 2010, pp. 678–681.

[24] M. Ganz, X. Yang, and G. Slabaugh, "Automatic segmentation of polyps in colonoscopic narrow-band imaging data," *IEEE Trans. Biomed. Eng.*, vol. 59, no. 8, pp. 2144–2151, Aug. 2012.

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

[26] A. V. Mamonov, I. N. Figueiredo, P. N. Figueiredo, and Y.-H. R. Tsai, "Automated polyp detection in colon capsule endoscopy," *IEEE Trans. Med. Imag.*, vol. 33, no. 7, pp. 1488–1502, Jul. 2014.

[27] S. Ameling, S. Wirth, D. Paulus, G. Lacey, and F. Vilarino, "Texture-based polyp detection in colonoscopy," *Bildverarbeitung für die Medizin*, vol. 2009, pp. 346–350, 2009.

[28] S. A. Karkanis, D. K. Iakovidis, D. E. Maroulis, D. A. Karras, and M. Tzivras, "Computer-aided tumor detection in endoscopic video using color wavelet features," *IEEE Trans. Inf. Technol. Biomed.*, vol. 7, no. 3, pp. 141–152, Sep. 2003.

[29] J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2015, pp. 3431–3440.

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

[31] M. Akbari et al., "Polyp segmentation in colonoscopy images using fully convolutional network," in *Proc. IEEE Eng. Med. Biol. Soc.*, 2018, pp. 69–72.

[32] X. Sun, P. Zhang, D. Wang, Y. Cao, and B. Liu, "Colorectal polyp segmentation by U-Net with dilation convolution," in *Proc. IEEE Int. Conf. Mach. Learn. Appl.*, 2019, pp. 851–858.

[33] 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.

[34] F. I. Diakogiannis, F. Waldner, P. Caccetta, and C. Wu, "ResUNet-a: A deep learning framework for semantic segmentation of remotely sensed data," *ISPRS J. Photogrammetry Remote Sens.*, vol. 162, pp. 94–114, 2020.

[35] B. Murugesan, K. Sarveswaran, S. M. Shankaranarayana, K. Ram, J. Joseph, and M. Sivaprakasam, "PSI-Net: Shape and boundary aware joint multi-task deep network for medical image segmentation," in *Proc. IEEE Eng. Med. Biol. Soc.*, 2019, pp. 7223–7226.

[36] Y. Fang, C. Chen, Y. Yuan, and K.-Y. Tong, "Selective feature aggregation network with area-boundary constraints for polyp segmentation," in *Proc. IEEE Int. Conf. Mach. Learn. Appl.*, Springer, 2019, pp. 302–310.

Authorized licensed use limited to: The Chinese University of Hong Kong CUHK(Shenzhen). Downloaded on September 26,2025 at 07:38:58 UTC from IEEE Xplore. Restrictions apply.

[37] R. Zhang et al., “Lesion-aware dynamic kernel for polyp segmentation,” in *Proc. IEEE Int. Conf. Mach. Learn. Appl.*, 2022, pp. 99–109.

[38] M. Wang et al., “An efficient multi-task synergetic network for polyp segmentation and classification,” *IEEE J. Biomed. Health Informat.*, vol. 28, no. 3, pp. 1228–1239, Mar. 2024.

[39] D. Kingma, T. Salimans, B. Poole, and J. Ho, “Variational diffusion models,” in *Proc. Adv. Neural Inf. Process. Syst.*, 2021, pp. 21696–21707.

[40] J. Wolleb, F. Bieder, R. Sandkühler, and P. C. Cattin, “Diffusion models for medical anomaly detection,” in *Proc. Int. Conf. Med. Image Comput. Comput.-Assist. Intervention*, 2022, pp. 35–45.

[41] J. Song, C. Meng, and S. Ermon, “Denoising diffusion implicit models,” in *Proc. Int. Conf. Learn. Representations*, 2020.

[42] P. Dhariwal and A. Nichol, “Diffusion models beat GANs on image synthesis,” in *Proc. Adv. Neural Inf. Process. Syst.*, 2021, pp. 8780–8794.

[43] A. Q. Nichol and P. Dhariwal, “Improved denoising diffusion probabilistic models,” in *Proc. Int. Conf. Mach. Learn.*, 2021, pp. 8162–8171.

[44] R. Rombach, A. Blattmann, D. Lorenz, P. Esser, and B. Ommer, “High-resolution image synthesis with latent diffusion models,” in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2022, pp. 10684–10695.

[45] X. Guo et al., “Accelerating diffusion models via pre-segmentation diffusion sampling for medical image segmentation,” in *Proc. IEEE Int. Symp. Biomed. Imag.*, 2023, pp. 1–5.

[46] Y. Du and I. Mordatch, “Implicit generation and modeling with energy based models,” in *Proc. Adv. Neural Inf. Process. Syst.*, 2019, pp. 3608–3618.

[47] J. Xu, S. Liu, A. Vahdat, W. Byeon, X. Wang, and S. De Mello, “Open-vocabulary panoptic segmentation with text-to-image diffusion models,” in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2023, pp. 2955–2966.

[48] J. Wei, S. Wang, and Q. Huang, “$F^3$ net: Fusion, feedback and focus for salient object detection,” in *Proc. Conf. Assoc. Advance. Artif. Intell.*, 2020, pp. 12321–12328.

[49] D. Vázquez et al., “A benchmark for endoluminal scene segmentation of colonoscopy images,” *J. Healthcare Eng.*, vol. 2017, 2017, Art. no. 4037190.

[50] J. Bernal, F. J. Sánchez, G. Fernández-Esparrach, D. Gil, C. Rodríguez, and F. Vilariño, “WM-DOVA maps for accurate polyp highlighting in colonoscopy: Validation vs. saliency maps from physicians,” *Computerized Med. Imag. Graph.*, vol. 43, pp. 99–111, 2015.

[51] J. Silva, A. Histace, O. Romain, X. Dray, and B. Granado, “Toward embedded detection of polyps in WCE images for early diagnosis of colorectal cancer,” *Int. J. Comput. Assist. Radiol. Surg.*, vol. 9, pp. 283–293, 2014.

[52] D. Jha et al., “Kvasir-SEG: A segmented polyp dataset,” in *Proc. Int. Conf. Multimedia Model.*, 2020, pp. 451–462.

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

[54] E. Sanderson and B. J. Matuszewski, “FCN-transformer feature fusion for polyp segmentation,” in *Proc. Annu. Conf. Med. Image Understanding Anal.*, 2022, pp. 892–907.

[55] D.-P. Fan, M.-M. Cheng, Y. Liu, T. Li, and A. Borji, “Structure-measure: A new way to evaluate foreground maps,” in *Proc. IEEE Int. Conf. Comput. Vis.*, 2017, pp. 4548–4557.

[56] D.-P. Fan, C. Gong, Y. Cao, B. Ren, M.-M. Cheng, and A. Borji, “Enhanced-alignment measure for binary foreground map evaluation,” in *Proc. Int. Joint Conf. Artif. Intell.*, 2018, pp. 698–704.

[57] R. Achanta, S. Hemami, F. Estrada, and S. Susstrunk, “Frequency-tuned salient region detection,” in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2009, pp. 1597–1604.

[58] R. Margolin, L. Zelnik-Manor, and A. Tal, “How to evaluate foreground maps?,” in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2014, pp. 248–255.

[59] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, “Grad-CAM: Visual explanations from deep networks via gradient-based localization,” in *Proc. IEEE Int. Conf. Comput. Vis.*, 2017, pp. 618–626.

[60] J. Wu et al., “MedSegDiff: Medical image segmentation with diffusion probabilistic model,” in *Proc. Med. Imag. Deep Learn.*, 2023.

Authorized licensed use limited to: The Chinese University of Hong Kong CUHK(Shenzhen). Downloaded on September 26,2025 at 07:38:58 UTC from IEEE Xplore. Restrictions apply.