Semantic Map Guided Synthesis of Wireless Capsule Endoscopy Images using Diffusion Models

Haejin Lee [1] **Jeongwoo Ju** [1] **Jonghyuck Lee** [1,2]
[2] Seoreu Co., Ltd.
Busan 46288, Korea
jhlee@seoreu.com

Yeoun Joo Lee [1,3]
[3] Department of Pediatrics
Pusan National University School of Medicine
Pusan National University Yangsan Hospital
Yangsan 50612, Korea
moonmissing@gmail.com

Heechul Jung [1,4]
[4] Department of Artificial Intelligence
Kyungpook National University
Daegu 41566, Korea
heechul@knu.ac.kr

[1] Captos Co., Ltd.
Yangsan 50602, Korea
{seareale,veryju,jhlee,yeounjoolee,heechul}@captos.co.kr

Abstract

Wireless capsule endoscopy (WCE) is a non-invasive method for visualizing the gastrointestinal (GI) tract, crucial for diagnosing GI tract diseases. However, interpreting WCE results can be time-consuming and tiring. Existing studies have employed deep neural networks (DNNs) for automatic GI tract lesion detection, but acquiring sufficient training examples, particularly due to privacy concerns, remains a challenge. Public WCE databases lack diversity and quantity. To address this, we propose a novel approach leveraging generative models, specifically the diffusion model (DM), for generating diverse WCE images. Our model incorporates semantic map resulted from visualization scale (VS) engine, enhancing the controllability and diversity of generated images. We evaluate our approach using visual inspection and visual Turing tests, demonstrating its effectiveness in generating realistic and diverse WCE images.

1 Introduction

WCE [14] is a non-invasive method for viewing the GI tract, aiding in GI tract disease diagnosis. However, reading WCE results can be time-consuming and lead to fatigue. Studies [26, 3, 33, 2] have worked on automatic GI tract lesion detection using deep neural networks (DNNs), which require extensive training examples like ImageNet [4]. However, privacy concerns make gathering enough examples challenging [31]. Morevover, public WCE databases [19, 25] lack diversity and sufficient examples. Generating data using generative models [7] may be a solution. Generative models, especially deep learning-based, excel in generating high-quality data in text [15], image [9], and audio [13, 21]. They address issues like class imbalance [24] and example augmentation [1, 10], relevant in the medical domain. Most works, including medical image datasets, rely on generative adversarial networks (GANs) [11]. Denoising diffusion probabilistic models (DDPMs) [12] capture
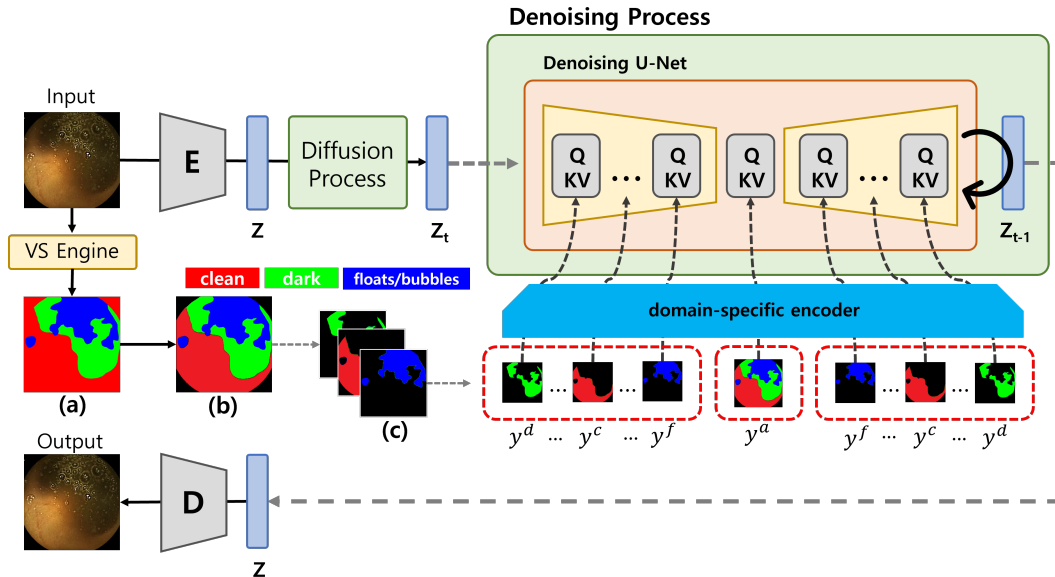
Figure 1: Flowchart for training the Diffusion Model for WCE image generation. We start by obtaining the segmentation map from the VS engine. This map is then divided into class-specific maps. These individual maps are subsequently integrated into the Encoder, Decoder, and middle layer of the Denoising U-Net.

real image variations effectively, potentially outperforming GANs [5]. Recent image synthesis efforts favor DDPM approaches [20, 18, 32]. In the medical domain, attention to WCE is limited compared to other modalities. Diamantis et al. [6, 7, 6, 8] and Valts et al. [29, 30] addressed WCE image synthesis, but more sophisticated techniques for diversity and controllability are needed. This study aims to generate diverse WCE images using a diffusion model with semantic segmentation map. We achieved controllability over the semantic map, which was readily obtained through the VS engine [16], enabling the generation of WCE images with desired semantic map. Additionally, as an WCE image is composed of semantic area (e.g. clean area, dark area, the other area where floating debris or bubbles occupy) such that one can easily notice those area at a glance. Also, considering the fact that lesion can only appear inside clean area and the rest area can exhibit different appearance in a various way, controlling semantic area is a key technique to achieve greater diversity.

Evaluation included visual inspection and visual turing test. Contributions: i) Pioneering use of DM for synthesized WCE images. ii) Incorporation of floats/bubbles and darkness area as controllable factors using VS results.

2 Methods

2.1 Kvasir-capsule Dataset

In this study, we used the Kvasir-Capsule dataset, which is the largest publicly available collection of WCE images captured by PillCam (Medtronic, Minneapolis, MN, USA). The dataset consists of 47,238 labeled images and 43 labeled videos, along with 74 unlabeled videos. It includes images corresponding to multiple organs, such as the stomach, small bowel (SB) and colon. We used all labeled images to train our latent diffusion model (LDM).

2.2 Segmentation Mask

Ju et al. [16] developed visualization scale (VS) engine designed for semantic segmentation to assess the clean mucosa area in small bowel images captured by WCE. The target classes for segmentation are darkness (absence of light), floats/bubbles (bubbles and floating debris), and clean (clear visual area). The overall performance, measured by the Dice index, was 0.9457. The comparison between physicians and the AI engine demonstrated good agreement in accessing the clean area [17].

In this study, we fed the images to the VS engine and obtained a semantic segmentation map. However, the VS engine mistakenly regarded the blank areas at the four corners of a WCE image as clean areas (Figure 1.(a)). To address this issue, we manually filtered out the corners from the semantic map and reassigned them as an extra class (Figure 1.(b)). Additionally, we split the single-channel semantic map into three channels, where each channel depicts a single semantic class (Figure 1.(c)). As a result, we obtained three channels based on the single semantic map.

2.3 Mask-conditioned Latent Diffusion Models

Rombach et al. [22] introduced LDM which is computationally faster version of vanilla DM while keeping synthesis performance. Moreover, one of key technique for generative model is modelling conditional distribution, $p(\cdot|\boldsymbol{y})$, they proposed conditioning mechanism with following loss function, $L_{DM} = \mathbb{E}_{\boldsymbol{z},\boldsymbol{y},\boldsymbol{\epsilon}\sim\mathcal{N}(\mathbf{0},\boldsymbol{I}),t}\left[\left\|\boldsymbol{\epsilon} - f_{\boldsymbol{\theta}}(\boldsymbol{z}_t,t,\tau_{\boldsymbol{\theta}}(\boldsymbol{y}))\right\|_2^2\right]$, where t denotes time uniformly sampled from $\{1,...,T\}$, $f_{\boldsymbol{\theta}}(*,t,\tau_{\boldsymbol{\theta}}(\boldsymbol{y}))$ is conditional denoising autoencoder, $\tau_{\boldsymbol{\theta}}(\boldsymbol{y})$ is a domain-specific encoder, $\boldsymbol{z}_t$ is the noisy variant of latent variable $\boldsymbol{z}$ at time t, and $\boldsymbol{y}$ is conditional variable. f_{θ} is basically implemented by introducing the backbone with U-Net [23] and cross-attention mechanism [28]. Moreover, feeding conditional variable via $\tau_{\boldsymbol{\theta}}$ into the intermediate layers of UNet is the core of modelling $p(\boldsymbol{z}|\boldsymbol{y})$. In contrast to existing methods where $\boldsymbol{y}$ is fixed as a single segmentation map, we divided $\boldsymbol{y}$ into multiple semantic maps. Specifically, $\boldsymbol{y}$ is composed of maps sampled from $\{\boldsymbol{y}^d, \boldsymbol{y}^c, \boldsymbol{y}^f, \boldsymbol{y}^a\}$, where $\boldsymbol{y}^d, \boldsymbol{y}^c, \boldsymbol{y}^f, \boldsymbol{y}^a$ denote the segmentation maps for dark, clean, floats/bubbles, and all areas respectively. When implementing the denoising U-Net, we sequentially fed samples from $\{\boldsymbol{y}^d, \boldsymbol{y}^c, \boldsymbol{y}^f\}$ into the U-Net encoder in a specified order, $\boldsymbol{y}^a$ into the middle layer, and the reversed samples (which is the exact reverse order of semantic maps for the encoder) into the decoder. By doing so, we expected U-Net to learn sequentially on how to generate each semantic area.

3 Results

VTT Five gastroenterologists with over 5 years of experience in capsule endoscopy interpretation conducted a visual Turing test using a total of 160 images, including 80 real and 80 fake images. The results showed an average accuracy of 0.64 in correctly identifying real images as real, and a 0.662 accuracy in incorrectly identifying fake images as real. This indicates that the quality of the generated WCE is perceived to be comparable to real images, as evidenced by the similar rates at which experts classified real and fake images as real during the assessment of image quality.

Visual Inspection Figure 2 displays images generated with varying random seeds when an arbitrary segmentation mask is provided (top row). It can be observed that different areas are created in accordance with the regions outlined in the segmentation mask, including dark, clean, and floats/bubbles regions. Notably, despite the fact that mask 5 represent semantic maps that are challenging to exist in reality (we created), our DM successfully generated corresponding areas in accordance with these maps.

4 Related Work

Generating WCE images Diamantis et al. [8] attempted WCE image generation with GANs. However, their study lacked extensive experiments, only offering binary classification results. Their subsequent work, EndoVAE [6], used a simple architecture but still lacked thorough experiments, focusing only on binary classification. Their latest work, TIDE [7], improved VAE by incorporating multiscale blocks (MSB) for high-quality image synthesis, outperforming GAN variants. Vats et al. [29] aimed to create a WCE atlas and simulate disease progression for educational purposes. They built an embedding space using StyleGAN2 due to limited annotated WCE data. Although their synthetic images were high-quality, attribute discovery involved human validation. However, our work has demonstrated the ability to control semantic areas, generating high-quality WCE images without the need for human intervention.

Diffusion-based Models Sohl-Dickstein et al. [27] introduced DM inspired by statistical physics. Ho et al. [12] demonstrated DM's capability for synthetic image generation. However, DM's high computational cost led to the development of LDM by Rombach et al. [22], achieving similar quality

with less resource consumption. LDM's innovation lies in utilizing DM in the latent space of a pretrained autoencoder and incorporating cross-attention layers for conditional image generation with text input. We chose LDM as our base WCE image generator.

5 Conclusion

In this study, we introduced a novel approach for generating diverse WCE images using a DM with semantic segmentation maps. By obtaining controllability over the semantic map, we demonstrated the ability to generate WCE images with desired appearances, eliminating the need for additional annotations. The evaluation, including visual inspection and a visual Turing test, confirmed the high quality and realism of the generated images. Additionally, we incorporated floats/bubbles and darkness area as controllable factors, enhancing the model's versatility. This work paves the way for improved WCE image synthesis, providing a valuable resource for medical research and education.
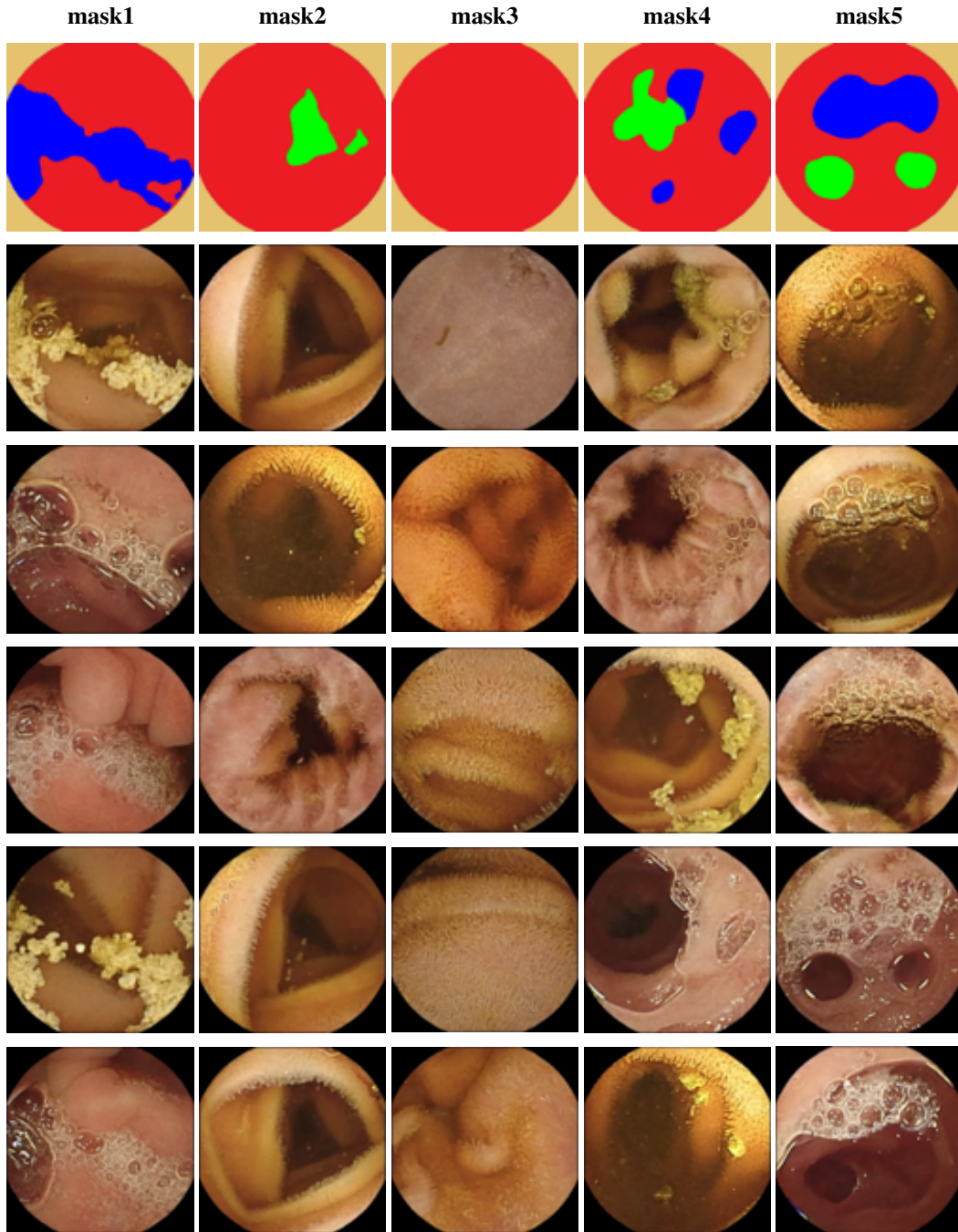
Figure 2: The segmentation maps used in WCE image generation (top row). The masks 1 to 3 were obtained from real images, while masks 4 to 5 were created by us. The generated images corresponding to each map are shown in the subsequent rows. **Beige-brown: blank area, red: clean area, green: dark area, blue: floats/bubbles area.**

References

[1]   Antreas Antoniou, Amos Storkey, and Harrison Edwards. "Data augmentation generative adversarial networks". In: *arXiv preprint arXiv:1711.04340* (2017).

[2]   Tomonori Aoki et al. "Automatic detection of various abnormalities in capsule endoscopy videos by a deep learning-based system: a multicenter study". In: *Gastrointestinal Endoscopy* 93.1 (2021), pp. 165–173.

[3]   Yiftach Barash et al. "Ulcer severity grading in video capsule images of patients with Crohn's disease: An ordinal neural network solution". In: *Gastrointestinal Endoscopy* 93.1 (2021), pp. 187–192.

[4]   Jia Deng et al. "Imagenet: A large-scale hierarchical image database". In: *2009 IEEE conference on computer vision and pattern recognition*. Ieee. 2009, pp. 248–255.

[5]   Prafulla Dhariwal and Alexander Nichol. "Diffusion models beat gans on image synthesis". In: *Advances in Neural Information Processing Systems* 34 (2021), pp. 8780–8794.

[6]   Dimitrios E Diamantis, Panagiota Gatoula, and Dimitris K Iakovidis. "EndoVAE: Generating Endoscopic Images with a Variational Autoencoder". In: *2022 IEEE 14th Image, Video, and Multidimensional Signal Processing Workshop (IVMSP)*. IEEE. 2022, pp. 1–5.

[7]   Dimitrios E Diamantis et al. "This Intestine Does Not Exist: Multiscale Residual Variational Autoencoder for Realistic Wireless Capsule Endoscopy Image Generation". In: *arXiv preprint arXiv:2302.02150* (2023).

[8]   Dimitris E Diamantis et al. "Towards the substitution of real with artificially generated endoscopic images for CNN training". In: *2019 IEEE 19th International Conference on Bioinformatics and Bioengineering (BIBE)*. IEEE. 2019, pp. 519–524.

[9]   Mohamed Elasri et al. "Image Generation: A Review". In: *Neural Processing Letters* 54.5 (2022), pp. 4609–4646.

[10] Maayan Frid-Adar et al. "GAN-based synthetic medical image augmentation for increased CNN performance in liver lesion classification". In: *Neurocomputing* 321 (2018), pp. 321–331.

[11] Ian Goodfellow et al. "Generative adversarial networks". In: *Communications of the ACM* 63.11 (2020), pp. 139–144.

[12] Jonathan Ho, Ajay Jain, and Pieter Abbeel. *Denoising Diffusion Probabilistic Models*. 2020. arXiv: 2006.11239 [cs.LG].

[13] Muhammad Huzaifah and Lonce Wyse. "Deep generative models for musical audio synthesis". In: *Handbook of Artificial Intelligence for Music: Foundations, Advanced Approaches, and Developments for Creativity* (2021), pp. 639–678.

[14] Gavriel Iddan et al. "Wireless capsule endoscopy". In: *Nature* 405.6785 (2000), pp. 417–417.

[15] Touseef Iqbal and Shaima Qureshi. "The survey: Text generation models in deep learning". In: *Journal of King Saud University-Computer and Information Sciences* 34.6 (2022), pp. 2515–2528.

[16] Jeong-Woo Ju et al. "Semantic Segmentation Dataset for AI-Based Quantification of Clean Mucosa in Capsule Endoscopy". In: *Medicina* 58.3 (2022), p. 397.

[17] Jeongwoo Ju et al. "Clean mucosal area detection of gastroenterologists versus artificial intelligence in small bowel capsule endoscopy". In: *Medicine* 102.6 (2023).

[18] Amirhossein Kazerouni et al. "Diffusion models for medical image analysis: A comprehensive survey". In: *arXiv preprint arXiv:2211.07804* (2022).

[19] Anastasios Koulaouzidis et al. "KID Project: an internet-based digital video atlas of capsule endoscopy for research purposes". In: *Endoscopy international open* 5.06 (2017), E477–E483.

[20] Gustav Müller-Franzes et al. "Diffusion Probabilistic Models beat GANs on Medical Images". In: *arXiv preprint arXiv:2212.07501* (2022).

[21] Aaron van den Oord et al. "Wavenet: A generative model for raw audio". In: *arXiv preprint arXiv:1609.03499* (2016).

[22] Robin Rombach et al. "High-resolution image synthesis with latent diffusion models". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2022, pp. 10684–10695.

[23] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. "U-net: Convolutional networks for biomedical image segmentation". In: *Medical Image Computing and Computer-Assisted Intervention–MICCAI 2015: 18th International Conference, Munich, Germany, October 5-9, 2015, Proceedings, Part III 18*. Springer. 2015, pp. 234–241.

[24] Vignesh Sampath et al. "A survey on generative adversarial networks for imbalance problems in computer vision tasks". In: *Journal of big Data* 8 (2021), pp. 1–59.

[25] Pia H Smedsrud et al. "Kvasir-Capsule, a video capsule endoscopy dataset". In: *Scientific Data* 8.1 (2021), p. 142. DOI: 10.1038/s41597-021-00920-z.

[26] Shelly Soffer et al. "Deep learning for wireless capsule endoscopy: a systematic review and meta-analysis". In: *Gastrointestinal endoscopy* 92.4 (2020), pp. 831–839.

[27] Jascha Sohl-Dickstein et al. "Deep unsupervised learning using nonequilibrium thermodynamics". In: *International Conference on Machine Learning*. PMLR. 2015, pp. 2256–2265.

[28] Ashish Vaswani et al. "Attention is all you need". In: *Advances in neural information processing systems* 30 (2017).

[29] Anuja Vats et al. "Evaluating clinical diversity and plausibility of synthetic capsule endoscopic images". In: *arXiv preprint arXiv:2301.06366* (2023).

[30] Anuja Vats et al. "This changes to that: Combining causal and non-causal explanations to generate disease progression in capsule endoscopy". In: *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE. 2023, pp. 1–5.

[31] Paul Voigt and Axel Von dem Bussche. "The eu general data protection regulation (gdpr)". In: *A Practical Guide, 1st Ed., Cham: Springer International Publishing* 10.3152676 (2017), pp. 10–5555.

[32] Ling Yang et al. "Diffusion models: A comprehensive survey of methods and applications". In: *arXiv preprint arXiv:2209.00796* (2022).

[33] Young Joo Yang. "The future of capsule endoscopy: The role of artificial intelligence and other technical advancements". In: *Clinical Endoscopy* 53.4 (2020), pp. 387–394.