# Generative Medical Segmentation

Jiayu Huo[1(✉)], Xi Ouyang[2], Sébastien Ourselin[1], and Rachel Sparks[1]

[1] School of Biomedical Engineering and Imaging Sciences (BMEIS), King's College London, London, UK
jiayu.huo@kcl.ac.uk
[2] Shanghai United Imaging Intelligence Co., Ltd., Shanghai, China

**Abstract.** Rapid advancements in medical image segmentation performance have been significantly driven by the development of Convolutional Neural Networks (CNNs) and Vision Transformers (ViTs). However, these models introduce high computational demands and often have limited ability to generalize across diverse medical imaging datasets. In this manuscript, we introduce Generative Medical Segmentation (GMS), a novel approach leveraging a generative model for image segmentation. Concretely, GMS employs a robust pre-trained Variational Autoencoder (VAE) to derive latent representations of both images and masks, followed by a mapping model that learns the transition from image to mask in the latent space. This process culminates in generating a precise segmentation mask within the image space using the pre-trained VAE decoder. The design of GMS leads to fewer learnable parameters in the model, resulting in a reduced computational burden and enhanced generalization capability. Our extensive experimental analysis across five public datasets in different medical imaging domains demonstrates GMS outperforms existing discriminative segmentation models and has remarkable domain generalization. Our experiments suggest GMS could set a new benchmark for medical image segmentation, offering a scalable and effective solution. GMS implementation and model weights are available at https://github.com/King-HAW/GMS

**Keywords:** Image segmentation · Generative model · Cross-domain generalization.

## 1 Introduction

Image segmentation plays a crucial role in the field of medical image analysis allowing an automated method to precisely delineate and separate different anatomical structures and pathological entities visible in medical images. Automated segmentation enables clinicians to obtain detailed visualizations and quantitative assessments of lesions and other structural anomalies facilitating computer-aided diagnosis, treatment planning, and the monitoring of disease progression [1,21,10].

Current deep learning models designed for the medical image segmentation task, such as U-Net [17] and its various adaptations [18,8], have significantly

advanced the field of medical imaging analysis. These models have been pivotal in enhancing the accuracy and efficiency of segmenting anatomical structures or abnormalities from various imaging modalities such as MRI and CT. Early deep learning models leverage convolutional neural networks (CNNs) to learn local patch representations from large amounts of labeled data. Despite their successes, CNN-based models often have a large number of parameters which may produce challenges in model training. Additionally, the limited receptive field of convolution operations makes it difficult for CNN-based models to learn global context information that can provide important guidance during segmentation. Moreover, CNN-based models sometimes struggle with generalizing to unseen domains, leading to substantial potential performance drops when the test dataset distribution is shifted from the training dataset distribution.

Vision transformer (ViT) [5] has recently been presented as a powerful alternative to CNN-based segmentation models in medical imaging analysis. ViT enables capturing global semantic information often overlooked by convolution operations. Transformer-based segmentation models, such as UCTransNet [22] and Swin-Unet [2], leverage the transformer architecture to treat images as sequences of patches, promoting the model to learn relationships across the entire image. Transformer-based models facilitate more holistic image analysis by integrating both local and global context information. Therefore, they can accurately segment anatomical structures or pathological changes in medical images, surpassing CNNs in certain domains. However, transformer-based models are required to be pre-trained on very large datasets to achieve optimal performance, which can be a major bottleneck given the scarcity of such datasets in the medical field. Additionally, the high computational complexity needed for the multi-head attention module poses practical challenges for real-time applications and deployment in environments with limited computational resources. Furthermore, due to the large number of parameters in transform-based models, there is an increased risk of overfitting when training on small datasets with the subsequent challenges of poor generalization to out-of-domain datasets.

Generative models, such as Generative Adversarial Networks (GANs) [6] and Variational Autoencoders (VAEs) [12], are often adopted as data augmentation techniques to improve the performance of segmentation models [7]. However, GANs are well-known to suffer from mode collapse and distorted outputs when the number of training samples is small [11]. Additionally, GANs can not guarantee that the distribution of synthetic images they create is similar to the distribution of real images. Image-to-image translation models have been presented for building segmentation models directly in a generative manner. To date, the performance of image-to-image models is well below state-of-the-art model performance [13]. Recently, BerDiff [4] deployed a diffusion model using an image as the condition for mask generation. Such an approach is less efficient for image segmentation because it requires repetitive denoising steps. Chen *et al.* [3] proposed a generative semantic segmentation (GSS) framework with a two-stage learning protocol. However, GSS still has a high computational cost as it trains

an image encoder to translate the input image into a latent prior distribution, requiring a large number of trainable parameters.

In this paper, we propose Generative Medical Segmentation (GMS) to model the segmentation task in a purely generative manner. GMS leverages a pre-trained VAE encoder to obtain a latent representation with semantic information and then only trains a latent mapping model to learn a transformation function from the image latent representation to the mask latent representation. The final segmentation mask in the image space is obtained by decoding the transformed latent representation using a pre-trained VAE decoder. Benefiting from the robust latent space of the pre-trained VAE model, GMS achieves the best performance among five public medical image segmentation datasets across different domains. Furthermore, we perform a cross-domain experiment to demonstrate that the inherent domain generalization ability of GMS is better than other domain generalization methods presented in the literature such as MixStyle [25].

## 2 Methodology

### 2.1 Overview of Generative Medical Segmentation

The Generative Medical Segmentation (GMS) model architecture is shown in Fig. 1. Given a 2D image $I$ and its segmentation mask $M$, we use a pre-trained encoder $\mathcal{E}$ to obtain the latent representations $Z_I$ and $Z_M$ of $I$ and $M$, respectively. $Z_I$ is used as input into the latent mapping model which is trained to predict an estimated latent representation of the mask $M$, corresponding to $\hat{Z}_M$. Finally, $\hat{Z}_M$ is decoded by a pre-trained decoder $\mathcal{D}$ to obtain the segmentation mask prediction $\hat{M}$ in the original image space. Note that the weights of the $\mathcal{E}$ and $\mathcal{D}$ are assumed to be pre-trained and are frozen during both model training and inference, which enables only updating the latent mapping model parameters during training. This reduces the number of trainable parameters to be much smaller than most other state-of-the-art deep learning segmentation models.

The choice of appropriate $\mathcal{E}$ and $\mathcal{D}$ to obtain a representative latent space for both input images and masks is critical for GMS performance. We use the weights of stable diffusion (SD) VAE [16] for $\mathcal{E}$ and $\mathcal{D}$. Since SD VAE was trained on a large natural image dataset [19], it has a rich and diverse latent information representation, leading to a strong zero-shot generalization ability even for medical images. SD VAE can achieve near-perfect image reconstruction, which enables the feasibility of training GMS.

SD VAE contains three down-sampling layers in $\mathcal{E}$ and three up-sampling layers in $\mathcal{D}$, which means the latent representation $Z$ is no longer a one-dimensional vector, but a 3D tensor with spatial information ($Z \in \mathbb{R}^{4\times\frac{H}{8}\times\frac{W}{8}}$ if $I \in \mathbb{R}^{3\times H\times W}$). Such design enables $Z$ to have a rich feature representation, improved reconstruction quality, and enhanced generalization.

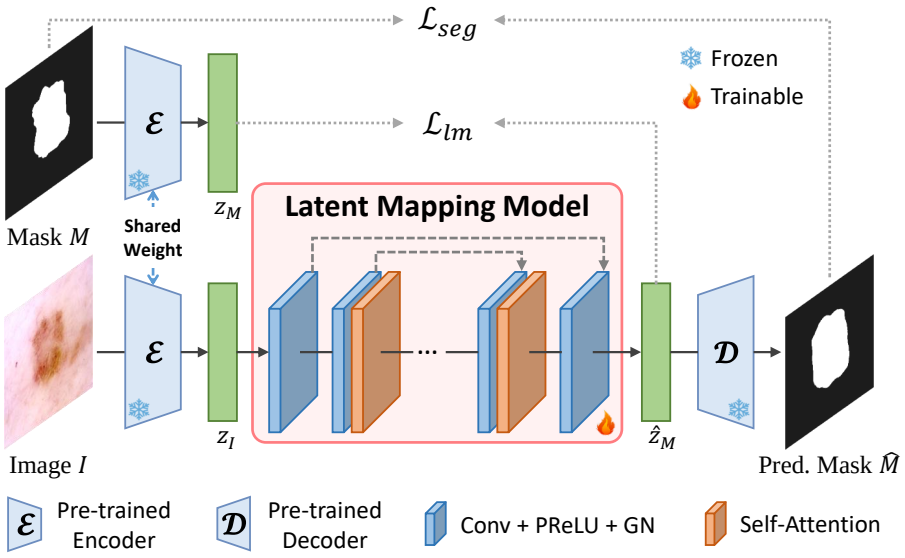


**Fig. 1.** GMS network architecture for 2D medical image segmentation. $\mathcal{E}$ and $\mathcal{D}$ represent a pre-trained SD VAE and weights are frozen. The latent mapping model does not contain down-sampling layers to prevent information loss.

### 2.2 Latent Mapping Model

The latent mapping model is the key component in GMS to learn the transformation function from $Z_I$ to $Z_M$. Instead of using a transformer-based model with high computational cost, we build an UNet-like latent mapping model with 2D convolutions. Since the input and the output are both latent representations, we do not include any down-sampling layers in the mapping model to avoid information loss. Note excluding down-sampling layers is not practical in the original UNet model because the receptive field of convolutional operations is greatly limited if no down-sampling layers are in the model. The latent mapping model structure is shown in the lower middle of Fig. 1. A 2D convolutional layer (Conv) followed by a PReLU activation function and group normalization (GN) layer forms the basic model block. Furthermore, we utilize the self-attention mechanism to better capture global semantic relationships and facilitate feature interaction. Due to the small spatial size of the latent space features, employing the self-attention mechanism results in only a minor increase in computational overhead. We also utilize skip connections to prevent vanishing gradients and the loss of semantic-relevant features.

### 2.3 Loss Functions

We utilize two loss functions for model training, a latent matching loss $\mathcal{L}_{lm}$ in the latent space and a segmentation loss $\mathcal{L}_{seg}$ in the image space. $\mathcal{L}_{lm}$ is formulated

to ensure the similarity between $Z_M$ and $\hat{Z}_M$. Specifically, $\mathcal{L}_{lm}$ is defined as:

$$\mathcal{L}_{lm} = \left\| Z_M - \hat{Z}_M \right\|_2^2. \tag{1}$$

$\mathcal{L}_{seg}$ enforces alignment between the predicted mask $\hat{M}$ and the ground truth mask $M$, even where $\hat{Z}_M$ deviates from $Z_M$. $\mathcal{L}_{seg}$ is defined as:

$$\mathcal{L}_{seg} = 1 - \frac{2 * \sum M * \hat{M}}{\sum M + \sum \hat{M}}, \tag{2}$$

where $*$ denotes element-wise multiplication. The final compound loss function used for model training is:

$$\mathcal{L} = \mathcal{L}_{lm} + \mathcal{L}_{seg}. \tag{3}$$

## 3 Experiments

### 3.1 Dataset

We evaluated the performance of GMS on five public datasets: BUS [24], BUSI [1], GlaS [20], HAM10000 [21] and Kvasir-Instrument [10]. BUS and BUSI are breast lesion ultrasound datasets that contain 163 and 647 images, respectively. GlaS is a colon histology segmentation challenge dataset divided into 85 images for training and 80 images for testing. HAM10000, referred to as HAM in the Tables, is a large dermatoscopic dataset with 10015 images and skin lesion segmentation masks. The Kvasir-Instrument dataset, referred to as Kvasir in the Tables, contains 590 endoscopic images with tool segmentation masks. For datasets not already divided, we randomly select 80% of the images for training and the remaining for testing.

### 3.2 Implementation Details

Our framework is implemented using PyTorch v1.13, and all model training was performed on an NVIDIA A100 40G GPU. We use AdamW [15] as the training optimizer. We utilize the cosine annealing learning rate scheduler to adjust the learning rate in each epoch with the initial learning rate set to $2e^{-3}$. For all experiments, the batch size was set to 8 and the total training epochs were 1000. The input image size is resized to $224 \times 224$, and on-the-fly data augmentations were performed during training including random flip, random rotation, and color jittering in the HSV domain. We set a threshold of 0.5 to change the predicted gray-scale masks to binary masks. We use the Dice coefficient (DSC) and Intersection over Union (IoU) to quantify segmentation performance.

**Table 1.** Trainable parameters for each model expressed in millions (M). Note that EGE-UNet is a well-designed lightweight model for medical image segmentation.

| | UNet | MultiRes UNet | ACC UNet | EGE UNet | Swin UNet | SME SwinUNet | UCTrans Net | GSS | **GMS** |
|---|---|---|---|---|---|---|---|---|---|
| Trainable Params (M) | 14.0 | 7.3 | 16.8 | 0.05 | 27.2 | 169.8 | 66.4 | 49.8 | 1.5 |

### 3.3 Experimental Results

We compared GMS to CNN-based (UNet [17], MultiResUNet [9], ACC-UNet [8] and EGE-UNet [18]), transformer-based (SwinUNet [2], SME-SwinUNet [23] and UCTransNet [22]), and generative (GSS [3]) segmentation models. We also compared against two domain generalization models (MixStyle [25] and DSU [14]) to show the inherent cross-domain generalization ability of GMS. Total trainable parameters for each model are shown in Table 1. Note only EGE-UNet has fewer parameters than GMS, and most have between $\times 10$ and $\times 100$ more parameters. We present the reconstruction results of SD VAE in the supplementary materials to show the suitability of SD VAE to encode images and masks.

**In-domain Segmentation** Table 2 presents in-domain segmentation performance, i.e. training and test set are from the same dataset. GMS achieves the highest scores in terms of both DSC and IoU across all evaluated datasets. Notably, generative segmentation models outperform all discriminative segmentation methods on the two breast ultrasound datasets (BUS and BUSI), which demonstrates the potential of generative approaches to enhance the precision and reliability of breast ultrasound lesion segmentation. For the other datasets, GMS outperforms CNN-based or transformer-based models, suggesting that generative models when carefully designed can provide stronger generalization and are suitable for a wide variety of segmentation tasks.

Qualitative results for each model for all five datasets are shown in Fig. 2. The original image and segmentation mask are the left two columns. Fig. 2 shows GMS segmentation masks are more consistent with the ground truth segmentation masks than other models. UNet and its variants are more likely to give false-positive or false-negative predictions. Additionally, GMS predictions are more likely to be connected, while methods such as SwinUNet predict isolated noise. The qualitative results suggest GMS not only improves segmentation accuracy but also predicts cleaner and more reliable segmentation masks.

**Cross-domain Segmentation** We evaluated all models on their ability to segment cross-domain images to demonstrate model domain generalization ability. Specifically, we train the model using the training set for one dataset (as with the in-domain experiment) but evaluate performance on the test set of a different dataset. This experiment was performed with the BUS and BUSI datasets interchangeably as training and test sets since they are the same modalities (breast

**Table 2.** Quantitative performance for in-domain segmentation. Best and second-best performances are bold and underlined, respectively. † indicates fewer trainable parameters than GMS.

| Model | BUS | | BUSI | | GlaS | | HAM | | Kvasir | |
|---|---|---|---|---|---|---|---|---|---|---|
| | DSC | IoU | DSC | IoU | DSC | IoU | DSC | IoU | DSC | IoU |
| UNet | 81.50 | 70.77 | 72.27 | 63.00 | 87.99 | 80.01 | 92.24 | 86.93 | 93.82 | 89.23 |
| MultiResUNet | 80.41 | 70.33 | 72.43 | 62.59 | 88.34 | 80.34 | 92.74 | 87.60 | 92.31 | 87.03 |
| ACC-UNet | 83.40 | 73.51 | 77.19 | 68.51 | <u>88.60</u> | <u>80.84</u> | 93.20 | 88.44 | <u>93.95</u> | <u>89.73</u> |
| EGE-UNet† | 72.79 | 61.96 | 75.17 | 60.23 | 83.25 | 71.31 | <u>93.90</u> | 88.50 | 92.65 | 86.30 |
| SwinUNet | 80.37 | 69.75 | 76.06 | 66.10 | 86.44 | 76.89 | 93.51 | 88.68 | 92.02 | 85.83 |
| SME-SwinUNet | 78.87 | 67.13 | 73.93 | 62.70 | 83.72 | 72.77 | 92.71 | 87.21 | 93.32 | 88.27 |
| UCTransNet | 83.44 | 73.74 | 76.55 | 67.50 | 87.17 | 78.80 | 93.45 | <u>88.73</u> | 93.27 | 88.48 |
| GSS | <u>84.86</u> | <u>77.58</u> | <u>79.56</u> | <u>71.22</u> | 87.41 | 79.17 | 92.92 | 87.98 | 93.66 | 89.15 |
| GMS (Ours) | **88.42** | **80.56** | **81.43** | **72.58** | **88.98** | **81.16** | **94.11** | **89.68** | **94.24** | **90.02** |

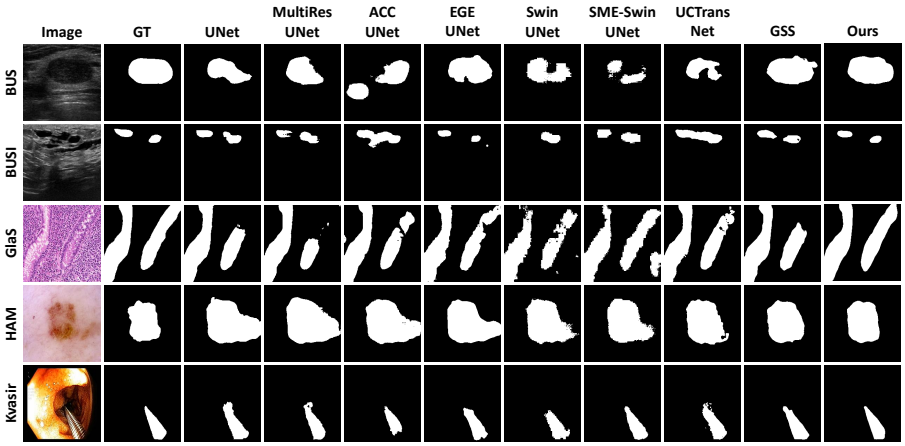


**Fig. 2.** Qualitative segmentation performance of each model for the five datasets.

ultrasound) but acquired from different centers and vendors. Therefore, the data distributions of the training and test sets are not aligned. Table 3 shows the quantitative performance for the cross-domain segmentation task. Our GMS model outperforms all other models, including the two domain generalization methods (MixStyle and DSU). Interestingly, generative segmentation models tend to have better performance than discriminative models when the training set is small (BUS to BUSI). The increase in performance we believe is due to the latent representations derived from the per-trained VAE being more domain-agnostic than learned parameters, which improves domain generalization. Additionally, GMS is more lightweight compared with the other generative model (GSS), which further reduces the likelihood of overfitting the model to the training set.

**Table 3.** Quantitative performance for cross-domain segmentation. A to B indicates A for training and B for testing. Best and second-best performances are bold and underlined, respectively.

| Model | BUSI to BUS | | BUS to BUSI | |
|---|---|---|---|---|
| | DSC | IoU | DSC | IoU |
| UNet | 62.99 | 56.63 | 53.83 | 44.09 |
| MultiResUNet | 61.53 | 52.76 | 56.25 | 46.18 |
| ACC-UNet | 64.60 | 57.23 | 47.80 | 39.24 |
| EGE-UNet† | 69.04 | 59.74 | 54.46 | 43.91 |
| SwinUNet | <u>78.38</u> | <u>68.42</u> | 57.47 | 46.30 |
| SME-SwinUNet | 74.78 | 63.51 | 58.28 | 47.06 |
| UCTransNet | 72.76 | 64.30 | 56.94 | 46.28 |
| MixStyle | 73.07 | 67.03 | 57.97 | 48.82 |
| DSU | 66.15 | 59.57 | 56.70 | 46.90 |
| GSS | 68.74 | 62.57 | <u>58.72</u> | <u>49.46</u> |
| GMS (Ours) | **80.31** | **71.99** | **61.60** | **53.09** |

**Table 4.** Quantitative segmentation performance on three datasets for ablation study using different loss functions.

| $\mathcal{L}_{lm}$ | $\mathcal{L}_{seg}$ | Dataset | DSC | IoU |
|---|---|---|---|---|
| ✓ | | BUSI | 80.25 | 71.26 |
| | ✓ | BUSI | 78.75 | 69.87 |
| ✓ | ✓ | BUSI | **81.43** | **72.58** |
| ✓ | | HAM | 93.92 | 89.41 |
| | ✓ | HAM | 93.64 | 88.99 |
| ✓ | ✓ | HAM | **94.11** | **89.68** |
| ✓ | | Kvasir | 92.93 | 88.28 |
| | ✓ | Kvasir | 93.00 | 88.47 |
| ✓ | ✓ | Kvasir | **94.24** | **90.02** |

**Ablation Study** The BUSI, HAM10000, and Kvasir-Instrument datasets were used to perform an ablation study on different loss function combinations. As shown in Fig. 4, the compound loss ($\mathcal{L}_{lm} + \mathcal{L}_{seg}$) always has the best segmentation performance regardless of dataset modality or size. Interestingly, different datasets have different supervision preferences. GMS only using $\mathcal{L}_{lm}$ for model training performs better on BUSI and HAM10000 datasets, which implies supervision in the latent space is more effective compared to the image space. However, GMS performance is better for $\mathcal{L}_{seg}$ when training on the Kvasir-Instrument dataset, indicating supervision in the image space is more important. The compound loss having the best performance suggests that supervision in the image and latent space are both important to maximize performance.

## 4 Conclusion

We presented Generative Medical Segmentation (GMS) to perform medical image segmentation. Unlike other methods where a discriminative model is trained, GMS leverages a powerful pre-trained VAE encoder to obtain latent representations of images and masks. Next, our novel lightweight latent mapping model learns a transformation function from image latent representations to mask latent representations. Finally, a pre-trained VAE decoder obtains a predicted segmentation mask in the image space using the predicted latent representation of the mask. Extensive experiments on five datasets show that GMS outperforms the state-of-the-art discriminative segmentation models such as ACC-UNet. Moreover, the domain generalization ability of GMS is stronger than even well-designed domain generalization models, like DSU and MixStyle, due to the

domain-agnostic latent embedding space used by GMS. One key limitation is that currently GMS can only segment 2D medical images, due to stable diffusion variational autoencoder (SD VAE) being used for the pre-trained encoding and decoding networks. In the future, we will explore extending GMS to 3D medical images by selecting an appropriate pre-trained 3D model and adapting the latent mapping model.

## References


1. Al-Dhabyani, W., Gomaa, M., Khaled, H., Fahmy, A.: Dataset of breast ultrasound images. Data in brief **28**, 104863 (2020)
2. Cao, H., Wang, Y., Chen, J., Jiang, D., Zhang, X., Tian, Q., Wang, M.: Swin-unet: Unet-like pure transformer for medical image segmentation. In: European conference on computer vision. pp. 205–218. Springer (2022)
3. Chen, J., Lu, J., Zhu, X., Zhang, L.: Generative semantic segmentation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 7111–7120 (2023)
4. Chen, T., Wang, C., Shan, H.: Berdiff: Conditional bernoulli diffusion model for medical image segmentation. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 491–501. Springer (2023)
5. Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., Houlsby, N.: An image is worth 16x16 words: Transformers for image recognition at scale. In: International Conference on Learning Representations (2021), `https://openreview.net/forum?id=YicbFdNTTy`
6. Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., Bengio, Y.: Generative adversarial nets. Advances in neural information processing systems **27** (2014)
7. Huo, J., Vakharia, V., Wu, C., Sharan, A., Ko, A., Ourselin, S., Sparks, R.: Brain lesion synthesis via progressive adversarial variational auto-encoder. In: International Workshop on Simulation and Synthesis in Medical Imaging. pp. 101–111. Springer (2022)
8. Ibtehaz, N., Kihara, D.: Acc-unet: A completely convolutional unet model for the 2020s. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 692–702. Springer (2023)
9. Ibtehaz, N., Rahman, M.S.: Multiresunet: Rethinking the u-net architecture for multimodal biomedical image segmentation. Neural networks **121**, 74–87 (2020)
10. Jha, D., Ali, S., Emanuelsen, K., Hicks, S.A., Thambawita, V., Garcia-Ceja, E., Riegler, M.A., de Lange, T., Schmidt, P.T., Johansen, H.D., et al.: Kvasir-instrument: Diagnostic and therapeutic tool segmentation dataset in gastrointestinal endoscopy. In: MultiMedia Modeling: 27th International Conference, MMM 2021, Prague, Czech Republic, June 22–24, 2021, Proceedings, Part II 27. pp. 218–229. Springer (2021)
11. Karras, T., Aittala, M., Hellsten, J., Laine, S., Lehtinen, J., Aila, T.: Training generative adversarial networks with limited data. Advances in neural information processing systems **33**, 12104–12114 (2020)
12. Kingma, D.P., Welling, M.: Auto-encoding variational bayes. arXiv preprint arXiv:1312.6114 (2013)

13. Li, D., Yang, J., Kreis, K., Torralba, A., Fidler, S.: Semantic segmentation with generative models: Semi-supervised learning and strong out-of-domain generalization. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 8300–8311 (2021)
14. Li, X., Dai, Y., Ge, Y., Liu, J., Shan, Y., DUAN, L.: Uncertainty modeling for out-of-distribution generalization. In: International Conference on Learning Representations (2022), https://openreview.net/forum?id=6HN7LHyzGgC
15. Loshchilov, I., Hutter, F.: Decoupled weight decay regularization. In: International Conference on Learning Representations (2019), https://openreview.net/forum?id=Bkg6RiCqY7
16. Rombach, R., Blattmann, A., Lorenz, D., Esser, P., Ommer, B.: High-resolution image synthesis with latent diffusion models. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 10684–10695 (2022)
17. Ronneberger, O., Fischer, P., Brox, T.: 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. pp. 234–241. Springer (2015)
18. Ruan, J., Xie, M., Gao, J., Liu, T., Fu, Y.: Ege-unet: an efficient group enhanced unet for skin lesion segmentation. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 481–490. Springer (2023)
19. Schuhmann, C., Beaumont, R., Vencu, R., Gordon, C., Wightman, R., Cherti, M., Coombes, T., Katta, A., Mullis, C., Wortsman, M., et al.: Laion-5b: An open large-scale dataset for training next generation image-text models. Advances in Neural Information Processing Systems **35**, 25278–25294 (2022)
20. Sirinukunwattana, K., Pluim, J.P., Chen, H., Qi, X., Heng, P.A., Guo, Y.B., Wang, L.Y., Matuszewski, B.J., Bruni, E., Sanchez, U., et al.: Gland segmentation in colon histology images: The glas challenge contest. Medical image analysis **35**, 489–502 (2017)
21. Tschandl, P., Rosendahl, C., Kittler, H.: The ham10000 dataset, a large collection of multi-source dermatoscopic images of common pigmented skin lesions. Scientific data **5**(1), 1–9 (2018)
22. Wang, H., Cao, P., Wang, J., Zaiane, O.R.: Uctransnet: rethinking the skip connections in u-net from a channel-wise perspective with transformer. In: Proceedings of the AAAI conference on artificial intelligence. vol. 36, pp. 2441–2449 (2022)
23. Wang, Z., Min, X., Shi, F., Jin, R., Nawrin, S.S., Yu, I., Nagatomi, R.: Smeswin unet: Merging cnn and transformer for medical image segmentation. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 517–526. Springer (2022)
24. Yap, M.H., Pons, G., Marti, J., Ganau, S., Sentis, M., Zwiggelaar, R., Davison, A.K., Marti, R.: Automated breast ultrasound lesions detection using convolutional neural networks. IEEE journal of biomedical and health informatics **22**(4), 1218–1226 (2017)
25. Zhou, K., Yang, Y., Qiao, Y., Xiang, T.: Mixstyle neural networks for domain generalization and adaptation. International Journal of Computer Vision pp. 1–15 (2023)

# Generative Medical Segmentation (Supplementary Materials)

No Author Given

No Institute Given

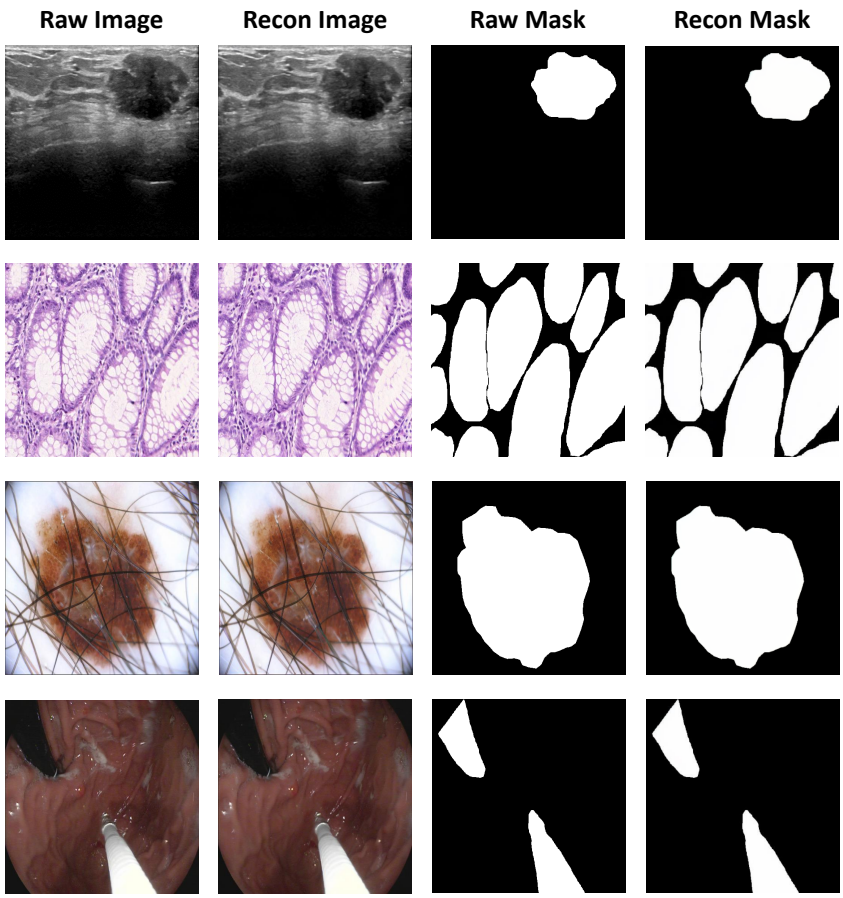


**Fig. 1.** Visualization results of original and reconstructed images and masks. Images and masks were input into the pre-trained Stable Diffusion (SD) VAE encoder to obtain latent representations and then passed through the pre-trained VAE decoder to get corresponding reconstructed images and masks. SD VAE achieves almost perfect reconstruction even for complex inputs. Dice scores between the original and reconstructed masks are above 99.5% which further confirms the conclusion of good reconstruction.