Generative Medical Segmentation

Jiayu Huo[1]*, Xi Ouyang[2], Sébastien Ourselin[1], Rachel Sparks[1]

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

Abstract

Rapid advancements in medical image segmentation performance have been significantly driven by the development of Convolutional Neural Networks (CNNs) and Vision Transformers (ViTs). These models follow discriminative pixel-wise classification learning paradigm and often have limited ability to generalize across diverse medical imaging datasets. In this manuscript, we introduce Generative Medical Segmentation (GMS), a novel generative approach to perform image segmentation. GMS employs a robust pre-trained vision foundation model to extract latent representations for images and corresponding ground truth masks, followed by a lightweight model that learns a mapping function from the image to the mask in the latent space. Once trained, the model can generate estimated segmentation masks using the pre-trained vision foundation model to decode the predicted latent mask representation back into image space. The design of GMS leads to fewer trainable parameters in the model, reducing the risk of overfitting and enhancing its generalization capability. Our experimental analysis across five open-source datasets in different medical imaging domains demonstrates GMS outperforms existing discriminative and generative segmentation models. Furthermore, GMS is able to generalize well across datasets of the same imaging modality from different centers. Our experiments suggest GMS offers a scalable and effective solution for medical image segmentation.

Code — https://github.com/King-HAW/GMS

Introduction

Image segmentation plays a crucial role in the field of medical image analysis enabling automated, precise delineation of anatomical and pathological structures. Automated segmentation enables clinicians to obtain detailed visualizations and quantitative assessments of lesions and other structural anomalies, facilitating computer-aided diagnosis, treatment planning, and monitoring disease progression, thereby enhancing the precision and efficacy of therapeutic interventions and improving patient outcomes (Al-Dhabyani et al. 2020; Tschandl, Rosendahl, and Kittler 2018; Jha et al. 2021).

Inaugural deep learning models designed for medical image segmentation, such as UNet (Ronneberger, Fischer, and Brox 2015) and its various adaptations (Ruan et al. 2023; Ibtehaz and Kihara 2023), have significantly advanced the field of medical imaging analysis. These early deep learning image segmentation models leverage convolution kernels to learn local patch representations from large amounts of labeled data. Despite their successes, models based on Convolutional Neural Networks (CNNs) often have a large number of trainable parameters which can introduce challenges in model training and increase the likelihood of overfitting when training datasets are small. Additionally, the limited receptive field of the convolution kernel makes it difficult for CNN-based models to learn global context information that can provide important guidance during image segmentation. Moreover, CNN-based models struggle with generalizing to unseen domains, leading to potentially substantial performance drops when the test dataset distribution is shifted from the training dataset distribution.

The Vision Transformer (ViT) (Dosovitskiy et al. 2021) has recently been presented as a powerful alternative to CNN-based segmentation models in medical imaging analysis. ViT can capture global semantic information that the convolution kernel is unable to represent. Transformer-based segmentation models, such as UCTransNet (Wang et al. 2022a) and Swin-Unet (Cao et al. 2022), leverage the transformer architecture to represent images as sequences of patches, enabling the model to learn relationships across the entire image. However, transformer-based models are required to be 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 costs needed for the multi-head attention module pose practical challenges for real-time applications and deployment in environments with limited computational resources. Furthermore, due to the large number of parameters in transformer-based models, there is an increased risk of overfitting when training on small datasets with subsequent challenges of poor generalization to out-of-domain datasets under such conditions.

Generative models, such as Generative Adversarial Networks (GANs) (Goodfellow et al. 2014) and Variational Autoencoders (VAEs) (Kingma and Welling 2013), are often adopted as data augmentation techniques to improve

*Corresponding author

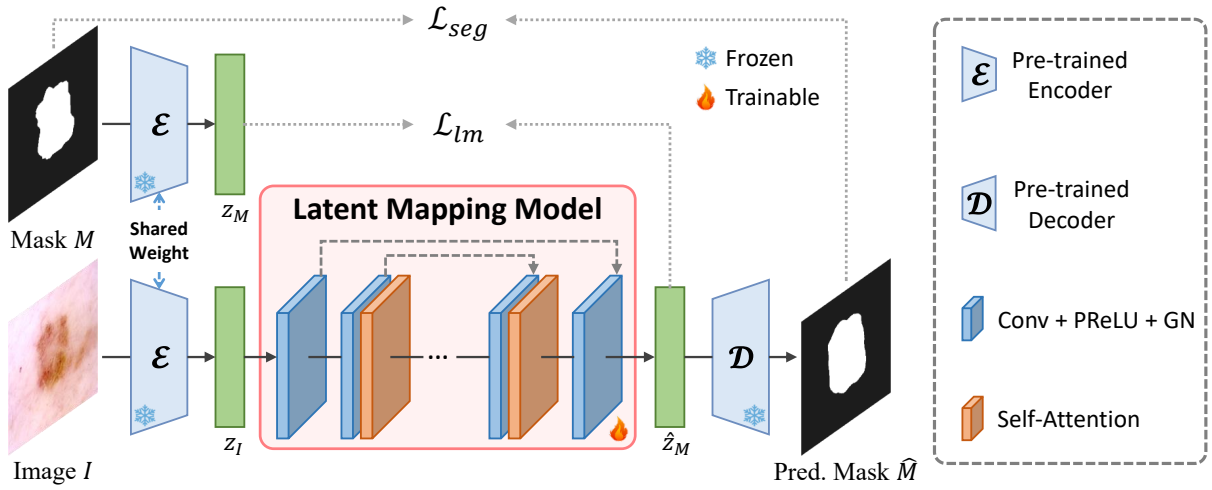

Figure 1: GMS network architecture for 2D medical image segmentation. $\mathcal{E}$ and $\mathcal{D}$ represent a pre-trained vision foundation model and weights are frozen. We utilize the model weights from the Stable Diffusion VAE for $\mathcal{E}$ and $\mathcal{D}$. The latent mapping model (orange box) contains convolution blocks and self-attention blocks but does not contain down-sampling layers. Such a design helps to preserve the spatial information in the input feature vectors. Conv indicates the 2D convolution operation, and GN represents the Group Normalization.

the performance of segmentation models (Huo et al. 2022). However, GANs suffer from mode collapse and may fail to give plausible outputs when the number of training samples is small (Karras et al. 2020). Image-to-image translation models have been used to perform image segmentation in a generative manner, where the image serves as the input and the mask as the prediction. To date, the performance of image-to-image models is well below state-of-the-art segmentation model performance (Li et al. 2021). Recently, MedSegDiff-V2 (Wu et al. 2024) utilized a diffusion model for medical image segmentation, where a condition model encoded images into the feature space for mask generation. However, diffusion-based approaches require repetitive denoising steps which lead to longer inference times. GSS (Chen et al. 2023) is a generative semantic segmentation framework designed for semantic image segmentation, where Vector Quantized Variational Autoencoder (VQ-VAE) (Van Den Oord, Vinyals et al. 2017) was employed to project the image and mask into a latent space, and an additional image encoder was designed and trained to match the latent distributions between the mask and image. However, GSS has high computational costs as the additional image encoder is complex, requiring a large number of trainable parameters, to translate the input image into a latent prior distribution.

In this paper, we present Generative Medical Segmentation (GMS) to perform image segmentation in a generative manner. GMS leverages a pre-trained image encoder to obtain latent representations containing semantic information for input images and masks, and then a latent mapping model learns 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 mask latent representation using a pre-trained image decoder paired with the pre-trained image encoder. In this approach, only the latent mapping model parameters are learned from the training dataset. The pre-trained image encoder and decoder are obtained from a vision foundation model trained on a large, general dataset. Therefore, the latent representations are more general to unseen data compared to models trained only on images for the desired specific task. We demonstrate GMS achieves the best performance on five open-source medical image segmentation datasets across different domains. Furthermore, we perform an external validation experiment to demonstrate that the inherent domain generalization ability of GMS is better than other domain generalization methods.

Related Works

Medical Image Segmentation

Medical image segmentation has experienced rapid advancements in the last decade due to the development of deep-learning techniques. The encoder-decoder architecture with skip connections enables accurate image segmentation by combining low-level and high-level features to perform pixel-wise prediction, making UNet (Ronneberger, Fischer, and Brox 2015) a benchmark method across various medical image segmentation tasks. Subsequent model enhancements such as MultiResUNet (Ibtehaz and Rahman 2020) and ACC-UNet (Ibtehaz and Kihara 2023) have been implemented using the basic UNet architecture and integrating the residual blocks or redesigning the hierarchical feature fusion pipeline to gain improved segmentation performance. nnU-Net (Isensee et al. 2021) established a guideline for tailoring the receptive field size of convolution kernels and network depth to specific tasks while incorporating extensive data augmentation during model training to improve segmentation performance.

The Vision Transformer (ViT) introduced the multi-head attention mechanism, which captures long-range feature dependencies across patches in the image, leading to stronger feature representations for image segmentation compared to CNN-based models. This ability to model relationships between distant pixels or features has proven highly beneficial for medical image segmentation, where understanding the broader context is often crucial to performing the task well. ViT-based segmentation models (Cao et al. 2022; Wang et al. 2022b,a) have competitive results compared against traditional CNN-based models.

Generative & Foundation Models

Generative models are commonly designed for image synthesis and image-to-image translation tasks. For image synthesis, GANs (Goodfellow et al. 2014) and VAEs (Kingma and Welling 2013) are often leveraged to generate more data for downstream model training (Huo et al. 2022; Chaitanya et al. 2021), especially in the context of medical image segmentation, as the cost of obtaining large, annotated medical imaging datasets is high. Recently, studies have explored diffusion models to create more training instances and alleviate data scarcity (Ye et al. 2023). However, the iterative denoising process in diffusion models results in a longer inference time compared to GAN or VAE-based approaches. For image-to-image translation, models developed on CNNs (Kong et al. 2021) or ViT (Liu et al. 2023) show satisfactory results on the MRI missing modality completion task. Currently, few generative models are designed for performing image segmentation directly. GSS (Chen et al. 2023) is the exception, this model employs VQ-VAE (Van Den Oord, Vinyals et al. 2017) to discretize image and mask pairs into a finite set of latent codes, which are then reconstructed back into the image space. An independent image encoder is trained to match the image latent codes to the mask latent codes.

Foundation models, such as Stable Diffusion (Rombach et al. 2022) and Segment Anything (SAM) (Kirillov et al. 2023), are trained on large-scale datasets and are designed to generalize across a wide range of tasks. These models are designed to serve as a versatile starting point for numerous tasks. Stable Diffusion utilizes a VAE to first encode the image into a latent space and leverages a UNet to iteratively denoise and reconstruct the latent embeddings, guiding the generation process towards a high-quality output. SAM is designed for image segmentation with a prompt that allows user interactions to adapt the model to various segmentation tasks with no or minimal fine-tuning. Together, these models exemplify the power and flexibility of foundation models in addressing diverse and complex tasks such as image segmentation.

Methodology

Architecture Overview

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

Image Tokenizer

The pre-trained encoder $\mathcal{E}$ and decoder $\mathcal{D}$ are treated as an image tokenizer as they map an input from the image space to the latent space ($\mathcal{E}$) or the latent to image space ($\mathcal{D}$). The choice of an appropriate, paired $\mathcal{E}$ and $\mathcal{D}$ to obtain a representative latent space for both input images and masks is critical for GMS performance. In this work, we use the weights of Stable Diffusion (SD) VAE (Rombach et al. 2022) for $\mathcal{E}$ and $\mathcal{D}$. Since SD-VAE was trained on a large natural image dataset (Schuhmann et al. 2022), 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 (Rombach et al. 2022).

SD-VAE is comprised of three down-sampling blocks in $\mathcal{E}$ and three up-sampling blocks in $\mathcal{D}$. The latent representation Z is a 3D tensor containing 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 strong feature representation, improved reconstruction quality and enhanced the generalization of the latent representation.

Latent Mapping Model (LMM)

The latent mapping model (LMM) is the key component in GMS to map from Z_I to Z_M. We build the LMM to be specifically lightweight by using 2D convolutions rather than transformer blocks that require many parameters for multi-head attention. Besides, we do not include any down-sampling layers in LMM to avoid spatial information loss. Note excluding down-sampling layers is not practical in the original UNet model because the receptive fields of the convolution operations are greatly limited if no down-sampling layers are used in the model. However, in the LMM they are not required as the latent representation has already been down-sampled by $\mathcal{E}$. Skip connections between convolutional layers are added to prevent vanishing gradients and the loss of semantic-relevant features.

The LMM architecture is shown in the lower middle of Figure 1 (orange box). Given Z_I, which is acquired from the pre-trained encoder $\mathcal{E}$, it first goes through two convolution blocks where each block consists of a 2D convolutional layer (Conv), a PReLU activation function, and group normalization (GN) layer to obtain the feature vector $\mathbf{F}$.

Next, a self-attention mechanism layer is added to better capture global semantic relationships and facilitate feature interaction within $\mathbf{F}$. Specifically, we use three inde-

Type	Model	Trainable Params (M)	BUS			BUSI		
			DSC↑	IoU↑	HD95↓	DSC↑	IoU↑	HD95↓
CNN	UNet	14.0	81.50	70.77	17.68	72.27	63.00	35.42
	MultiResUNet	7.3	80.41	70.33	19.22	72.43	62.59	34.19
	ACC-UNet	16.8	83.40	73.51	16.49	77.19	68.51	25.49
	nnUNet	20.6	<u>85.71</u>	<u>78.68</u>	<u>11.43</u>	79.45	70.99	<u>22.13</u>
	EGE-UNet†	0.05	72.79	61.96	27.73	75.17	60.23	29.51
Transformer	SwinUNet	27.2	80.37	69.75	20.49	76.06	66.10	28.69
	SME-SwinUNet	169.8	78.87	67.13	22.19	73.93	62.70	30.45
	UCTransNet	66.4	83.44	73.74	16.33	76.55	67.50	25.46
Generative	MedSegDiff-V2	129.4	83.23	74.36	17.02	71.32	62.73	38.47
	SDSeg	329.0	82.47	73.45	20.53	72.76	63.52	36.79
	GSS	49.8	84.86	77.58	22.42	<u>79.56</u>	<u>71.22</u>	28.20
	GMS (Ours)	1.5	**88.42**	**80.56**	**6.79**	**81.43**	**72.58**	**19.50**

Table 1: Quantitative segmentation performance on two ultrasound datasets. The best and second-best performances are bold and underlined, respectively. † indicates fewer trainable parameters than GMS.

Type	Model	GlaS			HAM10000			Kvasir-Instrument		
		DSC↑	IoU↑	HD95↓	DSC↑	IoU↑	HD95↓	DSC↑	IoU↑	HD95↓
CNN	UNet	87.99	80.01	18.45	92.24	86.93	13.74	93.82	89.23	8.71
	MultiResUNet	88.34	80.34	17.42	92.74	87.60	13.02	92.31	87.03	9.49
	ACC-UNet	<u>88.60</u>	<u>80.84</u>	<u>17.14</u>	93.20	88.44	10.83	93.91	89.73	8.74
	nnUNet	87.25	78.24	20.07	93.83	<u>89.32</u>	<u>9.43</u>	<u>93.95</u>	**90.20**	8.51
	EGE-UNet†	83.25	71.31	28.79	<u>93.90</u>	88.50	10.01	92.65	86.30	9.04
Transformer	SwinUNet	86.44	76.89	19.63	93.51	88.68	10.46	92.02	85.83	9.15
	SME-SwinUNet	83.72	72.77	26.23	92.71	87.21	12.53	93.32	88.27	8.91
	UCTransNet	87.17	78.80	20.79	93.45	88.73	10.91	93.27	88.48	8.84
Generative	MedSegDiff-V2	86.82	77.05	19.96	92.28	87.02	13.02	92.29	87.21	9.06
	SDSeg	86.76	76.23	21.41	92.54	87.53	12.29	91.23	86.54	9.38
	GSS	87.41	79.17	19.81	92.92	87.98	11.29	93.66	89.15	<u>7.25</u>
	GMS (Ours)	**88.98**	**81.16**	**16.32**	**94.11**	**89.68**	**9.32**	**94.24**	<u>90.02</u>	**7.03**

Table 2: Quantitative segmentation performance on three medical datasets of different modalities. The best and second-best performances are bold and underlined, respectively. † indicates fewer trainable parameters than GMS.

pendent convolution operations to generate query Q, key K, and value V, respectively:

$$Q = W_Q \cdot \mathbf{F} + b_Q, K = W_K \cdot \mathbf{F} + b_K, V = W_V \cdot \mathbf{F} + b_V. \quad (1)$$

Here W is the convolution kernel matrix and b is the learnable bias.

Then the self-attention for the query, key, and value is computed as:

$$\text{Attention}(Q, K, V) = \text{softmax}(\frac{QK^T}{\sqrt{d_k}})V, \quad (2)$$

where d_k denotes the feature channel of K, and softmax denotes the softmax normalization function. Due to the small spatial size of the $\mathbf{F}$, employing the self-attention mechanism allows for the efficient capture of long-range dependencies and interactions within the latent representations.

Loss Functions

Two loss functions are used to guide model training, a 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 enforce similarity between Z_M and $\hat{Z}_M$. Specifically, $\mathcal{L}_{lm}$ is defined as:

$$\mathcal{L}_{lm} = \left\| Z_M - \hat{Z}_M \right\|_2^2. \quad (3)$$

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

$$\mathcal{L}_{seg} = 1 - \frac{2 * \sum M \odot \hat{M}}{\sum M + \sum \hat{M}}, \quad (4)$$

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

$$\mathcal{L} = \mathcal{L}_{lm} + \mathcal{L}_{seg}. \quad (5)$$

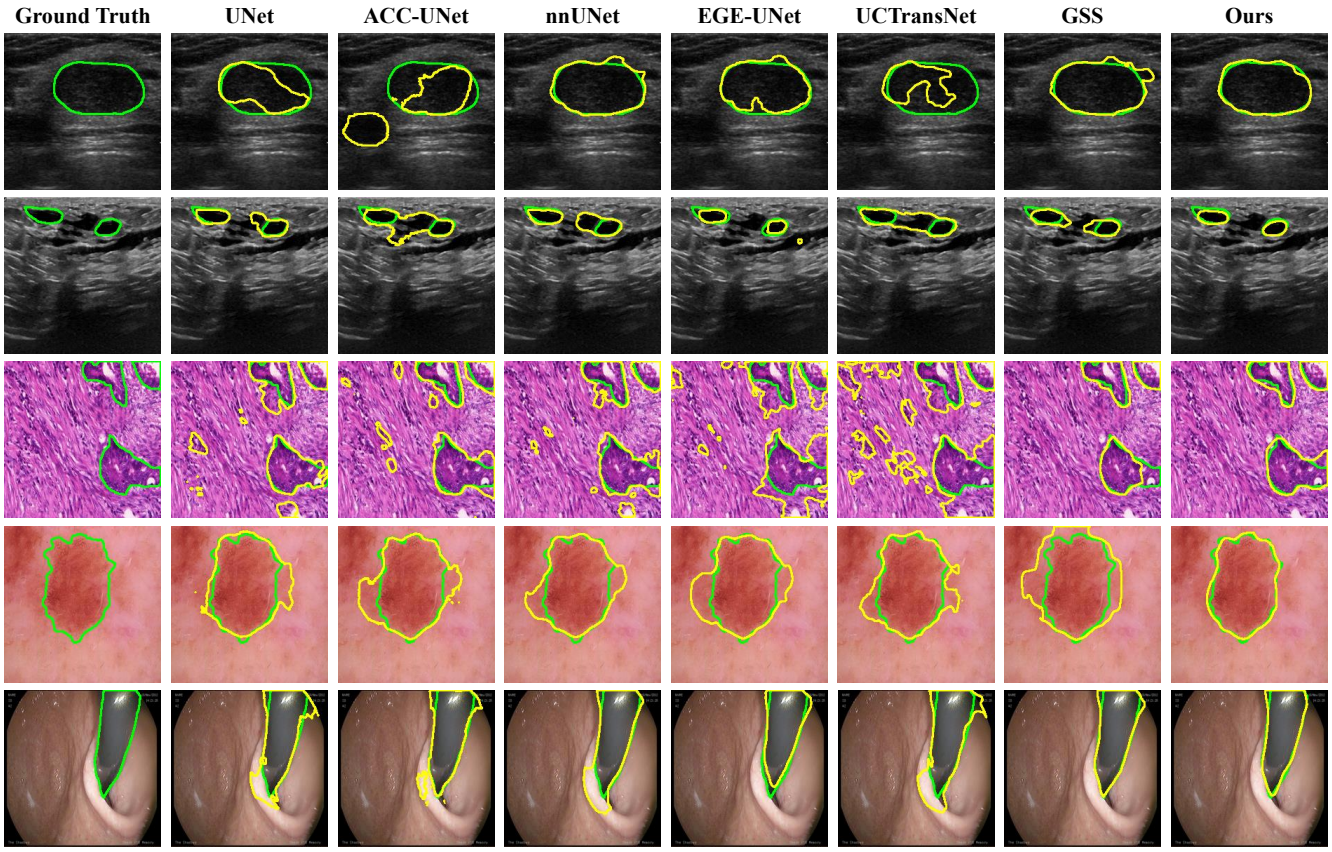

Figure 2: Exemplar segmentation results. From top to bottom are images from the BUS, BUSI, GlaS, HAM10000, and Kvasir-Instrument datasets. The green contours are the ground truth, and the yellow contours are the model predictions. Zoom in for more details.

Experiments

Datasets

We evaluated the performance of GMS on five public datasets: BUS, BUSI, GlaS, HAM10000, and Kvasir-Instrument. BUS (Yap et al. 2017) and BUSI (Al-Dhabyani et al. 2020) are breast lesion ultrasound datasets that contain 163 and 647 images, respectively. GlaS (Sirinukunwattana et al. 2017) is a colon histology segmentation challenge dataset divided into 85 images for training and 80 images for testing. HAM10000 (Tschandl, Rosendahl, and Kittler 2018) is a large dermatoscopic dataset that consists of 10015 images with skin lesion segmentation masks. The Kvasir-Instrument dataset (Jha et al. 2021) contains 590 endoscopic images with tool segmentation masks. For all datasets except GlaS, we randomly select 80% of the images for training and the remaining 20% for testing. We keep the official training and testing set split of GlaS.

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 (Loshchilov and Hutter 2019) 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 is resized to 224×224, 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 binarize the predicted values. We quantify segmentation performance using Dice coefficient (DSC), Intersection over Union (IoU), and Hausdorff Distance 95th percentile (HD95).

Comparison with State-of-the-Art Models

We compare GMS with other state-of-the-art methods to evaluate its performance, including CNN-based methods: UNet (Ronneberger, Fischer, and Brox 2015), MultiResUNet (Ibtehaz and Rahman 2020), ACC-UNet (Ibtehaz and Kihara 2023), nnUNet (Isensee et al. 2021) and EGE-UNet (Ruan et al. 2023); Transformer-based methods: Swin-UNet (Cao et al. 2022), SME-SwinUNet (Wang et al. 2022b) and UCTransNet (Wang et al. 2022a); and generative methods: MedSegDiff-V2 (Wu et al. 2024), SDSeg (Lin et al. 2024) and GSS (Chen et al. 2023)). We also compared against two domain generalization models: MixStyle (Zhou et al. 2023) and DSU (Li et al. 2022) to evaluate the inherent

Model	BUSI to BUS		BUS to BUSI	
	DSC↑	HD95↓	DSC↑	HD95↓
UNet	62.99	47.26	53.83	96.81
MultiResUNet	61.53	53.97	56.25	94.31
ACC-UNet	64.60	42.87	47.80	135.24
nnUNet	<u>78.39</u>	<u>20.53</u>	<u>59.13</u>	<u>89.32</u>
EGE-UNet†	69.04	34.63	54.46	105.23
SwinUNet	78.38	21.94	57.47	91.63
SME-SwinUNet	74.78	25.81	58.28	91.26
UCTransNet	72.76	28.47	56.94	94.32
MixStyle	73.07	26.52	57.97	93.54
DSU	66.15	40.03	56.70	95.31
MedSegDiff-V2	69.56	32.51	55.21	98.57
SDSeg	74.03	26.32	57.03	94.61
GSS	68.74	35.74	58.72	92.57
GMS (Ours)	**80.31**	**18.55**	**61.60**	**85.25**

Table 3: Quantitative performance for domain generalization segmentation. A to B indicates A for training and B for testing. Best and second-best performances are bold and underlined, respectively. † indicates fewer trainable parameters than GMS.

domain generalization ability of GMS.

Quantitative comparisons of all models on the two ultrasound datasets are presented in Table 1. GMS achieves the highest DSC, IoU, and HD95. We also present the trainable parameter of each model in Table 1, only EGE-UNet has fewer trainable parameters than GMS, and most models have between $\times 10$ and $\times 100$ more parameters. GMS achieves a 2.71% and 1.87% improvement in the DSC metric on the BUS and BUSI datasets, respectively, compared to the second-best model. Additionally, for IoU and HD95 metrics, our approach shows improvement of 1.88% and 4.64, respectively, over the second-best model. Transformer-based segmentation models do not show competitive results on these datasets, which indicates that the intrinsic long-range modeling capability of the transformer block may not be suitable for ultrasound images which lack chromatic information that often aids in distinguishing different tissues. The limited texture and low contrast in ultrasound images might reduce the effectiveness of the multi-head attention module in transformer-based models. nnUNet, as a powerful auto-configuration segmentation model, beats transformer-based models and even some generative models, and is the second-best model on the BUS dataset. Notably, segmentation performance is not correlated to the models' number of trainable parameters but benefits from plausible model design and the robust representations provided by the pre-trained vision foundation model. However, models (e.g. EGE-UNet) that contain too few trainable parameters, may lack the capacity to capture complex patterns and relationships in the images, leading to underfitting and poor performance on both datasets. In contrast, models (e.g. SME-SwinUNet and MedSegDiff-V2) with an excess of parameters can easily overfit the training dataset, memorizing rather than generalizing, which compromises performance on the test set.

Table 2 presents quantitative results on the other three datasets, where all images are RGB representations. GMS achieves the best segmentation performance except for the IoU metric on the Kvasir-Instrument dataset. It is worth noting that not all generative segmentation models outperform other discriminative models, which proves the importance of design when applying the generative model framework. MedSegDiff-V2 employs an encoder-decoder model to embed images as conditions for guiding the denoising step, yet its performance remains below CNN and transformer based models. SDSeg utilized the Stable Diffusion model to generate latent representations and further decode them as predicted masks. Additionally, SDSeg proposed a trainable encoder to embed the image into the latent space as the condition for the denoising step. This design does not maximize the use of the knowledge encapsulated in the pre-trained vision foundation model, which may account for its poorer performance. GMS outperforms both CNN-based and transformer-based models, suggesting that generative models when carefully designed can be suitable for a wide variety of segmentation tasks.

Domain Generalization Ability

We evaluated all models on their ability to segment images within the same modality but collected at different centers on different machines to demonstrate model domain generalization ability. Specifically, we train the model using the training set from one dataset but evaluate the performance on 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 ultrasound) but acquired from different centers and vendors. Therefore, the data distributions of the training and test sets are not aligned. Quantitative results are shown in Table 3, where GMS outperforms all other models in terms of DSC and HD95. In particular, nnUNet demonstrates powerful domain generalization abilities on both datasets, due to its network architecture and the use of data augmentation techniques. However, GMS surpasses nnUNet by around 2% for DSC. We also recruited two domain generalization methods (MixStyle and DSU) for comparison. MixStyle and DSU were implemented based on DeepLab-V3 (Chen et al. 2017) and employed the ResNet50 (He et al. 2016) as the encoder. GMS is better than two domain generalization methods, which demonstrates the powerful domain generalization ability of our model. The improvements achieved by our model are likely due to the latent representations derived from the pre-trained large vision model, which are domain-agnostic as it was trained on a large, general-purpose dataset. Additionally, GMS has fewer trainable parameters compared to the other generative models, which further reduces the likelihood of overfitting the model to the training set.

Qualitative Segmentation Results

Qualitative segmentation results for different models are shown in Figure 2. The yellow and green lines denote the contours of predictions and ground truth, respectively. In images acquired from the BUS and BUSI datasets (top two

$\mathcal{L}_{lm}$	$\mathcal{L}_{seg}$	BUSI			HAM10000			Kvasir		
		DSC↑	IoU↑	HD95↓	DSC↑	IoU↑	HD95↓	DSC↑	IoU↑	HD95↓
✓		80.25	71.26	21.57	93.92	89.41	9.95	92.93	88.28	10.21
	✓	78.75	69.87	24.78	93.64	88.99	10.27	93.00	88.47	10.68
✓	✓	**81.43**	**72.58**	**19.50**	**94.11**	**89.68**	**9.32**	**94.24**	**90.02**	**7.03**

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

Image Tokenizer	BUSI			HAM10000			Kvasir		
	DSC↑	IoU↑	HD95↓	DSC↑	IoU↑	HD95↓	DSC↑	IoU↑	HD95↓
VQ-VAE	79.23	70.34	24.57	92.77	87.61	13.34	92.47	88.31	9.73
SD-VAE	**81.43**	**72.58**	**19.50**	**94.11**	**89.68**	**9.32**	**94.24**	**90.02**	**7.03**

Table 5: Quantitative segmentation performance on three datasets using different image tokenizers.

Model	GPU Memory	GFLOPS	Infer Speed (samples / s)
UNet	2.02	25.13	36.65
UCTransNet	6.03	38.61	26.67
MedSegDiff-V2	26.32	972.66	0.31
SDSeg	29.80	655.34	8.36
GSS	18.11	476.28	9.02
GMS (Ours)	11.86	341.30	15.38

Table 6: Computational complexity analysis of GMS and other models.

rows), breast lesions show regular shapes, but the segmentation results of other models are often irregular causing over- or under-segmentation. GMS shows the most consistent results compared with the ground truth, which also proves the superiority of GMS. For the histology and dermatology images (the third and fourth rows), there are some regions with highly similar appearances to the target area, which leads to false positive segmentation results in the CNN and transformer-based models. However, GMS is still able to accurately segment those images with complex or misleading patterns. For the endoscopic image (last row), generative methods (GSS and our approach) give the most accurate predictions, which demonstrates the advantages of employing large pre-trained vision models for the segmentation task.

Sensitivity Analysis

Ablation Studies on Loss Function. We performed an ablation study on different loss function combinations for BUSI, HAM10000, and Kvasir-Instrument datasets. As shown in Table 4, the compound loss ($\mathcal{L}_{lm}$ + $\mathcal{L}_{seg}$) always has the best segmentation performance regardless of dataset size or modality. Interestingly, different datasets have different supervision preferences. GMS using only $\mathcal{L}_{lm}$ 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 supervision in the image and latent space are both important for achieving the best performance.

Image Tokenizer Effectiveness. We evaluated two pre-trained image tokenizers to assess their performances across three datasets. VQ-VAE (Van Den Oord, Vinyals et al. 2017) is a variant of VAE, incorporating a vector quantization step to generate discrete latent representations. Table 5 displays the results using VQ-VAE and SD-VAE (default tokenizer used in GMS) as the image tokenizers. SD-VAE improves DSC by up to 2.2% and reduces the HD95 by up to 5.07, indicating that SD-VAE is more suitable for image tokenization compared to VQ-VAE. The performance improvements also affirm the appropriateness of SD-VAE for handling diverse image segmentation tasks.

Computational Complexity We evaluated the computational complexity of each model on three metrics: GPU memory usage, GFLOPS, and inference speed. Results are shown in Table 6. Compared to CNN and Transformer-based models, GMS has larger GPU memory usage, higher GFLOPS, and slower inference speed since it contains many self-attention modules. However, GMS has less computational complexity than other generative segmentation models, underscoring its efficiency.

Conclusion

We present Generative Medical Segmentation (GMS), a generative approach to performing medical image segmentation. GMS leverages a pre-trained vision foundation model to obtain latent representations of both images and masks, and we design a lightweight latent mapping model that learns a mapping function from image latent representations to mask latent representations. Experiments on five datasets show that GMS outperforms the state-of-the-art discriminative and generative segmentation models. Moreover, we prove that the domain generalization ability of GMS is stronger than other well-designed domain generalization models. In the future, we will explore extending GMS to 3D medical images by selecting an appropriate pre-trained model for 3D images.

Acknowledgments

This work was supported by Centre for Doctoral Training in Surgical and Interventional Engineering at King's College London, the funding from the Wellcome Trust Award (218380/Z/19/Z), and by core funding from the Wellcome/EPSRC Centre for Medical Engineering [WT203148/Z/16/Z]. For the purpose of Open Access, the Author has applied a CC BY public copyright license to any Author Accepted Manuscript version arising from this submission.

References

Al-Dhabyani, W.; Gomaa, M.; Khaled, H.; and Fahmy, A. 2020. Dataset of breast ultrasound images. *Data in brief*, 28: 104863.

Cao, H.; Wang, Y.; Chen, J.; Jiang, D.; Zhang, X.; Tian, Q.; and Wang, M. 2022. Swin-unet: Unet-like pure transformer for medical image segmentation. In *European conference on computer vision*, 205–218. Springer.

Chaitanya, K.; Karani, N.; Baumgartner, C. F.; Erdil, E.; Becker, A.; Donati, O.; and Konukoglu, E. 2021. Semi-supervised task-driven data augmentation for medical image segmentation. *Medical Image Analysis*, 68: 101934.

Chen, J.; Lu, J.; Zhu, X.; and Zhang, L. 2023. Generative semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 7111–7120.

Chen, L.-C.; Papandreou, G.; Schroff, F.; and Adam, H. 2017. Rethinking atrous convolution for semantic image segmentation. arXiv:1706.05587.

Dosovitskiy, A.; Beyer, L.; Kolesnikov, A.; Weissenborn, D.; Zhai, X.; Unterthiner, T.; Dehghani, M.; Minderer, M.; Heigold, G.; Gelly, S.; Uszkoreit, J.; and Houlsby, N. 2021. An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. In *International Conference on Learning Representations*.

Goodfellow, I.; Pouget-Abadie, J.; Mirza, M.; Xu, B.; Warde-Farley, D.; Ozair, S.; Courville, A.; and Bengio, Y. 2014. Generative adversarial nets. *Advances in neural information processing systems*, 27.

He, K.; Zhang, X.; Ren, S.; and Sun, J. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 770–778.

Huo, J.; Vakharia, V.; Wu, C.; Sharan, A.; Ko, A.; Ourselin, S.; and Sparks, R. 2022. Brain Lesion Synthesis via Progressive Adversarial Variational Auto-Encoder. In *International Workshop on Simulation and Synthesis in Medical Imaging*, 101–111. Springer.

Ibtehaz, N.; and Kihara, D. 2023. Acc-unet: A completely convolutional unet model for the 2020s. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, 692–702. Springer.

Ibtehaz, N.; and Rahman, M. S. 2020. MultiResUNet: Rethinking the U-Net architecture for multimodal biomedical image segmentation. *Neural networks*, 121: 74–87.

Isensee, F.; Jaeger, P. F.; Kohl, S. A.; Petersen, J.; and Maier-Hein, K. H. 2021. nnU-Net: a self-configuring method for deep learning-based biomedical image segmentation. *Nature methods*, 18(2): 203–211.

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. 2021. 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*, 218–229. Springer.

Karras, T.; Aittala, M.; Hellsten, J.; Laine, S.; Lehtinen, J.; and Aila, T. 2020. Training generative adversarial networks with limited data. *Advances in neural information processing systems*, 33: 12104–12114.

Kingma, D. P.; and Welling, M. 2013. Auto-encoding variational bayes. arXiv:1312.6114.

Kirillov, A.; Mintun, E.; Ravi, N.; Mao, H.; Rolland, C.; Gustafson, L.; Xiao, T.; Whitehead, S.; Berg, A. C.; Lo, W.-Y.; et al. 2023. Segment anything. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 4015–4026.

Kong, L.; Lian, C.; Huang, D.; Hu, Y.; Zhou, Q.; et al. 2021. Breaking the dilemma of medical image-to-image translation. *Advances in Neural Information Processing Systems*, 34: 1964–1978.

Li, D.; Yang, J.; Kreis, K.; Torralba, A.; and Fidler, S. 2021. 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*, 8300–8311.

Li, X.; Dai, Y.; Ge, Y.; Liu, J.; Shan, Y.; and DUAN, L. 2022. Uncertainty Modeling for Out-of-Distribution Generalization. In *International Conference on Learning Representations*.

Lin, T.; Chen, Z.; Yan, Z.; Zheng, F.; and Yu, W. 2024. Stable Diffusion Segmentation for Biomedical Images with Single-step Reverse Process. arXiv:2406.18361.

Liu, J.; Pasumarthi, S.; Duffy, B.; Gong, E.; Datta, K.; and Zaharchuk, G. 2023. One model to synthesize them all: Multi-contrast multi-scale transformer for missing data imputation. *IEEE Transactions on Medical Imaging*, 42(9): 2577–2591.

Loshchilov, I.; and Hutter, F. 2019. Decoupled Weight Decay Regularization. In *International Conference on Learning Representations*.

Rombach, R.; Blattmann, A.; Lorenz, D.; Esser, P.; and Ommer, B. 2022. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 10684–10695.

Ronneberger, O.; Fischer, P.; and Brox, T. 2015. 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, 234–241. Springer.

Ruan, J.; Xie, M.; Gao, J.; Liu, T.; and Fu, Y. 2023. Ege-unet: an efficient group enhanced unet for skin lesion segmentation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, 481–490. Springer.

Schuhmann, C.; Beaumont, R.; Vencu, R.; Gordon, C.; Wightman, R.; Cherti, M.; Coombes, T.; Katta, A.; Mullis, C.; Wortsman, M.; et al. 2022. Laion-5b: An open large-scale dataset for training next generation image-text models. *Advances in Neural Information Processing Systems*, 35: 25278–25294.

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. 2017. Gland segmentation in colon histology images: The glas challenge contest. *Medical image analysis*, 35: 489–502.

Tschandl, P.; Rosendahl, C.; and Kittler, H. 2018. The HAM10000 dataset, a large collection of multi-source dermatoscopic images of common pigmented skin lesions. *Scientific data*, 5(1): 1–9.

Van Den Oord, A.; Vinyals, O.; et al. 2017. Neural discrete representation learning. *Advances in neural information processing systems*, 30.

Wang, H.; Cao, P.; Wang, J.; and Zaiane, O. R. 2022a. Uctransnet: rethinking the skip connections in u-net from a channel-wise perspective with transformer. In *Proceedings of the AAAI conference on artificial intelligence*, volume 36, 2441–2449.

Wang, Z.; Min, X.; Shi, F.; Jin, R.; Nawrin, S. S.; Yu, I.; and Nagatomi, R. 2022b. SMESwin Unet: Merging CNN and transformer for medical image segmentation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, 517–526. Springer.

Wu, J.; Ji, W.; Fu, H.; Xu, M.; Jin, Y.; and Xu, Y. 2024. Medsegdiff-v2: Diffusion-based medical image segmentation with transformer. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, 6030–6038.

Yap, M. H.; Pons, G.; Marti, J.; Ganau, S.; Sentis, M.; Zwiggelaar, R.; Davison, A. K.; and Marti, R. 2017. Automated breast ultrasound lesions detection using convolutional neural networks. *IEEE journal of biomedical and health informatics*, 22(4): 1218–1226.

Ye, J.; Ni, H.; Jin, P.; Huang, S. X.; and Xue, Y. 2023. Synthetic augmentation with large-scale unconditional pre-training. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, 754–764. Springer.

Zhou, K.; Yang, Y.; Qiao, Y.; and Xiang, T. 2023. Mixstyle neural networks for domain generalization and adaptation. *International Journal of Computer Vision*, 1–15.