Guided Latent Diffusion for Universal Medical Image Segmentation

Mattia Salsi[a], Yunying Wang[⋆a], Chen Hu[⋆b], Yueyue Hu[a], Hanchi Ren[*b], Jingjing Deng[a], and Xianghua Xie[b]

[a]Durham University, Durham, UK
[b]Swansea University, Swansea, UK

ABSTRACT

Deep learning based medical segmentation still presents a great challenge due to the lack of large-scale datasets in the medical domain. The existing publicly available datasets vary significantly in terms of imaging modalities and target anatomies. This paper presents a novel guided latent diffusion model for universal medical segmentation, capable of segmenting diverse anatomical structures using a single and unified architecture. Given a Contrastive Language-Image Pretraining (CLIP) embedding specifying the target anatomical structure, the proposed model leverages a collection of datasets covering the diverse structures which can segment any anatomical targets that are presented in the aggregated data. By performing diffusion fully in latent space, we achieve comparable results to pixel-space diffusion with significantly lower computational cost. The proposed methods demonstrates competitive performance against existing deep learning-based discriminative approaches on several benchmarks. Furthermore, it shows strong generalization capabilities on unseen datasets.

Keywords: Medical Image Segmentation, Denoising Diffusion Probabilistic Models, Contrastive Language-Image Pretraining

1. INTRODUCTION

Medical image segmentation is a technique used to automatically partition biomedical images into meaningful sub-structures such as organs, lesions, or pathologies. This process aids clinicians in identifying and delineating abnormalities, playing a crucial role in various medical applications, including radiotherapy. Current state-of-the-art models can be categorised into two variations: fully-convolutional networks (FCNs), and hybrid transformer-convolutional approaches. The most notable FCN architecture is the U-Net[1] which utilise a encoder-decoder structure to extract and transform feature maps from the input into a segmentation mask. The most successful of these approaches is nnU-Net,[2] a self-configuring U-Net framework that automatically adapts its architecture based on the training data-set fingerprint. Transformer based models introduce a Vision Transformer (ViT) backbone[3] into the architecture, employing self-attention to capture long-term dependencies that are neglected by convolutional layers. The most notable of these techniques is Swin-UNETR,[4] which replaces the encoder path of the traditional U-Net with Swin Transformers.

Challenges persist in training these models for the medical domain, primarily due to limited labeled datasets and high data variance. As a result, the research landscape has largely focused on specialist few-class segmentation models, with generalist multi-class models occupying a smaller representation. This practice conflicts with the established principle in deep learning that increasing the quantity and diversity of training data is key to improving model performance and generalization capabilities. Recent works investigating universal medical segmentation models[5,6] have found that generalist models consistently outperform their specialist counterparts, aligning with these expectations. Recently, there has been an increased focus on modifying these approaches towards universal architectures, capable of segmenting a comprehensive number of anatomical structures. Liu *et al.* presented CLIP-Driven Universal Model,[5] combining Swin-UNETR for volumetric feature extraction together with CLIP,[7] a joint vision-language model that combines an image encoder and text encoder to produce

⋆ Chen Hu and Yunying Wang contributed equally as the second authors.

a combined embedded representation of an image. The universal model generates class-specific decoding parameters using the CLIP text embeddings which specify the target anatomical structure, which are then used by the convolutional decoder to produce binary segmentation masks. Their approach currently ranks highest in both the MSD[8] and BTCV[9] benchmarks, highlighting the potential of universal models for medical image segmentation.

Parallel yet diverging evolution has taken place in the broader computer vision field, where diffusion models have established themselves as state-of-the-art solutions for various applications.[10] Diffusion Models[11] are a family of generative models that have achieved state-of-the-art performance in various computer vision tasks such as image and video synthesis.[10] However, their application to discriminative tasks, such as medical image segmentation, is still at an early stage, where these works largely focus on binary or few-class segmentation. Wolleb *et al.*[12] introduced the first model for brain tumor segmentation, highlighting their ability for ensemble generation and uncertainty visualisation as desirable properties for increasing model interpretability and encouraging clinical use. Subsequent works by Wu *et al.*[13] refined this approach, introducing architectural improvements such as separate image encoders and transformer-based feature aggregation. Further works have explored 3D architectures[14] and operating in latent space.[15] While these methods have shown promising results, they focus on training on individual datasets with single/few-class segmentation targets, and no ability for conditional sampling. Recent advancements in semi-supervised segmentation, such as IPixMatch,[16] have highlighted the importance of capturing inter-pixel dependencies to improve model performance, particularly in situations with limited labeled data. Inspired by these methods, our work aims to leverage latent space diffusion to achieve generalization across diverse anatomical structures while ensuring computational efficiency.

Motivated by the trend towards universal models and their state-of-the-art performance, this paper presents a guided latent diffusion model for universal medical segmentation via integrating a class-aware image encoder into the U-Net architecture. By performing diffusion entirely in latent space and leveraging autoencoders for segmentation masks and conditional images, we effectively reduce computational costs while maintaining competitive accuracy. We have collated and standardized multiple datasets, training a model capable of segmenting any anatomical structure covered in the aggregated data given user prompting. Finally, we provide a comparative evaluation against existing deep learning-based discriminative segmentation approaches across several benchmarks, underscoring the advantages of our proposed method.

2. THE PROPOSED METHOD

Figure 1 shows the overview of the proposed method. Given a collection of datasets $D = \{D_0, \ldots, D_M\}$, each dataset D_i consists of image-label pairs $\{(\mathbf{y}_1, \mathbf{x}_1), \ldots, (\mathbf{y}_N, \mathbf{x}_N)\}$, where N_i is the number of cases and $K_i = \{k_0, \ldots, k_n\}$ denotes the segmentation targets covered by dataset D_i. The goal is to learn the function $F_\theta(\mathbf{y}, k) = \mathbf{x}$. The model effectively treats multi-class segmentation as separate instances of binary segmentation, where $\mathbf{x}_n^{ijk} = 1$ only if $\mathbf{y}_n^{ijk}$ belongs to class k, and 0 otherwise. Two separate auto-encoders, ($\mathcal{E}_y$, $\mathcal{D}_y$) and ($\mathcal{E}_x$, $\mathcal{D}_x$) are used to map images $\mathbf{y}$ and segmentation masks $\mathbf{x}$ to a lower-dimensional latent space using the encoder $\mathcal{E} : \mathbb{R}^{H \times W} \rightarrow \mathbb{R}^{C_z \times H_z \times W_z}$, where C_z is the channel dimensionality of the latent space and $H_z = \frac{H}{f_z}$, $W_z = \frac{W}{f_z}$ are the spatial dimensions of the latent space after a down-scaling factor f_z. The decoder is then used to reconstruct the original input $\hat{\mathbf{y}} = \mathcal{D}_y(\mathbf{z}^{(y)})$ and $\hat{\mathbf{x}} = \mathcal{D}_x(\mathbf{z}^{(x)})$.

2.1 Mask and Image Encoder

The mask auto-encoder ($\mathcal{E}_x$, $\mathcal{D}_x$) is a vanilla auto-encoder,[18] trained by minimising a reconstruction loss that quantifies the fidelity of the reconstructed segmentation mask $\hat{\mathbf{x}}$ with respect to the original $\mathbf{x}$. The reconstruction loss is composed of two terms: (1) a pixel-wise binary cross entropy loss $\mathcal{L}_{BCE}$, and (2) a global spatial loss $\mathcal{L}_{DICE}$, defined respectively as:

$$\mathcal{L}_{BCE}(\mathbf{x}, \hat{\mathbf{x}}) := -\frac{1}{N} \sum_{i=1}^{N} \left[\mathbf{x}_i \log(\hat{\mathbf{x}}_i) + (1 - \mathbf{x}_i) \log(1 - \hat{\mathbf{x}}_i) \right] \tag{1}$$

$$\mathcal{L}_{DICE}(\mathbf{x}, \hat{\mathbf{x}}) := 1 - \frac{2 \sum_{i=1}^{N} \mathbf{x}_i \hat{\mathbf{x}}_i + \epsilon}{\sum_{i=1}^{N} \mathbf{x}_i + \sum_{i=1}^{N} \hat{\mathbf{x}}_i + \epsilon} \tag{2}$$

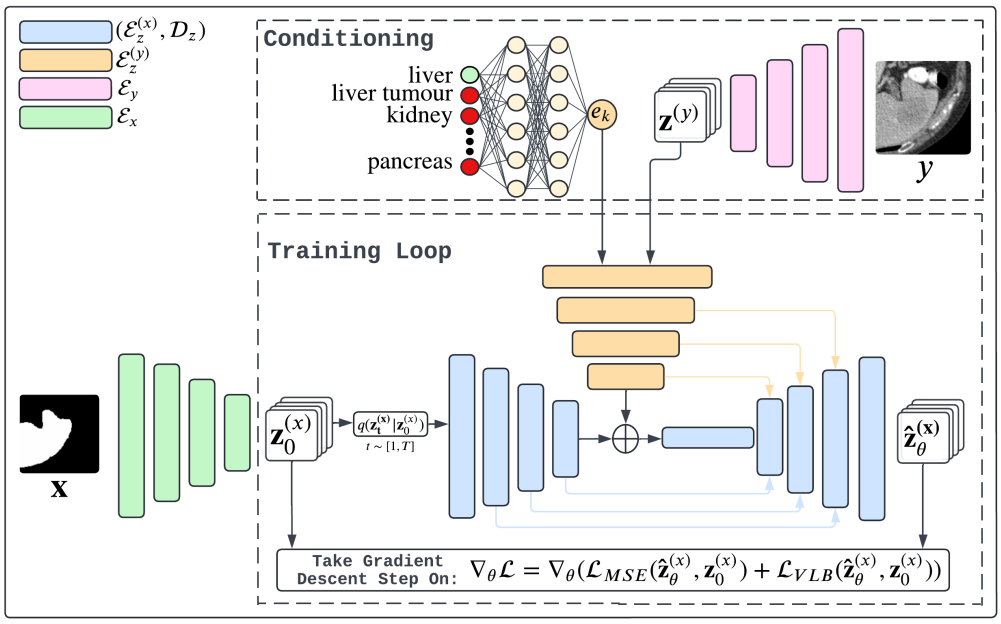

Figure 1: An overview of the architecture and training procedure of the proposed diffusion segmentation model. e_k is the class embedding vector for segmentation targets. The training loop includes adding noise, neural network-based denoising, and optimizing the model through backpropagation using $\mathcal{L}_{MSE}$ and $\mathcal{L}_{VLB}$.[17]

where ϵ is a term introduced for numerical stability.

We note that previous work by Rombach *et al.*[19] found it beneficial to diffuse across a regularised latent space, encoded by either a VAE or VQ-VAE. However, we observe that regularised latent spaces are redundant for encoding a binary segmentation mask, and increase the convergence time of the segmentation model. As such, we opt for a regular auto-encoder and utilise weight decay[20] which by extension enforces a lower variance latent space. The final loss function is then given by:

$$\mathcal{L}_{AE} := \mathcal{L}_{BCE} + \mathcal{L}_{DICE} + \lambda||w||^2 \tag{3}$$

where w are the auto-encoder model weights, and λ is a tune-able hyper-parameter controlling the strength of the regularisation term, that we set to $1e^{-5}$. We opt to utilise a variational-autoencoder (VAE)[21] as our image auto-encoder ($\mathcal{E}_y$, $\mathcal{D}_y$). The VAE training objective function is based on maximizing the Evidence Lower Bound (ELBO), which consists of two main components: (1) the expected log-likelihood of the reconstruction, and (2) the negative Kullback-Leibler (KL) divergence between the approximate posterior $q(\mathbf{z}^{(y)}|\mathbf{y})$ and the prior $p(\mathbf{z}^{(y)})$. This ensures that the approximate posterior is roughly modeled by the prior, which in our case we define as a multivariate Gaussian distribution $\mathcal{N}(\mathbf{0}, \mathbf{I})$. We additionally incorporate a learned perceptual similarity[22] loss defined as:

$$\mathcal{L}_{LPIPS}(\mathbf{y}, \hat{\mathbf{y}}) := \sum_l w_l \odot ||(\mathcal{F}_{\mathbf{y}}^l - \mathcal{F}_{\hat{\mathbf{y}}}^l)||_2^2 \tag{4}$$

where $\mathcal{F}_{\mathbf{y}} = \varphi(\mathbf{y})$, $\mathcal{F}_{\hat{\mathbf{y}}} = \varphi(\hat{\mathbf{y}})$ are multi-layer feature maps, and φ is a pre-trained VGG network. The final training objective for our VAE is then given by:

$$\mathcal{L}_{VAE} := \mathbb{E}\left[-\log(\mathbf{y}|\mathbf{z}^{(y)}) + \alpha\mathcal{L}_{LPIPS}\right] + \beta D_{KL}(q(\mathbf{z}^{(y)}|\mathbf{y}) \,||\, p(\mathbf{z}^{(y)})) \tag{5}$$

2.2 Diffusion Processes

The forward diffusion process gradually corrupts a latent segmentation mask $\mathbf{z}_0^{(x)}$, adding Gaussian noise according to a number of steps T and corresponding monotonically-increasing variance schedule $\beta_t \in [0, 1]$ as:

$$q(\mathbf{z}_t^{(x)}|\mathbf{z}_0^{(x)}) := \mathcal{N}(\mathbf{z}_t^{(x)}; \sqrt{\bar{\alpha}_t}\,\mathbf{z}_0^{(x)}, (1 - \bar{\alpha}_t)\mathbf{I}) \tag{6}$$

where $\alpha_t := 1 - \beta_t$ and $\bar{\alpha}_t = \prod_i^t \alpha_i$ are defined in order to allow us to compute $\mathbf{z}_t^{(x)}$ without the prior trajectory $\mathbf{z}_0^{(x)}, \dots \mathbf{z}_{t-1}^{(x)}$. At the end of the trajectory, for a well defined noise schedule β_t, $\mathbf{z}_T^{(x)}$ should be approximately Gaussian distributed.

The denoising process is estimated by our neural network, parameterized by learned parameters θ, and learns to reverse the forward process according to:

$$p_\theta(\mathbf{z}_{t-1}^{(x)}|\mathbf{z}_t^{(x)}) := \mathcal{N}(\mathbf{z}_{t-1}^{(x)}; \mu_\theta, \Sigma_\theta) \tag{7}$$

where $\mu_\theta = \mu_\theta(\mathbf{z}_t^{(x)}, \mathbf{z}^{(y)}, t, k)$ and $\Sigma_\theta = \Sigma_\theta(\mathbf{z}_t^{(x)}, \mathbf{z}^{(y)}, t, k))$ are approximated by our model, conditional on the latent image $\mathbf{z}^{(y)}$, denoising time-step t, and segmentation class k. To sample from our model, we can sample $\mathbf{z}_T \sim \mathcal{N}(0, \mathbf{I})$ and apply the denoising trajectory:

$$p_\theta(\mathbf{z}_0^{(x)}|\mathbf{z}_T^{(x)}) = p_\theta(\mathbf{z}_T^{(x)}) \prod_{t=1}^{T} p_\theta(\mathbf{z}_{t-1}^{(x)}|\mathbf{z}_t^{(x)}) \tag{8}$$

There are multiple ways that μ_θ can be parameterized. In practice, Ho *et al.*[23] found that training the model to predict the noise ϵ added by $q(\mathbf{z}_t^{(x)}|\mathbf{z}_0^{(x)})$ and parameterizing μ_θ as a function of ϵ_θ achieves the best results for image synthesis. However, given that we always condition on the conditional image $\mathbf{z}^{(y)}$, this provides a strong enough signal to be able to estimate $\mathbf{z}_0^{(x)}$ at any point of the diffusion process. We find that parameterizing μ_θ by predicting $\mathbf{z}_0^{(x)}$ leads to faster convergence, such as:

$$\mu_\theta = \frac{\sqrt{\alpha_t}(1 - \bar{\alpha}_{t-1})}{1 - \bar{\alpha}_t}\mathbf{z}_t^{(x)} + \frac{\beta_t\sqrt{\bar{\alpha}_{t-1}}}{1 - \bar{\alpha}_t}\mathbf{z}_\theta^{(x)}(\mathbf{z}_t^{(x)}, \mathbf{z}^{(y)}, t, k) \tag{9}$$

where $\hat{\mathbf{z}}_\theta^{(x)}(\mathbf{z}_t^{(x)}, \mathbf{z}^{(y)}, t, k)$ is parameterized by the model parameters θ to predict $\hat{\mathbf{z}}_0^{(x)}$, and learnt by optimising the following loss function:

$$\mathcal{L}_{MSE} := \mathbb{E}\left[||\mathbf{z}_0^{(x)} - \hat{\mathbf{z}}_\theta^{(x)}(\mathbf{z}_t^{(x)}, \mathbf{z}^{(y)}, t, k)||^2\right] \tag{10}$$

In summary, the proposed method achieves superior segmentation performance by leveraging a dual auto-encoder framework. Our approach not only improves reconstruction quality, as evidenced by a lower MSE, but also achieves enhanced consistency and robustness across multiple classes compared to baseline methods.

3. EXPERIMENT AND DISCUSSION

3.1 Dataset and Implementation

We collected a set of datasets (Table 1) focusing on CT imaging of the abdominal structure, covering a total of 29 segmentation targets (23 organs and 6 tumours). The aggregated dataset D was divided into a training and validation set using a 9:1 split. To ensure conformity between samples, we utilise the MONAI library[29] to re-space all volumes to $(1.5\text{mm}, 1.5\text{mm}, 2\text{mm})$ spacing, scale and normalise intensity, and apply foreground cropping and spatial padding. Since different datasets may have different label formats, we standardized all labels to make them consistent. We created a mapping table to align different labels across datasets, ensuring that the structures were represented the same way. Finally, labels were converted to a one-hot encoded format.

To address imbalanced datasets, we uniformly sample datasets $D_i \sim \mathcal{U}(D)$ and randomly select cases $c \in D_i$. For each case c, we extract B patches $\{(\mathbf{y}_1, \mathbf{x}_1), \dots, (\mathbf{y}_B, \mathbf{x}_B)\} \subset c$, where $\mathbf{y}, \mathbf{x} \in \mathbb{R}^{256\times256}$, using an oversampling technique where patches with active segmentation labels are sampled with higher probability. Each sample is

Table 1: Statistics of the aggregated dataset

Dataset	# of Cases	# of Structures
MSD (CT)[8]	945	9
BTCV[9]	30	12
KiTS 2023[24]	489	4
TCIA Pancreas[25]	76	1
AbdomenCT-1k[26]	722	5
AMOS (CT)[27]	500	15
WORD[28]	120	16

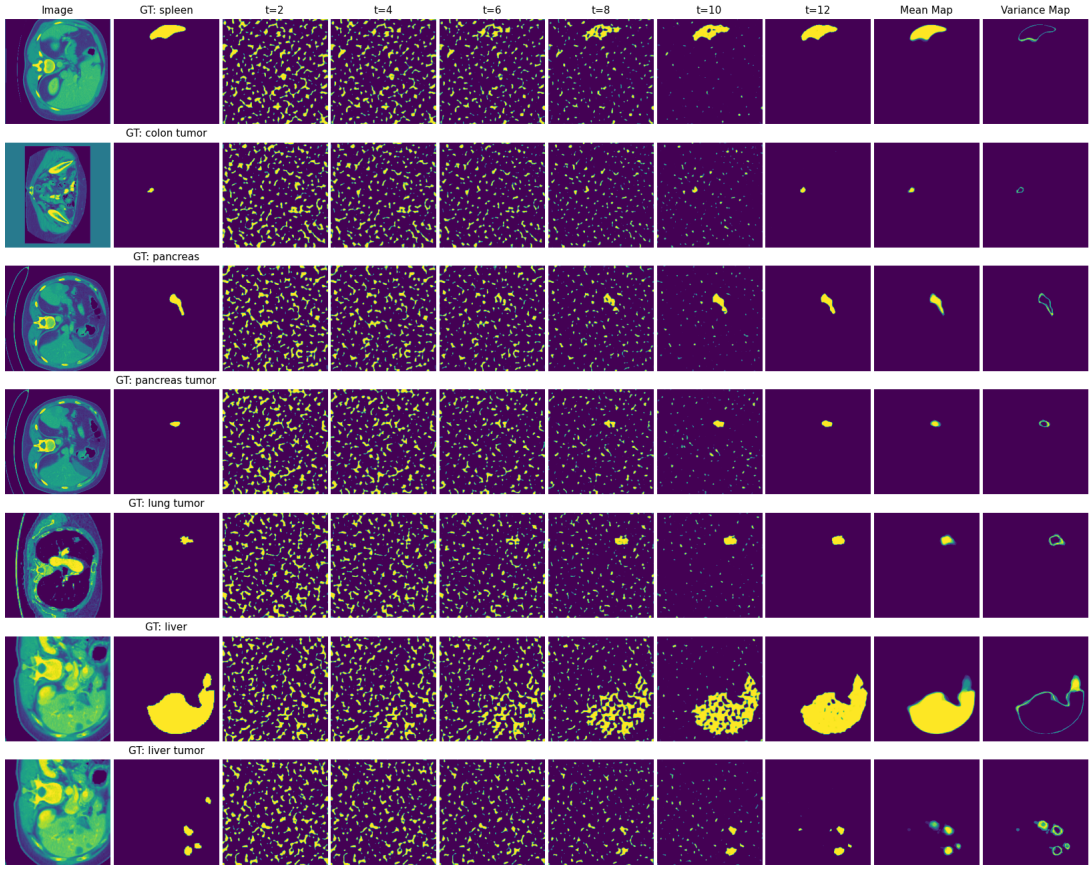

Figure 2: A set of predictions generated from our diffusion segmentation model.

noised using time-steps sampled using importance sampling, where time-steps are weighted according to the average term they contribute to the denoising loss. To perform inference on a 3D medical image $\mathbf{y} \in \mathbb{R}^{H \times W \times D}$ with target segmentation classes $K = \{k_1, \dots k_N\}$, we utilise the MONAI sliding window inference algorithm to process $\mathbf{y}$ as a series of 2D patches $\{\tilde{\mathbf{y}}_0, \dots, \tilde{\mathbf{y}}_M\}, \tilde{\mathbf{y}}_i \in \mathbb{R}^{256 \times 256}$. For each patch $\tilde{\mathbf{y}}_i$ and target $k_n \in K$, we forward to the model a tuple $(\mathbf{z}_T^{(x)}, \mathbf{z}_i^{(\tilde{y})}, k_n)$ where $\mathbf{z}_T^{(x)} \sim \mathcal{N}(0, \mathbf{I})$ and $\mathbf{z}_i^{(\tilde{y})} = \mathcal{E}_y(\tilde{\mathbf{y}}_i)$ is the latent conditional image. We then apply the denoising trajectory $p_\theta(\mathbf{z}_0^{(x)} | \mathbf{z}_T^{(x)})$ to sample $\mathbf{z}_{i,n}^{(x)}$, which corresponds to a latent segmentation mask for target k_n of patch $\tilde{\mathbf{y}}_\mathbf{i}$. Finally we utilise the decoder $\mathcal{D}_x$ to produce the full-size segmentation mask $\mathbf{x}_{i,n} = \mathcal{D}_x(\mathbf{z}_{i,n}^{(x)}) \in \mathbb{R}^{256 \times 256}$.

3.2 Experimental Result

Figure 2 shows a set of predictions sampled from our model. The first two columns show the conditional image and ground-truth segmentation mask, labeled by the target anatomical structure. The following columns show

Table 2: Benchmark on MSD (DICE %) by Model Configuration

Target	Emb.	Spl.	Liv.		HepVes.		Panc.		Lung Tmr.	Colon Tmr.
			Org.	Tmr.	Org.	Tmr.	Org.	Tmr.		
Noise $\boldsymbol{\epsilon_\theta}$	Learnt	92.2	92.3	42.9	45.0	20.6	69.6	23.3	13.2	16.8
	CLIP	92.7	94.8	43.4	32.8	20.4	62.3	17.6	15.1	20.7
Mask $\hat{\mathbf{z}}_\theta^{(x)}$	Learnt	94.2	**95.4**	55.2	45.5	**44.9**	67.3	**30.3**	35.4	23.8
	CLIP	**94.3**	94.2	**56.9**	**47.1**	41.5	**68.6**	**30.3**	**39.1**	**41.3**

the denoising trajectory of a single sample, and the final two columns show the mean and variance map obtained by ensembling various samples.

The training objective of this model centers on optimizing convergence speed and segmentation accuracy, particularly for small anatomical structures such as liver tumors, pancreas, intestines, and blood vessels. To achieve this, the model is designed to estimate the parameter μ_θ by directly predicting the segmentation mask $\hat{\mathbf{z}}_\theta^{(x)}$ instead of the conventional noise prediction $\boldsymbol{\epsilon_\theta}$. This adjustment leverages the conditionally strong signal from the image itself, allowing the model to focus on capturing finer details in segmentation. Given the limitations of regularized latent spaces, our model employs a standard autoencoder with weight decay. This setup reduces latent space variance and enhances computational efficiency without sacrificing segmentation precision. As shown in Table 2, this approach has proven advantageous, particularly for smaller structures, such as liver tumors(+13.5%) and pancreatic tumors(+7%), as evidenced by the significant improvements in Dice score.

3.2.1 Generalisability Performance

In order for medical models to be deployed at large-scale clinical use, they must be able to readily and accurately process images taken by varying machinery across several different hospitals.[30] This is a typical challenge faced by specialist models, whose limited training dataset renders them more susceptible to the intrinsic imaging noise produced by different machinery. As such, we evaluate the generalisability of our model on 3D-IRCADb,[31] an external dataset covering CT abdominal imaging that was not used in our training dataset (Table 3, and the metrics for the competing models are sourced from the work of Liu *et al.*[5]).

Our Universal Diffusion model achieves an average Dice score of 86.98% on the 3D-IRCADb dataset, ranking as the second-best model overall. While our model does not achieve the highest Dice scores, it consistently outperforms several specialized models, such as nnFormer and Swin UNETR. For instance, our model scores 93.97% on the left kidney, compared to nnFormer's 88.20% and Swin UNETR's 66.34%, illustrating its capacity to generalize well across various anatomical structures without task-specific training (see Table 3). However, our model struggles with certain structures, such as the pancreas (Dice score of 81.63%). This could be due to the limited representation of the pancreas in the dataset or its small size and complex anatomy, making it challenging for accurate segmentation. CLIP Universal model achieves the highest average Dice score of 91.62% across the structures, demonstrating exceptional generalisability. This superior performance is likely due to the model's use of CLIP embeddings, which integrate both anatomical and semantic context, enabling the model to better capture the intrinsic relationships among different anatomical structures.[7] The vision-language framework of CLIP provides robust contextual cues, allowing CLIP Universal to adapt effectively to new datasets with diverse imaging characteristics and handle variations introduced by different imaging equipment.[5]

The proposed model demonstrates significant potential in advancing medical image segmentation, offering both unique capabilities and practical benefits that set it apart from models like CLIP Universal. Unlike CLIP Universal, which relies heavily on computationally intensive, pre-trained CLIP embeddings, our model achieves high segmentation accuracy with a streamlined, diffusion-based architecture that does not require such pre-training. This design reduces complexity and resource demands, making deployment and maintenance far more feasible, especially in clinical settings with limited computational resources. By adopting a unified framework, our model can segment diverse anatomical structures across multiple datasets without specialized training, highlighting its generalizability and adaptability. It paves the way for developing future models capable of handling a broader range of tasks, including multi-modality and multi-organ segmentation. While CLIP Universal may be preferable when maximum segmentation accuracy is critical, our approach strikes an optimal balance between accuracy, efficiency, and flexibility, making it an ideal choice for diverse clinical applications where ease

Table 3: Generalisability Benchmark on 3D-IRCADb. Blue values indicate instances where ours achieved the second-best performance among the compared methods.

Model	Spl.	RKidney.	LKidney	Gallbl.	Liv.	Sto.	Panc.	Avg
SegResNet[32]	94.08	80.01	91.60	69.59	95.62	**89.53**	79.19	85.66
nnFormer[33]	93.75	88.20	90.11	62.22	94.93	87.93	78.90	85.14
UNesT[34]	94.02	84.90	**94.95**	68.58	95.10	89.28	79.94	86.68
TransBTS[35]	91.33	76.22	88.87	62.50	94.42	85.87	63.90	80.44
TransUNet[36]	94.09	82.07	89.92	63.07	95.55	89.12	79.53	84.76
UNETR[37]	92.23	91.28	94.19	56.20	94.25	86.73	72.56	83.92
Swin UNETR[4]	93.51	66.34	90.63	61.05	94.73	87.37	73.77	81.05
CLIP Universal[5]	**95.76**	**94.99**	94.42	**88.79**	**97.03**	89.36	**90.99**	**91.62**
Universal Diffusion (Ours)	93.89	**93.97**	93.47	**75.77**	94.93	75.17	81.63	**86.98**

of deployment and practical adaptability are priorities. Future work could explore adapting the model to other imaging modalities, such as MRI, and incorporating domain adaptation techniques to enhance generalizability across different medical domains.

4. CONCLUSION

In this paper, we presented a guided latent-diffusion model for universal segmentation. Our framework enables us to train a unified model across a collection of data-sets covering a number of diverse anatomical structures, and allows model prompting to guide the sampling process towards segmenting any target class covered in the training data-set. We show that a diffusion back-boned model is capable of effectively modelling the joint distribution of several anatomical structures within a single, shared architecture. Our approach produces competitive results to existing models, laying a strong foundation for further research in diffusion-backboned medical imaging models.

REFERENCES

[1] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation, 2015.

[2] Fabian Isensee, Paul F. Jaeger, and et. al. nnu-net: a self-configuring method for deep learning-based biomedical image segmentation. *Nature Methods*, 18(2):203–211, Feb 2021.

[3] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, and et. al. An image is worth 16x16 words: Transformers for image recognition at scale, 2021.

[4] Ali Hatamizadeh, Vishwesh Nath, Yucheng Tang, Dong Yang, Holger Roth, and Daguang Xu. Swin unetr: Swin transformers for semantic segmentation of brain tumors in mri images, 2022.

[5] Jie Liu, Yixiao Zhang, Jie-Neng Chen, and et. al. Clip-driven universal model for organ segmentation and tumor detection, 2023.

[6] Jun Ma, Yuting He, Feifei Li, Lin Han, Chenyu You, and Bo Wang. Segment anything in medical images. *Nature Communications*, 15(1):654, Jan 2024.

[7] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision, 2021.

[8] Michela Antonelli and et. al. The medical segmentation decathlon. *Nature Communications*, 13(1), July 2022.

[9] Bennett Landman, Zhoubing Xu, J Igelsias, Martin Styner, Thomas Langerak, and Arno Klein. Miccai multi-atlas labeling beyond the cranial vault–workshop and challenge. In *Proc. MICCAI Multi-Atlas Labeling Beyond Cranial Vault—Workshop Challenge*, volume 5, page 12, 2015.

[10] Ling Yang, Zhilong Zhang, Yang Song, Shenda Hong, Runsheng Xu, Yue Zhao, Wentao Zhang, Bin Cui, and Ming-Hsuan Yang. Diffusion models: A comprehensive survey of methods and applications, 2024.

[11] Jascha Sohl-Dickstein, Eric A. Weiss, Niru Maheswaranathan, and Surya Ganguli. Deep unsupervised learning using nonequilibrium thermodynamics, 2015.

[12] Julia Wolleb, Robin Sandkühler, Florentin Bieder, Philippe Valmaggia, and Philippe C. Cattin. Diffusion models for implicit image segmentation ensembles, 2021.
[13] Junde Wu, Wei Ji, Huazhu Fu, Min Xu, Yueming Jin, and Yanwu Xu. Medsegdiff-v2: Diffusion based medical image segmentation with transformer, 2023.
[14] Zhaohu Xing, Liang Wan, Huazhu Fu, Guang Yang, and Lei Zhu. Diff-unet: A diffusion embedded network for volumetric segmentation, 2023.
[15] Hung Vu Quoc, Thao Tran Le Phuong, and et. al. Lsegdiff: A latent diffusion model for medical image segmentation. In *Proceedings of the 12th International Symposium on Information and Communication Technology*, SOICT '23, page 456–462, New York, NY, USA, 2023. Association for Computing Machinery.
[16] Kebin Wu, Wenbin Li, and Xiaofei Xiao. Ipixmatch: Boost semi-supervised semantic segmentation with inter-pixel relation. In *2024 International Joint Conference on Neural Networks (IJCNN)*, pages 1–7, 2024.
[17] Alex Nichol and Prafulla Dhariwal. Improved denoising diffusion probabilistic models, 2021.
[18] D. E. Rumelhart, G. E. Hinton, and R. J. Williams. *Learning internal representations by error propagation*, page 318–362. MIT Press, Cambridge, MA, USA, 1986.
[19] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models, 2022.
[20] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization, 2019.
[21] Diederik P Kingma and Max Welling. Auto-encoding variational bayes, 2022.
[22] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *CVPR*, 2018.
[23] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models, 2020.
[24] Nicholas Heller and et. al. The kits21 challenge: Automatic segmentation of kidneys, renal tumors, and renal cysts in corticomedullary-phase ct, 2023.
[25] Holger R. Roth, Le Lu, Amal Farag, Hoo-Chang Shin, Jiamin Liu, Evrim Turkbey, and Ronald M. Summers. Deeporgan: Multi-level deep convolutional networks for automated pancreas segmentation, 2015.
[26] Jun Ma, Yao Zhang, Song Gu, Cheng Zhu, Cheng Ge, Yichi Zhang, Xingle An, Congcong Wang, Qiyuan Wang, Xin Liu, Shucheng Cao, Qi Zhang, Shangqing Liu, Yunpeng Wang, Yuhui Li, Jian He, and Xiaoping Yang. AbdomenCT-1K: Is abdominal organ segmentation a solved problem? *IEEE Trans. Pattern Anal. Mach. Intell.*, 44(10):6695–6714, October 2022.
[27] Yuanfeng Ji, Haotian Bai, Jie Yang, Chongjian Ge, Ye Zhu, Ruimao Zhang, Zhen Li, Lingyan Zhang, Wanling Ma, Xiang Wan, and Ping Luo. Amos: A large-scale abdominal multi-organ benchmark for versatile medical image segmentation, 2022.
[28] Xiangde Luo, Wenjun Liao, Jianghong Xiao, Jieneng Chen, Tao Song, Xiaofan Zhang, Kang Li, Dimitris N. Metaxas, Guotai Wang, and Shaoting Zhang. Word: A large scale dataset, benchmark and clinical applicable study for abdominal organ segmentation from ct image. *Medical Image Analysis*, 82:102642, 2022.
[29] M. Jorge Cardoso and et. al. Monai: An open-source framework for deep learning in healthcare, 2022.
[30] John Mongan, Linda Moy, and Charles E Kahn, Jr. Checklist for artificial intelligence in medical imaging (CLAIM): A guide for authors and reviewers. *Radiol. Artif. Intell.*, 2(2):e200029, March 2020.
[31] Luc Soler, Alexandre Hostettler, and et. al. 3d image reconstruction for comparison of algorithm database: A patient specific anatomical and medical image database. *IRCAD, Strasbourg, France, Tech. Rep*, 1, 2010.
[32] Andriy Myronenko. 3d mri brain tumor segmentation using autoencoder regularization, 2018.
[33] Hong-Yu Zhou, Jiansen Guo, Yinghao Zhang, Lequan Yu, Liansheng Wang, and Yizhou Yu. nnformer: Interleaved transformer for volumetric segmentation, 2022.
[34] Xin Yu, Qi Yang, Yinchi Zhou, and et. al. Unest: Local spatial representation learning with hierarchical transformer for efficient medical segmentation, 2023.
[35] Wenxuan Wang, Chen Chen, Meng Ding, Jiangyun Li, Hong Yu, and Sen Zha. Transbts: Multimodal brain tumor segmentation using transformer, 2021.
[36] Jieneng Chen, Yongyi Lu, Qihang Yu, Xiangde Luo, Ehsan Adeli, Yan Wang, Le Lu, Alan L. Yuille, and Yuyin Zhou. Transunet: Transformers make strong encoders for medical image segmentation, 2021.
[37] Ali Hatamizadeh, Yucheng Tang, Vishwesh Nath, Dong Yang, Andriy Myronenko, Bennett Landman, Holger Roth, and Daguang Xu. Unetr: Transformers for 3d medical image segmentation, 2021.