Supplementary Material

1. Related Work

1.1. Continuous Latent Space Models

Continuous latent space(CLS) models have gained significant popularity in the field of medical image segmentation. These models offer a flexible and powerful framework for capturing the complex and continuous variations present in medical images. Existing work can be divided into CNNs, transformers, and hybrid models.

CNN-based CLS models: Convolutional Neural Networks (CNNs) have emerged as the widely accepted standard for various computer vision applications. Image segmentation, a task that involves assigning class labels to individual pixels, has particularly benefited from the effectiveness of CNNs. Initial work in the field of image segmentation, such as Fully Convolutional Networks (FCN) [18], and SegNet [1], demonstrated the effectiveness of CNNs in this domain. FCN eliminated the need for fully connected layers and enabled pixel-wise segmentation. SegNet, on the other hand, introduced an encoder-decoder architecture that utilized pooling indices for efficient upsampling. Other notable work consist of DeepLab [7], [20], [36] which improves FCNs by increasing the receptive field and capturing contextual information. CNN models have also achieved remarkable success in medical imaging tasks, notably with the introduction of U-net [12], which inspired subsequent research on U-shaped encoder-decoder architectures [2, 14, 21, 37]. Notably, studies [2, 2, 14] have explored enhancing the encoder-decoder structure with dense skip connections, leading to improved performance in diverse medical domains. Furthermore, encoder-decoder arch. have also shown great success in Semi-SL [29–34]

Multi-head Cross-attention Mechanism emerges as a pivotal convergence point in both natural language processing (NLP) and computer vision domains, amalgamating the potency of multi-head attention and cross-attention mechanisms. It combines the strengths of multi-head attention, which is rooted in the Transformer model's mechanism for focusing on different parts of input, and cross-attention, which extends this capability to interactions between different data types. For instance, in tasks like image captioning and understanding relationships between images and text, the concept proves its utility. This innovation has led to improvements in machine translation, question answering, and text summarization as well, showcasing its potential to revolutionize the handling of diverse data. In our study, we harness Multi-Head Cross Attention to jointly model discrete and continuous latent spaces, capturing complementary fine and coarse-grained information. This is particularly critical in medical image segmentation.

Transformer-based CLS models Vision transformers [8] and their variants [16, 17, 23, 27, 35] have emerged as powerful models in computer vision, akin to the remarkable success of transformers in Natural Language Processing (NLP). These models leverage self-attention mechanisms to learn global information and have achieved impressive results in various visual tasks such as object classification [35], segmentation [4, 6, 16], and detection [3, 38]. Their end-to-end solutions demonstrate the versatility and effectiveness of vision transformers across different vision domains. For instance, Swin Transformer [17] introduces a hierarchical vision transformer that efficiently computes self-attention locally using a shifted windowing approach. CrossViT [5] proposes a dual-branch vision transformer followed by a cross-attention module, enabling richer feature representations while maintaining linear time complexity. These approaches have proven effective in improving performance. In addition to fully transformer-based models, recent methods like Swin-UNet [4] and TransUNet [6] adopt pure transformer architectures with a U-shaped design based on Swin Transformer for 2D segmentation tasks. More recent work such as FCT [25] and Transwnet [28] more accurately capture local and global information and improve medical segmentation performance.

Hybrid CLS models Hybrid models combine the capabilities of CNN and transformers models to capture local and global complementary features tackling the limitation of each. TransUNet [6] combines the strengths of both CNNs and transformers [10] to capture both low-level and high-level features, while UNETR [12] utilizes a transformer-based encoder and a CNN-based decoder for 3D segmentation tasks. More recent approach HiFormer [13]effectively incorporates both global and local information and utilizes a novel transformer-based fusing scheme to maintain feature richness and consistency for the task of 2D medical image segmentation.

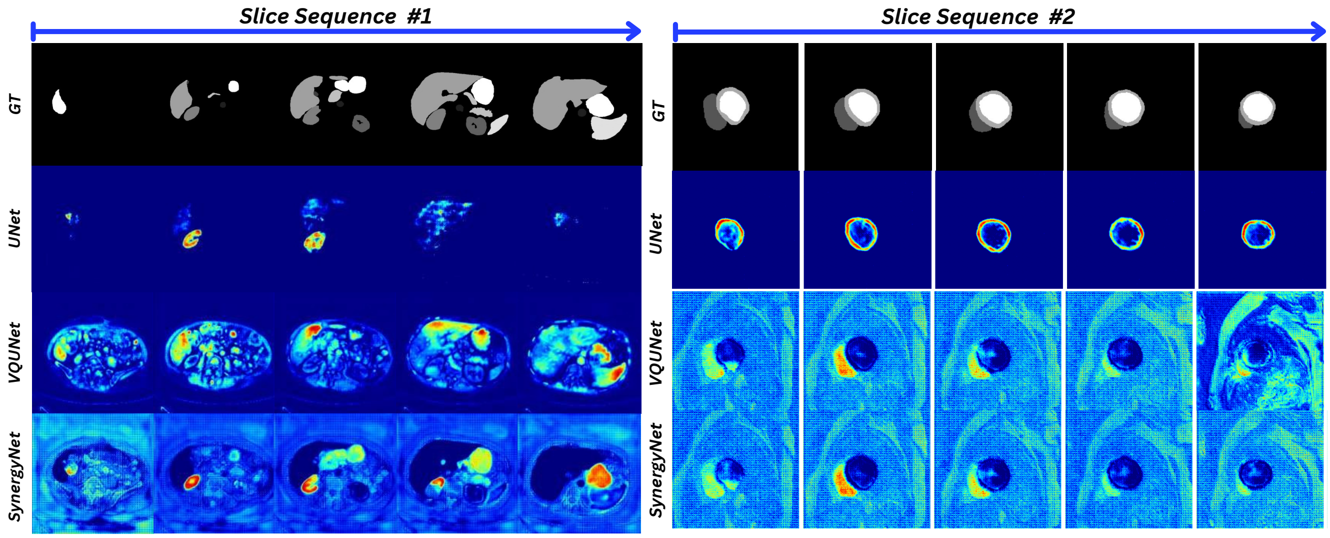

Figure 1. The Grad-CAM visualization demonstrates the characteristics of different models. A UNet with a confined receptive field excels at capturing essential local context, making it suitable for tasks like cardiac segmentation. However, it may overlook the imperative global context required for intricate organ segmentation. Conversely, vector quantization adeptly captures the global context but misses finer details like boundaries. As evident from the visualization, the proposed SynergyNet successfully combines the strengths of both local and global contexts due to the synergy between CLS and DLS components.

1.2. Discrete Latent Space models

Vector Quantization Vector quantization is a classical method for compressed coding that employs a codebook and quantization strategy. Typically using mean square error (MSE), it identifies similar patterns in the codebook to replace original input data. It's akin to discrete representation learning, using a one-hot vector coefficient. Research [11, 19, 24] demonstrates its impact on visual understanding and model robustness. Notably, VQVAE [26] leverages a codebook-based neural network for effective discrete feature distribution learning in images, widely adopted in generative models. In our study, we integrate VQVAE's discrete representation with continuous representation, enhancing medical image segmentation.

Discrete latent space (DLS) models have emerged as a promising approach in various domains, including computer vision and natural language processing. Unlike continuous latent space models, which utilize continuous variables, DLS models leverage discrete variables to represent latent features or concepts. However, the application of DLS in medical image segmentation remains an active and evolving research domain. For instance, Gangloff et al. [9] exploit DLS techniques for anomaly detection, while Jin et al. [15] employ a DLS-based model [26] as a regularizer for semantic segmentation of fundus retina images. Pinaya et al. [22] introduce VQUnet, a DLS-based approach for 3D anomaly detection and segmentation in brain imaging. Additionally, Santhirasekaram et al. [24] demonstrate the robustness and interpretability of vector quantization in semantic segmentation tasks. These studies collectively contribute to the ongoing exploration and advancement of DLS methods in the context of medical image segmentation.

2. Limitations:

While SynergyNet has demonstrated success, it can also be susceptible to issues inherited from its quantizer module, including limited scalability and sensitivity to hyperparameters. Additionally, relying solely on a strategy that selects the most similar codebook item to represent input might face limitations in capturing intricate data patterns, potentially leading to information loss. In cases like ACDC, where fixed ROIs require segmentation, continuous space can offer a more viable option due to vector quantization's structured sparsity property. Furthermore, we observe that CLS and DLS models experience false negatives due to poor inter-class dependencies, a problem partially addressed by SynergyNet. As a result, there is merit in conducting further research on SynergyNet.

3. Future Works:

Integrating SynergyNet with efficient architectures like Swin Transformer [4], HiFormer [13] and others shows promise for further advancements. Exploring SynergyNet's performance with unsupervised models is an intriguing research area that enables leveraging unlabeled data to enhance capabilities in medical image analysis. This holds the potential to improve efficiency and performance in this critical domain.

References

[1] Vijay Badrinarayanan, Alex Kendall, and Roberto Cipolla. Segnet: A deep convolutional encoder-decoder architecture for image segmentation. *IEEE transactions on pattern analysis and machine intelligence*, 39(12):2481–2495, 2017. 1

[2] Sijing Cai, Yunxian Tian, Harvey Lui, Haishan Zeng, Yi Wu, and Guannan Chen. Dense-unet: a novel multiphoton in vivo cellular image segmentation model based on a convolutional neural network. *Quantitative imaging in medicine and surgery*, 10(6):1275, 2020. 1

[3] Zhaowei Cai, Quanfu Fan, Rogerio S Feris, and Nuno Vasconcelos. A unified multi-scale deep convolutional neural network for fast object detection. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11–14, 2016, Proceedings, Part IV 14*, pages 354–370. Springer, 2016. 1

[4] Hu Cao, Yueyue Wang, Joy Chen, Dongsheng Jiang, Xiaopeng Zhang, Qi Tian, and Manning Wang. Swin-unet: Unet-like pure transformer for medical image segmentation. In *Proceedings of the European Conference on Computer Vision*, pages 205–218, 2022. 1, 2

[5] Chun-Fu Richard Chen, Quanfu Fan, and Rameswar Panda. Crossvit: Cross-attention multi-scale vision transformer for image classification. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 357–366, 2021. 1

[6] 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. *arXiv preprint arXiv:2102.04306*, 2021. 1

[7] Liang-Chieh Chen, George Papandreou, Iasonas Kokkinos, Kevin Murphy, and Alan L Yuille. Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. *IEEE transactions on pattern analysis and machine intelligence*, 40(4):834–848, 2017. 1

[8] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020. 1

[9] Hugo Gangloff, Minh-Tan Pham, Luc Courtrai, and Sébastien Lefèvre. Leveraging vector-quantized variational autoencoder inner metrics for anomaly detection. In *Proceedings of the International Conference on Pattern Recognition (ICPR)*, pages 435–441, 2022. 2

[10] Mark S Graham, Petru-Daniel Tudosiu, Paul Wright, Walter Hugo Lopez Pinaya, U Jean-Marie, Yee H Mah, James T Teo, Rolf Jager, David Werring, Parashkev Nachev, et al. Transformer-based out-of-distribution detection for clinically safe segmentation. In *International Conference on Medical Imaging with Deep Learning*, pages 457–476, 2022. 1

[11] Robert Gray. Vector quantization. *IEEE Assp Magazine*, 1(2):4–29, 1984. 2

[12] Ali Hatamizadeh, Yucheng Tang, Vishwesh Nath, Dong Yang, Andriy Myronenko, Bennett Landman, Holger R Roth, and Daguang Xu. Unetr: Transformers for 3d medical image segmentation. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, pages 574–584, 2022. 1

[13] Moein Heidari, Amirhossein Kazerouni, Milad Soltany, Reza Azad, Ehsan Khodapanah Aghdam, Julien Cohen-Adad, and Dorit Merhof. Hiformer: Hierarchical multi-scale representations using transformers for medical image segmentation. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 6202–6212, 2023. 1, 2

[14] Huimin Huang, Lanfen Lin, Ruofeng Tong, Hongjie Hu, Qiaowei Zhang, Yutaro Iwamoto, Xianhua Han, Yen-Wei Chen, and Jian Wu. Unet 3+: A full-scale connected unet for medical image segmentation. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1055–1059. IEEE, 2020. 1

[15] Ge Jin, Xu Chen, and Long Ying. Deep multi-task learning for an autoencoder-regularized semantic segmentation of fundus retina images. *Mathematics*, 10(24):4798, 2022. 2

[16] Tsung-Yi Lin, Piotr Dollár, Ross Girshick, Kaiming He, Bharath Hariharan, and Serge Belongie. Feature pyramid networks for object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2117–2125, 2017. 1

[17] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 10012–10022, 2021. 1

[18] Jonathan Long, Evan Shelhamer, and Trevor Darrell. Fully convolutional networks for semantic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3431–3440, 2015. 1

[19] Chengzhi Mao, Lu Jiang, Mostafa Dehghani, Carl Vondrick, Rahul Sukthankar, and Irfan Essa. Discrete representations strengthen vision transformer robustness. *arXiv preprint arXiv:2111.10493*, 2021. 2

[20] Hyeonwoo Noh, Seunghoon Hong, and Bohyung Han. Learning deconvolution network for semantic segmentation. In *Proceedings of the IEEE international conference on computer vision*, pages 1520–1528, 2015. 1

[21] Ozan Oktay, Jo Schlemper, Loic Le Folgoc, Matthew Lee, Mattias Heinrich, Kazunari Misawa, Kensaku Mori, Steven McDonagh, Nils Y Hammerla, Bernhard Kainz, et al. Attention u-net: Learning where to look for the pancreas. *arXiv preprint arXiv:1804.03999*, 2018. 1

[22] Walter HL Pinaya, Petru-Daniel Tudosiu, Robert Gray, Geraint Rees, Parashkev Nachev, Sebastien Ourselin, and M Jorge Cardoso. Unsupervised brain imaging 3d anomaly detection and segmentation with transformers. *Medical Image Analysis*, 79:102475, 2022. 2

[23] Bo-Kai Ruan, Hong-Han Shuai, and Wen-Huang Cheng. Vision transformers: state of the art and research challenges. *arXiv preprint arXiv:2207.03041*, 2022. 1

[24] Ainkaran Santhirasekaram, Avinash Kori, Mathias Winkler, Andrea Rockall, and Ben Glocker. Vector quantisation for robust segmentation. In *Proceedings of the 25th International Conference on Medical Image Computing and Computer Assisted Intervention (MICCAI 2022)*, pages 663–672, 2022. 2

[25] Athanasios Tragakis, Chaitanya Kaul, Roderick Murray-Smith, and Dirk Husmeier. The fully convolutional transformer for medical image segmentation. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 3660–3669, 2023. 1

[26] Aaron Van Den Oord, Oriol Vinyals, et al. Neural discrete representation learning. *Advances in neural information processing systems*, 30, 2017. 2

[27] Zhuofan Xia, Xuran Pan, Shiji Song, Li Erran Li, and Gao Huang. Vision transformer with deformable attention. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4794–4803, 2022. 1

[28] Yazhen Xie, Yanglin Huang, Yuan Zhang, Xuanya Li, Xiongjun Ye, and Kai Hu. Transwnet: Integrating transformers into cnns via row and column attention for abdominal multi-organ segmentation. In *ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2023. 1

[29] Chenyu You, Weicheng Dai, Fenglin Liu, Haoran Su, Xiaoran Zhang, Lawrence Staib, and James S Duncan. Mine your own anatomy: Revisiting medical image segmentation with extremely limited labels. *arXiv preprint arXiv:2209.13476*, 2022. 1

[30] Chenyu You, Weicheng Dai, Yifei Min, Fenglin Liu, Xiaoran Zhang, Chen Feng, David A Clifton, S Kevin Zhou, Lawrence Hamilton Staib, and James S Duncan. Rethinking semi-supervised medical image segmentation: A variance-reduction perspective. *arXiv preprint arXiv:2302.01735*, 2023. 1

[31] Chenyu You, Weicheng Dai, Yifei Min, Lawrence Staib, and James S Duncan. Bootstrapping semi-supervised medical image segmentation with anatomical-aware contrastive distillation. In *International Conference on Information Processing in Medical Imaging*, pages 641–653. Springer, 2023. 1

[32] Chenyu You, Weicheng Dai, Yifei Min, Lawrence Staib, and James S Duncan. Implicit anatomical rendering for medical image segmentation with stochastic experts. *arXiv preprint arXiv:2304.03209*, 2023. 1

[33] Chenyu You, Weicheng Dai, Yifei Min, Lawrence Staib, Jas Sekhon, and James S Duncan. Action++: Improving semi-supervised medical image segmentation with adaptive anatomical contrast. *arXiv preprint arXiv:2304.02689*, 2023. 1

[34] Chenyu You, Ruihan Zhao, Fenglin Liu, Siyuan Dong, Sandeep Chinchali, Ufuk Topcu, Lawrence Staib, and James Duncan. Class-aware adversarial transformers for medical image segmentation. *Advances in Neural Information Processing Systems*, 35:29582–29596, 2022. 1

[35] Li Yuan, Yunpeng Chen, Tao Wang, Weihao Yu, Yujun Shi, Zi-Hang Jiang, Francis EH Tay, Jiashi Feng, and Shuicheng Yan. Tokens-to-token vit: Training vision transformers from scratch on imagenet. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 558–567, 2021. 1

[36] Hengshuang Zhao, Jianping Shi, Xiaojuan Qi, Xiaogang Wang, and Jiaya Jia. Pyramid scene parsing network. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2881–2890, 2017. 1

[37] Zongwei Zhou, Md Mahfuzur Rahman Siddiquee, Nima Tajbakhsh, and Jianming Liang. Unet++: A nested u-net architecture for medical image segmentation. In *Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support: 4th International Workshop, DLMIA 2018, and 8th International Workshop, ML-CDS 2018, Held in Conjunction with MICCAI 2018, Granada, Spain, September 20, 2018, Proceedings 4*, pages 3–11. Springer, 2018. 1

[38] Xizhou Zhu, Weijie Su, Lewei Lu, Bin Li, Xiaogang Wang, and Jifeng Dai. Deformable detr: Deformable transformers for end-to-end object detection. *arXiv preprint arXiv:2010.04159*, 2020. 1