MARSeg: Enhancing Medical Image Segmentation with MAR and Adaptive Feature Fusion

Jeonghyun Hwang*[1], Seungyeon Rhee*[1], Minjeong Kim*[1], Thanaporn Viriyasaranon[2], and Jang-Hwan Choi[1,2](✉)

[1] Artificial Intelligence Convergence, Department of Artificial Intelligence and Software, Ewha Womans University, Seoul, Korea
{jeonghyunhwang,syeonrhee,gskmj20,choij}@ewha.ac.kr
[2] Department of Artificial Intelligence, Ewha Womans University, Seoul, Korea
thanaporn.v@ewhain.net

Abstract. Recent advances in Masked Autoregressive (MAR) models highlight their ability to preserve fine-grained details through continuous vector representations, making them highly suitable for tasks requiring precise pixel-level delineation. Motivated by these strengths, we introduce **MARSeg**, a novel segmentation framework tailored for medical images. Our method first pre-trains a MAR model on large-scale CT scans, capturing both global structures and local details without relying on vector quantization. We then propose a Generative Parallel Adaptive Feature Fusion (GPAF) module that effectively unifies spatial and channel-wise attention, thereby combining latent features from the pre-trained MAE encoder and decoder. This approach preserves essential boundary information while enhancing the robustness of organ and tumor segmentation. Experimental results on multiple CT datasets from the Medical Segmentation Decathlon (MSD) demonstrate that MARSeg outperforms existing state-of-the-art methods in terms of Dice Similarity Coefficient (DSC) and Intersection over Union (IoU), confirming its efficacy in handling complex anatomical and pathological variations. The code is available at https://github.com/Ewha-AI/MARSeg.

Keywords: Medical Image Segmentation · Masked Autoregressive · Adaptive Feature Fusion · CT Imaging

1 Introduction

Accurate segmentation of cancerous lesions and related anatomical structures from CT images is paramount for reliable diagnosis, treatment planning, and disease monitoring [16,17,6]. Such segmentation not only enables precise targeting in radiotherapy and surgical planning but also facilitates automated detection methods and reduces observer variability. However, the inherent heterogeneity of

* Equally contributed.

tumors, coupled with substantial variability in organ shapes, sizes, and imaging conditions, makes it challenging to achieve robust and generalizable performance [3,1,22].

Traditional deep learning methods, often based on convolutional neural networks (CNNs) or transformers, rely on large-scale annotated datasets and tend to struggle under data-scarce or complex imaging scenarios [15,19]. While CNN-based models excel at capturing local features, their limited receptive fields can hinder modeling of long-range dependencies. Transformer-based approaches leverage global self-attention but may lose fine-grained details that are crucial for precisely delineating complex organ boundaries. Consequently, even hybrid CNN–transformer strategies face difficulties in balancing broad contextual information with local detail.

Meanwhile, recent progress in generative models such as Generative Adversarial Networks (GANs) [7] and Variational Autoencoders (VAEs) [12] has showcased the capacity to learn rich, high-dimensional representations of medical images [11,23]. These models, often pre-trained on large datasets, capture complex anatomical features and spatial relationships, offering a promising resource for downstream tasks. Of particular relevance is the Masked Autoregressive (MAR) model [13], which combines a Masked Autoencoder (MAE) [8] with diffusion techniques. MAR has demonstrated exceptional capability in preserving high-resolution details, making it an attractive candidate for medical image segmentation.

Motivated by these developments, we introduce **MARSeg**, a novel multi-stage segmentation framework that leverages the powerful representational capacity of MAR. Specifically, we first perform large-scale pre-training on CT data to learn robust global and local features. We then propose a Generative Parallel Adaptive Feature Fusion (GPAF) module that integrates spatial and channel attention in parallel, effectively merging global context with fine-grained local details. During the segmentation stage, a portion of MAE component of the MAR model is frozen, while the image tokenizer, Adaptive Feature Fusion module, and segmentation head are fine-tuned. This two-stage strategy overcomes the inherent limitations of conventional approaches, leading to improved performance in challenging segmentation scenarios.

Concretely, our contributions can be summarized as follows:

- **Multi-Stage Framework:** We devise a two-step approach that first learns a robust and semantically rich representation via MAR pre-training on CT data, followed by a segmentation-specific fine-tuning step.
- **Generative Parallel Adaptive Feature Fusion Module:** We introduce a fusion module that combines parallel spatial and channel attention mechanisms, enabling effective integration of global and local information for precise boundary delineation.

To assess MARSeg's effectiveness, we conduct extensive experiments on CT datasets from the Medical Segmentation Decathlon (MSD) [2], encompassing four different organs and their respective tumors. Experimental results show that MARSeg not only balances global context and fine structural details but also

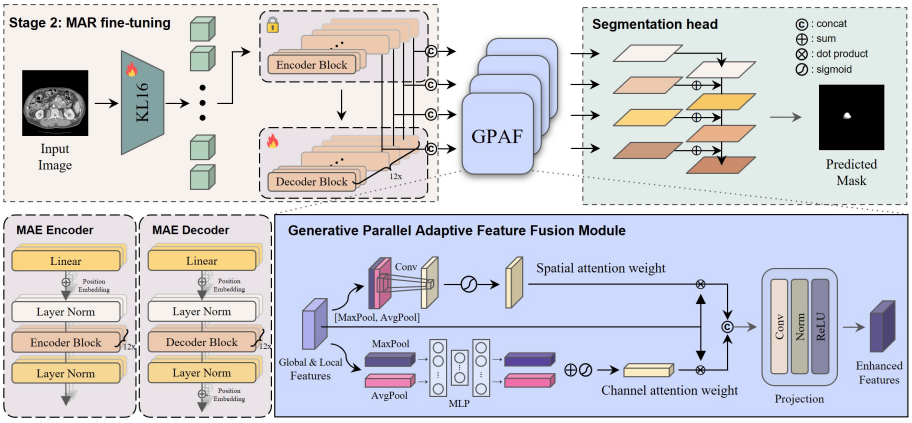

Fig. 1. Overall architecture of MARSeg. A 2D CT slice is first processed by the KL-16 encoder (tokenizer) and then passed to the Masked Autoregressive (MAR) model. The fused features from encoder-decoder outputs are refined using the Generative Parallel Adaptive Feature Fusion (GPAF) module, which applies spatial and channel attention. Finally, the resulting feature maps are fed into a segmentation head to produce the final mask.

achieves superior segmentation performance and domain robustness compared to existing state-of-the-art methods.

2 Method

MARSeg is designed as a multi-stage framework. In the first stage, we pre-train a MAR on large-scale CT images (Section 2.1). In the second stage, shown in Fig. 1, we fuse the pre-trained encoder and decoder features via our proposed module and fine-tune them for the final segmentation task (Sections 2.2 and 2.3).

2.1 MAR Pre-training

In the pre-training stage, a Masked Autoencoder (MAE) is integrated with a continuous vector-based diffusion process to form a generative model, referred to as MAR. Specifically, we employ the KL-16 tokenizer from the LDM framework [18], which is regularized by Kullback–Leibler divergence rather than traditional vector quantization [13]. This design choice effectively preserves both global anatomical structures and fine-grained local details while minimizing information loss.

The pre-training pipeline proceeds as follows:

1. **Latent Encoding.** Each CT slice is encoded into a latent representation $\mathbf{z}$ via the KL-16 tokenizer. This pre-trained tokenizer, kept fixed during pre-training, captures key tissue-related features of the input data.

2. **Masked Autoencoding.** A random subset of latent patches in $\mathbf{z}$ is masked, forcing the MAE encoder to learn a global context from incomplete data. The MAE decoder then reconstructs the missing patches, guided by the global cues from the encoder.
3. **Diffusion Refinement.** A diffusion network refines the reconstructed output by iteratively denoising it, enhancing structural fidelity and detail preservation. This step consolidates local fine-grained information within the broader anatomical context.

By training on a combination of pancreas and liver CT images from the Medical Segmentation Decathlon (MSD) dataset, the MAR model learns robust representations that capture both high-level semantic information and nuanced boundary details. These pre-trained weights serve as strong initialization for the subsequent segmentation stage, where only the KL-16 tokenizer and MAE decoder are fine-tuned, while the MAE encoder is kept frozen.

2.2 Generative Parallel Adaptive Feature Fusion Module

To effectively leverage the semantic information learned by the MAR model for segmentation, we propose a Generative Parallel Adaptive Feature Fusion (GPAF) module. Figure 1 (in Section 2) shows an overview of how the fused encoder-decoder features pass through this module before reaching the segmentation head.

Feature Preparation. Let O_{e_i} and O_{d_i} denote the MAE encoder and decoder outputs at layer i, respectively. Initially, each has shape

$$O_{e_i},\, O_{d_i} \;\in\; \mathbb{R}^{B\times C_i\times L_i}, \tag{1}$$

where B is the batch size, C_i is the number of channels at layer i, and L_i is the patch-sequence length. We reshape each tensor from a 1D patch sequence into a 2D spatial grid:

$$\tilde{O}_{e_i},\, \tilde{O}_{d_i} \;\in\; \mathbb{R}^{B\times C_i\times H_i\times W_i}, \quad \text{where } H_i \times W_i = L_i. \tag{2}$$

We then concatenate these reshaped tensors along the channel dimension and apply a 1×1 convolutional projection (denoted as *Proj*) to unify their channel sizes:

$$O_{c,i} = \mathit{Proj}\Big(\mathrm{Concat}[\tilde{O}_{e_i}, \tilde{O}_{d_i}]\Big) \;\in \mathbb{R}^{B\times d_i\times H_i\times W_i}, \tag{3}$$

where d_i is the new channel dimension after concatenation and projection.

Parallel Spatial and Channel Attention. Next, $O_{c,i}$ is fed into two parallel attention branches: spatial attention (SA) and channel attention (CA).

Spatial Attention (SA). We highlight important spatial regions by combining average and max pooling along the channel dimension, followed by a convolutional layer with a sigmoid activation:

$$O_{\mathrm{SA},i} \;=\; O_{c,i} \;\odot\; \sigma\Big(\mathrm{Conv}_{k\times k}\big[\mathrm{Mean}(O_{c,i}),\; \mathrm{Max}(O_{c,i})\big]\Big), \tag{4}$$

where $\odot$ denotes element-wise multiplication, and σ is the sigmoid function. This mechanism assigns higher weights to spatial positions deemed more relevant.

Channel Attention (CA). Simultaneously, we evaluate the importance of each channel by aggregating spatial information (e.g., using average and max pooling in the spatial domain):

$$O_{\mathrm{CA},i} \;=\; O_{c,i} \;\odot\; \sigma\Big(W_2\big(\mathrm{ReLU}\big(W_1(\mathrm{AvgPool}(O_{c,i}) \;+\; \mathrm{MaxPool}(O_{c,i}))\big)\big)\Big), \quad (5)$$

where W_1 and W_2 are learnable parameters. The channel-wise weights reflect each feature map's relative contribution to the overall representation.

Finally, the outputs of SA and CA are concatenated along the channel axis, and another 1×1 projection is applied:

$$O_{\mathrm{GPAF},i} \;=\; \mathit{Proj}\Big(\mathrm{Concat}[\,O_{\mathrm{SA},i},\, O_{\mathrm{CA},i}]\Big). \quad (6)$$

Thus, $O_{\mathrm{GPAF},i} \in \mathbb{R}^{B\times d_i'\times H_i\times W_i}$ merges both global and local cues while adaptively refining feature emphasis across spatial and channel dimensions.

2.3 Fine-Tuning Stage for Segmentation

After pre-training the MAR model, we pass the input image through the KL-16 encoder to obtain latent representations, which are then fed into the frozen MAE encoder and the trainable MAE decoder. In this phase, the KL-16 tokenizer and the MAE decoder are fine-tuned, whereas the MAE encoder remains fixed. This strategy ensures that the highly capable generative encoder retains its global context while adapting the decoder and tokenizer to the specific segmentation domain.

Feature Extraction. We extract the output features from the final four MAE encoder and decoder layers, denoted by

$$\{\tilde{O}_{e_i} \mid i \in \{9,10,11,12\}\}, \quad \{\tilde{O}_{d_i} \mid i \in \{9,10,11,12\}\}.$$

Each pair of encoder-decoder features $(\tilde{O}_{e_i}, \tilde{O}_{d_i})$ is fused via the proposed GPAF module (Section 2.2), forming a set of multi-scale features that encapsulate both high-level and fine-grained information.

FPN-based Segmentation Head. To convert these fused features into segmentation masks, we employ a simplified Feature Pyramid Network (FPN) [14]. Each fused feature map $O_{\mathrm{GPAF},i}$ is projected to a fixed channel size via a 1×1 lateral convolution and these projected maps are subsequently merged via element-wise summation with the fused feature maps in the chosen layers in top-down manner. Then we use consecutive 3×3 and 1×1 convolutions to produce the final segmentation logits. This design preserves both low-level cues and high-level semantic context, enabling MARSeg to achieve robust and precise boundary delineation under diverse anatomical variations.

3 Experiments

3.1 Experimental Setup

Dataset and Metric. For model training and validation, we utilize the publicly available dataset provided by the Medical Segmentation Decathlon (MSD) [2], a biomedical image analysis challenge designed to assess the generalizability of segmentation algorithms. Among its tasks, the MSD Pancreas dataset comprises 420 cases with a total of 26,719 portal-venous phase CT slices of patients who underwent surgical resection for pancreatic masses. To further expand our evaluation, we incorporate three additional MSD datasets:

- **MSD Liver:** Consisting of 201 CT scans (58,641 slices), this dataset requires segmenting both the liver and any tumors therein.
- **MSD Colon:** Containing 190 CT scans (13,486 slices) focused on primary colon cancer, it presents a challenge due to the heterogeneous appearance of lesions and surrounding structures.
- **MSD Spleen:** Comprising 61 CT scans (3,650 slices) from patients undergoing chemotherapy for liver metastases, it varies significantly in splenic size and shape.

Leveraging these diverse datasets enables a comprehensive evaluation of our model's segmentation capabilities under varied anatomical and pathological scenarios. All input images are resized to 256×256, and we split each dataset by patient in an 8:1:1 ratio for training, validation, and testing. We report performance using two widely adopted metrics in medical image segmentation: the Dice Similarity Coefficient (DSC) and the Intersection over Union (IoU).

Implementation Details. Our experiments are implemented in PyTorch and run on two NVIDIA RTX 6000 Ada Generation GPUs. We employ a combined loss function (Dice loss : cross-entropy loss = 1:2) and use the AdamW optimizer with a learning rate of 2.5×10^{-5}. Each model is trained for 200 epochs with a batch size of 32, requiring approximately one day of computation on the aforementioned GPUs for liver datasets.

For data preprocessing, we apply organ-specific clipping to remove irrelevant regions based on organ characteristics, followed by intensity normalization. The 3D volumes are then sliced into 2D images to form the final training dataset. This approach ensures that domain-specific information is retained while reducing input size and complexity.

3.2 Results and Analysis

We evaluate the performance of our proposed method, **MARSeg**, on four MSD datasets and compare it with various state-of-the-art segmentation methods. As shown in Table 1, MARSeg achieves the highest DSC and IoU scores, surpassing

Table 1. Comparison with state-of-the-art medical image segmentation methods on four MSD tasks. Best values are marked in **red-bold**, and second-best values are marked in blue-underline.

Methods	Panc Tumor		Liver Tumor		Colon Cancer		Spleen	
	DSC	IoU	DSC	IoU	DSC	IoU	DSC	IoU
SwinUNet [4]	48.27	31.81	69.16	52.85	41.28	26.00	91.00	83.49
TransUNet [5]	33.32	19.99	**75.77**	**61.00**	42.42	26.92	92.74	86.46
MTUNet [21]	38.01	23.46	69.87	53.69	19.60	10.86	94.66	89.86
MISSFormer [9]	35.98	21.94	64.46	47.56	10.83	5.73	93.98	88.65
nnUNet [10]	38.74	24.02	67.35	50.77	42.53	27.01	95.84	92.01
CFATransUnet [20]	35.25	21.39	65.55	48.75	12.86	6.87	93.20	87.27
MARSeg (Ours)	**49.45**	**32.84**	74.90	59.87	**51.70**	**34.86**	**96.08**	**92.56**

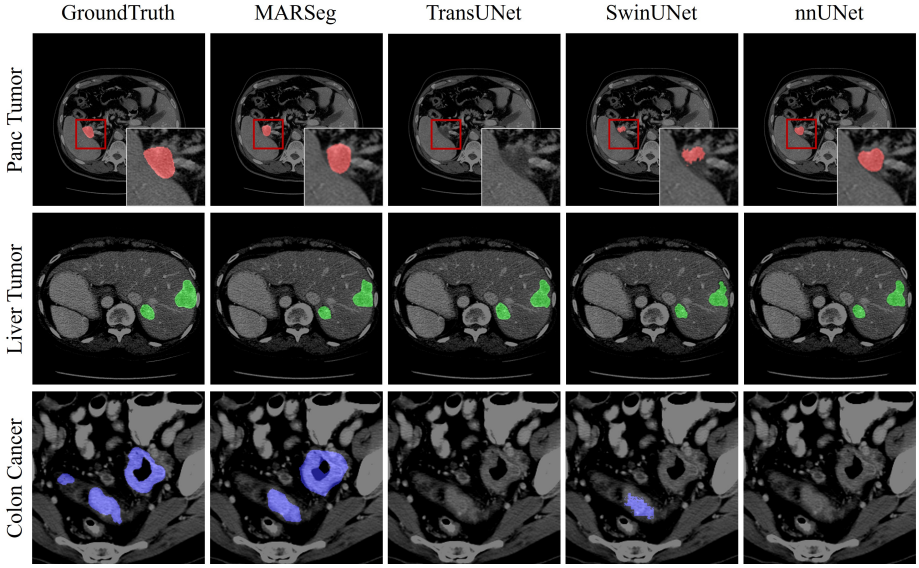

Fig. 2. Qualitative comparison of segmentation performance on sample CT slices. Columns (from left to right) show the ground truth (GT), MARSeg (ours), TransUNet, SwinUNet, and nnUNet. Red rectangles denote regions of interest, which are magnified in the lower-right inset.

other approaches in pancreatic tumor, colon cancer and spleen (organ) segmentation tasks. Although pancreatic tumor detection remains challenging, our model achieves the highest performance with a Dice score of 49.45%. In the case of liver tumor segmentation, our method performs comparably to state-of-the-art approaches. In the case of colon cancer—where most models struggle with segmentation—our model achieves a Dice score that is 21.57% higher and an IoU that is 29.06% higher compared to the second-best approach, nnUNet. Further-

more, for relatively small organs such as the spleen, our model achieves superior performance compared to other methods, attaining a Dice score of 96.08% and an IoU of 92.56%.

Figure 2 provides a visual comparison of segmentation results on the MSD dataset. MARSeg demonstrates superior boundary delineation and more accurate lesion segmentation, highlighting its ability to capture both fine-grained and global contextual features effectively.

Ablation Study. To compare the performance of the proposed GPAF module with conventional feature fusion methods, we conduct ablation studies on the MSD Pancreas and Spleen dataset. The results in Table 2 demonstrated that our GPAF module achieves superior performance over other feature fusion methods. Furthermore, we evaluate the effect of the selected feature layers from the generative model (MAR). Table 3 presents utilizing the last four layers of the encoder and decoder yields better performance. For the layers selection experiment, we exclusively employed the Cross-Entropy loss function.

Table 2. Ablation study for different feature fusion methods.

Feature Fusion	Spleen		Panc Tumor	
	DSC	IoU	DSC	IoU
w/o Fusion	95.39	91.19	37.75	23.27
Concat	95.67	91.71	46.36	30.17
Cross Attention	95.25	90.93	45.61	29.55
GPAF (Ours)	**96.16**	**92.62**	**48.81**	**32.28**

Table 3. Ablation study for different fusion layers with proposed module.

Layers	Spleen		Panc Tumor	
	DSC	IoU	DSC	IoU
[3,6,9,12]	95.97	92.26	38.92	24.16
[9,10,11,12]	**96.17**	**92.62**	**40.93**	**25.73**

4 Conclusion

In this work, we proposed **MARSeg**, a segmentation framework that leverages the powerful representational capabilities of a Masked Autoregressive (MAR) generative model. By pre-training on large-scale CT data, our approach captures both global structural patterns and fine-grained local details, addressing common challenges in medical image segmentation. We further introduced a Generative Parallel Adaptive Feature Fusion module that unifies encoder-decoder features through parallel spatial and channel attention, enhancing boundary delineation for organs and tumors.

Comprehensive experiments on various MSD datasets confirm that MARSeg consistently outperforms state-of-the-art methods, achieving higher Dice and IoU scores. These results indicate that our generative pre-training strategy and proposed fusion module can effectively integrate semantic richness with localized anatomical precision. Future directions include extending MARSeg to 3D

volumetric segmentation, investigating robust multi-institutional scenarios, and exploring semi-supervised or low-data regimes to further capitalize on the MAR model's generative strengths. Our MARSeg enhances consistency in diagnosis, treatment planning, and disease monitoring through accurate segmentation, and alleviate physician's workload.

Acknowledgments. This work was partly supported by the National Research Foundation of Korea (NRF) grant funded by the Korea government (MSIT) (No. RS-2025-02215813 and No. RS-2025-00520578); by the Technology development Program of MSS [S3146559]; and by Institute of Information & Communications Technology Planning & Evaluation (IITP) grant funded by the Korea government (MSIT) (No. RS-2022-00155966).

Disclosure of Interests. The authors have no competing interests to declare that are relevant to the content of this article.

References

1. Almotairi, S., Kareem, G., Aouf, M., Almutairi, B., Salem, M.A.M.: Liver tumor segmentation in ct scans using modified segnet. Sensors **20**(5) (2020)
2. Antonelli, M., Reinke, A., Bakas, S., Farahani, K., Kopp-Schneider, A., Landman, B.A., Litjens, G., Menze, B., Ronneberger, O., Summers, R.M., et al.: The medical segmentation decathlon. Nature communications **13**(1), 4128 (2022)
3. Balasubramanian, P.K., Lai, W.C., Seng, G.H., C, K., Selvaraj, J.: Apestnet with mask r-cnn for liver tumor segmentation and classification. Cancers **15**(2) (2023)
4. Cao, H., Wang, Y., Chen, J., Jiang, D., Zhang, X., Tian, Q., Wang, M.: Swin-unet: Unet-like pure transformer for medical image segmentation (2021), `https://arxiv.org/abs/2105.05537`
5. Chen, J., Lu, Y., Yu, Q., Luo, X., Adeli, E., Wang, Y., Lu, L., Yuille, A.L., Zhou, Y.: Transunet: Transformers make strong encoders for medical image segmentation (2021), `https://arxiv.org/abs/2102.04306`
6. DeSouza, N.M., van Der Lugt, A., Deroose, C.M., Alberich-Bayarri, A., Bidaut, L., Fournier, L., Costaridou, L., Oprea-Lager, D.E., Kotter, E., Smits, M., et al.: Standardised lesion segmentation for imaging biomarker quantitation: a consensus recommendation from esr and eortc. Insights into imaging **13**(1), 159 (2022)
7. Goodfellow, I.J., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., Bengio, Y.: Generative adversarial networks (2014), `https://arxiv.org/abs/1406.2661`
8. He, K., Chen, X., Xie, S., Li, Y., Dollár, P., Girshick, R.: Masked autoencoders are scalable vision learners (2021), `https://arxiv.org/abs/2111.06377`
9. Huang, X., Deng, Z., Li, D., Yuan, X.: Missformer: An effective medical image segmentation transformer (2021), `https://arxiv.org/abs/2109.07162`
10. Isensee, F., Jaeger, P.F., Kohl, S.A., Petersen, J., Maier-Hein, K.H.: nnu-net: a self-configuring method for deep learning-based biomedical image segmentation. Nature methods **18**(2), 203–211 (2021)
11. Kazeminia, S., Baur, C., Kuijper, A., van Ginneken, B., Navab, N., Albarqouni, S., Mukhopadhyay, A.: Gans for medical image analysis. Artificial Intelligence in Medicine **109**, 101938 (2020)

12. Kingma, D.P., Welling, M.: Auto-encoding variational bayes (2022), https://arxiv.org/abs/1312.6114
13. Li, T., Tian, Y., Li, H., Deng, M., He, K.: Autoregressive image generation without vector quantization (2024), https://arxiv.org/abs/2406.11838
14. Lin, T.Y., Dollár, P., Girshick, R., He, K., Hariharan, B., Belongie, S.: Feature pyramid networks for object detection (2017), https://arxiv.org/abs/1612.03144
15. Litjens, G., Kooi, T., Bejnordi, B.E., Setio, A.A.A., Ciompi, F., Ghafoorian, M., van der Laak, J.A., van Ginneken, B., Sánchez, C.I.: A survey on deep learning in medical image analysis. Medical Image Analysis **42**, 60–88 (2017)
16. Moghbel, M., Mashohor, S., Mahmud, R., Saripan, M.I.B.: Automatic liver tumor segmentation on computed tomography for patient treatment planning and monitoring. EXCLI journal **15**, 406 (2016)
17. Rekik, I., Allassonnière, S., Carpenter, T.K., Wardlaw, J.M.: Medical image analysis methods in mr/ct-imaged acute-subacute ischemic stroke lesion: Segmentation, prediction and insights into dynamic evolution simulation models. a critical appraisal. NeuroImage: Clinical **1**(1), 164–178 (2012)
18. Rombach, R., Blattmann, A., Lorenz, D., Esser, P., Ommer, B.: High-resolution image synthesis with latent diffusion models (2022), https://arxiv.org/abs/2112.10752
19. Roy, S., Koehler, G., Ulrich, C., Baumgartner, M., Petersen, J., Isensee, F., Jaeger, P.F., Maier-Hein, K.: Mednext: Transformer-driven scaling of convnets for medical image segmentation (2024)
20. Wang, C., Wang, L., Wang, N., Wei, X., Feng, T., Wu, M., Yao, Q., Zhang, R.: Cfatransunet: Channel-wise cross fusion attention and transformer for 2d medical image segmentation. Computers in Biology and Medicine **168**, 107803 (2024)
21. Wang, H., Xie, S., Lin, L., Iwamoto, Y., Han, X.H., Chen, Y.W., Tong, R.: Mixed transformer u-net for medical image segmentation. In: ICASSP 2022-2022 IEEE international conference on acoustics, speech and signal processing (ICASSP). pp. 2390–2394. IEEE (2022)
22. Yao, X., Song, Y., Liu, Z.: Advances on pancreas segmentation: a review. Multimedia Tools and Applications **79**(9), 6799–6821 (2020)
23. Yi, X., Walia, E., Babyn, P.: Generative adversarial network in medical imaging: A review. Medical Image Analysis **58**, 101552 (2019)