Class-Aware Adversarial Transformers for Medical Image Segmentation

Anonymous Author(s)
Affiliation
Address
email

Abstract

Transformers have made remarkable progress towards modeling long-range dependencies within the medical image analysis domain. However, current transformer-based models suffer from several disadvantages: (1) existing methods fail to capture the important features of the images due to the naive tokenization scheme; (2) the models suffer from information loss because they only consider single-scale feature representations; and (3) the segmentation label maps generated by the models are not accurate enough without considering rich semantic contexts and anatomical textures. In this work, we present `CASTformer`, a novel type of adversarial transformers, for 2D medical image segmentation. First, we take advantage of the pyramid structure to construct multi-scale representations and handle multi-scale variations. We then design a novel class-aware transformer module to better learn the discriminative regions of objects with semantic structures. Lastly, we utilize an adversarial training strategy that boosts segmentation accuracy and correspondingly allows a transformer-based discriminator to capture high-level semantically correlated contents and low-level anatomical features. Our experiments demonstrate that `CASTformer` dramatically outperforms previous state-of-the-art transformer-based approaches on three benchmarks, obtaining 2.54%-5.88% absolute improvements in Dice over previous models. Further qualitative experiments provide a more detailed picture of the model's inner workings, shed light on the challenges in improved transparency, and demonstrate that transfer learning can greatly improve performance and reduce the size of medical image datasets in training, making `CASTformer` a strong starting point for downstream medical image analysis tasks. Codes and models will be made available to public.

1 Introduction

Accurate and consistent measurements of anatomical features and functional information on medical images can greatly assist radiologists in making accurate and reliable diagnoses, treatment planning, and post-treatment evaluation [1]. Convolutional neural networks (CNNs) have been the de-facto standard for medical image analysis tasks. However, such methods fail in explicitly modeling long-range dependencies due to the intrinsic locality and weight sharing of the receptive fields in convolution operations. Such a deficiency in context modeling at multiple scales often yields sub-optimal segmentation capability in capturing rich anatomical features of variable shapes and scales (*e.g.*, tumor regions with different structures and sizes) [2, 3]. Moreover, using transformers has been shown to be more promising in computer vision [4–9] for utilizing long-range dependencies than other, traditional CNN-based methods. In parallel, transformers with powerful global relation

Submitted to 36th Conference on Neural Information Processing Systems (NeurIPS 2022). Do not distribute.

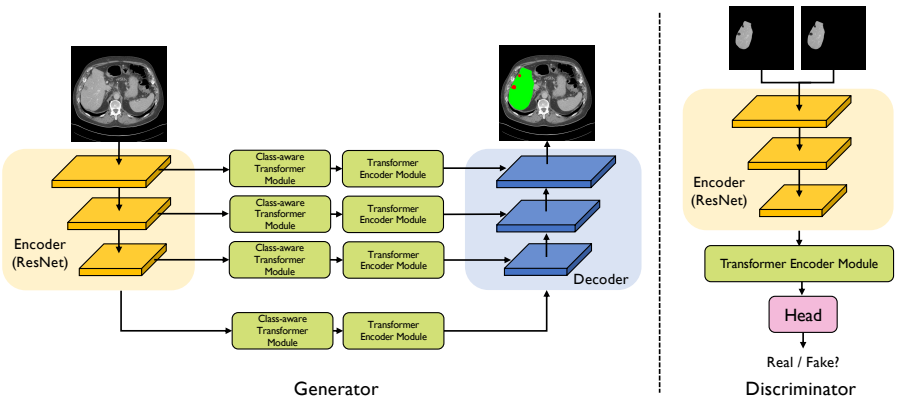

Figure 1: Our proposed `CASTformer` consists of a transformer-based generator (*i.e.*, `CATformer`) and a discriminator.

modeling abilities have become the standard starting point for training on a wide range of downstream medical imaging analysis tasks, such as image segmentation [7, 10–13], image synthesis [14–16], and image enhancement [17–20].

Medical image semantic segmentation can be formulated as a typical dense prediction problem, which aims at performing pixel-level classification on the feature maps. Recently, Chen *et al.* [7] introduced `TransUNet`, which inherits the advantage of both `UNet` [21] and `Transformers` [4], to exploit high-resolution informative representations in the spatial dimension by CNNs and the powerful global relation modeling by Transformers. Although existing transformer-based approaches have proved promising in the medical image segmentation task, there remain several formidable challenges, because (1) the model outputs a single-scale and low-resolution feature representation; (2) prior work mainly adopts the standard tokenization scheme, hard splitting an image into a sequence of image patches of size 16×16, which may fail to capture inherent object structures and the fine-grained spatial details for the downstream dense prediction task; (3) compared to the standard convolution, the transformer architecture requires a grid structure, and thus lacks the capability to localize regions that contain objects of interest instead of the uninteresting background; and (4) existing methods are usually deficient in ensuring the performance without capturing both global and local contextual relations among pixels. We argue that transformer-based segmentation models are not yet robust enough to replace CNN-based methods, and investigate several above-mentioned key challenges transformer-based segmentation models still face.

Inspired by recent success of vision transformer networks [3, 4, 22–28], we make a step towards a **more practical scenario** in which we only assume access to pre-trained models on public computer vision datasets, and a relatively small medical dataset, which we can use the weights of the pre-trained models to achieve higher accuracy in the medical image analysis tasks. These settings are particularly appealing as (1) such models can be easily adopted on typical medical datasets; (2) such a setting only requires limited training data and annotations; and (3) transfer learning typically leads to better performance [29–32]. Inspired by such findings, we propose several novel strategies for expanding its learning abilities to our setting, considering both multi-scale anatomical feature representations of interesting objects and transfer learning in the medical imaging domain.

First, we aim to model multi-scale variations by learning feature maps of different resolutions. Thus, we propose to incorporate the pyramid structure into the transformer framework for medical image segmentation, which enables the model to capture rich global spatial information and local multi-scale context information. Additionally, we consider medical semantic segmentation as a sequence-to-sequence prediction task. The standard patch tokenization scheme in [4] is an art—splitting it into several fixed-size patches and linearly embedding them into input tokens. Even if significant progress is achieved, model performance is likely to be sub-optimal. We address this issue by introducing a novel class-aware transformer module, drawing inspiration from a progressive sampling strategy in image classification [33], to adaptively and selectively learn interesting parts of objects. This

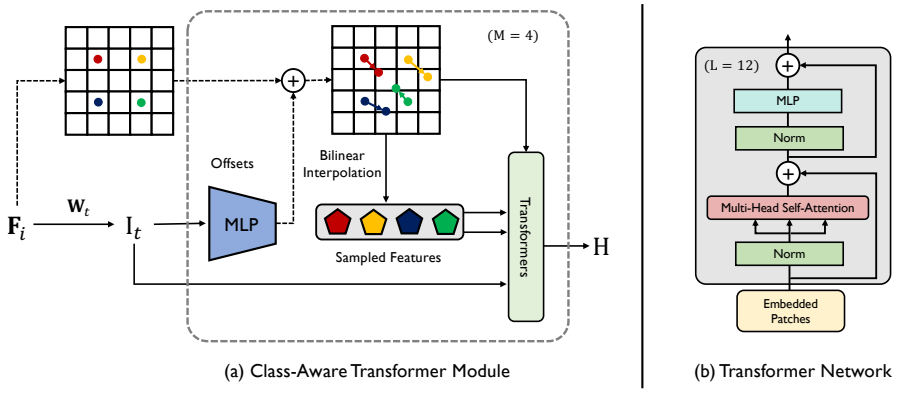

Figure 2: An illustration of (a) our class-aware transformer module and (b) transformer network. Our class-aware transformer module iteratively samples discriminative locations. A group of points is initialized in the regularly-spaced location. At each step, given the feature map $\mathbf{F}_i$, we iteratively update its sampling locations by adding them with the estimated offsets of the last step. Note that only 4 points are shown for a clear presentation, and there are more points in the actual implementation.

essentially allows us to obtain effective anatomical features from spatial attended regions within the medical images, so as to guide the segmentation of objects or entities.

Second, we adopt the idea of Generative Adversarial Networks (GANs) to improve segmentation performance and correspondingly enable a transformer-based discriminator to learn low-level anatomical features and high-level semantics. The standard GANs are not guaranteed to prioritize the most informative demonstrations on interesting anatomical regions, and mixing irrelevant regions (*i.e.*, background) creates inferior contexts, which drastically underperform segmentation performance [2, 34, 3]. Additionally, it is well-known that they are notoriously difficult to train and prone to model collapse [35]. Training vision transformers is also tedious and requires large amounts of labeled data, which largely limits the training quality. We use a more refined strategy, where, for each input, we combine it with the predicted segmentation mask to create the image with anatomical demonstrations. We also leverage the pre-trained checkpoints to compensate the need of large-dataset training, thereby providing a good starting point with more discriminative visual demonstrations. We refer to our approach as `CASTformer`, **c**lass-**a**ware adver**s**arial **t**rans**former**s: a strong transformer-based method for 2D medical image segmentation. Our contributions are summarized as follows:

- **Novel Network Architecture:** We make the first attempt to build a GAN using a transformer-based architecture for the 2D medical image segmentation task. We incorporate the pyramid structure into the generator to learn rich global and local multi-scale spatial representations, and also devise a novel class-aware transformer module by progressively learning the interesting regions correlated with semantic structures of images. To the best of our knowledge, we are the **first work** to explore these techniques in the context of medical imaging segmentation.
- **Better Understanding Inner Workings:** We conduct careful analyses to understand the model's inner workings, how the sampling strategy works, and how different training factors lead to the final performance. We highlight that it is more effective to progressively learn distinct contextual representations with the class-aware transformer module, resulting in more accurate and robust models that applied better to a variety of downstream medical image analysis tasks (See Appendix J and K).
- **Remarkable Performance Improvements:** `CASTformer` contributes towards a dramatic improvement across three datasets we evaluate on. For instance, we achieve Dice scores of 82.55% and 73.82% by obtaining gains up to 5.88% absolute improvement compared to prior methods on the Synapse multi-organ dataset. We illustrate the benefits of leveraging pre-trained models from the computer vision domain, and provide suggestions for future research that could be less susceptible to the confounding effects of training data from the natural image domain.

2 Related Work

CNN-based Segmentation Networks Before the emergence of transformer-based methods, CNNs were the *de facto* methods in medical image segmentation tasks [36–50]. For example, Ronneberger *et al.* [21] proposed a deep 2D `UNet` architecture, combining skip connections between opposing convolution and deconvolution layers to achieve promising performance on a diverse set of medical segmentation tasks. Han *et al.* [51] developed a 2.5D 24-layer Fully Convolutional Network (FCN) for liver segmentation tasks where the residual block was incorporated into the model. To further improve segmentation accuracy, Kamnitsas *et al.* [52] proposed a dual pathway 11-layer 3D CNN, and also employed a 3D fully connected conditional random field (CRF) [53] as an additional pairwise constraint between neighboring pixels for the challenging task of brain lesion segmentation.

Transformers in Medical Image Segmentation Recent studies [7, 10, 12, 13, 54–59] have focused on developing transformer-based methods for medical image analysis tasks. Recently, Chen *et al.* [7] proposed `TransUNet`, which takes advantage of both `UNet` and `Transformers`, to exploit high-resolution informative information in the spatial dimension by CNNs and the global dependencies by Transformers. Cao *et al.* [10] explored how to use a pure transformer for medical image analysis tasks. However, the results do not lead to better performance. These works mainly utilized hard splitting some highly semantically correlated regions without capturing the inherent object structures. In this work, beyond simply using the naive tokenization scheme in [4, 7], we aim at enabling the transformer to capture global information flow to estimate offsets towards regions of interest.

Transformer in Generative Adversarial Networks Adversarial learning has proved to be a very useful and widely applicable technique for learning generative models of arbitrarily complex data distributions in the medical domain. As the discriminator D differentiates between real and fake samples, the adversarial loss serves as the regularization constraint to enforce the generator G to predict more realistic samples. Inspired by such recent success [60–65], Jiang *et al.* [60] proposed to build a GAN pipeline with two pure transformer-based architectures in synthesizing high-resolution images. Esser *et al.* [25] first used a convolutional GAN model to learn a codebook of context-rich visual features, followed by transformer architecture to learn the compositional parts. Hudson *et al.* [3] proposed a bipartite self-attention on StyleGAN to propagate latent variables to the evolving visual features. Despite such success, it requires high computation costs due to the quadratic complexity, which fundamentally hinders its applicability to the real world. Besides the image generation task, we seek to take a step forward in tackling the challenging task of 2D medical image segmentation.

3 Method

Our proposed approach is presented in Figure 1. Given the input image $\mathbf{x} \in \mathbb{R}^{H \times W \times 3}$, similar to `TransUNet` architecture [7], our proposed generator network G, termed `CATformer`, is comprised of four key components: encoder (feature extractor) module, class-aware transformer module, transformer encoder module, and decoder module. As shown in Figure 1, our generator has four stages with four parallel subnetworks. All stages share a similar architecture, which contains a patch embedding layer, class-aware layer, and L_i Transformer encoder layers.

Encoder Module. Our method adopts a CNN-Transformer hybrid model design instead of using a pure transformer, which uses 40 convolutional layers, to generate multi-scale feature maps. Such a *convolutional stem* setting provides two advantages: (1) using *convolutional stem* helps transformers perform better in the downstream vision tasks [66–72]; (2) it provides high-resolution feature maps with parallel medium- and low-resolution feature maps to help boost better representations. In this way, we can construct the feature pyramid for the Transformers, and utilize the multi-scale feature maps for the downstream medical segmentation task. With the aid of feature maps of different resolutions, our model is capable of modeling multi-resolution spatially local contexts.

Hierarchical Feature Representation. Inspired by recent success in object detection [73], we deviate from `TransUNet` by generating a single-resolution feature map, and our focus is on extracting CNN-like multi-level features $\mathbf{F}_i$, where $i \in \{1, 2, 3, 4\}$, to achieve high segmentation accuracy by leveraging high-resolution features and low-resolution features. More precisely, in the first stage, we utilize the encoder module to obtain the dense feature map $\mathbf{F}_1 \in \mathbb{R}^{\frac{H}{2} \times \frac{W}{2} \times C_1}$, where $(\frac{H}{2}, \frac{W}{2}, C_1)$ is the spatial feature resolution and the number of feature channels. In a similar way, we can formulate the following feature maps as follows: $\mathbf{F}_2 \in \mathbb{R}^{\frac{H}{2} \times \frac{W}{2} \times (C_1 \cdot 4)}$, $\mathbf{F}_3 \in \mathbb{R}^{\frac{H}{4} \times \frac{W}{4} \times (C_1 \cdot 8)}$, and $\mathbf{F}_4 \in \mathbb{R}^{\frac{H}{8} \times \frac{W}{8} \times (C_1 \cdot 12)}$. Then, we divide $\mathbf{F}_1$ into $\frac{HW}{16^2}$ patches with the patch size P of $16 \times 16 \times 3$, and feed the flattened patches into a learnable linear transformation to obtain the patch embeddings of size $\frac{HW}{16^2} \times C_1$.

Class-Aware Transformer Module. The class-aware transformer module (CAT) is designed to adaptively focus on useful regions of objects (*e.g.*, the underlying anatomical features and structural information). Our CAT module is largely inspired by the recent success for image classification [33], but we deviate from theirs as follows: (1) we remove the vision transformer module in [33] to alleviate the computation and memory usage; (2) we use 4 separate Transformer Encoder Modules (TEM), which will be introduced below; (3) we incorporate M CAT modules on multi-scale representations to allow for contextual information of anatomical features to propagate into the representations. Our class-aware transformer module is an iterative optimization process. In particular, we apply the class-aware transformer module to obtain the sequence of tokens $\mathbf{I}_{M,1} \in \mathbb{R}^{C \times (n \times n)}$, where $(n \times n)$ and M are the number of samples on each feature map and the total iterative number, respectively. As shown in Figure 2, given the feature map $\mathbf{F}_1$, we iteratively update its sampling locations by adding them with the estimated offset vectors of the last step, which can be formulated as follows:

$$\mathbf{s}_{t+1} = \mathbf{s}_t + \mathbf{o}_t, \ t \in \{1, \ldots, M-1\}, \tag{1}$$

where $\mathbf{s}_t \in \mathbb{R}^{2 \times (n \times n)}$, and $\mathbf{o}_t \in \mathbb{R}^{2 \times (n \times n)}$ are the sampling location and the predicted offset vector at t-th step. Specifically, the $\mathbf{s}_1$ is initialized at the regularly spaced sampling grid. The i-th sampling location $\mathbf{s}_1^i$ is defined as follows:

$$\mathbf{s}_1^i = [\beta_i^y \tau_h + \tau_h/2, \beta_i^x \tau_w + \tau_w/2], \tag{2}$$

where $\beta_i^y = \lfloor i/n \rfloor$, $\beta_i^x = i - \beta_i^y * n$. The step sizes in the y (row index) and x (column index) directions denote $\tau_h = H/n$ and $\tau_w = W/n$, respectively. $\lfloor \cdot \rfloor$ is the floor operation. We can define the initial token on the input feature map in the following form: $\mathbf{I}_t^{'} = \mathbf{F}_i(\mathbf{s}_t)$, where $t \in \{1, \ldots, M\}$, and $\mathbf{I}_t^{'} \in \mathbb{R}^{C \times (n \times n)}$ denotes the initial sampled tokens at t-th step. We set the sampling function as the bilinear interpolation, since it is differentiable with respect to both the sampling locations $\mathbf{s}_t$ and the input feature map $\mathbf{F}_i$. We do an element-wise addition of the current positional embedding of the sampling locations, the initial sampled tokens, and the estimated tokens of the last step, and then we can obtain the output tokens at each step:

$$\begin{aligned} \mathbf{S}_t &= \mathbf{W}_t \mathbf{s}_t \\ \mathbf{V}_t &= \mathbf{I}_t^{'} \oplus \mathbf{S}_t \oplus \mathbf{I}_{t-1} \\ \mathbf{I}_t &= \text{Transformer}(\mathbf{V}_t), \ t \in \{1, \ldots, M\}, \end{aligned} \tag{3}$$

where $\mathbf{W}_t \in \mathbb{R}^{C \times 2}$ is the learnable matrix that embeds $\mathbf{s}_t$ to positional embedding $\mathbf{S}_t \in \mathbb{R}^{C \times (n \times n)}$, and $\oplus$ is the element-wise addition. Transformer$(\cdot)$ is the transformer encoder layer, as we will show in the following paragraphs. We can compute the estimated sampling location offsets as:

$$\mathbf{o}_t = \theta_t(\mathbf{I}_t), \ t \in \{1, \ldots, M-1\}, \tag{4}$$

where $\theta_t(\cdot) \in \mathbb{R}^{2 \times (n \times n)}$ is the learnable linear mapping for the estimated sampling offset vectors. It is worth noting that these operations are all differentiable, thus the model can be learned in an end-to-end fashion.

Transformer Encoder Module. The transformer encoder module (TEM) is designed to model long-range contextual information by aggregating global contextual information from the complete sequences of input image patches embedding. In implementations, the transformer encoder module

follows the architecture in ViT [4], which is composed of Multi-head Self-Attention (MSA), and MLP blocks, which can be formulated as:

$$\mathbf{E}_0 = [\mathbf{x}_p^1\mathbf{H}; \mathbf{x}_p^2\mathbf{H}; \cdots ; \mathbf{x}_p^N\mathbf{H}] + \mathbf{H}_{pos}, \tag{5}$$

$$\mathbf{E'}_i = \mathrm{MSA}(\mathrm{LN}(\mathbf{E}_{i-1})) + \mathbf{E}_{i-1}, \tag{6}$$

$$\mathbf{E}_i = \mathrm{MLP}(\mathrm{LN}(\mathbf{E'}_i)) + \mathbf{E'}_i, \tag{7}$$

where $i = 1 \dots M$, and $\mathrm{LN}(\cdot)$ is the layer normalization. $\mathbf{H} \in \mathbb{R}^{(P^2 \cdot C) \times D}$ and $\mathbf{H}_{pos} \in \mathbb{R}^{N \times D}$ denote the patch embedding projection and the position embedding.

Decoder Module. The decoder is designed to generate the segmentation mask based on four output feature maps of different resolutions. In implementations, rather than designing a hand-crafted decoder module that requires high computational demand, we incorporate a lightweight All-MLP decoder [74], and such a simple design allows us to yield a powerful representation much more efficiently. The decoder includes the following criteria: 1) the channel dimension of multi-scale features is unified through the MLP layers; 2) we up-sample the features to 1/4th and concatenate them together; 3) we utilize the MLP layer to fuse the concatenated features, and then predict the multi-class segmentation mask $\mathbf{y}'$ from the fused features.

Discriminator Network. We use the R50+ViT-B/16 hybrid model pre-trained on ImageNet from ViT [4] as a starting point for our discriminator design, in this case using the pre-trained strategies to learn effectively on the limited size target task data. Then, we simply apply a two-layer multi-layered perceptron (MLP) to make a prediction about the identity of the class-aware image. Following previous work [2], we first utilize the ground truth image $\mathbf{x}$ and the predicted segmentation mask $\mathbf{y}'$ to obtain the class-aware image $\tilde{\mathbf{x}}$ (*i.e.*, pixel-wise multiplication of $\mathbf{x}$ and $\mathbf{y}'$). It is important to note that this construction re-uses the pre-trained weights and does not introduce any additional parameters. D seeks to classify between real and fake samples [75]. G and D compete with each other through attempting to reach an equilibrium point of the minimax game. Using this structure enables the discriminator to model long-range dependencies, making it better assess the medical image fidelity. This also essentially endows the model with a more holistic understanding of the anatomical visual modality (categorical features).

Training Objective. As to the loss function and training configurations, we adopt the settings used in Wasserstein GAN (WGAN) [76], and use WGAN-GP loss [77]. We jointly use the segmentation loss [7, 13] and WGAN-GP loss to train G. Concretely, the segmentation loss includes the dice loss and cross-entropy loss. Hence, the training process of `CASTformer` can be formulated as:

$$\mathcal{L}_G = \lambda_1 \mathcal{L}_{\text{CE}} + \lambda_2 \mathcal{L}_{\text{DICE}} + \lambda_3 \mathcal{L}_{\text{WGAN-GP}}, \tag{8}$$

where λ_1, λ_2, λ_3 determine the importance of each loss term. See Appendix H for an ablation study.

4 Experimental Setup

Datasets. We experiment on multiple challenging benchmark datasets: Synapse[1], LiTS, and MP-MRI. More details can be found in Appendix A.

Implementation Details. We utilize the AdamW optimizer [78] in all our experiments. For training our generator and discriminator, we use a learning rate of $5e^{-4}$ with a batch size of 6, and train each model for 300 epochs for all datasets. We set the sampling number n on each feature map and the total iterative number M as 16 and 4, respectively. See Appendix B, C, H for details on the training configuration, model architecture and hyperparameters. We also adopt the input resolution and patch size P as 224×224 and 14, respectively. We set $\lambda_1 = 0.5$, $\lambda_2 = 0.5$, and $\lambda_3 = 0.1$ in this experiments. In the testing stage, we adopt four metrics to evaluate the segmentation performance: Dice coefficient (Dice), Jaccard Index (Jaccard), 95% Hausdorff Distance (95HD), and Average Symmetric Surface Distance (ASD). All our experiments are implemented in Pytorch 1.7.0. We train all models on a single NVIDIA GeForce RTX 3090 GPU with 24GB of memory.

[1] `https://www.synapse.org/#!Synapse:syn3193805/wiki/217789`

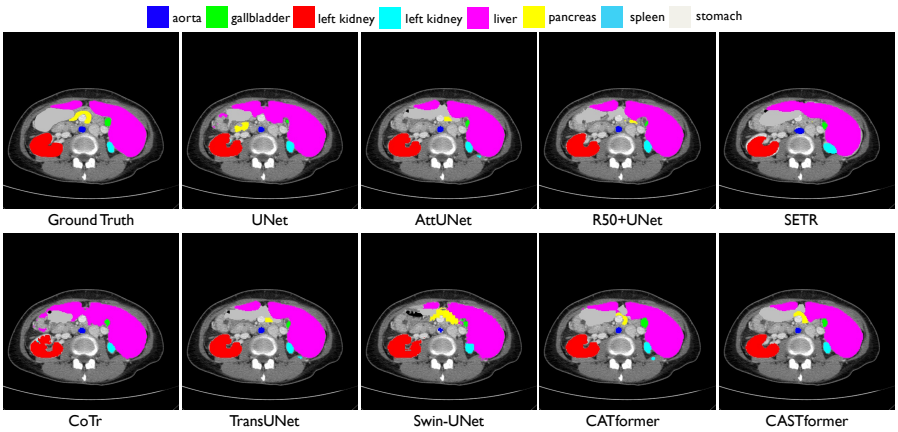

Figure 3: Visual comparisons with other methods on Synapse dataset. As observed, `CASTformer` achieves superior performance with detailed anatomical features and the boundary information of different organs.

Table 1: Quantitative segmentation results on the Synapse multi-organ CT dataset.

Framework		Average				Aorta	Gallbladder	Kidney (L)	Kidney (R)	Liver	Pancreas	Spleen	Stomach
Encoder	Decoder	DSC ↑	Jaccard ↑	95HD ↓	ASD ↓								
	`UNet` [21]	70.11	59.39	44.69	14.41	84.00	56.70	72.41	62.64	86.98	48.73	81.48	67.96
	`AttnUNet` [79]	71.70	61.38	34.47	10.00	82.61	61.94	76.07	70.42	87.54	46.70	80.67	67.66
`ResNet50`	`UNet` [21]	73.51	63.81	29.65	8.83	82.21	55.06	76.71	73.07	89.36	53.52	84.91	73.22
`ResNet50`	`AttnUNet` [79]	74.74	62.69	33.04	9.49	83.68	58.63	79.08	74.53	90.81	55.76	83.80	71.68
	`SETR` [80]	66.30	54.19	29.09	7.16	66.63	38.34	74.45	68.49	92.18	35.91	83.01	71.41
	`CoTr w/o CNN-encoder` [13]	54.82	42.49	69.58	20.37	63.22	37.86	67.10	60.61	88.48	15.46	60.74	45.12
	`CoTr` [13]	72.60	61.25	41.55	12.42	83.27	60.41	79.58	73.01	91.93	45.07	82.84	64.67
	`TransUNet` [7]	77.48	64.78	31.69	8.46	87.23	63.13	81.87	77.02	94.08	55.86	85.08	75.62
	`SwinUNet` [10]	76.33	65.64	27.16	8.32	85.47	66.53	83.28	79.61	94.29	56.58	90.66	76.60
•	`CATformer` (ours)	82.17	73.22	**16.20**	**4.28**	88.98	67.16	85.72	81.69	95.34	66.53	90.74	81.20
○	`CASTformer` (ours)	**82.55**	**74.69**	22.73	5.81	89.05	67.48	86.05	82.17	95.61	67.49	91.00	81.55

5 Results

We compare our approaches (*i.e.*, `CATformer` and `CASTformer`) with previous state-of-the-art transformer-based segmentation methods, including `UNet` [21], `AttnUNet` [79], `ResNet50+UNet` [21], `ResNet50+AttnUNet` [79], `SETR` [80], `CoTr w/o CNN-encoder` [13], `CoTr` [13], `TransUNet` [7], `SwinUnet` [10] on Synapse, LiTS, and MP-MRI datasets. More results are in Appendix D and E.

Experiments: Synapse Multi-organ. The quantitative results on the Synapse dataset are shown in Table 1. The results are visualized in Figure 3. It can be observed that our `CATformer` outperforms the previous best model by a large margin and achieves a $4.69\% - 8.44\%$ absolute improvement in Dice and Jaccard, respectively. Our `CASTformer` achieves the best performance of 82.55% and 74.69%, dramatically improving the previous state-of-the-art model (`TransUNet`) by $+5.07\%$ and $+9.91\%$, in terms of both Dice and Jaccard scores. This shows that the anatomical visual information is useful for the model to gain finer control in localizing local semantic regions. As also shown in Table 1, our `CASTformer` achieves absolute Dice improvements of $+2.77\%$, $+2.51\%$, $+1.35\%$, $+4.95\%$ on large organs (*i.e.*, left kidney, right kidney, liver, stomach) respectively. Such improvements demonstrate the effectiveness of learning the evolving anatomical features of the image, as well as accurately identifying the boundary information of large organs. We also observed similar trends that, compared to the previous state-of-the-art results, our `CASTformer` obtains 89.05%, 67.48%, 67.49% in terms of Dice on small organs (*i.e.*, aorta, gallbladder, pancreas) respectively, which yields big improvements of $+1.82\%$, $+0.95\%$, $+10.91\%$. This clearly demonstrates the superiority of our models, allowing for a spatially finer control over the segmentation process.

Experiments: LiTS. To further evaluate the effectiveness of our proposed approaches, we compare our models on the LiTS dataset. Experimental results on the LiTS CT dataset are summarized

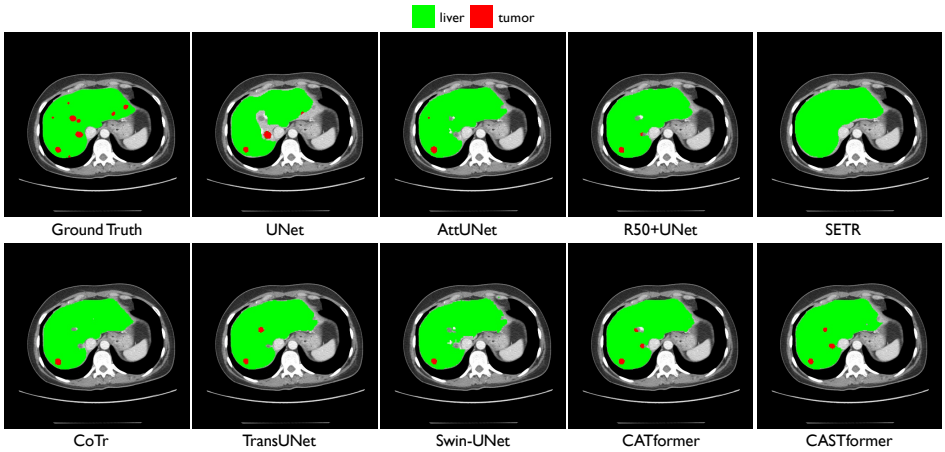

Figure 4: Visual comparisons with other methods on LiTS dataset. As observed, `CASTformer` achieves superior performance with detailed anatomical information (*e.g.*, the tumor regions in red).

in Appendix Table 6. As is shown, we observe that our `CATformer` yields a 72.39% Dice score, outperforming all other methods. Moreover, our `CASTformer` significantly outperforms all previous approaches, including the previous best `TransUNet`, and establishes a new state-of-the-art of 73.82% and 64.91% in terms of Dice and Jaccard, which are 5.88% and 4.66% absolute improvements better than `TransUNet`. For example, our `CASTformer` achieves the best performance of 95.88% Dice on the liver region by 2.48%, while it dramatically increases the result from 42.49% to 51.76% on the tumor region, demonstrating that our model achieves competitive performance on liver and tumor segmentation. As shown in Figure 4, our method is capable of predicting high-quality object segmentation, considering the fact that the improvement in such a setting is challenging. This demonstrates: (1) the necessity of adaptively focusing on the region of interests; and (2) the efficacy of semantically correlated information. Compared to previously high-performing models, our two approaches achieve significant improvements on all datasets, demonstrating their effectiveness.

6 Analysis

In this section, we conduct thorough analyses of our `CASTformer` in the following aspects: transfer learning, model components, effects of iteration number N (Appendix F), sampling number n (Appendix G), hyperparameter selection (Appendix H), different GAN-based loss functions (Appendix I), and gain a better understanding of the model's inner workings (Appendix J and K).

Transfer Learning. We consider whether we can leverage the pre-trained model commonly used in computer vision literature [4], to provide more evidence for the beneficial impact of our network performance. We use `CATformer` (G) as the baseline and evaluate all the settings on the Synapse multi-organ dataset. To put our results in perspective, we compare with six ways of using pre-trained R50+ViT-B/16 hybrid model from ViT [4] for transfer learning, namely (1) `CATformer`: w/o pre-trained and w/ pre-trained; (2) `CASTformer`: *both* w/o pre-trained, *only* w/ pre-trained D, *only* w/ pre-trained G, and *both* w/ pre-trained G and D.

Table 2: Effect of transfer learning in our `CATformer` and `CASTformer` on the Synapse multi-organ dataset.

Model	DSC	Jaccard	95HD	ASD
• `CATformer` (w/o pre-trained)	74.84	65.61	31.81	7.23
• `CATformer` (w/ pre-trained)	82.17	73.22	16.20	4.28
○ `CASTformer` (*both* w/o pre-trained)	73.64	62.68	42.77	11.76
○ `CASTformer` (*only* w/ pre-trained D)	78.87	69.36	30.54	9.17
○ `CASTformer` (*only* w/ pre-trained G)	81.46	71.80	27.36	6.91
○ `CASTformer` (*both* w/ pre-trained)	82.55	74.69	22.73	5.81

Table 3: Ablation on model component: Baseline; `CATformer` w/o CAT; `CATformer` w/o TEM; and `CATformer`.

Model	DSC	Jaccard	95HD	ASD
Baseline	77.48	64.78	31.69	8.46
• `CATformer` w/o CAT	80.09	70.56	25.62	7.30
• `CATformer` w/o TEM	81.35	72.66	24.43	7.17
• `CATformer`	82.17	73.22	16.20	4.28
○ `CASTformer`	82.55	74.69	22.73	5.81

The results are in Table 2. Overall, we observe that using "w/ pre-trained" leads to higher accuracy than "w/o pre-trained", with significant improvements for the smaller sizes of datasets, suggesting that using "w/ pre-trained" provides us a good set of initial parameters for the downstream tasks. With using pre-trained weights, `CATformer` outperforms the setting without using pre-trained weights by a large margin and achieves 7.33% and 7.61% absolute improvements in terms of Dice and Jaccard. `CASTformer` ("both w/ pre-trained") also yields big improvements ($+8.91\%$ and $+12.01\%$) in Dice and Jaccard. This suggests that `CASTformer` is better at both initializing from the pre-trained models and better at gathering the anatomical information in a more adaptive and selective manner. As shown in Table 2, surprisingly, there is a significant discrepancy between only using "w/ pre-trained D" and "w/ pre-trained G": for example, `CASTformer` achieves 78.87% in Dice with only w/ pre-trained D, while `CASTformer` achieves 81.46% if only G uses the pre-trained weights. This demonstrates that only using pre-trained weight in D might be the culprit for the exploitation of anatomical information.

Our results suggest that (1) utilizing pre-trained models in the computer vision domain can help the model quickly adapt to new downstream medical segmentation tasks *without* re-building billions of anatomical representations; (2) we find that leveraging pre-trained weights can further boost the performance because it can mitigate the discrepancy between training and inference; and (3) it also creates a possibility to adapt our model to the typical medical dataset with the smaller size.

Ablation of Model Components. Our key observation is that it is crucial to build high-quality anatomical representations through each model component. To show the strengths of our approach, we examine the following variants and inspect each key component on the Synapse multi-organ segmentation dataset: (1) **Baseline**: we remove the class-aware transformer module and the transformer encoder module in our `CATformer` as the baseline, similar to `TransUNet` defined in [7]; (2) `CATformer` **w/o CAT**: we *only* remove the class-aware transformer module in our `CATformer`; (3) `CATformer` **w/o TEM**: we *only* remove the transformer encoder module in our `CATformer`; (4) `CATformer`: this is our G model; and (5) `CASTformer`: this is our final model described in Section 3. Table 3 summarizes the results of all the variants. As is shown, we observe that compared to the baseline model, both `CATformer` w/o CAT and `CATformer` w/o TEM are able to develop a better holistic understanding of global shapes/structures and fine anatomical details, thus leading to large performance improvements ($+2.61\%$ and $+3.87\%$) in terms of Dice. Our results show the class-aware transformer module is useful in improving the segmentation performance, suggesting that the discriminative regions of medical images are particularly effective. Finally, one thing worth noticing is that incorporating both the class-aware transformer module and the transformer encoder module performs better than *only* using a single module, highlighting the importance of two modules in our `CATformer`.

7 Conclusion and Discussion of Broader Impact

In this work, we have introduced `CASTformer`, a simple yet effective type of adversarial transformers, for 2D medical image segmentation. The key insight is to integrate the multi-scale pyramid structure to capture rich global spatial information and local multi-scale context information. Furthermore, `CASTformer` also benefits from our proposed class-aware transformer module to progressively and selectively learn interesting parts of the objects. Lastly, the generator-discriminator design is used to boost segmentation performance and correspondingly enable the transformer-based discriminator to capture low-level anatomical features and high-level semantics. Comprehensive experiments demonstrate that our `CASTformer` outperforms the previous state-of-the-art on three popular medical datasets considerably. We conduct extensive analyses to study the robustness of our approach, and form a more detailed understanding of desirable properties in the medical domain (*i.e.*, transparency and data efficiency).

Overall, we hope that this model can serve as a solid baseline for 2D medical image segmentation and motivate further research in medical image analysis tasks. It also provides a new perspective on transfer learning in medical domain, and initially shed novel insights towards understanding neural network behavior. As such pattern is hard to quantify, we expect more mechanistic explanations for

clinical practise. We also plan to optimize the transformer-based architectures for the downstream medical image analysis tasks both in terms of data and model parameters.

Broader Impact. We acknowledge that our research will not pose significant risks of societal harm to society. Our work is scientific nature and will have the potential to positively contribute to a number of real-world clinical applications that establish high-quality and end-to-end medical image segmentation systems. We expect our approach to contribute to the grand goal of building more secured and trustworthy clinical AI.

References

[1] Mehrdad Moghbel, Syamsiah Mashohor, Rozi Mahmud, and M Iqbal Bin Saripan. Review of liver segmentation and computer assisted detection/diagnosis methods in computed tomography. *Artificial Intelligence*, 2017.

[2] Yuan Xue, Tao Xu, Han Zhang, L Rodney Long, and Xiaolei Huang. Segan: Adversarial network with multi-scale l 1 loss for medical image segmentation. *Neuroinformatics*, 2018.

[3] Drew A Hudson and Larry Zitnick. Generative adversarial transformers. In *International Conference on Machine Learning (ICML)*, 2021.

[4] 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. In *International Conference on Learning Representations (ICLR)*, 2020.

[5] Daquan Zhou, Bingyi Kang, Xiaojie Jin, Linjie Yang, Xiaochen Lian, Zihang Jiang, Qibin Hou, and Jiashi Feng. Deepvit: Towards deeper vision transformer. *arXiv preprint arXiv:2103.11886*, 2021.

[6] Aditya Desai, Zhaozhuo Xu, Menal Gupta, Anu Chandran, Antoine Vial-Aussavy, and Anshumali Shrivastava. Raw nav-merge seismic data to subsurface properties with mlp based multi-modal information unscrambler. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2021.

[7] 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. In *International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, 2021.

[8] Max Jaderberg, Karen Simonyan, Andrew Zisserman, et al. Spatial transformer networks. *Advances in neural information processing systems*, 28, 2015.

[9] Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022.

[10] 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. *arXiv preprint arXiv:2105.05537*, 2021.

[11] Wenxuan Wang, Chen Chen, Meng Ding, Hong Yu, Sen Zha, and Jiangyun Li. Transbts: Multimodal brain tumor segmentation using transformer. In *International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, 2021.

[12] Jeya Maria Jose Valanarasu, Poojan Oza, Ilker Hacihaliloglu, and Vishal M Patel. Medical transformer: Gated axial-attention for medical image segmentation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, 2021.

[13] Yutong Xie, Jianpeng Zhang, Chunhua Shen, and Yong Xia. Cotr: Efficiently bridging cnn and transformer for 3d medical image segmentation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, 2021.

[14] Lingke Kong, Chenyu Lian, Detian Huang, Yanle Hu, Qichao Zhou, et al. Breaking the dilemma of medical image-to-image translation. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2021.

[15] Nicolae-Catalin Ristea, Andreea-Iuliana Miron, Olivian Savencu, Mariana-Iuliana Georgescu, Nicolae Verga, Fahad Shahbaz Khan, and Radu Tudor Ionescu. Cytran: Cycle-consistent transformers for non-contrast to contrast ct translation. *arXiv preprint arXiv:2110.06400*, 2021.

[16] Onat Dalmaz, Mahmut Yurt, and Tolga Çukur. Resvit: Residual vision transformers for multi-modal medical image synthesis. *arXiv preprint arXiv:2106.16031*, 2021.

[17] Yilmaz Korkmaz, Salman UH Dar, Mahmut Yurt, Muzaffer Özbey, and Tolga Çukur. Unsupervised mri reconstruction via zero-shot learned adversarial transformers. *arXiv preprint arXiv:2105.08059*, 2021.

[18] Zhicheng Zhang, Lequan Yu, Xiaokun Liang, Wei Zhao, and Lei Xing. Transct: Dual-path transformer for low dose computed tomography. In *International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, 2021.

[19] Achleshwar Luthra, Harsh Sulakhe, Tanish Mittal, Abhishek Iyer, and Santosh Yadav. Eformer: Edge enhancement based transformer for medical image denoising. *arXiv preprint arXiv:2109.08044*, 2021.

[20] Dayang Wang, Zhan Wu, and Hengyong Yu. Ted-net: Convolution-free t2t vision transformer-based encoder-decoder dilation network for low-dose ct denoising. *arXiv preprint arXiv:2106.04650*, 2021.

[21] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, 2015.

[22] Han Zhang, Ian Goodfellow, Dimitris Metaxas, and Augustus Odena. Self-attention generative adversarial networks. In *International Conference on Learning Representations (ICLR)*, 2019.

[23] Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-end object detection with transformers. In *European Conference on Computer Vision (ECCV)*, 2020.

[24] Yanhong Zeng, Jianlong Fu, and Hongyang Chao. Learning joint spatial-temporal transformations for video inpainting. In *European Conference on Computer Vision (ECCV)*, 2020.

[25] Patrick Esser, Robin Rombach, and Bjorn Ommer. Taming transformers for high-resolution image synthesis. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021.

[26] 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. *arXiv preprint arXiv:2103.14030*, 2021.

[27] Hugo Touvron, Matthieu Cord, Matthijs Douze, Francisco Massa, Alexandre Sablayrolles, and Hervé Jégou. Training data-efficient image transformers & distillation through attention. *arXiv preprint arXiv:2012.12877*, 2020.

[28] Hugo Touvron, Matthieu Cord, Alexandre Sablayrolles, Gabriel Synnaeve, and Hervé Jégou. Going deeper with image transformers. *arXiv preprint arXiv:2103.17239*, 2021.

[29] S. Kevin Zhou, Hayit Greenspan, Christos Davatzikos, James S. Duncan, Bram Van Ginneken, Anant Madabhushi, Jerry L. Prince, Daniel Rueckert, and Ronald M. Summers. A review of deep learning in medical imaging: Imaging traits, technology trends, case studies with progress highlights, and future promises. *Proceedings of the IEEE*, 2021.

[30] Gonglei Shi, Li Xiao, Yang Chen, and S Kevin Zhou. Marginal loss and exclusion loss for partially supervised multi-organ segmentation. *Medical Image Analysis*, 2021.

[31] Qingsong Yao, Li Xiao, Peihang Liu, and S Kevin Zhou. Label-free segmentation of covid-19 lesions in lung ct. *IEEE Transactions on Medical Imaging*, 2021.

[32] Jiuwen Zhu, Yuexiang Li, Yifan Hu, Kai Ma, S Kevin Zhou, and Yefeng Zheng. Rubik's cube+: A self-supervised feature learning framework for 3d medical image analysis. *Medical Image Analysis*, 2020.

[33] Xiaoyu Yue, Shuyang Sun, Zhanghui Kuang, Meng Wei, Philip Torr, Wayne Zhang, and Dahua Lin. Vision transformer with progressive sampling. In *IEEE International Conference on Computer Vision (ICCV)*, 2021.

[34] Pauline Luc, Camille Couprie, Soumith Chintala, and Jakob Verbeek. Semantic segmentation using adversarial networks. *arXiv preprint arXiv:1611.08408*, 2016.

[35] Tim Salimans, Ian Goodfellow, Wojciech Zaremba, Vicki Cheung, Alec Radford, and Xi Chen. Improved techniques for training gans. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2016.

[36] Hayit Greenspan, Bram Van Ginneken, and Ronald M Summers. Guest editorial deep learning in medical imaging: Overview and future promise of an exciting new technique. *IEEE Transactions on Medical Imaging*, 2016.

[37] Geert Litjens, Thijs Kooi, Babak Ehteshami Bejnordi, Arnaud Arindra Adiyoso Setio, Francesco Ciompi, Mohsen Ghafoorian, Jeroen AWM van der Laak, Bram Van Ginneken, and Clara I Sánchez. A survey on deep learning in medical image analysis. *Medical Image Analysis*, 2017.

[38] Yizhe Zhang, Lin Yang, Jianxu Chen, Maridel Fredericksen, David P Hughes, and Danny Z Chen. Deep adversarial networks for biomedical image segmentation utilizing unannotated images. In *International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, 2017.

[39] Xiaomeng Li, Lequan Yu, Hao Chen, Chi-Wing Fu, and Pheng-Ann Heng. Semi-supervised skin lesion segmentation via transformation consistent self-ensembling model. *arXiv preprint arXiv:1808.03887*, 2018.

[40] Dong Nie, Yaozong Gao, Li Wang, and Dinggang Shen. Asdnet: Attention based semi-supervised deep networks for medical image segmentation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, 2018.

[41] Yicheng Wu, Yong Xia, Yang Song, Yanning Zhang, and Weidong Cai. Multiscale network followed network model for retinal vessel segmentation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, 2018.

[42] Lequan Yu, Shujun Wang, Xiaomeng Li, Chi-Wing Fu, and Pheng-Ann Heng. Uncertainty-aware self-ensembling model for semi-supervised 3d left atrium segmentation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, 2019.

[43] Yu Zeng, Yunzhi Zhuge, Huchuan Lu, and Lihe Zhang. Joint learning of saliency detection and weakly supervised semantic segmentation. In *IEEE International Conference on Computer Vision (ICCV)*, 2019.

[44] Gerda Bortsova, Florian Dubost, Laurens Hogeweg, Ioannis Katramados, and Marleen de Bruijne. Semi-supervised medical image segmentation via learning consistency under transformations. In *International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, 2019.

[45] Yicheng Wu, Yong Xia, Yang Song, Donghao Zhang, Dongnan Liu, Chaoyi Zhang, and Weidong Cai. Vessel-net: retinal vessel segmentation under multi-path supervision. In *International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, 2019.

[46] Chenyu You, Junlin Yang, Julius Chapiro, and James S. Duncan. Unsupervised wasserstein distance guided domain adaptation for 3d multi-domain liver segmentation. In *Interpretable and Annotation-Efficient Learning for Medical Image Computing*, pages 155–163. Springer International Publishing, 2020.

[47] Linfeng Yang, Rajarshi P Ghosh, J Matthew Franklin, Simon Chen, Chenyu You, Raja R Narayan, Marc L Melcher, and Jan T Liphardt. Nuset: A deep learning tool for reliably separating and analyzing crowded cells. *PLoS computational biology*, 2020.

[48] Chenyu You, Ruihan Zhao, Lawrence Staib, and James S Duncan. Momentum contrastive voxel-wise representation learning for semi-supervised volumetric medical image segmentation. *arXiv preprint arXiv:2105.07059*, 2021.

[49] Chenyu You, Yuan Zhou, Ruihan Zhao, Lawrence Staib, and James S Duncan. Simcvd: Simple contrastive voxel-wise representation distillation for semi-supervised medical image segmentation. *IEEE Transactions on Medical Imaging*, 2022.

[50] Krishna Chaitanya, Ertunc Erdil, Neerav Karani, and Ender Konukoglu. Contrastive learning of global and local features for medical image segmentation with limited annotations. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2020.

[51] Yuexing Han, Xiaolong Li, Bing Wang, and Lu Wang. Boundary loss-based 2.5 d fully convolutional neural networks approach for segmentation: A case study of the liver and tumor on computed tomography. *Algorithms*, 2021.

[52] Konstantinos Kamnitsas, Christian Ledig, Virginia FJ Newcombe, Joanna P Simpson, Andrew D Kane, David K Menon, Daniel Rueckert, and Ben Glocker. Efficient multi-scale 3d cnn with fully connected crf for accurate brain lesion segmentation. *Medical Image Analysis*, 2017.

[53] John Lafferty, Andrew McCallum, and Fernando CN Pereira. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In *International Conference on Machine Learning (ICML)*, 2001.

[54] Ali Hatamizadeh, Yucheng Tang, Vishwesh Nath, Dong Yang, Andriy Myronenko, Bennett Landman, Holger Roth, and Daguang Xu. Unetr: Transformers for 3d medical image segmentation. *arXiv preprint arXiv:2103.10504*, 2021.

[55] Fabian Isensee, Paul F Jaeger, Simon AA Kohl, Jens Petersen, and Klaus H Maier-Hein. nnu-net: a self-configuring method for deep learning-based biomedical image segmentation. *Nature methods*, 2021.

[56] Sixiao Zheng, Jiachen Lu, Hengshuang Zhao, Xiatian Zhu, Zekun Luo, Yabiao Wang, Yanwei Fu, Jianfeng Feng, Tao Xiang, Philip HS Torr, et al. Rethinking semantic segmentation from a sequence-to-sequence perspective with transformers. *arXiv preprint arXiv:2012.15840*, 2020.

[57] Georg Hille, Shubham Agrawal, Christian Wybranski, Maciej Pech, Alexey Surov, and Sylvia Saalfeld. Joint liver and hepatic lesion segmentation using a hybrid cnn with transformer layers. *arXiv preprint arXiv:2201.10981*, 2022.

[58] 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 *IEEE Winter Conference on Applications of Computer Vision (WACV)*, 2022.

[59] Yucheng Tang, Dong Yang, Wenqi Li, Holger Roth, Bennett Landman, Daguang Xu, Vishwesh Nath, and Ali Hatamizadeh. Self-supervised pre-training of swin transformers for 3d medical image analysis. *arXiv preprint arXiv:2111.14791*, 2021.

[60] Yifan Jiang, Shiyu Chang, and Zhangyang Wang. Transgan: Two transformers can make one strong gan. *arXiv preprint arXiv:2102.07074*, 2021.

[61] Yu Zeng, Zhe Lin, and Vishal M Patel. Sketchedit: Mask-free local image manipulation with partial sketches. *arXiv preprint arXiv:2111.15078*, 2021.

[62] Yu Zeng, Zhe Lin, Huchuan Lu, and Vishal M Patel. Cr-fill: Generative image inpainting with auxiliary contextual reconstruction. In *IEEE International Conference on Computer Vision (ICCV)*, 2021.

[63] Dor Arad Hudson and Larry Zitnick. Compositional transformers for scene generation. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2021.

[64] Long Zhao, Zizhao Zhang, Ting Chen, Dimitris N Metaxas, and Han Zhang. Improved transformer for high-resolution gans. *arXiv preprint arXiv:2106.07631*, 2021.

[65] Yanhong Zeng, Huan Yang, Hongyang Chao, Jianbo Wang, and Jianlong Fu. Improving visual quality of image synthesis by a token-based generator with transformers. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2021.

[66] Zihang Dai, Hanxiao Liu, Quoc V Le, and Mingxing Tan. Coatnet: Marrying convolution and attention for all data sizes. *arXiv preprint arXiv:2106.04803*, 2021.

[67] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask r-cnn. In *Proceedings of the IEEE international conference on computer vision*, pages 2961–2969, 2017.

[68] Liang-Chieh Chen, Yukun Zhu, George Papandreou, Florian Schroff, and Hartwig Adam. Encoder-decoder with atrous separable convolution for semantic image segmentation. In *European Conference on Computer Vision (ECCV)*, 2018.

[69] Chao Peng, Xiangyu Zhang, Gang Yu, Guiming Luo, and Jian Sun. Large kernel matters–improve semantic segmentation by global convolutional network. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017.

[70] Fisher Yu and Vladlen Koltun. Multi-scale context aggregation by dilated convolutions. *arXiv preprint arXiv:1511.07122*, 2015.

[71] Tete Xiao, Piotr Dollar, Mannat Singh, Eric Mintun, Trevor Darrell, and Ross Girshick. Early convolutions help transformers see better. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2021.

[72] Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. Focal loss for dense object detection. In *IEEE International Conference on Computer Vision (ICCV)*, 2017.

[73] Tsung-Yi Lin, Piotr Dollár, Ross Girshick, Kaiming He, Bharath Hariharan, and Serge Belongie. Feature pyramid networks for object detection. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017.

[74] Enze Xie, Wenhai Wang, Zhiding Yu, Anima Anandkumar, Jose M Alvarez, and Ping Luo. Segformer: Simple and efficient design for semantic segmentation with transformers. *arXiv preprint arXiv:2105.15203*, 2021.

[75] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2014.

[76] Martin Arjovsky, Soumith Chintala, and Léon Bottou. Wasserstein generative adversarial networks. In *International Conference on Machine Learning (ICML)*, 2017.

[77] Ishaan Gulrajani, Faruk Ahmed, Martin Arjovsky, Vincent Dumoulin, and Aaron Courville. Improved training of wasserstein gans. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2017.

[78] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *International Conference on Learning Representations (ICLR)*, 2019.

[79] Jo Schlemper, Ozan Oktay, Michiel Schaap, Mattias Heinrich, Bernhard Kainz, Ben Glocker, and Daniel Rueckert. Attention gated networks: Learning to leverage salient regions in medical images. *Medical Image Analysis*, 2019.

[80] Sixiao Zheng, Jiachen Lu, Hengshuang Zhao, Xiatian Zhu, Zekun Luo, Yabiao Wang, Yanwei Fu, Jianfeng Feng, Tao Xiang, Philip HS Torr, et al. Rethinking semantic segmentation from a sequence-to-sequence perspective with transformers. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021.

[81] Ekin D Cubuk, Barret Zoph, Jonathon Shlens, and Quoc V Le. Randaugment: Practical automated data augmentation with a reduced search space. In *CVPR Workshops*, 2020.

[82] Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jonathon Shlens, and Zbigniew Wojna. Rethinking the inception architecture for computer vision. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.

[83] Hongyi Zhang, Moustapha Cisse, Yann N Dauphin, and David Lopez-Paz. mixup: Beyond empirical risk minimization. In *International Conference on Learning Representations (ICLR)*, 2018.

[84] Sangdoo Yun, Dongyoon Han, Seong Joon Oh, Sanghyuk Chun, Junsuk Choe, and Youngjoon Yoo. Cutmix: Regularization strategy to train strong classifiers with localizable features. In *IEEE International Conference on Computer Vision (ICCV)*, 2019.

[85] Boris T Polyak and Anatoli B Juditsky. Acceleration of stochastic approximation by averaging. *SIAM Journal on Control and Optimization*, 1992.

[86] Mario Lucic, Karol Kurach, Marcin Michalski, Sylvain Gelly, and Olivier Bousquet. Are gans created equal? a large-scale study. *arXiv preprint arXiv:1711.10337*, 2017.

[87] Xudong Mao, Qing Li, Haoran Xie, Raymond YK Lau, Zhen Wang, and Stephen Paul Smolley. Least squares generative adversarial networks. In *IEEE International Conference on Computer Vision (ICCV)*, 2017.

[88] Jifeng Dai, Haozhi Qi, Yuwen Xiong, Yi Li, Guodong Zhang, Han Hu, and Yichen Wei. Deformable convolutional networks. In *IEEE International Conference on Computer Vision (ICCV)*, 2017.

[89] 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.

[90] Zhuofan Xia, Xuran Pan, Shiji Song, Li Erran Li, and Gao Huang. Vision transformer with deformable attention. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4794–4803, 2022.

[91] Alex Zwanenburg, Martin Vallières, Mahmoud A Abdalah, Hugo JWL Aerts, Vincent Andrearczyk, Aditya Apte, Saeed Ashrafinia, Spyridon Bakas, Roelof J Beukinga, Ronald Boellaard, et al. The image biomarker standardization initiative: standardized quantitative radiomics for high-throughput image-based phenotyping. *Radiology*, 295(2):328–338, 2020.

[92] Joost JM Van Griethuysen, Andriy Fedorov, Chintan Parmar, Ahmed Hosny, Nicole Aucoin, Vivek Narayan, Regina GH Beets-Tan, Jean-Christophe Fillion-Robin, Steve Pieper, and Hugo JWL Aerts. Computational radiomics system to decode the radiographic phenotype. *Cancer research*, 77(21):e104–e107, 2017.

[93] Ahmet Saygılı. A new approach for computer-aided detection of coronavirus (covid-19) from ct and x-ray images using machine learning methods. *Applied Soft Computing*, 105:107323, 2021.

[94] Feng Shi, Liming Xia, Fei Shan, Bin Song, Dijia Wu, Ying Wei, Huan Yuan, Huiting Jiang, Yichu He, Yaozong Gao, et al. Large-scale screening to distinguish between covid-19 and community-acquired pneumonia using infection size-aware classification. *Physics in medicine & Biology*, 66(6):065031, 2021.

Checklist

1. For all authors...
   (a) Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope? [Yes]
   (b) Did you describe the limitations of your work? [Yes] See Section 7, and Broader Impact on Page 9.
   (c) Did you discuss any potential negative societal impacts of your work? [Yes] See Section 7, and Broader Impact on Page 9.
   (d) Have you read the ethics review guidelines and ensured that your paper conforms to them? [Yes]
2. If you are including theoretical results...
   (a) Did you state the full set of assumptions of all theoretical results? [N/A]
   (b) Did you include complete proofs of all theoretical results? [N/A]
3. If you ran experiments...
   (a) Did you include the code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL)? [Yes] See Sections 4 and supplemental material.
   (b) Did you specify all the training details (e.g., data splits, hyperparameters, how they were chosen)? [Yes] See Sections 4 and supplemental material.
   (c) Did you report error bars (e.g., with respect to the random seed after running experiments multiple times)? [Yes] See Sections 4 and supplemental material.
   (d) Did you include the total amount of compute and the type of resources used (e.g., type of GPUs, internal cluster, or cloud provider)? [Yes] See Sections 4 and supplemental material.
4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets...
   (a) If your work uses existing assets, did you cite the creators? [Yes]
   (b) Did you mention the license of the assets? [Yes]
   (c) Did you include any new assets either in the supplemental material or as a URL? [No]
   (d) Did you discuss whether and how consent was obtained from people whose data you're using/curating? [Yes] See Sections 4 and supplemental material.
   (e) Did you discuss whether the data you are using/curating contains personally identifiable information or offensive content? [Yes] See Sections 4 and supplemental material.
5. If you used crowdsourcing or conducted research with human subjects...
   (a) Did you include the full text of instructions given to participants and screenshots, if applicable? [N/A]
   (b) Did you describe any potential participant risks, with links to Institutional Review Board (IRB) approvals, if applicable? [N/A]
   (c) Did you include the estimated hourly wage paid to participants and the total amount spent on participant compensation? [N/A]

Appendix to
Class-Aware Adversarial Transformers for Medical Image Segmentation

Sec. A provides additional details on the training datasets.

Sec. B provides additional details on the implementation.

Sec. C provides additional model details.

Sec. D provides more quantitative results on the LiTS dataset.

Sec. E provides more experimental results on the MP-MRI dataset.

Sec. F provides additional analysis on Iteration Number N.

Sec. G provides additional analysis on sampling number n.

Sec. H provides additional details on the hyperparameter selection.

Sec. I provides additional GAN-based loss details.

Sec. J takes a deeper look and understand our proposed class-aware transformer module.

Sec. K forms a better understand the attention mechanism in our proposed `CASTformer`.

A Datasets

Synapse: Synapse multi-organ segmentation dataset includes 30 abdominal CT scans with 3779 axial contrast-enhanced abdominal clinical CT images. Each CT volume consists of 85 $\sim$ 198 slices of 512$\times$512 pixels, with a voxel spatial resolution of $([0.54 \sim 0.54] \times [0.98 \sim 0.98] \times [2.5 \sim 5.0])\text{mm}^3$. The dataset is randomly divided into 18 volumes for training (2212 axial slices), and 12 for validation. For each case, 8 anatomical structures are aorta, gallbladder, spleen, left kidney, right kidney, liver, pancreas, spleen, stomach.

LiTS: MICCAI 2017 Liver Tumor Segmentation Challenge (LiTS) includes 131 contrast-enhanced 3D abdominal CT volumes for training and testing. The dataset is assembled by different scanners and protocols from seven hospitals and research institutions. The image resolution ranges from 0.56mm to 1.0mm in axial and 0.45mm to 6.0mm in z direction. The dataset is randomly divided into 100 volumes for training, and 31 for testing.

MP-MRI: Multi-phasic MRI dataset is an in-house dataset including multi-phasic MRI scans of 20 local patients with HCC, each of which consisted of T1 weighted DCE-MRI images at three-time points (pre-contrast, arterial phase, and venous phases). Three images are mutually registered to the arterial phase images, with an isotropic voxel size of 1.00 mm. The dataset is randomly divided into 48 volumes for training, and 12 for testing.

B More Implementation Details

The training configuration and hyperparameter settings are summarized in Table 4.

C Model Architecture

We present the detailed architecture of `CATformer`'s encoding pipeline in Section 5. We use input/output names to indicate the direction of the data stream. `CATformer` applies independent class-aware attention on 4 levels of features extracted by the `ResNetV2` model. Each feature level L-k is processed by `CATformer`-k, consisting of 4 blocks of class-aware transformer modules, followed by 12 layers of transformer encoder modules. Outputs from all four feature levels are fed into the decoder pipeline to generate the segmentation masks.

Table 4: Training configuration and hyperparameter settings.

Training Config	Hyperparameter
Optimizer	AdamW
Base learning rate	5e-4
Weight decay	0.05
Optimizer momentum	β_1, β_2=0.9, 0.999
Batch size	6
Training epochs	300
Learning rate schedule	cosine decay
Warmup epochs	5
Warmup schedule	linear
Randaugment [81]	(9, 0.5)
Label smoothing [82]	0.1
Mixup [83]	0.8
Cutmix [84]	1.0
Gradient clip	None
Exp. mov. avg. (EMA) [85]	None

Table 5: Architecture configuration of `CATformer`

`CATformer`					
Stage	Layer	Input Name	Input Shape	Output Name	Output Shape
`Encoder`	`ResNetV2`	Original Image	$224 \times 224 \times 3$	RN-L1 RN-L2 RN-L3 RN-L4	$112 \times 112 \times 64$ $56 \times 56 \times 256$ $28 \times 28 \times 512$ $14 \times 14 \times 1024$
`CATformer-1`	CAT$\times$4 TEM$\times$12	RN-L1 CAT-1	$112 \times 112 \times 64$ $(28 \times 28) \times 64$	CAT-1 F1	$(28 \times 28) \times 64$ $(28 \times 28) \times 64$
`CATformer-2`	CAT$\times$4 TEM$\times$12	RN-L2 CAT-2	$56 \times 56 \times 256$ $(28 \times 28) \times 256$	CAT-2 F2	$(28 \times 28) \times 256$ $(28 \times 28) \times 256$
`CATformer-3`	CAT$\times$4 TEM$\times$12	RN-L3 CAT-3	$28 \times 28 \times 512$ $(28 \times 28) \times 512$	CAT-3 F3	$(28 \times 28) \times 512$ $(28 \times 28) \times 512$
`CATformer-4`	CAT$\times$4 TEM$\times$12	RN-L4 CAT-4	$14 \times 14 \times 768$ $(14 \times 14) \times 768$	CAT-4 F4	$(14 \times 14) \times 768$ $(14 \times 14) \times 768$

D More Experiments: LiTS

Experimental results are summarized in Table 6.

Table 6: Quantitative segmentation results on the LiTS dataset.

Framework		Average				Liver	Tumor
Encoder	Decoder	DSC ↑	Jaccard ↑	95HD ↓	ASD ↓		
`UNet` [21]		62.88	54.64	57.59	27.74	88.27	37.49
`AttnUNet` [79]		66.03	58.49	31.34	16.15	92.26	39.81
`ResNet50`	`UNet` [21]	65.25	58.09	27.97	10.02	93.78	36.73
`ResNet50`	`AttnUNet` [79]	66.22	59.27	31.47	10.41	93.26	39.18
`SETR` [80]		54.79	49.21	36.34	15.04	91.69	17.90
`CoTr` w/o `CNN-encoder` [13]		53.35	47.11	55.82	22.99	85.25	21.45
`CoTr` [13]		62.67	55.43	34.75	15.84	89.43	35.92
`TransUNet` [7]		67.94	60.25	29.32	12.45	93.40	42.49
`SwinUNet` [10]		65.53	57.84	36.45	16.52	92.15	38.92
• `CATformer` (ours)		72.39	62.76	**22.38**	11.57	94.18	49.60
○ `CASTformer` (ours)		**73.82**	**64.91**	23.35	**10.16**	95.88	51.76

E More Experiments: MP-MRI

Experimental results are summarized in Table 7. Overall, `CATformer` and `CASTformer` outperform the previous results in terms of Dice and Jaccard. Compared to SETR, our `CATformer` and

Table 7: Quantitative segmentation results on the MP-MRI dataset.

Framework		Average			
Encoder	Decoder	DSC ↑	Jaccard ↑	95HD ↓	ASD ↓
UNet [21]		88.38	79.42	39.23	11.14
AttnUNet [79]		89.79	81.51	30.13	7.85
ResNet50	UNet [21]	91.51	84.39	15.38	4.53
ResNet50	AttnUNet [79]	91.43	84.24	14.14	4.24
SETR [80]		92.39	85.89	7.66	3.79
CoTr w/o CNN-encoder [13]		85.21	74.49	44.25	12.58
CoTr [13]		90.06	81.94	28.91	7.89
TransUNet [7]		92.08	85.36	23.17	6.03
SwinUNet [10]		92.07	85.32	7.62	3.88
• CATformer (ours)		94.17	86.50	**6.55**	3.33
○ CASTformer (ours)		**94.93**	**87.81**	8.29	**3.02**

CASTformer perform 1.78% and 2.54% higher in Dice, respectively. We also find CASTformer performs better than CATformer, which suggests that using discriminator can make the model better assess the medical image fidelity. Figure 5 shows qualitative results, where our CATformer and CASTformer provide better anatomical details than all other methods. This clearly demonstrates the superiority of our models. All these experiments are conducted using the same hyperparameters in our CASTformer.

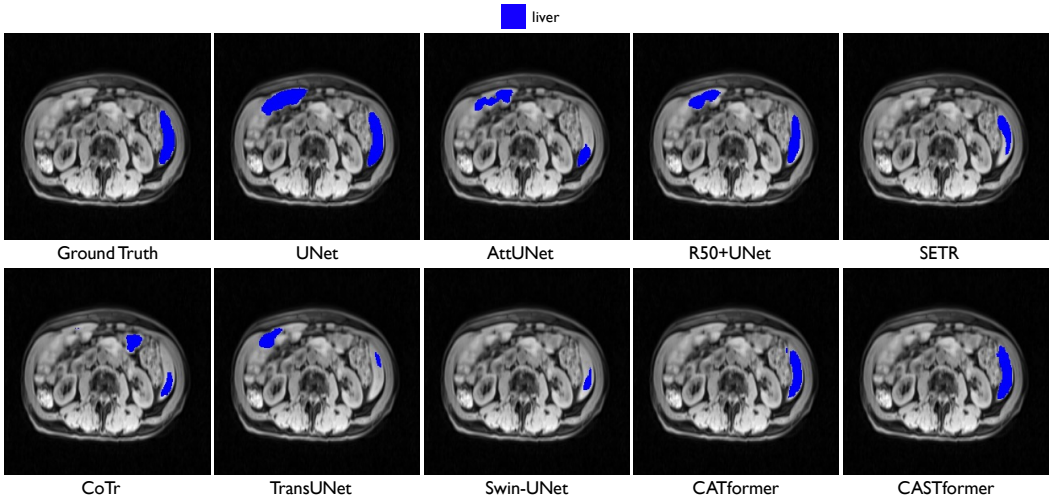

Figure 5: Visual comparisons with other methods on MP-MRI dataset.

F Effect of Iteration Number N

We explore the effect of different iteration number N in Figure 6 (a). Note that in the case of $N = 1$, the sampling locations will not be updated. We find that more iterations of sampling clearly improve network performance in Dice and Jaccard. However, we observe that the network performance does not further increase from $N = 4$ to $N = 6$. In our study, we use $N = 4$ for the class-aware transformer module.

G Effect of Sampling Number n

We further evaluate the effect of sampling number n of the class-aware transformer module in Figure 6 (b). Empirically, we observe that results are generally well correlated when we gradually increase the size of n. As is shown, the network performance is optimal when $n = 16$.

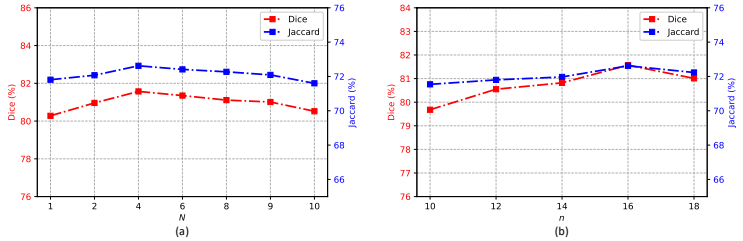

Figure 6: Effects of the iteration number N and the sampling number n in the class-aware transformer module. We report Dice and Jarrcd of `CATformer` on the Synapse multi-organ dataset.

H Hyperparameter Selection

We carry out grid-search of $\lambda_1, \lambda_2, \lambda_3 \in \{0.0, 0.1, 0.2, 0.5, 1.0\}$. As shown in Figure 7, with a carefully tuned hyperparameters $\lambda_1 = 0.5$, $\lambda_2 = 0.5$, and $\lambda_3 = 0.1$, such setting performs generally better than others.

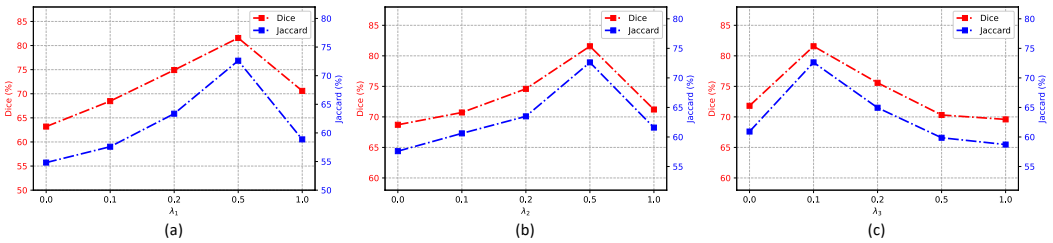

Figure 7: Effects of hyperparameters $\lambda_1, \lambda_2, \lambda_3$. We report Dice and Jarrcd of `CASTformer` on the Synapse multi-organ dataset.

I Importance of Loss Functions

One main argument for the discriminator is that modeling long-range dependencies and acquiring a more holistic understanding of the anatomical visual information can contribute to the improved capability of the generator. Besides the WGAN-GP loss [77], the minimax (MM) GAN loss [75], the Non-Saturating (NS) GAN loss [86], and Least Squares (LS) GAN Loss [87] are also commonly used as adversarial training. We test these alternatives and find that, in most cases, using WGAN-GP loss achieves comparable or higher performance than other loss functions. In addition, models trained using MM-GAN loss perform comparably to those trained using LS-GAN loss. In particular, our approach outperforms the second-best LS-GAN loss [87] by 1.10 and 2.49 points in Dice and Jaccard scores on the Synapse multi-organ dataset. It demonstrates the effectiveness of the WGAN-GP loss in our `CASTformer`.

Table 8: Ablation on Loss Function: MM-GAN loss [75]; NS-GAN loss [86]; LS-GAN loss [87]; and WGAN-GP loss [77].

Model	DSC	Jaccard	95HD	ASD
MM-GAN loss [75]	81.19	71.76	20.75	5.90
NS-GAN loss [86]	80.02	70.47	26.06	6.96
LS-GAN loss [87]	81.45	72.20	20.39	6.49
WGAN-GP loss [77]	82.55	74.69	22.73	5.81

J Visualization of Learned Sampling Location

To gain more insight into the evolving sampling locations learned by our proposed class-aware transformer module, we visualize the predicted offsets in Figure 8. We can see that particular

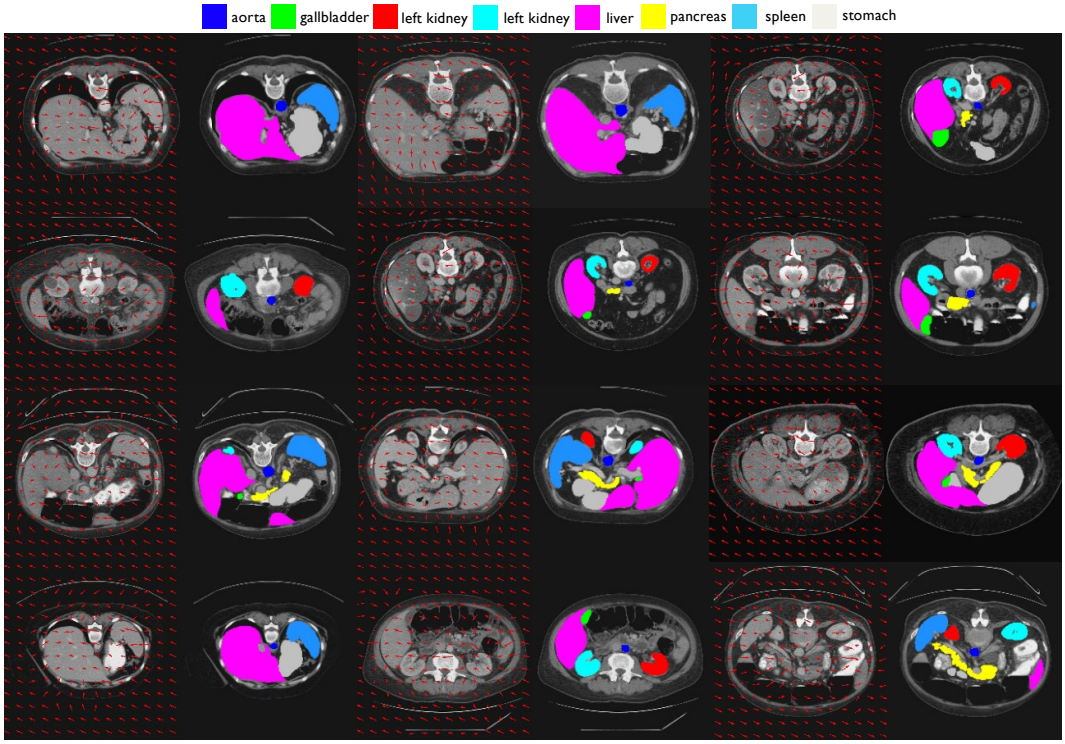

Figure 8: Visualization of sampled locations in the proposed class-aware transformer module.

sampling points around objects tend to attend to coherent segmented regions in terms of anatomical similarity and proximity. As is shown, we show the classes with the highly semantically correlated regions, indicating that the model coherently attends to anatomical concepts such as liver, right/left kidney, and spleen. These visualizations also illustrate how it behaves adaptively and distinctively to focus on the content with highly semantically correlated discriminative regions (*i.e.*, different organs). These findings can thereby suggest that our design can aid the `CATformer` to exercise finer control emphasizing anatomical features with the intrinsic structure at the object granularity. As is indicated (Figure 8 last column), we also find evidence that our model is prone to capture some small object cases (*e.g.*, pancreas, aorta, gallbladder). We hypothesize that it is because they contain more anatomical variances, which makes the model more difficult to exploit.

K Vision Transformer Visualization

In this section, we visualize the first 12 class-aware transformer layers on sequences of 28×28 feature patches in the encoder pipeline. In Figure 9, we plot the attention probabilities from a single patch over different layers and heads. Each row corresponds to one CAT layer; each column corresponds to an attention head. As we go deeper into the network, we are able to observe three kinds of attention behaviors as further discussed below.

Attend to similar features: In the first group of layers (layer 1 through 4), the attention probability is spread across a relatively large group of patches. Notably, these patches correspond to areas in the image with similar color and texture to the query patch. These more primitive attention distributions indicate that the class-awareness property has not yet been established.

Attend to the same class and its boundary: In the middle layers of the transformer model, most noticeable in the 5th and 6th layers, the attention probabilities start to concentrate on areas that share the same class label as the query patch (layer 5-2). In some other instances, the model attends to the boundary of the current class (layer 5-3, 5-6).

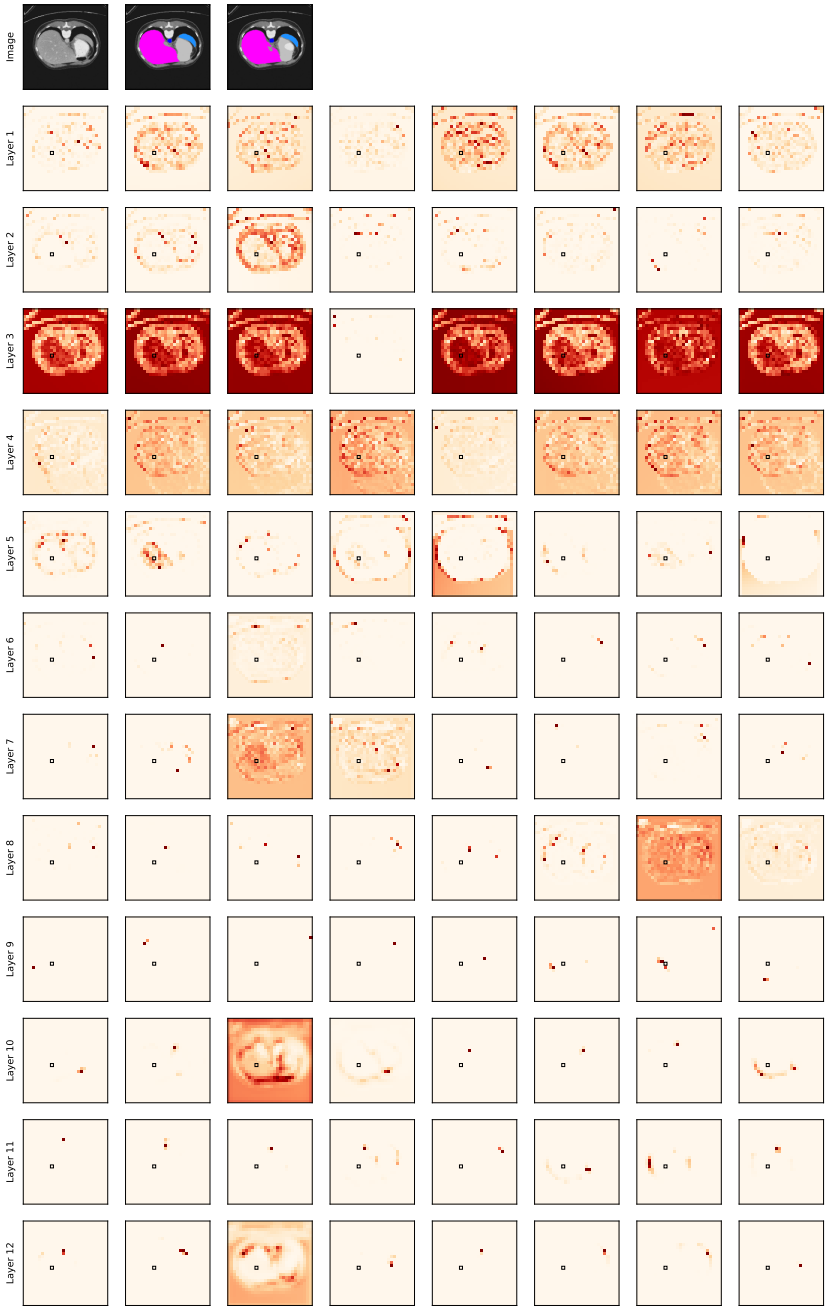

Figure 9: Attention probability of our 12 class-aware transformer layers, each with 8 heads. The black box marks the query patch. The input image, ground truth and predicted label are shown on the first row.

Attend to other classes: In the deeper layers of the model, the attention probability mainly concentrates on other classes. This clearly demonstrates persuasive evidence that the model establishes class awareness, which is helpful in the downstream medical segmentation tasks.

L More Ablations on Decoder Modules

In this section, we explore another state-of-the-art backbone proposed by Lin *et al.* [73], termed Feature Pyramid Network (FPN). FPN utilizes a top-down pyramid with lateral connections to construct

the semantically strong multi-scale feature pyramid from a single-scale input. The major differences between FPN and our work are as follows:

- The former utilizes a CNN-based decoder (FPN [73]), and ours uses an All-MLP-based decoder. In particular, our motivation comes from the observation that the attention of lower layers tends to be local, and those of the higher layers are highly non-local [74]. As the decoder design plays an important role in determining the semantic level of the latent representations [9] and Transformers have the larger receptive fields compared to CNNs, how to use large receptive fields to include context information is the key issue [74, 70, 69, 68]. Prior work [74] suggests that the use of MLP-based decoder design can be a very effective tool in learning additional contextual information to build powerful representations. The key idea is to essentially take benefits of the Transformer-induced features by leveraging the local attention at the lower layers and highly non-local (global) attention at the higher layers to formulate the powerful representations [74]. To this end, we utilize an MLP-based decoder instead of a CNN-based decoder to preserve more contextual information, specifically for medical imaging data, including more anatomical variances.

- We devise the class-aware transformer module to progressively learn interesting anatomical regions correlated with semantic structures of images, so as to guide the segmentation of objects or entities. We study the model's qualitative behavior through learnable sampling locations inside the class-aware module in Figure 8. As indicated, sampling locations are adaptively adjusted according to the interesting regions.

The table below shows the comparision results of using an FPN decoder, MLP-based decoder, and the class-aware transformer (CAT) module, all of which include the backbone feature extractor (ResNet50), on the Synapse multi-organ CT dataset. All the experiments are conducted under the same experimental setting in Section 4. As we can see, adopting the MLP-based decoder can outperform the state-of-the-art FPN decoder in terms of DSC, Jaccard, 95HD, and ASD, respectively. Similarly, incorporating the CAT module can also consistently improve the segmentation performance by a large margin on the Synapse multi-organ CT dataset. The results prove the robustness of our MLP-based decoder and the effectiveness of our proposed CAT module for medical image segmentation.

Table 9: Ablation on Decoder Modules: FPN decoder [73]; MLP-based decoder; and Class-Aware Transformer (CAT) module.

Encoder	Decoder	DSC	Jaccard	95HD	ASD
ResNet50 w/o CAT	FPN	74.64	63.91	29.54	8.81
ResNet50 w/ CAT	FPN	78.11	65.63	28.06	8.08
ResNet50 w/o CAT	MLP	80.09	70.56	25.62	7.30
ResNet50 w/ CAT	MLP	82.17	73.22	16.20	4.28

M More Ablations on Segmentation Losses

To deal with the imbalanced medical image segmentation, Lin *et al.* [72] proposed Focal loss in terms of the standard cross entropy to address the extreme foreground-background class imbalance by focusing on the hard pixel examples. The table below shows the results of the loss function. We follow $\gamma = 2$ in the original paper. As we can see, the setting using Focal loss and the other (*i.e.*, Dice + Cross-Entropy) achieve similar performances.

Table 10: Ablation on Segmentation Losses: Focal loss [72]; Dice loss; and Cross-Entropy loss.

Model	DSC	Jaccard	95HD	ASD
Focal loss [72]	82.08	73.52	16.14	4.99
Dice + Focal loss [72]	81.88	72.94	16.52	5.00
Dice + Cross-Entropy loss (ours)	82.17	73.22	16.20	4.28

N More Ablations on Sampling Modules

In this section, we investigate the effect of recent state-of-the-art sampling modules [88–90]. However, the motivation and the sampling strategy are different from these works [88–90]. Our motivation comes from the accurate and reliable clinical diagnosis that rely on the meaningful radiomic features from the correct "region of interest" instead of other irrelevant parts [91–94]. The process of extracting different radiomic features from medical images is done in a progressive and adaptive manner [92, 93].

`DCN` [88] proposed to learn 2D spatial offsets to enable the CNN-based model to generalize the capability of regular convolutions. Because CNNs only have limited receptive fields compared to Transformers, `DCN` focuses on local information around a certain point of interest. In contrast, our `CATformer`/`CASTformer` take benefits of the Transformer-induced features by leveraging the local attention at the lower layers and highly non-local (global) attention at the higher layers to formulate the powerful representations.

`Deformable DETR` [89] incorporated the deformation attention to focus on a sparse set of keys (*i.e.*, global keys are not shared among visual tokens). This is particularly useful for its original experiment setup on object detection. Since there are only a handful of query features corresponding to potential object classes, deformable DETR learns different attention locations for each class. In contrast, our approach aims at refining the anatomical tokens for medical image segmentation. To this end, we proposed to iteratively and adaptively focus on the most discriminative region of interests. This essentially allows us to obtain effective anatomical features from spatial attended regions within the medical images, so as to guide the segmentation of objects or entities

`DAT` [90] introduced deformable attention to make use of global information (*i.e.*, global keys are shared among visual tokens) by placing a set of the supporting points uniformly on the feature maps. In contrast, our approach introduces an iterative and progressive sampling strategy to capture the most discriminative region and avoid over-partition anatomical features.

The table below shows the comparison results between `DCN` [88], `Deformable DETR` [89], `DAT` [90], and ours (`CATformer`/`CASTformer`) on the Synapse multi-organ CT dataset. As we can see, our approach (*i.e.*, `CATformer`/`CASTformer`) can outperform existing state-of-the-art models, *i.e.*,`DCN` [88], and `Deformable DETR`.

Table 11: Ablation on Sampling Module: `DCN` [88], `Deformable DETR` [89], `DAT` [90], and ours (`CATformer`/`CASTformer`).

Model	DSC	Jaccard	95HD	ASD
`DCN` [88]	73.19	62.81	33.46	10.22
`Deformable DETR` [89]	79.13	66.58	30.21	8.65
`DAT` [90]	80.34	68.15	26.14	7.76
`CATformer` (ours)	82.17	73.22	16.20	4.28
`CASTformer` (ours)	82.55	74.69	22.73	5.81

O More Ablations on Architecture Baselines

In this section, we conduct the ablation study on the Synapse multi-organ CT dataset to compare our approach with the recent state-of-the-art architecture (`SwinUnet`) [10]. The table below shows the results of our proposed architecture (*e.g.*, Swin-class-aware transformer (Swin-CAT) module, multi-scale feature extraction module) are superior compared to the other state-of-the-art method on the Synapse multi-organ CT dataset. All the experiments are conducted under the same experimental setting in Section 4. For brevity, we refer our `CATformer` and `CASTformer` using `SwinUnet` as the baseline model to `Swin-CATformer` and `Swin-CASTformer`. As we can see, using `SwinUnet` as the baseline, the following observations can be drawn: (1) "w/ pre-trained" consistently achieves significant performance gains compared to the "w/o pre-trained", which demonstrates the effectiveness of the pre-training strategy; (2) we can find that incorporating the adversarial training can boost the segmentation performance, which suggests the effectiveness of the adversarial training strategy;

and (3) our Swin-CASTformer with different modules can also achieves consistently improved performance. The results prove the superiority of our proposed method on the medical image segmentation task.

Table 12: Effect of transfer learning in our Swin-CATformer and Swin-CASTformer on the Synapse multi-organ dataset.

Model	DSC	Jaccard	95HD	ASD
• Swin-CATformer (w/o pre-trained)	76.82	65.44	29.58	8.58
• Swin-CATformer (w/ pre-trained)	80.19	70.61	22.66	6.02
○ Swin-CASTformer (*both* w/o pre-trained)	71.67	61.08	43.01	13.21
○ Swin-CASTformer (*only* w/ pre-trained D)	76.55	64.27	34.62	12.13
○ Swin-CASTformer (*only* w/ pre-trained G)	77.12	65.39	30.99	11.00
○ Swin-CASTformer (*both* w/ pre-trained)	80.49	71.19	23.94	6.91

Table 13: Ablation on model component: Baseline; Swin-CATformer w/o Swin-CAT; Swin-CATformer w/o multi-scale feature extraction; and Swin-CASTformer.

Model	DSC	Jaccard	95HD	ASD
Baseline	76.33	65.64	27.16	8.32
• Swin-CATformer w/o Swin-CAT	77.76	68.47	25.26	7.15
• Swin-CATformer w/o multi-scale feature extraction	78.45	78.26	24.94	7.08
• Swin-CATformer	80.19	70.61	22.66	6.02
○ Swin-CASTformer	80.49	71.19	23.94	6.91