This ICCV workshop paper is the Open Access version, provided by the Computer Vision Foundation. Except for this watermark, it is identical to the accepted version; the final published version of the proceedings is available on IEEE Xplore.

Transformers Pay Attention to Convolutions Leveraging Emerging Properties of ViTs by Dual Attention-Image Network

Yousef Yeganeh[1,3] * Azade Farshad[1,3] * Peter Weinberger[1] * Seyed-Ahmad Ahmadi[4]
Ehsan Adeli[5] Nassir Navab[1,2]
[1]Technical University of Munich [2]Johns Hopkins University [3]MCML [4]NVIDIA
[5]Stanford University

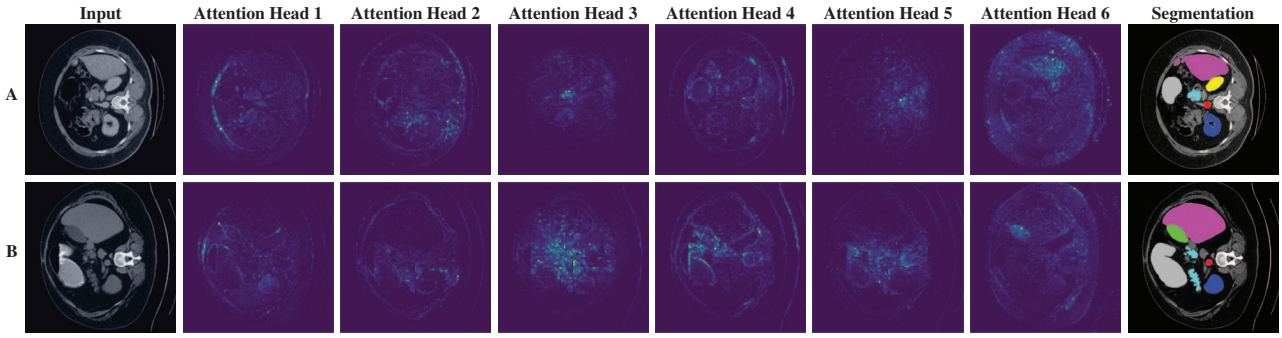

Figure 1. **Attention Map Visualization.** Given an image, we generate the attention map visualizations using a pretrained DINO [11]. The attention maps bear observable resemblance to the ground truth segmentation maps. Inpired by this phenomenon, we innovatively pass these segmentation map visualizations along with the input image to the segmentation network to generate the semantic segmentation prediction.

Abstract

Although purely transformer-based architectures pretrained on large datasets are introduced as foundation models for general computer vision tasks, hybrid models that incorporate combinations of convolution and transformer blocks showed state-of-the-art performance in more specialized tasks. Nevertheless, despite the performance gain of both pure and hybrid transformer-based architectures compared to convolutional networks, their high training cost and complexity make it challenging to use them in real scenarios. In this work, we propose a novel and simple architecture based on only convolutional layers and show that by just taking advantage of the attention map visualizations obtained from a self-supervised pretrained vision transformer network, complex transformer-based networks, and even 3D architectures are outperformed with much fewer computation costs. The proposed architecture is composed of two encoder branches with the original image as input in one branch and the attention map visualizations of the same image from multiple self-attention heads from a pre-trained DINO model in the other branch. The results of our experiments on medical imaging datasets show that the extracted attention map visualizations from the attention heads of a pre-trained transformer architecture combined with the image provide strong prior knowledge for a pure CNN architecture to outperform CNN-based and transformer-based architectures. Project Page: dai-net.github.io

1. Introduction

Medical image segmentation aims to highlight critical parts of an image, such as organs and tumors, in various modalities (e.g., CT, MR, etc), for the purpose of clinical diagnosis. The cost and availability of annotation and segmentation by experts necessitates automatic methods, and deep-learning based techniques have given a significant boost to the field in recent years. The next leap in performance is likely to stem from methods that can leverage large amounts of unannotated data in a self-supervised manner, and uncover the underlying structure and knowledge in the data to improve segmentation of medical images. The pipelines in medical image segmentation networks commonly follow an encoder-decoder architecture resembling a pixel-to-pixel mapping from image to segmentation maps.

One of the earliest and most successful representatives

*The authors contributed equally to this work

of this approach is U-Net [65], which uses CNN blocks and skip-connections from encoder to decoder at different resolution levels. Many works proposed CNN architectures that were inspired by or expanded upon U-Net [33, 70, 69, 37]. Until today though, most CNN-based approaches are outperformed by ensembles of vanilla U-Net architectures as proposed in nnU-Net [38], which showed that image preprocessing heuristics may have a larger effect than architectural improvements. The inductive bias of CNNs is beneficial for faster convergence. However, it can also cause the model to saturate faster and miss complex underlying relations. Therefore, more complex networks like attention-based models [15] were explored for medical image segmentation.

Transformers [76] introduced attention layers for natural language processing (NLP), and were recently adapted to Vision Transformers (ViT) for different computer vision tasks [24]. Some models for medical image segmentation employed purely transformer-based architectures [16, 8]. Others followed a hybrid approach and incorporated transformer blocks in the encoder [12, 85, 83, 16], decoder [51, 52], both [80, 97, 54] or other network parts [15, 93, 46]. Although transformer-based designs lead to a lack of inductive bias and a wider field of view compared to CNNs, they require a large amount of data for training. Further, the memory requirement of attention layers grows quadratically with the number of image patches, which leads to higher computational resources compared to CNNs, particularly during training. Interpretability is also another crucial criteria, especially in sensitive scenarios like medical applications, and despite some efforts to improve it [14, 42], investigating the interpratibility of attention-based architectures is more challenging compared to CNNs [40].

The lack of carefully annotated data in medical imaging is another challenge to adopting transformers for medical image segmentation. As a consequence, the adoption of self-supervised learning methods for segmentation tasks has been proposed [8, 16]. There are several approaches proposed for pretraining Transformers. The most common one is introduced for natural language processing through defining pretext tasks [76]. Although adopting this approach for images showed promising results [24], the DINO model [11] which was recently adopted for self-supervised training of vision transformers, could be a better fit for computer vision tasks. DINO follows a student-teacher scheme for distilling the knowledge of the teacher branch to the student branch. The teacher branch receives a larger field of view, and the teacher's knowledge is propagated to the student in an unsupervised manner. Despite the difference in training of DINO [11] and ViT [24], both reported an emergence of meaningful shapes in their attention map visualizations.

In this work, we take advantage of these attention map visualizations extracted from the pretrained DINO model for the generation of pairs of image-attention map visualization. We hypothesize that these visualizations encode additional information that can be leveraged for downstream tasks such as semantic segmentation. Our goal is to leverage the benefits of transformers in medical image segmentation without having to bear their limitations and high complexities. Therefore, we generate pairs of image-attention map visualizations to combine the perspective and richness of extracted features in transformers with the simplicity and effectiveness of CNNs in medical image segmentation.

As it can be observed in Figure 1, each head resembles certain parts of the segmentation maps; thus the model is enforced to capture all the available information by incorporation of the generated attention-map visualizations as input data. To achieve this goal, we propose **DAINet** (**D**ual **A**ttention-**I**mage Network), a novel method for incorporating the knowledge of the transformer into a simple CNN architecture. For this, we propose two architecture variations that take the image and attention-map visualization pairs an input. First, we concatenate the input image with the attention map visualizations at channel level and feed these to the segmentation network. This variation is limited by the entangled representation learning of the visualizations and spatial features that need to be extracted from the image. Therefore, we propose a second variant with two distinct encoders. The first encoder receives the input image and the attention visualizations are fed to the second encoder. This simple modification enforces the network to extract more meaningful features that may correlate better with the predicted segmentation maps. We verify this with an ablation study of skip connections from the attention encoder to the decoder of the segmentation network. In summary, our main contributions in this work are:

- We demonstrate that transformers trained in a self-supervised manner could capture essential information, which can be leveraged for downstream tasks such as semantic segmentation. To the best of our knowledge, this is the first work that directly incorporates these information for the image segmentation task.

- We introduce a simple yet effective CNN-based architecture to employs the attention maps visualizations along the original image for medical image segmentation. This potentially allows for more investigatory measures like interpretability that has been widely studied for CNNs and is cruical for medical applications.

- We show the effectiveness of our approach in two well-known organ segmentation datasets and compare its performance to other well-known segmentation techniques based on transformers and CNNs.

2. Related Work

Deep learning has immensely affected medical image segmentation [88, 86, 28, 2, 87]. Among earlier architectures, Long *et al*. [57] introduced an image-image mapping based on Fully Convolutional Network. Later, U-Net revolutionized medical image segmentation with its encoder-decoder architecture connecting in a bottleneck and skip-connections between encoder and decoder components [65]; since then, U-Net has been primarily used as a benchmark in medical image segmentation. Some other architectures also proposed incorporating the components of U-Net, e.g., [70, 69, 25], which most of them use U-Net-based skip connections. Generalization is improved by shortening the gap between the encoder and decoder semantic maps [98]. Adding residual connections for each encoder and decoder block and dividing the input image into patches with a weighted map for each patch as input to the model is also investigated [82]. Diakogiannis *et al*. [22] utilized U-Net backbone, combined with residual connections, atrous convolutions, pyramid scene parsing pooling and multi-tasking inference, and Jha *et al*. [41] explored the advantages of residual blocks combined with the squeeze and excitation blocks. V-Net [59] and 3D U-Net [99] utilized similar architecture for 3D medical image segmentation [66]. U-Net architecture use-cases also expanded to other medical image analysis tasks such as computer-aided diagnosis [71, 73, 72, 53, 23], image denoising [89, 58], image registration [4, 34], and it is also utilized in diffusion models for state-of-the-art image generation [36, 26]. Apart from the traditional U-Net form, Y-Net [27], a segmentation network with two encoders and one shared decoder has been recently proposed for medical image segmentation. Y-Net introduces a spectral encoder that extracts frequency domain features and shows superior performance. The spectral encoder consists of Fast Fourier convolutional (FFC) blocks with fast Fourier transform at their core.

Convolutional layers have inherent inductive bias, that limit the ability of the network to make spatial relation in different parts of the input image. Some works employed sequence-to-sequence modules like RNN and LSTM in medical image segmentation [35, 68]. Dilated convolutions [90], as the name suggests, tried to modify convolutional layers to have a broader view, and they were used by many methods to preserve the spatial size of the feature map and to enlarge the receptive field [17, 95, 92]. The receptive field can also be enlarged by using larger kernels [60] and enriched with contextual information by using kernels of multiple scales [18]. Hardnet [13] employs a Harmonic Densely Connected Network that is shown to be highly effective in many real-time tasks, such as classification and object detection. Wang *et al*. used attention maps in each feature map of the encoder-decoder block to enlarge the representation between farther areas in the image [78] and it was extended by the incorporation of scalar gates in the attention layer [75]. In the same work, they proposed Local-Global training strategy or LoGo that aims to consider specialized branches for local and global views.

2.1. Vision Transformers

Transformer architecture was first introduced for machine translation and thereafter became the dominant backbone architecture for Natural Language Processing; it is constructed mainly by attention-blocks in the form of encoder-decoder, which utilized the GPUs much more efficiently compared to other sequence-sequence blocks like LSTM and RNN [76]. Depending on the task, many architectures adopted the original encoder-decoder form [49, 47, 19] and some only used the encoder part [21, 55, 45, 5] or decoder part [61, 20, 84]. Inspired by the success of transformers, many attempted to bring transformers to vision tasks [79, 9, 63, 94], and Dosovitskiy *et al*. [24] inspired by the relative simplicity of BERT architecture [21], introduced vision transformers (ViT) as the state-of-the-art in image classification tasks; They also explored the concept of hybrid architectures that uses CNNs to generate embeddings, and other works like CvT [81] improved the ViT performance by incorporating CNNs. LeViT [31] used low-resolution attention maps and combined them with CNNs to improve the inference time. Many other works extend ViT to other tasks like video classification [7] and ViLT [43] extract and combine features from text and image for better inference. To fit the transformer architecture for tasks other than classification, hybrid architectures were proposed. For object detection, DETR [10] uses a CNN block as the entry backbone for feature extraction and bounding box detection and uses the encoder-decoder transformer block for prediction of the labels. For semantic segmentation, SETR [96] transformers are implemented to extract features in the encoder for sequentializing images without using the traditional FCN [57]. Ranftl *et al*. [64] leverage vision transformers in place of CNNs as a backbone for dense prediction tasks.

2.1.1 Self-Supervised Learning in Vision Transformers

Self-supervised learning to pretrain transformers initially introduced as pretext task in NLP by masking random embedding vectors and optimizing the model to recover them. In the same way, ViT [24] also uses a simple pretext approach, by randomly masking patches of the input image, and asking the model to predict its average color. In another work, Atito *et al*. [3] utilize group mask model learning (GMML) for pretext pretraining of vision transformers to improve the simple masking approach. SelfPatch [91] is also a pretext technique that aims to learn better patch-level representations. On the other hand, DINO [11] follows a student-teach scheme that is more common in computer vision. EsViT [50] also exploits Knowledge Distillation with a fixed teacher network and a

student network that is continuously updated in an attempt to minimize a loss function. Caron *et al*. [11] claimed in DINO that with this self-supervised approach, the extracted features in self-supervised vision transformers contain meaningful and intuitive information about the image, which does not appear as explicit as supervised vision transformers or CNNs, which also inspired this work to benefit from such information in medical image segmentation. Ge *et al*. [30] also presented a pipeline to use the priors of transformers in a CNN network; it has two branch architectures: one for the CNN, and the other for the transformer that guides the CNN branch in a self-supervised manner, and both branches are trained simultaneously.

2.2. Transformers for medical image segmentation

A pure transformer-based model for medical image segmentation was introduced by Karimi *et al*. [16] for 3D medical image segmentation, and Swin-UNet [8] based on Swin-Transformers [56] showed its effectiveness on ACDC [6] and other datasets; however, hybrid architectures were more researched. In most of such architectures, similar to U-Net, an encoder-decoder shape is followed. Transformer blocks are utilized in different parts of such networks. Swin UNETR [74], Trans Claw-UNet [12], Claw-UNet [85] and LeViT-UNet [83] adopted transformers for feature extraction in the encoder. TransUNet [16] with a similar approach showed superior performance on synapse dataset [48] and ACDC [6]. Fewer works explored Transformers in only decoder parts [51, 52]. On the other hand, UTNet [80], nnFormer [97] and Dual Swin Transformer UNet (DS-TransUNet) [54] incorporated the transformers along CNNs in both encoder and decoder part. TransAttUNet [15] employed guided attentions in skip-connections to provide more expressive representation. Axial Fusion Transformer UNet (AFTer-UNet) [93] implement a fusion layer with axial fusion layers, and SegTHOR [46] also suggested another type of fusion layer.

3. Method

We present a simple architecture for semantic segmentation adapted from the well-known U-Net [65]. Unlike other complicated methods, rather than incorporating transformer components in our architecture, we propose extracting attention map visualizations from the multi-head attention layer of the last block in a pre-trained vision transformer model such as DINO [11] and feeding them to the segmentation network in addition to the input image. Our framework consists of two steps: 1) self-supervised training of the DINO model to obtain the attention map visualizations for images, 2) Training our proposed segmentation model using the segmentation map annotations, input images, and their corresponding attention map visualizations from step 1. Figure 2 shows an overview of our proposed segmentation network in step (2), while the input to the lower branch is obtained from step (1).

3.1. Definitions

Given a dataset $\mathcal{D}$ of input images $x \in \mathbb{R}^{H\times W\times C}$ with height H, width W and number of input channels C, and their corresponding segmentation map annotations $y \in \mathbb{R}^{H\times W\times N}$, where $\{x, y\} \in \mathcal{D}$. The number of classes in our dataset is defined by N. The goal of our segmentation model parameterized by θ is to predict the semantic segmentation maps $\hat{y} = \theta(x)$. We denote the pre-trained DINO [11] model by ϕ. We extract the attention map visualization from each self-attention head in DINO and denote them by ν_i where $i \in \{1, ..., h\}$ defines the head index and h defines the total number of heads. Therefore, the predicted segmentation map becomes a function of ν in addition to x, leading to $\hat{y} = \theta(x, \nu)$.

3.2. DINO

As mentioned earlier, DINO [11] is a vision transformer-based architecture that is trained in a self-supervised manner without labels. DINO consists of a teacher (parameterized by ϕ_t) model and a student model (parameterized by ϕ_s) and is trained using self-distillation. The teacher and the student share the same architecture and receive two augmented cropped views x_1, x_2 from the input image x. The architecture of ϕ is based on a vision transformer [24] (ViT), followed by a projection head with the dimension K that outputs the probability distributions P_s, P_t of the student and teacher models, respectively; where $P(x)^{(j)}$ is the probability distribution of input image patch $x^{(j)}$ in Equation 1:

$$P(x)^{(j)} = \frac{\exp(\phi(x)^{(j)}/\tau)}{\sum_{k=1}^{K} \exp(\phi(x)^{(k)}/\tau)}. \tag{1}$$

Each network has its own parameters ϕ_s, ϕ_t and temperature τ_s, τ_t, which yields P_s, P_t using Equation 1. The temperature parameter defines the sharpness of the probability distribution P. The backbone model (ViT-S/16) receives a grid of image patches $x^{(j)}$ with resolution 16×16 as input. A set of embeddings are generated by feeding the image patches to a linear layer, and are then followed by a learnable token with the goal of aggregating the information from the whole grid sequence. The embeddings are passed to a standard Transformer network which is a sequence of self-attention and feed-forward layers with skip connections. There are in total h self-attention heads in the last block of the network, that generate the attention map visualizations $\nu_{i,j}$ for $i \in (1, ..., h)$ of image patch $x^{(j)}$.

DINO Optimization The student model parameters are updated by applying stochastic gradient descent and minimizing the cross-entropy loss between the features from the

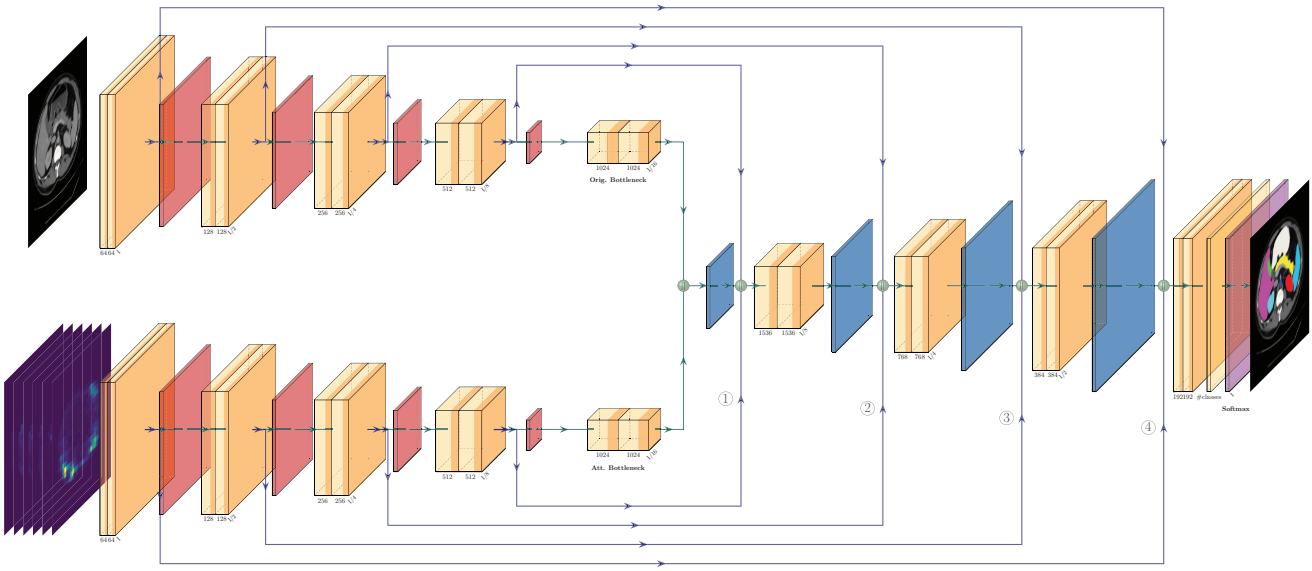

Figure 2. **DAINet Architecture.** (Convolution, ReLu + Batch Norm, Max. Pooling, Softmax, Transposed Convolution, Concatenation.)

student and teacher (Equation 2):

$$\mathcal{L}_{st} = \frac{-P_t(x_1)\log P_s(x_2)}{2} + \frac{-P_t(x_2)\log P_s(x_1)}{2}. \quad (2)$$

The teacher model parameters are optimized via an exponential moving average of the students' parameters using:

$$\phi_t \leftarrow \lambda\phi_t + (1-\lambda\phi_s), \quad (3)$$

where λ defines a a cosine scheduler from 0.996 to 1.

3.3. Semantic Segmentation

The main contribution of our method lies in showing that the self-attention visualizations in a self-supervised pre-trained transformer-based architecture can be employed without actually using the transformer components in the main model architecture, thus leading to a simple architecture, similar to U-Net, that uses additional features as input to perform the segmentation task more effectively. We also define a switching mechanism that allows further customization for keeping or removing skip connections.

We present two versions of our segmentation model. The first version simply adapts the U-Net [65] model by increasing the number of input channels in θ and concatenating the input image x with the attention map visualizations ν. This modification is done by changing the number of input channels in the initial layer of the U-Net from C to $C+h$ (number of attention heads).

We hypothesize that since the input image and the attention map visualizations are from different domains, introducing an extra encoder to extract the features from the attention map visualizations would lead to better feature modulation. Furthermore, we believe that the attention map visualizations from a model encode more valuable information than the original image, which would have higher correlation with the final segmentation map; thus facilitating the possibility of the model to assign higher weights to such features in a separate encoder.

Therefore, we propose a network with two encoders and a shared decoder for predicting the semantic segmentation map. The first encoder E_x gets the original image x as input, while the second encoder E_ν, which has h input channels, receives the attention map visualizations ν of the different heads. The features extracted by these two encoders are concatenated at the bottleneck and then fed to the shared decoder D.

Since the skip connections from the encoders provide a direct connection to the decoder features and, as a result, the predicted segmentation map, we explore the settings of employing the skip connections at different points. These are denoted by switches (①, ②, ③, ④) in our experiments defining whether the skip connection exists at the specified point or not. Figure 2 depicts the DAINet network with all switches ① = ② = ③ = ④ = 1.

Losses To optimize our segmentation model, we employ the combined cross-entropy, and dice loss [39]. The dice-coefficient loss (Equation 5) has high flexibility towards class imbalance, while the cross-entropy loss (Equation 4) helps with the curve smoothing [39].

$$\ell_{CE}(\hat{y}, y) = -\frac{1}{N}\sum_{n=0}^{N} y_n \log(\hat{y}_n), \quad (4)$$

Table 1. Quantitative results of our segmentation model compared to SOTA on Synapse Dataset [48]

Method	DSC (%) ↑	HD95 (mm) ↓	Aorta	Gallbladder	Kidney (L)	Kidney (R)	Liver	Pancreas	Spleen	Stomach
V-Net* [59]	68.81	-	75.34	51.87	77.10	80.75	87.84	40.05	80.56	56.98
DARR* [29]	69.77	-	74.74	53.77	72.31	73.24	94.08	54.18	89.90	45.96
R50 U-Net* [1]	74.68	36.87	84.18	62.84	79.19	71.29	93.35	48.23	84.41	73.92
R50 AttnUNet* [67]	75.57	36.97	55.92	63.91	79.20	72.71	93.56	49.37	87.19	74.95
U-Net# [65]	76.85	39.70	89.07	69.72	77.77	68.60	93.43	53.98	86.67	75.58
AttnUNet# [67]	77.77	36.02	89.55	68.88	77.98	71.11	93.57	58.04	87.30	75.75
R50 ViT# [24]	71.29	32.87	73.73	55.13	75.80	72.20	91.51	45.99	81.99	73.95
ViT* [24]	61.50	39.61	44.38	39.59	67.46	62.94	89.21	43.14	75.45	69.78
TransUNet [16]	77.48	31.69	87.23	63.13	81.87	77.02	94.08	55.86	85.08	75.62
TransClaw U-Net [12]	78.09	26.38	85.87	61.38	84.83	79.36	94.28	57.65	87.74	73.55
MT-UNet [77]	78.59	26.59	87.92	64.99	81.47	77.29	93.06	59.46	87.75	76.81
U-Net †	79.52	33.99	89.64	69.73	82.79	77.26	93.50	61.71	84.15	77.36
Swin-UNet [8]	79.13	21.55	85.47	66.53	83.28	79.61	94.29	56.58	90.66	76.60
TransCASCADE [62]	82.68	17.34	86.63	68.48	87.66	**84.56**	94.43	65.33	90.79	**83.52**
DAINet (Ours)	**84.26**	**13.79**	**89.66**	**72.47**	**87.89**	83.90	**95.34**	**67.61**	**93.74**	83.48

$$\ell_{DICE}(\hat{y}, y) = 1 - \frac{2y\hat{y} + \epsilon}{y + \hat{y} + \epsilon}, \tag{5}$$

where ϵ is added to the numerator and denominator for numerical stability. The total loss (Equation 6) for the segmentation model then is the sum of dice and cross-entropy loss:

$$\mathcal{L}_{seg} = \frac{1}{2}\ell_{CE} + \frac{1}{2}\ell_{DICE}. \tag{6}$$

4. Experiments

In this section we present the results of our experiments on two publicly benchmarks for medical image segmentation namely the ACDC [6] and Synapse [48] datasets. These datasets hold images of MR and CT belonging to different sets of organs with different sizes and intensities with challenging structures and shapes for the evaluation of our proposed networks. As follows, we shortly refer to the public benchmarks used in this work, then we present the experimental setup and implementation details. Finally, we demonstrate the results of our experiments compared against state-of-the-art (SOTA) methods and after that an ablation study of different settings of our proposed method. We used the average dice score and Hausdorff distance for the evaluation, which are the standard metrics in medical image segmentation.

4.1. Datasets

Synapse The Synapse dataset [48] is a multi-organ segmentation dataset of abdominal CT images. We follow the same experimental protocol to MT-UNet [77] for training and evaluation of our model. In total, 30 abdominal CT scans and their corresponding semantic segmentation maps, belonging to eight abdominal organs (aorta, gallbladder, spleen, left kidney, right kidney, liver, pancreas, spleen, and stomach) are adopted. Each CT volume in the dataset has varying number of between 85 to 198 slices with a resolution of 512×512. Similar to [77], we employ 18 volumes for training and 12 volumes for testing.

ACDC The Automated Cardiac Diagnosis Challenge (ACDC) dataset [6] consists of cardiac MR images from 150 patients. Similar to Synapse, we follow the same experimental protocol as MT-UNet [77] and TransUNet [16], which utilize the data of 100 out of the 150 patients in this dataset. For each data sample in the dataset, the data for two modalities of end-diastole (ED) and end-systole (ES) are provided. The annotations provided in the dataset provide semantic segmentation maps belonging to three regions, left ventricle (LV), right ventricle (RV), and myocardium (Myo). The train / validation and test splits follow the same setting as previous work with 70, 10, and 20 samples respectively. The slices in this dataset have a resolution of 352×352 and each volume has between 7 and 17 slices.

4.2. Experimental Setup

We follow the same training and evaluation scheme as [77], if not otherwise stated. We use the Adam optimizer [44] for training the model. The initial learning rate is set to $1e{-}4$, with a step scheduler to gradually decrease the learning rate in each iteration of the training with a rate of 0.9. The model has a weight decay of $1e{-}4$ for regularization. The maximum number of training epochs is set to 300, while we apply early stopping for our model based on the validation dice score. The batch size was set to 16 and the image resolution was 224×224. For augmentation purposes, we apply random flipping and rotations to the images during the training. We report the dice score value for different organs and regions and the average dice score for both datasets. For the Synapse dataset [48], we also report the Hausdorff distance (HD95) between the predicted and ground truth segmentation maps.

To extract the attention map visualizations for both ACDC and Synapse datasets, we employed the pre-trained DINO [11] model on ImageNet and fine-tuned it on the corresponding dataset. We adopt the ViT-S/16 model and fine-tune it on images with a resolution of 256×256, with a batch size of 64 for 800 epochs and with a learning rate of $1e{-}4$ on the training set of each dataset. To employ attention map visualizations with the same resolution as the input images for the segmentation task, we downsample the attention map visualization to 224×224. The ViT-S/16 model has 6 attention heads, which would also set the number of input channels to our attention encoder to 6.

4.3. Results

We present quantitative and qualitative results of our method compared against the state-of-the-art as follows. Then, we show an ablation study of the components of our architecture. The quantitative results of our model compared to the SOTA on the ACDC dataset is presented in Table 2. To have a fair comparison against the SOTA, we train and evaluate the UNet model which is the most similar architecture to ours. The values reported in Table 1 are obtained from each original paper unless specified. The UNet † model has the same configuration as Swin-Unet. The results marked with * are obtained from [16] and # from [8]. All the experiments demonstrate the superiority of DAINet to comparable previous work. DAINet outperforms the complex transformer-based architectures such as MT-UNet [77] and Swin-UNet [8] by a large margin in terms of dice score. The qualitative results in Figure 3 show that DAINet predicts the segmentation map more accurately compared to U-Net.

4.3.1 Comparison to SOTA

We show the results of our experiments on the Synapse dataset [48] both quantitatively in Table 1 and qualitatively in Figure 3. As it can be seen in Table 1, DAINet outperforms Swin-Unet in terms of the average dice score and HD95 by 5.13 percent and 7.76 points, respectively. The results show that the performance differences among the compared methods could be dependent on the shape and the size of the segmented anatomy. It can be observed that for the larger organs, the results have less variations. For instance in Table 1, we see that the dice score in the liver has the highest value among abdomen organs, and despite the fact that our model outperformed others, the improvement is marginal and only around 1.05%. We also observe that the models have much more variations in segmenting organs with less compact forms or irregular shapes. For instance, in the case of Aorta, our proposed method, U-Net and Attn-Unet have similar performances while transformer based-architectures have lower performance in terms of dice score. This could be due to the existence of skip connections in the U-Net family that directly enrich the decoder's inference with image priors. Additionally, our ablation study in Table 3, offers some evidence that the placement of skip-connections in the U-Net family could be a potential factor in segmentation of different organs as well. It can be seen that with lowering the effect of skip connections in the attention map branch of the last layer, the performance of *LV* improves, which is the only label Table 2 at which U-Net is marginally better.

Table 2. Comparison of our method against related work on the ACDC [6] dataset (* obtained from [16], and † trained by us).

Method	RV	Myo	LV	DSC (%)
R50 U-Net* [1]	84.62	84.52	93.68	87.60
R50 AttnUNet* [67]	83.27	84.33	93.53	86.90
ViT-CUP* [24]	80.93	78.12	91.17	83.41
R50 ViT* [24]	82.51	83.01	93.05	86.19
TransUNet [16]	86.67	87.27	95.18	89.71
Swin-Unet [8]	85.77	84.42	94.03	88.07
MT-UNet [77]	86.64	89.04	95.62	90.43
U-Net† [65]	89.67	89.27	**95.76**	91.57
TransCASCADE [62]	89.14	**90.25**	95.50	91.63
DAINet (Ours)	**90.53**	89.52	95.63	**91.90**

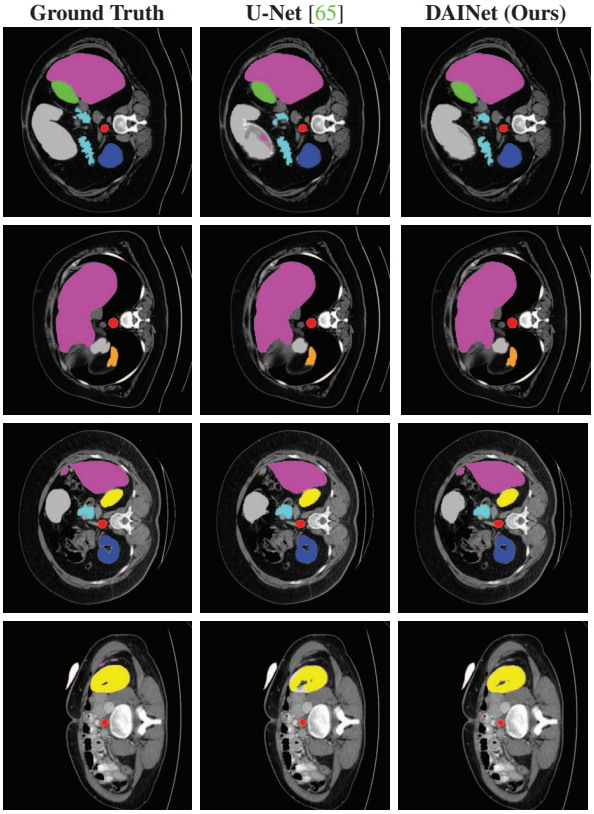

Figure 3. **Some qualitative results on comparison of DAINet against U-Net on the test data of the Synapse dataset [48].** (Aorta, Gallbladder, Left Kidney, Right Kidney, Liver, Spleen, Pancreas, and Stomach)

4.3.2 Ablation Study

In Table 3, we present an ablation study of the proposed method. We analyze and discuss the role of different skip connections by gradually adding them from the first / last block to the bottleneck, as well as feeding the attention maps to a single encoder together with the image as follows.

Single Encoder In this experiment, we present the results of the model with a single encoder, where the attention map visualizations are simply concatenated together with the input image to a single encoder. The basis for proposing a dual encoder network for our method is that the original image and the attention maps are from different domains and allowing their features to be extracted by two separate encoders provide richer features at the bottleneck. This can be seen in Table 3, as it shows that the dual encoder architecture achieves 1.5% higher dice score compared to the single one.

Residual Connections The results of the ablation study in Table 3 show that having all four skip connections from the attention encoder to the decoder give the best overall performance; however, this is not the case for all of the regions. We observe that the removing the last skip connection has an advantage in *LV*, and we speculate that this effect might be due to the fact that some shapes, maybe because of simplicity, benefit from using the priors that are directly obtained from the shapes and the priors from transformers makes the inference more complex; therefore, lowering the effect of the skip-connection in the last layer of the attention encoder yields better performance

Computational Cost We have compared the number of parameters that are used in transformer-based and state-of-the-art models in medical image segmentation. We observed that the number of parameters in some of the models is comparably large, and in turn, the training cost of the model is expected to be higher. On the other hand it is assumed that in most cases, the transformer block requires more training to capture the inductive bias as is reported by Dosovitskiy [24]. It can be therefore considered that this affects the training speed and cost even further.

Table 3. Ablation study of our model on ACDC [6]

Encoder	Att. Skip-Connection ①	②	③	④	RV	Myo	LV	DSC (%)
Single	-	-	-	-	88.69	87.22	95.25	90.38
Dual	0	0	0	0	89.83	89.29	95.67	91.60
	0	0	0	1	90.23	89.27	95.65	91.71
	0	1	1	1	90.39	89.42	**95.75**	91.85
	1	1	1	1	**90.53**	**89.52**	95.63	**91.90**

5. Limitations

Our proposed pipeline receives pairs of image-attention map visualizations, which do not change during the training. The transformer block in our method can also be fine-tuned for different datasets for further improvement, but on the other hand, other models that incorporate the transformer blocks into their architecture are easier to be fine-tuned. If the generated attention-map visualizations capture irrelevant features, as seen in Figure 1, especially visible in Attention Head 3 in sample B, it might give incorrect bias for the segmentation task.

Ideally, to demonstrate the full potential of these architectures, utilizing a ViT/DINO foundational model in 3D, pre-trained in a self-supervised manner on large amounts of unannotated medical images is preferred. However, such a model does not exist, and training our own backbone would require significant amounts of data and computational resources. Instead, we use this work as an opportunity to investigate how to optimally incorporate a 2D DINO backbone for medical image segmentation and showcase its effectiveness. In our results, we are able to show that our two simple 2D CNN architectures utilizing DINO attention maps are sufficient to outperform many reference and SOTA approaches, even those trained in 3D. In fact, our 2D method is also outperformed a 3D hybrid ViT-CNN architectures that is pre-trained on a large data [32] on HD95 metric (ours: 13.79 compared to 20.53), whose attention maps are not directly usable in our approach as those from 2D DINO. This work, therefore, opens the path for new possibilities when novel backbone architectures become available.

6. Conclusion

In this work, we presented a simple yet effective model for semantic segmentation of medical images called DAINet. With this architecture, we showed that the self-attention map visualizations in transformers trained in a self-supervised manner, such as DINO could capture meaningful features that can be directly used as input for improving medical image segmentation. We used a model (here DINO) pre-trained on the ImageNet dataset and fine-tuned on our target dataset, and showed its effectiveness of the extracted features in medical image segmentation. Unlike other methods that incorporate the transformer blocks in the main architecture, our architecture does not depend on the transformer blocks in the run-time, and it can achieve state-of-the-art performance on two medical image segmentation benchmarks. We also presented an ablation study on different customization of the architecture. The proposed method can open the path for future architecture designs that aim to be lightweight and interpretable, yet take advantage of the representation power of transformers in their pipeline.

References

[1] Saruar Alam, Nikhil Kumar Tomar, Aarati Thakur, Debesh Jha, and Ashish Rauniyar. Automatic polyp segmentation using u-net-resnet50. *arXiv preprint arXiv:2012.15247*, 2020. 6, 7

[2] Mehdi Astaraki, Francesca De Benetti, Yousef Yeganeh, Iuliana Toma-Dasu, Örjan Smedby, Chunliang Wang, Nassir Navab, and Thomas Wendler. Autopaint: A self-inpainting method for unsupervised anomaly detection. *arXiv preprint arXiv:2305.12358*, 2023. 3

[3] Sara Atito, Muhammad Awais, and Josef Kittler. Sit: Self-supervised vision transformer. *arXiv preprint arXiv:2104.03602*, 2021. 3

[4] Guha Balakrishnan, Amy Zhao, Mert R Sabuncu, John Guttag, and Adrian V Dalca. Voxelmorph: a learning framework for deformable medical image registration. *IEEE transactions on medical imaging*, 38(8):1788–1800, 2019. 3

[5] Iz Beltagy, Matthew E Peters, and Arman Cohan. Longformer: The long-document transformer. *arXiv preprint arXiv:2004.05150*, 2020. 3

[6] Olivier Bernard, Alain Lalande, Clement Zotti, Frederick Cervenansky, Xin Yang, Pheng-Ann Heng, Irem Cetin, Karim Lekadir, Oscar Camara, Miguel Angel Gonzalez Ballester, et al. Deep learning techniques for automatic mri cardiac multi-structures segmentation and diagnosis: is the problem solved? *IEEE transactions on medical imaging*, 37(11):2514–2525, 2018. 4, 6, 7, 8

[7] Gedas Bertasius, Heng Wang, and Lorenzo Torresani. Is space-time attention all you need for video understanding? In *Proceedings of the International Conference on Machine Learning (ICML)*, July 2021. 3

[8] 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, 2021. 2, 4, 6, 7

[9] Yue Cao, Jiarui Xu, Stephen Lin, Fangyun Wei, and Han Hu. Gcnet: Non-local networks meet squeeze-excitation networks and beyond. In *Proceedings of the IEEE/CVF international conference on computer vision workshops*, pages 0–0, 2019. 3

[10] Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-end object detection with transformers, 2020. 3

[11] Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9650–9660, 2021. 1, 2, 3, 4, 7

[12] Yao Chang, Hu Menghan, Zhai Guangtao, and Zhang Xiao-Ping. Transclaw u-net: Claw u-net with transformers for medical image segmentation. *arXiv preprint arXiv:2107.05188*, 2021. 2, 4, 6

[13] Ping Chao, Chao-Yang Kao, Yu-Shan Ruan, Chien-Hsiang Huang, and Youn-Long Lin. Hardnet: A low memory traffic network. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 3552–3561, 2019. 3

[14] Hila Chefer, Shir Gur, and Lior Wolf. Transformer interpretability beyond attention visualization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 782–791, 2021. 2

[15] Bingzhi Chen, Yishu Liu, Zheng Zhang, Guangming Lu, and David Zhang. Transattunet: Multi-level attention-guided u-net with transformer for medical image segmentation. *arXiv preprint arXiv:2107.05274*, 2021. 2, 4

[16] Jieneng Chen, Yongyi Lu, Qihang Yu, Xiangde Luo, Ehsan Adeli, Yan Wang, Le Lu, Alan L Yuille, and Yuyin Zhou. Transunet: Transformers make strong encoders for medical image segmentation. *arXiv preprint arXiv:2102.04306*, 2021. 2, 4, 6, 7

[17] Liang-Chieh Chen, George Papandreou, Iasonas Kokkinos, Kevin Murphy, and Alan L Yuille. Semantic image segmentation with deep convolutional nets and fully connected crfs. *arXiv preprint arXiv:1412.7062*, 2014. 3

[18] Liang-Chieh Chen, George Papandreou, Iasonas Kokkinos, Kevin Murphy, and Alan L Yuille. Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. *IEEE transactions on pattern analysis and machine intelligence*, 40(4):834–848, 2017. 3

[19] Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. Unsupervised cross-lingual representation learning at scale. *arXiv preprint arXiv:1911.02116*, 2019. 3

[20] Zihang Dai, Zhilin Yang, Yiming Yang, Jaime Carbonell, Quoc V Le, and Ruslan Salakhutdinov. Transformer-xl: Attentive language models beyond a fixed-length context. *arXiv preprint arXiv:1901.02860*, 2019. 3

[21] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics. 3

[22] Foivos I Diakogiannis, François Waldner, Peter Caccetta, and Chen Wu. Resunet-a: A deep learning framework for semantic segmentation of remotely sensed data. *ISPRS Journal of Photogrammetry and Remote Sensing*, 162:94–114, 2020. 3

[23] Jia Ding, Aoxue Li, Zhiqiang Hu, and Liwei Wang. Accurate pulmonary nodule detection in computed tomography images using deep convolutional neural networks. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 559–567. Springer, 2017. 3

[24] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020. 2, 3, 4, 6, 7, 8

[25] Azade Farshad, Anastasia Makarevich, Vasileios Belagiannis, and Nassir Navab. Metamedseg: Volumetric meta-learning for few-shot organ segmentation. In *MICCAI Workshop on*

Domain Adaptation and Representation Transfer, pages 45–55. Springer, 2022. 3

[26] Azade Farshad, Yousef Yeganeh, Yu Chi, Chengzhi Shen, Björn Ommer, and Nassir Navab. Scenegenie: Scene graph guided diffusion models for image synthesis. *arXiv preprint arXiv:2304.14573*, 2023. 3

[27] Azade Farshad, Yousef Yeganeh, Peter Gehlbach, and Nassir Navab. Y-net: A spatiospectral dual-encoder network for medical image segmentation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 582–592. Springer, 2022. 3

[28] Azade Farshad, Yousef Yeganeh, and Nassir Navab. Learning to learn in medical applications: A journey through optimization. In *Meta-Learning with Medical Imaging and Health Informatics Applications*, pages 3–25. Elsevier, 2023. 3

[29] Shuhao Fu, Yongyi Lu, Yan Wang, Yuyin Zhou, Wei Shen, Elliot Fishman, and Alan Yuille. Domain adaptive relational reasoning for 3d multi-organ segmentation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 656–666. Springer, 2020. 6

[30] Chongjian Ge, Youwei Liang, Yibing Song, Jianbo Jiao, Jue Wang, and Ping Luo. Revitalizing cnn attention via transformers in self-supervised visual representation learning. *Advances in Neural Information Processing Systems*, 34:4193–4206, 2021. 4

[31] Benjamin Graham, Alaaeldin El-Nouby, Hugo Touvron, Pierre Stock, Armand Joulin, Hervé Jégou, and Matthijs Douze. Levit: a vision transformer in convnet's clothing for faster inference. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 12259–12269, 2021. 3

[32] Ali Hatamizadeh, Vishwesh Nath, Yucheng Tang, Dong Yang, Holger R Roth, and Daguang Xu. Swin unetr: Swin transformers for semantic segmentation of brain tumors in mri images. In *Brainlesion: Glioma, Multiple Sclerosis, Stroke and Traumatic Brain Injuries: 7th International Workshop*. Springer, 2022. 8

[33] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016. 2

[34] Mattias P Heinrich. Closing the gap between deep and conventional image registration using probabilistic dense displacement networks. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 50–58. Springer, 2019. 3

[35] Mohammad Hesam Hesamian, Wenjing Jia, Xiangjian He, and Paul Kennedy. Deep learning techniques for medical image segmentation: achievements and challenges. *Journal of digital imaging*, 32(4):582–596, 2019. 3

[36] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in Neural Information Processing Systems*, 33:6840–6851, 2020. 3

[37] Gao Huang, Zhuang Liu, Laurens Van Der Maaten, and Kilian Q Weinberger. Densely connected convolutional networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4700–4708, 2017. 2

[38] Fabian Isensee, Paul F. Jaeger, Simon A. A. Kohl, Jens Petersen, and Klaus H. Maier-Hein. nnU-net: a self-configuring method for deep learning-based biomedical image segmentation. *Nature Methods*, 18(2):203–211, Dec. 2020. 2

[39] Shruti Jadon. A survey of loss functions for semantic segmentation. In *2020 IEEE Conference on Computational Intelligence in Bioinformatics and Computational Biology (CIBCB)*, pages 1–7. IEEE, 2020. 5

[40] Sarthak Jain and Byron C Wallace. Attention is not explanation. *arXiv preprint arXiv:1902.10186*, 2019. 2

[41] Debesh Jha, Pia H Smedsrud, Michael A Riegler, Dag Johansen, Thomas De Lange, Pål Halvorsen, and Håvard D Johansen. Resunet++: An advanced architecture for medical image segmentation. In *2019 IEEE International Symposium on Multimedia (ISM)*, pages 225–2255. IEEE, 2019. 3

[42] Sangwon Kim, Jaeyeal Nam, and Byoung Chul Ko. Vit-net: Interpretable vision transformers with neural tree decoder. In *International Conference on Machine Learning*, pages 11162–11172. PMLR, 2022. 2

[43] Wonjae Kim, Bokyung Son, and Ildoo Kim. Vilt: Vision-and-langsuage transformer without convolution or region supervision. In *International Conference on Machine Learning*, pages 5583–5594. PMLR, 2021. 3

[44] Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In Yoshua Bengio and Yann LeCun, editors, *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*, 2015. 6

[45] Nikita Kitaev, Łukasz Kaiser, and Anselm Levskaya. Reformer: The efficient transformer. *arXiv preprint arXiv:2001.04451*, 2020. 3

[46] Zoé Lambert, Caroline Petitjean, Bernard Dubray, and Su Kuan. Segthor: Segmentation of thoracic organs at risk in ct images. In *2020 Tenth International Conference on Image Processing Theory, Tools and Applications (IPTA)*, pages 1–6. IEEE, 2020. 2, 4

[47] Guillaume Lample and Alexis Conneau. Cross-lingual language model pretraining. *arXiv preprint arXiv:1901.07291*, 2019. 3

[48] Bennett Landman, Zhoubing Xu, J Igelsias, Martin Styner, T Langerak, and Arno Klein. Miccai multi-atlas labeling beyond the cranial vault–workshop and challenge, 2015. 4, 6, 7

[49] Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Ves Stoyanov, and Luke Zettlemoyer. Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. *arXiv preprint arXiv:1910.13461*, 2019. 3

[50] Chunyuan Li, Jianwei Yang, Pengchuan Zhang, Mei Gao, Bin Xiao, Xiyang Dai, Lu Yuan, and Jianfeng Gao. Efficient self-supervised vision transformers for representation learning. *arXiv preprint arXiv:2106.09785*, 2021. 3

[51] Shaohua Li, Xiuchao Sui, Xiangde Luo, Xinxing Xu, Yong Liu, and Rick Siow Mong Goh. Medical image segmentation using squeeze-and-expansion transformers. *arXiv preprint arXiv:2105.09511*, 2021. 2, 4

[52] Yijiang Li, Wentian Cai, Ying Gao, and Xiping Hu. More than encoder: Introducing transformer decoder to upsample. *arXiv preprint arXiv:2106.10637*, 2021. 2, 4

[53] Fangzhou Liao, Ming Liang, Zhe Li, Xiaolin Hu, and Sen Song. Evaluate the malignancy of pulmonary nodules using the 3-d deep leaky noisy-or network. *IEEE transactions on neural networks and learning systems*, 30(11):3484–3495, 2019. 3

[54] Ailiang Lin, Bingzhi Chen, Jiayu Xu, Zheng Zhang, and Guangming Lu. Ds-transunet: Dual swin transformer u-net for medical image segmentation. *arXiv preprint arXiv:2106.06716*, 2021. 2, 4

[55] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*, 2019. 3

[56] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 10012–10022, 2021. 4

[57] Jonathan Long, Evan Shelhamer, and Trevor Darrell. Fully convolutional networks for semantic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3431–3440, 2015. 3

[58] Qing Lyu, Chenyu You, Hongming Shan, and Ge Wang. Super-resolution mri through deep learning. *arXiv preprint arXiv:1810.06776*, 2018. 3

[59] Fausto Milletari, Nassir Navab, and Seyed-Ahmad Ahmadi. V-net: Fully convolutional neural networks for volumetric medical image segmentation. In *2016 fourth international conference on 3D vision (3DV)*, pages 565–571. IEEE, 2016. 3, 6

[60] Chao Peng, Xiangyu Zhang, Gang Yu, Guiming Luo, and Jian Sun. Large kernel matters–improve semantic segmentation by global convolutional network. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4353–4361, 2017. 3

[61] Alec Radford, Karthik Narasimhan, Tim Salimans, Ilya Sutskever, et al. Improving language understanding by generative pre-training. 2018. 3

[62] Md Mostafijur Rahman and Radu Marculescu. Medical image segmentation via cascaded attention decoding. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 6222–6231, 2023. 6, 7

[63] Prajit Ramachandran, Niki Parmar, Ashish Vaswani, Irwan Bello, Anselm Levskaya, and Jon Shlens. Stand-alone self-attention in vision models. *Advances in Neural Information Processing Systems*, 32, 2019. 3

[64] René Ranftl, Alexey Bochkovskiy, and Vladlen Koltun. Vision transformers for dense prediction. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 12179–12188, 2021. 3

[65] O. Ronneberger, P.Fischer, and T. Brox. U-net: Convolutional networks for biomedical image segmentation. In *Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, volume 9351 of *LNCS*, pages 234–241. Springer, 2015. (available on arXiv:1505.04597 [cs.CV]). 2, 3, 4, 5, 6, 7

[66] Abhijit Guha Roy, Shayan Siddiqui, Sebastian Pölsterl, Azade Farshad, Nassir Navab, and Christian Wachinger. Few-shot segmentation of 3d medical images. In *Meta-Learning with Medical Imaging and Health Informatics Applications*, pages 161–183. Elsevier, 2023. 3

[67] 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*, 53:197–207, 2019. 6, 7

[68] Iram Shahzadi, Tong Boon Tang, Fabrice Meriadeau, and Abdul Quyyum. Cnn-lstm: cascaded framework for brain tumour classification. In *2018 IEEE-EMBS Conference on Biomedical Engineering and Sciences (IECBES)*, pages 633–637. IEEE, 2018. 3

[69] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014. 2, 3

[70] Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, and Andrew Rabinovich. Going deeper with convolutions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1–9, 2015. 2, 3

[71] Hao Tang, Daniel R Kim, and Xiaohui Xie. Automated pulmonary nodule detection using 3d deep convolutional neural networks. In *Biomedical Imaging (ISBI 2018), 2018 IEEE 15th International Symposium on*, pages 523–526. IEEE, 2018. 3

[72] Hao Tang, Xingwei Liu, and Xiaohui Xie. An end-to-end framework for integrated pulmonary nodule detection and false positive reduction. In *2019 IEEE 16th International Symposium on Biomedical Imaging (ISBI 2019)*, pages 859–862. IEEE, 2019. 3

[73] Hao Tang, Chupeng Zhang, and Xiaohui Xie. Nodulenet: Decoupled false positive reductionfor pulmonary nodule detection and segmentation. *arXiv preprint arXiv:1907.11320*, 2019. 3

[74] 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. 4

[75] Jeya Maria Jose Valanarasu, Poojan Oza, Ilker Hacihaliloglu, and Vishal M Patel. Medical transformer: Gated axial-attention for medical image segmentation. *arXiv preprint arXiv:2102.10662*, 2021. 3

[76] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017. 2, 3

[77] Hongyi Wang, Shiao Xie, Lanfen Lin, Yutaro Iwamoto, Xian-Hua Han, Yen-Wei Chen, and Ruofeng Tong. Mixed transformer u-net for medical image segmentation. In *ICASSP 2022-2022 IEEE International Conference on Acoustics,*

Speech and Signal Processing (ICASSP), pages 2390–2394. IEEE, 2022. 6, 7

[78] Huiyu Wang, Yukun Zhu, Bradley Green, Hartwig Adam, Alan Yuille, and Liang-Chieh Chen. Axial-deeplab: Stand-alone axial-attention for panoptic segmentation. In *European Conference on Computer Vision*, pages 108–126. Springer, 2020. 3

[79] Song Wang, Xin Guo, Yun Tie, Ivan Lee, Lin Qi, and Ling Guan. Graph-Based Safe Support Vector Machine for Multiple Classes. *IEEE Access*, 6:28097–28107, 2018. 3

[80] Sinong Wang, Belinda Z Li, Madian Khabsa, Han Fang, and Hao Ma. Linformer: Self-attention with linear complexity. *arXiv preprint arXiv:2006.04768*, 2020. 2, 4

[81] Haiping Wu, Bin Xiao, Noel Codella, Mengchen Liu, Xiyang Dai, Lu Yuan, and Lei Zhang. Cvt: Introducing convolutions to vision transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 22–31, 2021. 3

[82] Xiao Xiao, Shen Lian, Zhiming Luo, and Shaozi Li. Weighted res-unet for high-quality retina vessel segmentation. In *2018 9th international conference on information technology in medicine and education (ITME)*, pages 327–331. IEEE, 2018. 3

[83] Guoping Xu, Xingrong Wu, Xuan Zhang, and Xinwei He. LeViT-UNet: make faster encoders with transformer for medical image segmentation. *arXiv preprint arXiv:2107.08623*, 2021. 2, 4

[84] Zhilin Yang, Zihang Dai, Yiming Yang, Jaime Carbonell, Russ R Salakhutdinov, and Quoc V Le. Xlnet: Generalized autoregressive pretraining for language understanding. *Advances in neural information processing systems*, 32, 2019. 3

[85] Chang Yao, Jingyu Tang, Menghan Hu, Yue Wu, Wenyi Guo, Qingli Li, and Xiao-Ping Zhang. Claw u-net: A unet variant network with deep feature concatenation for scleral blood vessel segmentation. In *CAAI International Conference on Artificial Intelligence*, pages 67–78. Springer, 2021. 2, 4

[86] Yousef Yeganeh, Azade Farshad, Johann Boschmann, Richard Gaus, Maximilian Frantzen, and Nassir Navab. Fedap: Adaptive personalization in federated learning for non-iid data. In *International Workshop on Distributed, Collaborative, and Federated Learning*, pages 17–27. Springer, 2022. 3

[87] Yousef Yeganeh, Azade Farshad, Goktug Guevercin, Amr Abu-zer, Rui Xiao, Yongjian Tang, Ehsan Adeli, and Nassir Navab. Scope: Structural continuity preservation for medical image segmentation. *arXiv preprint arXiv:2304.14572*, 2023. 3

[88] Yousef Yeganeh, Azade Farshad, Nassir Navab, and Shadi Albarqouni. Inverse distance aggregation for federated learning with non-iid data. In *Domain Adaptation and Representation Transfer, and Distributed and Collaborative Learning*, pages 150–159. Springer, 2020. 3

[89] Chenyu You, Linfeng Yang, Yi Zhang, and Ge Wang. Low-Dose CT via Deep CNN with Skip Connection and Network in Network. In *Developments in X-Ray Tomography XII*, volume 11113, page 111131W. International Society for Optics and Photonics, 2019. 3

[90] Fisher Yu and Vladlen Koltun. Multi-scale context aggregation by dilated convolutions. *arXiv preprint arXiv:1511.07122*, 2015. 3

[91] Sukmin Yun, Hankook Lee, Jaehyung Kim, and Jinwoo Shin. Patch-level representation learning for self-supervised vision transformers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8354–8363, 2022. 3

[92] Hang Zhang, Kristin Dana, Jianping Shi, Zhongyue Zhang, Xiaogang Wang, Ambrish Tyagi, and Amit Agrawal. Context encoding for semantic segmentation. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, pages 7151–7160, 2018. 3

[93] Yanci Zhang, Tianming Du, Yujie Sun, Lawrence Donohue, and Rui Dai. Form 10-q itemization. In *Proceedings of the 30th ACM International Conference on Information and Knowledge Management (CIKM '21)*, New York, NY, USA, 2021. Association for Computing Machinery. 2, 4

[94] Hengshuang Zhao, Jiaya Jia, and Vladlen Koltun. Exploring self-attention for image recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10076–10085, 2020. 3

[95] Hengshuang Zhao, Jianping Shi, Xiaojuan Qi, Xiaogang Wang, and Jiaya Jia. Pyramid scene parsing network. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2881–2890, 2017. 3

[96] Sixiao Zheng, Jiachen Lu, Hengshuang Zhao, Xiatian Zhu, Zekun Luo, Yabiao Wang, Yanwei Fu, Jianfeng Feng, Tao Xiang, Philip H.S. Torr, and Li Zhang. Rethinking semantic segmentation from a sequence-to-sequence perspective with transformers. In *CVPR*, 2021. 3

[97] Hong-Yu Zhou, Jiansen Guo, Yinghao Zhang, Lequan Yu, Liansheng Wang, and Yizhou Yu. nnformer: Interleaved transformer for volumetric segmentation. *arXiv preprint arXiv:2109.03201*, 2021. 2, 4

[98] Zongwei Zhou, Md Mahfuzur Rahman Siddiquee, Nima Tajbakhsh, and Jianming Liang. Unet++: A nested u-net architecture for medical image segmentation. In *Deep learning in medical image analysis and multimodal learning for clinical decision support*, pages 3–11. Springer, 2018. 3

[99] Özgün Çiçek, Ahmed Abdulkadir, Soeren S. Lienkamp, Thomas Brox, and Olaf Ronneberger. 3d u-net: Learning dense volumetric segmentation from sparse annotation, 2016. 3