Medical Image Segmentation via Cascaded Attention Decoding

Md Mostafijur Rahman Radu Marculescu
The University of Texas at Austin
{mostafijur.rahman, radum}@utexas.edu

Abstract

Transformers have shown great promise in medical image segmentation due to their ability to capture long-range dependencies through self-attention. However, they lack the ability to learn the local (contextual) relations among pixels. Previous works try to overcome this problem by embedding convolutional layers either in the encoder or decoder modules of transformers thus ending up sometimes with inconsistent features. To address this issue, we propose a novel attention-based decoder, namely CASCaded Attention DEcoder (CASCADE), which leverages the multi-scale features of hierarchical vision transformers. CASCADE consists of i) an attention gate which fuses features with skip connections and ii) a convolutional attention module that enhances the long-range and local context by suppressing background information. We use a multi-stage feature and loss aggregation framework due to their faster convergence and better performance. Our experiments demonstrate that transformers with CASCADE significantly outperform state-of-the-art CNN- and transformer-based approaches, obtaining up to 5.07% and 6.16% improvements in DICE and mIoU scores, respectively. CASCADE opens new ways of designing better attention-based decoders.

1. Introduction

Medical image segmentation is one of the critical steps in pre-treatment diagnoses, treatment planning, and post-treatment assessments of various diseases. Medical image segmentation can be formulated as a dense prediction problem which performs pixel-wise classification and creates segmentation maps of lesions or organs. Convolutional neural networks (CNNs) have been widely used for medical image segmentation tasks [24, 37, 15, 22, 23, 10]. Specifically, UNet [24] has shown remarkable performance in medical image segmentation due to producing high-resolution segmentation maps aggregating multi-stage features using skip connections. Due to the sophisticated encoder-decoder architecture of UNet, a few variants of UNet, such as UNet++ [37], UNet 3+ [15], DC-UNet [22] have demonstrated impressive performance in medical image segmentation. Despite the satisfactory performance of CNN-based methods, they have limitations in learning the long-range dependencies among pixels due to the spatial context of the convolution operation [2]. To overcome this limitation, some works [23, 6, 10] incorporate attention modules in their architectures to enhance the feature map for better pixel-level classification of medical images. Although these attention-based methods achieve improved performance (due to capturing salient features), they still suffer from capturing insufficient long-range dependencies.

The recent progress in vision transformers [9] overcomes the above limitation in capturing long-range dependencies, particularly for medical image segmentation [3, 2, 8, 30]. Transformers rely on an attention-based network architecture; they were first introduced for sequence-to-sequence prediction in natural language processing (NLP) [28]. Transformers use self-attention to learn correlations among all the input tokens that enable them to capture long-range dependencies. Following the success of transformers in NLP, the vision transformer [9] divides an image into non-overlapping patches which are fed into the transformer module with positional embeddings. More recently, hierarchical vision transformers, such as Swin transformer [20] with window-based attention and pyramid vision transformer (PVT) [31] with spatial reduction attention have been introduced to reduce the computational costs. These hierarchical vision transformers are effective for medical image segmentation tasks [2, 8, 30]. However, the self-attention used in transformers limits their ability to learn local (contextual) relations among pixels [7, 16].

Recently, SegFormer [35], UFormer [33] and PVTv2 [32] try to overcome this limitation by embedding convolution layers in transformers. Although these architectures can partly learn the local (contextual) relations among pixels, they i) have limited discrimination ability due to embedding convolution layer directly between fully-connected layers of the feed-forward network, and ii) do not properly aggregate the multi-stage features generated by the hierarchical encoder. Considering these issues, we introduce a novel CASCaded Attention DEcoder (CASCADE) which

leverages the hierarchical representation of vision transformers. CASCADE fuses (with skip connections) and refines features using attention gates (AGs) and convolutional attention modules (CAMs), respectively. Due to using hierarchical transformers as a backbone network and aggregating multi-stage features using attention-based convolutional modules, CASCADE captures both global and local (contextual) relationships among pixels. Our contributions are summarized as follows:

- **Novel Network Architecture:** We introduce a novel hierarchical cascaded attention-based decoder (CASCADE) for 2D medical image segmentation which takes advantage of the multi-stage feature representation of vision transformers while learning multiscale and multiresolution spatial representations. We build our decoder using a novel convolutional attention module which suppresses unnecessary information. Additionally, we incorporate skip connections with attention-gated fusion which also suppresses irrelevant regions and highlights salient features. To the best of our knowledge, we are the first to propose this type of decoder for medical image segmentation.
- **Multi-stage Loss Optimization and Feature Aggregation:** We aggregate and optimize multiple losses from different stages of the hierarchical decoder. Our empirical analysis shows that multistage loss enables faster convergence of models accuracy and improves decoder performance. We also produce the final segmentation map incorporating multi-resolution features which puts more confidence on salient features.
- **Versatile and Improved Performance:** We empirically show that CASCADE can be used with any hierarchical vision encoder (e.g., PVT [32], TransUNet [3]) while significantly improving the performance of 2D medical image segmentation. When compared against multiple baselines, CASCADE produces new state-of-the-art (SOTA) results on ACDC, Synapse multi-organ, and Polyp segmentation benchmarks.

2. Related Work

We divide the related work into three parts, i.e., vision transformers, attention mechanisms, and medical image segmentation; these are described next.

2.1. Vision transformers

Dosovitskiy et al. [9] first introduce the vision transformer (ViT), which achieves outstanding performance due to capturing long-range dependencies among the pixels. While early vision transformers were computationally expensive, recent works have tried to further enhance ViT in several ways. Touvron et al. [27] introduce DeiT which tries to minimize the computational cost for ViT using data-efficient training strategies. Liu et al. [20] develop the Swin transformer using a sliding window attention mechanism. In SegFormer, Xie et al. [35] introduce a Mix-FFN module for encoding better positional information and an efficient self-attention mechanism for reducing the computational costs. SegFormer is also a hierarchical transformer where image patches are merged to preserve the local continuity among patches. Wang et al. [31] propose a pyramid vision transformer (PVT) where the computational cost is reduced using a spatial reduction attention mechanism. In PVTv2, Wang et al. [32] improve the performance of PVT by incorporating a linear complexity attention layer, an overlapping patch embedding, and a convolutional feed-forward network.

Although vision transformers have shown excellent promise, their performance is limited when trained on small datasets. This limitation makes the transformers difficult to train for applications like medical image segmentation with small amounts of data. We try to overcome this limitation by using pretrained transformer backbones in large datasets (like ImageNet); indeed, previous studies [8, 30] have found that pretrained transformer weights on other non-medical large datasets boost the performance of medical image segmentation tasks.

2.2. Attention mechanisms

Oktay et al. [23] introduce a low-cost attention gate module for U-shaped architectures to fuse features with skip-connections; this helps the model focus on the relevant information in the image. Chen et al. [6] propose a reverse attention module to explore the missing detail information which results in high resolution and accurate outputs. Hu et al. [14] introduce a squeeze-and-excitation block using global average-pooled features to compute channel attention; this identifies the important feature maps for learning and then enhances them. Although channel attention can identify which feature map to focus on, it lacks the ability to identify where to focus. To supplement the channel attention block, Chen et al. [4] propose a spatial attention block to better focus on a feature map. Woo et al. [34] introduce a convolutional block attention module (CBAM) utilizing both channel and spatial attention to capture where and on which feature to focus in a feature map. Their experiments show that channel attention followed by spatial attention produces the best results.

Due to the additive advantage of CBAM with negligible overhead, we incorporate channel attention followed by spatial attention in our CAM. The CAM differs from CBAM in the design of the block itself and in how the blocks are used. Firstly, our CAM consists of channel attention, spatial attention, and a convolutional block, while CBAM consists of only channel attention and spatial at-

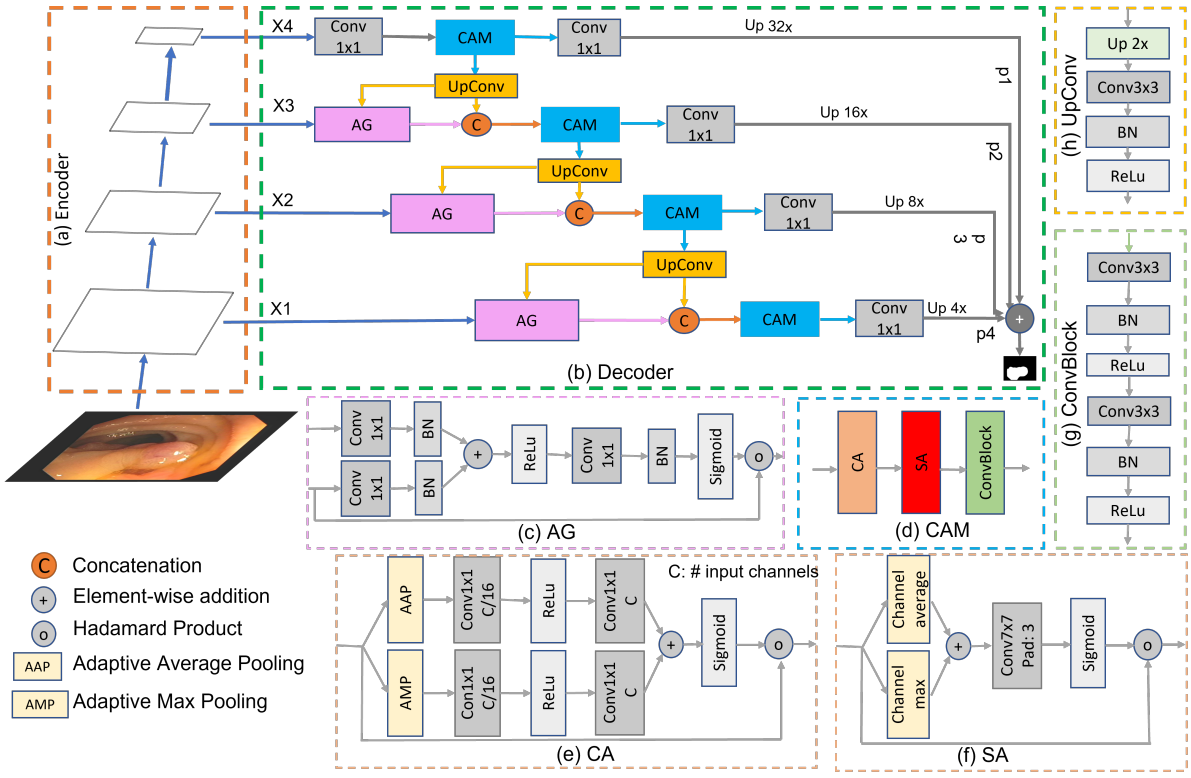

Figure 1. PVT-CASCADE network architecture. (a) PVTv2-b2 Encoder backbone with four stages, (b) CASCADE decoder, (c) Attention gate (AG), (d) Convolutional attention module (CAM), (e) Channel attention (CA), (f) Spatial attention (SA), (g) ConvBlock, (h) UpConv. X1, X2, X3, and X4 are the output features of the four stages of hierarchical encoder backbones. p1, p2, p3, and p4 are output feature maps from four stages of our decoder.

tention. Secondly, CBAM is placed in each convolutional block of both encoder and decoder, while the CAM module appears only in the decoder.

2.3. Medical image segmentation

Medical image segmentation is a dense prediction task that classifies the pixels of organs or lesions in a given medical image (e.g., CT, MRI, endoscopy, OCT, etc.) [3, 8]. UNet [24] and its variants [37, 15, 22, 23] are widely used in medical image segmentation tasks because of their better performance and sophisticated architecture. UNet [24] is an encoder−decoder architecture where features from the encoder are aggregated with upsampled features of the decoder using skip connections to produce high-resolution segmentation maps. Zhou et al. [37] introduce UNet++ where the encoder-decoder sub-networks are linked using nested and dense skip connections. Huang et al. [15] propose UNet 3+ utilizing full-scale skip connections including intra-connections among the decoder blocks. Lou et al. [22] introduce a dual channel UNet (DC−UNet) architecture that utilizes the multi-resolution convolution block and residual path in skip connections. Following the progress of computer vision, the ResNet architecture [13] has been generally adopted as the backbone for medical image segmentation. The pyramid pooling and dilated convolution [5] are also used for lesion and organ segmentation [12, 11].

Nowadays, transformer-based methods have also shown great success in medical image segmentation [3, 2, 19, 8, 30]. Chen et al. [3] proposed TransUNet which uses a hybrid CNN- transformer encoder to capture long-range dependencies and a cascaded CNN upsampler as a decoder to capture local contextual relations among pixels. In contrast, we propose a new attention-based cascaded decoder which shows a significant performance boost when used on top of the encoder. Li et al. [19] introduce TFCNs by combining transformer and fully convolutional DenseNet to propagate semantic features and filter out non-semantic features. Cao et al. [2] proposed Swin-Unet, which is a pure transformer architecture based on Swin transformer [20]. Swin-Unet uses transformers in both the encoder and decoder, which does not lead to performance improvement.

Recent studies incorporate different attention mechanisms with CNN [23, 10, 36] and transformer-based architectures [8, 30] for medical image segmentation. Fan et al. [10] adopt reverse attention [6] for polyp segmentation. Zhang et al. [36] utilize squeeze-and-excitation attention

[14] for segmenting vessels in retina images. Dong et al. [8] adopt a CBAM [34] attention block in their decoder; they use the CBAM block only with the low-level features from the first layer of the PVTv2 which limits the ability to refine all multi-stage features. In contrast, we incorporate the AG to fuse features with skip connection and use a CAM module in all of our decoder blocks.

3. Method

We first introduce the transformer backbones and our proposed CASCADE decoder. We then describe two different transformer-based architectures (TransCASCADE and PVT-CASCADE) incorporating our proposed decoder.

3.1. Transformer backbones

To ensure enough generalization and multi-scale feature processing abilities for medical image segmentation, we use the pyramid transformer, as well as the hybrid CNN-transformer (instead of only CNN) as the encoder. Specifically, we adopt the encoder design of PVTv2 [32] (Figure 1(a)) and TransUNet [3]. PVTv2 uses the convolution operation instead of the patch embedding module of the traditional transformer to consistently capture the spatial information. TransUNet utilizes a transformer on top of CNN to capture both global and spatial relationships among features. Our proposed decoder is flexible and easy to adopt with other hierarchical backbone networks.

3.2. CASCaded Attention DEcoder (CASCADE)

Existing transformer-based models have limited (local) contextual information processing ability among pixels. As a result, the transformer-based model faces difficulties in locating the more discriminating local features. To address this issue, we propose a novel attention-based cascaded multi-stage feature aggregation decoder, CASCADE, for pyramid features.

As shown in Figure 1(b), CASCADE consists of the UpConv block to upsample the features, the AG for cascaded feature fusion, and the CAM to robustly enhance the feature maps. We have four CAM blocks for the four stages of pyramid features from the encoder backbone and three AGs for three skip connections. To aggregate the multi-scale features, we first combine the upsampled features from the previous decoder block with the features from the skip connections using AG. Then, we concatenate the fused features with the upsampled features from the previous layer. Afterward, we process the concatenated features using our CAM module for pixel grouping and suppressing background information using both channel and spatial attention. Finally, we send the output from each CAM layer to a prediction head and aggregate four different predictions to produce the final segmentation map.

3.2.1 Attention gate (AG)

AGs are used to progressively suppress features in irrelevant background regions by adopting a grid-attention technique where the gating signal is based on the spatial information of the image [23]. More specifically, the gating signal used to aggregate each skip connection fuses the multi-stage features which increase the spatial resolution of the query signal. Like Attention UNet [23], we use additive attention to obtain the gating coefficient because of its better performance compared to multiplicative attention. The additive attention gate $AG(\cdot)$ is given in Equations 1 and 2:

$$q_{att}(g,x) = \sigma_1(BN(C_g(g) + BN(C_x(x)))) \quad (1)$$

$$AG(g,x) = x * \sigma_2(BN(C(q_{att}(g,x)))) \quad (2)$$

where $\sigma_1(\cdot)$ and $\sigma_2(\cdot)$ correspond to ReLU and Sigmoid activation function, respectively. $C_g(\cdot)$, $C_x(\cdot)$, and $C(\cdot)$ represent channel-wise 1×1 convolution operation. $BN(\cdot)$ is the batch normalization operation. g and x are the upsampled and skip connection features, respectively.

3.2.2 Convolutional attention module (CAM)

We use the convolutional attention modules to refine the feature maps. CAM consists of a channel attention [14] ($CA(\cdot)$), a spatial attention [4] ($SA(\cdot)$), and a convolutional block ($ConvBlock$) as in Equation 3:

$$CAM(x) = ConvBlock(SA(CA(x))) \quad (3)$$

where x is the input tensor and $CAM(\cdot)$ represents the convolutional attention module.

Channel Attention (CA): Channel attention identifies *which feature maps* to focus on (and then refine them). The channel attention $CA(\cdot)$ is defined using Equation 4:

$$CA(x) = \sigma_2(C_2(\sigma_1(C_1(P_m(x)))) + C_2(\sigma_1(C_1(P_a(x))))) \circledast x \quad (4)$$

where $\sigma_2(\cdot)$ is the Sigmoid activation. $P_m(\cdot)$ and $P_a(\cdot)$ denote adaptive maximum pooling and adaptive average pooling, respectively. $C_1(\cdot)$ is a convolutional layer with 1×1 kernel size to reduce the channel dimension 16 times. σ_1 is a ReLU activation layer and $C_2(\cdot)$ is another convolutional layer to recover the original channel dimension. $\circledast$ is the Hadamard product.

Spatial Attention (SA): Spatial attention determines *where* to focus in a feature map and then enhances those features. The spatial attention $SA(\cdot)$ is given in Equation 5:

$$SA(x) = \sigma(C(C_m(x) + C_a(x))) \circledast x \quad (5)$$

where $\sigma(\cdot)$ is a Sigmoid activation function. $C_m(\cdot)$ and $C_a(\cdot)$ represent the maximum and average values obtained along

the channel dimension, respectively. $C(\cdot)$ is a 7×7 convolutional layer with padding 3 to enhance spatial contextual information (as in [8]).

ConvBlock: The ConvBlock is used to further enhance the features generated using our CA and SA operations. ConvBlock consists of two 3×3 convolution layers each followed by a batch normalization layer and a ReLU activation layer. $ConvBlock(\cdot)$ is formulated as Equation 6:

$$ConvBlock(x) = \sigma(BN(C(\sigma(BN(C(x)))))) \quad (6)$$

where σ is the ReLU activation layer, $BN(\cdot)$ represents batch normalization, and $C(\cdot)$ is a 3×3 convolution layer.

3.2.3 UpConv

UpConv progressively upsamples the features of the current layer to match the dimension to the next skip connection. Each UpConv layer consists of an UpSampling $UP(\cdot)$ with scale-factor 2, a 3×3 convolution $Conv(\cdot)$, a batch normalization $BN(\cdot)$, and a ReLU activation layers. The $UpConv(\cdot)$ can be formulated as Equation 7:

$$UpConv(x) = ReLU(BN(Conv(Up(x)))) \quad (7)$$

3.3. Multi-stage loss and feature aggregation

We use four prediction heads for the four stages of hierarchical encoders. We compute the final prediction map using additive aggregation as in Equation 8:

$$output = w \times p1 + x \times p2 + y \times p3 + z \times p4 \quad (8)$$

where $p1$, $p2$, $p3$, and $p4$ are the feature maps of four prediction heads, and w, x, y, and z are the weights for individual prediction heads. In our experiments, we set all w, x, y, and z to 1.0. We get the final prediction output by applying the Sigmoid activation for binary segmentation and Softmax activation for multi-class segmentation.

However, we compute the loss for each prediction head separately and then aggregate them using equation 9:

$$loss = \alpha \times loss_{p1} + \beta \times loss_{p2} + \gamma \times loss_{p3} + \zeta \times loss_{p4} \quad (9)$$

where $loss_{p1}$, $loss_{p2}$, $loss_{p3}$, and $loss_{p4}$ are the losses for four different prediction heads, and α, β, γ, and ζ are the weights for the loss of individual prediction heads. In our experiments, we set all α, β, γ, and ζ to 1.0.

3.4. Overall architecture

We utilize two different hierarchical backbone encoder networks such as PVTv2 [32] and TransUNet [3] for our experiments. In the case of TransUNet, we only use their hybrid CNN-transformer backbone encoder network. By utilizing the PVTv2-b2 (Standard) encoder, we create the PVT-CASCADE architecture. To adopt PVTv2-b2, we first extract the features (X1, X2, X3, and X4) from four layers and feed them (i.e., X4 in the upsample path and X3, X2, X1 in the skip connections) into our CASCADE decoder as shown in Figure 1(a-b). Then our CASCADE decoder processes them and produces four prediction feature maps for the four stages of the encoder network. Afterward, we aggregate the prediction feature maps using Equation 8 to produce the final prediction feature map. Finally, we apply the Sigmoid activation for binary segmentation and Softmax for multi-class segmentation tasks. Besides, we introduce TransCASCADE architecture by adopting the backbone encoder network of TransUNet. We follow similar steps in our TransCASCADE architecture. These two architectures achieve SOTA performance on Synapse multi-organ segmentation, ACDC, and several polyp segmentation benchmarks. Details are given in the experimental section.

4. Experiments

In this section, we first compare the results of our proposed CASCADE decoder with SOTA methods to demonstrate the superiority of our proposed method. Then, we carry out ablation studies to evaluate the effectiveness of our CASCADE decoder.

4.1. Datasets and evaluation metrics

Synapse multi-organ dataset. The Synapse multi-organ dataset[1] has 30 abdominal CT scans with 3779 axial contrast-enhanced abdominal CT images. Each CT scan consists of 85-198 slices of 512×512 pixels, with a voxel spatial resolution of ([0:54-0:54] $\times$ [0:98-0:98]$\times$[2:5-5:0])mm^3. Following TransUNet [3], we divide the dataset randomly into 18 scans for training (2212 axial slices), and 12 for validation. We segment 8 anatomical structures, such as aorta, gallbladder (GB), left kidney (KL), right kidney (KR), liver, pancreas (PC), spleen (SP), and stomach (SM). **ACDC dataset.** The ACDC dataset[2] consists of 100 cardiac MRI scans collected from different patients. Each scan contains three organs, right ventricle (RV), left ventricle (LV), and myocardium (Myo). Following TransUNet [3], we use 70 cases (1930 axial slices) for training, 10 for validation, and 20 for testing. **Polyp datasets.** CVC-ClinicDB [1] contains 612 images, which are extracted from 31 colonoscopy videos. Kvasir includes 1,000 polyp images, which are collected from the polyp class in the Kvasir-SEG dataset [17]. Following the settings in PraNet [10], we adopt the same 900 and 548 images from CVC-ClinicDB and Kvasir datasets as the training set, and the remaining 64 and 100 images are employed as the respective testsets. To evaluate the generalization performance, we test the model on three unseen datasets, namely EndoScene [29], ColonDB [26], and ETIS-LaribDB [25]. These three testsets are collected

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

[2] https://www.creatis.insa-lyon.fr/Challenge/acdc/

Architectures	Average				Aorta	GB	KL	KR	Liver	PC	SP	SM
	DICE↑	HD95↓	mIoU↑	ASD↓								
UNet [24]	70.11	44.69	59.39	14.41	84.00	56.70	72.41	62.64	86.98	48.73	81.48	67.96
AttnUNet [23]	71.70	34.47	61.38	10.00	82.61	61.94	76.07	70.42	87.54	46.70	80.67	67.66
R50+UNet [3]	74.68	36.87	–	–	84.18	62.84	79.19	71.29	93.35	48.23	84.41	73.92
R50+AttnUNet [3]	75.57	36.97	–	–	55.92	63.91	79.20	72.71	93.56	49.37	87.19	74.95
SSFormerPVT [30]	78.01	25.72	67.23	4.56	82.78	63.74	80.72	78.11	93.53	61.53	87.07	76.61
PolypPVT [8]	78.08	25.61	67.43	4.89	82.34	66.14	81.21	73.78	94.37	59.34	88.05	79.4
TFCNs [19]	75.63	30.63	64.69	5.29	**88.23**	59.18	80.99	73.12	92.02	54.24	88.36	68.9
TransUNet [3]	77.61	26.9	67.32	4.66	86.56	60.43	80.54	78.53	94.33	58.47	87.06	75
SwinUNet [2]	77.58	27.32	66.88	4.7	81.76	65.95	82.32	79.22	93.73	53.81	88.04	75.79
PVT-CASCADE (Ours)	81.06	20.23	70.88	3.61	83.01	**70.59**	82.23	80.37	94.08	64.43	90.1	**83.69**
TransCASCADE (Ours)	**82.68**	**17.34**	**73.48**	**2.83**	86.63	68.48	**87.66**	**84.56**	**94.43**	**65.33**	**90.79**	83.52
Improve TransUNet	5.07	9.56	6.16	1.83	0.07	8.05	7.12	6.03	0.1	6.86	3.73	8.52

Table 1. Results of Synapse multi-organ segmentation. Only DICE scores are reported for individual organs. R50+UNet and R50+AttnUNet adopt a pre-trained ResNet50 backbone network. We reproduce the results of UNet, AttnUNet, SSFormerPVT, Polyp-PVT, TFCNs, TransUNet, and SwinUNet with the default experimental settings of TransUNet. ↑ denotes higher the better, ↓ denotes lower the better. All CASCADE results are averaged over five runs. The best results are in bold.

Architectures	Avg DICE	RV	Myo	LV
R50+UNet [3]	87.55	87.10	80.63	94.92
R50+AttnUNet [3]	86.75	87.58	79.20	93.47
ViT+CUP [3]	81.45	81.46	70.71	92.18
R50+ViT+CUP [3]	87.57	86.07	81.88	94.75
TransUNet [3]	89.71	86.67	87.27	95.18
SwinUNet [2]	88.07	85.77	84.42	94.03
PVT-CASCADE (Ours)	91.46	88.9	89.97	95.50
TransCASCADE (Ours)	**91.63**	**89.14**	**90.25**	**95.50**

Table 2. Results on ACDC dataset. DICE scores are reported for individual organs. We reproduce the results of SwinUNet. All CASCADE results are averaged over five runs.

from different medical centers. In other words, the data from these three sources are not used to train our model. EndoScene, ColonDB, and ETIS-LaribDB contain 60, 380, and 196 images, respectively.

Evaluation metrics. We use DICE, mean intersection over union (mIoU), 95% Hausdorff Distance (95HD), and Average surface distance (ASD) as the evaluation metrics in our experiments on Synapse Multi-organ dataset. Following existing methods, we use only DICE scores for the ACDC dataset. For the experiments on polyp segmentation, we use DICE and mIoU as the evaluation metrics.

4.2. Implementation details

All our experiments are implemented in Pytorch 1.11.0. We train all models on a single NVIDIA RTX A6000 GPU with 48GB of memory. We utilize the pre-trained weights on ImageNet for backbone networks. We use AdamW optimizer [21] with learning rate and weight decay of 1e-4.

Synapse Multi-organ dataset. Following TransUNet [3], we use a batch size of 24 and train each model maximum of 150 epochs. We use the input resolution and patch size P as 224×224 and 16, respectively. We employ random flipping and rotation for data augmentation. The combined cross-entropy and DICE loss are used as the loss function. **ACDC dataset.** For the ACDC dataset, we train each model for a maximum of 150 epochs with a batch size of 12. We set the input resolution and patch size P as 224×224 and 16, respectively. Random flipping and rotation are applied for data augmentation. We use the combined cross-entropy and DICE loss function. **Polyp datasets.** Following Polyp-PVT [8], we use a batch size of 16 and train each model maximum of 100 epochs. We resize the image to 352×352 and use a similar multi-scale $\{0.75, 1.0, 1.25\}$ training strategy with a gradient clip limit of 0.5 as Polyp-PVT. We use the combined weighted IoU and weighted BCE loss function.

4.3. Results

We compare our architectures (i.e., PVT-CASCADE and TransCASCADE) with SOTA CNN and transformer-based segmentation methods on Synapse Multi-organ, ACDC, and Polyp (i.e., Endoscene [29], CVC-ClinicDB [1], Kvasir [17], ColonDB [26], ETIS-LaribDB [25]) datasets. More results are available in the supplementary materials.

4.3.1 Experimental results on Synapse dataset

We demonstrate the performance of different CNN and transformer-based methods in Table 1. As shown in Table 1, transformer-based models have superior performance compared to CNN-based models. Our proposed CASCADE decoder improves the average DICE, mIoU, and HD95

Architectures	EndoScene		CVC-ClinicDB		Kvasir		ColonDB		ETIS-LaribDB	
	DICE	mIoU	DICE	mIoU	DICE	mIoU	DICE	mIoU	DICE	mIoU
UNet [24]	71.0	62.7	82.3	75.5	81.8	74.6	51.2	44.4	39.8	33.5
UNet++ [37]	70.7	62.4	79.4	72.9	82.1	74.3	48.3	41.0	40.1	34.4
PraNet [10]	87.1	79.7	89.9	84.9	89.8	84.0	71.2	64.0	62.8	56.7
UACANet-L [18]	88.21	80.84	91.07	86.7	90.83	85.95	72.57	65.41	63.89	56.87
SSFormerPVT [30]	89.46	82.68	92.88	88.27	91.11	86.01	79.34	70.63	78.03	70.1
PolypPVT [8]	88.71	81.89	93.08	88.28	91.23	86.3	80.75	71.85	78.67	70.97
PVT-CASCADE (Ours)	**90.47**	**83.79**	**94.34**	**89.98**	**92.58**	**87.76**	**82.54**	**74.53**	**80.07**	**72.58**
Improve SSFormerPVT	1.01	1.11	1.46	1.71	1.47	1.75	3.2	3.9	2.04	2.48
Improve PolypPVT	1.76	1.9	1.26	1.7	1.35	1.46	1.79	2.68	1.4	1.61

Table 3. Results on polyp segmentation datasets. Training on combined Kvasir [17] and CVC-ClinicDB [1] trainset. The results of UNet, UNet++ and PraNet are taken from [10]. We reproduce the results of PolypPVT, SSFormerPVT, and UACANet using their public source code with default settings. All PVT-CASCADE results are averaged over five runs. The best results are in bold.

Components			EndoScene		CVC-ClinicDB		Kvasir		ColonDB		ETIS-LaribDB	
Cascaded	AG	CAM	DICE	mIoU	DICE	mIoU	DICE	mIoU	DICE	mIoU	DICE	mIoU
No	No	No	88.41	81.47	91.82	87.12	91.09	86.13	77.86	69.43	77.04	68.47
Yes	No	No	89.11	82.32	93.54	88.95	91.98	87.05	81.30	73.21	78.16	69.97
Yes	Yes	No	89.25	82.57	93.61	89.04	92.45	87.57	81.72	73.67	79.27	71.38
Yes	No	Yes	89.39	82.79	93.88	89.31	92.20	87.28	82.11	74.09	79.57	71.73
Yes	Yes	Yes	**90.47**	**83.79**	**94.34**	**89.98**	**92.58**	**87.76**	**82.54**	**74.53**	**80.07**	**72.58**

Table 4. Quantitative results of different components of CASCADE with PVTv2-b2 backbone. Training on combined Kvasir and CVC-ClinicDB trainset and testing on five testsets (i.e., Endoscene, CVC-ClinicDB, Kvasir, ColonDB, ETIS-LaribDB). All results are averaged over five runs. The best results are in bold.

scores of TransUNet by 5.07%, 6.16%, and 9.56, respectively. TransCASCADE achieves the best average DICE (82.67%), mIoU (73.48%), HD95 (17.34), and ASD (2.83) scores among all other methods. Moreover, TransCASCADE demonstrates significant performance improvements in both small and large organ segmentation. For small organs, 8.05%, 7.12%, and 6.03% improvements in gallbladder, left kidney, and right kidney, respectively. For large organs, 8.52%, 6.86%, and 3.73% improvements in stomach, pancreas, and spleen, respectively. This is because CASCADE captures both long-range dependencies and local contextual relations among pixels. Due to using attention, CASCADE better refines the feature maps and produces stronger feature representations than other decoders. The lower HD95 scores indicate that our CASCADE decoder can better locate the boundary of organs.

4.3.2 Experimental results on ACDC dataset

We evaluate the performance of our method on the MRI images of the ACDC dataset. Table 2 presents the average DICE scores of our PVT-CASCADE and TransCASCADE along with other SOTA methods. Our TransCASCADE achieves the highest average DICE score of 91.63% improving about 2% over TransUNet though we share the same encoder. Our PVT-CASCADE gains 91.46% DICE score which is also better than all other methods. Besides, our TransCASCADE has an improvement of 2.5 - 3% DICE score in challenging organs RV and Myo segmentation.

4.3.3 Experimental results on Polyp datasets

We evaluate the performance and generalizability of our CASCADE decoder on five different polyp segmentation test sets among which three are completely unseen datasets collected from different labs. Table 3 displays the DICE and mIoU scores of SOTA methods along with our CASCADE decoder. From Table 3, we can show that CASCADE significantly outperforms all other methods achieving 2.04 - 3.2% and 2.5 - 3.9% improvement in DICE and mIoU scores in unseen test sets over the previous best model using the same pre-trained transformer backbone. It is noteworthy that CASCADE outperforms the best CNN-based model UACANet by a large margin on unseen datasets (i.e., 16.2% and 10% DICE score improvement in ETIS-LaribDB and ColonDB, respectively). Therefore, we can conclude that due to using transformers as a backbone network and our attention-based CASCADE decoder, PVT-CASCADE inherits the merits of transformers, CNNs, and attentions which makes them highly generalizable for unseen datasets.

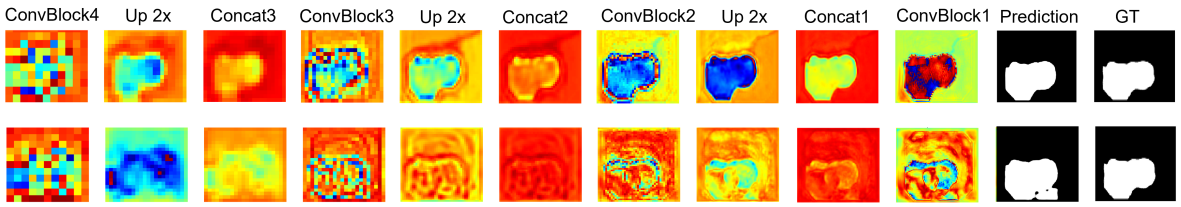

Figure 2. CASCADE vs. Cascaded Upsampler (CUP) features. First and second rows present CASCADE and CUP features, respectively. We put only the similar layers feature for our CASCADE decoder for fair comparisons. Layers are numbered based on their corresponding transformer layer number. In both cases, we use the ImageNet pretrained PVTv2-b2 backbone as the encoder.

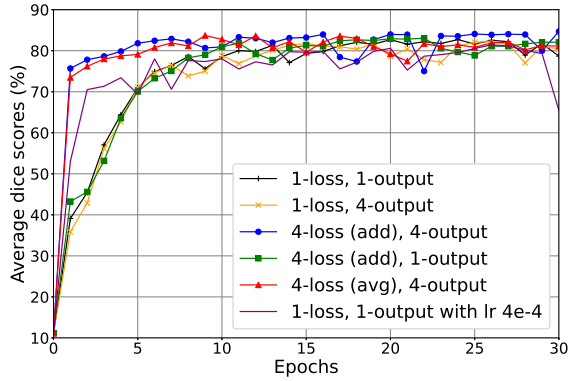

Figure 3. Multi-stage loss and output aggregation vs. single loss and output. We plot the average DICE scores of five testsets (i.e., Endoscene, CVC-ClinicDB, Kvasir, ColonDB, ETIS-LaribDB) vs. # epochs in six different loss and output aggregation settings.

4.4. Ablation study

Effective enhancement/refinement of features. We visualize the features of our CASCADE, as well as Cascaded Upsampler (CUP) [3] in Figure 2. We compute the average of all channels in the feature map and then produce the heatmap using OpenCV-Python. It is evident from Figure 2 that the attention mechanism used in our CASCADE helps identify, enhance, and group the pixels better than CUP.

Effectiveness of different parts of CASCADE. We carry out ablation studies on the Polyp datasets to evaluate the effectiveness of the different components of our proposed CASCADE decoder. We use the same PVTv2-b2 backbone pre-trained on ImageNet and the same experimental settings for polyp datasets in all experiments. We remove different modules such as AGs and CAM from the CASCADE decoder and compare the results. It is evident from the Table 4 that the cascaded structure of the decoder improves performance over the non-cascaded decoder. AG and CAM modules also help improve performance. However, the use of both AG and CAM modules produces the best performance in all test datasets.

Faster learning of multi-stage loss and output fusions. We add the loss and output from four stages of our CASCADE decoder to get the overall loss and final segmentation map. Figure 3 plots the average DICE score across five datasets for each epoch. The graph contains six different loss and output aggregation settings such as "1-loss, 1-output", "1-loss, 4-output", "4-loss, 1-output", "4-loss (add), 4 output", "4-loss (avg), 4 output", and "1-loss, 1 output with learning rate 4e-4". It is evident from the graph that "4-loss (add), 4 output" and "4-loss (avg), 4 output" achieve 74 - 75% DICE scores in the first epoch, and these settings gain more than 82% DICE score within 5 epochs. On the other hand, other losses and output aggregations have a DICE score of around 35 - 53%, and these settings achieve 71% DICE score within 5 epochs. We can also see from the graph that "4-loss (add), 4 output" shows the best performance, achieving 84.67% average DICE score. Therefore, we can conclude that aggregation of multi-stage loss and output leverages multi-scale features that help to produce accurate and high-resolution segmentation outputs.

5. Conclusion

In this paper, we have proposed a novel attention-based decoder for hierarchical feature aggregation, which has robust generalization and learning ability; these are crucial for medical image segmentation. We believe that CASCADE has great potential to improve deep learning performance in other medical image segmentation tasks. Moreover, experiments demonstrate that CASCADE effectively enhances transformer features and incorporates spatial relationships among pixels (e.g., improves baseline TransUNet by 5.07% DICE and 6.16% mIoU in Synapse Multi-organ segmentation). Experimental results demonstrate that CASCADE can locate the organs or lesions well (e.g., improves HD95 score by 9.56) due to using attention in the decoding process. Therefore, our decoder can be further used to enhance the transformer feature for general computer vision and highly generalizable medical applications.

Acknowledgment

This work is supported, in part, by NSF grant CNS 2007284.

References

[1] Jorge Bernal, F Javier Sánchez, Gloria Fernández-Esparrach, Debora Gil, Cristina Rodríguez, and Fernando Vilariño. Wm-dova maps for accurate polyp highlighting in colonoscopy: Validation vs. saliency maps from physicians. *Computerized medical imaging and graphics*, 43:99–111, 2015.

[2] 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.

[3] 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.

[4] Long Chen, Hanwang Zhang, Jun Xiao, Liqiang Nie, Jian Shao, Wei Liu, and Tat-Seng Chua. Sca-cnn: Spatial and channel-wise attention in convolutional networks for image captioning. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5659–5667, 2017.

[5] 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.

[6] Shuhan Chen, Xiuli Tan, Ben Wang, and Xuelong Hu. Reverse attention for salient object detection. In *Proceedings of the European conference on computer vision (ECCV)*, pages 234–250, 2018.

[7] Xiangxiang Chu, Zhi Tian, Bo Zhang, Xinlong Wang, Xiaolin Wei, Huaxia Xia, and Chunhua Shen. Conditional positional encodings for vision transformers. *arXiv preprint arXiv:2102.10882*, 2021.

[8] Bo Dong, Wenhai Wang, Deng-Ping Fan, Jinpeng Li, Huazhu Fu, and Ling Shao. Polyp-pvt: Polyp segmentation with pyramid vision transformers. *arXiv preprint arXiv:2108.06932*, 2021.

[9] 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.

[10] Deng-Ping Fan, Ge-Peng Ji, Tao Zhou, Geng Chen, Huazhu Fu, Jianbing Shen, and Ling Shao. Pranet: Parallel reverse attention network for polyp segmentation. In *International conference on medical image computing and computer-assisted intervention*, pages 263–273. Springer, 2020.

[11] Shuanglang Feng, Heming Zhao, Fei Shi, Xuena Cheng, Meng Wang, Yuhui Ma, Dehui Xiang, Weifang Zhu, and Xinjian Chen. Cpfnet: Context pyramid fusion network for medical image segmentation. *IEEE transactions on medical imaging*, 39(10):3008–3018, 2020.

[12] Zaiwang Gu, Jun Cheng, Huazhu Fu, Kang Zhou, Huaying Hao, Yitian Zhao, Tianyang Zhang, Shenghua Gao, and Jiang Liu. Ce-net: Context encoder network for 2d medical image segmentation. *IEEE transactions on medical imaging*, 38(10):2281–2292, 2019.

[13] 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.

[14] Jie Hu, Li Shen, and Gang Sun. Squeeze-and-excitation networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7132–7141, 2018.

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

[16] Md Amirul Islam, Sen Jia, and Neil DB Bruce. How much position information do convolutional neural networks encode? *arXiv preprint arXiv:2001.08248*, 2020.

[17] Debesh Jha, Pia H Smedsrud, Michael A Riegler, Pål Halvorsen, Thomas de Lange, Dag Johansen, and Håvard D Johansen. Kvasir-seg: A segmented polyp dataset. In *International Conference on Multimedia Modeling*, pages 451–462. Springer, 2020.

[18] Taehun Kim, Hyemin Lee, and Daijin Kim. Uacanet: Uncertainty augmented context attention for polyp segmentation. In *Proceedings of the 29th ACM International Conference on Multimedia*, pages 2167–2175, 2021.

[19] Zihan Li, Dihan Li, Cangbai Xu, Weice Wang, Qingqi Hong, Qingde Li, and Jie Tian. Tfcns: A cnn-transformer hybrid network for medical image segmentation. *arXiv preprint arXiv:2207.03450*, 2022.

[20] 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.

[21] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017.

[22] Ange Lou, Shuyue Guan, and Murray Loew. Dc-unet: rethinking the u-net architecture with dual channel efficient cnn for medical image segmentation. In *Medical Imaging 2021: Image Processing*, volume 11596, pages 758–768. SPIE, 2021.

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

[24] 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*, pages 234–241. Springer, 2015.

[25] Juan Silva, Aymeric Histace, Olivier Romain, Xavier Dray, and Bertrand Granado. Toward embedded detection of polyps in wce images for early diagnosis of colorectal can-

cer. *International journal of computer assisted radiology and surgery*, 9(2):283–293, 2014.

[26] Nima Tajbakhsh, Suryakanth R Gurudu, and Jianming Liang. Automated polyp detection in colonoscopy videos using shape and context information. *IEEE transactions on medical imaging*, 35(2):630–644, 2015.

[27] Hugo Touvron, Matthieu Cord, Matthijs Douze, Francisco Massa, Alexandre Sablayrolles, and Hervé Jégou. Training data-efficient image transformers & distillation through attention. In *International Conference on Machine Learning*, pages 10347–10357. PMLR, 2021.

[28] 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.

[29] David Vázquez, Jorge Bernal, F Javier Sánchez, Gloria Fernández-Esparrach, Antonio M López, Adriana Romero, Michal Drozdzal, and Aaron Courville. A benchmark for endoluminal scene segmentation of colonoscopy images. *Journal of healthcare engineering*, 2017, 2017.

[30] Jinfeng Wang, Qiming Huang, Feilong Tang, Jia Meng, Jionglong Su, and Sifan Song. Stepwise feature fusion: Local guides global. *arXiv preprint arXiv:2203.03635*, 2022.

[31] Wenhai Wang, Enze Xie, Xiang Li, Deng-Ping Fan, Kaitao Song, Ding Liang, Tong Lu, Ping Luo, and Ling Shao. Pyramid vision transformer: A versatile backbone for dense prediction without convolutions. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 568–578, 2021.

[32] Wenhai Wang, Enze Xie, Xiang Li, Deng-Ping Fan, Kaitao Song, Ding Liang, Tong Lu, Ping Luo, and Ling Shao. Pvt v2: Improved baselines with pyramid vision transformer. *Computational Visual Media*, 8(3):415–424, 2022.

[33] Zhendong Wang, Xiaodong Cun, Jianmin Bao, Wengang Zhou, Jianzhuang Liu, and Houqiang Li. Uformer: A general u-shaped transformer for image restoration. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 17683–17693, 2022.

[34] Sanghyun Woo, Jongchan Park, Joon-Young Lee, and In So Kweon. Cbam: Convolutional block attention module. In *Proceedings of the European conference on computer vision (ECCV)*, pages 3–19, 2018.

[35] Enze Xie, Wenhai Wang, Zhiding Yu, Anima Anandkumar, Jose M Alvarez, and Ping Luo. Segformer: Simple and efficient design for semantic segmentation with transformers. *Advances in Neural Information Processing Systems*, 34:12077–12090, 2021.

[36] Zhijie Zhang, Huazhu Fu, Hang Dai, Jianbing Shen, Yanwei Pang, and Ling Shao. Et-net: A generic edge-attention guidance network for medical image segmentation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 442–450. Springer, 2019.

[37] Z Zhou, MMR Siddiquee, N Tajbakhsh, and J Liang. A nested u-net architecture for medical image segmentation. arxiv 2018. *arXiv preprint arXiv:1807.10165*.