Hierarchical Transformer: Bring Scale To Your Attention

Xiaoya Tang; Bodong Zhang; Tolga Tasdizen

Background Vision Transformer(ViT) is a pioneering work of adapting transformer from Nature Language Processing(NLP) to visual domain. It uncovered the huge potential of transformer-based models in image analysis, and led a wave of research on visual-oriented transformers.

- ViT follows the structure of the vanilla transformer's encoder. It tokenizes the image into 14x14 (varies with input size) equal-sized sequences by using a 16x16 grid[1].
- The self attention mechanism models the 196 patch embeddings as a fully-connected graph, giving the transformers the non-local ability of global perception.[2]

Multi-head Attention Multiple independent embeddings can be learned in multi-head attention, which has been proved to be beneficial to capture dependencies of various ranges in sequences. Each head represents a learnable linear projection of query, key and values, endowing the attention mechanism with the ability of modeling from different representation subspaces.

Hierarchical transformer A drawback of ViTs is their lack of inductive bias. This bias mainly includes locality and hierarchical representations, which are commonly observed in Convolutional Neural Networks (CNNs). On the other hand, visual entities are generally at different scales and it can vary substantially in images, while all embedded patches in vanilla ViT are at a fixed scale. The ignorance of scales may bring failure or non-improvement of transformer based models on vision tasks. There has been quite a bit of work trying to solve this problem with transformers. Here we introduce a novel method to bring needed inductive bias into vision transformer, namely, hierarchical transformer.

Our model includes three main ideas:

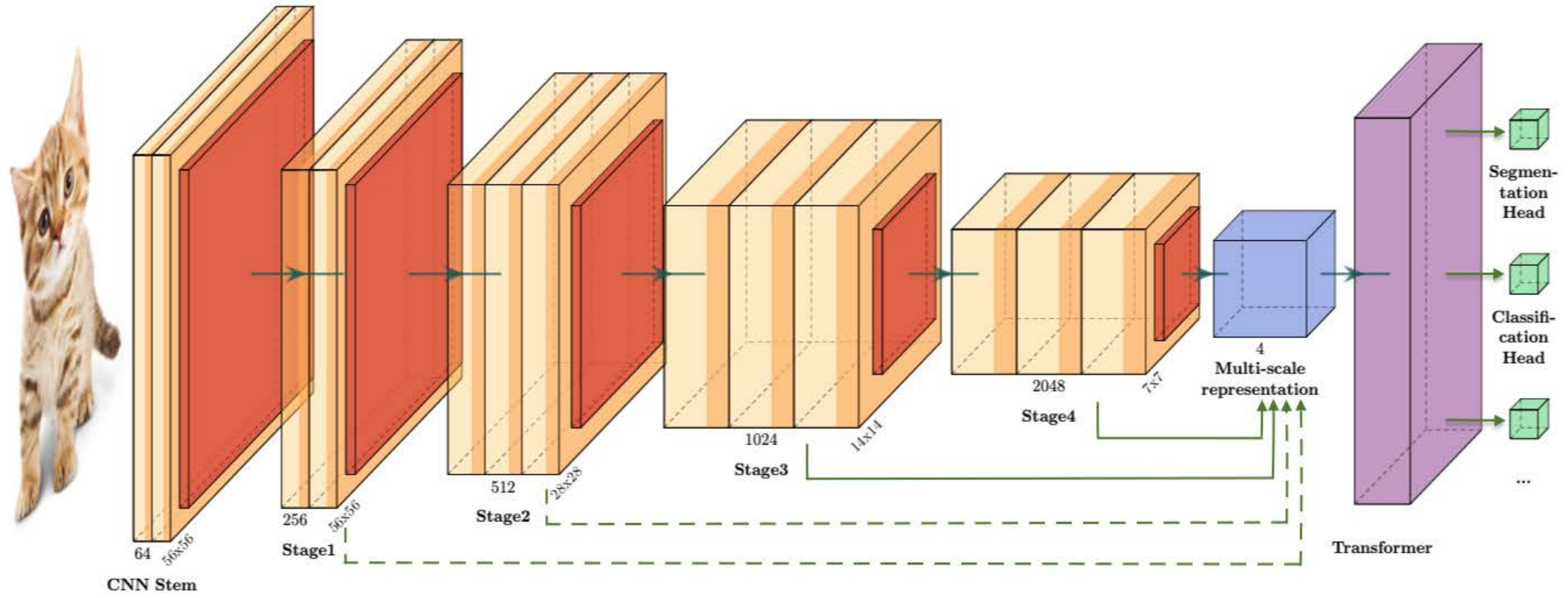

Figure 1 The framework of our Hierarchical transformer

- A CNN backbone is used to produce hierarchical representations for an image from different stages of it. Our novel patch tokenization enables leveraging hierarchical representations into multi-scale embeddings, instead of embeddings at a single scale.
- A multi-head attention mechanism on the dimension of scales before the global attention. We refer to this local attention as scale attention, which enables the model to capture multi-scale information needed for downstream tasks inside each patch.
- A scale token learns crucial information from channels, enhancing the ability to gather more multi-scale information.

The proposed model integrates inductive bias for ViT without auxiliary loss or additional training. It outperforms the CNN baseline and hybrid ViTs on small -sized datasets, which demonstrates it can better learn the needed information for downstream tasks.

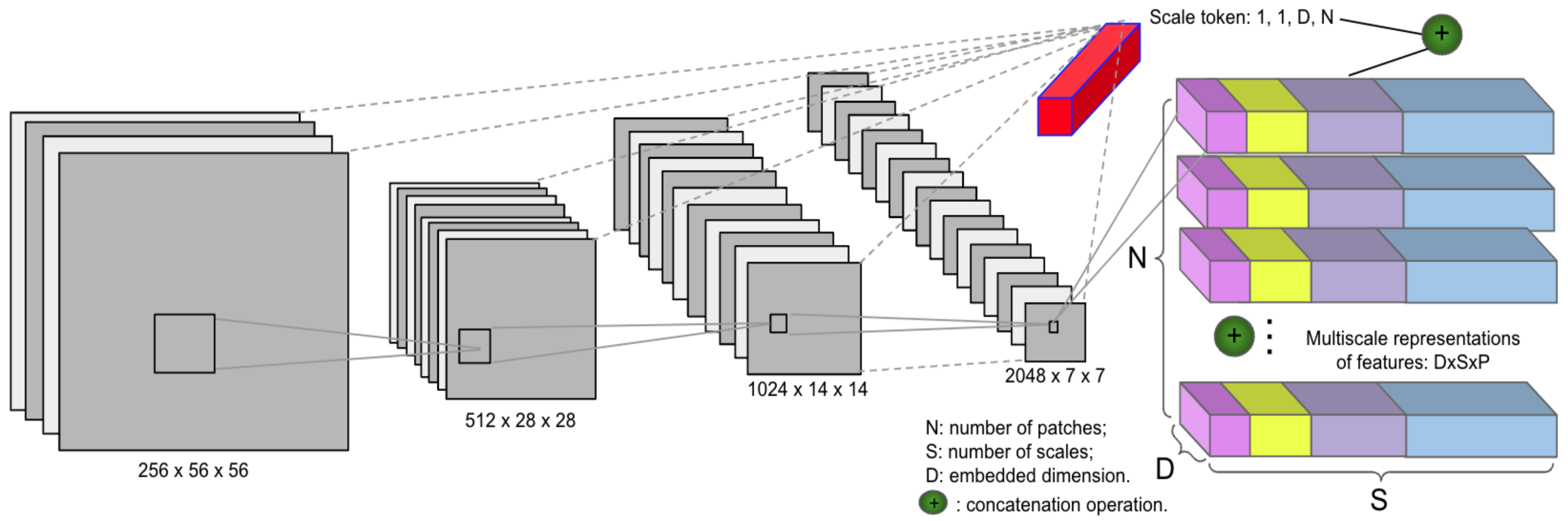

Figure 2 The multi-scale patch tokenization

Backbone Our framework is adaptive to various convolutional backbones with multiple stages. We tried two variations of ResNet in the experiments of image classification. For segmentation, we used the U-Net which consistently performed well on the task. Our model should be able to do various tasks including but not limited to these two tasks.

Multi-scale Patch Tokenization To make sure the patches from different stages to focus on the corresponding regions in the original image X∈R^{HxW}, we first tokenize the features from different scales

$$E_{\mathbf{i}} \in \mathbb{R}^{\frac{H}{2^{i+1}} \times \frac{W}{2^{i+1}} \times D_i} \left(i \in S\{1, 2, 3, 4\} \right)$$

by reshaping them into token sequences with patch size *P, 2P, 4P, 8P.* The channel dimensions are all transformed to the same by a single projection, for example, D = 768. Thus, we concatenate the token sequences of all scales,

$$\mathrm{T}_{\mathrm{S}} = \mathrm{concat}\left[\left(\mathrm{T}_1, \mathrm{T}_2, \mathrm{T}_3, \mathrm{T}_4 \right)\right], \mathrm{T}_{\mathrm{S}} \in \mathbb{R}^{\mathrm{D} \times \mathrm{N} \times \mathrm{S}}$$

Scale Token We use simple convolutional layers to gather channel-wise information from the CNN backbone.

$$S = \phi\left[\left(C_1, C_2, C_3, C_4 \right)\right], C_i \in MaxPooling\left(T_i \right)$$

Attention Module Our model includes scale attention and patch attention modules. The scale attention can be a plug-in module with respective tokenization to any transformer-based model.

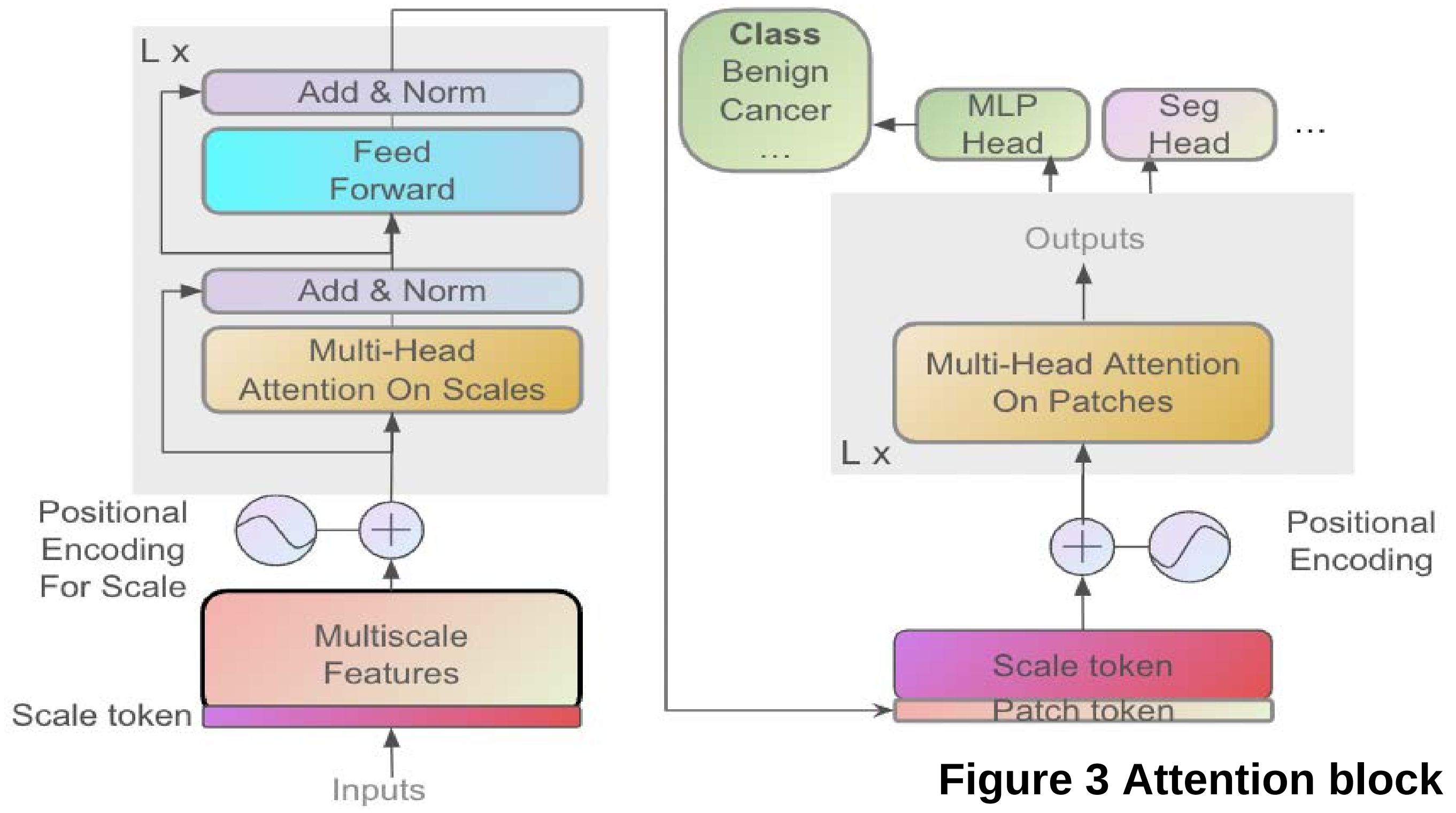

Figure 3 Attention block

Image Classification

Dataset	SVHN	UTAH Kidney Cancer
Train Size	$61,257$	$37,138$
Validation Size	12000	$8,717$
Test Size	$26,032$	$9,174$
Image Size	32×32	224×224
No. Classes	10	4

Image Segmentation

Dataset	MoNuSeg
Train Size	24
Validation Size	6
Test Size	14
Image Size	224×224

Experiments Results

Method	UTAH Kidney Cancer (224x224)	SVHN (32x32)
	Test Accuracy(%)	Test Accuracy(%)
ResNet50	85.58	96.21
ResNet50(pretrained) + ViT-B	80.45	95.25
ResNet50(pretrained) + ViT(24 blocks)	80.15	95.36
ResNet50(pretrained) + Ours	**87.21**	**96.53**
ResNet18	85.92	95.60
ResNet18(pretrained) + ViT-B	82.35	95.21
ResNet18(pretrained) + ViT(24 blocks)	86.39	95.12
ResNet18(pretrained) + Ours	**88.37**	**96.27**

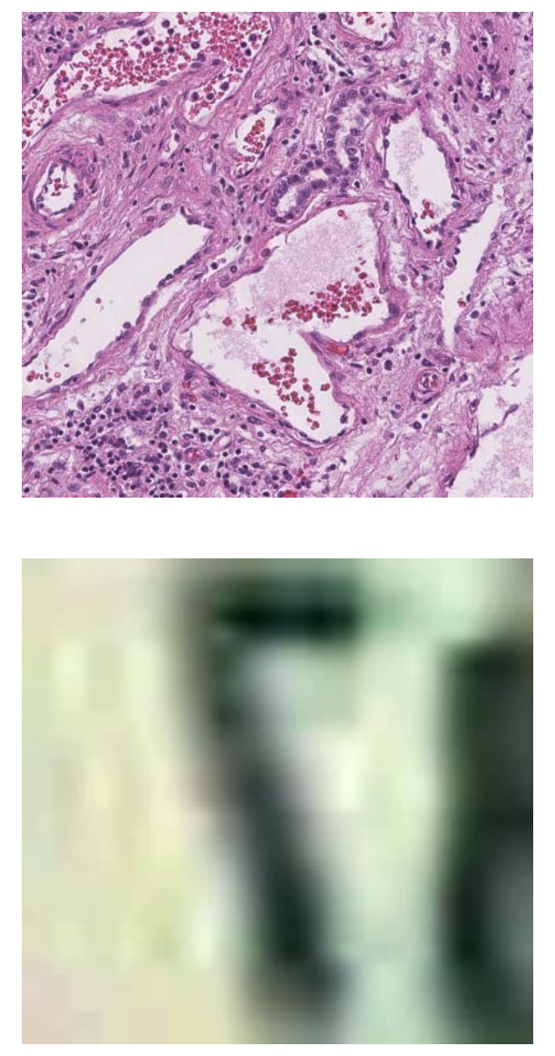

Figure 4 Sample images

Method	MoNuSeg	
	Dice(%)	IoU(%)
U-Net	76.45	62.86
UNet++	77.01	63.04
AttUNet	76.67	63.47
MRUNet	78.22	64.83
TransUNet	78.53	65.05
MedT	77.46	63.37
Swin-Unet	77.69	63.77
Ours	**78.78**	**65.16**

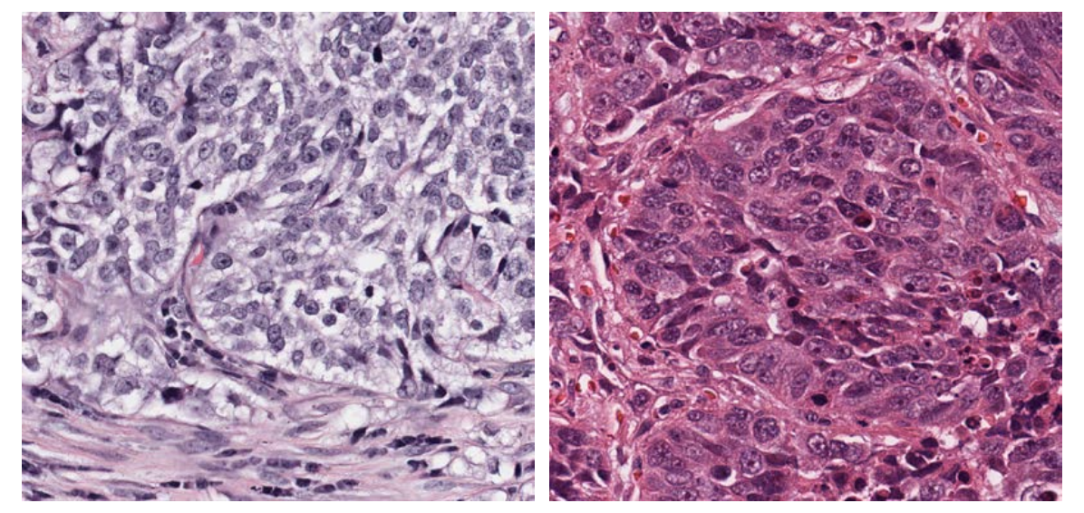

Figure 4 Sample images in MoNuSeg

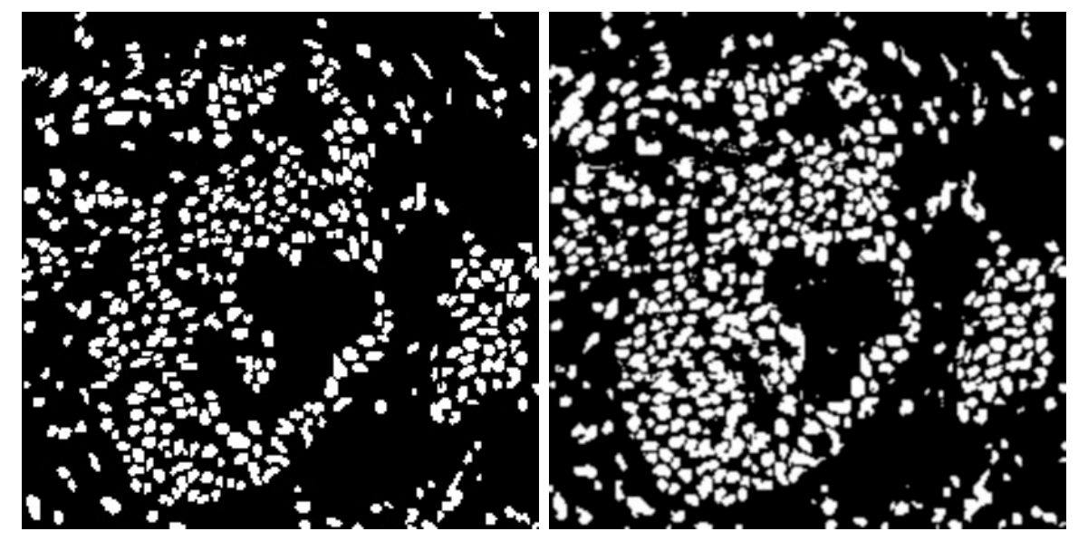

Figure 5 Ground Truth(Left) Predicted(Right)

[1] Dosovitskiy, Alexey, et al. "An image is worth 16x16 words: Transformers for image recognition at scale." arXiv preprint arXiv:2010.11929 (2020).
[2] Xu, Peng, Xiatian Zhu, and David A. Clifton. "Multimodal learning with transformers: A survey." IEEE Transactions on Pattern Analysis and Machine Intelligence (2023).

THE UNIVERSITY OF UTAH
SCI
www.sci.utah.edu