MaxViT-UNet: Multi-Axis Attention for Medical Image Segmentation

Abdul Rehman Khan[1, 2], Asifullah Khan[1, 2, 3*]

[1] Pattern Recognition Lab, Department of Computer & Information Sciences, Pakistan Institute of
Engineering & Applied Sciences, Nilore, Islamabad, 45650, Pakistan

[2] PIEAS Artificial Intelligence Center (PAIC), Pakistan Institute of Engineering & Applied Sciences,
Nilore, Islamabad, 45650, Pakistan

[3] Center for Mathematical Sciences, Pakistan Institute of Engineering & Applied Sciences, Nilore,
Islamabad, 45650, Pakistan

Corresponding Author: *Asifullah Khan, asif@pieas.edu.pk

ABSTRACT

Convolutional Neural Networks (CNNs) have made significant strides in medical image analysis in recent years. However, the local nature of the convolution operator may pose a limitation for capturing global and long-range interactions in CNNs. Recently, Transformers have gained popularity in the computer vision community and also medical image segmentation due to their ability to process global features effectively. The scalability issues of self-attention mechanism and lack of the CNN-like inductive bias may have limited their adoption. Therefore, hybrid Vision transformers (CNN-Transformer), exploiting advantages of both Convolution and Self-attention Mechanisms, have gained importance. In this work, we present MaxViT-UNet, an Encoder-Decoder based hybrid vision transformer (CNN-Transformer) for medical image segmentation. The proposed Hybrid Decoder, based on MaxViT-block, is designed to harness the power of both the convolution and self-attention mechanisms at each decoding stage with nominal computational burden. The inclusion of multi-axis self-attention, within each decoder stage, significantly enhances the discriminating capacity between the object and background regions, and thereby helps in improving the segmentation efficiency. In the Hybrid Decoder block, the fusion process commences by integrating the upsampled lower level decoder features, obtained through transpose convolution, with the skip-connection features derived from the hybrid encoder. Subsequently, the fused features undergo refinement through the utilization of a multi-axis attention mechanism. The proposed decoder block is repeated multiple times to progressively segment the nuclei regions. Experimental results on MoNuSeg18 and MoNuSAC20 dataset demonstrates the effectiveness of the proposed technique. Our MaxViT-UNet outperformed the previous CNN-based (UNet) and Transformer-based (Swin-UNet) techniques by a considerable margin on both of the standard datasets. The implementation and trained weights are publicly available at github (https://github.com/PRLAB21/MaxViT-UNet).

Keywords Image Segmentation · Cancer Diagnostics · Medical Image Analysis · CNN-Transformer · MaxViT · Multi-Axis Attention · Sparse Attention · Vision Transformer · UNet Architecture · Auto Encoder

1 Introduction

Image segmentation has primary importance in the field of medical image analysis, which involves identifying and delineating specific structures or regions of interest from medical images [1]. Specifically, nuclei segmentation is a crucial task that involves identifying the boundaries of nuclei cells in microscopic histopathology images [1]. Nuclei segmentation plays a crucial role in enhancing the precision and dependability of medical diagnosis and treatment. By enabling precise detection and quantification of nuclei, it can aid in the development of personalized treatment plans and improve patient outcomes.

In recent years, deep learning algorithms have demonstrated outstanding performance across various applications [2, 3, 4], particularly in image segmentation [5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16]. They have proven to be highly effective, delivering both accuracy and efficiency in solving this challenging task. The ability of Convolutional Neural Networks to automatically capture low-level to high-level features hierarchically from the dataset renders them highly valuable for medical image analysis [17]. Particularly, the application of deep CNNs in medical image segmentation is growing rapidly. Deep learning models such as Fully Convolutional Neural Network (FCNN) [5] and UNet-like encoder-decoder architectures [6, 18, 19, 20, 21, 7] have emerged as powerful tools in medical image segmentation and have shown superior performance compared to traditional methods [22, 23, 24, 25]. Such networks utilize a combination of convolutional and down-sampling layers in the encoder and up-sampling layers in the decoder to extract deep features and fuse high-resolution features to achieve pixel-level semantic predictions. They can effectively extract features and learn complex patterns from medical images, making them particularly suitable for medical image analysis tasks such as segmentation.

More recently, vision transformers [26, 27, 28, 7, 29] have also emerged as a powerful tool for the medical domain and have shown impressive results on a variety of segmentation tasks in medical imaging. The self-attention mechanism of vision transformers allow them to focus on relevant image regions while suppressing irrelevant features, giving them potential to significantly improve the segmentation accuracy of medical images. It helps in effectively extracting the most important features in an image and capture the long-range dependencies between them [26]. Recently, Swin-UNet [7] adapted the Swin Transformer for medical image segmentation by utilizing the Shifted-Window attention mechanism in the decoder. Their architecture is purely transformer-based and lacks the inductive bias of convolutions. Hybrid approaches [30, 31, 32, 33, 34, 35, 36] try to tackle this problem by using convolutions along with self-attention in their encoder but either they suffer from quadratic nature of self-attention or use convolution-based decoders. To the best of our knowledge, Hybrid Decoder has never gained much attention for medical image segmentation tasks before. Our proposed idea utilizes hybrid technique in both encoder and decoder and uses multi-axis attention with linear time complexity with respect to image dimensions.

Inspired by the impressive outcomes achieved by Multi-Axis self-attention (Max-SA) in previous research [37], we have harnessed its potential for medical image segmentation and introduced a novel architecture called MaxViT-UNet, which adopts a UNet-style framework. By incorporating Max-SA, we have enhanced the multi-head self-attention mechanism, enabling the extraction of local and global level features in a computationally efficient manner. The hybrid design of the MaxViT-UNet encoder and decoder enables the generation of contextually rich features at higher levels and noise-free features at lower levels, which are crucial for achieving accurate medical image segmentation. The main outcomes of the proposed work are as follows:

1. A new framework named MaxViT-UNet is proposed, a comprehensive hybrid image segmentation framework based on the encoder-decoder architecture similar to UNet. The framework comprises two key components: MaxViT-Encoder and Proposed Hybrid Decoder. The encoder and the proposed deocder are hybrid modules utilizing the MaxViT-block and skip connections for effective feature processing.

2. The proposed Hybrid Decoder block, Hybrid Decoder, consists of two steps: (1) In merging step, up-sampled features from higher semantic level and encoder features from high spatial level are simply concatenated, (2) In fusion step, MaxViT-blocks are utlized for efficiently processing and fusing concatenated information.

3. The proposed Hybrid Decoder, made by simply repeating the decoder block, is designed to be parameter efficient and computationally lightweight without compromising on the segmentation performance. The local and global attention throughout the proposed decoder helps in discarding irrelevant features for high quality image segmentation.

4. The effectiveness of the proposed MaxViT-UNet has been demonstrated through experiments conducted on multiple datasets. Additionally, the ablation results indicate the promising prospects of utilizing a Hybrid Decoder for medical image segmentation.

The next sections are organized as follows: In Section 2, the recent CNN and Transformer-based techniques in medical image segmentation are highlighted. The proposed MaxViT-UNet details are discussed in detail in Section 3. The dataset and results are provided in Section 4, and the last section concludes the paper.

2 Related Works

The conventional approaches for medical image segmentation primarily relied on morphological operations (erosion, dilation, opening, and closing), or color, contour, and watershed-based techniques and traditional machine learning [22, 23, 24, 25]. These approaches do not generalize well and suffer from different sources of variations in medical

images such as variation in nuclei shape across various organs and tissue types, variation in color across crowded and sparse nuclei, variation in imaging equipment and hospitals/clinics protocol [1].

2.1 CNN Based Techniques

In deep learning based medical image segmentation, one of the earlier approaches includes FCN (fully convolutional network) [5]. Although FCN out-performed conventional techniques, its pooling operation resulted in loss of texture and edge information, necessary for segmentation. Therefore, Ronneberger et al. [6] proposed an encoder-decoder structure called UNet by improving the idea of FCN. Its U-shaped architecture connected the encoder and decoder at various stages via skip-connections to compensate for the semantic loss. The simple and unique architecture of UNet provided superior performance and led to the development of many variants in different medical image domains for segmentation. MultiResUNet [18] replaced skip-connections with residual paths to extract semantic information at multiple scales. M-Net [38] captured multi-level semantic details by injecting rich multi-scale input features to different layers and processing them through a couple of downsampling and upsampling layers. UNet++ [19] proposed a new variant of UNet that incorporates dense connections nested together to effectively represent the fine-grained object information. DenseRes-Unet [39] used dense bottleneck layer in Unet architecture for nuclei segmentation. Recently, the idea of attention mechanism has been applied for enhancing the segmentation performance of medical systems: AttentionUNet [20] improved the segmentation performance of medical images using soft attention by introducing an attention gate module. The CA-Net [40] merged the Spital, Channel, and Scale attentions into one comprehensive attention mechanism. Attention Assisted UNet [41] also improved the attention mechanism for accurate segmentation of sclera images. NucleiSegNet [42] utilized attention in the decoding stage. Some notable work for end-to-end 3D medical image segmentation includes: 3D UNet [21] which replaced 2D convolution with 3D convolution. V-net [43] also improved the UNet with the help of 3D convolution and proposed a dice loss for better segmentation masks. Recently, the idea of channel boosting based CNNs (CB-CNN) have also emerged for medical image segmentation tasks [9, 10, 11, 44], where diverse feature spaces from multiple encoders are fused together for improving the quality of segmentation models.

2.2 ViT Based Techniques

The CNN-based U-shaped networks are effective but the convolution operation captures only the local information and discards global information. To prevent misclassification in segmentation, it is crucial to learn the long-range dependency between background and mask pixels. However, building deeper networks or using larger convolution kernels to capture long-range dependencies results in an explosion of parameters and training process becomes costly. To address these issues, Dosovitskiy et al. [26] introduced Vision Transformers (ViT) with a multi-headed self-attention mechanism capable of capturing long-range dependencies in computer vision tasks. After the success of ViT in natural images and large datasets, medical image processing is also evolving with transformer-based techniques. One of the first techniques that combined transformer and UNet is TransUNet [27]. Later, to efficiently handle smaller sized medical image datasets, Medt [28] improved the self-attention mechanism using a gated axial attention module. Swin-Unet [7], a pure transformer architecture, adopted Swin Transformer into a U-shaped encoder-decoder segmentation framework to extract local and global semantic features in a hierarchical fashion. UCTransNet [29] replaced the skip-connection with the channel transformer (CTrans) module. Karimi et. al. [45] applied self-attention between neighboring image patches by modifying the MHSA mechanism of vision transformers. Despite the excellent performance in multiple image segmentation tasks. Vision Transformers suffer from the problem of computational overload, which has never been solved.

2.3 Hybrid Based Techniques

The Transformer-based architecture outperforms CNN in capturing long-range dependencies, but it has a drawback of lacking interaction with surrounding feature information due to the division of images into fixed-size patches. Surpassing the current performance of medical image segmentation systems is difficult using transformer-based, or CNN-based UNet architectures. To overcome this challenge and enhance segmentation performance, researchers have combined Transformer with CNNs by utilizing the self-attention mechanism along with convolution operation to create a lightweight hybrid image segmentation framework [30, 31, 32, 33]. Chen et al. [27] proposed a strong encoder, by combining Transformer with CNN, for segmenting 2D medical images. Claw UNet [34] used the complementarity of Transformer and CNN to create hybrid blocks in the encoder for multi-organ dataset segmentation. Multi-Compound Transformer [35] attained state-of-the-art results in six different benchmarks by integrating semantic information of hierarchical scales into a unified framework. Zhou et al [36] applied CNN and transformer blocks in a crosswise manner to achieve better performance.

Above mentioned hybrid techniques still suffer from the quadratic nature of self-attention mechanism and are unable to properly utilize the local feature extraction capability of CNNs with global feature extraction capability of Transformers. In the proposed technique we interleaved convolution and self-attention efficiently at each stage and the linear nature of the multi-axis attention mechanism utilized makes our technique fast and robust for medical image segmentation tasks.

3 Proposed Methodology for MaxViT-UNet Framework

3.1 Architecture Overview of MaxViT-UNet Framework

The proposed MaxViT-UNet consists of an encoder, bottleneck layer, proposed decoder, and skip connections. Figure 1 presents the complete architectural details of the proposed methodology. Throughout our encoder-decoder architecture, we utilized the identical MaxViT block structure, consisting of a parameter-efficient `MBConv` [46] and scalable `Max-SA` mechanisms [37]. The stem stage of the encoder ($S0$) downsamples the input image of shape $C \times H \times W$ into $64 \times \frac{H}{4} \times \frac{W}{4}$ using `Conv3 × 3` layers. The input sequentially passes through four encoder stages $S1$ to $S4$. In each stage, the number of feature channels are doubled ($64, 128, 256, 512$) and the spatial size is reduced by half ($\frac{1}{4}, \frac{1}{8}, \frac{1}{16}, \frac{1}{32}$), creating hierarchical features like UNet. The first `MBConv` block in each stage is responsible for doubling the input channels using `Conv1 × 1` layer and halving the spatial size using `Depthwise Conv3 × 3` layer. The last encoder layer, also called bottleneck, contains contextually rich features and provides a bridge from encoder to decoder.
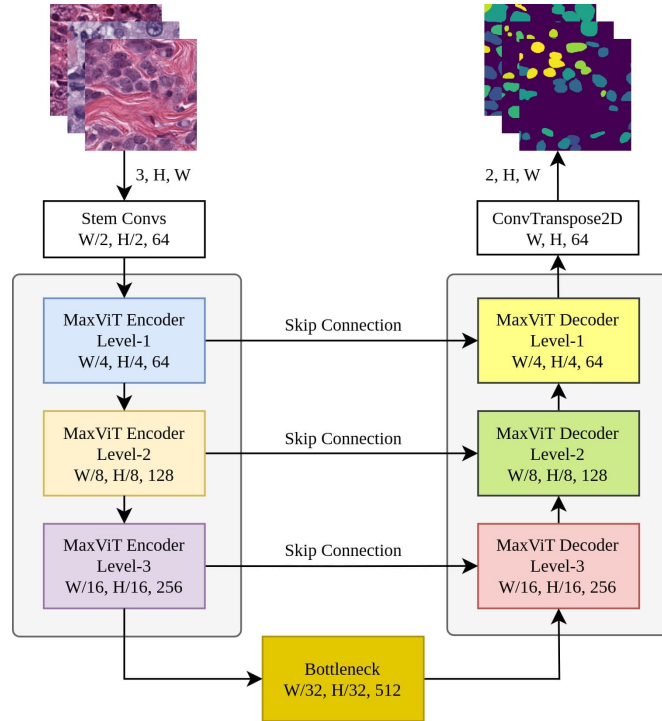

Figure 1: Encoder-Decoder architecture of the Proposed MaxViT-UNet. The MaxViT encoder generates hierarchical features at four scales. The proposed decoder, just like UNet, first upscales the bottom-level features, merges them with skip-connection features, and applies the MaxViT-block a couple of times to produce an output mask image.

The proposed Hybrid Decoder is composed of Transpose Convolutions and MaxViT-blocks. Inspired by UNet [6], it is symmetric in architecture, meaning the decoder is made up of three stages $D1$ to $D3$, matching with $S1$ to $S3$ stages of encoder respectively. The down-sampling operation in the encoder causes the loss of spatial information. To overcome this information, at each stage of the decoder, contextually rich features from the lower decoder stage are concatenated with locality-rich features of the encoder transferred through skip-connection. In contrast to the `Conv1 × 1` layer in the encoder that shrinks the spatial size, a Transpose Convolution layer is used for up-sampling the feature maps of the previous stage. The concatenated features are transformed through a couple of MaxViT-blocks before passing to the next stage. The MaxSA processing helps in reducing the noise information and simultaneously modeling the local-global differences between background and nuclei pixels. After the feature processing stages of the decoder, the feature maps of shape $64 \times 4 \times 4$ are up-sampled four times to match the dimensions of the output mask

Table 1: Configuration of the Proposed MaxViT-UNet as an Encoder-Decoder architecture.

Encoder Level	Ouput Size	MaxViT Encoder	
Stem	(64, 128, 128)	Conv(k=3, s=2) Conv(k=3, s=1)	
S1	(64, 64, 64)	MBConv(E=4, R=4) Window-Rel-MSA(P=8, H=2) Grid-Rel-MSA(G=8, H=2)	× 2
S2	(128, 32, 32)	MBConv(E=4, R=4) Window-Rel-MSA(P=8, H=2) Grid-Rel-MSA(G=8, H=2)	× 2
S3	(256, 16, 16)	MBConv(E=4, R=4) Window-Rel-MSA(P=8, H=2) Grid-Rel-MSA(G=8, H=2)	× 2/5
S4	(512, 8, 8)	MBConv(E=4, R=4) Window-Rel-MSA(P=8, H=2) Grid-Rel-MSA(G=8, H=2)	× 2
Decoder Level	Ouput Size	Hybrid Decoder	
D1	(64, 64, 64)	ConvTranspose(k=2, s=2) MBConv(E=4, R=4) Window-Rel-MSA(P=8, H=2) Grid-Rel-MSA(G=8, H=2)	× 2
D2	(128, 32, 32)	MBConv(E=4, R=4) Window-Rel-MSA(P=8, H=2) Grid-Rel-MSA(G=8, H=2)	× 2
D3	(256, 16, 16)	MBConv(E=4, R=4) Window-Rel-MSA(P=8, H=2) Grid-Rel-MSA(G=8, H=2)	× 2

with that of the input image and true mask $H \times W$. The last convolution layer reduces the channels from 64 to 2 to generate the pixel-level segmentation probabilities. The proposed MaxViT-UNet, though hybrid in nature, consists of only 27.29 million parameters, lighter than UNet with 29.06 million parameters. Table 1 summarizes the architectural configurations of the MaxViT-UNet.

3.2 MaxViT Block

The hybrid MaxViT-block effectively blends the multi-axis attention mechanism with convolution, as shown in Figure 2. It is based on the observation that convolution complements transformer attention by improving the generalization and the training speed of the network [47]. To this end, `MBConv` sub-block [46], containing squeeze-and-excitation (SE) [48] attention, is used for feature processing before applying the multi-axis self-attention mechanism. Another advantage of the `MBConv` layer is that its `depth-wise convolutions` act as conditional position encoding (CPE) [49], removing the need for explicit positional encoding layers. In `MBConv` the expansion for inverted bottleneck layer was set to 4 and shrink rate for squeeze-excitation layer was set to 0.25. After the `MBConv` layer, the block and grid self-attentions are stacked sequentially to model the local and global feature interactions simultaneously in a single block. Following the good design practices [26, 7], the MaxViT block contains `LayerNorm` [50], `Feed-Forward Networks` (FFNs) [26, 7], and skip-connections in `MBConv`, block and grid attentions sub-blocks.

Let **x** represents input feature tensor, the `MBConv` block without downsampling is given as:

$$\mathbf{x} = \mathbf{x} + \texttt{Proj}(\texttt{SE}(\texttt{DWConv} \downarrow (\texttt{Conv}(\texttt{Norm}(\mathbf{x}))))) \tag{1}$$

where `Norm` represents `BatchNorm` layer [51], `Conv` is expanding layer consisting of `Conv1 × 1`, `BatchNorm` and `GELU` [52] activation function. `DWConv` is processing layer consisting of `Depthwise Conv3 × 3`, `BatchNorm` and `GELU`. `SE` represents Squeeze-Excitation layer [48], while `Proj` reduces the number of channels using `Conv1 × 1`.

5

In each stage, the first `MBConv` downsample the input $\mathbf{x}$ using `Depthwise Conv3 × 3` with stride of 2, while the residual connection consists of pooling and channel projection layers:

$$\mathbf{x} = \text{Proj}(\text{MaxPool}(\mathbf{x})) + \text{Proj}(\text{SE}(\text{DWConv} \downarrow (\text{Conv}(\text{Norm}(\mathbf{x}))))) \tag{2}$$

3.3 Max-SA: Multi-Axis Self-Attention

The self-attention introduced by transformer [53] and utilized by vision transformer [26] fall into the category of dense attention mechanism due to its quadratic complexity. Considering the effectiveness of sparse approaches for self-attention [54, 55], Tu et al. [37] introduced an efficient and scalable self-attention module called multi-axis self-attention, which decomposes the original self-attention into sparse forms. (1) Window Attention for blocked local feature extraction and (2) Grid Attention for dilated global feature processing. Max-SA provides linear complexity without losing locality information.

The blocked local or window attention follows the idea of Swin Transformer [7]. Let $\mathbf{X}$ represent a feature tensor of shape $C \times H \times W$. The window partition layer reshapes $\mathbf{X}$ into shape $(N, P \times P, C)$, where $N = \frac{H}{P} \times \frac{W}{P}$ represents the total number of non-overlapping local windows, each of spatial shape $P \times P$ and channel dimension C. Each local window is passed through standard multi-head self-attention (MHSA) to model the local interactions. Finally, the window reverse layer reshapes $\mathbf{X}$ back to $C \times H \times W$.

In order to model global interactions, Max-SA incorporates grid attention, a simple and effective way of obtaining global relations in a sparse manner. The grid partition layer reshapes $\mathbf{X}$ into shape $(G \times G, N, C)$, to obtain $G \times G$ number of global windows, each having dynamic spatial size represented with $N = \frac{H}{G} \times \frac{W}{G}$, and channel dimension C. Each local window is passed through standard multi-head self-attention (MHSA) to model the local interactions. Finally, the window reverse layer reshapes $\mathbf{X}$ back to $C \times H \times W$. The utilization of the grid-attention on the decomposed grid axis enables the global mixing of spatial tokens through dilated operations. With consistent window and grid sizes, the Max-SA mechanism ensures linear complexity in relation to the spatial size.

The CNNs are famous for their location equivariance inductive bias, a property lacked by vanilla self-attention mechanism [26, 56]. Max-SA attention blocks adopted the pre-normalized relative self-attention [57] as a solution to this problem. The relative self-attention mechanism [57, 58, 59, 7] adds a learnable relative bias to the attention weights, which proved to improve the self-attention mechanism.

Max-SA allows global-local feature interactions on various feature resolutions throughout the encoder and decoder architecture. In the proposed encoder-decoder architecture, both the window and grid sizes were fixed to 8 to make it compatible with 256×256 image size, used for training and testing. The number of attention heads were set to 32 for all attention blocks.

3.4 MaxViT-UNet Encoder

The encoder of the proposed MaxViT-UNet framework is made of MaxViT architecture [37], by simply stacking `MBConv` and `Max-SA` modules alternatively in a hierarchical fashion. Unlike the MaxViT [37], where the number of blocks and channel dimensions are increased per stage to scale up the model. We used two MaxViT blocks per stage, to obtain a small and efficient encoder. Additionally, the third stage was repeated 2 times and 5 times for MoNuSeg18 and MoNuSAC20 datasets, respectively. The four stages of our encoder produce hierarchical feature representation just like UNet. MaxViT takes advantage of the local-global receptive fields via convolution and local-global attention mechanisms, throughout the encoder from earlier to deeper stages, and shows better generalizability and model capacity. The last stage of the encoder is named bottleneck as it contains semantic-rich features and provides a bridge from encoder to decoder.

3.5 Proposed Hybrid Decoder

The proposed Hybrid Decoder is designed by stacking layers of MaxViT-block in a hierarchical architecture, with a `Transpose Conv` layer at the start of each stage, as shown in Figure 2. Similar to the encoder, we created a parameter-efficient decoder by using only two MaxViT blocks per stage. The decoder also enjoys the global and local receptive fields at all stages and is able to better reconstruct output masks as compared to previous approaches. Similar to Swin-UNet [7], our decoder contains three stages that are connected with the corresponding top three stages of the encoder. Inside a single decoder block, features from the bottom layers are passed through the `Transpose Conv` layer for up-sampling and matching their shape with features from skip-path. The up-sampled features are concatenated with the corresponding skip-connection features to obtain semantic and spatial-rich features. The MaxViT blocks further enhance them using `MBConv`, `local attention` and `global attention` sub-block. Let $\mathbf{y^{(i-1)}}$ the represents features coming from the previous decoder stage having dimension $C \times H \times W$, and $\mathbf{x^{(i)}}$ represent features coming

from skip-connection at the same stage having dimensions $C \times 2H \times 2W$, then the first block of each decoder stage can be formulated as:

$$\mathbf{y}^{(i)} = \text{UpConv}(\mathbf{y}^{(i\text{-}1)}) \tag{3}$$

$$\mathbf{x} = \mathbf{x} + \text{GridAttn}(\text{BlockAttn}(\text{MBConv}(\mathbf{y}^{(i)}|\mathbf{x}^{(i)}))) \tag{4}$$

where `UpConv` consists of `BatchNorm` layer [51] and `Mish` activation function [60], the | operator represents `concatenation`, and `MBConv`, `BlockAttn`, and `GridAttn` are sub-blocks of MaxViT-block for feature processing.
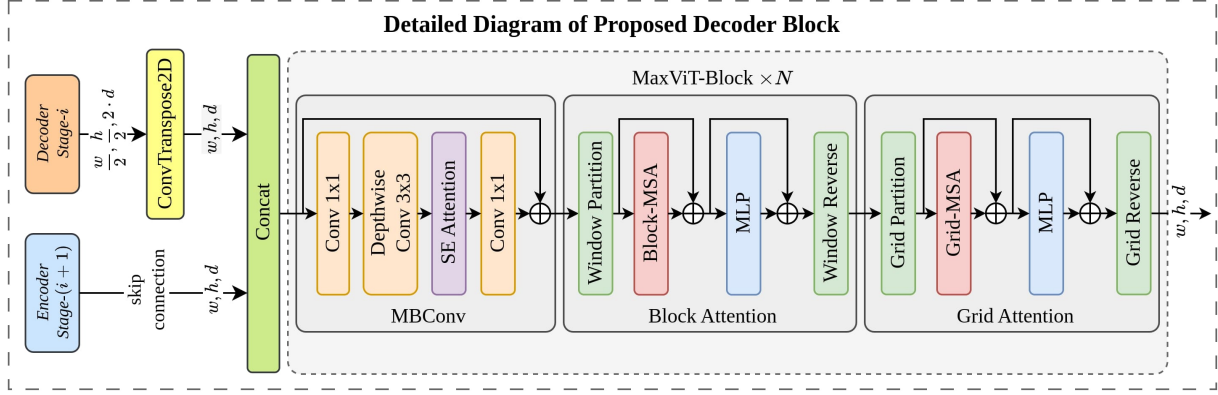

Figure 2: Detailed architecture of the proposed Hybrid Decoder block. Features from the ith decoder stage are upscaled using `ConvTranspose2D` layer to match with size of the $(i+1)^{th}$ encoder stage coming from skip-connection. After concatenation (`concat`) operation, the MaxViT-block is used a couple of times to merge the features efficiently.

3.6 Loss Functions of the Proposed MaxViT-UNet

The models are penalized using the composite loss function consisting of `Cross-Entropy` and `Dice` loss functions with loss weights of $\lambda 1 = 1$ and $\lambda 2 = 3$ respectively in all the experiments. Both `Cross-Entropy` and `Dice` losses are calculated pixel-wise. In `dice` loss, only the nuclei class was considered and the background was ignored to force the model to focus on nuclei regions more strongly. Given true mask $\mathbf{y}$ and predicted mask $\hat{\mathbf{y}}$ images, the mathematical form of both the loss functions is as follows:

$$\text{Loss}(\mathbf{y}, \hat{\mathbf{y}}) = \lambda 1 * \text{CELoss}(\mathbf{y}, \hat{\mathbf{y}}) + \lambda 2 * \text{DiceLoss}(\mathbf{y}, \hat{\mathbf{y}}) \tag{5}$$

$$\text{CELoss}(\mathbf{y}, \hat{\mathbf{y}}) = - \sum_{i=1}^{H \times W} (\mathbf{y}_i * \log(\hat{\mathbf{y}_i})) \tag{6}$$

$$\text{DiceLoss}(\mathbf{y}, \hat{\mathbf{y}}) = 1 - \sum_{c=1}^{C} 2 \times \frac{|\mathbf{y}_c \cap \hat{\mathbf{y}}_c|}{|\mathbf{y}_c| + |\hat{\mathbf{y}}_c|} \tag{7}$$

In `CELoss`, $\mathbf{y}_i$ represents the $\mathbf{i^{th}}$ pixel of true mask image and $\hat{\mathbf{y}}_i$ represents the $\mathbf{i^{th}}$ pixel of predicted mask image, and summation is performed over all pixels ($H \times W$) to accumulate the error for a complete image. In `DiceLoss`, $\mathbf{y}_c$ represents the $\mathbf{c^{th}}$ class channel of true mask image, and $\hat{\mathbf{y}}_c$ represents the $\mathbf{c^{th}}$ class channel of predicted mask image, and summation is performed over all classes (C) to accumulate the error for all classes and all pixels.

4 Experiments

In order to show the effectiveness of the proposed MaxViT-UNet segmentation framework, extensive experiments were performed on medical image segmentation tasks. The following section provides the details of the dataset used, than pre-processing steps, working environment used, hyper-parameters and performance metrics are discusses. Finally, the quantitative and qualitative results of the proposed MaxViT-UNet are compared with previous image segmentation techniques.

4.1 Dataset Description

Over the past few years, many challenges have been organized for medical image segmentation tasks to progress the research in this field. In order to show the effectiveness of our proposed MaxViT-UNet framework, we choose to work with MoNuSeg 2018 [1] and MoNuSAC 2020 [61] challenge datasets. Table 2 summarized the details of both datasets. Both datasets have their own difficulties and deals with different level of problems. The next few sections highlight the details of both datasets.

Table 2: Summary of the datasets used to train and evaluate the proposed MaxViT-UNet

Dataset	Classes	Subset	Images	Nuclei	Organs
MoNuSeg18	Background, Nuclei	Train	30	21,623	Breast, Liver, Kidney, Prostate, Bladder, Colon, Stomach
		Test	14	7,223	Breast, Kidney, Prostate, Bladder, Colon, Lung, Brain
MoNuSAC20	Epithelial, Lymphocytes, Macrophoges, Neutrophils	Train	46	31,411	Breast, Kidney, Lung, Prostate
		Test	25	15,498	Breast, Kidney, Lung, Prostate

4.1.1 MoNuSeg18

The MoNuSeg 2018 challenge provided a challenging dataset [1] comprising images from 7 different organs: (1) breast, (2) colon, (3) bladder, (4) stomach, (5) kidney, (6) liver, and (7) prostate. Also, images acquired from 18 different hospitals, practicing different staining techniques and image acquisition equipment, add another source of variation and ensures the diversity of nuclear appearances. The training data consists of 30 tissue images (1000×1000 resolution), 7 validation images, and 14 test images. The training dataset consists of 21623 annotated manually nuclear boundaries. For each selected individual patient from TCGA [62], an image was extracted from a distinct whole slide image (WSI) that was scanned at 40× magnification. Sub-images were selected from regions containing a high density of nuclei. To ensure diversity in the dataset, only one crop per WSI and patient was included. The test comprises 14 images spanning 5 organs common with the training set: (1) breast, (2) colon, (3) bladder, (4) kidney, (5) liver, and 2 organs different from the testing set: (1) lung, (2) brain, to make the test set more challenging. The test set contains approximately 7,223 annotated nuclear boundaries.

4.1.2 MoNuSAC20

The MoNuSAC20 dataset [61] was designed to be representative of various organs and nucleus types relevant to tumor research. Specifically, it included Lymphocytes, Epithelial, Macrophages, and Neutrophils. The training data consisted of cropped whole slide images (WSIs) obtained from 32 hospitals and 46 patients from TCGA [62] data portal, scanned at a 40× magnification. The dataset provides nuclei class labels along with nuclear boundary annotations. The testing data followed a similar preparation procedure but included annotations for ambiguous regions. These are regions with faint nuclei, unclear boundaries, or where the true class is not confirmed by annotators. The testing data comprised 25 patient samples from 19 different hospitals, with 14 hospitals overlapping with the training dataset.

4.2 Dataset Pre-processing

4.2.1 MoNuSeg18

For the MoNuSeg18 dataset [1], 256×256 dimension patches (images and masks) from 1000×1000 images were extracted to use for training and testing purposes of segmentation models. It was also ensured that training patches remain in the training set and testing patches remain in the testing set, to avoid leaking of the testing set and faulty evaluation metrics. Various augmentation techniques are utilized in order to increase the size of the dataset, including *RandomAffine*, *PhotoMetricDistortion*, *Random Horizontal* and *Vertical Flip* with 0.5 flip probability. The step-by-step outcome of these pre-processing steps are shown in figure 3 for MoNuSeg18 [1] dataset. Considering the modality differences between ImageNet and histopathology images, we calculated normalization parameters (`mean=[171.31, 119.69, 157.71]`, `std=[56.04, 59.61, 47.69]`) and used them for image normalization during training and testing phases.

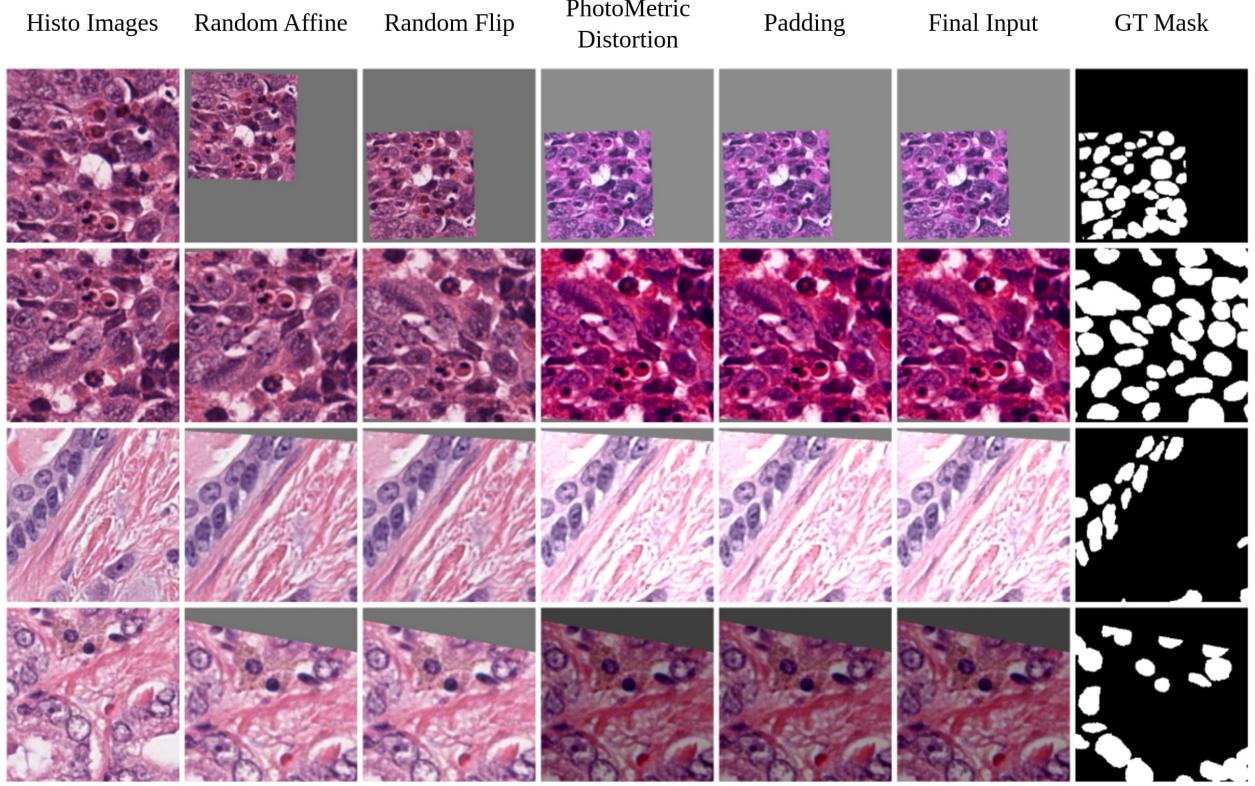Figure 3: Data Pre-processing Pipeline visualized for MoNuSeg18 dataset. From left to right: Original Image (resized to 256×256,

4.2.2 MoNuSAC20

For the MoNuSAC20 dataset [61], the same pre-processing was applied as for the MoNuSeg18, i.e. 256×256 dimension patches (images and masks) were extracted for training and testing. The same augmentation techniques were applied to increase the dataset size and robustness of the model. For MoNuSAC20, the ImageNet normalization parameters `mean=[123.675, 116.28, 103.53]`, `std=[58.395, 57.12, 57.375]` were utilized during training and testing phases as they obtained good results on this dataset.

4.3 Working Environment

For the implementation, training, and evaluation of our proposed MaxViT-UNet and baseline models, we used MMSegmentation [63] (v0.24.1) and PyTorch [64] (v1.12.1) frameworks. We used conda (v4.12.0) for setting up our environment. All the trainings were done using NVIDIA DGX Station with 4 Tesla V100 GPUs, 120GB GPU memory, 256 GB RAM and Intel Xeon E5-2698 CPU.

4.4 Training Details of the Proposed MaxViT-UNet

For fast training speed, we utilized the distributed training provided by MMSegmentation [63]. The batch size for single GPU was set to 4, with effective batch size of 16 due to distributed training on 4 GPUs. We experimented with SGD, Adam [72], AdaBelief [73], and AdamW [74] optimizers, and found the results of AdamW [74] better than others. For all final trainings, AdamW was used with a learning-rate of 0.005, weight decay of 0.01, and values of betas were set to (0.9, 0.999) for optimizing our model through backpropagation.

4.5 Performance Metrics

The MoNuSeg18 and MoNuSAC20 datasets were evaluated using `Dice` and `IoU` evaluation metrics. Both the `Dice` and `IoU` are widely used segmentation metrics that produce values between 0 and 1. The `Dice` is equivalent to `F1-Score` in

Table 3: Comparative results of the proposed MaxViT-UNet framework with previous techniques on MoNuSeg 2018 Challenge Dataset

Method	Dice	IoU
U-Net [6]	0.8185	0.6927
U-net++ [19]	0.7528	0.6089
AttentionUnet [20]	0.7620	0.6264
MultiResUnet [18]	0.7754	0.6380
Bio-net [65]	0.7655	0.6252
TransUnet [27]	0.7920	0.6568
ATTransUNet [31]	0.7916	0.6551
MedT [28]	0.7924	0.6573
UCTransnet [29]	0.7987	0.6668
FSA-Net [66]	0.8032	0.6699
MBUTransUNet [67]	0.8160	0.6902
Swin-Unet [7]	0.7956	0.6471
MaxViT-UNet (Proposed)	**0.8378**	**0.7208**

Table 4: Comparative results of the proposed MaxViT-UNet framework with previous techniques on MoNuSAC 2020 Challenge Dataset

Method	Dice	IoU
UNet [6]	0.7197	0.5874
Hover-net [68]	0.7626	-
Dilated Hover-net w/o ASPP [68]	0.7571	-
Dilated Hover-net w/ ASPP [68]	0.7718	-
MulVerNet [69]	0.7660	-
NAS-SCAM [70]	0.6501	-
PSPNet [71]	0.7893	0.6594
Swin-Unet [7]	0.4689	0.3924
MaxViT-UNet (Proposed)	**0.8215**	**0.7030**

image segmentation tasks, and $\texttt{IoU}$ is also referred to as Jaccard Index. Mathematically, $\texttt{Dice}$ and $\texttt{IoU}$ are defined as:

$$\texttt{Dice}(\mathbf{y}, \hat{\mathbf{y}}) = \sum_{c=1}^{C} 2 \times \frac{|\mathbf{y}_c \cap \hat{\mathbf{y}_c}|}{|\mathbf{y}_c| + |\hat{\mathbf{y}_c}|} \tag{8}$$

$$\texttt{IoU}(\mathbf{y}, \hat{\mathbf{y}}) = \sum_{c=1}^{C} \frac{|\mathbf{y}_c \cap \hat{\mathbf{y}_c}|}{|\mathbf{y}_c \cup \hat{\mathbf{y}_c}|} \tag{9}$$

where $\mathbf{y}_c$ represents the $\mathbf{c^{th}}$ class channel of true mask image and $\hat{\mathbf{y}_c}$ represents the $\mathbf{c^{th}}$ class channel of predicted mask image. The summation is performed over all classes (C) to accumulate the evaluation metric for a complete image.

4.6 Results and Discussions

The proposed MaxViT-UNet is compared with previous techniques on both the MoNuSeg18 and MoNuSAC20 datasets. The following sections discuss the experimental results on each dataset in detail.

4.6.1 Performance Evaluation of the Proposed MaxViT-UNet

The comparative results of the proposed MaxViT-UNet along with the previous methodologies are presented in Table 3 on MoNuSeg18 dataset and Table 4 on MoNuSAC20 dataset. For comparison on both datasets, UNet and Swin-UNet were trained using MMSegmentation [63] with same hyper-parameters as the proposed technique. For MoNuSeg18 dataset, we performed binary semantic segmentation. Whereas, the MoNuSAC20 challenge contains four types of nuclei, so we performed multi-class semantic segmentation for MoNuSAC20 dataset. The proposed MaxViT-UNet beats the previous techniques by a large margin on both the datasets and proves the significance of the hybrid encoder-decoder
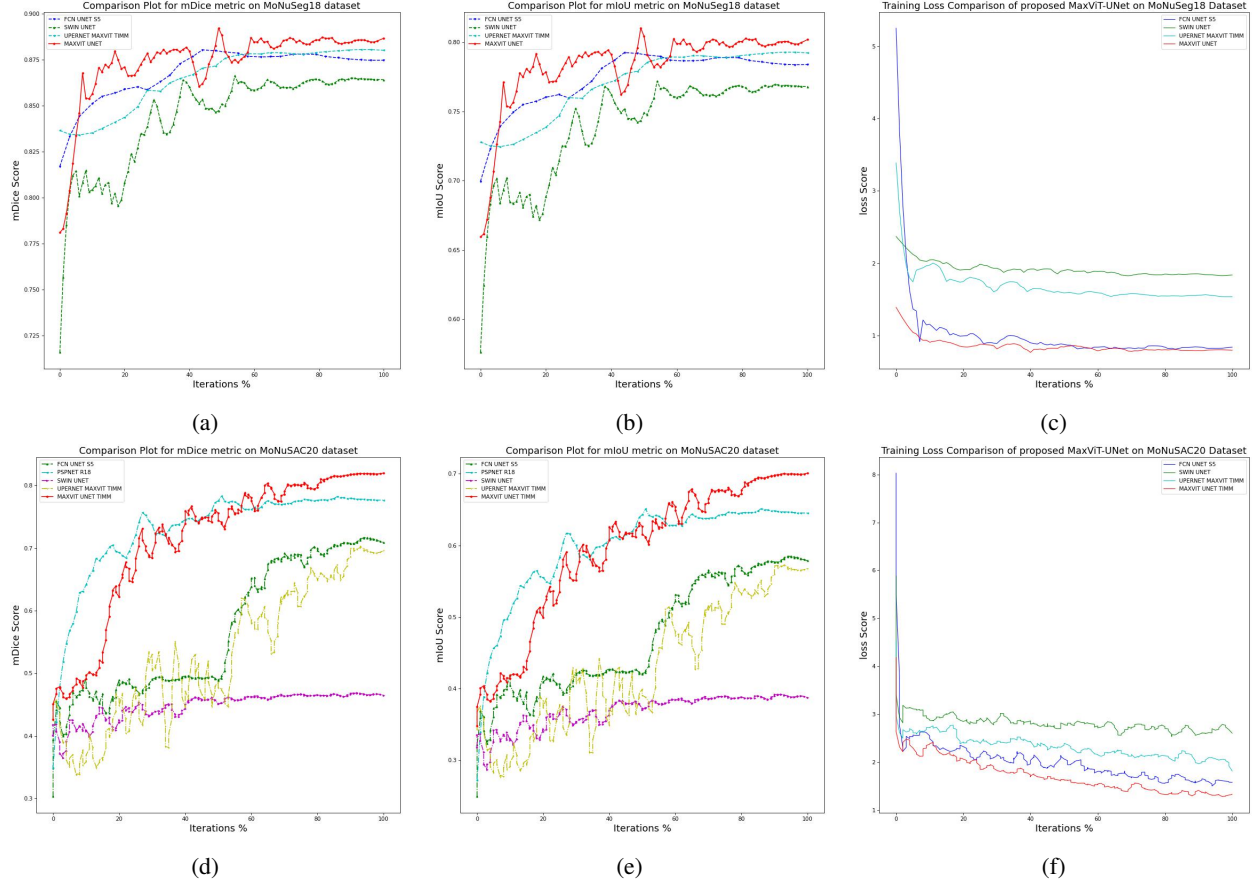
Figure 4: Comparitive plots of the proposed MaxViT-UNet with previous techniques on MoNuSeg18 and MoNuSAC20 challenge datasets. Top row displays the (a) Dice, (c) IoU, and (e) Training Loss on MoNuSeg18 dataset, whereas bottom row displays the (b) Dice, (d) IoU, and (f) Training Loss on MoNuSAC20 dataset.
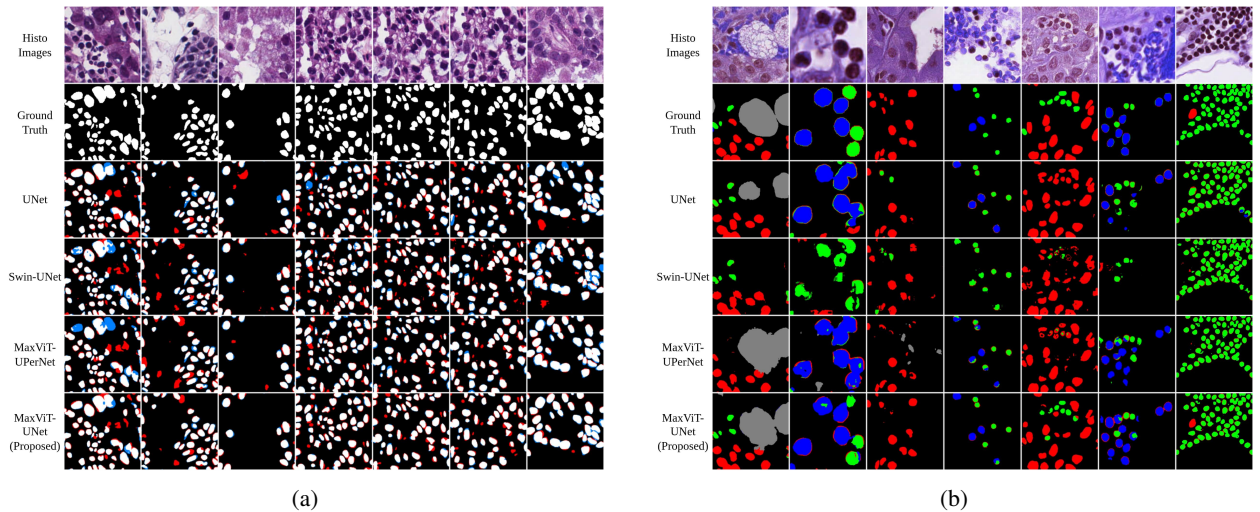

Figure 5: Qualitative comparison of the proposed MaxViT-UNet with recent techniques on (a) MoNuSeg18 dataset; the white, red and blue colors represent true-positive, false positive and false negative predictions respectively. (a) MoNuSAC20 dataset; different colors represent the four classes of nuclei in the dataset.

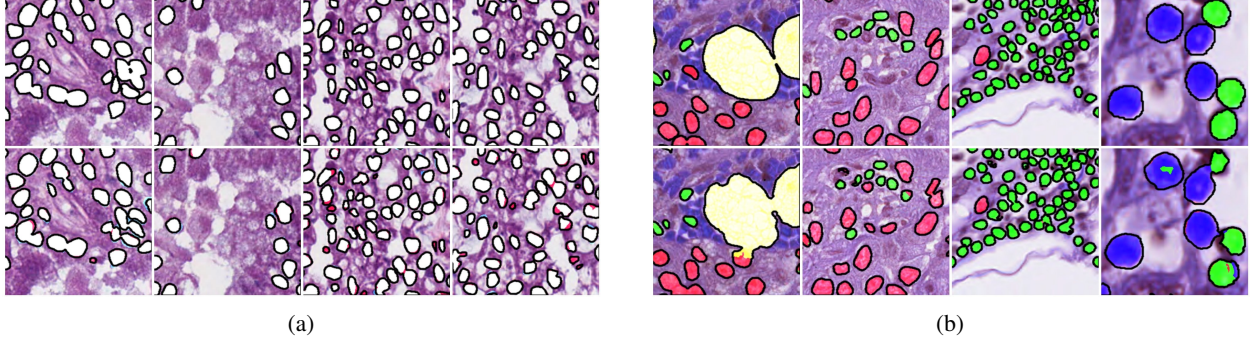(a)   (b)

Figure 6: Comparison of True masks (top rows) and MaxViT-UNet predicted masks (bottom rows) overlaid on the histopathology images on a) MoNuSeg18 dataset; the white, red and blue colors represent true-positive, false positive and false negative predictions respectively. (a) MoNuSAC20 dataset; different colors represent the four classes of nuclei in the dataset.

architecture. It surpassed the CNN-based UNet [6] by 2.36% Dice score and 4.06% IoU score; and Transformer-based Swin-UNet [7] by 5.31% Dice score and 11.40% IoU score on MoNuSeg18 dataset. Whereas, on MoNuSAC20 dataset, the proposed framework surpassed UNet [6] by 14.14% Dice and 19.68% IoU scores; and Swin-UNet [7] by large margin on Dice and IoU metrics as evident from Table 4. The large improvement in both mDice and mIoU scores shows the significance of hybrid encoder-decoder architecture. Figure 4 shows the learning curve plots of mDice (mean Dice), mIoU (mean IoU) and training loss on both MoNuSeg18 and MoNuSAC20 challenge datasets. The proposed MaxViT-UNet framework is represented by red curve lines in all the six metric plots. The baselines are shown with different colors that are same throughout all the metric plots. The training loss curve of the proposed MaxViT-UNet is very stable and lower than the baselines showing the stability and convergence of the proposed framework. The qualitative results on diverse images is presented on MoNuSeg18 dataset in Figure 5a and MoNuSAC20 dataset in Figure 5b. The masks generated by MaxViT are less prone to error as compared to Swin-UNet [7] and vanilla UNet [6], and it also gives relatively accurate boundaries. Figures 6a and 6b compares the ground truth mask images and predicted mask images of the proposed MaxViT-UNet overlaid on histopathology images. For MoNoSeg18 dataset 6a, white color represents the true predicted regions whereas red and blue colors highlight the erroneous regions. And for MoNuSAC20 dataset 6b, four different colors represent four types of nuclei classes in the dataset.

4.6.2 Ablation Study of the Proposed MaxViT-UNet

To inspect the efficacy of the proposed Hybrid Decoder, MaxViT encoder was trained with UPerHead Decoder [75] network. The UPerHead decoder takes advantage of the Pyramid Pooling Module (PPM) [71] in the bottleneck and all of the decoding stages are based on convolution layer. Table 5 shows the comparative results on Dice and IoU metric for MoNuSeg18 and MoNuSAC20 datasets and proves the effectiveness of the proposed Hybrid Decoder. The large margin for multi-class problem of MoNuSAC20 highlights that the proposed Hybrid Decoder has better capacity at handling and distinguishing between regions of multiple classes. The hybrid nature of proposed decoder may provide it with the ability to exploit large to small contexts at each scale. The symmetric decoder is designed to be used standalone in UNet like encoder-decoder frameworks, and the improved performance shows that it has potential for generating accurate segmentation masks with other types of encoders as well.

Table 5: Ablation study results of the proposed MaxViT-UNet Decoder

Method	Dataset	Image Size	Dice	IoU
MaxViT with UPerNet Decoder	MoNuSeg18	(256, 256)	0.8176	0.6914
MaxViT-UNet (Proposed)	MoNuSeg18	(256, 256)	**0.8378**	**0.7208**
MaxViT with UPerNet Decoder	MoNuSAC20	(256, 256)	0.7148	0.5828
MaxViT-UNet (Proposed)	MoNuSAC20	(256, 256)	**0.8215**	**0.7030**

12

5 Conclusion

In this work, we presented an encoder-decoder based hybrid vision transformer architecture, dubbed MaxViT-UNet, for medical image segmentation. The hybrid nature of MaxViT encoder provides locally-globally rich hierarchical features at multiple scales. The proposed Hybrid Decoder effectively utilizes the Hybrid Decoder, based on MaxViT-block, to generate output segmentation masks accurately. The basic unit of the proposed segmentation technique is MaxViT-block, which consists of MBConv-based convolution followed by an efficient and scalable multi-axis attention mechanism (Max-SA). The proposed Hybrid Decoder is lightweight, computationally efficient and designed as a plug-and-play module in UNet-like encoder-decoder architectures. Experiments on the MoNuSeg18 and MoNuSAC20 dataset reveals the effectiveness of the MaxViT hybrid encoder and proposed Hybrid Decoder. Our approach surpassed the previous approaches, specifically CNN-based (UNet) and Transformer-based (Swin-UNet), by a large margin on both datasets in terms of Dice, IoU and PQ metrics.

In this study, we dealt with the segmentation problem of 2d histopathology images only. Possible future direction will be to find the effectiveness of the proposed framework and Hybrid Decoder on other 2d/3d imaging modalities and real-life datasets. The concepts of channel boosting and ensembelling of the deep neural networks can be investigated in future for developing more robust and generalized segmentation techniques.

Acknowledgements

References

[1] Neeraj Kumar, Ruchika Verma, Deepak Anand, Yanning Zhou, Omer Fahri Onder, Efstratios Tsougenis, Hao Chen, Pheng-Ann Heng, Jiahui Li, Zhiqiang Hu, et al. A multi-organ nucleus segmentation challenge. *IEEE transactions on medical imaging*, 39(5):1380–1391, 2019.

[2] Umm-e-Hani Tayyab, Faiza Babar Khan, Muhammad Hanif Durad, Asifullah Khan, and Yeon Soo Lee. A survey of the recent trends in deep learning based malware detection. *Journal of Cybersecurity and Privacy*, 2(4):800–829, 2022.

[3] Anabia Sohail, Bibi Ayisha, Irfan Hameed, Muhammad Mohsin Zafar, and Asifullah Khan. Deep neural networks based meta-learning for network intrusion detection. *arXiv preprint arXiv:2302.09394*, 2023.

[4] Asifullah Khan, Saddam Hussain Khan, Mahrukh Saif, Asiya Batool, Anabia Sohail, and Muhammad Waleed Khan. A survey of deep learning techniques for the analysis of covid-19 and their usability for detecting omicron. *Journal of Experimental & Theoretical Artificial Intelligence*, pages 1–43, 2023.

[5] 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.

[6] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *Medical Image Computing and Computer-Assisted Intervention–MICCAI 2015: 18th International Conference, Munich, Germany, October 5-9, 2015, Proceedings, Part III 18*, pages 234–241. Springer, 2015.

[7] 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. In *European Conference on Computer Vision*, pages 205–218. Springer, 2022.

[8] Zunaira Rauf, Anabia Sohail, Saddam Hussain Khan, Asifullah Khan, Jeonghwan Gwak, and Muhammad Maqbool. Attention-guided multi-scale deep object detection framework for lymphocyte analysis in ihc histological images. *Microscopy*, 72(1):27–42, 2023.

[9] Momina Liaqat Ali, Zunaira Rauf, Abdur Rehman Khan, and Asifullah Khan. Channel boosting based detection and segmentation for cancer analysis in histopathological images. In *2022 19th International Bhurban Conference on Applied Sciences and Technology (IBCAST)*, pages 1–6. IEEE, 2022.

[10] Abdullah Aziz, Anabia Sohail, Labiba Fahad, Muhammad Burhan, Noorul Wahab, and Asifullah Khan. Channel boosted convolutional neural network for classification of mitotic nuclei using histopathological images. In *2020 17th International Bhurban Conference on Applied Sciences and Technology (IBCAST)*, pages 277–284. IEEE, 2020.

[11] Anabia Sohail, Asifullah Khan, Humaira Nisar, Sobia Tabassum, and Aneela Zameer. Mitotic nuclei analysis in breast cancer histopathology images using deep ensemble classifier. *Medical image analysis*, 72:102121, 2021.

[12] Saddam Hussain Khan, Asifullah Khan, Yeon Soo Lee, Mehdi Hassan, and Woong Kyo Jeong. Segmentation of shoulder muscle mri using a new region and edge based deep auto-encoder. *Multimedia Tools and Applications*, 82(10):14963–14984, 2023.

[13] Asifullah Khan, Zunaira Rauf, Anabia Sohail, Abdul Rehman, Hifsa Asif, Aqsa Asif, and Umair Farooq. A survey of the vision transformers and its cnn-transformer based variants. *arXiv preprint arXiv:2305.09880*, 2023.

[14] Baihua Zhang, Shouliang Qi, Yanan Wu, Xiaohuan Pan, Yudong Yao, Wei Qian, and Yubao Guan. Multi-scale segmentation squeeze-and-excitation unet with conditional random field for segmenting lung tumor from ct images. *Computer Methods and Programs in Biomedicine*, 222:106946, 2022.

[15] Quoc Dang Vu and Jin Tae Kwak. A dense multi-path decoder for tissue segmentation in histopathology images. *Computer methods and programs in biomedicine*, 173:119–129, 2019.

[16] Yatong Liu, Yu Zhu, Ying Xin, Yanan Zhang, Dawei Yang, and Tao Xu. Mestrans: Multi-scale embedding spatial transformer for medical image segmentation. *Computer Methods and Programs in Biomedicine*, 233:107493, 2023.

[17] Asifullah Khan, Anabia Sohail, Umme Zahoora, and Aqsa Saeed Qureshi. A survey of the recent architectures of deep convolutional neural networks. *Artificial intelligence review*, 53:5455–5516, 2020.

[18] Nabil Ibtehaz and M Sohel Rahman. Multiresunet: Rethinking the u-net architecture for multimodal biomedical image segmentation. *Neural networks*, 121:74–87, 2020.

[19] Zongwei Zhou, Md Mahfuzur Rahman Siddiquee, Nima Tajbakhsh, and Jianming Liang. Unet++: Redesigning skip connections to exploit multiscale features in image segmentation. *IEEE transactions on medical imaging*, 39(6):1856–1867, 2019.

[20] Ozan Oktay, Jo Schlemper, Loic Le Folgoc, Matthew Lee, Mattias Heinrich, Kazunari Misawa, Kensaku Mori, Steven McDonagh, Nils Y Hammerla, Bernhard Kainz, Ben Glocker, and Daniel Rueckert. Attention u-net: Learning where to look for the pancreas. In *Medical Imaging with Deep Learning*, 2018.

[21] Özgün Çiçek, Ahmed Abdulkadir, Soeren S Lienkamp, Thomas Brox, and Olaf Ronneberger. 3d u-net: learning dense volumetric segmentation from sparse annotation. In *Medical Image Computing and Computer-Assisted Intervention–MICCAI 2016: 19th International Conference, Athens, Greece, October 17-21, 2016, Proceedings, Part II 19*, pages 424–432. Springer, 2016.

[22] Xiaodong Yang, Houqiang Li, and Xiaobo Zhou. Nuclei segmentation using marker-controlled watershed, tracking using mean-shift, and kalman filter in time-lapse microscopy. *IEEE Transactions on Circuits and Systems I: Regular Papers*, 53(11):2405–2414, 2006.

[23] Mitko Veta, Paul J Van Diest, Robert Kornegoor, André Huisman, Max A Viergever, and Josien PW Pluim. Automatic nuclei segmentation in h&e stained breast cancer histopathology images. *PloS one*, 8(7):e70221, 2013.

[24] Andy Tsai, Anthony Yezzi, William Wells, Clare Tempany, Dewey Tucker, Ayres Fan, W Eric Grimson, and Alan Willsky. A shape-based approach to the segmentation of medical imagery using level sets. *IEEE transactions on medical imaging*, 22(2):137–154, 2003.

[25] Karsten Held, E Rota Kops, Bernd J Krause, William M Wells, Ron Kikinis, and H-W Muller-Gartner. Markov random field segmentation of brain mr images. *IEEE transactions on medical imaging*, 16(6):878–886, 1997.

[26] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations*, 2021.

[27] 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.

[28] Jeya Maria Jose Valanarasu, Poojan Oza, Ilker Hacihaliloglu, and Vishal M Patel. Medical transformer: Gated axial-attention for medical image segmentation. In *Medical Image Computing and Computer Assisted Intervention–MICCAI 2021: 24th International Conference, Strasbourg, France, September 27–October 1, 2021, Proceedings, Part I 24*, pages 36–46. Springer, 2021.

[29] Haonan Wang, Peng Cao, Jiaqi Wang, and Osmar R Zaiane. Uctransnet: rethinking the skip connections in u-net from a channel-wise perspective with transformer. In *Proceedings of the AAAI conference on artificial intelligence*, volume 36, pages 2441–2449, 2022.

[30] Yundong Zhang, Huiye Liu, and Qiang Hu. Transfuse: Fusing transformers and cnns for medical image segmentation. In *Medical Image Computing and Computer Assisted Intervention–MICCAI 2021: 24th International Conference, Strasbourg, France, September 27–October 1, 2021, Proceedings, Part I 24*, pages 14–24. Springer, 2021.

[31] Xuewei Li, Shuo Pang, Ruixuan Zhang, Jialin Zhu, Xuzhou Fu, Yuan Tian, and Jie Gao. Attransunet: An enhanced hybrid transformer architecture for ultrasound and histopathology image segmentation. *Computers in Biology and Medicine*, 152:106365, 2023.

[32] Jianyuan Guo, Kai Han, Han Wu, Yehui Tang, Xinghao Chen, Yunhe Wang, and Chang Xu. Cmt: Convolutional neural networks meet vision transformers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12175–12185, 2022.

[33] Jiashi Li, Xin Xia, Wei Li, Huixia Li, Xing Wang, Xuefeng Xiao, Rui Wang, Min Zheng, and Xin Pan. Next-vit: Next generation vision transformer for efficient deployment in realistic industrial scenarios. *arXiv preprint arXiv:2207.05501*, 2022.

[34] Chang Yao, Menghan Hu, Qingli Li, Guangtao Zhai, and Xiao-Ping Zhang. Transclaw u-net: claw u-net with transformers for medical image segmentation. In *2022 5th International Conference on Information Communication and Signal Processing (ICICSP)*, pages 280–284. IEEE, 2022.

[35] Yuanfeng Ji, Ruimao Zhang, Huijie Wang, Zhen Li, Lingyun Wu, Shaoting Zhang, and Ping Luo. Multi-compound transformer for accurate biomedical image segmentation. In *Medical Image Computing and Computer Assisted Intervention–MICCAI 2021: 24th International Conference, Strasbourg, France, September 27–October 1, 2021, Proceedings, Part I 24*, pages 326–336. Springer, 2021.

[36] 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.

[37] Zhengzhong Tu, Hossein Talebi, Han Zhang, Feng Yang, Peyman Milanfar, Alan Bovik, and Yinxiao Li. Maxvit: Multi-axis vision transformer. In *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XXIV*, pages 459–479. Springer, 2022.

[38] Huazhu Fu, Jun Cheng, Yanwu Xu, Damon Wing Kee Wong, Jiang Liu, and Xiaochun Cao. Joint optic disc and cup segmentation based on multi-label deep network and polar transformation. *IEEE transactions on medical imaging*, 37(7):1597–1605, 2018.

[39] Iqra Kiran, Basit Raza, Areesha Ijaz, and Muazzam A Khan. Denseres-unet: Segmentation of overlapped/clustered nuclei from multi organ histopathology images. *Computers in Biology and Medicine*, 143:105267, 2022.

[40] Ran Gu, Guotai Wang, Tao Song, Rui Huang, Michael Aertsen, Jan Deprest, Sébastien Ourselin, Tom Vercauteren, and Shaoting Zhang. Ca-net: Comprehensive attention convolutional neural networks for explainable medical image segmentation. *IEEE transactions on medical imaging*, 40(2):699–711, 2020.

[41] Caiyong Wang, Yunlong Wang, Yunfan Liu, Zhaofeng He, Ran He, and Zhenan Sun. Sclerasegnet: An attention assisted u-net model for accurate sclera segmentation. *IEEE Transactions on Biometrics, Behavior, and Identity Science*, 2(1):40–54, 2019.

[42] Shyam Lal, Devikalyan Das, Kumar Alabhya, Anirudh Kanfade, Aman Kumar, and Jyoti Kini. Nucleisegnet: robust deep learning architecture for the nuclei segmentation of liver cancer histopathology images. *Computers in Biology and Medicine*, 128:104075, 2021.

[43] 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.

[44] Momina Liaqat Ali, Zunaira Rauf, Asifullah Khan, Anabia Sohail, Rafi Ullah, and Jeonghwan Gwak. Cb-hvtnet: A channel-boosted hybrid vision transformer network for lymphocyte assessment in histopathological images. *arXiv preprint arXiv:2305.09211*, 2023.

[45] Davood Karimi, Serge Didenko Vasylechko, and Ali Gholipour. Convolution-free medical image segmentation using transformers. In *Medical Image Computing and Computer Assisted Intervention–MICCAI 2021: 24th International Conference, Strasbourg, France, September 27–October 1, 2021, Proceedings, Part I 24*, pages 78–88. Springer, 2021.

[46] Mark Sandler, Andrew Howard, Menglong Zhu, Andrey Zhmoginov, and Liang-Chieh Chen. Mobilenetv2: Inverted residuals and linear bottlenecks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4510–4520, 2018.

[47] Tete Xiao, Mannat Singh, Eric Mintun, Trevor Darrell, Piotr Dollár, and Ross Girshick. Early convolutions help transformers see better. *Advances in Neural Information Processing Systems*, 34:30392–30400, 2021.

[48] 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.

[49] Xiangxiang Chu, Zhi Tian, Bo Zhang, Xinlong Wang, and Chunhua Shen. Conditional positional encodings for vision transformers. In *The Eleventh International Conference on Learning Representations*, 2023.

[50] Jimmy Lei Ba, Jamie Ryan Kiros, and Geoffrey E Hinton. Layer normalization. *arXiv preprint arXiv:1607.06450*, 2016.

[51] Sergey Ioffe and Christian Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In *International conference on machine learning*, pages 448–456. pmlr, 2015.

[52] Dan Hendrycks and Kevin Gimpel. Gaussian error linear units (gelus). *arXiv preprint arXiv:1606.08415*, 2016.

[53] 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.

[54] Zhengzhong Tu, Hossein Talebi, Han Zhang, Feng Yang, Peyman Milanfar, Alan Bovik, and Yinxiao Li. Maxim: Multi-axis mlp for image processing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5769–5780, 2022.

[55] Long Zhao, Zizhao Zhang, Ting Chen, Dimitris Metaxas, and Han Zhang. Improved transformer for high-resolution gans. *Advances in Neural Information Processing Systems*, 34:18367–18380, 2021.

[56] Kai Han, An Xiao, Enhua Wu, Jianyuan Guo, Chunjing Xu, and Yunhe Wang. Transformer in transformer. *Advances in Neural Information Processing Systems*, 34:15908–15919, 2021.

[57] Zihang Dai, Hanxiao Liu, Quoc V Le, and Mingxing Tan. Coatnet: Marrying convolution and attention for all data sizes. *Advances in Neural Information Processing Systems*, 34:3965–3977, 2021.

[58] Peter Shaw, Jakob Uszkoreit, and Ashish Vaswani. Self-attention with relative position representations. *arXiv preprint arXiv:1803.02155*, 2018.

[59] Yifan Jiang, Shiyu Chang, and Zhangyang Wang. Transgan: Two pure transformers can make one strong gan, and that can scale up. *Advances in Neural Information Processing Systems*, 34:14745–14758, 2021.

[60] Diganta Misra. Mish: A self regularized non-monotonic activation function. *arXiv preprint arXiv:1908.08681*, 2019.

[61] Ruchika Verma, Neeraj Kumar, Abhijeet Patil, Nikhil Cherian Kurian, Swapnil Rane, Simon Graham, Quoc Dang Vu, Mieke Zwager, Shan E Ahmed Raza, Nasir Rajpoot, et al. Monusac2020: A multi-organ nuclei segmentation and classification challenge. *IEEE Transactions on Medical Imaging*, 40(12):3413–3423, 2021.

[62] TCGA. Network data. http://cancergenome.nih.gov/, 2006.

[63] MMSegmentation Contributors. Openmmlab semantic segmentation toolbox and benchmark, 2020.

[64] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Kopf, Edward Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. Pytorch: An imperative style, high-performance deep learning library. In *Advances in Neural Information Processing Systems 32*, pages 8024–8035. Curran Associates, Inc., 2019.

[65] Tiange Xiang, Chaoyi Zhang, Dongnan Liu, Yang Song, Heng Huang, and Weidong Cai. Bio-net: learning recurrent bi-directional connections for encoder-decoder architecture. In *Medical Image Computing and Computer Assisted Intervention–MICCAI 2020: 23rd International Conference, Lima, Peru, October 4–8, 2020, Proceedings, Part I 23*, pages 74–84. Springer, 2020.

[66] Bangcheng Zhan, Enmin Song, and Hong Liu. Fsa-net: Rethinking the attention mechanisms in medical image segmentation from releasing global suppressed information. *Computers in Biology and Medicine*, page 106932, 2023.

[67] JunBo Qiao, Xing Wang, Ji Chen, and MingTao Liu. Mbutransnet: multi-branch u-shaped network fusion transformer architecture for medical image segmentation. *International Journal of Computer Assisted Radiology and Surgery*, pages 1–8, 2023.

[68] Ji Wang, Lulu Qin, Dan Chen, Juan Wang, Bo-Wei Han, Zexuan Zhu, and Guangdong Qiao. An improved hover-net for nuclear segmentation and classification in histopathology images. *Neural Computing and Applications*, pages 1–15, 2023.

[69] Vi Thi-Tuong Vo and Soo-Hyung Kim. Mulvernet: Nucleus segmentation and classification of pathology images using the hover-net and multiple filter units. *Electronics*, 12(2):355, 2023.

[70] Zuhao Liu, Huan Wang, Shaoting Zhang, Guotai Wang, and Jin Qi. Nas-scam: Neural architecture search-based spatial and channel joint attention module for nuclei semantic segmentation and classification. In *Medical Image Computing and Computer Assisted Intervention–MICCAI 2020: 23rd International Conference, Lima, Peru, October 4–8, 2020, Proceedings, Part I 23*, pages 263–272. Springer, 2020.

[71] 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.

[72] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

[73] Juntang Zhuang, Tommy Tang, Yifan Ding, Sekhar C Tatikonda, Nicha Dvornek, Xenophon Papademetris, and James Duncan. Adabelief optimizer: Adapting stepsizes by the belief in observed gradients. *Advances in neural information processing systems*, 33:18795–18806, 2020.

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

[75] Tete Xiao, Yingcheng Liu, Bolei Zhou, Yuning Jiang, and Jian Sun. Unified perceptual parsing for scene understanding. In *Proceedings of the European conference on computer vision (ECCV)*, pages 418–434, 2018.