The following publication H. Xie et al., "SATR: A Structure-Affinity Attention-Based Transformer Encoder for Spine Segmentation," 2024 IEEE International Symposium on Biomedical Imaging (ISBI), Athens, Greece, 2024, pp. 1-5 is available at https://doi.org/10.1109/ISBI56570.2024.10635612.

SATR: A STRUCTURE-AFFINITY ATTENTION-BASED TRANSFORMER ENCODER FOR SPINE SEGMENTATION

Hao Xie[1] *Zixun Huang*[1] *Frank H. F. Leung*[1] *N. F. Law*[1] *Yakun Ju*[3] *Yong-Ping Zheng*[2] *Sai Ho Ling*[4]

[1]Department of Electrical and Electronic Engineering, [2]Department of Biomedical Engineering
The Hong Kong Polytechnic University, Hong Kong, China
[3]School of Electrical and Electronic Engineering, Nanyang Technological University, Singapore
[4]School of Electrical and Data Engineering, University of Technology Sydney, NSW, Australia

ABSTRACT

In digital histopathology, spine segmentation on ultrasound images plays a vital role, especially as a pre-processing filter to measure spine deformity and diagnose scoliosis automatically. This segmentation task remains challenging owing to the lack of consideration of high spatial correlation for different bone features. In this paper, in order to encode the rich prior knowledge regarding their structural attributes and spatial relationships, we propose a novel structure-affinity attention-based transformer encoder (SATR) to segment spine. It exploits the hierarchical architecture to output multi-scale feature representations. Meanwhile, the constraint on spine structural information enhances the feature usability of the network and consequently improves the segmentation accuracy. The comparative experiments verify that SATR achieves promising performance on spine segmentation as compared with other state-of-the-art candidates, which makes it conveniently replace the backbone networks for intelligent scoliosis assessment.

Index Terms— Spine Segmentation, Structure-Affinity Attention, Transformer Architecture, Scoliosis Diagnosis

1. INTRODUCTION

Scoliosis is medically defined as a lateral curvature of the spine exceeding 10 degrees. The demographic group at the highest risk for scoliosis is that of adolescents who are on their bone development stage [1]. Adolescent Idiopathic Scoliosis (AIS) accounts for approximately 85% of all scoliosis cases [2]. In contrast, Adult Degenerative Scoliosis (ADS) presents as another form of coronal spine deformity, affects the elderly population without a scoliosis history [3].

Owing to the fact that bone is the tissue with the highest acoustic impedance, ultrasound imaging can be used to visualize and locate the bone surface in surgical operations [4]. For faster diagnosis and better visualization of the spine structure, Volume Projection Imaging (VPI) was proposed to analyse the intensity of all voxels and form coronal 2D images [5]. As a pre-analyzing step for automatic measurement of spine deformity, spine segmentation from ultrasound VPI images provides the basis for intelligent scoliosis diagnosis.

In recent years, with the increasing attention to artificial intelligence (AI), Convolutional Neural Networks (CNNs), particularly convolutional encode-decoder architectures [6], have been applied to extract bone features from ultrasound images in an automated manner. Currently, the UNet model [7] has become the de-facto standard for accurate medical image segmentation. This further motivated researchers to develop extensions for more effective spine segmentation [8], [9]. However, a common characteristic of the aforementioned work is that they are heavily based on the CNN structure, which suffers from a weak global representation learning capability.

To make up for the above deficiency, exploration has been made on self-attention mechanism [10], which enables a single feature from any position to perceive features of all the other positions. Transformers [11] rely on global self-attention mechanisms and were introduced to computer vision tasks, called Vision Transformers (ViT), to serve as an alternative to CNNs for image classification [12]. This first pure transformer was applied directly to sequences of *image patches* and obtained comparable and even better results in some tasks than CNNs, such as semantic segmentation [13], [14]. However, pure transformer-based methods have not been widely applied in medical image segmentation due to the much higher resolution of pixels in medical images that requires dense prediction at the pixel level. Moreover, ViT outputs single-scale low-resolution features instead of multi-scale ones. To overcome these limitations, Xie *et al.* proposed Segformer [15], a hierarchical architecture that enables the encoder to generate both high-resolution fine features and low-resolution coarse features. Swin Transformer [16] constructed a hierarchical representation by starting from small-sized patches and gradually merging neighboring patches in deeper transformer layers.

Different bone features show high spatial correlation and only appears in some regions in the image. However, only limited exploration has been made to utilize the structural at-

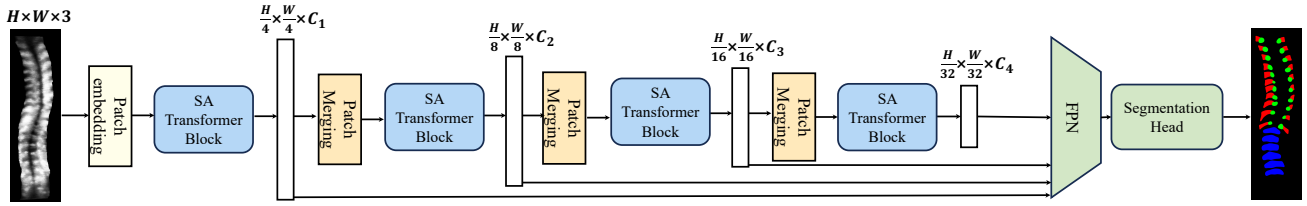

Fig. 1. Proposed SATR framework for spine segmentation consists of two main modules: Hierarchical SA transformer blocks to extract high/low resolution features; and a segmentation network (feature pyramid network (FPN) + head).

tributes and spatial relationships of different bones. For spine segmentation, the strong prior knowledge of shapes and positions of the spine bones deserves to be analysed.

In this paper, we propose a novel structure-affinity hierarchical transformer (SATR) framework to segment spine in ultrasound images more effectively. Same as Swin Transformer, we refine the pyramidal structure to produce high/low-resolution attention maps. The image patches are encoded with larger fields of view compared to conventional CNNs, and are decoded by taking local and global dependency information into account. Critically, in order to encode prior knowledge on the structure of the spine bones into the semantic representations, we utilize the characteristic of capturing semantic-level affinity in the self-attention mechanism [17] and design a structure-affinity attention (SAA) layer embedded in the structure-affinity (SA) transformer block to enrich the learned bone features. We apply this attention layer in each encoder scale to model the multi-resolution feature representation, with the same feature map resolutions as those of typical CNNs. As a result, the proposed architecture could replace the backbone networks for spine segmentation.

Our experimental results demonstrate that the proposed SAA layer is found adept in the transformer encoder. The proposed SATR framework can recognize the spine bones more effectively, which significantly leads to a stable and better spine segmentation performance quantitatively and qualitatively. To summarize, our major contributions are as follows:

- A novel structure-affinity self-attention mechanism to produce structure-affinity feature representations.
- Embedding SAA layer to the hierarchical architecture to propose a novel hierarchical transformer encoder.
- Showing SATR's effectiveness for spine segmentation in ultrasound VPI images, surpassing other transformer-based methods on scoliosis data.

2. METHODOLOGY

An overview of the structure-affinity hierarchical transformer framework is presented in Fig. 1. Given a spine ultrasound VPI image with size $H \times W \times 3$, we first split it into patches of size 4×4 by a patch embedding module. Then, it is projected to an arbitrary dimension C_1, (empirically $C_1 = 192$). Second, four transformer blocks with designed structure-affinity attention layer are applied on these image patches to get multi-level features with output resolution $\{\frac{H}{4} \times \frac{W}{4}, \frac{H}{8} \times \frac{W}{8}, \frac{H}{16} \times \frac{W}{16}, \frac{H}{32} \times \frac{W}{32}\}$. This encoder produces a hierarchical representation, then they are passed into the decoder to predict the segmentation mask.

2.1. Hierarchical Transformer Encoder

The encoder adopts stacked SA transformer blocks and patch merging layers to produce a hierarchical representation as the network gets deeper. By this means, our encoder generates multi-level multi-scale features given an input image. Specifically, after each SA transformer block, we perform patch merging to merge neighbouring patches and obtain a hierarchical feature map f^i with a resolution of $\frac{H}{2^{i+1}} \times \frac{W}{2^{i+1}} \times C_i$, where $i \in \{1, 2, 3, 4\}$, and $C_{i+1} = 2C_i$. This operation can be easily implemented by "nn.Conv2D" in PyTorch. Practically, each layer down-samples the feature representation while up-sampling the channel dimension by a factor of 2. Ultimately, these maps provide both low-resolution coarse features and high-resolution fine features for the segmentation head.

2.2. Structure-Affinity (SA) Transformer Block

Our SA transformer block is built based on regular window (W-) and shifted window (SW-) multi-head self-attention (MSA). The block is composed of two successive transformer sub-blocks as in [16], but a structure-affinity attention (SAA) layer (see details in Sec. 2.3) is employed at the end of each sub-block for the further processing of feature maps to produce spine bone affinity. As illustrated in Fig. 2, each sub-block contains LayerNorm (LN) and 2-layer multi-layer perceptron (MLP). Accordingly, the SA Transformer Block is formulated as:

$$\begin{aligned} \hat{\boldsymbol{f}}^{\boldsymbol{i}} &= \text{W-MSA}(\text{LN}(\boldsymbol{f}^{\boldsymbol{i-1}})) + \boldsymbol{f}^{\boldsymbol{i-1}}, \\ \boldsymbol{f}^{\boldsymbol{i}} &= \text{SAA}(\text{MLP}(\text{LN}(\hat{\boldsymbol{f}}^{\boldsymbol{i}})) + \hat{\boldsymbol{f}}^{\boldsymbol{i}}), \\ \hat{\boldsymbol{f}}^{\boldsymbol{i+1}} &= \text{SW-MSA}(\text{LN}(\boldsymbol{f}^{\boldsymbol{i}})) + \boldsymbol{f}^{\boldsymbol{i}}, \\ \boldsymbol{f}^{\boldsymbol{i+1}} &= \text{SAA}(\text{MLP}(\text{LN}(\hat{\boldsymbol{f}}^{\boldsymbol{i+1}})) + \hat{\boldsymbol{f}}^{\boldsymbol{i+1}}) \end{aligned} \tag{1}$$

where $\hat{\boldsymbol{f}}^{\boldsymbol{i}}$ and $\boldsymbol{f}^{\boldsymbol{i}}$ denote the output features of the (S)W-MSA module and the SAA layer for block $\boldsymbol{i}$, $\boldsymbol{i} \in \{1, 2, 3, 4\}$.

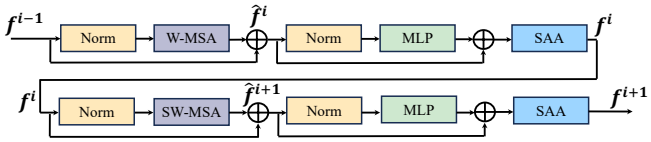

Fig. 2. An illustration of the proposed SA Transformer Block. W-MSA and SW-MSA are multi-head self-attention modules with regular and shifted windows respectively. "MLP" indicates multi-layer perceptron.

2.3. Structure-Affinity Self-Attention Mechanism

In spine images, three spine bones are typically identified: rib, thoracic process, and lump. These bone features exhibit a relatively consistent shape and position across spine images, thereby harboring valuable prior knowledge regarding their structural attributes and spatial relationships. To capture this rich information, we propose a structure-affinity attention layer, which is embedded into the SA transformer block. This layer is designed to acquire and encode prior knowledge into attention maps, subsequently yielding affinity for different spine bones. We employ four attention maps to collect the structural knowledge after taking the categories of bone features and background region into account. This approach enhances the concentration of contextual information of bone features and effectively learns affinity from attention across spine images.

Fig. 3 shows the details of structure-affinity attention layer for the enhancement of spine segmentation. The input is a feature map $\boldsymbol{f_s} \in \mathbb{R}^{C\times H\times W}$, where C represents the number of channels, and H and W correspond to the height and width of the input, respectively. Firstly, we adopt a convolution layer with a kernel size of 1×1 and the reshape operation to generate the query and key representation, denoted as $\boldsymbol{q} \in \mathbb{R}^{C^{'}\times(HW)}$ and $\boldsymbol{k} \in \mathbb{R}^{N\times(HW)}$, respectively. It is worth noting that the biggest difference between the convolution φ and convolution θ is the reduction scale of output channel dimension. C is reduced to $C^{'} = \frac{C}{4}$ via convolution φ in order to reduce the computational complexity. Meanwhile, N represents the number of classes for segmentation, which includes three spine bone features and the background region, i.e., $N = 4$

For the key representation $\boldsymbol{k}$, we have reduced its number of channel to 4, matching the number of classes N through convolution θ. That means each channel map can be regarded as a class-specific spatial response, enabling the precise description of features related to either one foreground spine bone or the background region. Essentially, the self-attention mechanism functions as a directed graphical model [17], where the affinity matrix aligns with the attention map, as points sharing the same structural knowledge are assumed to obtain the equal affinity. Consequently, we produce a novel structure-affinity key representation $\hat{\boldsymbol{k}} \in \mathbb{R}^{(HW)\times N}$ serving as the value representation for pixel-pair in the conventional

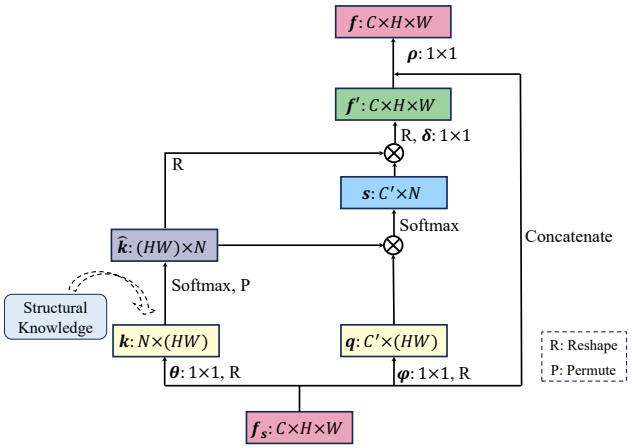

Fig. 3. Proposed structure-affinity attention layer. The feature maps are shown as the shape of their tensors. Proper reshaping or permuting is performed on specific rows. "⊗" denotes matrix multiplication.

self-attention computational process. A matrix multiplication between $\hat{\boldsymbol{k}}$ and $\boldsymbol{q}$ is applied, followed by a softmax layer for proper normalization, to generate the attentive affinity matrix $\boldsymbol{s} \in \mathbb{R}^{C^{'}\times N}$. Next, another matrix multiplication between $\boldsymbol{s}$ and the reshaped $\hat{\boldsymbol{k}}$ leads to the re-estimated structure-affinity features $\boldsymbol{f}^{'} = \delta(\boldsymbol{s} \times \hat{\boldsymbol{k}}) \in \mathbb{R}^{C\times H\times W}$.

This method allows comprehensive acquisition of structural knowledge across spine bones, facilitated by a affinity matrix because the features are directly synthesized with the structure-affinity key representation $\hat{\boldsymbol{k}}$. Ultimately, we concatenate the feature map with the original input, followed by a convolutional mapping ρ to obtain the final output $\boldsymbol{f}$. This propagation process optimally exploits the high-affinity regions of the spine bones while mitigating the influence of wrongly activated areas in ultrasound VPI images.

3. EXPERIMENTS

3.1. Dataset

The dataset is composed of 109 ultrasound VPI images, which are collected from 109 subjects (82 females and 27 males) with varying degrees of spine deformity using the Scolioscan system (Model SCN801, Telefield Medical Imaging Ltd, Hong Kong). Each VPI image is acquired by projecting 3D ultrasound scanning of the whole spine region into a 2D coronal plane. Accurate ground-truth segments are manually labelled by ultrasound experts. The dataset is split into two sets: a training set with 80 samples, and a testing set with 29 samples. To ensure uniformity, all images are resized to dimensions of 2048×512. During the training stage, we further crop the image size to 512×512 pixels as the input of the transformer encoder for effective feature learning. In the testing stage, the resized samples are passed into the segmentation model to generate the segmentation mask.

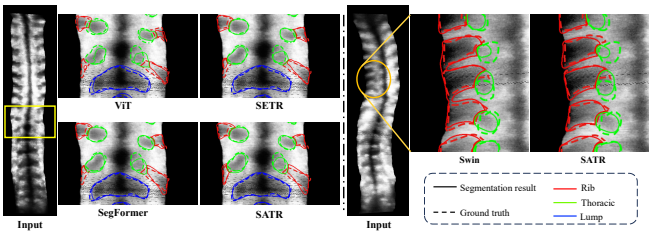

Fig. 4. Visual comparisons of the spine bone segmentation results based on different methods. The segmented rib, thoracic process, and lump are annotated in red, green and blue. The area around the boundary of the thoracic and lumbar region is highlighted in the yellow box, and the orange circle marks the fine structure of rib.

3.2. Implementation Details

We establish our proposed framework based on PyTorch and MMSegmentation. The segmentation head is built based on the same settings in [18]. In the hierarchical transformer encoder, the number of heads in the SA transformer block increases by $\{6, 12, 24, 48\}$ as the network go deeper. The number of channels C in the SAA layer is represented by C_i, $i \in \{1, 2, 3, 4\}$. During training, we build a mini-batch with 4 training samples. The model is trained on a single NVIDIA RTX4090 GPU for 1.6×10^5 iterations by the AdamW optimizer. The learning rate is initialized to 6×10^{-5} and gradually reduced to the minimum leaning rate 0, based on a poly schedule. The weight decay is set to 10^{-2} for regularization.

3.3. Quantitative and Qualitative Results

We evaluate the performance of our proposed SATR framework against UNet [7], the recently proposed convolution-based network of SEAM [9] for ultrasound VPI images, and the state-of-the-art transformer architectures. Table 1 presents the comparative results on scoliosis dataset. According to the results, SATR outperforms both CNN and Transformer-based approaches on nearly all the evaluation metrics. We can even observe a significant improvement of over 2% on the pixel accuracy in the rib and thoracic regions. However, it is worth noting that the evaluation metric of IoU at the area of thoracic process is not satisfactory as compared with one SOTA Transformer-based framework, Swin [16]. We consider that the strong noise in this region confuses the structure-affinity attention layer to discriminate the specific bone features. Broadly, the CNN-based works exceed pure transformer-based methods owing to the limitation of single-scale low-resolution features. However, the introduction of hierarchical architecture enhances the representative ability of attention-based modules. As SATR achieves the best scores, it can be a more preferable approach thanks to its hierarchical architecture and structure-affinity self-attention mechanism.

Furthermore, we provide a qualitative visualization of the spine bone segmentation results in Fig. 4. It can be observed that our proposed SATR provides more accurate and smoother shape of each spine bone at the area around the boundary of the thoracic and lumbar region, which is more similar to the ground-truth segment. Specifically, as compared with Swin [16], the application of SAA layer distinguishes the fine structure of the rib bone features. At the root of the ribs, Swin obfuscates the edge line of each rib, while SATR enables the clear and accurate appearance of segmented bone area in the boundary of the image, although the occupied area is small.

Table 1. Performance Comparison of Different Methods on Spine Segmentation Task, where Dice:Dice Score(%), IoU:Intersection over Union(%) and Acc:Pixel Accuracy(%)

Modules	Rib			Thoracic			Lump			Ave.		
	Dice	IoU	Acc	Dice	IoU	Acc	Dice	IoU	Acc	Dice	IoU	Acc
UNet [7]	78.38	65.92	80.28	77.45	63.39	77.30	85.85	75.52	88.24	80.86	68.28	81.94
SEAM [9]	77.79	65.83	79.72	76.36	64.24	72.34	84.40	76.52	87.91	79.52	69.68	79.99
ViT [12]	77.86	63.74	73.27	75.86	60.79	72.33	79.20	65.56	75.75	82.08	70.44	79.54
SETR [13]	80.34	67.14	80.29	78.46	64.56	77.49	86.48	76.18	86.99	85.44	75.27	85.32
SegFormer [15]	79.13	65.47	78.03	77.81	63.67	78.82	81.56	68.87	78.28	83.67	72.66	82.93
Swin [16]	80.47	67.43	80.56	78.89	**65.41**	77.99	86.53	76.31	88.20	85.57	75.59	85.91
SATR (add)	80.74	67.70	80.89	78.98	65.26	78.19	86.89	76.82	88.35	85.80	75.80	85.98
SATR (Ours)	**80.92**	**67.95**	**82.20**	**79.14**	65.31	**80.15**	**87.03**	**77.04**	**89.54**	**85.81**	**75.81**	**86.59**

3.4. Ablation Study

At the end of structure-affinity attention layer, a "concatenate" operation is adopted between the structure-affinity feature $\boldsymbol{f}'$ and the original input $\boldsymbol{f_s}$, followed by a convolution layer. In the ablation study, we employ the "add" operation instead of the above ones to explore the influence of different feature fusion methods on the model performance. The final output representation $\boldsymbol{f}$ is directly obtained without the convolution, denoted as "SATR (add)". The results observed from Table 1 show that although the refined SATR framework surpasses Swin Transformer by a certain margin, the whole performance still lags behind owing to the fact that the rich global context of each channel is captured via the "concatenate" operation, enhancing the representation capability of some important channel maps.

4. CONCLUSION

This paper presents a structure-affinity attention-based transformer, SATR, which produces a hierarchical bone feature representation for effective spine segmentation. Specifically, in order to make full use of the structural information of spine bone, we propose the structure-affinity attention layer, embedding it into the hierarchical transformer encoder. The quantitative and qualitative results show that our model SATR achieves more accurate segmentation results on the spine ultrasound images, significantly surpassing previous transformer-based methods. We hope that our framework can serve as a solid baseline to segment spine for automatic scoliosis diagnosis and motivate further research in the future.

5. COMPLIANCE WITH ETHICAL STANDARDS

This study was performed in line with the principles of the Declaration of Helsinki. Approval was granted by the Ethics Committee of The Hong Kong Polytechnic University (06 Sep 2018/HSEARS20180906005).

6. ACKNOWLEDGMENT

This work was supported by a grant from the Research Grants Council of the Hong Kong Special Administrative Region, China, under Project B-Q86J.

7. REFERENCES

[1] Halima Shakil, Zaheen A Iqbal, and Ahmad H Al-Ghadir, "Scoliosis: review of types of curves, etiological theories and conservative treatment," *Journal of back and musculoskeletal rehabilitation*, vol. 27, no. 2, pp. 111–115, 2014.

[2] John P Horne, Robert Flannery, and Saif Usman, "Adolescent idiopathic scoliosis: diagnosis and management," *American family physician*, vol. 89, no. 3, pp. 193–198, 2014.

[3] Zhibo Song, Zhaoquan Zhang, et al., "Mid-and long-term comparison analysis of two approaches for the treatment of level iii or higher lenke–silva adult degenerative scoliosis: Radical or limited surgery?," *Orthopaedic Surgery*, vol. 14, no. 9, pp. 2006–2015, 2022.

[4] Ilker Hacihaliloglu, "Ultrasound imaging and segmentation of bone surfaces: A review," *Technology*, vol. 5, no. 02, pp. 74–80, 2017.

[5] Chung-Wai James Cheung, Guang-Quan Zhou, Siu-Yin Law, et al., "Ultrasound volume projection imaging for assessment of scoliosis," *IEEE transactions on medical imaging*, vol. 34, no. 8, pp. 1760–1768, 2015.

[6] 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*, 2015, pp. 3431–3440.

[7] 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*. Springer, 2015, pp. 234–241.

[8] Zixun Huang, Li-Wen Wang, et al., "Bone feature segmentation in ultrasound spine image with robustness to speckle and regular occlusion noise," in *2020 IEEE International Conference on Systems, Man, and Cybernetics (SMC)*. IEEE, 2020, pp. 1566–1571.

[9] Rui Zhao, Zixun Huang, Tianshan Liu, et al., "Structure-enhanced attentive learning for spine segmentation from ultrasound volume projection images," in *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2021, pp. 1195–1199.

[10] Xiaolong Wang, Ross Girshick, Abhinav Gupta, and Kaiming He, "Non-local neural networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 7794–7803.

[11] Ashish Vaswani, Noam Shazeer, Niki Parmar, et al., "Attention is all you need," *Advances in neural information processing systems*, vol. 30, 2017.

[12] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, et al., "An image is worth 16x16 words: Transformers for image recognition at scale," *arXiv preprint arXiv:2010.11929*, 2020.

[13] Sixiao Zheng, Jiachen Lu, Hengshuang Zhao, et al., "Rethinking semantic segmentation from a sequence-to-sequence perspective with transformers," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2021, pp. 6881–6890.

[14] Li Zhang, Jiachen Lu, Sixiao Zheng, et al., "Vision transformers: From semantic segmentation to dense prediction," *arXiv preprint arXiv:2012.15840v3*, 2021.

[15] Enze Xie, Wenhai Wang, Zhiding Yu, et al., "Segformer: Simple and efficient design for semantic segmentation with transformers," *Advances in Neural Information Processing Systems*, vol. 34, pp. 12077–12090, 2021.

[16] Ze Liu, Yutong Lin, Yue Cao, et al., "Swin transformer: Hierarchical vision transformer using shifted windows," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2021, pp. 10012–10022.

[17] Lixiang Ru, Yibing Zhan, Baosheng Yu, et al., "Learning affinity from attention: End-to-end weakly-supervised semantic segmentation with transformers," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 16846–16855.

[18] Tete Xiao, Yingcheng Liu, Bolei Zhou, et al., "Unified perceptual parsing for scene understanding," in *Proceedings of the European conference on computer vision (ECCV)*, 2018, pp. 418–434.