This is the Pre-Published Version.

The following publication H. Xie, Z. Huang, F. H. F. Leung, Y. Ju, Y. -P. Zheng and S. H. Ling, "A Structure-Affinity Dual Attention-based Network to Segment Spine for Scoliosis Assessment," 2023 IEEE International Conference on Bioinformatics and Biomedicine (BIBM), Istanbul, Turkiye, 2023, pp. 1567-1574 is available at https://doi.org/10.1109/BIBM58861.2023.10385419.

A Structure-Affinity Dual Attention-based Network to Segment Spine for Scoliosis Assessment

1st Hao Xie
The Hong Kong Polytechnic University
Hong Kong, China
carry-h.xie@connect.polyu.hk

2nd Zixun Huang
The Hong Kong Polytechnic University
Hong Kong, China
zixun.huang@connect.polyu.hk

3rd Frank H. F. Leung
The Hong Kong Polytechnic University
Hong Kong, China
frank-h-f.leung@polyu.edu.hk

4th Yakun Ju*
The Hong Kong Polytechnic University
Hong Kong, China
kelvin.yakun.ju@gmail.com

5th Yong-Ping Zheng
The Hong Kong Polytechnic University
Hong Kong, China
yongping.zheng@polyu.edu.hk

6th Sai Ho Ling
University of Technology Sydney
NSW, Australia
steve.ling@uts.edu.au

***Abstract*—Ultrasound volume projection imaging has shown great promise to visualize spine features and diagnose scoliosis thanks to its harmlessness, cheapness, and efficiency. The key to measuring spine deformity and assessing scoliosis is to accurately segment the spine bone features. In this paper, we propose a novel structure-affinity dual attention-based network (SADANet) for effective spine segmentation. Global channel attention module and spatial criss-cross attention module are combined in a parallel manner to generate rich global context of spine images. Meanwhile, we present a structure-affinity strategy to encode the structural knowledge of spine bones into the semantic representations. By this means, the network can capture both contextual and structural information. Experiments show that our proposed algorithm achieves promising performance on spine segmentation as compared with other state-of-the-art candidates, which makes it an appealing approach for intelligent scoliosis assessment.**

***Index Terms*—Spine Segmentation, Structure-Affinity Dual Attention, Ultrasound volume Projection Imaging, Intelligent scoliosis diagnosis**

I. Introduction

Scoliosis is a medical condition in which the spinal cord gets severely deformed over time. It not only affects the appearance and cardiopulmonary function of the patient, but can also be a cause of psychological impact [1]. Currently, the common practice of scoliosis diagnosis involves measuring the Cobb Angle via radiography [2]. However, the ionizing radiation of X-rays is harmful to the patients. Specifically, radiographic measurements are required not only preoperatively but also postoperatively during the whole treatment [3]. A radiation-free measurement alternative to X-rays suitable for mass screening has become crucial.

Since bone is the tissue with the highest acoustic impedance in human tissues, ultrasound imaging can be used to visualize and locate the bone surface in surgical operations and clinical procedures [4]. Thanks to its advantages, such as no ionizing radiation, low cost, and real-time operation, ultrasound imaging is increasingly becoming a popular imaging method and is widely accepted as a safer alternative to fluorescence imaging [5]. In clinical scoliosis diagnosis, experts need to view hundreds of ultrasound images in a sequence of the whole spine region. This process is tedious and time-consuming [6]. For faster diagnosis and better visualization of the spine structure, Volume Projection Imaging (VPI) was proposed to analyse the intensity of all voxels in the ultrasound sequence and form coronal 2D images [7]. However, ultrasound images suffer from low contrast and tends to contain speckle noise. The quality of ultrasound images is inconsistent owing to variations in imaging equipment and scanning operators [8]. The diagnosis requires the examiners to have extensive experience. Yet, the subjective factors behind personal experience are inevitable in manual scoliosis diagnosis. Therefore, the analysis of ultrasound-based bone feature extraction should

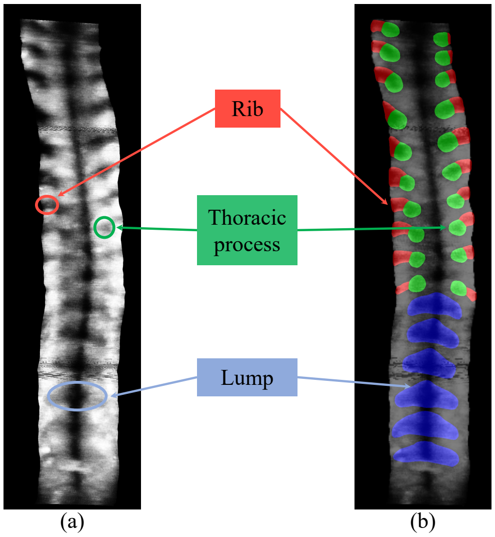

Fig. 1. An illustration of spine segmentation from ultrasound VPI images. (a) One ultrasound VPI image from a patient with scoliosis; (b) Different bone features in the spine image. The red and green regions represent the segmented rib and thoracic process. In the lumbar region, the lump, which is formed by the combined shadow of the partial bilateral inferior articular process, laminae, and the superior articular process of the inferior vertebrae, is annotated in blue.

[*] Yakun Ju is the corresponding author.

© 2023 IEEE. Personal use of this material is permitted. Permission from IEEE must be obtained for all other uses, in any current or future media, including reprinting/republishing this material for advertising or promotional purposes, creating new collective works, for resale or redistribution to servers or lists, or reuse of any copyrighted component of this work in other works.

better be fully automatic.

In recent years, with the increasing attention to artificial intelligence (AI) and deep learning, some medical image processing techniques have been applied to the diagnosis of scoliosis. As a pre-analyzing step for intelligent scoliosis diagnosis, automatic spine segmentation from ultrasound VPI images provides the basis for the measurement of spine deformity. The extraction of spine bone features is shown in Fig. 1. Currently, Convolutional Neural Networks (CNNs) have become the de-facto standard for accurate medical image segmentation. Fully Convolutional Networks (FCNs) [24], particularly convolutional encode-decoder networks, have drawn much attention [37], [40]. Great efforts have been made to investigate effective backbone architectures [9], [10], [43] and learning algorithms [11], [12], [44]. Owing to the nature of the convolutional operation, it cannot capture the long-range dependencies across different features but only obtains the local receptive fields [13] and short-range contextual information, which imposes a great adverse effect to networks owing to insufficient understanding of surrounded contextual information. To make up for the above deficiency, exploration has been made on self-attention mechanism [14], [39], [41], which enables a single feature from any position to perceive features of all the other positions. Motivated by the effectiveness of self-attention mechanism, Fu *et al.* [25] proposed to combine channel attention to capture the channel-wise interactions. The dual attention schemes [26], [27], [38], [45] have been shown to improve performance on different vision tasks.

Specifically, different bone features show high spatial correlation, and only appear in some regions in the ultrasound image. For spine segmentation, the strong prior knowledge of shapes and positions of the spine bones deserved to be considered. Motivated by the above discussion, we propose a novel structure-affinity dual spatial-channel attention network (SADANet) to effectively segment the bone feature in an ultrasound spine image. First, 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 [32]. We also propose a structure-affinity attention (SAA) module, and embed it as an auxiliary task into the spine segmentation network to enrich the learned bone features for more effective spine segmentation.

Furthermore, we introduce a dual attention mechanism to extract channel and spatial-wise dependencies across bone features. We propose a global channel attention (GCA) module and a spatial criss-cross attention (SCA) module, which are at the end of the backbone. GCA module is used to capture the global context of each channel and pay more attention to some important channels. SCA module is an enhancement of criss-cross attention [29], which only has sparse connections ($H+W-1$) for each position in the feature maps, where $H \times W$ denotes the spatial dimension of input feature maps. Specifically, the channel and spatial attention module are integrated in a parallel manner to capture bone feature dependencies in the channel and spatial dimensions respectively. This dual attention block can enhance the inter-class discrimination and intra-class responsiveness, and further extract long-range contextual representations by capturing the full-image spine information.

The resultant model can more effectively localize and recognize the spine bones in ultrasound images. Through experiments and studies, the proposed SAA module is found adept in training the spine segmentation network. Our proposed SADANet is beneficial to both the visual quality and segmentation accuracy of the spine bone features, achieving a stable and better performance than other state-of-the-art segmentation algorithms on ultrasound images.

The main contributions of this paper are summarized as follows:

- We employ the dual spatial-channel attention block to enhance the representative ability of feature maps by capturing rich global context and making an effective use of the multi-channel space for feature representation.
- We consider the structural information of different bones and propose a structure-affinity attention module as an auxiliary module to produce the structure-affinity contextual representations for more effective spine segmentation.
- We integrate three attention modules and propose a novel spine segmentation network SADANet, which provides better spine segmentation results on ultrasound images in terms of quantity and quality.

II. RELATED WORK

A. Spine Segmentation with Ultrasound

In the traditional measurement of scoliosis with tracked ultrasound, experts need to mark different bone features, including rib, thoracic process, and lump, in the VPI images. Many algorithms based on machine learning and deep learning were proposed to extract bone features in ultrasound images automatically [42]. Berton *et al.* [15] utilized an LDA classifier to extract the spinous process and acoustic shadow. However, when assessing scoliosis with ultrasound, the expert should consider not only the spinous processes but also the transverse process and the laminae [16]. The output in [15] cannot be effectively used to estimate the spine deformity. Based on volume projection imaging technique, more reliable approaches were proposed to compute the spine deformity using the paired thoracic processes and lumbar vertebrae. In [17], [18], it has been suggested that the transverse process (TP) measurement method can be used to measure spinal deformation. The methodology of TP measurement is to detect the bone features in an ultrasound scan. Recently, UNet [19] has been widely used in medical image segmentation tasks owing to its superior performance. It was utilized to segment all the bone features based on 2D ultrasound spine images automatically in an end-to-end manner. However, since the segmentation is based on 2D transverse images that are processed independently, the reconstructed images are of low quality and contain many incoherent structures. Huang *et al.* [20] introduced a total variance loss function into the UNet architecture to address the occlusion issue in VPI images. Banerjee *et al.* [28] proposed a lightweight UNet to perform effective spine segmentation with a low computational burden. Zhao *et al.* [35] introduced a structure supervision to the representation learning. These motivate us to investigate a more efficient learning strategy in this paper that can perform spine bone segmentation.

B. Attention Model

Attention model can capture long-range dependencies and is widely used for various tasks. In particular, the work [21] is the first to propose the self-attention mechanism to draw global dependencies of inputs and applies it in machine translation. Meanwhile, attention model are increasingly applied in the image/vision field. Wang *et al.* [14] proposed a non-local module to generate the huge attention map by calculating the correlation matrix between each spatial point in the feature maps, then guided a contextual information aggregation. In image segmentation, Ding *et al.* [22] presented a hierarchical attention network for effective medical image segmentation. Expectation-Maximization (EM) Attention network [23] aggregated the EM attention into the attentive learning framework to enhance the semantic representations. Fu *et al.* [25] proposed DANet for scene segmentation, where a channel and a position attention module were integrated at the end of a dilated FCN to model the semantic dependencies in both position and channel dimensions. Different from previous works, we refine the criss-cross attention module [29] to ensure that each position in the feature maps is sparsely connected with other ones which are in the same row and column, leading to fewer weights of the predicted attention map. Meanwhile, we propose the SAA module as an auxiliary module to produce the structure-affinity representations and achieve the effect of structure-affinity.

III. METHODOLOGY

In this section, we present the details of the proposed framework with structure-affinity for spine segmentation, including the different learning strategies for different attention modules. We first overview the whole architecture of SADANet, and then introduce the detailed design for each attention module. Finally, we describe how to integrate them together with an appropriate learning strategy for further refinement.

A. Network Architecture

The proposed network architecture is shown in Fig. 2. An input image of spine passes through a pretrained residual network [30], which is employed as the backbone of segmentation model, to produce feature representations with the spatial size of $H \times W$ for pixel-wise prediction. A convolution layer with a kernel size of 3×3 is applied for dimension reduction. Then the features are fed into attention modules.

First, a structure-affinity attention (SAA) module (see details in Sec. III-B) is adopted to produce spine bone affinity by encoding the structural knowledge of different bone regions into the key representation. Then, we design a dual spatial-channel attention block, which is made up of global channel attention (GCA, see details in Sec. III-C) and spatial criss-cross attention (SCA, see details in Sec. III-D) modules, to draw global context over local features. For the GCA module, the channel dependencies between any two channel maps are captured with a weighted sum of all channel maps to enhance the contrast of the features in different channels. Meanwhile, for the SCA module, the feature at a certain position is updated via aggregating features only in horizontal and vertical directions. Thus, two consecutive SCA modules are stacked to harvest full-image contextual information from all pixels, which greatly reduces the complexity in time and space.

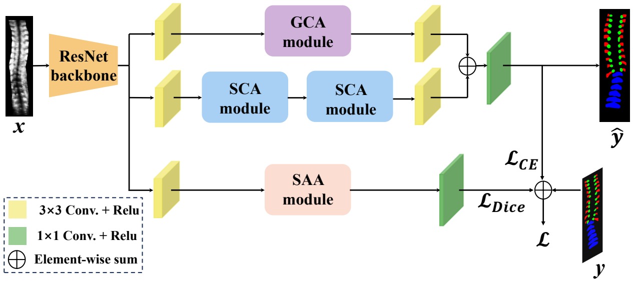

Fig. 2. An illustration of the proposed SADANet. $\boldsymbol{x}$ and $\boldsymbol{y}$ are the input image and its ground-truth segments respectively, $\hat{\boldsymbol{y}}$ denotes the predicted segmentation mask. $\mathcal{L}$ is the constraint of the estimation results in the forward process.

Finally, we transform the outputs of the dual attention block by a convolution layer and perform an element-wise summing to accomplish feature fusion. The last convolution layer with the kernel size of 1×1 is utilized to generate the final prediction map. It is worth noting that the SAA module, as an auxiliary decoder head, only affects the training stage. It outputs the spine bone feature classification results with a specific learning strategy to assist the loss function calculation and optimize the spine segmentation model during the training.

B. Structure-Affinity Attention Module

In a spine image, there are usually three different spine bones, namely rib, thoracic process, and lump. Owing to their relatively uniform shape and position in different spine images, the spine bones contain strong prior knowledge of shapes and positions on the structure. The SAA module is proposed to learn and encode the knowledge into attention maps, producing spine bone affinity, under the supervision of the ground-truth spine images. Considering the categories of bone features and background, we need four attention maps to contain the structural knowledge in order to make contextual information of bone features more concentrated and achieve the effect of affinity on the structure.

As illustrated in Fig. 3, consider a feature map $\boldsymbol{f} \in \mathbb{R}^{C\times H\times W}$, where C, H, W are the number of channels, height and width of the input respectively. We first feed it into a convolution layer with the kernel size of 1×1 to generate the query and the key representation, $\boldsymbol{q} = \theta(\boldsymbol{f}) \in \mathbb{R}^{C'\times H\times W}$ and $\boldsymbol{k} = \varphi(\boldsymbol{f}) \in \mathbb{R}^{N\times H\times W}$. $C = 1024$ is reduced to $C' = \frac{C}{4}$ to reduce the computational complexity. N denotes the number

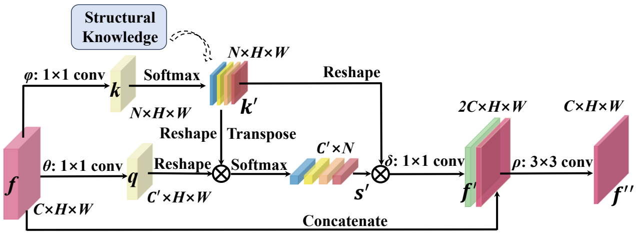

Fig. 3. The details of Structure-Affinity Attention module for spine segmentation

of classes which needs to be segmented, including three spine bone features and the background, i.e., $N = 4$

It is worth noting that in order to introduce structural information into the key representation $\boldsymbol{k}$, the number of channel is reduced to 4, which is equal to the number of classes N. Channel maps can be treated as the multi-spatial responses, and each channel map represents a class-specific spatial response. That means each channel in the key representation can describe the features of one foreground spine bone information or the background, and we can directly produce the bone structure affinity by self-attention mechanism. Essentially, the self-attention mechanism is a kind of directed graphical model [31], while the affinity matrix is usually consistent with attention map since points sharing the same structural knowledge are supposed to be equal. Thus, we produce a novel structure-affinity key representation $\boldsymbol{k}^{'} = Softmax(\boldsymbol{k}) \in \mathbb{R}^{N\times H\times W}$ as the value representation for pixel-pair in conventional self-attention, and perform a matrix multiplication between the transpose of $\boldsymbol{k}^{'}$ and the reshape of $\boldsymbol{q}$ to generate the attentive affinity matrix $\boldsymbol{s}^{'} \in \mathbb{R}^{C^{'}\times N}$ as follows:

$$\boldsymbol{s}^{'} = Softmax(\boldsymbol{q} \times \boldsymbol{k}^{'\boldsymbol{T}}) \tag{1}$$

where the softmax layer performs the normalization as shown in Eq. (1). Then, we perform a matrix multiplication again between $\boldsymbol{s}^{'}$ and the reshape of $\boldsymbol{k}^{'}$ to generate the re-estimated structure-affinity features $\boldsymbol{f}^{'} \in \mathbb{R}^{C\times H\times W}$:

$$\boldsymbol{f}^{'} = \delta(\boldsymbol{s}^{'} \times \boldsymbol{k}^{'}) \tag{2}$$

In this way, the structural knowledge of different spine bones is fully learned by the reliable affinity matrix, because the features are directly synthesized with the structure-affinity key representation $\boldsymbol{k}^{'}$. Finally, we adapt the concatenate operation between the feature map and the original input and pass it through a convolutional mapping ρ to obtain the final output representation $\boldsymbol{f}^{''}$. The propagation process makes full use of the similarity of the spine bone with high affinity and dampens the wrongly activated regions in ultrasound images.

C. Global Channel Attention module

Since each channel of a high-level feature can be regarded as a specific-class response, and some relatively important channels usually have similar spatial response, we build a GCA module to capture the rich global context of each channel and enhance the representation capability of some important channel maps.

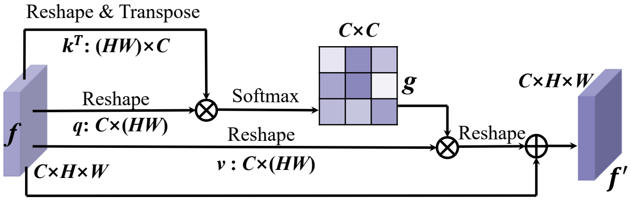

Fig. 4. The details of Global Channel Attention module

We illustrate the global channel attention module in Fig. 4. For input feature $\boldsymbol{f} \in \mathbb{R}^{C\times H\times W}$, where $C = 2048$ for the dual spatial-channel attention, we directly reshape $\boldsymbol{f}$ to $\boldsymbol{q} \in \mathbb{R}^{C\times(HW)}$ and perform a multiplication operation between $\boldsymbol{q}$ and the transpose of $\boldsymbol{f}$, $\boldsymbol{k}^{\boldsymbol{T}} \in \mathbb{R}^{(HW)\times C}$ to obtain the channel-wise similarity map. Then, we utilize a softmax layer on it to generate the channel-wise attention map $\boldsymbol{g} \in \mathbb{R}^{C\times C}$:

$$\boldsymbol{g} = Softmax(\boldsymbol{q} \times \boldsymbol{k}^{\boldsymbol{T}}) \tag{3}$$

In addition, we perform a matrix multiplication between the channel dependency matrix $\boldsymbol{g}$ and the reshape of $\boldsymbol{f}$, $\boldsymbol{v} \in \mathbb{R}^{C\times(HW)}$, and reshape the GCA-enhanced features to $\mathbb{R}^{C\times H\times W}$. The final output $\boldsymbol{f}^{'}$ is obtained by an element-wise summation operation with the feature map $\boldsymbol{f}$. It integrates the global context of each channel map and boosts the representation capability for some important channel maps.

D. Spatial Criss-cross Attention module

Owing to high spatial correlation of spine bone, discriminant feature representations are essential to localize bone contextual information and segment the spine effectively. To model the full-image contextual dependencies over local feature representations using lightweight computation, we introduce an SCA module to capture the similarity of any two correspondences in the horizontal and vertical directions and enhance the pixel-wise representative ability.

The architecture of spatial criss-cross attention module is illustrated in Fig. 5. For the feature maps $\boldsymbol{f} \in \mathbb{R}^{C\times H\times W}$, we place a 1×1 convolutional layer to generate two new feature maps $\boldsymbol{q}$, $\boldsymbol{k} \in \mathbb{R}^{C^{'}\times H\times W}$ respectively. $C^{'}$ is the number of channel, which is less than C for dimension reduction.

Furthermore, we perform an einsum operation, which can be defined as summing up the product of the elements of feature maps $\boldsymbol{q}$, $\boldsymbol{k}$ along the specified dimensions using a notation based on the Einstein summation convention ($\rightarrow$), to obtain the attention maps $\boldsymbol{h}^{\boldsymbol{T}} \in \mathbb{R}^{H\times W\times H}$ and $\boldsymbol{w} \in \mathbb{R}^{H\times W\times W}$ along the horizontal and vertical direction respectively. Then, we concatenate and apply a Softmax layer on them to get the spatial attention maps $\boldsymbol{s} \in \mathbb{R}^{H\times W\times(H+W-1)}$.

$$\begin{aligned}\boldsymbol{s} = Softmax((&einsum(``chw, ciw \rightarrow whi", \boldsymbol{q}, \boldsymbol{k}))^{\boldsymbol{T}} \\ &+einsum(``chw, chj \rightarrow hwj", \boldsymbol{q}, \boldsymbol{k}))\end{aligned} \tag{4}$$

where $chw, ciw \rightarrow whi$ and $chw, chj \rightarrow hwj$ in Eq. (4) mean the change of the spatial dimensions with the einsum operation on the query and key representations. Consequently, any two

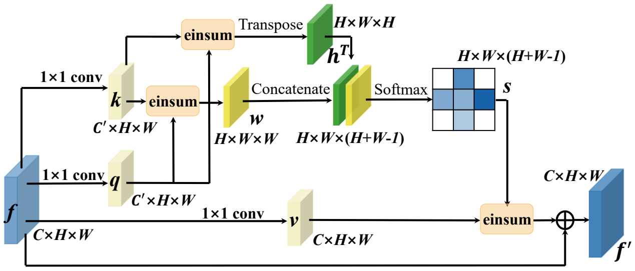

Fig. 5. The details of Spatial Criss-cross Attention module

points in the horizontal and vertical directions with the feature correlation matrix $\boldsymbol{s}$ promote each other if they are similar, and suppress each other if different. Meanwhile, another new feature map $\boldsymbol{v} \in \mathbb{R}^{C\times H\times W}$ is obtained by applying a convolutional layer with 1×1 filter, which is used to perform a einsum operation with spatial attention map $\boldsymbol{s}$ and generate the attention enhanced features. Then an element-wise operation is performed between the attention enhanced features and input features $\boldsymbol{f}$ to construct the output $\boldsymbol{f}' \in \mathbb{R}^{C\times H\times W}$ of the SCA module.

Despite the fact that the SCA module can capture contextual information in the horizontal and vertical directions, the connections between one pixel and its surrounding ones that are not in the criss-cross path are still absent. We stack two consecutive SCA modules to gain a global contextual view from all positions. This architecture makes up for the deficiency of criss-cross attention that cannot obtain the dense contextual information from all pixels, and achieve a more accurate segmentation performance for the spine bone with the cost of a minor computation complexity.

E. Loss Function

The target of spine segmentation is to classify different bone areas in the ultrasound image. In order to enhance the classification ability for each pixel, we choose the Cross Entropy (CE) loss $\mathcal{L}_{CE}$ to calculate the classification error of each pixel. Given a training pair $(\boldsymbol{x}, \boldsymbol{y})$, where $\boldsymbol{x}$ and $\boldsymbol{y}$ are the input image and its ground-truth segments respectively, and the predicted segmentation mask $\hat{\boldsymbol{y}}$, the CE loss is defined as:

$$\mathcal{L}_{CE} = -\frac{\sum_{i=1}^{H}\sum_{j=1}^{W}\sum_{c=1}^{N} y_{i,j}\log\hat{y}_{i,j}}{N\times H\times W} \tag{5}$$

where $H\times W$ is the total number of pixels in the original spine image, $\hat{y}_{i,j}$ and $y_{i,j}$ are the predicted output and the ground truth to the position in $(i,\ j)$, N is the number of classes, including three spine bone features and the background.

However, different bone features show high spatial correlation, and only appears in some regions in the ultrasound image. To effectively encode the structural knowledge, we also introduce Dice coefficient loss $\mathcal{L}_{Dice}$ to ensure the slight region-based segmentation:

$$\mathcal{L}_{Dice} = 1 - 2\frac{\sum_{i=1}^{H}\sum_{j=1}^{W}(\hat{y}\times y)}{\sum_{i=1}^{H}\sum_{j=1}^{W}(\hat{y}^2+y^2)} \tag{6}$$

The overall objective function is formulated as follows:

$$\mathcal{L} = \mathcal{L}_{CE}(\hat{y}, y) + \lambda\mathcal{L}_{Dice}(\hat{y}, y) \tag{7}$$

where λ is a hyperparameter to balance the smoothness constraint, which is empirically set as λ = 3.0.

IV. Experiments

A. Datasets and Evaluation Metrics

In our experiments, the dataset is collected from 3D ultrasound scanning in the whole spine region using the Scolioscan system (Model SCN801, Telefield Medical Imaging Ltd, Hong Kong). We utilize volume projection imaging (VPI) technique to generate 109 ultrasound VPI images from 109 patients (82 females and 27 males) with different degrees of scoliosis. The bone features are labelled by medical experts to serve as the ground-truth segments. We randomly divide the dataset into a training set and a testing set with 80 and 29 samples respectively. All images are resized to 512 $\times$ 2048 pixels. In the training process, patches of size 256 $\times$ 512 are extracted from the resized training set. In the testing process, the resized images are input to the segmentation model to produce the segmentation mask, which keep the original resolution for assessment.

To evaluate the performance of our proposed network, we employ the widely used metrics of Dice score (Dice),

TABLE I
Quantitative segmentation results in terms of Dice score (Dice)(%), Intersection over Union (IoU)(%), and Pixel Accuracy(%) based on Different Bone Regions

Methods	Rib			Thoracic			Lump			Ave.		
	Dice	IoU	Acc	Dice	IoU	Acc	Dice	IoU	Acc	Dice	IoU	Acc
UNet [9]	77.37	63.46	76.37	74.70	59.94	72.39	82.21	70.26	83.28	78.09	64.56	77.35
FPN [34]	77.19	62.85	72.51	76.63	62.11	74.37	86.52	76.25	87.85	84.17	73.54	82.89
RSNU [20]	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 [35]	77.79	65.83	79.72	76.36	**64.24**	72.34	84.40	76.52	87.91	79.52	69.68	79.99
PSANet [36]	78.42	64.50	77.79	77.14	62.78	75.09	86.68	76.49	88.24	84.64	74.17	84.38
CCNet [29]	77.62	63.42	75.26	76.89	62.46	75.26	86.49	76.19	87.97	84.32	73.73	83.89
DANet [25]	78.56	64.70	77.77	77.29	62.98	77.19	85.49	74.65	88.69	84.39	73.76	84.89
SADANet (Ours)	**78.82**	65.04	79.64	**77.64**	63.45	**79.70**	**86.96**	**76.93**	**90.66**	**84.90**	**74.54**	**86.15**

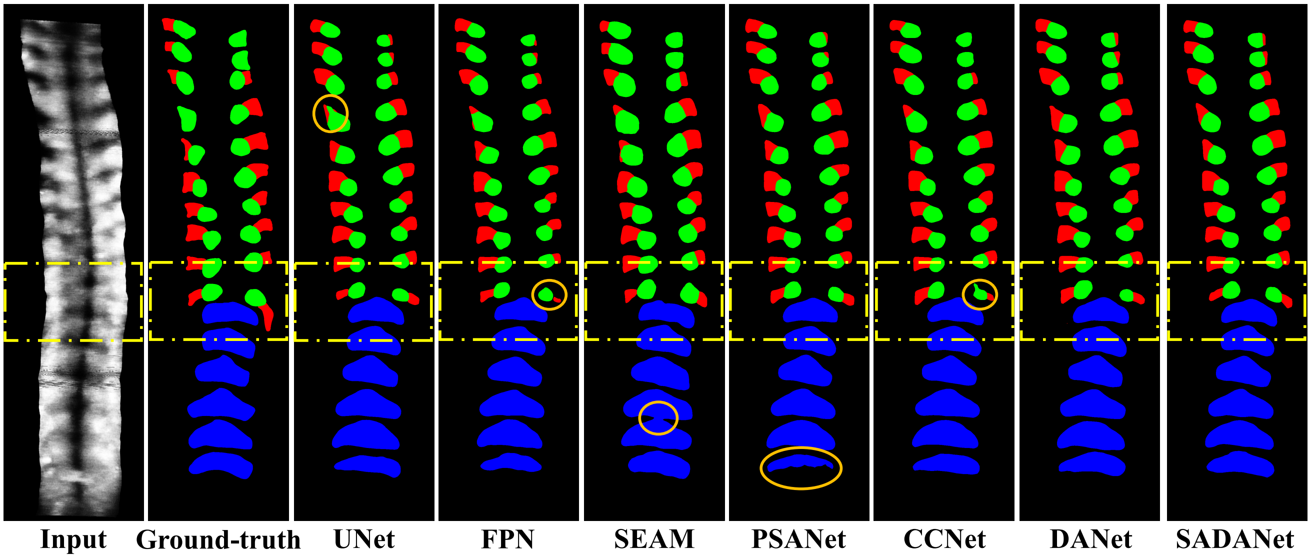

Fig. 6. A visualization of the spine bone segmentation results based on different segmentation methods. The segmented rib, thoracic process, and lump are annotated in red, green and blue. The areas around the boundary of the thoracic and lumbar region are highlighted in the yellow boxes, and the orange circles mark the defect parts of the predictions.

Intersection over Union (IoU) and Pixel Accuracy (Acc), which are formulated as follows:

$$\text{Dice} = \frac{2TP}{2TP + FP + FN} \tag{8}$$

$$\text{IoU} = \frac{TP}{TP + FP + FN} \tag{9}$$

$$\text{Acc} = \frac{TP + TN}{TP + TN + FP + FN} \tag{10}$$

where $TP, TN, FP,$ and FN refer to true positive, true negative, false positive, and false negative points, respectively.

B. Implementation Details

We implement our proposed framework based on PyTorch and MMSegmentation. During the training, we employ data augmentation including vertical flip, horizontal flip and random rotation. We build a mini-batch with 4 training samples. In the SAA module, the number of channels C is set to 1024, while in the dual spatial-channel attention block, it is set to 2048. The network is trained by the Adam optimizer for 1.6×10^5 iterations with the learning rate initialized to 10^{-3} and gradually decreased to 5×10^{-6}, based on the cosine annealing strategy [33]. The weight decay is set to 5×10^{-4} for regularization. We train the network on a single NVIDIA GeForce RTX4090 GPU.

C. Experimental Results

We test the effectiveness of our proposed SADANet for spine segmentation by comparing it with other state-of-the-art segmentation methods under the same setting and experimental environment for training and testing. They include the benchmark methods of UNet [9], FPN [34] for medical image segmentation; the recently proposed methods of RSNU [20], SEAM [35] especially for ultrasound VPI image, and the state-of-the-art attention-based methods of PSANet [36], CCNet [29], and DANet [25]. The quantitative segmentation results are reported in Table I. It is clear that the proposed SADANet surpasses all the benchmark methods [9], [34] by a large margin on all the evaluation metrics. This shows the effectiveness of the proposed network architecture for spine bone segmentation. Comparing with the methods designed for ultrasound images [20], [35], we can observe a significant improvement of over 4% on the average metrics. However, it can not be ignored that SADANet does not obtain a better evaluation metric of IoU in some specific bone features, i.e. rib and thoracic. We consider the reason to be that the strong noise in the ultrasound VPI images limits the representative ability of attention-based modules to capture the discriminative features for spine segmentation. More importantly, our proposed structure-affinity dual attention-based method outperforms other attention-based algorithms on nearly all the evaluation metrics, especially surpasses a lot on the pixel accuracy and achieves about 79.64%, 79.70%, and 90.66% for Rib, Thoracic, and Lump respectively. Thus, SADANet is desirable for VPI image enhancement and spine segmentation in clinical applications.

To further demonstrate the advantages of the proposed method, we visualize one sample from the testing set with different spine segmentation algorithms. The results are shown in Fig. 6. It can be observed that the benchmark methods, UNet [9] and FPN [34] produce unsatisfactory results in the connection area between the rib and thoracic process. The attention-based method PSANet [36] and SEAM [35], especially for ultrasound VPI image segmentation, predict a false mask and have a bad performance on the segmentation of the lumbar vertebra. Moreover, without considering the dual spatial-channel attention mechanism, CCNet [29] tends to obtain incorrect segmentation result at the area around

TABLE II
ABLATION PERFORMANCES OF SINGLE ATTENTION MODULE ON THE NETWORK ARCHITECTURE IN TERMS OF DIFFERENT BONE REGIONS

Modules	Rib			Thoracic			Lump			Ave.		
	Dice	IoU	Acc	Dice	IoU	Acc	Dice	IoU	Acc	Dice	IoU	Acc
DANet [25]	78.56	64.70	77.77	77.29	62.98	77.19	85.49	74.65	88.69	84.39	73.76	84.89
$\sim$ w/o SCA	78.36	64.42	78.66	77.52	63.29	76.76	86.51	76.23	89.84	84.46	73.90	85.32
$\sim$ w/o SAA	78.06	64.02	77.70	77.06	62.68	77.98	86.80	76.68	88.29	84.72	74.29	85.00
SADANet (Ours)	**78.82**	**65.04**	**79.64**	**77.64**	**63.45**	**79.70**	**86.96**	**76.93**	**90.66**	**84.90**	**74.54**	**86.15**

TABLE III
ABLATION PERFORMANCES OF SPATIAL CRISS-CROSS ATTENTION MODULE IN COMPUTATIONAL COST. **BOLD** INDICATES THE LOWEST ONE.

Methods	Flops (G)	Params (M)	Ave.		
			Dice	IoU	Acc
DANet [25]	200.67	51.03	84.39	73.76	84.89
$\sim$ w/o SAA	**199.35**	**49.81**	84.72	74.29	85.00
SADANet (Ours)	258.69	65.19	84.90	74.54	86.15

the boundary of the thoracic and lumbar region. Compared with these methods, the segmentation mask by SADANet are more similar to ground-truth segments, owing to the effect of structure-affinity realized by our proposed method. SADANet can locate and recognize spine bones more accurately and preserve the shape of each bone.

D. Ablation Studies

1) Effect of single attention module on the whole network: The proposed SADANet is based on structure-affinity dual spatial-channel attention network. To validate the effect of single attention module, we test the effectiveness of different network combinations under the conditions of without Spatial Criss-cross Attention (SCA) module or Structure-Affinity Attention (SAA) module, denoted as "$\sim$ w/o SCA" and "$\sim$ w/o SAA", respectively. It is worth noting that in terms of "$\sim$ w/o SCA", we employ the state-of-the-art dual attention method DANet [25] for segmentation and introduce SAA module to enrich the learned bone features. On the other hand, we adopt SCA module to replace the original position attention module in DANet (i.e., $\sim$ w/o SAA). The ablation experiment results are shown in Table II. It can be seen from the second row that structure-affinity attention module considers the structural information and obtains significant segmentation results at the area of thoracic process and lumbar, with a great increase of over 0.3% in terms of Dice Score and Intersection over Union. However, the evaluation metrics in the rib region are not satisfactory as compared with DANet, considering that the rib bone features are located in the boundary of the image, where the occupied area is small, restricting the representative ability of the SAA module. Meanwhile, the proposed SCA module also contributes a lot on the average evaluation metrics and improve more than 2% on the metric of IoU in the lumbar vertebra, compared with DANet.

2) Reduction of Complexity with Spatial Criss-cross Attention Module: SCA module has sparse connections $(H+W-1)$ for each position in the feature maps. By stacking two consecutive SCA modules, this can model full-image contextual dependencies using lightweight computation. To verify the benefits from the dual spatial-channel and spatial criss-cross attention mechanism, we adopt number of network parameters (Params) and floating point operations per second (Flops) to measure the computational cost of different network architectures. As shown in Table III, after introducing the SCA module to replace the original position attention module in DANet (i.e., $\sim$ w/o SAA), the computational and memory complexity is slightly lower than the baseline method, DANet. Meanwhile, the refined network achieves certain improvement in terms of the effectiveness of spine segmentation.

TABLE IV
ABLATION PERFORMANCES OF HYPERPARAMETER SETTINGS

Hyperparameter	Rib			Thoracic			Lump			Ave.		
	Dice	IoU	Acc	Dice	IoU	Acc	Dice	IoU	Acc	Dice	IoU	Acc
λ = 0	78.67	64.84	78.08	77.41	63.14	76.31	86.82	76.72	89.00	84.85	74.48	84.92
λ = 1.0	78.39	64.46	77.33	77.33	63.04	76.90	87.11	77.16	88.75	84.80	74.42	84.82
λ = 2.0	78.36	64.43	78.23	77.61	63.41	76.84	**87.20**	**77.31**	89.23	84.89	**74.55**	85.13
λ = 2.5	**78.82**	**65.05**	78.32	77.22	62.89	75.29	86.71	76.54	87.37	84.78	74.37	84.38
λ = 3.0	**78.82**	65.04	**79.64**	**77.64**	**63.45**	**79.70**	86.96	76.93	**90.66**	**84.90**	74.54	**86.15**
λ = 4.0	78.52	64.64	77.91	77.28	62.97	76.37	86.81	76.69	88.04	84.74	74.32	84.67
λ = 5.0	78.33	64.39	77.87	76.97	62.56	75.66	86.75	76.60	88.70	84.59	74.10	84.62

3) Determination of Weight Balance Parameter: SADANet adopts an appropriate loss function to optimize the process of spine segmentation, which is formulated as Eq. (7). λ is a weight parameter to balance two loss function terms. As tabulated in Table IV, we change the hyperparameter value to observe the effect of different weight parameters on the final quantitative spine segmentation results. It can be observed that when the value of λ increases from zero to five, the evaluation metrics achieve the best in the range of two to three, owing to the increasing weight of dice coefficient loss, which can effectively encode the structural knowledge of different bone features. However, when the weight is more than three, the performance becomes worse considering that it restraints the classification ability of the cross entropy loss function for each pixel. Thus, under the premise of considering the performance of the model, we set λ = 3.0 to carry out the subsequent network training.

V. CONCLUSION

In this paper, we have presented a structure-affinity dual spatial-channel attention network for effective spine segmentation, which adopts global channel attention module and spatial criss-cross attention module in a parallel manner to capture global dependencies in the channel and spatial dimensions respectively. Specifically, in order to enhance the structural information of spine bone into the semantic representation, we propose the structure-affinity attention module, integrating it as an auxiliary module with a segmentation network. The ablation studies and comparisons demonstrate that our method SADANet significantly improves the accuracy of the model, showing promising performance in terms of the balance between parameters, computational complexity and segmentation results, which makes it a potential solution to automatic scoliosis diagnosis in the future.

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.

References

[1] William P. Bunnell, "The natural history of idiopathic scoliosis before skeletal maturity," *Spine*, vol. 11, no. 8, pp. 773-776, 1986.

[2] J. R. Cobb, "Outline for the study of scoliosis," *Instructional course lecture*, 1948.

[3] O. M. Uddin, R. Haque, P. A. Sugrue, *et al.*, "Cost minimization in treatment of adult degenerative scoliosis," *Journal of Neurosurgery: Spine*, vol. 23, no. 6, pp. 798-806, 2015.

[4] I. Hacihaliloglu, "Ultrasound imaging and segmentation of bone surfaces: A review," *Technology*, vol. 5, no. 2, pp. 74-80, 2017.

[5] S. Y. Ng, J. Bettany-Saltikov, "Suppl-9, M5: Imaging in the diagnosis and monitoring of children with idiopathic scoliosis," *The open orthopaedics journal*,vol. 11, p. 1500, 2017.

[6] T. Ungi, F. King, M. Kempston, *et al.*, "Spinal curvature measurement by tracked ultrasound snapshots," *Ultrasound in medicine & biology*, vol. 40, no. 2, pp. 447-454, 2014.

[7] James. C. W. Cheung, G. Q. Zhou, S. Y. Law, *et al.*, "Ultrasound volume projection imaging for assessment of scoliosis," *IEEE transactions on medical imaging*, vol. 34, no. 8, pp. 1760-1768, 2015.

[8] H. Steiner, A. Staudach, *et al.*, "Diagnostic techniques: Three-dimensional ultrasound in obstetrics and gynaecology: technique, possibilities and limitations," *Human Reproduction*, pp. 1773-1778, 1994.

[9] Olaf Ronneberger, Philipp Fischer, *et al.*, "U-net: Convolutional networks for biomedical image segmentation," in *Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, 2015, pp. 234-241.

[10] A. H. Shahin, K. Amer, Mustafa A. Elattar, "Deep convolutional encoder-decoders with aggregated multi-resolution skip connections for skin lesion segmentation," in *IEEE International Symposium on Biomedical Imaging (ISBI)*, 2019, pp. 451-454.

[11] H. Xu, S. Geng, *et al.*, "Combining cgan and mil for hotspot segmentation in bone scintigraphy," *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2020, pp. 1404-1408.

[12] C. Chen, Q. Dou, H. Chen, *et al.*, "Semantic-aware generative adversarial nets for unsupervised domain adaptation in chest x-ray segmentation," in *International Workshop on Machine Learning in Medical Imaging*, Springer, 2018, pp. 143-151.

[13] W. Luo, Y. Li, R. Urtasun, *et al.*, "Understanding the effective receptive field in deep convolutional neural networks," *Advances in neural information processing systems*, vol. 29, 2016.

[14] X. Wang, R. Girshick, A. Gupta, *et al.*, "Non-local neural networks," *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018, pp. 7794-7803.

[15] F. Berton, F. Cheriet, M. C. Miron, *et al.*, "Segmentation of the spinous process and its acoustic shadow in vertebral ultrasound images," *Computers in biology and medicine*, vol. 72, pp. 201-211, 2016.

[16] W. Chen, *et al.*, "Ultrasound imaging of spinal vertebrae to study scoliosis," *Open Journal of Acoustics*, vol. 2, no. 3, pp. 95-103, 2012.

[17] Y. Wong, K. K. Lai, Y. Zheng, *et al.*, "Is radiation-free ultrasound accurate for quantitative assessment of spinal deformity in idiopathic scoliosis (IS): a detailed analysis with EOS radiography on 952 patients," *Ultrasound in medicine & biology*, vol. 45, no. 11, pp. 2866-2877, 2019.

[18] R. C. Brink, S. P. Wijdicks, I. N. Tromp, *et al.*, "A reliability and validity study for different coronal angles using ultrasound imaging in adolescent idiopathic scoliosis," *The Spine Journal*, vol.18, no.6, pp. 979-985, 2018.

[19] T. Ungi, H. Greer, et al. *et al.*, "Automatic spine ultrasound segmentation for scoliosis visualization and measurement," *IEEE Transactions on Biomedical Engineering*, vol. 67, no. 11, pp. 3234-3241, 2020.

[20] Z. Huang, L. W. Wang, Frank. H. F. Leung, *et al.*, "Bone feature segmentation in ultrasound spine image with robustness to speckle and regular occlusion noise," in *IEEE International Conference on Systems, Man, and Cybernetics (SMC)*, 2020, pp. 1566-1571.

[21] A. Vaswani, N. Shazeer, N. Parmar, *et al.*, "Attention is all you need," *Advances in neural information processing systems*, vol. 30, 2017.

[22] F. Ding, G. Yang, J. Liu, *et al.*, "Hierarchical attention networks for medical image segmentation," *arXiv preprint arXiv:1911.08777*, 2019.

[23] X. Li, Z. Zhong, *et al.*, "Expectation-maximization attention networks for semantic segmentation," in *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2019, pp. 9167-9176.

[24] J. Long, E. Shelhamer, T. Darrell, "Fully convolutional networks for semantic segmentation," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015, pp. 3431-3440.

[25] J. Fu, J. Liu, H. Tian, *et al.*, "Dual attention network for scene segmentation," *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019, pp. 3146-3154.

[26] L. Mou, Y. Zhao, H. Fu, *et al.*, "CS^2-Net: Deep learning segmentation of curvilinear structures in medical imaging," *Medical image analysis*, vol.67, p. 101874, 2021.

[27] X. Liu, G. Xiao, L. Dai, *et al.*, "SCSA-Net: Presentation of two-view reliable correspondence learning via spatial-channel self-attention," *Neurocomputing*, vol. 431, pp. 137-147, 2021.

[28] S. Banerjee, J. Lyu, Z. Huang, *et al.*, "Light-convolution Dense selection U-net (LDS U-net) for ultrasound lateral bony feature segmentation," *Applied Sciences*, vol. 11, no. 21, pp. 10180, 2021.

[29] Z. Huang, X. Wang, L. Huang, *et al.*, "Ccnet: Criss-cross attention for semantic segmentation," *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2019, pp. 603-612.

[30] K. He, X. Zhang, S. Ren, *et al.*, "Deep residual learning for image recognition," *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016, pp. 770-778.

[31] P. Veličković, G. Cucurull, *et al.*, "Graph Attention Networks," in *International Conference on Learning Representations (ICLR)*, 2018.

[32] L. Ru, Y. Zhan, B. Yu, *et al.*, "Learning affinity from attention: End-to-end weakly-supervised semantic segmentation with transformers," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022, pp. 16846-16855.

[33] Ilya Loshchilov, Frank Hutter, "SGDR: Stochastic gradient descent with warm restarts," in *International Conference on Learning Representations (ICLR)*, 2016, pp. 1-16.

[34] T. Y. Lin, P. Dollár, R. Girshick, *et al.*, "Feature pyramid networks for object detection," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017, pp. 2117-2125.

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

[36] H. Zhao, Y. Zhang, S. Liu, *et al.*, "Psanet: Point-wise spatial attention network for scene parsing," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018, pp. 267-283.

[37] J. Xiao, X. Jiang, *et al.*, "Online Video Super-Resolution with Convolutional Kernel Bypass Grafts," *IEEE Transactions on Multimedia*, 2023.

[38] Y. Ju, M. Jian, C. Wang, *et al.*, "Estimating High-resolution Surface Normals via Low-resolution Photometric Stereo Images," *IEEE Transactions on Circuits and Systems for Video Technology*, 2023.

[39] C. Zhang, J. Su, *et al.*, "Efficient Inductive Vision Transformer for Oriented Object Detection in Remote Sensing Imagery," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 61, pp. 1-20, 2023.

[40] J. Xiao, T. Liu, R. Zhao, *et al.*, "Balanced distortion and perception in single-image super-resolution based on optimal transport in wavelet domain," *Neurocomputing*, pp. 408-420, 2021.

[41] Y. Ju, K. M. Lam, J. Xiao, *et al.*, "Efficient Feature Fusion for Learning-Based Photometric Stereo," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2023, pp. 1-5.

[42] Z. Huang, R. Zhao, Frank. H. F. Leung, *et al.*, "Joint spine segmentation and noise removal from ultrasound volume projection images with selective feature sharing," *IEEE Transactions on Medical Imaging*, 2022, pp. 1610-1624.

[43] J. Xiao, W. Jia, K. M. Lam, "Feature redundancy mining: Deep lightweight image super-resolution model," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2021, pp. 1620-1624.

[44] C. Zhang, T. Liu, J. Xiao, *et al.*, "Boosting Object Detectors via Strong-Classification Weak-Localization Pretraining in Remote Sensing Imagery," *IEEE Transactions on Instrumentation and Measurement*, vol. 72, pp. 1-20, 2023.

[45] Y. Ju, J. Dong, S. Chen, "Recovering surface normal and arbitrary images: A dual regression network for photometric stereo," *IEEE Transactions on Image Processing*, 2021, pp. 3676-3690.