Seg4Reg+: Consistency Learning between Spine Segmentation and Cobb Angle Regression

Yi Lin[1✉], Luyan Liu[1], Kai Ma[1], and Yefeng Zheng[1]

Tencent Jarvis Lab, Shenzhen, China
`linyi.pk@gmail.com`

Abstract. Automated methods for Cobb angle estimation are of high demand for scoliosis assessment. Existing methods typically calculate the Cobb angle from landmark estimation, or simply combine the low-level task (e.g., landmark detection and spine segmentation) with the Cobb angle regression task, without fully exploring the benefits from each other. In this study, we propose a novel multi-task framework, named Seg4Reg+, which jointly optimizes the segmentation and regression networks. We thoroughly investigate both local and global consistency and knowledge transfer between each other. Specifically, we propose an attention regularization module leveraging class activation maps (CAMs) from image-segmentation pairs to discover additional supervision in the regression network, and the CAMs can serve as a region-of-interest enhancement gate to facilitate the segmentation task in turn. Meanwhile, we design a novel triangle consistency learning to train the two networks jointly for global optimization. The evaluations performed on the public AASCE Challenge dataset demonstrate the effectiveness of each module and superior performance of our model to the state-of-the-art methods.

1 Introduction

Adolescent idiopathic scoliosis (AIS) causes a structural, lateral, rotated curvature of the spine that arises in children at or around puberty [12]. The Cobb angle, derived from a anterior-posterior X-ray and measured by selecting the most tilted vertebra, is the primary means for clinical diagnosing of AIS. However, manual measurement of the Cobb angle requires professional radiologists to carefully identify each vertebra and measure angle, which is time-consuming and could suffer from a large inter-/intra-observer variety. Hence, it is needed to provide an accurate and robust method for quantitative measurement of Cobb angle automatically.

Numerous computer-aided methods have been developed for automated Cobb angle estimation. Conventional methods utilized active contour model [2], customized filter [1] and charged-particle models [8] for spine segmentation to calculate the Cobb angle, which are computationally expensive and the unclear spine boundary will result in inaccurate estimations. Recently, deep learning based methods [3,6,9,13,15,17] have been proposed to consolidate the tasks of vertebral landmark detection with Cobb angle estimation to improve the robustness
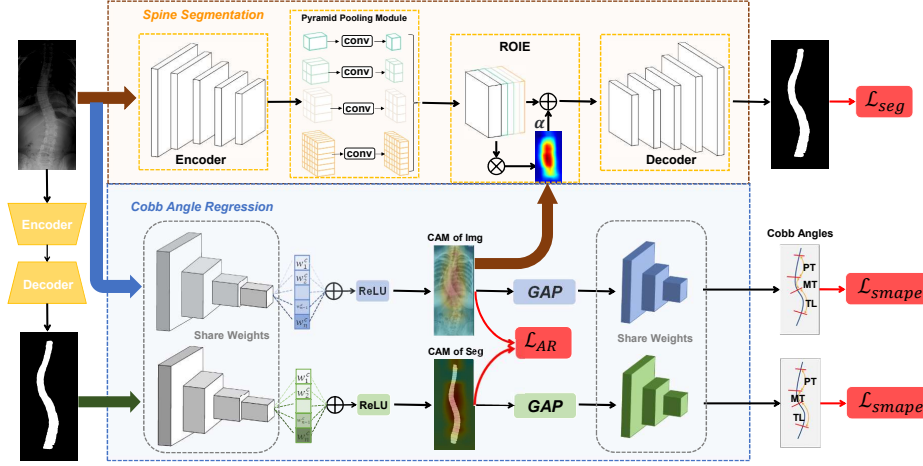
Fig. 1. An overview of the proposed framework.

of spinal curvature assessment. And Seg4Reg [7], which won the 1st place in the AASCE[1] challenge, regarded the segmentation results of spine as the input of the regression network for Cobb angle estimation. Its superior performance owns to the segmentation mask of spine, which retains the shape information and filters the distractions (e.g., artifacts and local contrast variation). However, the performance of this method depends heavily on the segmentation results.

Although the above methods have achieved great success, their applications to Cobb angle estimation suffer from three limitations: 1) The methods [3,6,9,13,15,17] relying on the landmark coordinates are susceptible since a small error in landmark coordinates may cause a huge mistake in angle estimation; 2) the two-stage frameworks [3,6,7,17] often suffers from the error accumulation; and 3) the cascaded networks[3,6,7] cannot guarantee a global optimum.

In this paper, we propose a novel consistency learning framework, named Seg4Reg+, which incorporates segmentation into the regression task, as shown in Fig. 1. The segmentation task extracts representative features for the regression task by an **attention regularization** (AR) module with auxiliary constraint on the class activation map (CAM). And the regression task is able to provide specific hints for the segmentation task by a **region-of-interest enhancement** (ROIE) gate to force the segmentation network to pay more attention to the important area. To reach the global optimum, we further design a novel **triangle consistency learning** scheme for end-to-end training. In summary, our main contributions are as follows:

---

[1] https://aasce19.grand-challenge.org

- We propose a novel consistency learning framework, named Seg4Reg+, incorporating segmentation and regression tasks with an AR module and ROIE gate to boost the performance of both tasks.
- We design a triangle consistency learning scheme for end-to-end training.
- Extensive evaluations on the public AASCE Challenge dataset demonstrate the effectiveness of each module and superior performance of our model to the state-of-the-art methods.

2 Method

As illustrated in Figure 1, the proposed Seg4Reg+ model consists of a segmentation network N_S and a regression network N_R. We first pre-train the two networks separately for approximately optimized results to speed up the training process. Then the two networks are boosted by each other by the ROIE gate and AR module. In addition, a novel training strategy named triangle consistency learning scheme is designed for end-to-end training. In the following, we first introduce the proposed ROIE gate and AR module, then illustrate the triangle consistency learning scheme.

ROIE Gate. We first train N_S to roughly segment the spine region. We adopt the same network as Seg4Reg [7] for a fair comparison. Specifically, we modify the PSPNet [18] by replacing the pooling layer with the dilated convolution in the pyramid pooling module and take ResNet-50 [4] as the backbone. The objective function is the weighted Dice loss and cross-entropy loss:

$$\mathcal{L}_{\text{seg}}(s(x), y) = \sum_i \left(1 - \frac{2 \times s(x_i)y_i}{s(x_i) + y_i}\right) + \lambda \sum_i (-y_i \log(s(x_i))), \qquad (1)$$

where x denotes the input image, and $s(x_i)$ and y_i denote the label of pixel x_i of prediction and ground truth, respectively, and λ denotes the hyperparameter weighting the two losses.

To boost the performance of N_S with N_R, the ROIE gate is designed as an attention mechanism to transfer the specific hints from N_R to N_S. The proposed ROIE gate is inspired by CAM, which is the most common technique in weakly supervised segmentation methods [19]. We expect that the CAM of N_R can incorporate the refined prior information about the spine area into the segmentation process. In addition, the value of each pixel on the CAM represents its significance to the regression output, which in turn guides N_S to pay more attention to the important areas (i.e., the most tilted vertebra endplates). Specifically, we treat CAM of N_R as attention map, and perform a matrix multiplication between CAM and the feature map $f_m(x)$ from the middle layer of N_S. Then, we multiply the result by a scalar parameter α and perform an element-wise sum operation with the feature $f_m(x)$ to obtain the final output $f'_m(x)$ as follows:

$$f'_m(x) = \alpha(C(r(x)) \circ f_m(x)) + f_m(x), \qquad (2)$$

where $C(\cdot)$ is a CAM that indicates the discriminative part with respect to the regression results, α is a learnable parameter which is initialized as 0, $\circ$ denotes
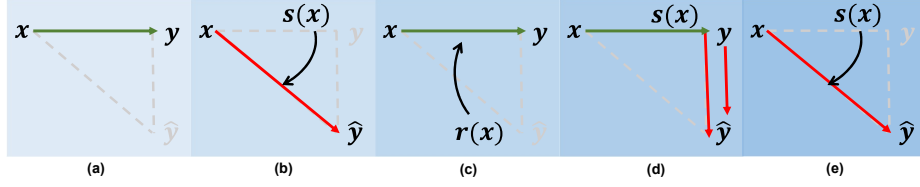
Fig. 2. The training strategy of triangle consistency learning: (a) training N_S, (b) training N_R with AR, (c) fine-tuning N_S with ROIE, (d) fine-tuning N_S by the SMAPE loss, and (e) fine-tuning N_R with refined segmentation.

multiplication function. It can be inferred from Equation (2) that the resulting feature $f'_m(x)$ combines global contextual view and selectively aggregates contexts according to the CAM, thus improving intra-class compact and semantic consistency.

AR Module. For N_r, we modify the state-of-the-art classification network by replacing the last convolutional layer with the output channel corresponding to three clinically relevant Cobb angles: proximal thoracic (PT), main thoracic (MT) and thoracolumbar/lumbar (TL). And the activation function in the last layer is set to the sigmoid function. Here, we design a novel objective function based on symmetric mean absolute percentage error, named SMAPE loss:

$$\mathcal{L}_{\text{SMAPE}}(r(x), \hat{y}) = \frac{\sum_{i=1}^{n} |\hat{y}_i - r(x_i)|}{\sum_{i=1}^{n} |\hat{y}_i + r(x_i) + \epsilon|}, \tag{3}$$

where $\hat{y}_i$ and $r(x_i)$ denote the ground truth and prediction of ith angle of the total $n = 3$ Cobb angles (i.e., PT, MT and TL), and ϵ is a smooth factor.

To boost the performance of N_R with N_S, the AR module is designed to explore the hidden state representation of the N_R via classification activation mapping (CAM) to force it to focus on the spine area. Specifically, to integrate regularization on N_R, we expand the N_R into a shared-weight Siamese structure. One branch takes the concatenation of the raw image and its corresponding segmentation mask as input, and the other directly takes the segmentation as input. The output activation maps from two branches are regularized by mean absolute error to guarantee the consistency of CAMs, and the regression network in consequence is forced to focus on the spine area. The objective function is:

$$\mathcal{L}_{\text{AR}} = \|C(x, s(x)) - C(s(x))\|_1, \tag{4}$$

Triangle Consistency Learning. Inspired by the inference-path invariance theory [16] which declares that inference paths with the same endpoints, but different intermediate domains, yield similar results. The segmentation process is essentially an auxiliary task for the regression task, thus the regression network has the potential to optimize the segmentation network. Based on this assumption, we design a novel training strategy named as triangle consistency learning for end-to-end training, are shown in Figure 2. The details of the proposed training strategy are shown in Algorithm 1, which can be divided into five processes:

Algorithm 1 Triangle Consistency Learning.

Require:

    Input image, $x \in X$;

    Ground truth of segmentation mask of spine, $y \in Y$;

    Ground truth of three Cobb angles (PT, MT and TL), $\hat{y} \in \hat{Y}$;

Ensure:

    $s(\cdot)$: Segmentation task of network N_S with parameter θ_1;

    $r(\cdot)$: Regression task of network N_R with parameter θ_2;

    $C(\cdot)$: Class activation map (CAM) generated by N_R;

    \\Training basic N_S.

1: **while** stopping criterion not met **do**

2:    Compute the segmentation loss $\mathcal{L}_{\text{seg}}(s(x), y)$ with Equation (1);

3:    Update parameters θ_1 of N_S by backpropagation;

4: **end while**

    \\Training N_R with AR.

5: **while** stopping criterion not met **do**

6:    Compute the SMAPE loss $\mathcal{L}_{\text{SMAPE}}(r(x, s(x)), \hat{y})$ with Equation (3);

        \\$(x, s(x))$ means concatenation of raw image and its segmentation mask.

7:    Compute the SMAPE loss $\mathcal{L}_{\text{SMAPE}}(r(s(x)), \hat{y})$ with Equation (3);

8:    Compute the attention regularization loss $\mathcal{L}_{\text{AR}} = \|C(x, s(x)) - C(s(x))\|_1$;

9:    Update the regression network parameters θ_2;

10: **end while**

    \\Fine-tuning N_S with ROIE.

11: **while** stopping criterion not met **do**

12:    Add local consistency constraints to N_S with Equation (2);

13:    Update the parameters θ_1 of N_S;

14: **end while**

    \\Fine-tuning N_S by SMAPE loss.

15: **while** stopping criterion not met **do**

16:    Compute the regression loss $\mathcal{L}_{\text{SMAPE}}(r(s), r(s(x, C(x))))$;

17:    Freeze the θ_2 and update the θ_1;

18: **end while**

    \\Fine-tuning N_R with refined segmentation.

19: Repeat steps 5-10

20: **return** θ_1 and θ_2

---

1) training basic N_S (steps 1-4); 2) training N_R with AR (steps 5-10); 3) fine-tuning N_S with ROIE (steps 11-14); 4) fine-tuning N_S with SMAPE loss on the regression output of $s(x)$ and y (steps 15-18); and 5) fine-tuning N_R with refined segmentation (step 19). In the fourth process, we compute the SMAPE loss for the regression outputs of segmentation results and the corresponding ground truth, denoted $r(s(x))$ and $r(y)$, respectively. Then we freeze the parameters of N_R and optimize N_S by the backpropagation. In this way, we can generate more suitable segmentation mask for N_R, superior to traditional segmentation.

Table 1. The ablation study for each part of Seg4Reg+. AR: attention regularization module, ROIE: region-of-interest enhancement, TCL: triangle consistency learning, Img: raw image as input, and Seg: segmentation mask.

| Baseline	AR	ROIE	TCL	Img	Seg	MAE	SMAPE (%)
✓				✓		6.34, 7.77, 8.01	12.32
✓	✓			✓		4.55, 5.75, 5.92	9.39
✓	✓	✓		✓		4.21, 5.32, **5.22**	9.32
✓	✓	✓	✓	✓		**4.01, 5.16**, 5.51	**9.17**
✓					✓	6.51, 6.22, 7.17	10.95
✓	✓				✓	4.03, 6.38, 5.80	9.52
✓	✓	✓			✓	**3.61**, 4.90, 5.53	9.01
✓	✓	✓	✓		✓	5.13, **4.73, 5.24**	**8.92**
✓	✓	✓	✓	✓	✓	**3.88, 4.62, 4.99**	**8.47**

3   Experiments

Data and Implementation Details. We use the public dataset of MICCAI 2019 AASCE Challenge [13], which consists of 609 spinal anterior-posterior X-ray images to evaluate our method. The dataset is divided by the provider into 481 images for training and 128 images for testing. We evaluate the proposed method using two metrics, symmetric mean absolute percent error (SMAPE) and mean absolute error (MAE). And we evaluate the segmentation results using five performance metrics including the Jaccard index (JA), Dice coefficient (Dice), pixel-wise accuracy (pixel-AC), pixel-wise sensitivity (pixel-SE), and pixel-wise specificity (pixel-SP).

We pre-process the data before inputting it into our network. First, we resize the image to [512, 256]. Then, we linearly transform the Cobb angles into [0, 1], and augment the dataset by randomly flipping, rotating ($-45°$, $45°$), and rescaling with the factor between (0.85, 1.25). We train the N_S for 90 epochs using ADAM optimization with learning rate 1×10^{-4} and weight decay 1×10^{-5}. And for regression, we train the network for 200 epochs in total with learning rate 1×10^{-3} and weight decay 1×10^{-5}.

Ablation Studies. To verify the effectiveness of each module in the proposd Seg4Reg+ approach, we conduct a set of experiments for ablation study, which are shown in Table 1. We first test with the raw image as input, the AR module has a 2.93% improvement in SMAPE compared to baseline, which validates that the AR module can gain from segmentation and benefit for the regression task. By combining the ROIE gate, the performance can be further improved by 0.07% (the improvement with the segmentation mask as input is more significant with 0.51% boost in SMAPE), which demonstrates N_R can boost N_S in turn. And applying triangle consistency learning could improve the SMAPE by 0.15%. Similar trend can also be observed when we test with segmentation results as input. Finally, when we concatenate the raw image and segmentation mask together as input, the performance can by further improved up to 8.47%.
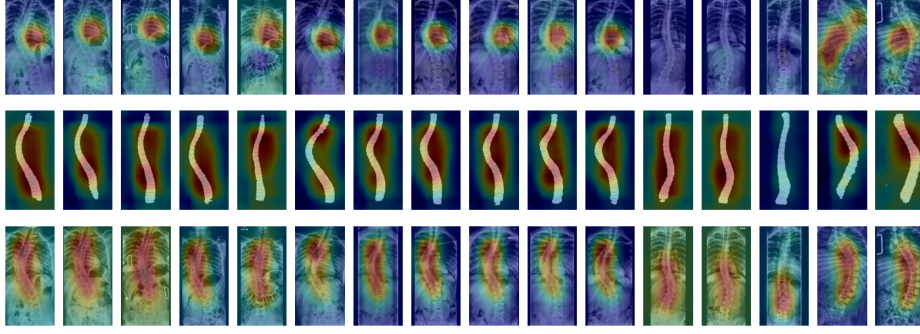
Fig. 3. The visualization of CAMs with and without the AR module. The top and middle rows show the result of baseline method (without AR) using raw image and segmentation mask as input, respectively. And the bottom row shows the CAMs generated from the baseline method with AR.
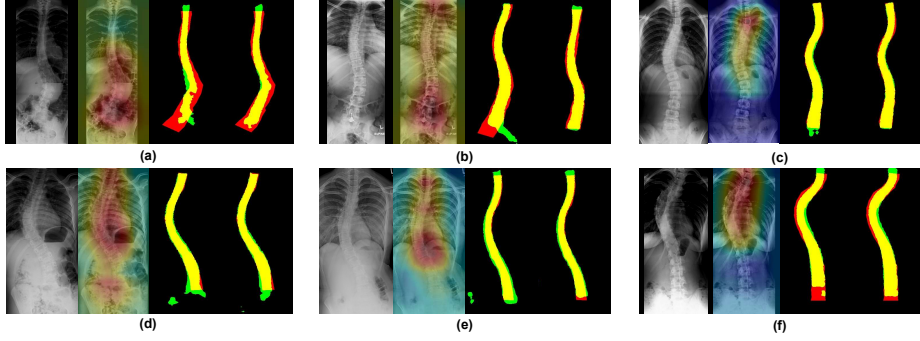
Fig. 4. Examples of segmentation results. Each group shows the original image, CAM, and segmentations without and with ROIE, respectively. The yellow mask is true positive, red mask is false negative, and green mask is false positive.

Effectiveness of AR Module. Figure 3 shows the results of the baseline method and the AR module. From the first row in Figure 3, we can see several drawbacks of the CAM generated by the baseline method: 1) focusing only on the local region (e.g., columns 1-11), 2) the relative inferior ability of feature extraction (e.g., columns 12-14), and 3) the vulnerability to the blurred images (e.g., columns 15-16). And the bottom row in Figure 3 shows that with the AR module, N_R can make more precise prediction with more reasonable perspective view. Particularly, in column 14, our AR module predicts more precise attention area while the baseline method generates very weak attention maps.

Effectiveness of ROIE Gate. We evaluate both qualitative and quantitative segmentation results of our ROIE module. Table 2 shows that applying the ROIE gate to the origin segmentation model promotes JA from 75.47% to 77.86%, and the visualization results in Figure 4 show that the ROIE gate is helpful for false

Table 2. Comparison of segmentation performance (%) of ROIE with other methods.

Methods	JA	Dice	pixel-AC	pixel-SE	pixel-SP
w/o CAMs	75.47	86.02	94.83	86.27	98.21
MDC [11]	75.91	86.31	**95.73**	**89.08**	98.12
D-Net [5]	76.99	87.00	95.02	86.61	98.46
MB-DCNN [14]	76.17	86.47	95.12	87.18	98.19
Ours	**77.86**	**87.55**	95.49	88.06	**98.42**

Table 3. Comparison of regression performance with the state-of-the-art methods.

Methods	A-Net [3]	L-Net [3]	B-Net [13]	PFA [10]	VF [6]		Seg4Reg [7]		Ours	
Backbone	-	-	-	res50	UNet	MNet	res18	eff_b1	res18	eff_b1
MAE	8.58	10.46	9.31	6.69	3.90	**3.51**	6.63	3.96	4.50	3.73
SMAPE (%)	20.35	26.94	23.44	12.97	8.79	7.84	10.95	7.64	8.47	**7.32**

positive reduction. Furthermore, we compare different ways of transferring CAM to the segmentation network, i.e., MDC [11], D-Net [5] and MB-DCNN [14]. However, simply delivering CAMs to N_S, these methods would be affected by inaccurate CAMs. And our ROIE gate alleviates this problem by fusing the CAM adaptively with a learnable parameter. For a fair comparison, all of these methods use the same segmentation architecture. It shows that our ROIE gate improves JA by 0.9% and 1.7% over D-Net and MB-DCNN, respectively.

Comparison with State-of-the-art. We compare the proposed method with state-of-the-art methods including A-Net [3], L-Net [3], BoostNet [13], PFA [10], VF [6], and Seg4Reg [7], which won the 1st place in AASCE challenge. We follow the same experiment setting with VF [6], and the performance of all competing methods is adopted from the original publications for a fair comparison.

As depicted in Table 3, promising results are observed in predicting Cobb angle using the proposed framework. The SMAPE is 7.32% with our Seg4Reg+ framework, whereas PFA, VF-MNet and Seg4Reg achieve 12.97%, 7.84% and 7.64%, respectively. The MAE of our method is 3.73, slightly inferior to VF-MNet. We believe the superior performance of our model in SMAPE is more convincing as SMAPE is the only evaluation metric in the AASCE challenge.

4   Conclusion

In this paper, we proposed a novel Seg4Reg+ model for automated Cobb angle estimation. Our Seg4Reg+ model incorporates segmentation into regression task via an attention regularization module and a region-of-interest enhancement gate to boost the performance of both tasks. Further, the two networks are integrated with global consistency learning for global optimization. Experimental results demonstrated the effectiveness of each module and the superior performance of our model to the state-of-the-art methods.

References

1. Anitha, H., Karunakar, A., Dinesh, K.: Automatic extraction of vertebral endplates from scoliotic radiographs using customized filter. Biomedical Engineering Letters **4**(2), 158–165 (2014)
2. Anitha, H., Prabhu, G.: Automatic quantification of spinal curvature in scoliotic radiograph using image processing. Journal of Medical Systems **36**(3), 1943–1951 (2012)
3. Chen, B., Xu, Q., Wang, L., Leung, S., Chung, J., Li, S.: An automated and accurate spine curve analysis system. IEEE Access **7**, 124596–124605 (2019)
4. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 770–778 (2016)
5. Hong, S., Noh, H., Han, B.: Decoupled deep neural network for semi-supervised semantic segmentation. arXiv preprint arXiv:1506.04924 (2015)
6. Kim, K.C., Yun, H.S., Kim, S., Seo, J.K.: Automation of spine curve assessment in frontal radiographs using deep learning of vertebral-tilt vector. IEEE Access **8**, 84618–84630 (2020)
7. Lin, Y., Zhou, H.Y., Ma, K., Yang, X., Zheng, Y.: Seg4Reg networks for automated spinal curvature estimation. In: International Workshop and Challenge on Computational Methods and Clinical Applications for Spine Imaging. pp. 69–74. Springer (2019)
8. Sardjono, T.A., Wilkinson, M.H., Veldhuizen, A.G., van Ooijen, P.M., Purnama, K.E., Verkerke, G.J.: Automatic Cobb angle determination from radiographic images. Spine **38**(20), E1256–E1262 (2013)
9. Sun, H., Zhen, X., Bailey, C., Rasoulinejad, P., Yin, Y., Li, S.: Direct estimation of spinal Cobb angles by structured multi-output regression. In: International Conference on Information Processing in Medical Imaging. pp. 529–540. Springer (2017)
10. Wang, J., Wang, L., Liu, C.: A multi-task learning method for direct estimation of spinal curvature. In: International Workshop and Challenge on Computational Methods and Clinical Applications for Spine Imaging. pp. 113–118. Springer (2019)
11. Wei, Y., Xiao, H., Shi, H., Jie, Z., Feng, J., Huang, T.S.: Revisiting dilated convolution: A simple approach for weakly- and semi-supervised semantic segmentation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 7268–7277 (2018)
12. Weinstein, S.L., Dolan, L.A., Cheng, J.C., Danielsson, A., Morcuende, J.A.: Adolescent idiopathic scoliosis. The Lancet **371**(9623), 1527–1537 (2008)
13. Wu, H., Bailey, C., Rasoulinejad, P., Li, S.: Automatic landmark estimation for adolescent idiopathic scoliosis assessment using BoostNet. In: International Conference on Medical Image Computing and Computer Assisted Intervention. pp. 127–135. Springer (2017)
14. Xie, Y., Zhang, J., Xia, Y., Shen, C.: A mutual bootstrapping model for automated skin lesion segmentation and classification. IEEE Transactions on Medical Imaging **39**(7), 2482–2493 (2020)

15. Yi, J., Wu, P., Huang, Q., Qu, H., Metaxas, D.N.: Vertebra-focused landmark detection for scoliosis assessment. In: IEEE 17th International Symposium on Biomedical Imaging. pp. 736–740. IEEE (2020)
16. Zamir, A.R., Sax, A., Cheerla, N., Suri, R., Cao, Z., Malik, J., Guibas, L.J.: Robust learning through cross-task consistency. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 11197–11206 (2020)
17. Zhang, C., Wang, J., He, J., Gao, P., Xie, G.: Automated vertebral landmarks and spinal curvature estimation using non-directional part affinity fields. Neurocomputing (2021)
18. Zhao, H., Shi, J., Qi, X., Wang, X., Jia, J.: Pyramid scene parsing network. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 2881–2890 (2017)
19. Zhou, B., Khosla, A., Lapedriza, A., Oliva, A., Torralba, A.: Learning deep features for discriminative localization. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 2921–2929 (2016)