2021 43rd Annual International Conference of the
IEEE Engineering in Medicine & Biology Society (EMBC)
Oct 31 - Nov 4, 2021. Virtual Conference

Heterogeneous Consistency Loss for Cobb Angle Estimation*

Yue Guo[1], Yanmei Li[2], Wenhao He[1,3], Haitao Song[1]

Abstract— **Cobb angle is the most common quantification of the spine deformity called scoliosis. Recently, automatic Cobb angle estimation has become popular with either semantic segmentation networks or landmark detectors. However, such methods can not perform robustly when some vertebrae have ambiguous appearances in X-ray images. To alleviate the above problem, we propose a multi-task model that simultaneously outputs semantic masks and keypoints of vertebrae. When training this model, we propose a heterogeneous consistency loss function to enhance the consistency between keypoints and semantic masks. Extensive experiments on anterior-posterior (AP) X-ray images from AASCE MICCAI 2019 Challenge demonstrate that our method significantly reduces Cobb angle estimation errors and achieves state-of-the-art performances.**

Clinical relevance— **This work shows that a multi-task model has some potential to measure Cobb angles in more challenging situations, and we can directly integrate it into an auxiliary clinical diagnosis system to assist doctors more effectively for subsequent treatments.**

I. INTRODUCTION

Scoliosis is a sideways curvature of the spine that mostly happens among teens, and it may cause back pain, leg numbness, tiredness, and even breathing and heart problems. The standard quantification of scoliosis is Cobb angle, which is measured between a tangent of the upper endplate from the upper vertebra and the other one of the lower endplate from the lower vertebra. Since manual Cobb angle measurement is time-consuming and largely depends on the doctor's experience, automatic methods in X-ray or Moire images have been focused on in many recent works [1], [2], [3], [4].

Accurate measurement of Cobb angles remains challenging due to the ambiguity and variability of vertebrae in X-ray images. Only relying on predictions in a single task, large offsets of vertebrae from models trained with limited data are inevitable, and their unexpected occurrences may lead to erroneous Cobb angles. Therefore, multiple tasks can be integrated into a Cobb angle estimator, and outputs from their branches should be correlated and consistent even with very different representations.

In this paper, we propose a hybrid method that provides both keypoints and semantic masks of vertebrae in X-ray images. This method utilizes the interaction between the above two branches in multi-task learning and is motivated by the intuition that predictions for an identical vertebra in different tasks should have more similar shapes, we propose a novel heterogeneous consistency loss function to supervise the model training procedure. The end-to-end trained model is evaluated on anterior-posterior (AP) X-ray images from AASCE MICCAI 2019 Challenge [5] and achieves superior performance compared to other state-of-the-art methods.

*This work is supported by National Key R&D Program of China (2018YFB1306302, 2018YFB1306300, and 2018YFB1306500).

*Corresponding author: Wenhao He.

[1]Yue Guo, Wenhao He, and Haitao Song are with Institute of Automation, Chinese Academy of Sciences, Beijing, China. guoyue2013@ia.ac.cn

[2]Yanmei Li is with Beijing College of Finance and Commerce, Beijing, China.

[3]Wenhao He is also with School of Artificial Intelligence, University of Chinese Academy of Sciences, Beijing, China.

II. METHOD

A. Keypoint Estimation

Keypoints of vertebrae are used to compute Cobb angles. The same as the keypoint estimation strategy from the prior work [6], there are t vertebrae (17 in this task) and $t \times 4$ keypoints (top-left, top-right, bottom-right, and bottom-right for each vertebra) in every X-ray image. Similar to CenterNet [7], vertebrae are firstly separated by t center points, each of which keypoints is later computed using their offsets.

Our keypoint estimator consists of both the keypoint localization branch and the semantic segmentation branch, as shown in Figure 1.

1) Feature Map: HRNet is integrated to extract convolutional features because it outperforms others such as U-Net and Hourglass in many computer vision tasks, based on its parallel multi-resolution fusions [8]. Consequently, given a $n_h \times n_w$ X-ray image, feature maps from four streams in multiple resolutions are extracted and channel-wisely concatenated to the highest resolution $\lfloor \frac{n_h}{4} \rfloor \times \lfloor \frac{n_w}{4} \rfloor$.

2) Keypoint: Branches for predicting center heatmaps, center offsets, corner offsets, and semantic masks are built, each of which contains two convolutional layers.

3) Center Heatmap: The Gaussian kernel is applied to generate heatmaps of the ground truth center points, and the element-wise maximum is computed if Gaussians of several central points overlap. Therefore, the center heatmap has only one channel to distinguish the vertebra from the background.

4) Center Offset: Center offsets are provided to reduce the localization errors resulted from feature map downsampling. Given a center point $p_k = (x_c, y_c)$ in the input image, its resolution is reduced to $(\lfloor \frac{x_c}{4} \rfloor, \lfloor \frac{y_c}{4} \rfloor)$ in the corresponding feature maps. Consequently, its center offset becomes $(\frac{x_c}{4} - \lfloor \frac{x_c}{4} \rfloor, \frac{y_c}{4} - \lfloor \frac{y_c}{4} \rfloor)$. As a result, the center offset map consists of $t \times 2$ channels, since there are t vertebrae, and two-dimensional coordinates of one center locate a vertebra in a single-channel map.

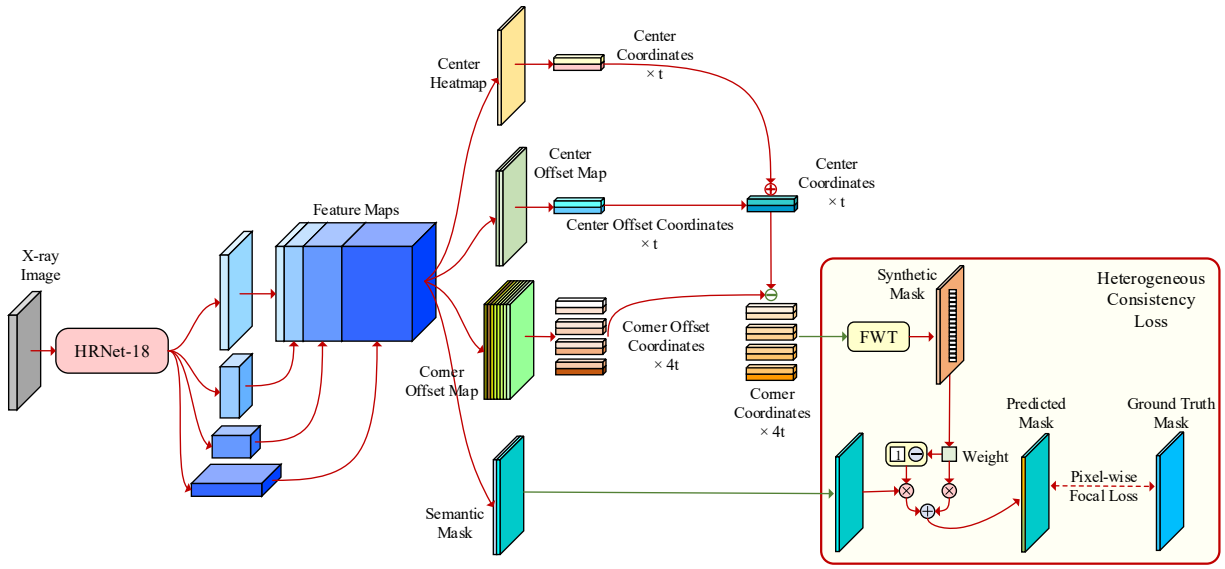


Fig. 1. The proposed keypoint estimation framework: given an input X-ray image, convolutional feature maps are extracted with HRNet-18 and are subsequently used to output the center heatmap, the center offset map, the corner offset map, and the semantic mask. t peaks determine center points in the center heatmap and corresponding values in the center offset map. Corner offset coordinates and center locations are finally combined to compute the output corner coordinates.

5) Corner Offset: Corner offsets are built to recover absolute locations of corner points in an image. Therefore, the corner offset map contains 4×2 channels, because two-dimensional coordinates of four corners form a vertebra in an image.

6) From Map to Keypoint: Top t peaks which values are no less than their eight-connected neighbors in the center heatmap are selected: $\{p_1, p_2, ..., p_t\}, p_i = (x_i, y_i)$, and according to offsets in the center offset maps: $\{\Delta p_1, \Delta p_2, ..., \Delta p_t\}, \Delta p_i = (\Delta x_i, \Delta y_i)$, locations of center points become $\{p_1 + \Delta p_1, p_2 + \Delta p_2, ..., p_t + \Delta p_t\}$.

7) Semantic Segmentation: The semantic segmentation task is added to make feature maps focus on the spine more directly, compared to keypoint regression.

B. Heterogeneous Consistency Loss

Convolutional features of vertebrae are captured and implicitly expressed by the output branches, and even they should be consistent even in different representations. Therefore, we propose a heterogeneous consistency loss to harmonize the differences brought by various representations.

A sequence $\{v_1, v_2, ..., v_t\}$ contains t vertebrae, where $v_i = \{p_{r,1}, p_{r,2}, p_{r,3}, p_{r,4}\}$ denotes i-th vertebra with four corner points. The output mask for training a keypoint estimator consists of a synthetic mask generated from keypoints and a semantic mask directly from the segmentation branch:

$$\begin{aligned} M_{kpt} &= FWT(v_1, v_2, ..., v_t) \\ M_{prd} &= \alpha_{csc} M_{kpt} + (1 - \alpha_{csc}) M_{msk} \end{aligned} \tag{1}$$

where M_{kpt}, M_{msk}, and M_{prd} are the synthetic mask, the semantic mask, and the predicted mask, respectively. FWT represents a simple image processing function to fill all the pixels inside every vertebra with white color given a completely black template, and α_{csc} is a weight to balance the above two masks.

Therefore, we define the heterogeneous consistency loss $L_{sp,csc}$ with a focal loss for pixel-wise spine classification.

C. Multi-task Loss

To train the keypoint estimator end-to-end in this task, we define an overall loss function for optimizing the weights of all the branches:

$$\begin{aligned} L_{total} &= L_{ctr,hm} + L_{ctr,off} + L_{cnr,off} + \alpha_{seg} L_{sp,csc} \\ &= L_{ctr,fcl} + L_{ctr,1} + L_{cnr,1} + \alpha_{seg} L_{sp,fcl} \end{aligned} \tag{2}$$

where $L_{ctr,hm}$, $L_{ctr,off}$, $L_{cnr,off}$, and $L_{sp,csc}$ denote losses for center heatmap, center offset, corner offset, and spine segmentation. Specifically, Focal loss $L_{\cdot,fcl}$ [9] and L1 loss $L_{\cdot,1}$ are applied for these tasks. α_{seg} balances the spine segmentation loss and other keypoint localization losses.

III. EXPERIMENTS

A. Dataset

Dataset 16 in AASCE MICCAI 2019 Challenge is used for evaluation. Concretely, every spine contains 17 vertebrae, and a vertebra has four keypoints. All the keypoints are manually labeled and divided by doctors to prevent patients from appearing in both training and test subsets. Consequently, there are 481 AP X-ray images for training and 128 counterparts for the test. Finally, the sub-challenge organizers provide Cobb angles using fixed rules.

B. Implementation Details

A vertebra-focused landmark detection network [6] is selected as our baseline, and ResNet-34 [12] with pre-trained weights is its backbone while HRNet-18 is ours. Due to the

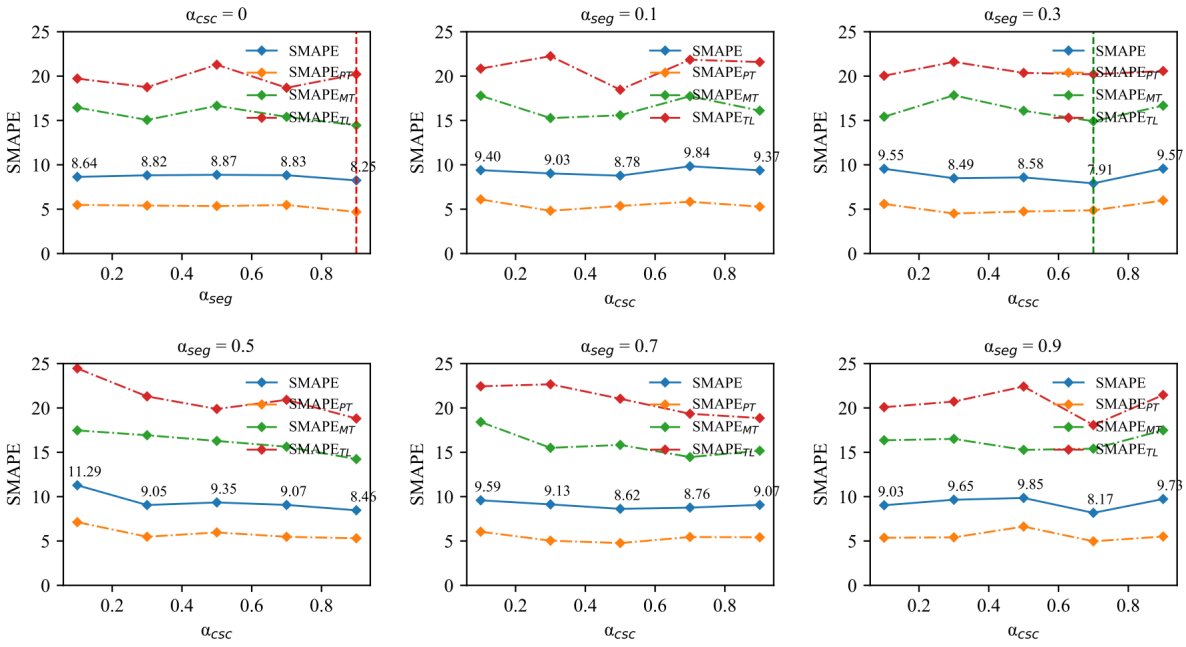


Fig. 2. Results of keypoint estimators using different parameters: models perform better if the synthetic mask from keypints and the semantic mask is properly fused. Consequently, SMAPE from the model trained with a heterogeneous consistency loss can reach 7.91, which is lower than that with a typical semantic segmentation loss.

TABLE I

COBB ANGLE ESTIMATION RESULTS ON TEST SET.

Method	Input Resolution	SMAPE↓	$SMAPE_{PT}$ ↓	$SMAPE_{MT}$ ↓	$SMAPE_{TL}$ ↓	MSE ↓	FPS↑
Multi-view Extrapolation Net [10]	512×256	23.43	16.38	30.27	35.61	77.94	11.40
Residual U-Net [11]	1024×512	16.48	9.71	25.97	33.01	74.07	2.38
Landmark Detection Network [6]	1024×512	10.81	6.26	18.04	23.42	**50.11**	5.65
Landmark Detection Network [6]	1024×512	9.71	6.22	15.39	22.39	61.90	**14.28**
Keypoint Estimator	1024×512	9.97	6.65	**14.94**	21.27	56.04	12.46
Ours	1024×512	**8.62**	**4.76**	15.83	**21.04**	52.72	12.33

small size of the dataset, we apply 15-fold cross-validation to the training set. Epochs and batch size for model training are respectively 50 and 2. The optimizer is Adam [13] with the initial learning rate 1.25×10^{-4}. When the model has a smaller loss on the current split validation subset than all the past ones, its weights are saved or covered. In addition, n_h and n_w are respectively set as 1024 and 512.

IV. EVALUATION METRICS

To evaluate the performance of a Cobb angle estimator, we use Symmetric Mean Absolute Percentage Error (SMAPE):

$$SMAPE = \frac{1}{n_I} \sum_{i=1}^{n_I} \frac{\sum_{c=1}^{3} |\hat{a}_{i,c} - a_{i,c}|}{\sum_{c=1}^{3} (\hat{a}_{i,c} + a_{i,c})} \tag{3}$$

where $\hat{a}$ and a denote respectively the predicted Cobb angle and the ground truth one. The test set has n_I images, then three Cobb angles in Proximal Thoracic (PT), Main Thoracic (MT), and ThoracoLumbar (TL) are evaluated in every image, and we abbreviate them as $SMAPE_{PT}$, $SMAPE_{MT}$, and $SMAPE_{TL}$.

To measure the distance between predicted keypoints and ground truth ones, we apply Mean Squared Error (MSE):

$$MSE = \frac{1}{n_p} \sum_{i=1}^{n_p} (\hat{p}_i - p_i)^2 \tag{4}$$

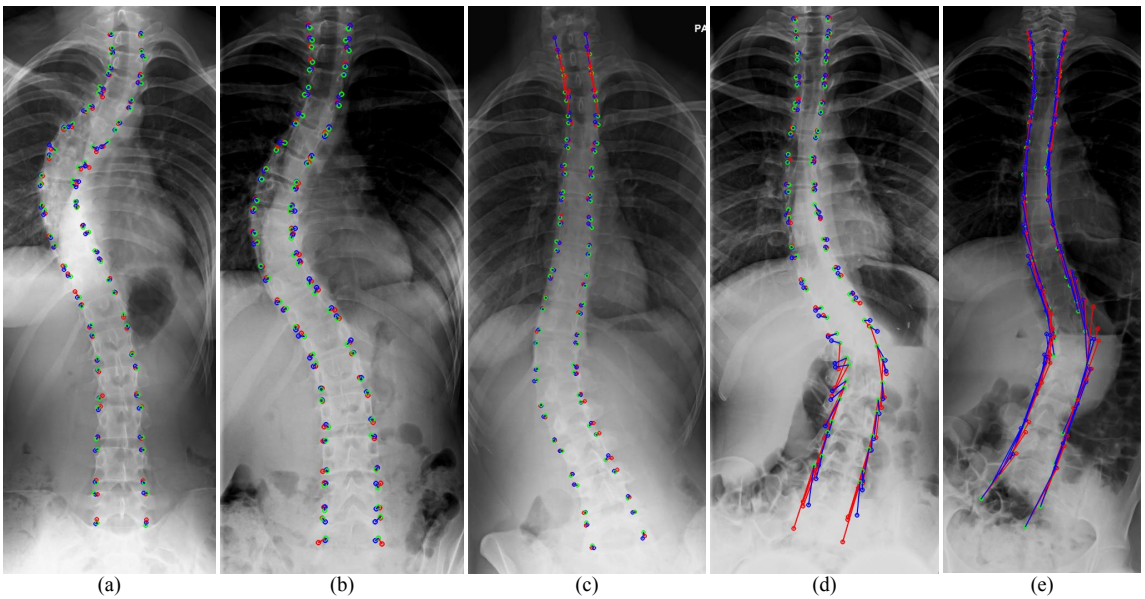
where the test set has n_p keypoints, and p_i is the i-th keypoint.

V. ABLATION STUDIES

We take ablation studies for our Cobb angle estimator, as shown in Table I and Figure 2. In Table I, the results of "Multi-view Extrapolation Net", "Residual U-Net", and "Landmark Detection Network" at the top three rows come from previous works. To make fair comparisons, using the same implementation details, "Landmark Detection Network" is chosen and re-experimented, and its variant with a different backbone HRNet-18 is called "Keypoint Estimator".

A. Effectiveness of HRNet

Keypoint locations are significantly improved using HRNet-18, compared to ResNet-34. Specifically, MSE of "Keypoint Estimator" drops from 61.90 to 56.04, compared to this metric of "Landmark Detection Network", although SMAPE and FPS slightly decrease by 0.26 and 1.82.



(a) (b) (c) (d) (e)

Fig. 3. Keypoint predictions on test images: red, blue, and green circles are keypoints from the baseline, our method, and ground truth annotations. Accordingly, red and blue lines represent distances from baseline predictions or ours to ground truth labels. From Subfigure (a) to Subfigure (d), we can properly estimate Cobb angles using both methods; even large offsets of keypoints exist in Subfigure (e), our method can still maintain a relatively correct spine structure.

B. Effectiveness of Semantic Segmentation

The semantic segmentation slightly affects the performances of keypoint regression, but it significantly helps Cobb angle estimation. MSE of "Keypoint Estimator" increases from 61.90 to 61.98, but SMAPE reduces from 9.97 to 8.25.

C. Effectiveness of Heterogeneous Consistency Loss

Estimators trained with the heterogeneous consistency loss outperform models without it if such consistency loss and common semantic segmentation loss are properly fused. It achieves the lowest SMAPE 7.91 not surprisingly because this consistency loss is extended exactly from the semantic segmentation one. However, its MSE increases to 64.19, which indicates that Cobb angle estimation does not always perform monotonously with corner point regression. Therefore, we select the model with SMAPE 8.62 and MSE 52.72 for the multi-task balance.

VI. CONCLUSION

In summary, we present a new loss function that enhances the consistency between keypoints and semantic masks for Cobb angle estimation. Our keypoint estimator is built on successful feature extraction layers and landmark detection networks. The heterogeneous consistency loss is simple and effective without slowing down the inference speed, and it pays more attention to the global spine structure described from keypoints. Extensive experimental results on X-ray images from AASCE MICCAI 2019 Challenge demonstrate its potentials to train more accurate Cobb angle estimators.

REFERENCES

[1] A. Safari, H. Parsaei, A. Zamani, et al., "A semi-automatic algorithm for estimating cobb angle," *J. Biomed.Phys. Eng.*, vol. 9, no. 3, pp. 317–326, 2019.

[2] B. Chen, Q. Xu, L. Wang, et al., "An automated and accurate spine curve analysis system," *IEEE Access*, vol. 7, pp. 124596–124605, 2019.

[3] B. Khanal, L. Dahal, P. Adhikari, et al., "Automatic cobb angle detection using vertebra detector and vertebra corners regression," *arXiv preprint arXiv:1910.14202*, 2019.

[4] R. Choi, K. Watanabe, H. Jingguiji, et al., "Cnn-based spine and cobb angle estimator using moire images," *IIEEJ Trans. Image Electron. Vis. Comput.*, vol. 5, no. 2, pp. 135–144, 2017.

[5] H. Wu, C. Bailey, P. Rasoulinejad, et al., "Automatic landmark estimation for adolescent idiopathic scoliosis assessment using boostnet," *IEEE Conf. Med. Image. Comput. Comput. Assist. Interv.*, pp. 127–135, 2017.

[6] J. Yi, P. Wu, Q. Huang, et al., "Vertebra-focused landmark detection for scoliosis assessment," *Int. J. Biomed. Imaging*, pp. 736–740, 2020.

[7] X. Zhou, D. Wang, and P. Krähenbühl, "Objects as points," *arXiv preprint arXiv:1904.07850*, 2019.

[8] J. Wang, K. Sun, T. Cheng, et al., "Deep high-resolution representation learning for visual recognition," *IEEE Trans. Pattern Anal. Mach. Intell.*, pp. 1–1, 2020.

[9] T. Lin, P. Goyal, R. Girshick, et al., "Video panoptic segmentation," *IEEE Int. Conf. Comput. Vis.*, pp. 2999–3007, 2017.

[10] L. Wang, Q. Xu, S. Leung, et al., "Accurate automated cobb angles estimation using multi-view extrapolation net," *Med. Image Anal.*, vol. 58, pp. 101542, 2019.

[11] M. H. Horng, C. P. Kuok, M. J. Fu, et al., "Cobb angle measurement of spine from x-ray images using convolutional neural network," *Comput. Math. Methods Med.*, vol. 6357171, pp. 1–18, 2019.

[12] K. He, X. Zhang, S. Ren, et al., "Deep residual learning for image recognition," *IEEE Conf. Comput. Vis. Pattern Recognit.*, pp. 770–778, 2016.

[13] D. P. Kingma and L. J. Ba, "Adam: A method for stochastic optimization," *IEEE Int. Conf. Learn. Representations.*, 2015.