Progressively Correcting Soft Labels via Teacher Team for Knowledge Distillation in Medical Image Segmentation

Yaqi Wang[1,2], Peng Cao[1,2,3,(✉)], Qingshan Hou[1,2], Linqi Lan[1,2], Jinzhu Yang[1,2,3], Xiaoli Liu[1], and Osmar R. Zaiane[4]

[1] Computer Science and Engineering, Northeastern University, Shenyang, China
[2] Key Laboratory of Intelligent Computing in Medical Image of Ministry of Education, Northeastern University, Shenyang, China
[3] National Frontiers Science Center for Industrial Intelligence and Systems Optimization, Shenyang, China
caopeng@mail.neu.edu.cn
yangjinzhu@cse.neu.edu.cn
[4] Alberta Machine Intelligence Institute, University of Alberta, Edmonton, Canada

Abstract. State-of-the-art knowledge distillation (KD) methods aim to capture the underlying information within the teacher and explore effective strategies for knowledge transfer. However, due to challenges such as blurriness, noise, and low contrast inherent in medical images, the teacher's predictions (soft labels) may also include false information, thus potentially misguiding the student's learning process. Addressing this, we pioneer a novel correction-based KD approach (PLC-KD) and introduce two assistants for perceiving and correcting the false soft labels. More specifically, the false-pixel-aware assistant targets global error correction, while the boundary-aware assistant focuses on lesion boundary errors. Additionally, a similarity-based correction scheme is designed to forcefully rectify the remaining hard false pixels. Through this collaborative effort, the *teacher team* (comprising a teacher and two assistants) progressively generates more accurate soft labels, ensuring the "all-correct" final soft labels for student guidance during KD. Extensive experimental results demonstrate that the proposed PLC-KD framework attains superior performance to state-of-the-art methods on three challenging medical segmentation tasks.

Keywords: Knowledge Distillation · Medical Image Segmentation · Label Correction.

1 Introduction

Knowledge distillation (KD) [3], as a popular model compression technique, has been widely applied in various fields of deep learning. This traditional paradigm concentrates on transferring knowledge from a heavy model (teacher) to a light one (student), which aims to obtain a student closer to the teacher but with fewer

parameters. Existing KD methods focus on capturing underlying knowledge from the teacher and devising effective transfer strategies, e.g., response-based [3,9,16], feature-based [7,15,6], and relation-based [6,11,4,13,14] approaches. However, in challenging tasks such as medical image segmentation, the sophisticated teacher network inevitably produces false predictions, thereby potentially misleading the student during the KD process. Intuitively, correcting teachers' false predictions (soft labels) can enhance student learning. Unfortunately, none of the existing KD works has delved into this aspect. A straightforward strategy is to forcibly correct these false predictions to fixed correct values, but it leads to a sub-optimal model since it ignores the reasonable confidence of soft labels which is crucial for knowledge transfer during KD. In this work, we offer a principled investigation of how the correction of soft labels can be applied in KD. After a thorough analysis, we pinpoint three factors contributing to the teacher's unsatisfactory performance, ranging from easy to difficult: 1) global false pixels produced by the teacher network, 2) blur boundary pixels due to the low contrast between lesions/tissues and the background, and 3) hard pixels that are difficult for segmentation models to discern. Taking these into consideration, three questions naturally arise: 1) how to encourage the teacher network to concentrate on false-predicted pixels, 2) how to enhance teachers' awareness of lesion boundaries, thereby averting misguided boundary guidance, and 3) how to correct hard pixels that can not be corrected easily, ensuring the "all-correct" soft labels?

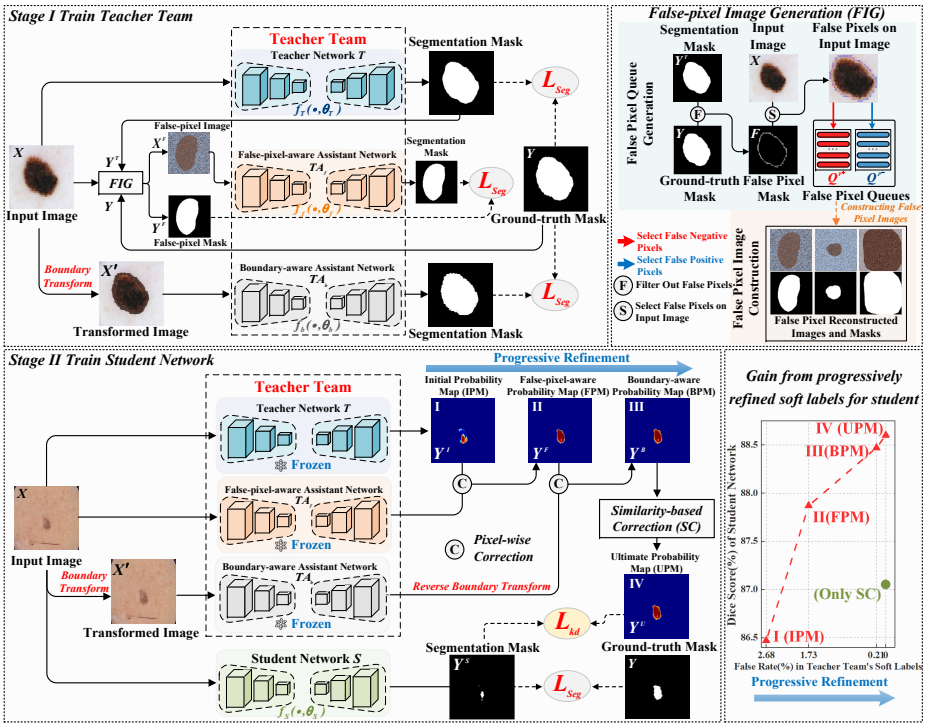
To overcome these challenges, we propose a **P**rogressive soft **L**abel **C**orrection **K**nowledge **D**istillation framework, termed PLC-KD, for medical image segmentation. In particular, the PLC-KD framework encompasses three components for enhancing soft labels: 1) the False-pixel Image Generation (FIG) module for generating synthetic images utilizing pixels mispredicted by the teacher network, 2) the Boundary Transform (BT) operation for generating images with easily discernible lesion boundaries, and 3) the Similarity-based Correction (SC) scheme for forcibly correcting hard false pixels. To this end, a **teacher team** is established, comprising a teacher with two assistants. Firstly, the teacher network generates an initial probability map (soft label), serving as the foundation for guiding the student. Subsequently, the false-pixel-aware assistant, trained on the FIG-generated false-pixel images, aims to global error correction, while the boundary-aware assistant, trained on the BT-generated boundary transform images, further refines segmentation boundary details. The entire teacher team collaborates to generate as accurate soft labels as possible. Finally, the SC scheme is applied to forcefully correct the remaining hard false pixels. Following such corrections, the obtained "all-correct" soft labels are available to guide the student's learning, thus leading to better students.

To the best of our knowledge, this is the first work to rethink knowledge distillation from the perspective of soft label correction. In summary, our contributions can be outlined as follows. (1) We provide an insightful view to study vanilla KD by correcting the teacher's soft labels and propose a novel PLC-KD framework to correct the teacher's soft labels progressively and hence promote better students. (2) We establish a collaborative teacher team to generate soft

labels for student learning. The two assistants, trained specifically with generated false-pixel and boundary transform images, aim to enhance the capabilities in both global error correction and boundary refinement, which ultimately leads to accurate soft labels. (3) Our PLC-KD framework has undergone rigorous quantitative and qualitative evaluations on three medical image segmentation benchmark datasets. The results prove that PLC-KD significantly outperforms the previous best methods.

2 Methodology

The proposed PLC-KD framework comprises a **teacher team** and a student network S. The teacher team includes a teacher network T and two assistants: a false-pixel-aware assistant network TA_f and a boundary-aware assistant network TA_b, where TA_f and TA_b share the same architecture as T. TA_f is trained with images generated by the FIG module, endowing it with global error correction capabilities and complementing the mistakes made by T. While TA_b is trained with images generated by the BT operation, which concentrates on the perception of segmentation boundaries. The entire teacher team collaborates



Pixel-wise Correction: *Replace the false predicted values in the current probability map with the corresponding correctly predicted values from another probability map.*
Correctly predicted: *Predicted values between 0.5 and 1 are considered correct for positive pixels. For negative pixels, the opposite criterion applies.*

Fig. 1. The overall architecture of PLC-KD.

to progressively generate more accurate soft labels. Under this framework, the pipeline of knowledge distillation is composed of two major stages (see Fig 1). First, we train the teacher networks T and assistants: TA_f and TA_b. Next, we utilize the pre-trained teacher team to guide the learning process of the student network. Notably, although both the "all-correct" ultimate probability map Y^U and the ground-truth mask Y contain fully accurate labels, Y^U provides soft labels that can transfer extra knowledge to the student.

2.1 PLC-KD Pipline

Stage 1: Train the teacher team. Initially, the teacher network T is directly trained using ground-truth mask Y. Subsequently, the false-pixel image generation module is employed to construct false-pixel images and their corresponding masks based on the false-predicted pixels by T, which are then utilized to train the false-pixel-aware assistant network TA_f. Notably, the FIG module comprises two components, "False Pixel Queue Generation" and "False Pixel Image Construction", both of which are performed sequentially during training. Finally, the boundary-aware assistant network TA_b is trained with boundary transform generated images.
Stage 2: Train the student network. The input image X first goes through the teacher network T to obtain the initial probability map Y^I. Assistants TA_f, TA_b, and the SC scheme, then progressively refine this probability map to obtain the "all-correct" ultimate probability map Y^U, which serves as the guidance for the student network S. Note that the designed SC scheme computes the cosine similarity between the misclassified pixels and correctly predicted pixels. Then, the value of the most similar and correct pixel from Y^B is assigned to the corresponding false pixel to obtain the ultimate probability map Y^U.

2.2 False-pixel Image Generation (FIG)

To comprehensively rectify false predictions for the teacher network, we propose the false-pixel image generation module, which involves "False Pixel Queue Generation" and "False Pixel Image Construction", as illustrated in Fig. 1. Notably, only the former operation is performed during the initial training phase until the false pixel queues are filled. Subsequently, both operations are alternately performed.

False Pixel Queue Generation. Given an input image $X \in \mathbb{R}^{H\times W\times C}$, the pre-trained teacher network T produces a segmentation mask Y^T. Afterward, false predicted pixels are filtered out by comparing Y^T with the corresponding ground-truth mask Y, resulting in the false pixel mask F. The procedure can be expressed as $F = f_T(X;\theta_T)\,\textcircled{F}\,Y$, where $f_T(\cdot,\theta_T)$ represents the teacher network and its parameters, while $\textcircled{F}$ denotes the filtering of false pixels. Subsequently, pixels corresponding to F in the original input image X are selected and divided into false negative pixels $P^+ = \{p_1^+, p_2^+, ..., p_m^+\}$ and false positive pixels $P^- =$

$\{p_1^-, p_2^-, ..., p_n^-\}$. Finally, k pixels are randomly selected from P^+ and P^- and added to the corresponding false negative pixel queue $Q^{P^+} \in \mathbb{R}^{M \times D}$ and false positive pixel queue $Q^{P^-} \in \mathbb{R}^{N \times D}$, respectively, where M and N represent the queue lengths, and $D = 3$ denotes the three RGB channels.

False Pixel Image Construction. Next, we elaborate on false pixel image construction. To enhance the performance for false predicted pixels in the teacher network T (i.e., pixels in queues Q^{P^+} and Q^{P^-}), we sample positive and negative pixels from Q^{P^+} and Q^{P^-}, respectively, for constructing synthetic images and their corresponding ground truth masks. Specifically, to simulate real medical images, we initialize a blank image $x_i^r \in \mathbb{R}^{H \times W \times C}$, and randomly select a ground-truth lesion mask as the lesion region with its pixel values obtained randomly from Q^{P^+}. Conversely, other regions serve as background regions, with pixel values randomly obtained from Q^{P^-}. In this way, we reconstruct images $X^r = \{x_1^r, x_2^r, ..., x_b^r\}$ and their corresponding ground truth masks $Y^r = \{y_1^r, y_2^r, ..., y_b^r\}$, where b denotes the batch size. Subsequently, we employ the false-pixel-aware assistant network TA_f to learn from these reconstructed images.

2.3 Boundary Transform (BT) Operation

The false-pixel-aware assistant network TA_f effectively corrects errors produced by T but overlooks the error-prone segmentation boundaries. To further enhance the boundary perception of the teacher's soft label, we propose a simple plug-and-play boundary transform operation. As shown in Fig. 2(f), the grayscale histogram distribution in the inner region exhibits significant discrepancy from the background region, while the distribution of the boundary region is closer to the background region, with potential overlap. Besides, convolution operations can capture local patterns due to their inherent kernel design [1]. Hence, the pixels near the lesion boundary are more likely to be misclassified as background regions. Inspired by this, we conjecture that swapping pixels near the lesion

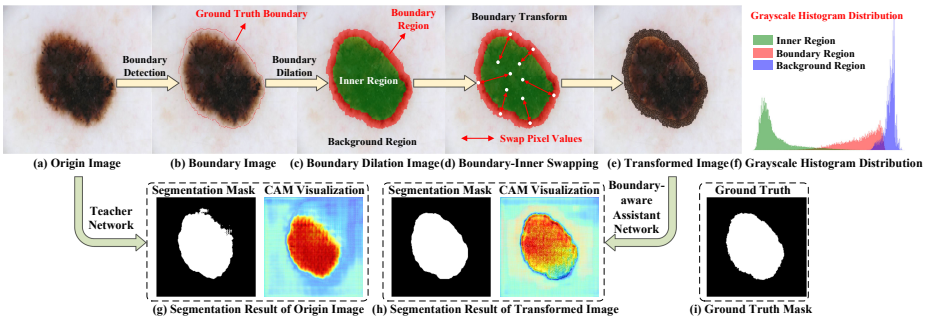


Fig. 2. Boundary Transform Operation.

boundary with those from the lesion interior can promote the model's ability to perceive boundary pixels. In this way, for a given image X, we first detect the boundaries of the ground truth lesion region (see Fig. 2(b)). To further identify the boundary region, a dilation operation is applied to expand δ pixels from the boundaries towards the interior of the lesion. Therefore, the remaining region within the lesion is considered the inner region, as depicted in Fig. 2(c). To further enhance the prediction accuracy of boundary pixels, we randomly sample an equal number of pixels from the boundary and inner regions and exchange their pixel values to create a transformed image X' for training the boundary-aware assistant network TA_b. Following this operation, the boundary pixels within the inner region exhibit a lower segmentation difficulty compared to the lesion boundary because they are surrounded by homogeneous lesion regions. As a result, the well-trained assistant network TA_b aims at correcting boundary errors, leading to better segmentation results (Fig. 2(g, h))). Finally, it is employed to correct boundary errors in Y_F and generate the boundary-aware probability map Y_B (Fig. 1).

2.4 Overall Framework

Overall, given the training dataset $\mathcal{D} = \{(X_i, Y_i)_{i=1}^N\}$, where X_i and Y_i denote the i_{th} image and the corresponding ground truth mask, and N is the number of samples. The segmentation losses for the teacher team (corresponding to T, TA_f, and TA_b) are formulated as $\mathcal{L}_T = \frac{1}{N}\sum_{i=1}^{N}\mathcal{L}_{Seg}(f_T(X_i,\theta_T),Y_i); \mathcal{L}_F = \frac{1}{M}\sum_{i=1}^{M}\mathcal{L}_{Seg}(f_f(X_i^r,\theta_f),Y_i^r); \mathcal{L}_B = \frac{1}{N}\sum_{i=1}^{N}\mathcal{L}_{Seg}(f_b(X_i',\theta_b),Y_i)$, where M is the number of reconstructed images and $\mathcal{L}_{Seg}$ can be any supervised semantic segmentation loss. Note that we chose the combination of cross entropy loss and dice loss in our experiments, as it has a compelling performance in medical image segmentation. Meanwhile, the student loss function is expressed as:

$$\mathcal{L}_S = \frac{1}{N}\sum_{i=1}^{N}\{\mathcal{L}_{Seg}(f_S(X_i,\theta_S),Y_i) + \lambda\mathcal{L}_{KD}(f_S(X_i,\theta_S),Y_i^U)\}, \tag{1}$$

where $\mathcal{L}_{KD}$ represents the distillation loss, and we adopt the KL divergence paradigm. The parameter λ is a regularization parameter to balance the segmentation and KD losses. It is worth noting that the teacher networks are discarded in the inference phase after training.

3 Experiments

Dataset and Implementation Details. We conduct extensive experiments on three different medical image segmentation tasks to evaluate the proposed method. (1) For skin lesion segmentation, we validate our method on the ISIC-2018 dataset [10], which includes 2,694 images for training and 1,000 images for testing with their corresponding annotations. (2) For polyp segmentation, we evaluate the proposed method on the public CVC-EndoSceneStill dataset [12],

Table 1. Comparison with state-of-the-art KD methods on three datasets, highlighting the best and second-best scores with bold and underlined, respectively.

				ISIC		CVC-EndoSceneStill		IDRiD-OD	
				Dice(%)↑	IoU(%)↑	Dice(%)↑	IoU(%)↑	Dice(%)↑	IoU(%)↑
Teacher: TransUNet [2]				88.99±0.24	81.52±0.42	79.53±3.27	71.64±3.42	90.97±3.07	84.82±4.01
Student: UNet [8]				86.40±0.51	77.92±0.70	40.75±5.58	29.07±4.54	88.27±2.12	82.09±2.25
Methods	Response	Feature	Relation						
KD [3]	✓	✗	✗	86.47±0.43	77.97±0.64	63.95±2.17	55.01±2.46	88.31±3.52	81.30±4.95
FitNet [7]	✓	✓	✗	86.96±0.19	78.70±0.21	65.76±3.31	56.20±3.16	88.44±2.57	81.66±3.47
AT [15]	✓	✓	✗	87.02±0.35	78.85±0.54	67.79±2.71	58.79±2.30	89.14±2.69	82.80±3.31
EMKD [6]	✓	✓	✓	87.11±0.51	78.95±0.75	65.15±4.03	55.95±4.27	<u>90.07±2.25</u>	<u>83.76±3.19</u>
CWD [9]	✓	✗	✗	87.09±0.46	78.90±0.67	<u>68.87±2.92</u>	<u>60.15±3.54</u>	89.61±2.29	82.86±3.47
SPKD [11]	✓	✗	✓	86.58±0.44	78.09±0.65	63.22±4.16	53.73±3.70	88.86±3.07	82.27±4.46
PA [4]	✓	✗	✓	84.33±0.90	75.25±1.14	62.81±3.48	51.39±2.05	88.42±1.16	82.17±3.55
IFVD [13]	✓	✗	✓	86.71±0.44	78.32±0.70	65.23±1.52	56.14±1.57	87.81±1.26	81.11±1.02
CIRKD [14]	✓	✗	✓	86.96±0.42	78.72±0.68	66.17±3.15	56.73±3.72	89.29±1.74	82.49±2.33
DKD [16]	✓	✗	✗	<u>87.14±0.18</u>	<u>79.02±0.24</u>	68.16±3.85	59.01±3.63	89.36±2.36	82.87±3.38
PLC-KD(Ours)	✓	✗	✗	**88.60±0.51**	**81.63±0.60**	**71.22±1.51**	**62.15±1.88**	**92.13±0.58**	**86.56±0.82**

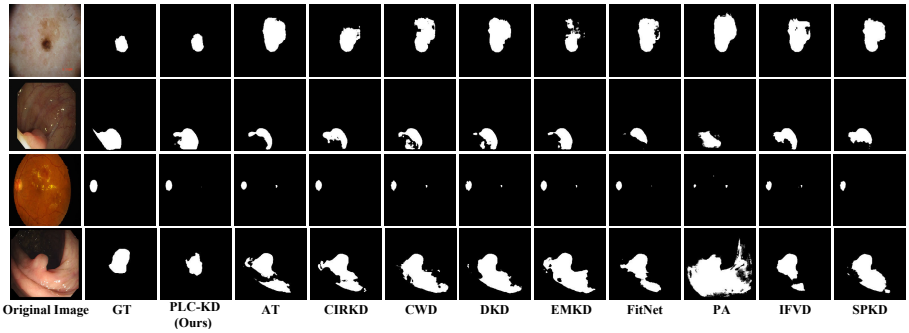


Fig. 3. Qualitative examples on three datasets.

containing 612 colonoscopy images. The dataset is split into 547 training frames, 183 validation frames, and 182 testing frames. (3) For Optic Disk segmentation, we evaluate the proposed method on the IDRiD dataset [5], which contains 81 fundus images with pixel-level annotations. The partition of the training set and testing set has been fixed, i.e., 54 images for training and the rest 27 images for testing. **More training details are in the supplementary material.**

Comparison with the State-of-the-Art Methods. We provide qualitative and quantitative comparisons with state-of-the-art KD methods on the well-performed network TranUNet [2] (teacher) and the common medical image segmentation network UNet [8] (student). Contenders encompass response-based methods [3,9,16], feature-based methods [7,15,6], and relation-based methods [6,11,4,13,14]. The first two categories aim to minimize the divergence between the teacher and student's output logits and intermediate feature maps, respectively. Conversely, the last category aims to capture implicit knowledge embedded in the teacher network. Our method belongs to the simplest first category without requiring complex intermediate feature map knowledge or relational

Table 2. Ablation Study of the different component combinations on three datasets.

Models	Components			ISIC		EndoScene		IDRiD-OD	
	TA_f	TA_b	SC	Dice(%)↑	IoU(%)↑	Dice(%)↑	IoU(%)↑	Dice(%)↑	IoU(%)↑
$VanillaKD$	✗	✗	✗	86.47±0.43	77.97±0.64	63.95±2.17	55.01±2.46	88.31±3.52	81.30±4.95
$T + SC$	✗	✗	✓	87.01±0.47	78.85±0.65	65.46±4.20	54.74±4.82	89.06±1.29	82.20±1.70
$T + TA_f$	✓	✗	✗	87.87±0.13	80.32±0.13	69.82±1.15	59.97±1.79	89.78±1.17	83.11±1.89
$T + TA_b$	✗	✓	✗	87.93±0.35	80.43±0.58	70.19±1.67	60.97±2.16	90.22±1.08	83.68±1.22
$T + TA_f + TA_b$	✓	✓	✗	88.47±0.37	81.48±0.79	70.86±1.24	61.88±1.41	91.63±1.34	85.16±1.70
PLC-KD	✓	✓	✓	**88.60±0.51**	**81.63±0.60**	**71.22±1.51**	**62.15±1.88**	**92.13±0.58**	**86.56±0.82**

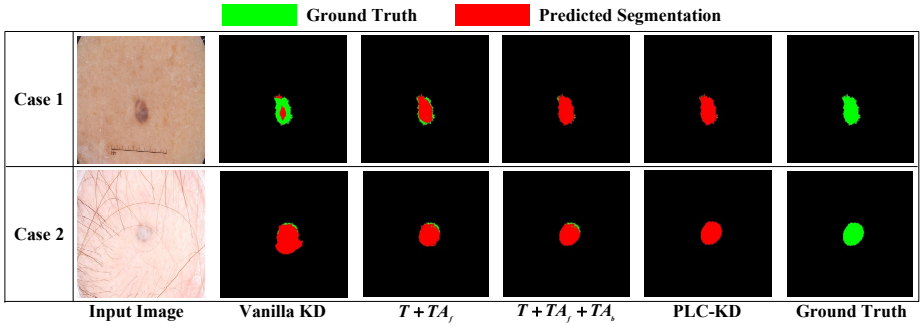


Fig. 4. The visualization of student segmentation results with progressive correction of soft labels, where the models correspond to those in Table 2. Case 1: correcting false negative pixels; Case 2: correcting false positive pixels.

knowledge. As reported in Tabel 1, PLC-KD substantially outperforms all previous methods and sets new state-of-the-art for both the Dice and IoU on three datasets. Moreover, as shown in Fig. 3, our method generates more accurate results by recovering finer segmentation details. These results imply that ensuring the correctness of soft labels can provide more precise segmentation knowledge, thereby enhancing the student's lesion perception.

Ablation Study. We conduct an ablation study in terms of soft label correction components. Tabel 2 reports the performance improvements over the baseline. Initially, we confirm that exclusively enforcing corrections to soft labels ($T + SC$) fails to achieve satisfactory performance due to the inherent lack of reasonable confidence in its generation of soft labels. It shows a trend that the segmentation performance improves when the components, including the assistant networks TA_f, TA_b, and SC scheme are incorporated into the baseline, and again confirms the necessity of the "all-correct" soft labels. Taking the ISIC dataset as an example, the incorporation of TA_f network ($T + TA_f$) results in an improvement of 1.40% and 2.35% in terms of Dice and IoU, which demonstrates the contribution of false-pixel-aware network correction. Building upon this, the incorporation of TA_b network ($T + TA_f + TA_b$) further improves performance by 0.6% and 1.16%, indicating the effectiveness of lesion boundary correction in enhancing the student network. Ultimately, the SC scheme boosts

performance to 88.60 in Dice and 81.63 in IoU. Moreover, the visualization effect in Fig. 4 further demonstrates the efficacy in correcting false negative and false positive pixels with progressive correction of soft labels. Additionally, similar performance improvements are observed across other datasets, underscoring the effectiveness and complementarity of each component for soft label correction.

4 Conclusion

We present a novel viewpoint to correct the false soft labels for knowledge distillation in medical image segmentation. Our main idea lies in relearning synthetic images produced by teacher errors and correcting them progressively, thus ensuring the transfer of "all-correct" knowledge to the student. Extensive experiments demonstrate the effectiveness of this idea and show that the proposed PLC-KD can achieve state-of-the-art performance. We hope this paper will contribute to future correction-based knowledge distillation research.

Acknowledgments

This research was supported by the National Natural Science Foundation of China (No.62076059) and the Science and Technology Joint Project of Liaoning province (2023JH2/101700367, ZX20240193).

References

1. Bragagnolo, L., Rezende, L., da Silva, R., Grzybowski, J.: Convolutional neural networks applied to semantic segmentation of landslide scars. Catena **201**, 105189 (2021)
2. Chen, J., Lu, Y., Yu, Q., Luo, X., Adeli, E., Wang, Y., Lu, L., Yuille, A.L., Zhou, Y.: Transunet: Transformers make strong encoders for medical image segmentation. arXiv preprint arXiv:2102.04306 (2021)
3. Hinton, G., Vinyals, O., Dean, J.: Distilling the knowledge in a neural network. arXiv preprint arXiv:1503.02531 (2015)
4. Liu, Y., Chen, K., Liu, C., Qin, Z., Luo, Z., Wang, J.: Structured knowledge distillation for semantic segmentation. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 2604–2613 (2019)
5. Porwal, P., Pachade, S., Kokare, M., Deshmukh, G., Son, J., Bae, W., Liu, L., Wang, J., Liu, X., Gao, L., et al.: Idrid: Diabetic retinopathy–segmentation and grading challenge. Medical image analysis **59**, 101561 (2020)
6. Qin, D., Bu, J.J., Liu, Z., Shen, X., Zhou, S., Gu, J.J., Wang, Z.H., Wu, L., Dai, H.F.: Efficient medical image segmentation based on knowledge distillation. IEEE Transactions on Medical Imaging **40**(12), 3820–3831 (2021)
7. Romero, A., Ballas, N., Kahou, S.E., Chassang, A., Gatta, C., Bengio, Y.: Fitnets: Hints for thin deep nets. arXiv preprint arXiv:1412.6550 (2014)
8. Ronneberger, O., Fischer, P., Brox, T.: 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. pp. 234–241. Springer (2015)

9. Shu, C., Liu, Y., Gao, J., Yan, Z., Shen, C.: Channel-wise knowledge distillation for dense prediction. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 5311–5320 (2021)
10. Sirinukunwattana, K., Pluim, J.P., Chen, H., Qi, X., Heng, P.A., Guo, Y.B., Wang, L.Y., Matuszewski, B.J., Bruni, E., Sanchez, U., et al.: Gland segmentation in colon histology images: The glas challenge contest. Medical image analysis **35**, 489–502 (2017)
11. Tung, F., Mori, G.: Similarity-preserving knowledge distillation. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 1365–1374 (2019)
12. Vázquez, D., Bernal, J., Sánchez, F.J., Fernández-Esparrach, G., López, A.M., Romero, A., Drozdzal, M., Courville, A., et al.: A benchmark for endoluminal scene segmentation of colonoscopy images. Journal of healthcare engineering **2017** (2017)
13. Wang, Y., Zhou, W., Jiang, T., Bai, X., Xu, Y.: Intra-class feature variation distillation for semantic segmentation. In: Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part VII 16. pp. 346–362. Springer (2020)
14. Yang, C., Zhou, H., An, Z., Jiang, X., Xu, Y., Zhang, Q.: Cross-image relational knowledge distillation for semantic segmentation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 12319–12328 (2022)
15. Zagoruyko, S., Komodakis, N.: Paying more attention to attention: Improving the performance of convolutional neural networks via attention transfer. arXiv preprint arXiv:1612.03928 (2016)
16. Zhao, B., Cui, Q., Song, R., Qiu, Y., Liang, J.: Decoupled knowledge distillation. In: Proceedings of the IEEE/CVF Conference on computer vision and pattern recognition. pp. 11953–11962 (2022)