Paced-Curriculum Distillation with Prediction and Label Uncertainty for Image Segmentation

Mobarakol Islam[1†], Lalithkumar Seenivasan[2†], S P Sharan[3],
V.K. Viekash[4], Bhavesh Gupta[5], Ben Glocker[1] and
Hongliang Ren[2,6∗]

Abstract

Purpose: In curriculum learning, the idea is to train on easier samples first and gradually increase the difficulty, while in self-paced learning, a pacing function defines the speed to adapt the training progress. While both methods heavily rely on the ability to score the difficulty of data samples, an optimal scoring function is still under exploration. **Methodology:** Distillation is a knowledge transfer approach where a teacher network guides a student network by feeding a sequence of random samples. We argue that guiding student networks with an efficient curriculum strategy can improve model generalization and robustness. For this purpose, we design an uncertainty-based paced curriculum learning in self-distillation for medical image segmentation. We fuse the prediction uncertainty and annotation boundary uncertainty to develop a novel paced-curriculum distillation (P-CD). We utilize the teacher model to obtain prediction uncertainty and spatially varying label smoothing with Gaussian kernel to generate segmentation boundary uncertainty from the annotation. We also investigate the robustness of our method by applying various types and severity of image perturbation and corruption.

∗Corresponding to Hongliang Ren hlren@ieee.org.

† Equal contribution

[1]Dept. of Computing, Imperial College London, United Kingdom.

[2]Dept. of Biomedical Engineering, National University of Singapore, Singapore.

[3]Dept. of Electronics and Communication Engineering, National Institute of Technology Tiruchirappalli, India.

[4]Dept. of Instrumentation and Control Engineering, National Institute of Technology Tiruchirappalli, India.

[5]Dept. of Mechanical Engineering, Indian Institute of Technology, Delhi, India

[6]Dept. of Electrical Engineering and Shun Hing Institute of Advanced Engineering, The Chinese University of Hong Kong.

Results: The proposed technique is validated on two medical datasets of breast ultrasound image segmentation and robot-assisted surgical scene segmentation and achieved significantly better performance in terms of segmentation and robustness. **Conclusion:** P-CD improves the performance and obtains better generalization and robustness over the dataset shift. While curriculum learning requires extensive tuning of hyper-parameters for pacing function, the level of performance improvement suppresses this limitation.

Keywords: Curriculum Learning, Boundary Uncertainty, Distillation, Segmentation

1 Introduction

Inspired by the cognitive process of humans and animals, numerous curriculum learning (CL) [1, 2] and self-paced learning (SPL) [3, 4] techniques are introduced in machine learning. While both the CL and SPL techniques heavily rely on difficulty measurement techniques and ranking among the samples, an ideal ranking method is rarely available. Several methods attempt to design difficulty measurement techniques using confidence score [5], multi-raters agreement [6], and feature smoothing [2] in classification tasks. However, most of these works focus on classification tasks where it is easy to drop the harder images (samples). In segmentation or pixel-wise classification, it is more complex to build a curriculum strategy where spatial dependency is crucial among the local structures. Specifically in medical imaging, object boundary label is always ambiguous, even by human experts. In this work, we design a novel curriculum strategy using prediction and annotation boundary uncertainty in self-distillation for medical image segmentation.

Most recently, CL and SPL techniques have proven superior with distillation in vision and language classification tasks. In terms of scoring techniques, the confidence score uses to rank the input examples [5], the mirror decent technique uses to obtain sequential optimization [7], fixed lineups to increase the difficulty [8], Sentence length and coherence between dialog pairs use to determine sample complexity [9] and less imbalance subset uses for self-paced expert selection and self-paced instance selection [10] to build curriculum or self-paced knowledge distillation. However, most of these works focus on classification tasks where measuring sample difficulty without considering spatial correlation and local structure should not be compatible with semantic segmentation. Unlike distillation, both uncertainty and confidence-based sample ranking techniques are found to be the most common scoring method to design curriculum learning due to simplicity. For example, in the applications of classification [11, 12], object localization [3], neural machine translation [13, 14] and reinforcement learning [15]. Nonetheless, the effectiveness of pixel-wise uncertainty-based curriculum strategy is less clear in semantic segmentation.

The underlying spatial dependencies among the semantic object limit the correctness of pixel-wise difficulty measurement using predicted uncertainty independently. In addition, there is obvious uncertainty about anatomical structure in medical imaging which can be considered as difficult pixels regardless of predicted uncertainty (PU) [16–18].

In this work, we design paced-curriculum distillation (P-CD) by fusing prediction uncertainty (PU) and annotation boundary uncertainty (BU) to determine pixel-wise difficulty scores for self-paced curriculum strategy. We use a teacher network to obtain the prediction uncertainty and spatially varying label smoothing [17] with Gaussian kernel to generate segmentation boundary uncertainty from the annotation. In P-CD, curriculum strategy allows guiding the student network to learn the easier pixels at the beginning of training and gradually increases the difficulty level and self-paced learning controls the speed of student network learning by feeding a certain ratio of samples with epochs. We validate the proposed method with two medical image segmentation datasets of robot-assisted surgical scene segmentation (Robotic Surgery) [19] and breast ultrasound tumor segmentation (BUS) [20]. We conduct extensive robustness experiments by corrupting the datasets with various perturbation techniques in five different severity levels. Our contributions can be summarized as follows:

– Introduce a novel paced-curriculum distillation (P-CD) that fuses prediction uncertainty and annotation boundary uncertainty to determine pixel-wise difficulty score, including structural dependency.
– P-CD validates on two medical image segmentation datasets and the results suggest the superiority of the model over the baselines.
– P-CD also demonstrates robust performance on the progressive perturbations with different severity levels.

2 Proposed Method

2.1 Preliminaries

2.1.1 Distillation

In distillation, besides training the student model against the true label, it is also trained to mimic the teacher model's prediction probabilities [21]. This is achieved by training the model on both the cross-entropy loss ($loss_{CE}$) against the true label and the KL divergence loss ($loss_{KL}$) between the student and teacher logits [22]. Training on $loss_{KL}$ enforces the student model to mimic similar feature representations to the teacher model. The self-distillation (SD) loss ($loss_{SD}$) is theorized as: $loss_{SD} = loss_{CE} + loss_{KL}$. If the image size is [HxW] then $loss_{SD}$ will be a matrix of the same size ($loss_{SD} \in \mathbb{R}^{H \times W}$) for pixel-wise loss before averaging in segmentation task.
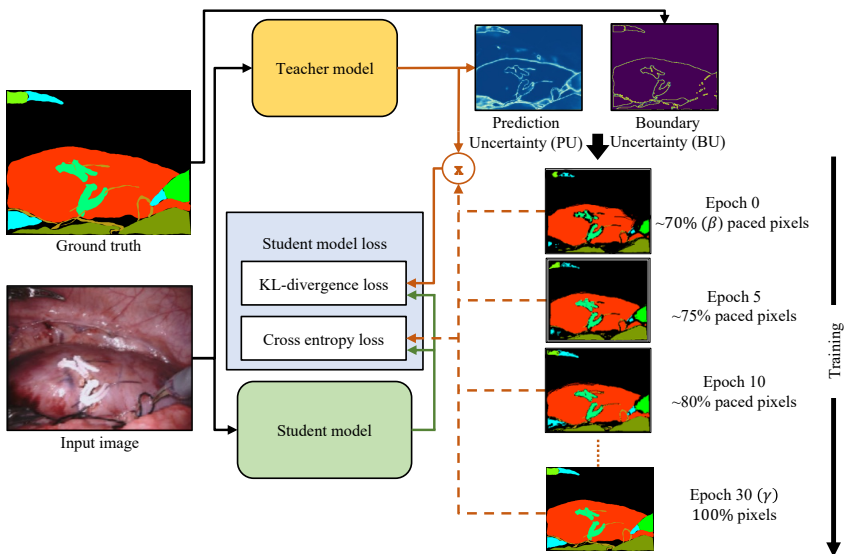
2.1.2 Uncertainty

Prediction uncertainty can interpret the confidence level of a model to recognize a sample. Alternatively, an easier sample produces lower uncertainty or higher confidence during the prediction. Therefore, prediction uncertainty (PU) can be easily determined as:

$$PU = (1 - confidence) \tag{1}$$

To generate boundary uncertainty (BU) from label, spatially varying label smoothing (SVLS) [17] design a weight matrix, with the 3×3 Gaussian kernel of the standard deviation, $\sigma = 1$ and modify the kernel to make sure center is equal to the sum of the neighbouring points and convolves across the OneHot label (OHLabel).

$$BU = 1 - SVLS^{weight}(\sigma) \circledast OHLabel \tag{2}$$

In this work, we utilize the SVLS [17] to obtain BU and fuse it with PU to design a novel curriculum scheme in the distillation process.

Fig. 1 Paced-curriculum distillation (P-CD), an uncertainty-based paced-curriculum learning in self-distillation for segmentation. Given the input image and label, the input is first propagated through the teacher model. Then calculate the pixel-wise difficulty score with the fusion of the predicted uncertainty from the teacher and label boundary uncertainty. Overall, curriculum learning (CL) guides the student to learn easier to harder pixels during training and self-paced learning controls the speed of the student network learning by feeding a certain ratio of pixels with epochs.

2.2 Paced-Curriculum Distillation (P-CD)

Curriculum learning, in most cases, is implemented on classification tasks as it is easier to discard tough images (samples). However, choosing which images to drop gets complicated for segmentation tasks as some pixels could be easier to learn than others in every image. One way to address this is to find the average difficulty of the pixels in an image that could be used as a parameter to quantify the difficulty of an image. However, this leads to a new problem as most images consist of pixels of varying difficulty. Eventually, this would still result in model learning both easy and hard pixels simultaneously, resulting in sub-optimal model convergence. To address this, we reformulate segmentation tasks as pixel-wise classification with images depicting mini-batches of pixels.

We design Paced-Curriculum Distillation (P-CD) by fusing prediction uncertainty from the teacher networks and boundary uncertainty from labels to determine pixel-wise difficulty for curriculum strategy. During distillation, we control the loss calculation by allowing easier pixels to learn at the beginning of the training and gradually including harder pixels over the training, as shown in Fig 1. We also design a pacing function to guide the suitable learning speed by determining the pixel ratio to calculate loss over training epochs. Overall procedures can be formulated as below.

2.2.1 Confidence Calibration

As confidence score is one of the major criteria in our curriculum design, we calibrate the confidence score on the teacher model to obtain accurate uncertainty. We apply, temperature scaling [23], a commonly used calibration technique to calibrate the model. In temperature scaling, an optimal temperature T is used to scale the logits z as (z/T) to obtain calibrated confidence.

2.2.2 Curriculum Objective Function

The effect of our curriculum directly influences the objective function of the student network during distillation. We drop the pixels with low confidence or high uncertainty in loss calculation at the initial stage of the training. For this purpose, we determine pixel-wise uncertainty weight matrices for both prediction uncertainty (W_{PU}) from calibrated teacher model (as equation 1) and boundary uncertainty (W_{BU}) from annotation label (as equation 2) by applying a threshold μ. This initial value of the threshold and increment criteria depend on the initial ratio of the pixels and end epoch of the curriculum learning (more details are in section 2.2.3). If μ is the uncertainty threshold (which determine by the pacing function), then binary weights matrices W_{PU} and W_{BU} can be calculated as-

$$W_{PU} \text{ or } W_{BU} = \begin{cases} 0, & \text{if } W_{PU} \text{ or } W_{BU} >= \mu \\ 1, & \text{otherwise} \end{cases} \tag{3}$$

If the pixel-wise vanilla distillation loss matrix $loss_{SD} \in \mathbb{R}^{H \times W}$, prediction and boundary uncertainty weight matrices $W_{PU} \in \mathbb{R}^{H \times W}$ and $W_{BU} \in \mathbb{R}^{H \times W}$,

then our paced-curriculum distillation loss matrix (before averaging) will be-

$$loss_{PCD} = W_{BU} \odot W_{PU} \odot loss_{KD} \tag{4}$$

2.2.3 Pacing Function

We design the pacing function based on two important criteria such as the initial ratio of data, β, and the end epoch of curriculum training, γ. Based on β, we can determine the initial uncertainty threshold μ_{init} from the teacher network. Then, threshold update μ_{update} can be obtained from the following equation-

$$\mu_{update} = \frac{1 - \mu_{init}}{\gamma/E_{interval}} \tag{5}$$

$$\mu = \mu_{init} + \mu_{update} \tag{6}$$

$$\mu = min(\mu, 1) \tag{7}$$

Where $E_{interval}$ is the epoch interval, we want to update the uncertainty to include harder pixels over the training period. We use $E_{interval} = 5$ by following [2]. When the uncertainty threshold, $\mu = 1$, then the student model will consider all the pixels to calculate loss during training.

3 Experiments

3.1 Datasets

3.1.1 Robotic Surgery

The MICCAI18 instrument segmentation challenge [19] dataset consists of 15 video sequences for training and 4 video sequences for testing. Each sequence consists of 149 frames which are resized to 224 x 224 in this work. For pixel-wise segmentation, the dataset consists of 12 classes of Kidney Parenchyma, Covered Kidney, Instrument Shaft, Instrument Clasper, Instrument Wrist, Thread, Suturing Needle, US Probe, Small Intestine, Suction Instrument, Clamps, and Background.

3.1.2 BUS

Breast ultrasound tumor segmentation (BUS) [20] is a publicly available dataset[1] that consists of 780 breast ultrasound images and corresponding tumor labels. All these images and labels are resized to 224×224 and split to 80/20 ratio for the training and validation purposes of this work.

---

[1]https://www.kaggle.com/datasets/aryashah2k/breast-ultrasound-images-dataset

3.2 Implementation Details

We first train some of the independent and identically distributed (IID) models like vanilla UNet [24], LinkNet [25], DeepLabv3+ [26] as teacher models for rest of the experiment. We then train the SD and proposed the P-CD method. As accurate uncertainty or confidence is crucial to determine pixel difficulty, we calibrate each teacher model using the commonly used technique temperature scaling [23]. Overall, the models are trained on Adam optimizer with a batch size of 16, and the number of epochs is taken to be 150.

4 Results

Table 1 Performance comparison of proposed P-CD over the model train on vanilla SD and i.i.d. for three state-of-the-art architectures of UNet, DeepLabV3+, and LinkNet.

Methods		Robotic surgery			BUS		
	Metrics	i.i.d.	SD	P-CD	i.i.d.	SD	P-CD
UNet [24]	DSC	0.6251	0.6445	**0.6644**	0.7620	0.7676	**0.7897**
	IoU	0.5827	0.6010	**0.6237**	0.6921	0.6919	**0.7191**
	Precision	0.6890	0.7212	**0.7535**	0.8696	**0.9092**	0.9042
DeepLabV3+ [26]	DSC	0.6656	0.6858	**0.6929**	0.7382	0.7611	**0.7722**
	IoU	0.6237	0.6437	**0.6503**	0.6556	0.6762	**0.6860**
	Precision	0.7511	**0.7829**	0.7803	0.8389	**0.8557**	0.8328
LinkNet [25]	DSC	0.6442	0.6614	**0.6715**	0.7261	0.7753	**0.7777**
	IoU	0.6071	0.6236	**0.6332**	0.6547	0.6962	**0.7045**
	Precision	0.7486	0.7419	**0.7707**	0.8448	0.6962	**0.8703**

The proposed method, P-CD, has been extensively validated on three state-of-the-art architectures of UNet [24], DeepLabV3+ [26], and LinkNet [25] and two medical image segmentation datasets and three commonly used evaluation metrics of e dice similarity coefficient (DSC), mean intersection over union (IoU), and precision. The results are presented in the Table 1 and Fig. 2.

From Table. 1, it is observed that the student model trained on the proposed P-CD outperforms both independent and identically distributed (i.i.d.) training model and vanilla SD version. Our method obtains DSC improvement of around 2-4% over i.i.d. for all the architectures on both datasets. Our proposed curriculum strategy significantly improves all three model performances across both datasets. weUnpaired one-tailed T-test is also conducted to check the significance of improvement for (i) SD over i.i.d, (ii) P-CD over SD, and (iii) P-CD over i.i.d on both datasets. Based on p-values (Table 2), an improvement from employing P-CD over i.i.d is found to be very significant ($p < 0.05$)

Table 2 Significant test (p-value) on DSC and IoU improvement when trained using (i) i.i.d vs SD, (ii) P-CD vs SD, (iii) and P-CD vs i.i.d, on both the BUS and Robotic surgery datasets.

| Dataset	DSC			IoU		
	i.i.d vs SD	SD vs P-CD	i.i.d vs P-CD	i.i.d vs SD	SD vs P-CD	i.i.d vs P-CD
BUS	**0.042**	**0.073**	**0.016**	0.104	0.127	**0.042**
Robotic surgery	0.161	0.224	**0.048**	0.173	0.212	**0.046**
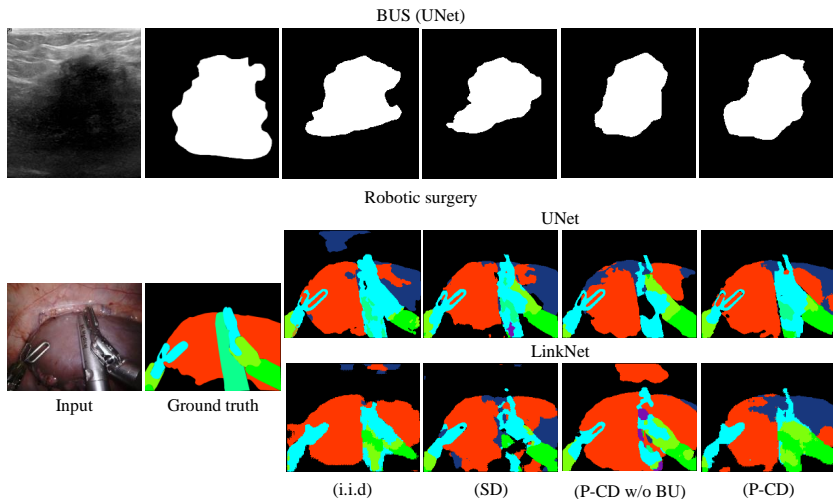
Fig. 2 Qualitative visualization of our P-CD or P-CD w/o BU against the ground truth (GT), teacher model, and student model trained using self-distillation (SD). The colors red, dark blue, light blue, and green mean kidney-parenchyma, covered-kidney, instrument-clasper, and wrist, respectively, in the robotic surgery dataset. In the BUS dataset, breast tumors are annotated as white color.

on both datasets in terms of DSC and IoU. Improvement from employing P-CD over SD is also found to be significant (p <0.1) on the BUS dataset in terms of DSC. Similar prediction trends are also observed in qualitative visualization in Fig 2. A slight enhancement in boundary prediction is noticed in the visualization. The performance is also compared with the top-3 participants in the challenge [19] and our method outperforms with a remarkable margin (Fig. 3(a)). Interestingly, Top1 performing method of the challenge is also utilized DeepLabV3+, which yields 2-3% higher mIoU by integrating our proposed P-CD. It is clear evidence that our technique is flexible and can be integrated with any state-of-the-art network to improve performance.
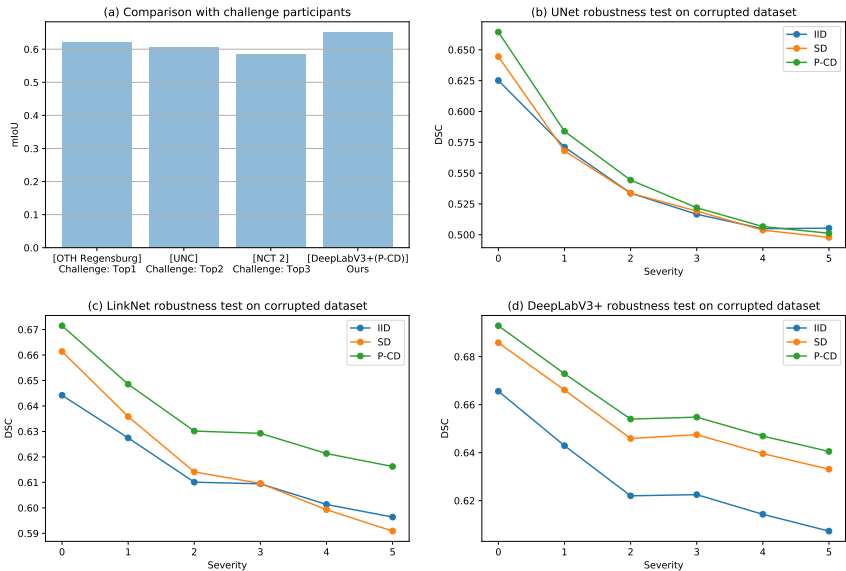
Fig. 3 Comparison and robustness studies: (a) Comparison with the challenge, (b) Robustness test on the Robotic surgery dataset with UNet model, (c) Robustness test on the Robotic surgery dataset with LinkNet model, (d) Robustness test on the Robotic surgery dataset with DeepLabv3+ model.
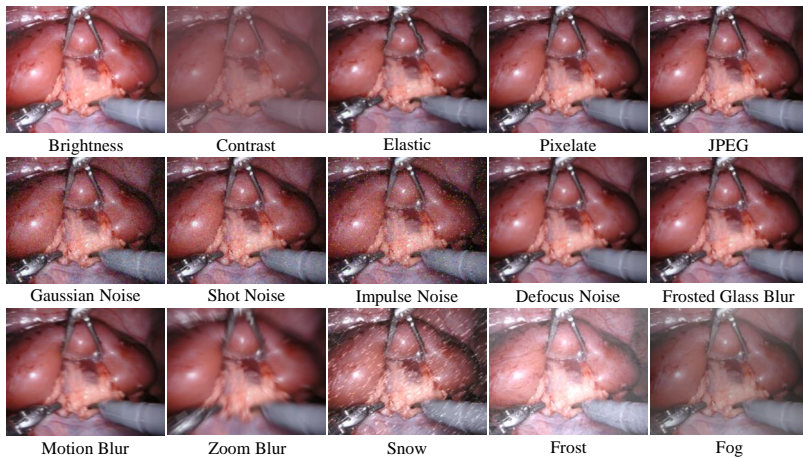
Fig. 4 A surgical scene perturbed in 15 types with severity level set to 3 to test the model robustness.

4.1 Robustness Test

Models trained on curriculum are argued to be robust to noises [27, 28]. The robustness of our P-CD trained student models to common perturbations is quantified against models trained using i.i.d, KD technique is quantified by introducing corruptions and perturbations of varied severity by following [29].

There are 15 different perturbation techniques with 5 severity levels. For example, noises such as Gaussian, shot, and impulse; blurring techniques such as defocus, glass, motion, zoom; weather corruptions such as snow, frost, and fog; and digital corruptions such as brightness, contrast, elastic, pixel, and JPEG. The severity levels are controlled by increasing the perturbation scale for each technique. For example, to add Gaussian noise with severity levels 1 to 5, the standard deviation of the Gaussian noise increase as $[0.04, 0.06, 0.08, 0.09, 0.10]$, respectively. The implementation of the perturbation techniques and severity levels are adopted from the reference work [29] and official repository [2]. When segmenting a surgical field, the model must be robust to bleeding/tissue occlusion and occlusion from smoke. When the instrument or tissue is occluded/partially occluded by other instruments or bleeding or dead tissue or saline, the presence of other instrument features, translucent fluid, or dead tissues could significantly change the feature representation of the boundaries and result in boundary uncertainty. As the robotic surgery dataset lacks the annotation (corruption/severity of corruption) to quantify the model's robustness against them, we relate robustness against (a) weather {fog} to smoke occlusion, (b) noise {gaussian} and blurring {defocus, glass, motion, zoom} to saline release, and (C) digital {pixel} and noise {gaussian} to instrument occlusion.

4.2 Ablation Study

4.2.1 W and WO Boundary Uncertainty

Table 1 presents the significant performance improvement with our P-CD. We also investigate the impact of model uncertainty and boundary uncertainty separately in Fig. 5(a) & (b). Although there is a clear prediction enhancement on curriculum learning with only model uncertainty, boundary uncertainty provides an additional benefit by difficulty measurement in curriculum strategy. This observation is further validated in Table 3, where, both LinkNet [25] and DeepLabV3+ [26] trained using P-CD outperforms LinkNet [25] and DeepLabV3+ [26] trained using P-CD w/o BU.

4.2.2 β vs γ Study

Initial β and γ values determine the initial μ value and dictate its increment during the training regime. As μ is used to pace the model learning, the effect on the model's performance from β and γ is significant. We study and report the performance of the student models based on various β vs γ combinations on the BUS and the Robotic surgery datasets (Fig. 6(a) & (b)). While high performance is observed with $beta = 90\%$ on the Robotic surgery dataset, it is not universal, as high performance on the BUS dataset is observed with $beta = 50\%$. Therefore, the best β vs γ could change for each dataset.

---

[2]https://github.com/hendrycks/robustness

Table 3 Performance comparison of models trained using P-CD w & w/o BU for Robotic surgery dataset.

Model			Dice	IoU	Precision
LinkNet [25]	P-CD w/o BU		0.6432	0.6058	0.7161
	P-CD		**0.6715**	**0.6332**	**0.7707**
DeepLabV3+ [26]	P-CD w/o BU		0.6822	0.6428	**0.7902**
	P-CD		**0.6929**	**0.6503**	0.7803
UNet [24]	P-CD w/o BU		0.6594	0.6187	**0.7545**
	P-CD		**0.6644**	**0.6237**	0.7535
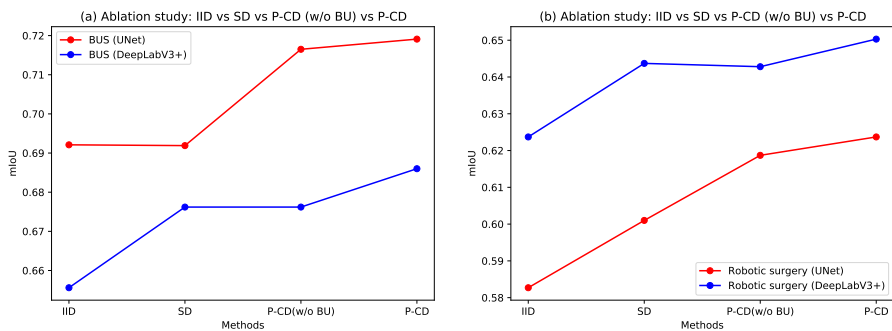
Fig. 5 Ablation studies: Performance improvements of UNet and DeepLabV3+ on (a) the BUS dataset and (b) the Robotic surgery dataset.
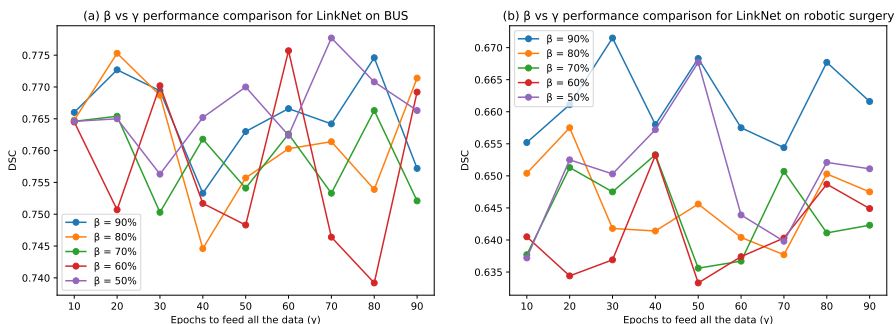
Fig. 6 β vs γ Study: β and γ tuning for LinkNet model on (a) the BUS dataset and (b) the Robotic surgery dataset.

5 Discussion and Conclusion

We present paced-curriculum distillation (P-CD), a model-agnostic curriculum-based distillation technique for semantic segmentation in medical imaging. We design a curriculum strategy by combining predicted uncertainty

from a teacher model and label boundary uncertainty to measure pixel difficulty levels. Incorporating boundary uncertainty with prediction uncertainty firms the complex boundary pixels might assign wrong confidence in the less robust teacher model. We empirically show the superior performance of our method with two medical image segmentation datasets and extensive ablation studies. We also investigate the robustness of our method by corrupting validation data with 15 various types of image corruption and perturbation techniques. The results suggest that P-CD improves the performance and obtains better generalization and robustness over the dataset shift. In addition to overall performance improvement, the results from Table 1 and Fig. 2 demonstrate that P-CD is also able to predict better object boundaries, which is very crucial in clinical applications. We observe that curriculum learning requires accurate hyper-parameters for pacing function, which leads to extensive tuning sometimes. However, the level of performance improvement suppresses this limitation. As future work, multi-raters disagreement [6] can be integrated with P-CD to obtain a more accurate pixel difficulty score for robust curriculum distillation.

Declarations

5.1 Code availability

This work is implemented using the PyTorch framework and the codes are available at: github.com/mobarakol/P-CD.

5.2 Funding

5.3 Ethics approval

All procedures performed in studies involving human participants were in accordance with the ethical standards of the institutional and/or national research committee.

5.4 Consent to participate

This article does not contain patient data.

5.5 Conflict of interest

The authors declare that they have no conflict of interest.

References

[1] Bengio, Y., Louradour, J., Collobert, R., Weston, J.: Curriculum learning. In: Proceedings of the 26th Annual International Conference on Machine Learning, pp. 41–48 (2009)

[2] Sinha, S., Garg, A., Larochelle, H.: Curriculum by smoothing. Advances in Neural Information Processing Systems **33**, 21653–21664 (2020)

[3] Kumar, M., Packer, B., Koller, D.: Self-paced learning for latent variable models. Advances in neural information processing systems **23**, 1189–1197 (2010)

[4] Jiang, L., Meng, D., Zhao, Q., Shan, S., Hauptmann, A.G.: Self-paced curriculum learning. In: Twenty-Ninth AAAI Conference on Artificial Intelligence (2015)

[5] Zhao, H., Sun, X., Dong, J., Dong, Z., Li, Q.: Knowledge distillation via instance-level sequence learning. Knowledge-Based Systems **233**, 107519 (2021)

[6] Wei, J., Suriawinata, A., Ren, B., Liu, X., Lisovsky, M., Vaickus, L., Brown, C., Baker, M., Nasir-Moin, M., Tomita, N., Torresani, L., Wei, J., Hassanpour, S.: Learn like a pathologist: Curriculum learning by annotator agreement for histopathology image classification. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), pp. 2473–2483 (2021)

[7] Shi, W., Song, Y., Zhou, H., Li, B., Li, L.: Follow your path: a progressive method for knowledge distillation. In: Joint European Conference on Machine Learning and Knowledge Discovery in Databases, pp. 596–611 (2021). Springer

[8] Ye, D., Chen, G., Zhang, W., Chen, S., Yuan, B., Liu, B., Chen, J., Liu, Z., Qiu, F., Yu, H., *et al.*: Towards playing full moba games with deep reinforcement learning. Advances in Neural Information Processing Systems **33**, 621–632 (2020)

[9] Zhu, Q., Chen, X., Wu, P., Liu, J., Zhao, D.: Combining curriculum learning and knowledge distillation for dialogue generation. In: Findings of the Association for Computational Linguistics: EMNLP 2021, pp. 1284–1295 (2021)

[10] Xiang, L., Ding, G., Han, J.: Learning from multiple experts: Self-paced knowledge distillation for long-tailed classification. In: European Conference on Computer Vision, pp. 247–263 (2020). Springer

[11] Wang, Y., Gan, W., Yang, J., Wu, W., Yan, J.: Dynamic curriculum learning for imbalanced data classification. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 5017–5026 (2019)

[12] Chang, H.-S., Learned-Miller, E., McCallum, A.: Active bias: Training more accurate neural networks by emphasizing high variance samples. Advances in Neural Information Processing Systems **30** (2017)

[13] Zhou, Y., Yang, B., Wong, D.F., Wan, Y., Chao, L.S.: Uncertainty-aware curriculum learning for neural machine translation. In: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, pp. 6934–6944 (2020)

[14] Platanios, E.A., Stretcu, O., Neubig, G., Poczos, B., Mitchell, T.M.: Competence-based curriculum learning for neural machine translation. In: International Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, pp. 1162–1172 (2019)

[15] Zhang, Y., Abbeel, P., Pinto, L.: Automatic curriculum learning through value disagreement. Advances in Neural Information Processing Systems **33**, 7648–7659 (2020)

[16] Lee, H.J., Kim, J.U., Lee, S., Kim, H.G., Ro, Y.M.: Structure boundary preserving segmentation for medical image with ambiguous boundary. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 4817–4826 (2020)

[17] Islam, M., Glocker, B.: Spatially varying label smoothing: Capturing uncertainty from expert annotations. In: International Conference on Information Processing in Medical Imaging, pp. 677–688 (2021). Springer

[18] Tang, P., Yang, P., Nie, D., Wu, X., Zhou, J., Wang, Y.: Unified medical image segmentation by learning from uncertainty in an end-to-end manner. Knowledge-Based Systems **241**, 108215 (2022)

[19] Allan, M., Kondo, S., Bodenstedt, S., Leger, S., Kadkhodamohammadi, R., Luengo, I., Fuentes, F., Flouty, E., Mohammed, A.K., Pedersen, M., Kori, A.K., Varghese, A., Krishnamurthi, G., Rauber, D., Mendel, R., Palm, C., Bano, S., Saibro, G., Shih, C., Chiang, H., Zhuang, J., Yang, J., Iglovikov, V., Dobrenkii, A., Reddiboina, M., Reddy, A., Liu, X., Gao, C., Unberath, M., Azizian, M., Stoyanov, D., Maier-Hein, L., Speidel, S.: 2018 robotic scene segmentation challenge. arXiv preprint arXiv:2001.11190 (2020)

[20] Al-Dhabyani, W., Gomaa, M., Khaled, H., Fahmy, A.: Dataset of breast ultrasound images. Data in brief **28**, 104863 (2020)

[21] Hinton, G., Vinyals, O., Dean, J.: Distilling the knowledge in a neural network. arXiv preprint arXiv:1503.02531 **2**(7) (2015)

[22] Kim, K., Ji, B., Yoon, D., Hwang, S.: Self-knowledge distillation with progressive refinement of targets. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 6567–6576 (2021)

[23] Guo, C., Pleiss, G., Sun, Y., Weinberger, K.Q.: On calibration of modern neural networks. In: International Conference on Machine Learning, pp. 1321–1330 (2017). PMLR

[24] Ronneberger, O., Fischer, P., Brox, T.: U-net: Convolutional networks for biomedical image segmentation. In: International Conference on Medical Image Computing and Computer-assisted Intervention, pp. 234–241 (2015). Springer

[25] Chaurasia, A., Culurciello, E.: Linknet: Exploiting encoder representations for efficient semantic segmentation. In: 2017 IEEE Visual Communications and Image Processing (VCIP), pp. 1–4 (2017). IEEE

[26] Chen, L.-C., Zhu, Y., Papandreou, G., Schroff, F., Adam, H.: Encoder-decoder with atrous separable convolution for semantic image segmentation. In: Proceedings of the European Conference on Computer Vision (ECCV), pp. 801–818 (2018)

[27] Gong, T., Zhao, Q., Meng, D., Xu, Z.: Why curriculum learning & self-paced learning work in big/noisy data: A theoretical perspective. Big Data & Information Analytics **1**(1), 111 (2016)

[28] Wu, X., Dyer, E., Neyshabur, B.: When do curricula work? arXiv preprint arXiv:2012.03107 (2020)

[29] Hendrycks, D., Dietterich, T.: Benchmarking neural network robustness to common corruptions and perturbations. arXiv preprint arXiv:1903.12261 (2019)