# Source-Free Domain Adaptive Fundus Image Segmentation with Class-Balanced Mean Teacher

Longxiang Tang[1], Kai Li[2(✉)], Chunming He[1], Yulun Zhang[3], and Xiu Li[1(✉)]

[1] Tsinghua Shenzhen International Graduate School,
Tsinghua University, Shenzhen, China
{lloong.x,chunminghe19990224}@gmail.com, li.xiu@sz.tsinghua.edu.cn
[2] NEC Laboratories America, Princeton, USA
li.gml.kai@gmail.com
[3] ETH Zurich, Zürich, Switzerland
yulun100@gmail.com

**Abstract.** This paper studies source-free domain adaptive fundus image segmentation which aims to adapt a pretrained fundus segmentation model to a target domain using unlabeled images. This is a challenging task because it is highly risky to adapt a model only using unlabeled data. Most existing methods tackle this task mainly by designing techniques to carefully generate pseudo labels from the model's predictions and use the pseudo labels to train the model. While often obtaining positive adaption effects, these methods suffer from two major issues. First, they tend to be fairly unstable - incorrect pseudo labels abruptly emerged may cause a catastrophic impact on the model. Second, they fail to consider the severe class imbalance of fundus images where the foreground (e.g., cup) region is usually very small. This paper aims to address these two issues by proposing the Class-Balanced Mean Teacher (CBMT) model. CBMT addresses the unstable issue by proposing a weak-strong augmented mean teacher learning scheme where only the teacher model generates pseudo labels from weakly augmented images to train a student model that takes strongly augmented images as input. The teacher is updated as the moving average of the instantly trained student, which could be noisy. This prevents the teacher model from being abruptly impacted by incorrect pseudo-labels. For the class imbalance issue, CBMT proposes a novel loss calibration approach to highlight foreground classes according to global statistics. Experiments show that CBMT well addresses these two issues and outperforms existing methods on multiple benchmarks.

**Keywords:** Source-free domain adaptation · Fundus image · Mean teacher

# 1    Introduction

Medical image segmentation plays an essential role in computer-aided diagnosis systems in different applications and has been tremendously advanced in the past few years [6,12,19,22]. While the segmentation model [3,10,11,21] always requires sufficient labeled data, unsupervised domain adaptation (UDA) approaches have been proposed, learning an adaptive model jointly with unlabeled target domain images and labeled source domain images [9], for example, the adversarial training paradigm [2,8,13,14,16].

Although impressive performance has been achieved, these UDA methods may be limited for some real-world medical image segmentation tasks where labeled source images are not available for adaptation. This is not a rare scenario because medical images are usually highly sensitive in privacy and copyright protection such that labeled source images may not be allowed to be distributed. This motivates the investigation of source-free domain adaptation (SFDA) where adapts a source segmentation model trained on labeled source data (in a private-protected way) to the target domain only using unlabeled data.

A few recent SFDA works have been proposed. OSUDA [17] utilizes the domain-specific low-order batch statistics and domain-shareable high-order batch statistics, trying to adapt the former and keep the consistency of the latter. SRDA [1] minimizes a label-free entropy loss guided with a domain-invariant class-ratio prior. DPL [4] introduces pixel-level and class-level pseudo-label denoising schemes to reduce noisy pseudo-labels and select reliable ones. U-D4R [27] applies an adaptive class-dependent threshold with the uncertainty-rectified correction to realize better denoising.

Although these methods have achieved some success in model adaptation, they still suffer from two major issues. <u>First</u>, they tend to be fairly unstable. Without any supervision signal from labeled data, the model heavily relies on the predictions generated by itself, which are always noisy and could easily make the training process unstable, causing catastrophic error accumulation after several training epochs as shown in Fig. 1(a). Some works avoid this problem by only training the model for very limited iterations (only 2 epochs in [4,27]) and selecting the best-performing model during the whole training process for testing. However, this does not fully utilize the data and it is non-trivial to select the best-performing model for this unsupervised learning task. <u>Second</u>, they failed to consider the severe foreground and background imbalance of fundus images where the foreground (e.g., cup) region is usually very small (as shown in Fig. 1(b)). This oversight could also lead to a model degradation due to the dominate background learning signal.

In this paper, we propose the Class-Balanced Mean Teacher (CBMT) method to address the limitations of existing methods. To mitigate the negative impacts of incorrect pseudo labels, we propose a weak-strong augmented mean teacher learning scheme which involves a teacher model and a student model that are both initialized from the source model. We use the teacher to generate pseudo label from a weakly augmented image, and train the student that takes strongly augmented version of the same image as input. We do not train the teacher
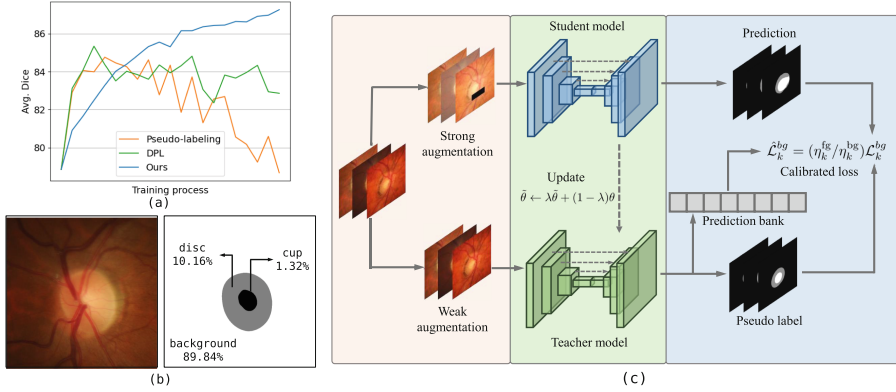
**Fig. 1.** (a) Training curve of vanilla pseudo-labeling, DPL [4] and our approach. (b) Fundus image and its label with class proportion from the *RIM-ONE-r3* dataset. (c) Illustrated framework of our proposed CBMT method.

model directly by back-propagation but update its weights as the moving average of the student model. This prevents the teacher model from being abruptly impacted by incorrect pseudo labels and meanwhile accumulates new knowledge learned by the student model. To address the imbalance between foreground and background, we propose to calibrate the segmentation loss and highlight the foreground class, based on the prediction statistics derived from the global information. We maintain a prediction bank to capture global information, which is considered more reliable than that inside one image.

Our contributions can be summarized as follows: (1) We propose the weak-strong augmented mean teacher learning scheme to address the stable issue of existing methods. (2) We propose the novel global knowledge-guided loss calibration technique to address the foreground and background imbalance problem. (3) Our proposed CBMT reaches state-of-the-art performance on two popular benchmarks for adaptive fundus image segmentation.

## 2   Method

Source-Free Domain Adaptive (SFDA) fundus image segmentation aims to adapt a source model $h$, trained with $N_S$ labeled source images $\mathcal{S} = \{(X_i, Y_i)\}_{i=1}^{N_S}$, to the target domain using only $N_T$ unlabeled target images $\mathcal{T} = \{X_i\}_{i=1}^{N_T}$. $Y_i \in \{0,1\}^{H \times W \times C}$ is the ground truth, and $H$, $W$, and $C$ denote the image height, width, and class number, respectively. A vanilla pseudo-labeling-based method generates pseudo labels $\hat{y} \in \mathbb{R}^C$ from the sigmoided model prediction $p = h(x)$ for each pixel $x \in X_i$ with source model $h$:

$$\hat{y}_k = \mathbb{1}\left[p_k > \gamma\right], \tag{1}$$

where $\mathbb{1}$ is the indicator function and $\gamma \in [0,1]$ is the probability threshold for transferring soft probability to hard label. $p_k$ and $y_k$ is the $k$-th dimension of $p$

and $y$, respectively, denoting the prediction and pseudo label for class $k$. Then $(x, \hat{y})$ is utilized to train the source model $h$ with binary cross entropy loss:

$$L_{bce} = \mathbb{E}_{x \sim X_i}[\hat{y} \log(p) + (1 - \hat{y}) \log(1 - p)] \tag{2}$$

Most existing SFDA works refine this vanilla method by proposing techniques to calibrate $p$ and get better pseudo label $\hat{y}$, or measure the uncertainty of $p$ and apply a weight when using $\hat{y}$ for computing the loss [4, 27]. While achieving improved performance, these methods still suffer from the unstable issue because noisy $\hat{y}$ will directly impact $h$, and the error will accumulate since then the predictions of $h$ will be used for pseudo labeling. Another problem with this method is that they neglect the imbalance of the foreground and background pixels in fungus images, where the foreground region is small. Consequently, the second term in Eq. (2) will dominate the loss, which is undesirable.

Our proposed CBMT model addresses the two problems by proposing the weak-strong augmented mean teacher learning scheme and the global knowledge-guided loss calibration technique. Figure 1(c) shows the framework of CBMT.

## 2.1 Weak-Strong Augmented Mean Teacher

To avoid error accumulation and achieve a robust training process, we introduce the weak-strong augmented mean teacher learning scheme where there is a teacher model $h_t$ and a student model $h_s$ both initialized from the source model $h$. We generate pseudo labels with $h_t$ and use the pseudo labels to train $h_s$. To enhance generalization performance, we further introduce a weak-strong augmentation mechanism that feeds weakly and strongly augmented images to the teacher model and the student model, respectively.

Concretely, for each image $X_i$, we generate a weakly-augmented version $X_i^w$ by using image flipping and resizing. Meanwhile, we generate a strongly-augmented version $X_i^s$. The strong augmentations we used include a random eraser, contrast adjustment, and impulse noises. For each pixel $x^w \in X_i^w$, we generate pseudo label $\hat{y}^w = h_t(x)$ by the teacher model $h_t$ with Eq. (1). Then, we train the student model $h_s$ with

$$\mathcal{L} = \mathbb{E}_{x^s \sim X_i^s, \hat{y}^w}[\tilde{\mathcal{L}}_{bce}], \tag{3}$$

where $\tilde{\mathcal{L}}_{bce}$ is the refined binary cross entropy loss which we will introduce later. It is based on Eq. (2) but addresses the fore- and back-ground imbalance problem.

The weakly-strong augmentation mechanism has two main benefits. First, since fundus image datasets are always on a small scale, the model could easily get overfitted due to the insufficient training data issue. To alleviate it, we enhance the diversity of the training set by introducing image augmentation techniques. Second, learning with different random augmentations performs as a consistency regularizer constraining images with similar semantics to the same class, which forms a more distinguishable feature representation.

We update the student model by back-propagating the loss defined in Eq. (3). But for the teacher model, we update it as the exponential moving average (EMA) of the student model as,

$$\tilde{\theta} \leftarrow \lambda\tilde{\theta} + (1 - \lambda)\theta, \tag{4}$$

where $\tilde{\theta}$, $\theta$ are the teacher and student model weights separately. Instead of updating the model with gradient directly, we define the teacher model as the exponential moving average of students, which makes the teacher model more consistent along the adaptation process. With this, we could train a model for a relatively long process and safely choose the final model without accuracy validation. From another perspective, the teacher model can be interpreted as a temporal ensemble of students in different time steps [18], which enhances the robustness of the teacher model.

## 2.2   Global Knowledge Guided Loss Calibration

For a fundas image, the foreground object (e.g., cup) is usually quite small and most pixel will the background. If we update the student model with Eq. (2), the background class will dominate the loss, which dilutes the supervision signals for the foreground class. The proposed global knowledge guided loss calibration technique aims to address this problem.

A naive way to address the foreground and background imbalance is to calculate the numbers of pixels falling into the two categories, respectively, within each individual image and devise a loss weighting function based on the numbers. This strategy may work well for the standard supervised learning tasks, where the labels are reliable. But with pseudo labels, it is too risky to conduct the statistical analysis based on a single image. To remedy this, we analyze the class imbalance across the whole dataset, and use this global knowledge to calibrate our loss for each individual image.

Specifically, we store the predictions of pixels from all images and maintain the mean loss for foreground and background as,

$$\eta_k^{\text{fg}} = \frac{\sum_i \mathcal{L}_{i,k} \cdot \mathbb{1}[\hat{y}_{i,k} = 1]}{\sum_i \mathbb{1}[\hat{y}_{i,k} = 1]}; \quad \eta_k^{\text{bg}} = \frac{\sum_i \mathcal{L}_{i,k} \cdot \mathbb{1}[\hat{y}_{i,k} = 0]}{\sum_i \mathbb{1}[\hat{y}_{i,k} = 0]} \tag{5}$$

where $\mathcal{L}$ is the segmentation loss mentioned above, and "fg" and "bg" represent foreground/background. The reason we use the mean of the loss, rather than the number of pixels, is that the loss of each pixel indicates the "hardness" of each pixel according to the pseudo ground truth. This gives more weight to those more informative pixels, thus more global knowledge is considered.

With each average loss, the corresponding learning scheme could be further calibrated. We utilize the ratio of $\eta_k^{\text{fg}}$ to $\eta_k^{\text{bg}}$ to weight background loss $\mathcal{L}_k^{bg}$:

$$\tilde{\mathcal{L}}_{bce} = \mathbb{E}_{x \sim X_i, k \sim C}[\hat{y}_k \log(p_k) + \eta_k^{\text{fg}}/\eta_k^{\text{bg}}(1 - \hat{y}_k) \log(1 - p_k)] \tag{6}$$

The calibrated loss ensures fair learning among different classes, therefore alleviating model degradation issues caused by class imbalance.

Since most predictions are usually highly confident (very close to 0 or 1), they are thus less informative. We need to only include pixels with relatively large loss scales to compute mean loss. We realize this by adopting constraint threshold $\alpha$ to select pixels: $\frac{|f(x_i)-\gamma|}{|\hat{y}_i-\gamma|} > \alpha$, where $\alpha$ is set by default to 0.2. $\alpha$ represents the lower bound threshold of normalized prediction, which can filter well-segmented uninformative pixels out.

## 3    Experiments

**Implementation Details**[1]. We apply the Deeplabv3+ [5] with MobileNetV2 [23] backbone as our segmentation model, following the previous works [4,26,27] for a fair comparison. For model optimization, we use Adam optimizer with 0.9 and 0.99 momentum coefficients. During the source model training stage, the initial learning rate is set to 1e-3 and decayed by 0.98 every epoch, and the training lasts 200 epochs. At the source-free domain adaptation stage, the teacher and student model are first initialized by the source model, and the EMA update scheme is applied between them for a total of 20 epochs with a learning rate of 5e-4. Loss calibration parameter $\eta$ is computed every epoch and implemented on the class cup. The output probability threshold $\gamma$ is set as 0.75 according to previous study [26] and model EMA update rate $\lambda$ is 0.98 by default. We implement our method with PyTorch on one NVIDIA 3090 GPU and set batch size as 8 when adaptation.

**Datasets and Metrics**. We evaluate our method on widely-used fundus optic disc and cup segmentation datasets from different clinical centers. Following previous works, We choose the REFUGE challenge training set [20] as the source domain and adapt the model to two target domains: RIM-ONE-r3 [7] and Drishti-GS [24] datasets for evaluation. Quantitatively, the source domain consists of 320/80 fundus images for training/testing with pixel-wise optic disc and cup segmentation annotation, while the target domains have 99/60 and 50/51 images. Same as [26], the fundus images are cropped to $512 \times 512$ as ROI regions.

We compare our CBMT model with several state-of-the-art domain adaptation methods, including UDA methods BEAL [26] and AdvEnt [25] and SFDA methods: SRDA [1], DAE [15] and DPL [4]. More comparisons with U-D4R [27] under other adaptation settings could be found in supplementary materials. General metrics for segmentation tasks are used for model performance evaluation, including the Dice coefficient and Average Symmetric Surface Distance (ASSD). The dice coefficient (the higher the better) gives pixel-level overlap results, and ASSD (the lower the better) indicates prediction boundary accuracy.

### 3.1    Experimental Results

The quantitative evaluation results are shown in Table 1. We include the without adaptation results from [4] as a lower bound, and the supervised learning results

---

[1] The code can be found in https://github.com/lloongx/SFDA-CBMT.

**Table 1.** Quantitative results of comparison with different methods on two datasets, and the best score for each column is highlighted. - means the results are not reported by that method, ± refers to the standard deviation across samples in the dataset. S-F means source-free.

| Methods | S-F | Optic Disc Segmentation | | Optic Cup Segmentation | |
|---|---|---|---|---|---|
| | | Dice[%] ↑ | ASSD[pixel] ↓ | Dice[%] ↑ | ASSD[pixel] ↓ |
| *RIM-ONE-r3* | | | | | |
| W/o DA [4] | | 83.18±6.46 | 24.15±15.58 | 74.51±16.40 | 14.44±11.27 |
| Oracle [26] | | 96.80 | – | 85.60 | – |
| BEAL [26] | × | 89.80 | – | 81.00 | – |
| AdvEnt [25] | × | 89.73±3.66 | 9.84±3.86 | 77.99±21.08 | **7.57±4.24** |
| SRDA [1] | ✓ | 89.37±2.70 | 9.91±2.45 | 77.61±13.58 | 10.15±5.75 |
| DAE [15] | ✓ | 89.08±3.32 | 11.63±6.84 | 79.01±12.82 | 10.31±8.45 |
| DPL [4] | ✓ | 90.13±3.06 | 9.43±3.46 | 79.78±11.05 | 9.01±5.59 |
| **CBMT(Ours)** | ✓ | **93.36±4.07** | **6.20±4.79** | **81.16±14.71** | 8.37±6.99 |
| *Drishti-GS* | | | | | |
| W/o DA [4] | | 93.84±2.91 | 9.05±7.50 | 83.36±11.95 | 11.39±6.30 |
| Oracle [26] | | 97.40 | – | 90.10 | – |
| BEAL [26] | × | 96.10 | – | **86.20** | – |
| AdvEnt [25] | × | 96.16±1.65 | 4.36±1.83 | 82.75±11.08 | 11.36±7.22 |
| SRDA [1] | ✓ | 96.22±1.30 | 4.88±3.47 | 80.67±11.78 | 13.12±6.48 |
| DAE [15] | ✓ | 94.04±2.85 | 8.79±7.45 | 83.11±11.89 | 11.56±6.32 |
| DPL [4] | ✓ | 96.39±1.33 | **4.08±1.49** | 83.53±17.80 | 11.39±10.18 |
| **CBMT(Ours)** | ✓ | **96.61±1.45** | **3.85±1.63** | 84.33±11.70 | **10.30±5.88** |

**Table 2.** Ablation study results of our proposed modules on the *RIM-ONE-r3* dataset. P-L means vanilla pseudo-labeling method. * represents the accuracy is manually selected from the best epoch. The best results are highlighted.

| P-L | EMA | Aug. | Calib. | Avg. Dice ↑ | Avg. ASSD ↓ |
|---|---|---|---|---|---|
| ✓ | | | | 64.19 (84.68*) | 15.11 (9.67*) |
| ✓ | ✓ | | | 83.63 | 8.51 |
| ✓ | ✓ | ✓ | | 84.36 | 8.48 |
| ✓ | ✓ | | ✓ | 86.04 | 8.26 |
| ✓ | ✓ | ✓ | ✓ | **87.26** | **7.29** |

from [26] as an upper bound, same as [4]. As shown in the table, both two quantitative metric results perform better than previous state-of-the-art SFDA methods and even show an improvement against traditional UDA methods on

**Table 3.** Loss calibration weight with different thresholds $\alpha$ on *RIM-ONE-r3* dataset. Our method is robust to the hyper-parameter setting.

| $\alpha$ | 0 | 0.1 | 0.2 | 0.3 | 0.4 | 0.5 | 0.6 | 0.7 | 0.8 | 0.9 |
|---|---|---|---|---|---|---|---|---|---|---|
| $\eta_k^{\mathrm{fg}}/\eta_k^{\mathrm{bg}}$ | 2.99 | 0.24 | 0.24 | 0.24 | 0.24 | 0.24 | 0.23 | 0.23 | 0.22 | 0.22 |

some metrics. Especially in the RIM-ONE-r3 dataset, our CBMT gains a great performance increase than previous works (dice gains by 3.23 on disc), because the domain shift issue is severer here and has big potential for improvement.

Moreover, CBMT alleviates the need for precise tuning of hyper-parameters. Here we could set a relatively long training procedure (our epoch number is 10 times that of [4,27]), and safely select the last checkpoint as our final result without concerning about model degradation issue, which is crucial in real-world clinical source-free domain adaptation application.

### 3.2   Further Analyses

**Ablation Study.** In order to assess the contribution of each component to the final performance, we conducted an ablation study on the main modules of CBMT, as summarized in Table 2. Note that we reduced the learning rates by a factor of 20 for the experiments of the vanilla pseudo-labeling method to get comparable performance because models are prone to degradation without EMA updating. As observed in quantitative results, the EMA update strategy avoids the model from degradation, which the vanilla pseudo-labeling paradigm suffers from. Image augmentation and loss calibration also boost the model accuracy, and the highest performance is achieved with both. The loss calibration module achieves more improvement in its solution to class imbalance, while image augmentation is easy to implement and plug-and-play under various circumstances.

**Hyper-parameter Sensitivity Analysis.** We further investigate the impact of different hyper-parameter. Figure 2(a) presents the accuracy with different EMA update rate parameters $\lambda$. It demonstrates that both too low and too high update rates would cause a drop in performance, which is quite intuitive: a higher $\lambda$ leads to inconsistency between the teacher and student, and thus teacher can hardly learn knowledge from the student; On the other hand, a lower $\lambda$ will always keep teacher and student close, making it degenerated to vanilla pseudo-labeling. But within a reasonable range, the model is not sensitive to update rate $\lambda$.

To evaluate the variation of the loss calibration weight $\eta_k^{\mathrm{fg}}/\eta_k^{\mathrm{bg}}$ with different constraint thresholds $\alpha$, we present the results in Table 3. As we discussed in Sect. 2.2, most pixels in an image are well-classified, and if we simply calculate with all pixels (i.e. $\alpha = 0$), as shown in the first column, the mean loss of background will be severely underestimated due to the large quantity of zero-loss pixel. Besides, as $\alpha$ changes, the calibration weight varies little, indicating the robustness of our calibration technique to threshold $\alpha$.
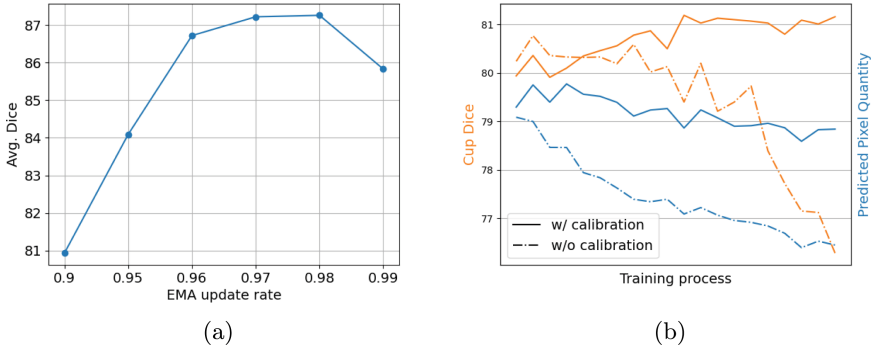
**Fig. 2.** (a) Model performance with different EMA update rate $\lambda$ setting. (b) Training curves with and without our proposed loss calibration scheme.

**The Effectiveness of Loss Calibration to Balance Class.** The class imbalance problem can cause misalignment in the learning processes of different classes, leading to a gradual decrease of predicted foreground area. This can ultimately result in model degradation. As shown in Fig. 2(b), neglecting the issue of class imbalance can cause a significant drop in the predicted pixel quantity of the class "cup" during training, and finally leads to a performance drop. Loss calibration performs a theoretical investigation and proposes an effective technique to alleviate this issue by balancing loss with global context.

## 4    Conclusion

In this work, we propose a class-balanced mean teacher framework to realize robust SFDA learning for more realistic clinical application. Based on the observation that model suffers from degradation issues during adaptation training, we introduce a mean teacher strategy to update the model via an exponential moving average way, which alleviates error accumulation. Meanwhile, by investigating the foreground and background imbalance problem, we present a global knowledge guided loss calibration module. Experiments on two fundus image segmentation datasets show that CBMT outperforms previous SFDA methods.

## References

1. Bateson, M., Kervadec, H., Dolz, J., Lombaert, H., Ben Ayed, I.: Source-Relaxed domain adaptation for image segmentation. In: MICCAI 2020. LNCS, vol. 12261, pp. 490–499. Springer, Cham (2020). https://doi.org/10.1007/978-3-030-59710-8_48

2. Cai, J., Zhang, Z., Cui, L., Zheng, Y., Yang, L.: Towards cross-modal organ translation and segmentation: a cycle-and shape-consistent generative adversarial network. Med. Image Anal. **52**, 174–184 (2019)

3. Carion, N., Massa, F., Synnaeve, G., Usunier, N., Kirillov, A., Zagoruyko, S.: End-to-end object detection with transformers. In: Vedaldi, A., Bischof, H., Brox, T., Frahm, J.-M. (eds.) ECCV 2020. LNCS, vol. 12346, pp. 213–229. Springer, Cham (2020). https://doi.org/10.1007/978-3-030-58452-8_13

4. Chen, C., Liu, Q., Jin, Y., Dou, Q., Heng, P.-A.: Source-free domain adaptive fundus image segmentation with denoised pseudo-labeling. In: de Bruijne, M., et al. (eds.) MICCAI 2021. LNCS, vol. 12905, pp. 225–235. Springer, Cham (2021). https://doi.org/10.1007/978-3-030-87240-3_22

5. 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)

6. Drozdzal, M., Vorontsov, E., Chartrand, G., Kadoury, S., Pal, C.: The importance of skip connections in biomedical image segmentation. In: Carneiro, G., et al. (eds.) LABELS/DLMIA -2016. LNCS, vol. 10008, pp. 179–187. Springer, Cham (2016). https://doi.org/10.1007/978-3-319-46976-8_19

7. Fumero, F., Alayón, S., Sanchez, J.L., Sigut, J., Gonzalez-Hernandez, M.: Rim-one: an open retinal image database for optic nerve evaluation. In: 2011 24th International Symposium on Computer-based Medical Systems (CBMS), pp. 1–6. IEEE (2011)

8. Gadermayr, M., Gupta, L., Appel, V., Boor, P., Klinkhammer, B.M., Merhof, D.: Generative adversarial networks for facilitating stain-independent supervised and unsupervised segmentation: a study on kidney histology. IEEE Trans. Med. Imaging **38**(10), 2293–2302 (2019)

9. Ganin, Y., et al.: Domain-adversarial training of neural networks. J. Mach. Learn. Res. **17**(1), 2096–2030 (2016)

10. He, C., et al.: Camouflaged object detection with feature decomposition and edge reconstruction. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 22046–22055 (2023)

11. He, C., et al.: Weakly-supervised concealed object segmentation with SAM-based pseudo labeling and multi-scale feature grouping. arXiv preprint: arXiv:2305.11003 (2023)

12. Ibtehaz, N., Rahman, M.S.: MultiResUNet: rethinking the u-net architecture for multimodal biomedical image segmentation. Neural Netw. **121**, 74–87 (2020)

13. Javanmardi, M., Tasdizen, T.: Domain adaptation for biomedical image segmentation using adversarial training. In: 2018 IEEE 15th International Symposium on Biomedical Imaging (ISBI 2018), pp. 554–558. IEEE (2018)

14. Kamnitsas, K., et al.: Unsupervised domain adaptation in brain lesion segmentation with adversarial networks. In: Niethammer, M., et al. (eds.) IPMI 2017. LNCS, vol. 10265, pp. 597–609. Springer, Cham (2017). https://doi.org/10.1007/978-3-319-59050-9_47

15. Karani, N., Erdil, E., Chaitanya, K., Konukoglu, E.: Test-time adaptable neural networks for robust medical image segmentation. Med. Image Anal. **68**, 101907 (2021)

16. Li, K., Zhang, Y., Li, K., Fu, Y.: Adversarial feature hallucination networks for few-shot learning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 13470–13479 (2020)

17. Liu, X., Xing, F., Yang, C., El Fakhri, G., Woo, J.: Adapting off-the-shelf source Segmenter for target medical image segmentation. In: de Bruijne, M., et al. (eds.) MICCAI 2021. LNCS, vol. 12902, pp. 549–559. Springer, Cham (2021). https://doi.org/10.1007/978-3-030-87196-3_51

18. Liu, Y.C., et al.: Unbiased teacher for semi-supervised object detection. arXiv preprint: arXiv:2102.09480 (2021)

19. Milletari, F., Navab, N., Ahmadi, S.A.: V-Net: fully convolutional neural networks for volumetric medical image segmentation. In: 2016 Fourth International Conference on 3D Vision (3DV), pp. 565–571. IEEE (2016)

20. Orlando, J.I., Fu, H., Breda, J.B., van Keer, K., Bathula, D.R., Diaz-Pinto, A., Fang, R., Heng, P.A., Kim, J., Lee, J., et al.: Refuge challenge: a unified framework for evaluating automated methods for glaucoma assessment from fundus photographs. Med. Image Anal. **59**, 101570 (2020)

21. Ren, S., He, K., Girshick, R., Sun, J.: Faster R-CNN: towards real-time object detection with region proposal networks. In: Advances in Neural Information Processing Systems, vol. 28 (2015)

22. Ronneberger, O., Fischer, P., Brox, T.: U-Net: convolutional networks for biomedical image segmentation. In: Navab, N., Hornegger, J., Wells, W.M., Frangi, A.F. (eds.) MICCAI 2015. LNCS, vol. 9351, pp. 234–241. Springer, Cham (2015). https://doi.org/10.1007/978-3-319-24574-4_28

23. Sandler, M., Howard, A., Zhu, M., Zhmoginov, A., Chen, L.C.: Mobilenetv 2: inverted residuals and linear bottlenecks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 4510–4520 (2018)

24. Sivaswamy, J., Krishnadas, S., Chakravarty, A., Joshi, G., Tabish, A.S., et al.: A comprehensive retinal image dataset for the assessment of glaucoma from the optic nerve head analysis. JSM Biomed. Imaging Data Pap. **2**(1), 1004 (2015)

25. Vu, T.H., Jain, H., Bucher, M., Cord, M., Pérez, P.: ADVENT: adversarial entropy minimization for domain adaptation in semantic segmentation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 2517–2526 (2019)

26. Wang, S., Yu, L., Li, K., Yang, X., Fu, C.-W., Heng, P.-A.: Boundary and entropy-driven adversarial learning for fundus image segmentation. In: Shen, D., et al. (eds.) MICCAI 2019. LNCS, vol. 11764, pp. 102–110. Springer, Cham (2019). https://doi.org/10.1007/978-3-030-32239-7_12

27. Xu, Z., et al.: Denoising for relaxing: unsupervised domain adaptive fundus image segmentation without source data. In: Wang, L., Dou, Q., Fletcher, P.T., Speidel, S., Li, S. (eds.) MICCAI 2022. Lecture Notes in Computer Science, vol. 13435, pp. 214–224. Springer, Cham (2022). https://doi.org/10.1007/978-3-031-16443-9_21