Seamless Iterative Semi-Supervised Correction of Imperfect Labels in Microscopy Images

Marawan Elbatel*, Christina Bornberg*, Manasi Kattel, Enrique Almar, Claudio Marrocco, and Alessandro Bria

University of Cassino and Southern Lazio, Cassino, Italy
marawan-kefah-fathi_elbatel@etu.u-bourgogne.fr

Abstract. In-vitro tests are an alternative to animal testing for the toxicity of medical devices. Detecting cells as a first step, a cell expert evaluates the growth of cells according to cytotoxicity grade under the microscope. Thus, human fatigue plays a role in error making, making the use of deep learning appealing. Due to the high cost of training data annotation, an approach without manual annotation is needed. We propose *Seamless Iterative Semi-Supervised correction of Imperfect labels (SISSI)*, a new method for training object detection models with noisy and missing annotations in a semi-supervised fashion. Our network learns from noisy labels generated with simple image processing algorithms, which are iteratively corrected during self-training. Due to the nature of missing bounding boxes in the pseudo labels, which would negatively affect the training, we propose to train on dynamically generated synthetic-like images using seamless cloning. Our method successfully provides an adaptive early learning correction technique for object detection. The combination of early learning correction that has been applied in classification and semantic segmentation before and synthetic-like image generation proves to be more effective than the usual semi-supervised approach by $> 15\%$ AP and $> 20\%$ AR across three different readers. Our code is available at https://github.com/marwankefah/SISSI.

Keywords: Label Correction · Cell Detection · Semi-Supervised Object Detection

1 Introduction

Testing medical devices with animals have a long tradition according to ISO 10993 [1]. Since 2017 the ISO 10993 has gradually evolved towards implementing alternative test methods. One of the in-vitro methods is the testing of cytotoxicity, described in the ISO 10993-5 [3]. Cell experts analyze cell growth of a fibroblast cell line such as L929 with the help of a microscope. The acceptance criteria for medical devices is 50% of dead cells (grade 2 criteria). If there are more than 50% dead cells, the medical device is not allowed to enter the market.

* Co-first authors

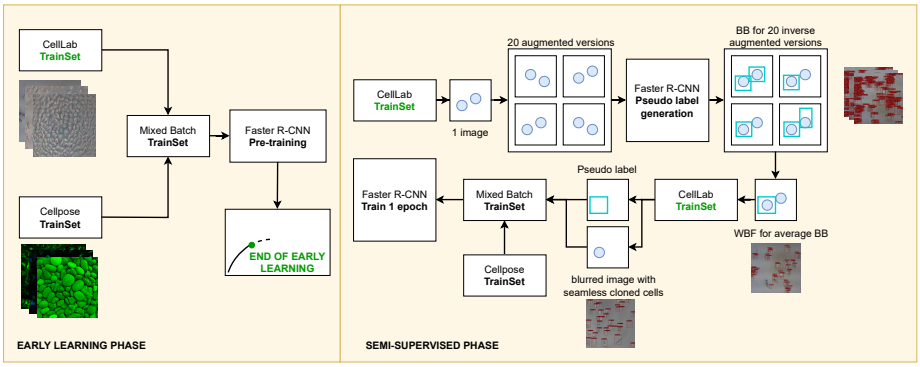

Fig. 1. Overall scheme of SISSI framework.

In this context, deep learning can serve as a second opinion since human error in the workplace is costly and dependent on the level of fatigue; the greater the level of fatigue, the higher the risk of errors occurring. Especially in the borderline cases of grade 2, the cell expert needs to be able to obtain a second opinion that is independent of human fatigue. Deep learning has shown substantial benefits in different life science and pharma applications such as chemo-informatics, computational genomics, and biomedical imaging such as cell segmentation [12] and seems to be a promising supplement to cytotoxicity grading. In the first instance, cells need to be detected, and in future work, an intuitive way of classifying cells into dead or alive needs to be found.

When dealing with imperfect datasets, problems including (partly) missing, inaccurate, or wrong labels arise. To handle imperfect datasets in object detection/segmentation tasks, one can leverage unlabelled (self/semi-supervised) or external labelled (transfer learning) data, regularise training, learn with class labels, and revisit loss functions (sparse/noisy labels) [15].

2 Related Works

Previous work has studied imperfect datasets, including semantic segmentation, instance segmentation, and object detection [18,19,4]. [18] propose a pipeline for semantic semi-supervised segmentation that separates pixels of a pseudo labelled image into reliable and unreliable. [6] propose Adaptive Early Learning Correction (ADELE) for semantic segmentation, with a supervised early-learning phase and subsequently a label correction phase. [8] propose a label mining pipeline for missing annotations using co-teaching for instance segmentation. [19] propose to generate masks with the Circle Hough Transform (CHT) and iteratively create pseudo labels with self-training for images where CHT failed. [20] propose to use a background calibration loss inspired by focal loss for object detection with missing annotations. [4] propose only annotating one instance

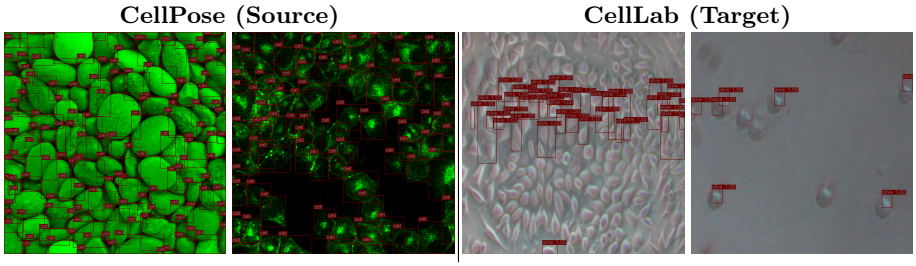

Fig. 2. Examples of the the labelled source and the noisy labelled target datasets.

per category in an image and iteratively generating pseudo-labels. [2] propose an object detector to handle noisy labels, masking the negative sample loss in the box predictor to avoid the harm of false-negative labels.

Though advances in dealing with imperfect datasets have been made, the problem of dealing with datasets having partly missing labels that are additionally noisy in object detection tasks remains.

We propose SISSI (Seamless Iterative Semi-Supervised correction of Imperfect labels) for training object detection models with noisy and missing labels in a semi-supervised fashion, see Fig. 1. We perform several experiments with mixed-batch training, self-training with iterative label correction, synthetic-like image generation, and altering the starting point of self-training (ADELE vs. validation loss).

3 Materials and Methods

3.1 Datasets

Microscopy images of fibroblast (L929) were aquired using a Nikon Eclipse TS 100 microscope and the OPTOCAM-I camera. This trainset (**CellLab** dataset) consists of 224 images, and their noisy annotations are generated with simple image processing pipelines such as Circle Hough Transform, Watershed, and Edge Detection. A detailed description of the initial weak label generation is shown in Fig. 5 in Appendix A. The CellLab testset consists of five images (640×480) annotated by three cell experts. Three readers annotated five images independently, resulting in (reader 1) 552, (reader 2) 565, and (reader 3) 477 annotated cells for the five images. In order to perform domain adaptation and enhance our weak and noisy labelled CellLab dataset, we use the labelled **Cellpose** [14] dataset. It consists of a large variety of fluorescent markers and image modalities, as well as natural images that can be segmented into repetitive structures/blobs. The Cellpose dataset is used for training (45,215 cells on 539 images) and validation (7,195 cells on 68 images). We extract bounding boxes from the segmentation masks for our detection task. We show examples of both datasets in Fig. 2.

3.2 Overall Framework

SISSI integrates a range of image processing and deep learning methods to make iterative label correction possible.

The **early learning phase** consists of a mixed-batch training combining the CellLab and Cellpose training datasets. We train the Faster R-CNN model in a supervised fashion with a Balanced Gradient Contribution [11], mixed-batch training, of target dataset with initial noisy annotations and source dataset until a memorisation phase on the noisy annotations is reached. We determine the end of early learning with a deceleration point based on the AP_{50} curve between the weak ground truth of the CellLab dataset and the model output.

In the following **semi-supervised phase** for each cycle, first, we apply label correction, followed by mixed-batch training with the pseudo labels and synthetic-like images (excluding undetected cells) of the CellLab dataset combined with the original Cellpose dataset. Pseudo-label generation uses test-time augmentation and weighted boxes fusion to generate confident bounding boxes. Since some cells are not detected, their appearance in the original image will confuse the network while training. Thus, we generate dynamically synthetic-like images for continual training. The overall scheme of SISSI framework is shown in Fig. 1.

3.3 Determining the Start of the Semi-Supervised Phase

While training with mixed-batch training, we notice a two-stage learning phenomenon previously noted in classification and semantic segmentation: in an early learning phase, the network fits the clean annotations; then, the network start memorising the initial noisy annotations [7,6]. To find the optimal point that represents when the memorisation phase starts, we adopt a method, ADELE [6], that has been used in previous works in the context of semantic segmentation. In our work, we rely on the deceleration of the AP_{50} training curve of the model output and the initial noisy annotated dataset, CellLab, to decide when to stop trusting the initial noisy annotations and generate pseudo labels. See Fig. 7 in Appendix B for the AP_{50} training curve with the point representing when the memorisation phase starts.

3.4 Pseudo Label Generation

Pseudo label generation is a technique where a pre-trained neural network generates labels for unlabelled data or updates labels for noisy labeled data [16]. We generate pseudo labels to update the noisy annotations of the CellLab dataset during the semi-supervised phase. Self-training networks have the disadvantage of being unable to correct their own mistakes. Therefore biased and wrong labels can be amplified. To filter potential bounding boxes, we integrate two techniques, test-time augmentation (TTA) [17] and weighted boxes fusion (WBF) [13]. We average predictions generated with TTA while considering the confidence score

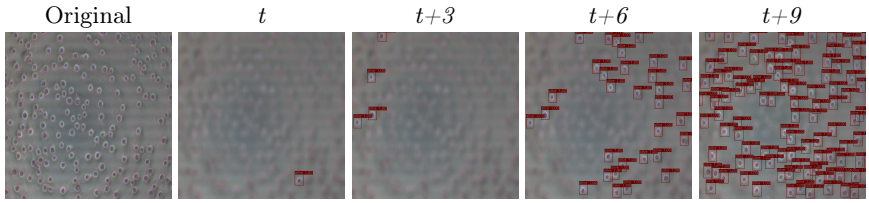

Fig. 3. Example of a synthetic-like image with weak blurring in training epochs (t).

of each bounding box in a WBF manner:

$$X_{1,2} = \frac{\sum_{i=1}^{T} C_i \cdot X_{1,2_i}}{\sum_{i=1}^{T} C_i}, \tag{1}$$

where T is the number for bounding boxes assigned to a single object in a cluster, $X_{1,2}$ (or $Y_{1,2}$) is the average start and end point on the x (or y) axis. This yields the average of the bounding box coordinates $X_{1,2_i}$ (or $Y_{1,2_i}$), weighted with the confidence score C_i for each bounding box.

3.5 Synthetic-like image adaptation according to pseudo labels

Undetected cells in the pseudo labels would affect the further training negatively. When the network localises true objects that are not present in the pseudo-labels, the network is penalised for those objects that are true. To solve this problem, we propose to generate synthetic-like images dynamically according to the pseudo labels generated for the CellLab dataset, see Fig. 3. To remove unlabeled cells in the training image in order not to confuse the network, we clone all the detected cells of the pseudo label (source) onto a strongly/weakly Gaussian blurred image (target). To avoid discontinuities between the target and the source, we mix edge textures with the seamless cloning algorithm (mixing gradient) [9].

4 Experiments

4.1 Implementation Details

The backbone of our Faster R-CNN is a ResNet-50, pre-trained on the MS COCO dataset [5]. We set hyperparameters according to existing Fast/Faster R-CNN work [10]. We do not freeze any layer to allow the gradient to propagate through the early layers.

We train the models using the Stochastic Gradient Descent (SGD) optimiser with a momentum of 0.9, weight decay of 0.0002, and learning rate of 0.001. We use a batch size of 8, with an equal number of images randomly chosen from the CellLab and Cellpose datasets, and resize the images to 512×512. We perform simple augmentations: channel shuffle, Gaussian blurring, horizontal flip, vertical flip, and shift-scale-rotate. For test-time augmentation used for label

Algorithm 1 Pseudocode for iterative self-training with SISSI, prediction (p), target (t), bounding boxes (bbs).

Require: $CLimg, CLbbs_t, CPimg, CPbbs_t$ ▷ CellLab and Cellpose datasets
Require: $NN(img)$ ▷ Faster R-CNN
Require: $E \leftarrow this.self\_training\_epoch$
  for each pseudo_batch B in E **do** ▷ Pseudo label generation
    $CLbbs_p[n], scores[n] \leftarrow NN(TTA(CLimg \in B))$ ▷ Generate boxes with TTA
    $CLbbs_t \leftarrow (\sum_{n=1}^{N} scores[n] \cdot CLbbs_p[n]) / \sum_{n=1}^{N} scores[n]$ ▷ Filter bbs with WBF
    $update\_dataset(CLbbs_t)$ ▷ Update final pseudo label
  end for
  for each mixed_batch B in E **do** ▷ Training
    $CLimg\_crops[n] \leftarrow crop(CLimg, CLbbs)$ ▷ Synthetic image generation
    $CLimg\_blur \leftarrow blur(CLimg)$
    $CLimg\_synth \leftarrow seamless\_clone(CLimg\_blur, CLimg\_crops[n])$
    $CLbbs_p, CPbbs_p \leftarrow NN(CLimg\_synth, CPimg \in B)$ ▷ Prediction
    $CLCP\_loss \leftarrow loss([CLbbs_p, CPbbs_p], [CLbbs_t, CPbbs_t])$ ▷ Loss calculation
    $CLCP\_loss.backprop()$
  end for
  $E.next()$

correction, we use a combination of scaling ([0.8, 0.9, 1, 1.1, 1.2]) and augmentations, vertical flipping, horizontal flipping, horizontal+vertical flipping, or no flipping. We end up with 20 versions of the same image. For background blurring in the synthetic-like image generation, we use Gaussian blurring with kernels of (21, 21) and kernels (12, 32), referred to as weak(W) and strong(S) background blurring respectively.

The datasets are used as follows. Mixed-batch training is applied in both the early supervised and semi-supervised learning phases, combining the CellLab and Cellpose training sets. With the start of self-training, labels and synthetic-like images for the CellLab dataset are updated in each following epoch. To perform validation for hyperparameter tuning, we use the Cellpose dataset since only five manually annotated images are available in the CellLab dataset, which all are used as a testset. For estimating the end of early learning, the weak training labels of CellLab are compared to the model output as proposed in ADELE. We calculate deceleration by the relative change in the derivative of the AP_{50} curve, and if it is above a certain threshold, 0.9, then label correction starts.

4.2 Evaluation Metrics and Results

In Table 1, we report three versions of AP, and AR over the CellLab testset. The metrics include the Pascal VOC metric (AP_{50}), as well as COCO evaluation metrics [5] (AP_{75}, and AP and AR averaged over different IoU thresholds). Bold numbers denote the best performance for each of the three cell experts' annotations.

We present the detection performance of different experiments on the CellLab testset. The Baseline model is first trained in a supervised fashion with

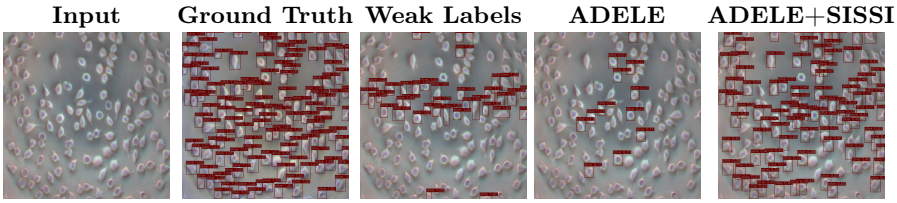

Fig. 4. Demonstration of improvement of results with our proposed method.

mixed-batch training of CellLab and Cellpose datasets. Aiming to correct and complete the labels, we perform early-stopping based on the validation loss of the source dataset, Cellpose, and apply self-training with test-time augmentation (TTA) and weighted boxes fusion (WBF) to iteratively update the pseudo labels. The following two experiments SISSI(W) and SISSI(S) additionally use synthetic-like image generation for the CellLab images, based on the pseudo labels generated with TTA and WBF. The weak (W) background blurring achieved better results than the baseline, while strong (S) background blurring has worse performance. Plain A, is an additional experiment similar to baseline setting but with an additional algorithm (ADELE) to find an optimal starting point for label correction. This shows an increase of about 10% across all readers in the AR metrics, compared to the baseline model, while the AP is lower.

Two versions of the final pipeline show the best results. It combines all steps (1) supervised learning with mixed-batch training of CellLab and Cellpose till a memorisation point is reached, (2) iteratively applying pseudo-label generation for the CellLab dataset with test-time augmentation and weighted boxes fusion, (3) generation of synthetic-like images according to the pseudo labels, and (4) training the network for an Epoch with mixed-batch training of the Cellpose dataset and the pseudo labels and synthetic-like images of the CellLab dataset. Pseudo code for the loop of 2-4 can be seen in Algorithm 1. Incorporating ADELE with our label correction and synthetic-like image generation method with strong blurring increases the AP by at least 15% and AR by at least 20% compared to the baseline across all cell experts. On Fig. 4, we show an example where SISSI successfully improves the detection results of Plain A (ADELE) experiment. Examples of a training image with its pseudo labels for different epochs (t) and experiments can be seen in Fig. 8 in Appendix B.

5 Discussion

5.1 Findings

The experiments made clear that both the start of label correction and the amount of background information appearing in images during training impact the results. When starting label correction too early, during early learning, the network is not confident enough to detect all objects in the image; thus, correcting initially noisy annotations at this stage results in a high rate of missing

Table 1. Results of Cell Detection on the CellLab testset.

	Annotator 1				Annotator 2				Annotator 3			
Pipeline	AP_{50}	AP_{75}	AP	AR	AP_{50}	AP_{75}	AP	AR	AP_{50}	AP_{75}	AP	AR
Baseline	45.6	15.7	21.3	32.7	44.3	16.1	20.2	31.0	**58.6**	28.1	29.7	41.8
SISSI(W)	52.8	23.2	25.9	37.7	49.5	21.3	24.4	34.6	58.5	29.0	29.7	42.0
SISSI(S)	40.3	8.3	16.2	32.8	38.4	7.2	15.2	31.8	46.6	9.7	19.2	37.8
Plain A	38.7	18.9	19.2	43.9	35.8	18.6	18.7	43.1	41.2	23.8	22.9	50.6
A+SISSI(W)	43.1	37.1	36.0	**60.3**	45.1	38.5	37.4	**58.8**	47.6	42.8	41.4	**66.9**
A+SISSI(S)	**54.9**	**49.0**	**43.2**	57.6	**51.2**	**45.1**	**39.7**	54.3	58.5	**55.5**	**47.9**	64.9

targets. Training a network on images with high missing targets without SISSI (Plain A) increases the uncertainty of the network compared to label correction in a later memorization phase with fewer missing targets, the baseline.

In the basic SISSI approach, where label correction is started on a model chosen based on the validation loss of the external Cellpose dataset, weak background blurring worked better than a strongly blurred background. We believe this phenomenon appears because the neural network has learned more contextual information in the memorisation stage and requires the background information.

On the other hand, starting label correction when the early learning phase ends, according to ADELE, strong blurring shows better results than weak blurring. The information about the background is less important. This can be an advantage in synthetic-like image generation because figuring out how to preserve contextual information seems less critical.

5.2 Limitations

The success of SISSI may be dependent on the stopping criteria and the training phase, early learning/memorisation phase. When the annotations in the image are too noisy, the network may not encompass the early learning phase as in previous works, ADELE. It may be unable to learn the task to produce new pseudo labels for further training. The effect of blurring during different training phases needs more empirical research for verification. SISSI is a simple approach that works with only one class of interest to detect. Blurring with multi-object needs further modification in future works. We use SISSI in these experiments with Faster R-CNN, which is more robust and friendly for the missing label scenario than other detection networks.

5.3 Conclusion

This paper presents a method to train object detection models with noisy and missing annotations with semi-supervised learning by proposing a novel technique. We use dynamically generated synthetic-like images using seamless cloning for further training the network after pseudo-label generation. We utilize a domain adaptation technique, Balanced Gradient Contribution, to generate stable

gradient directions and mitigate the so noisy annotation problem for our semi-supervised training. Finally, we evaluate our method for the cell detection task with various training procedures and show its improvement over the usual semi-supervised approach. Our method, SISSI, can be added on top of any detection network, and it also helps other methods like ADELE to be leveraged for object detection. In the future, we will adapt our method to work with multi-object detection and explore SISSI with different detection networks. Moreover, we will explore our method for different medical detection tasks and integrate our network to help cell experts with the grading task.

Acknowledgements We would like to acknowledge Oesterreichisches Forschungsinstitut für Chemie und Technik (OFI) for CellLab images and test set annotation.

References

1. Anderson, J.M.: Future challenges in the in vitro and in vivo evaluation of biomaterial biocompatibility . Regenerative Biomaterials **3**(2), 73–77 (03 2016). https://doi.org/10.1093/rb/rbw001, https://doi.org/10.1093/rb/rbw001
2. Gao, J., Wang, J., Dai, S., Li, L.J., Nevatia, R.: Note-rcnn: Noise tolerant ensemble rcnn for semi-supervised object detection. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV) (October 2019)
3. ISO: Iso 10993-5: 2009-biological evaluation of medical devices-part 5: Tests for in vitro cytotoxicity (2009)
4. Li, H., Pan, X., Yan, K., Tang, F., Zheng, W.S.: Siod: Single instance annotated per category per image for object detection. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 14197–14206 (2022)
5. Lin, T.Y., Maire, M., Belongie, S., Bourdev, L., Girshick, R., Hays, J., Perona, P., Ramanan, D., Zitnick, C.L., Dollár, P.: Microsoft coco: Common objects in context (2014), http://arxiv.org/abs/1405.0312
6. Liu, S., Liu, K., Zhu, W., Shen, Y., Fernandez-Granda, C.: Adaptive early-learning correction for segmentation from noisy annotations. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 2606–2616 (2022)
7. Liu, S., Niles-Weed, J., Razavian, N., Fernandez-Granda, C.: Early-learning regularization prevents memorization of noisy labels. Advances in Neural Information Processing Systems **33** (2020)
8. Lyu, F., Yang, B., Ma, A.J., Yuen, P.C.: A segmentation-assisted model for universal lesion detection with partial labels. In: de Bruijne, M., Cattin, P.C., Cotin, S., Padoy, N., Speidel, S., Zheng, Y., Essert, C. (eds.) Medical Image Computing and Computer Assisted Intervention – MICCAI 2021. pp. 117–127. Springer International Publishing, Cham (2021)
9. Pérez, P., Gangnet, M., Blake, A.: Poisson image editing. ACM Trans. Graph. **22**(3), 313–318 (jul 2003). https://doi.org/10.1145/882262.882269, https://doi.org/10.1145/882262.882269
10. Ren, S., He, K., Girshick, R., Sun, J.: Faster r-cnn: Towards real-time object detection with region proposal networks. Advances in neural information processing systems **28** (2015)

11. Ros, G., Stent, S., Fernández Alcantarilla, P., Watanabe, T.: Training constrained deconvolutional networks for road scene semantic segmentation. CoRR (04 2016)
12. Siegismund, D., Tolkachev, V., Heyse, S., Sick, B., Duerr, O., Steigele, S.: Developing deep learning applications for life science and pharma industry. Drug research **68**(06), 305–310 (2018)
13. Solovyev, R., Wang, W., Gabruseva, T.: Weighted boxes fusion: Ensembling boxes from different object detection models. Image and Vision Computing **107**, 104117 (2021)
14. Stringer, C., Pachitariu, M.: Cellpose 2.0: how to train your own model. bioRxiv (2022). https://doi.org/10.1101/2022.04.01.486764, https://www.biorxiv.org/content/early/2022/04/05/2022.04.01.486764
15. Tajbakhsh, N., Jeyaseelan, L., Li, Q., Chiang, J.N., Wu, Z., Ding, X.: Embracing imperfect datasets: A review of deep learning solutions for medical image segmentation. Medical Image Analysis **63**, 101693 (2020). https://doi.org/https://doi.org/10.1016/j.media.2020.101693, https://www.sciencedirect.com/science/article/pii/S136184152030058X
16. Triguero, I., García, S., Herrera, F.: Self-labeled techniques for semi-supervised learning: taxonomy, software and empirical study. Knowledge and Information systems **42**(2), 245–284 (2015)
17. Wang, G., Li, W., Aertsen, M., Deprest, J., Ourselin, S., Vercauteren, T.: Aleatoric uncertainty estimation with test-time augmentation for medical image segmentation with convolutional neural networks. Neurocomputing **338**, 34–45 (2019)
18. Wang, Y., Wang, H., Shen, Y., Fei, J., Li, W., Jin, G., Wu, L., Zhao, R., Le, X.: Semi-supervised semantic segmentation using unreliable pseudo-labels. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 4248–4257 (June 2022)
19. Xiong, H., Liu, S., Sharan, R.V., Coiera, E., Berkovsky, S.: Weak label based bayesian u-net for optic disc segmentation in fundus images. Artificial Intelligence in Medicine **126**, 102261 (2022)
20. Zhang, H., Chen, F., Shen, Z., Hao, Q., Zhu, C., Savvides, M.: Solving missing-annotation object detection with background recalibration loss. ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) pp. 1888–1892 (2020)