Learning from Partial Label Proportions for Whole Slide Image Segmentation

Shinnosuke Matsuo[1], Daiki Suehiro[1], Seiichi Uchida[1], Hiroaki Ito[2],
Kazuhiro Terada[2], Akihiko Yoshizawa[2], and Ryoma Bise[1]

[1] Kyushu University, Fukuoka, Japan
[2] Kyoto University Hospital, Kyoto, Japan

Abstract. In this paper, we address the segmentation of tumor subtypes in whole slide images (WSI) by utilizing incomplete label proportions. Specifically, we utilize 'partial' label proportions, which give the proportions among tumor subtypes but do not give the proportion between tumor and non-tumor. Partial label proportions are recorded as the standard diagnostic information by pathologists, and we, therefore, want to use them to realize the segmentation model that can classify each WSI patch into one of the tumor subtypes or non-tumor. We call this problem "learning from partial label proportions (LPLP)" and formulate the problem as a weakly supervised learning problem. Then, we propose an efficient algorithm for this challenging problem by decomposing it into two weakly supervised learning subproblems: multiple instance learning (MIL) and learning from label proportions (LLP). These subproblems are optimized efficiently in an end-to-end manner. The effectiveness of our algorithm is demonstrated through experiments conducted on two WSI datasets. This code is available at https://github.com/matsuo-shinnosuke/LPLP.

Keywords: Learning from partial label proportions · Whole slide image segmentation · Weakly-supervised learning.

1 Introduction

Automated segmentation of whole slide images (WSI) by tumor subtypes is crucial in aiding pathologists with accurate tumor diagnosis. Since WSI often has an extremely high resolution, its segmentation task can be treated as a patch-wise classification task [24]. In other words, our goal is to classify each patch from the cell region into one of the C tumor subtype classes or the non-tumor class.

A unique situation of automated WSI segmentation is that we can easily collect label proportions among C tumor subtype classes for each WSI in many pathological diagnoses. As shown in Fig. 1(a), when $C = 2$, pathologists' interest in their diagnosis is determining the proportion among two tumor subtypes, such as $0.8 : 0.2$. For instance, in clinical practices like chemotherapy [25], the PDL1 positive rate [12, 19, 28], and the diagnosis of lung tissue after surgery [5, 16], the determination of a condition's grade depends on the proportion of tumor subtypes excluding non-tumor areas.

For realizing patch-wise classification, however, the proportions only among tumor classes are ambiguous information because they do not care about the non-tumor region.
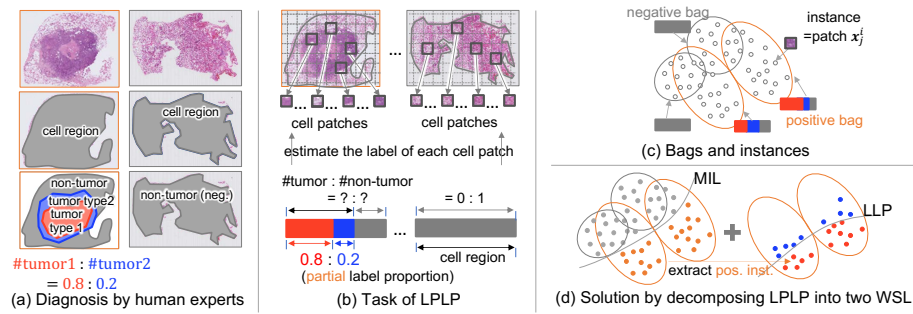
Fig. 1. Overview of our work.

In other words, we can only know the *partial* label proportion. Assume $(0, 8, 0.2)$ shows the proportion between the two tumor subtypes. Note that the proportion of the non-tumor region is unknown. Then, if the proportion between tumor and non-tumor is found to be $0.5 : 0.5$, the actual label proportion is $(0, 4, 0.1, 0.5)$.

This work tackles a new machine-learning problem, *learning from partial label proportions* (LPLP). As shown in Fig. 1(b), by using partial label proportions of WSI images, we train a cell patch classifier to classify a patch into one of $C + 1$ classes. This problem is not solvable without any hint of non-tumor patches; this is simply because the partial label proportions do not give any information about non-tumors. Fortunately, in practical WSI diagnosis scenarios, pathologists often encounter "totally healthy" WSIs, like the right WSI in Fig. 1(a). Since this WSI contains no tumor cells, all its cell patches are guaranteed as non-tumor patches. As detailed in the following, we will fully utilize such WSIs to make LPLP a solvable problem.

As shown in Fig. 1(c), LPLP is formulated by instances, bags, and partial label proportions. An instance corresponds to each cell patch and thus belongs to one of the $C + 1$ classes. Each bag can be positive or negative; a *positive bag* contains instances of tumor patches and, in general, also contains instances of non-tumor patches. In contrast, a *negative bag* contains only instances of non-tumor. A healthy WSI forms a negative bag. A bag is associated with a partial label proportion. For example, if $C = 2$ and the tumor region contains 80% of tumor type1 (red) and 20% of type2 (blue), then the partial label proportion of tumor subtype classes is given as $(0.8, 0.2)$. In contrast, we do not know the proportion of the non-tumor class except when the bag is negative.

The highlight of this paper is that we decompose LPLP into two weakly supervised learning (WSL) subproblems: multiple instance learning (MIL) and learning from label proportions (LLP) as shown in Fig. 1(d). MIL is known as a WSL problem using binary bag labels (positive or negative). In our MIL as the subproblem, we aim to classify the instances for positive (tumor) or negative (non-tumor) as shown in Fig. 1(d) left. LLP is a WSL problem using label proportions for *all* classes. To complement the lack of the proportion information of negative instances, we use the inference of the above MIL. More precisely, our LLP as subproblem is for the subbags of the positive bags, which contains only positive instances as shown in Fig. 1(d) right. Note that these two WSLs

Table 1. Our problem setting and conventional WSL.

	LPLP (Ours)	MIL	WSL other than MIL
Labels	**Partial** label proportions	Bag-level Pos/Neg label	Scribbles, etc.
Labeling cost (↓)	Low (easily available)	Low (easily available)	Relatively High (with extra-annotation)
Objective	Segmentation	Bag-level classification	Segmentation
Comparability	**No fair comparison is possible due to the above differences**		
Existing methods	**None**	[8, 20] etc.	[1, 3] etc.

are not disjoint; we design a neural network that can simultaneously solve both MIL and LLP problems end-to-end.

2 Related work

Weakly-Supervised Learning (WSL) for Whole Slide Images (WSI). WSLs for WSI are summarized in Table 1. The most representative learning problem in WSL for WSI is multiple instance learning (MIL), where the positive (tumor) or negative (non-tumor) label is given to a bag [7, 8, 10, 13, 20]. These studies' objective differs from ours; most MIL methods aim for bag-level classification but do not segment subtypes (instance-level classification). To use an AI system in clinical, showing segmentation results is also important for explainability. Recently, the segmentation from bag-level labels has been challenged by [18, 21]; however, these methods assume the binary class (tumor or non-tumor) and thus does not work for multiple classes, including tumor subtypes. Generally, classifying subtypes is significantly more challenging than binary classification due to the similar appearances of the tumor subtypes.

Several WSL approaches other than MIL have also been proposed. For example, SEGGIN [1] trains a network using scribbles, which can be considered as partial instance (pixel)-level labels, and HISTOSEGNET [3] performs pixel-level WSI segmentation from patch-level labels. These methods do not use the proportions of tumor subtypes in a WSI and require extra annotations. In contrast, the proportions are recorded as diagnosis information as mentioned in the introduction and, thus, are easily available. To the authors' best knowledge, this paper is the first work to deal with the partial label proportion.

Learning from label proportions (LLP). Learning from label proportions (LLP) [2, 22, 6, 9, 11, 15, 17, 23, 26, 29, 30] is a WSL task given bags and their label proportions. For example, if a bag contains 150, 30, and 20 instances of classes 1, 2, and 3, respectively, the proportion is $(0.75, 0.15, 0.10)$. The goal of LLP is to obtain an instance classifier from bags and their label proportions.

The most methods of LLP are based on a proportion loss to train a deep neural network, which predicts the class of an instance. The proportion loss is a bag-level cross-entropy between the ground truth and the label proportions estimated by averaging the probability outputs in a bag. These methods require proportions of 'all' the class. In other words, applying LLP to this LPLP problem requires the proportion between tumor and non-tumor regions, which incurs additional labeling costs.
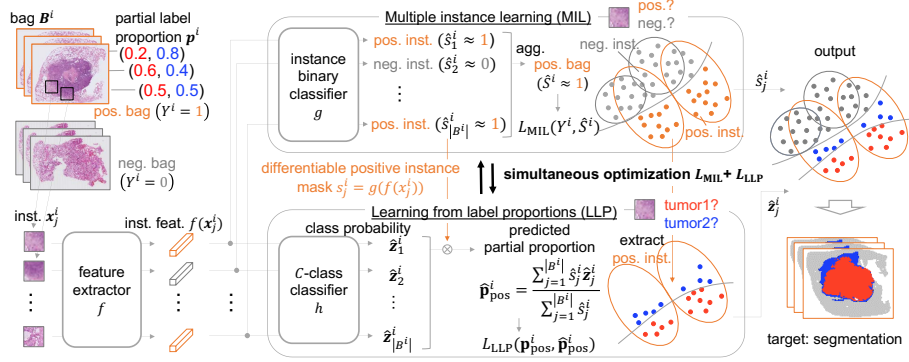
Fig. 2. Overall architecture of the proposed model.

3   Learning from partial label proportions (LPLP)

LPLP is a novel and "more weakly" supervised learning problem than LLP because the label proportion of a certain class is always unknown in all bags. Hereafter, we call the unknown class "the negative class" and the other C "positive classes." Assume n training (positive) bags $\mathcal{B}_{\mathrm{pos}} = \{\boldsymbol{B}^i\}_{i=1}^n$, each of which contains instances of $C+1$ classes $\{\boldsymbol{x}_j^i\}_{j=1}^{|\boldsymbol{B}^i|}$. LPLP aims to train an instance classifier for $(C+1)$ classes by using the partial label proportions, which are the label proportions of positive classes, $\boldsymbol{p}^i = (p_1^i, \ldots, p_C^i) \in \Delta^C, i = 1, \ldots, n$, where Δ^C is the C-probability simplex. Note that the proportion between the positive classes and the negative class is unknown.

Solving the above LPLP problem is impossible as it is because no information is given for the negative class. Fortunately, negative bags $\mathcal{B}_{\mathrm{neg}} = \{\boldsymbol{B}^i\}_{i=n+1}^m$ that *only* contain instances of the negative class are available from easily prepared healthy WSIs. The negative bags can give a hint to discriminate the negative instances from the positive instances. All m training bags are one of these two types, i.e. $\mathcal{B} = \mathcal{B}_{\mathrm{pos}} + \mathcal{B}_{\mathrm{neg}} = \{\boldsymbol{B}^i\}_{i=1}^m$.

In summary, the problem setup of LPLP can be rewritten as follows: given positive training set $\{\boldsymbol{B}^i, Y^i, \boldsymbol{p}^i\}_{i=1}^n$ and negative training set $\{\boldsymbol{B}^i, Y^i\}_{i=n+1}^m$, training an instance classifier for $(C+1)$ classes, composing C positive classes and negative class. Here, the bag-level binary label $Y^i \in \{0, 1\}$, takes 0 if the bag $\boldsymbol{B}^i$ is a negative bag, otherwise 1.

4   Simultaneous learning of MIL and LLP for LPLP

Using additionally obtained the bag-level binary class information, we decompose the challenging WSL problem LPLP into two simpler WSL problems: multiple instance learning (MIL) and learning from label proportion (LLP). As shown in Fig. 2, the proposed method consists of the MIL and LLP modules. In the MIL module, given bag-level binary labels, it aims to train an instance-level binary classifier g that classifies an

instance x_j^i into positive or negative. LLP aims to train an instance classifier h for C positive classes using a set of bags only containing positive instances extracted by the MIL module and their partial proportions.

If we simply apply this two-stage framework, the classification performance depends on the first step (MIL); the errors in MIL estimation might affect the second step (LLP). This is because the positive proportion is shifted if the bag contains noisy negative instances.

To avoid this problem, we simultaneously optimize both MIL and LLP end-to-end. The MIL module classifies an instance into a positive or negative instance. Then, the positive instance is further classified into either of the C positive classes while ignoring negative classes by multiplying the outputs of the LLP module by the positive instance masks. It enables to training of both modules simultaneously.

MIL module. In the MIL module of our method, we utilize mi-Net (output-aggregation-MIL) [27], which estimates the bag-level label by aggregating the outputs of the feature extractor f and the instance classifier g for each instance in a bag. Moreover, the classifier g generates the differentiable positive instance scores (masks) $\hat{s}_j^i = g(f(x_j^i)) \in [0, 1]$, which complement the lack of negative proportions in positive bags for the LLP module. The differentiable positive instance scores are close to 1 for positive instances and 0 for negative instances. All of the instance scores $\{\hat{s}_j^i\}_{j=1}^{|\boldsymbol{B}^i|}$ in a bag are aggregated using an aggregation function, such as the mean-pooling, max-pooling, and LogSum-Exp (LSE)-pooling, to obtain the bag score $\hat{S}^i$. A bag-level binary cross-entropy is used for the loss function of the MIL task as follows:

$$\mathcal{L}_{\mathrm{MIL}}(Y^i, \hat{S}^i) = -Y^i \log \hat{S}^i - (1 - Y^i) \log(1 - \hat{S}^i). \tag{1}$$

In the output-aggregation MIL, although the network is trained by bag-level labels, the instance-level classifier can be trained. Let us consider a case using mean-pooling as the aggregation function; when the bag-level label is 0, the network is trained so that the output for an instance becomes 0, and when it's 1, positive instances are close to 1, while negative instances tend to keep 0 to make consistent with learning from the negative bags.

LLP module with the positive instance mask. The LLP module aims to further classify instances, which are selected as positive by MIL, into C classes. To train the C positive class classifier h, we use positive bags, their proportions for positive instances, and their instance scores estimated by MIL. Since the given proportion is only for positive classes and the bag $\boldsymbol{B}^i$ may contain negative instances, selecting only positive instances from the bag is required to compute the proportion loss. As discussed above, to make the positive instance selection differentiable in joint learning and mitigate the adverse effect of MIL errors, we use a soft mask, which gives weight to an instance based on the positive score estimated by MIL, instead of a hard selection.

Specifically, the C-class classifier h outputs the class probabilities $\hat{\mathbf{z}}_j^i = (\hat{z}_{j,1}^i, \ldots, \hat{z}_{j,C}^i) \in \Delta^C$ of the instance $\boldsymbol{x}_j^i$, where $\hat{z}_{j,c}^i = h(f(\boldsymbol{x}_j^i))_c$ is the probability for class c and $\sum_c z_{j,c}^i = 1$. Then, the label proportions $\hat{\boldsymbol{p}}^i \in \Delta^C$ of bag $\boldsymbol{B}^i$ is estimated by the summation of multiplying the pos/neg instance score and the class probability as

follows:

$$\hat{\boldsymbol{p}}^i = \frac{\sum_{j=1}^{|B^i|} \hat{s}_j^i \hat{\mathbf{z}}_j^i}{\sum_{j=1}^{|B^i|} \hat{s}_j^i}. \tag{2}$$

The estimated proportion can be considered as the proportion of the positive classes since the negative instances are ignored by multiplying the binary instance score, which is close to 0 if the MIL estimation is correct. This soft mask may mitigate the adverse effect for the LLP by false positives and negatives in the MIL estimation during training.

The proportion loss between the estimated proportion $\hat{\boldsymbol{p}}^i$ and the ground truth $\boldsymbol{p}^i$ of the positive classes is defined as:

$$\mathcal{L}_{\mathrm{LLP}}(\boldsymbol{p}^i, \hat{\boldsymbol{p}}^i) = -\sum_{c=1}^{C} p_c^i \log \hat{p}_c^i. \tag{3}$$

Simultaneous learning of MIL and LLP. All the modules f, g, h are trained by the entire loss function $\mathcal{L}$ defined by the sum of the losses for LLP and MIL: $\mathcal{L}_{\mathrm{LLP}}(\boldsymbol{p}^i, \hat{\boldsymbol{p}}^i) + w_{\mathrm{MIL}} \, \mathcal{L}_{\mathrm{MIL}}(Y^i, \hat{S}^i)$, where the hyperparameter w_{MIL} is the weight of $\mathcal{L}_{\mathrm{MIL}}$.
Inference. In inference, an instance is classified into $C + 1$ class by the two-stage approach using the trained networks. First, the MIL module classifies an instance into a positive or negative instance. Then, when the instance is classified as positive, the instance is further classified into either of the C positive classes.

5   Experiments

Comparative methods. We evaluated the following two methods as oracles of instance-level supervised learning (SL) and learning from "complete" label proportions: 1) cross-entropy loss (CE), which is a conventional classification method trained by a cross-entropy loss, assuming the class label is given for each instance as training data; and 2) proportion loss (PL), which is a conventional standard method for LLP that uses proportion loss, assuming the proportions for all classes are given as training data.

The existing WSL methods are not feasible for comparing our method directly in the same problem setup as shown in Table 1. We thus prepared two baselines: 3) partial proportions loss (PPL), which is an extension of the proportion loss that calculates loss only for positive classes, i.e., it uses the estimation results of the $(C + 1)$-class instance classifier trained by LLP and ignores the loss for the negative class; and 4) two-stage approach, which combines MIL and LLP. Unlike our method, the MIL is trained independently from LLP. These represent non-trivial extensions of existing methods and serve as strong baselines.
Implementation details. In all methods, the feature extractor f is composed of ResNet 18, and the instance binary classifier g and C-class classifier h are composed of linear layers. The learning rate was set to 0.0003, and the optimizer was Adam. The batch size was set to 16 bags. The model was trained until the validation loss did not update for 30 epochs. The hyperparameter w_{MIL} was set to 0.01. An optimal aggregation function

Table 2. Accuracy and mIoU on pathological image datasets.

			CRC100K		Chemotherapy	
Setting	Given label	Method	Acc.[%] ↑	mIoU[%] ↑	Acc.[%] ↑	mIoU[%] ↑
SL	Instance label	CE	99.22±0.12	97.98±0.34	86.31±2.10	70.62±3.44
LLP	(Comp.) label proportion	PL	99.02±0.14	97.50±0.30	84.53±2.30	63.44±5.19
LPLP	Partial label proportion	PPL	96.29±0.65	90.82±1.50	72.08±6.20	48.01±8.98
		Two-stage	83.96±3.61	61.34±8.11	50.49±6.85	27.26±3.16
		Ours	**98.53±0.39**	**96.32±0.96**	**74.10±3.07**	**52.01±3.77**

in MIL was chosen from mean-pooling, max-pooling, and LSE-pooling using validation data. As performance metrics, we used instance-level accuracy (Acc) and mean intersection over Union (mIoU) [4].

Dataset. To show the effectiveness of our method in pathological image segmentation, we evaluated our method using one public WSI dataset, CRC100K [14], and one private WSI dataset, Chemotherapy. As preliminary experiments, we also evaluated our method using CIFAR10 and SVHN, which is detailed in the supplementary material.

CRC100K contains $100,000$ image patches from WSIs with nine-class annotations. In the experiment, we treated cancer-associated stroma and colorectal adenocarcinoma epithelium as two tumor classes and merged the remaining seven healthy classes as a non-tumor class because the proportions of tumor classes are only given in a real scenario. Each bag comprises 32 samples whose label proportion was randomly determined. The 32 instances were randomly extracted from the dataset according to the label proportion. 500 positive and 500 negative bags were prepared for training and validation. Other 100 positive and 100 negative bags were prepared for the test subset.

In Chemotherapy, there are three classes: non-tumor (negative), tumor-bed (positive), and residual-tumor (positive). The viable tumor rate (the residual-tumor region size divided by the sum of the regions of tumor-bed and residual-tumor) is diagnosed in clinical. It indicates that the proportion of the non-tumor region is unknown, and the proportions of the positive classes are known as the partial proportion labels. The chemotherapy dataset comprises 143 positive WSIs and 10 negative WSIs, which form bags. To show the effectiveness of our method, we prepared the segmentation masks for quantitative evaluation. Patients are separated in training, validation, and testing.

Quantitative evaluation results. Table 2 shows the accuracy and mIoU on the two pathological datasets: CRC100K and chemotherapy. "Setting" and "Given label" indicate the problem scenario and the type of labels used for training, respectively. In CRC100K, despite the challenging LPLP scenario, our method achieved a competitive performance with the oracles (SL and LLP) results, which used more informative labels as training data. In particular, the difference with LLP using "complete" label proportions was slight. Our method can successfully estimate the negative instances by MIL, and thus, the classification performance of positive classes was also achieved at the same level as LLP. Next, our method was evaluated with two baseline methods in the same scenario (LPLP). In both datasets, our method outperformed the baseline methods. Our method was superior to PPL. We consider that the extension of the conventional proportion loss is insufficient for LPLP because it is difficult to train the discriminative
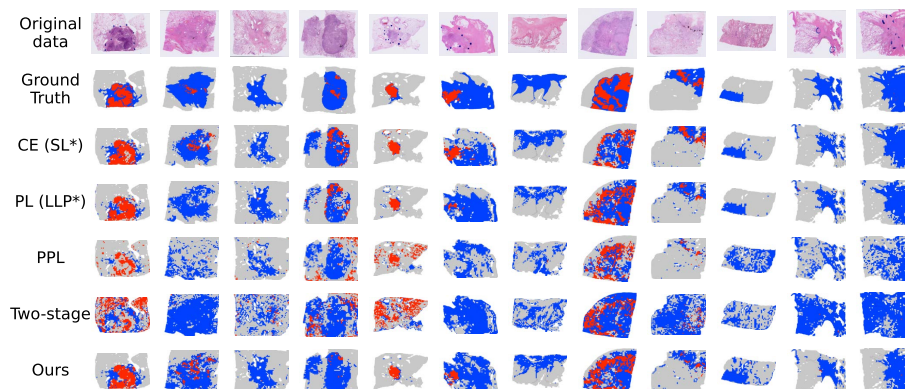
Fig. 3. Segmentation results.

Table 3. Instances-level binary classification results by the standard MIL and our method.

	CRC100K		Chemotherapy	
	Acc.[%] ↑	mIoU[%] ↑	Acc.[%] ↑	mIoU[%] ↑
MIL	85.14±3.78	68.88±7.35	47.09±7.05	26.24±7.22
Ours	**97.98±1.13**	**95.57±2.43**	**69.44±6.58**	**51.16±5.96**

features of negative instances from only the negative bags. Moreover, compared to the two-stage method, it showed the effectiveness of our joint learning of MIL and LLP performed via differentiable positive-instance masks.

Effectiveness of joint learning for the performance of MIL. Table 3 shows the performances of binary classification by MIL in the two-stage method and our method. Our method significantly improved both the Acc and mIoU in two datasets. It indicates that our simultaneous learning is synergistic, i.e., the loss from LLP has a good effect on the feature extractor f and the binary classifier g of MIL.

Qualitative evaluation results. Fig. 3 shows the segmentation results on Chemotherapy : original image, ground truth, segmentation results from 1) CE, 2) PL, 3) PPL, 4) Two-stage, and ours from the top to the bottom row, respectively. Gray indicates non-tumor areas, blue represents tumor-bed, and red signifies residual-tumor. Our method's segmentation results were more accurate than the baseline methods in the LPLP scenario. In addition, our method was competitive with SL and LLP, although a part of the label proportions was not given.

6   Conclusion

This paper first formulated the problem of learning from *partial* label proportions (LPLP) for the WSI segmentation task with the tumor subtype proportions. We then decomposed the LPLP problem into two WSL problems, MIL and LLP, and proposed their joint and end-to-end optimization method. The experimental results indicated that very weak but easily available supervision of the partial label proportions shows a similar

performance to stronger supervision on the CRC100K dataset. Even on the chemotherapy dataset, our method outperformed other possible methods of LPLP.

Disclosure of interests. The authors have no competing interests to declare that are relevant to the content of this article.

References

1. Anklin, V., Pati, P., Jaume, G., Bozorgtabar, B., Foncubierta-Rodríguez, A., Thiran, J.P., Sibony, M., Gabrani, M., Goksel, O.: Learning whole-slide segmentation from inexact and incomplete labels using tissue graphs. In: MICCAI (2021)
2. Ardehaly, E.M., Culotta, A.: Co-training for demographic classification using deep learning from label proportions. In: ICDMW. pp. 1017–1024 (2017)
3. Chan, L., Hosseini, M.S., Rowsell, C., Plataniotis, K.N., Damaskinos, S.: Histosegnet: Semantic segmentation of histological tissue type in whole slide images. In: ICCV (2019)
4. Csurka, G., Larlus, D., Perronnin, F.: What is a good evaluation measure for semantic segmentation? In: CVPR (2013)
5. Dacic, S., Travis, W., Giltnane, J., Abel, J., Kos, F., Hilz, S., Hennek, S., Fujimoto, J., Sholl, L., Khalil, F., Ritter, J., D'Apuzzo, M., Lee, J., Rusch, V., Carbone, D., Nicholas, A., Johnson, A., Schulze, K., Kris, M., Wistuba, I.: Artificial intelligence (AI)–powered pathologic response (PathR) assessment of resection specimens after neoadjuvant atezolizumab in patients with non-small cell lung cancer: Results from the lcmc3 study. Journal of Clinical Oncology **39**, 106–106 (2021)
6. Gabriel, D.A., Neil, Z., Marco, C., Lucas, B., Jean-Philippe, V.: Deep multi-class learning from label proportions. In: arXiv preprint arXiv:1905.12909 (2020)
7. Hashimoto, N., Fukushima, D., Koga, R., Takagi, Y., Ko, K., Kohno, K., Nakaguro, M., Nakamura, S., Hontani, H., Takeuchi, I.: Multi-scale domain-adversarial multiple-instance cnn for cancer subtype classification with unannotated histopathological images. In: CVPR (2020)
8. Javed, S.A., Juyal, D., Padigela, H., Taylor-Weiner, A., Yu, L., Prakash, A.: Additive mil: Intrinsically interpretable multiple instance learning for pathology. In: NeurIPS (2022)
9. Kubo, S., Matsuo, S., Suehiro, D., Terada, K., Ito, H., Yoshizawa, A., Bise, R.: Theoretical proportion label perturbation for learning from label proportions in large bags. In: ECAI (2024)
10. Li, B., Li, Y., Eliceiri, K.W.: Dual-stream multiple instance learning network for whole slide image classification with self-supervised contrastive learning. In: CVPR. pp. 14318–14328 (2021)
11. Liu, J., Wang, B., Qi, Z., Tian, Y., Shi, Y.: Learning from label proportions with generative adversarial networks. In: NeurIPS (2019)
12. Liu, J., Zheng, Q., Mu, X., Zuo, Y., Xu, B., Jin, Y., Wang, Y., Tian, H., Yang, Y., Xue, Q., Huang, Z., Chen, L., Gu, B., Hou, X., Shen, L., Guo, Y., li, Y.: Automated tumor proportion score analysis for pd-l1 (22c3) expression in lung squamous cell carcinoma. Scientific Reports **11** (08 2021)
13. Lv, Z., Yan, R., Lin, Y., Wang, Y., Zhang, F.: Joint region-attention and multi-scale transformer for microsatellite instability detection from whole slide images in gastrointestinal cancer. In: MICCAI (20202)

14. Macenko, M., Niethammer, M., Marron, J.S., Borland, D., Woosley, J.T., Guan, X., Schmitt, C., Thomas, N.E.: A method for normalizing histology slides for quantitative analysis. In: ISBI. pp. 1107–1110 (2009)

15. Matsuo, S., Bise, R., Uchida, S., Suehiro, D.: Learning from label proportion with online pseudo-label decision by regret minimization. In: ICASSP (2023)

16. Moreira, A.L., Ocampo, P.S., Xia, Y., Zhong, H., Russell, P.A., Minami, Y., Cooper, W.A., Yoshida, A., Bubendorf, L., Papotti, M., et al.: A grading system for invasive pulmonary adenocarcinoma: a proposal from the international association for the study of lung cancer pathology committee. Journal of Thoracic Oncology **15**(10), 1599–1610 (2020)

17. Okuo, T., Nishimura, K., Ito, H., Terada, K., Yoshizawa, A., Bise, R.: Proportion estimation by masked learning from label proportion. In: MICCAI Workshop DALI (2023)

18. Qu, L., Wang, M., Song, Z., et al.: Bi-directional weakly supervised knowledge distillation for whole slide image classification. In: NeurIPS (2022)

19. Roach, C., Zhang, N., Corigliano, E., Jansson, M., Toland, G., Ponto, G., Dolled-Filhart, M., Emancipator, K., Stanforth, D., Kulangara, K.: Development of a companion diagnostic pd-l1 immunohistochemistry assay for pembrolizumab therapy in non–small-cell lung cancer. Applied Immunohistochemistry & Molecular Morphology **24**(6), 392 (2016)

20. Shao, Z., Bian, H., Chen, Y., Wang, Y., Zhang, J., Ji, X., et al.: Transmil: Transformer based correlated multiple instance learning for whole slide image classification. In: NeurIPS. pp. 2136–2147 (2021)

21. Silva-Rodríguez, J., Schmidt, A., Sales, M.A., Molina, R., Naranjo, V.: Proportion constrained weakly supervised histopathology image classification. Computers in Biology and Medicine **147**, 105714 (2022)

22. Takanori, A., Matsuo, S., Suehiro, D., Bise, R.: Mixbag: Bag-level data augmentation for learning from label proportions. In: ICCV (2023)

23. Tokunaga, H., Iwana, B.K., Teramoto, Y., Yoshizawa, A., Bise, R.: Negative pseudo labeling using class proportion for semantic segmentation in pathology. In: European Conference on Computer Vision (2020)

24. Tokunaga, H., Teramoto, Y., Yoshizawa, A., Bise, R.: Adaptive weighting multi-field-of-view cnn for semantic segmentation in pathology. In: Computer Vision and Pattern Recognition (2019)

25. Travis, W.D., Dacic, S., Wistuba, I., Sholl, L., Adusumilli, P., Bubendorf, L., Bunn, P., Cascone, T., Chaft, J., Chen, G., et al.: Iaslc multidisciplinary recommendations for pathologic assessment of lung cancer resection specimens after neoadjuvant therapy. Journal of Thoracic Oncology **15**(5), 709–740 (2020)

26. Tsa, K.H., Lin, H.T.: Learning from label proportions with consistency regularization. In: ACML (2020)

27. Wang, X., Yan, Y., Tang, P., Bai, X., Liu, W.: Revisiting multiple instance neural networks. Pattern Recognition **74**, 15–24 (2018)

28. Widmaier, M., Wiestler, T., Walker, J., Barker, C., Scott, M.L., Sekhavati, F., Budco, A., Schneider, K., Segerer, F.J., Steele, K., Rebelatto, M.C.: mparison of continuous measures across diagnostic pd-l1 assays in non-small cell lung cancer using automated image analysis. Modern Pathology **33** (03 2020)

29. Yang, H., Zhang, W., Lam, W.: A two-stage training framework with feature-label matching mechanism for learning from label proportions. In: ACML. pp. 1461–1476 (2021)

30. Yong, S., Jiabin, L., Bo, W., Zhiquan, Q., YingJie, T.: Deep learning from label proportions with labeled samples. In: Neural Networks. pp. 73–81 (2020)