Decoupled Gradient Harmonized Detector for Partial Annotation: Application to Signet Ring Cell Detection

Tiancheng Lin[a,1], Yuanfan Guo[a,1], Canqian Yang[a], Jiancheng Yang[a] and Yi Xu[a,*]

[a] *Shanghai Jiao Tong University, School of Electronic Information and Electrical Engineering*

ARTICLE INFO

Keywords:
Computer-aided Detection
Signet Ring Cell
Partially Annotated Object Detection
DGHM

ABSTRACT

Early diagnosis of signet ring cell carcinoma dramatically improves the survival rate of patients. Due to lack of public dataset and expert-level annotations, automatic detection on signet ring cell (SRC) has not been thoroughly investigated. In MICCAI DigestPath2019 challenge, apart from foreground (SRC region)-background (normal tissue area) class imbalance, SRCs are partially annotated due to costly medical image annotation, which introduces extra label noise. To address the issues simultaneously, we propose Decoupled Gradient Harmonizing Mechanism (DGHM) and embed it into classification loss, denoted as DGHM-C loss. Specifically, besides positive (SRCs) and negative (normal tissues) examples, we further decouple noisy examples from clean examples and harmonize the corresponding gradient distributions in classification respectively. Without whistles and bells, we achieved the 2nd place in the challenge. Ablation studies and controlled label missing rate experiments demonstrate that DGHM-C loss can bring substantial improvement in partially annotated object detection.

1. Introduction

Signet ring cell carcinoma (SRCC) is a form of highly malignant adenocarcinoma. Nuclei push against cell membranes creating a typical signet ring cell (SRC), which has a large mucin vacuole filling the cytoplasm [13]. SRCC tumors are mostly found in stomach and less frequently in breast, bladder and other organs [3]. Since the prognosis for patients with SRCC is extremely poor, early diagnosis and aggressive surgical intervention are of essential importance. However, manual analysis of digital pathology image is labor-intensive and time-consuming, which has become a bottleneck to diagnosis. Therefore, computer-aided technical for SRC detection, as an ancillary study, is promising and highly-demanded.

MICCAI DigestPath2019 is the first challenge and first public dataset on SRC detection. Before this challenge, automatic algorithms on SRC detection have not been thoroughly investigated. However, perfect annotations are practically impossible due to various appearances of SRCs, as well as scattered/overcrowded regions. Thus, the dataset in the challenge is partially annotated and the task is partially annotated object detection (PAOD). As can be seen in Figure 1, in the abnormal pathology (AP) image, green and yellow bounding boxes denote annotated and unannotated SRCs respectively, while normal pathology (NP) image contains no SRCs at all. In conventional detection task [33, 26], there are only positive and negative examples. However, when it comes to PAOD, some positive examples will be incorrectly labeled as negative, which introduces label noise. Directly training with noisy labels will cause adverse impact to the classification accuracy [28, 47].

One solution, proposed by the challenge organizers, is a semi-supervised learning framework for label correction [23], which consists of three steps: initial fully-supervised training, self-training and co-training. Experiments demonstrate that not only is annotation quality improved, extra unlabeled images are also better explored. Instead of pseudo label generation, we tackle PAOD in a loss correction paradigm. The basic idea is to decouple noisy examples from clean examples, disentangling the partial annotation into two parts of noisy-supervised learning and full-supervised learning. However, there are two intrinsic issues: noisy example overfitting and hard example under-learning. Several studies address these two problems respectively. On one hand, Huber loss [14] and GHM-C loss [21] consider hard examples mostly as outliers and reduce their loss contribution. Thus, noisy example overfitting can be prevented, which is suitable for noisy-supervised learning. But they also sacrifice the ability of models to learn hard examples well. On the other hand, online hard example mining [34] and focal loss [26] put more emphasizes on hard examples. Hence, hard example under-learning can be addressed, which is appropriate for full-supervised learning. Unfortunately, these methods further aggravate the overfitting of noisy examples.

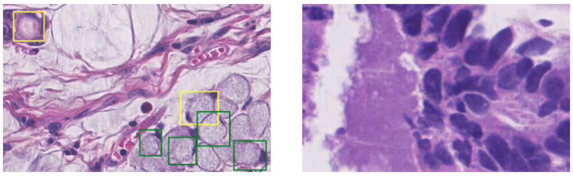
Inspired by GHM [21], we propose Decoupled Gradient Harmonising Mechanism (DGHM) for PAOD, addressing the problems of hard-example under-learning and noisy-example overfitting at the same time. We apply DGHM by decoupling noisy examples from clean examples, harmonizing their gradient distributions separately and managing their outliers differently. For the classification branch of detection, DGHM is embedded into cross entropy (CE) loss, forming DGHM-C loss. Particularly, hard examples in noisy part will be explicitly down-weighted to prevent noisy example overfitting, while the opposite operation is performed on clean part to address hard example under-learning. For the regression branch, since regression loss is only calculated over positive examples, we take smooth-L_1 loss for simplicity.

[*] Corresponding author

ltc19940819@sjtu.edu.cn (T. Lin); gyfastas@sjtu.edu.cn (Y. Guo); charles.young@sjtu.edu.cn (C. Yang); jekyll4168@sjtu.edu.cn (J. Yang); xuyi@sjtu.edu.cn (Y. Xu)

ORCID(s): 0000-0002-5835-8545 (Y. Guo)

[1] These authors are contributed equally to this work



Abnormal Pathology

Normal Pathology

Figure 1: Examples of abnormal and normal pathology image provided in our dataset. Signet ring cell only exists in AP image. The green bounding boxes are provided annotation while the yellow one might be signet ring cell but not annotated.

Experiments show that with the help of DGHM, the performance of PAOD is substantially improved. The main contributions of our work are summarized as follows:

1. We decouple noisy examples from clean examples and disentangle partial annotation into noisy-supervised learning and full-supervised learning.
2. We propose the DGHM and DGHM-C loss for PAOD, making the model robust to noisy examples and focus on hard case of clean examples simultaneously.
3. We conduct comprehensive experiments to demonstrate the effectiveness of DGHM-C loss, which shows an improvement over other loss functions.

2. Related Work

2.1. Object Detection

Object detection is one of the most fundamental and important task in the field of computer vision, which involves classifying and locating the objects. Deep neural networks have made great progress in object detection [33, 25, 11, 10, 5]. Generally, detectors are divided into one-stage and two-stage according to whether to skip the region proposal stage as well as anchor-based and anchor-free depending on whether to use predefined anchors which serve as hypotheses over possible locations of ground-truth bounding boxes. Here, we classify detectors into full-supervised, weakly supervised and noisy-supervised with regard to the quality of labels.

Full-supervised Detection In fully-supervised object detection (FSOD), focal loss [26] and GHM-C loss [21] are proposed to tackle the class imbalance problem in positive/ negative and hard/easy examples. Guided Anchoring [42] leverages semantic features to generate high-quality proposals and boosts the detection performance. Libra RCNN [29] designs an overall balanced learning, consisting in sample level, feature level and object level, which can be generalized to both one-stage and two-stage detectors. However, these methods highly depend on fine-grained instance-level bounding boxes.

Weakly Supervised Detection Weakly supervised object detection (WSOD) requires only image level annotation. WSDNN [4] designs a specific framework with two streams of recognition and detection, and the final proposals are jointly supervised by the two streams. Based on WSDNN, contextual region [19] and multi-stage classifier refinement [37] are introduced for further improvement. Recently, a min-entropy latent model [41] is proposed to discover latent objects and minimize the localization randomness for WSOD with recurrent learning algorithm. Moreover, Multiple Instance Learning (MIL), as a common method for WSOD, is recently further improved in C-MIL [40] by alleviating the non-convexity problem. Some works also use additional annotations or data to improve the performance, e.g. domain adaption [17, 22], semantic knowledge transfer [38] and so on.

Noisy-supervised Detection Different from WSOD, in noisy-supervised detection, models are trained with partial bounding-box annotations, which is called PAOD in this paper. In [9], a self-attention mechanism is adopted to handle the partial annotation problem in Nuclei detection task. A hybrid supervised learning framework [45] is proposed to solve the missing label problem by generating pseudo labels during training. Moreover, soft sampling strategy [44] is established to alleviate the effect of overcrowded false negatives, while giving no guarantee to scattered false negatives. Co-occurrence loss is introduced in [1] to solve the problem of sparsely verified classes, where hierarchical and spatial relationships between multiple classes are exploited to alleviate noisy example overfitting. It imposes the label consistency constraint on the subclass and its parent class, where the performance depends on the high-quality labels of the parent classes. In the DigestPath2019 challenge, however, noisy labels only exist in the class of SRC due to partial annotation in the AP images. Meanwhile, clean labels are available for the annotated SRCs and the NP images. Accordingly, this paper contributes to both reducing the overfitting problem for the noisy labels of SRC in the AP images and the under-learning problem of hard examples with the clean labels.

2.2. Learning with Noisy Examples

Learning with noisy examples is a common situation especially in medical imaging community where manual annotation could be expensive and needs expert knowledge [6]. In the existing literature, the solutions can be roughly classified into two types: label correction and robust loss function.

Label Correction Some previous works tackle the noisy label problem by removing or correcting the noisy labels during training. In [18, 46, 39, 20], auxiliary networks are introduced to correct the missing labels. For example, in [20], a feature encoder that learns the class representative embedding vectors (called "class-prototype") is utilized to decide whether the label is correct or not by comparing with the prototype. MentorNet [18] pre-trains an extra network for clean label selection. In [46], an extra module called MetaCleaner is used to assign scores for samples labeled as same category and hallucinate a clean representation based on the scores. Recently, specific training strategies such as self-training [23, 36, 32] and curriculum learning [12, 7, 27] are proposed for learning under presence of noise. In [23], co-training and self-training are both used to iteratively generate pseudo labels, which shows great performance in SRC detection task. Knowledge distillation is first proposed in [16] and explored in [24] to handle the noisy label problem, which utilizes knowledge learned from an extra clean dataset to generate soft labels for a larger noisy dataset. However, all these methods require either auxiliary network architectures, complex training procedure or additional supervision, which makes it hard to be deployed in real-world applications.

Robust Loss Function Robust loss function has been widely employed in learning with noisy examples. Backward [30] and Forward [30] modify the loss function depending on the noise transition matrix. Nonetheless, the accurate estimation of transition matrix is not promised. Various regularizers, like label smoothing [35] and confidence penalty [31], are introduced to prevent noisy example overfitting at the expense of hard example under-learning. Moreover, Symmetric Cross Entropy (SCE) [43] combines CE with a noise tolerant term, so that both hard example learning and noisy example robustness are improved. However, directly using robust loss in SRC detection is unreasonable, because partial annotation in AP leads to the coexistence of noisy examples and clean examples.

Therefore, we establish DGHM and DGHM-C loss for PAOD. Without any auxiliary, the problems of hard example under-learning and noisy example overfitting can be tackled simultaneously by decoupling noisy examples from clean ones. The proposed DGHM-C loss can be directly applied to SRC detection in MICCAI DigestPath2019 challenge.

3. Method

3.1. Problem Description

In MICCAI DigestPath2019, automatic algorithms are demanded for SRC detection with partial annotation. Using assignment rule from [26], label $p_i^* \in \{0, 1\}$ is assigned to the candidate anchor with index i, where '0' denotes negative examples while '1' represents positive examples. As shown in Figure 2, the green boxes and yellow boxes represent annotated and unannotated SRCs, respectively. The red boxes indicate several candidate anchors, denoted as x_i. Suppose anchor x_1 has an Intersection-over-Union(IoU) overlap with one green box over 0.5, then it is labeled as '1'. In contrast, anchor x_2 is labeled as '0', since their IoU overlap is less than 0.5. The anchors in NP images, i.e. x_4 and x_5, will be all labeled as 0, since there is no SRCs existed. However, though anchor x_3 has an IoU with one yellow box over 0.5, it will be mislabeled as '0'. Actually, yellow box is missed due to partial annotation of SRCs in AP images. Thus it can be seen, for PAOD, labeling anchors only based on IoU is inaccurate for classification. Specifically, label noise will be only introduced to the negative examples of AP images. Both the positive examples of AP images and negative examples spreading over NP images are clean examples with experts' manual annotation.

Fortunately, the noisy anchors can be easily located by introducing an extra attribute into the dataset. Generally, more attributes contribute to better describing the dataset, which may further improve the performance of downstream tasks. Back to the partial annotation problem, we assign image-wise label to each anchor as an extra attribute. By combining the attribute with anchor-wise label, noisy examples can be decoupled from clean examples, as can be seen in Figure 2. Specifically, negative examples from AP images (AP_N) consists of the noisy class, because both correctly labeled anchors such as x_2 and incorrectly labeled anchors such as x_3 exist in these examples, since only a small portion of SRCs are annotated. Meanwhile, positive examples from AP images (AP_P) and negative ones from NP images (NP_N) consists of the clean class. Since all annotated cells are indeed SRCs and NP images contain no SRCs at all.

With the attribute of image-wise label a_i, the training data space can be defined as $S = \{(x_i, p_i^*, a_i)\}_{i=1}^N$, where N is the total number of anchors and p_i^* is the given label for example with index i. Noisy and clean data space can also be obtained by $S_n = \{(x_i, p_i^*, a_i) \mid p_i^* = 0 \text{ and } a_i = 1\}_{i=1}^{N_n}$ and $S_c = \{(x_i, p_i^*, a_i) \mid p_i^* = 1 \text{ or } a_i = 0\}_{i=1}^{N_c}$, subject to $N_n + N_c = N$, $S_n \cap S_c = \varnothing$ and $S_n \cup S_c = S$, where N_n and N_c are the number of noisy and clean anchors. We assume that there is a corresponding ideal label $\hat{p^*}$, thus noisy and clean examples could be described as:

$$P(\hat{p_i^*} \neq p_i^*) = \begin{cases} \eta, & \text{if } x_i \in S_n \\ 0, & \text{if } x_i \in S_c. \end{cases} \tag{1}$$

We denote $P(\hat{p_i^*} \neq p_i^*)$ as the probability that the given label is incorrect where η is label corruption ratio, which is nearly proportional to the number of unannotated cells. For clean anchors the label corruption ratio is 0.

3.2. Classification Loss

Object detection includes two sub tasks: classification and bounding boxes regression[33, 10, 5]. To alleviate the problem mentioned above, our efforts are mainly focused on the classification sub task. As a matter of fact, our method is established based on focal loss and GHM-C loss, which are recently introduced to address the class imbalance from different aspects. Both of them are based on CE loss, which is defined as

$$L_{CE}(p, p^*) = \begin{cases} -\log(p), & p^* = 1 \\ -\log(1-p), & p^* = 0. \end{cases} \tag{2}$$

In eq.(2), $p^* \in \{0, 1\}$ denotes the given ground-truth label and $p \in [0, 1]$ denotes the probability of the class with label $p^* = 1$ according to the model estimation. The gradient norm g is defined as

$$g = |p - p^*| = \begin{cases} 1-p, & p^* = 1 \\ p, & p^* = 0. \end{cases} \tag{3}$$

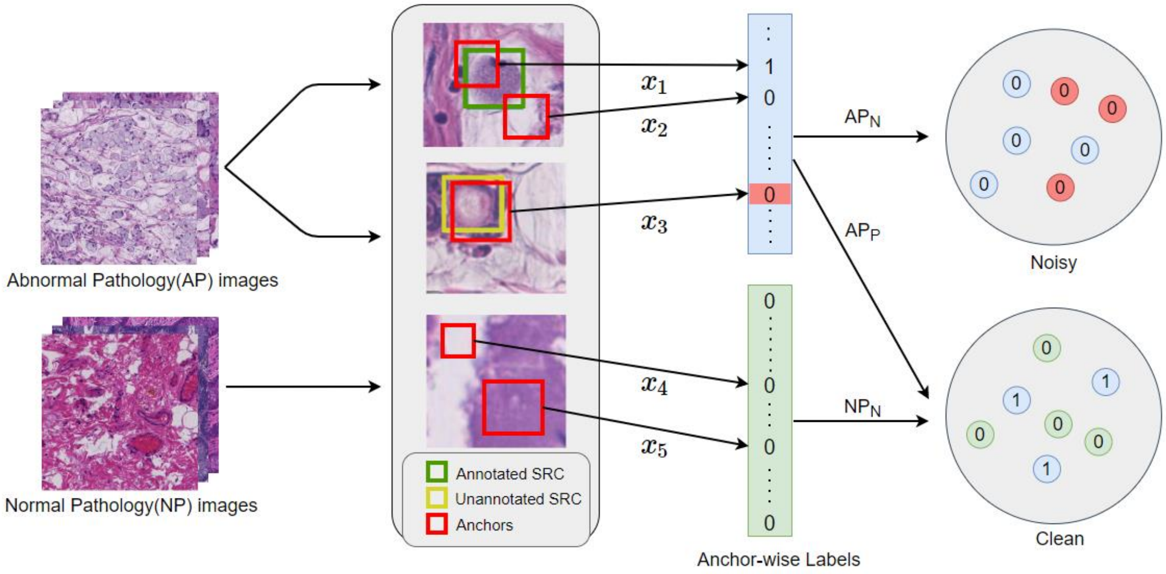


Figure 2: Illustration of noisy labels introduced in partially annotated SRC detection. Since anchors (such as x_3) matching unannotated bounding boxes (yellow) are falsely labeled as negative examples, negative examples in AP images (AP_N) are noisy, where some are actually positive. Since annotated bounding boxes (green) are indeed SRCs and it is certain that no SRCs in NP images, positive examples in AP images (AP_P) and negative examples in NP images(NP_N) are clean examples.

Correspondingly, focal loss can be formulated as

$$L_{focal} = \frac{1}{N}\sum_{i=1}^{N} -\alpha_t g_i^{\gamma} L_{CE}(p_i, p_i^*). \tag{4}$$

Focal loss down-weights the contribution of easy examples by adding a weighting factor α_t and a modulating factor g^{γ}, where $\gamma \geq 0$ is a tunable focusing factor. For notation convenience, α_t can be simplified analogously to g and assigned with $\alpha \in [0, 1]$ for foreground and $1-\alpha$ for background. Besides easy examples, GHM-C loss further down-weights the contribution of outliers regarding to gradient density (GD) [21]. GD denotes the number of examples in specific regions. GD function and the weighted parameters β are defined as

$$GD(g_i) = \frac{1}{l_{\epsilon}(g_i)} \sum_{k=1}^{N} \delta_{\epsilon}(g_k, g_i), \tag{5}$$

$$\beta_i = \frac{N}{GD(g_i)}, \tag{6}$$

where g_k, g_i denote the gradient norm for the example with index k and i respectively and $\delta_{\epsilon}(g_k, g_i)$ is a function indicating whether g_k is located in the region centered at g_i with the valid length $l_{\epsilon}(g_i) = max(1, g_i + \frac{1}{2}\epsilon) - min(0, g_i - \frac{1}{2}\epsilon)$. Thus, GHM-C loss function is

$$L_{GHM-C} = \frac{1}{N}\sum_{i=1}^{N} \beta_i L_{CE}\left(p_i, p_i^*\right). \tag{7}$$

Recently, Symmetric Cross Entropy(SCE) loss is introduced in [43] to handle the hard example under-learning and the noisy example overfitting problem in classification. SCE loss is defined as the combination of CE and reversed CE.

$$L_{SCE}(p, p^*) = \alpha_{SCE} L_{CE}(p, p^*) + \beta_{SCE} L_{CE}(p^*, p), \tag{8}$$

where α_{SCE} and β_{SCE} are weighting factors to balance between CE and the noise-robust term.

3.3. Motivation

Intuitively, we can train a model with only clean examples. However, experimental results suffer from high false positive rate. That is mainly because model is not able to learn how to discriminate background (normal tissue areas) from foreground (SRC) in AP images without supervision on AP_N. In fact, the normal tissue areas in AP images present a much different appearance from those in NP images. Consequently, the diversity of the background will not be fully learned in the training using only a part of dataset. However, if training with both clean and noisy examples, a dilemma arises when applying a unified function such as CE loss, focal loss or GHM-C loss. To illustrate the dilemma, we take a converged model trained with CE loss as an example. Different from [21], we figure out the gradient norm distribution of

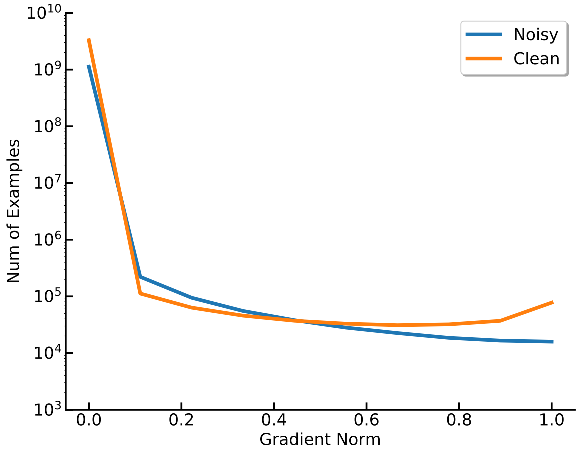


Figure 3: An illustration of the gradient norm distribution of a converged model trained with Cross Entropy loss. The y-axis is in log scale.

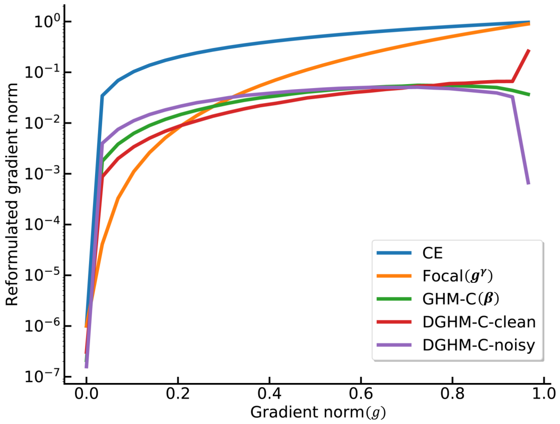


Figure 4: Reformulated gradient norm of different loss functions. The y-axis is in log scale. As shown, DGHM-C down-weights the outliers of noisy examples while up-weights those of the clean ones.

the model in Figure 3, where the noisy and clean examples are counted in two independent histograms. Outliers (examples with large gradient norm) exist in both histograms yet with different meanings. In the histogram of clean examples, outliers are the hard examples. In contrast, the outliers in the histogram of noisy examples are very likely to be mislabeled examples. Moreover, outliers in clean examples are relatively more than that in noisy examples, which indicates that the model under learns the hard examples but overfits the noisy examples. Therefore, we propose DGHM to conquer noisy and clean examples separately, providing two different operations on the outliers corresponding to two kinds of examples.

3.4. DGHM

We propose DGHM for different outlier operations in PAOD. To be specific, we decouple noisy examples from clean ones and calculate their gradient norm distribution respectively, as shown in Figure 3. Then, the GD function is reformulated as

$$GD(g_i) = \begin{cases} \frac{1}{l_\epsilon(g)}(\sum_{k=1}^{N_c} \delta_\epsilon(g_k, g_i)), & x_i \in S_c \\ \frac{1}{l_\epsilon(g)}(\sum_{k=1}^{N_n} \delta_\epsilon(g_k, g_i)), & x_i \in S_n. \end{cases} \tag{9}$$

After calculating the gradient norm in a decoupled way, the gradient density harmonizing parameter is defined as

$$\beta_i = \frac{N}{GD(g_i)^{\gamma_i}}, \tag{10}$$

$$\gamma_i = \begin{cases} \mu_n, & g_i \geq \lambda, x_i \in S_n \\ \mu_c, & g_i \geq \lambda, x_i \in S_c \\ 1, & \text{otherwise}, \end{cases} \tag{11}$$

where outlier threshold $\lambda \in [0, 1]$.

We embed DGHM into CE loss, then DGHM-C loss is formulated as

$$L_{DGHM-C} = \frac{1}{MN} \sum_{i=1}^{N} \beta_i L_{CE}\left(p_i, p_i^*\right), \tag{12}$$

where M is the number of gradient norm distributions. It should be noted that modulating factor $\gamma_i \geq 0$ in eq.(11) can control the contribution of outliters in the DGHM-C loss function eq.(12), i.e. $\mu_n \geq 1$ is chosen to down-weight the outliers in S_n to avoid overfitting the noisy examples, and $\mu_c \leq 1$ is considered to up-weight the outliers in S_c to avoid under-learning of clean examples. After applying harmonizing parameter on gradient norm, the reformulated gradient norm of different losses are shown in Figure 4. It can be seen that the curve of GHM-C loss and two curves of DGHM-C loss, i.e. DGHM-C-clean and DGHM-C-noisy, are of the similar trend before outliers. However, outliers in DGHM-C-clean are relatively up-weighted, which is opposite to GHM-C loss and different from focal loss. On the other hand, outliers in DGHM-C-noisy are explicitly down-weighted, while only slightly down-weighting can be observed in GHM-C loss.

4. Experiment

4.1. Dataset

The dataset is provided by MICCAI Digestpath2019 challenge. A total of 90 patients' 450 pathology images are provided, where 78 AP images with SRC annotation and 372 NP images. The pathology images are from 2 organs, including gastric mucosa and intestine. All pathology images are stained by hematoxylin and eosin (H & E) and scanned at ×40, with the size of 2,000 × 2,000. Each SRC is annotated by a rectangle bounding box tightly surrounding the cell with a total number of 15,000 cells annotated. It should be noted that the SRCs are partially annotated, where pathologists can guarantee that the annotated cells are indeed SRCs. The unannotated regions of AP images are also expected to contain SRCs. It is also guaranteed that NP images contains no SRCs.

4.2. Implement Details

RetinaNet [26] with backbone of ImageNet pre-trained ResNet-18 [15] is used in our experiments. We follow the setting in [26] except the basic areas of anchors, which are tuned for our task from 16^2 to 256^2. All models are trained with Adam optimizer with initial learning rate of 0.0001, and the learning rate is decreased by 0.1 at 9th and 12th epoch. Since the provided images are too large to feed into the memory, sliding windows strategy is used to decompose the original images into patches of 800 × 800 pixels with stride of 300 pixels. The patches are then randomly cropped to 600-pixel width and height, followed by horizontal flip and rotation for data augmentation. Models are trained on a single 1080Ti GPU with mini-batch size of 8. To balance the positive and negative samples during training, 1:3 ratio of positive to negative examples is used when constructing a mini-batch.

The whole available images are randomly split into 5 subsets with around 16 positive images and 72 negative images each fold. We perform 5-fold cross validation in the comparison experiment of different loss functions. For other experiments, performance on a randomly chosen subset is evaluated as the result.

4.3. Evaluation

Recall and precision are usually used for object detection evaluation [8]. In PAOD, however, precision is not reliable because of distinct label noise. Therefore, same as the challenge, instance-level recall, normal region false positives (NFPs) and FROC [2] are used in our experiments with requirement of precision over 0.2. Instance-level recall at IoU of 0.3 is computed among AP images, which is defined as

$$Recall = \frac{TP}{TP + FN}, \tag{13}$$

where *TP* is the number of correctly detected positive examples and *FN* is missed ones. The definition of precision is

$$Precision = \frac{TP}{TP + FP}, \tag{14}$$

where *FP* is the number of negative examples predicted as positive. However, precision is not reliable, because some examples in *TP* are wrongly attributed as *FP* due to partial annotation. Instead, since NP images contain no SRCs, NFPs is considered to measure the wrong predictions of models.

$$NFPs = \max(100 - W, 0), \tag{15}$$

where *W* is the average number of predicted boxes in NP images. To access recall and NFPs comprehensively, we further consider FROC [23], which is defined as average recall given confidence threshold at NFPs in S = [1, 2, 4, 8, 16, 32].

$$FROC = \frac{1}{N} \sum_{i \in S} recall^{NFPs=i}. \tag{16}$$

Moreover, we further verify the recall on training data. In controlled label missing rate experiments, only a part of annotations are used for training and the rest are manually removed. We denote the recall on training and removed annotations as T-recall and R-recall respectively. Intuitively, T-recall reveals whether model has already fit the training data well, while R-recall shows whether unannotated objects can be detected. To some extent, these two metrics can be used to evaluate whether the hard example under-learning and noisy example overfitting are really addressed.

4.4. Ablation Study

For a comprehensive understanding of each term in DGHM-C loss, a series of ablation studies are conducted to assess the sensitivity of outlier operations, i.e. the modulating factor μ_n, μ_c and outlier threshold λ. With $\mu_n = \mu_c = 1$, only decoupling is applied without any outlier operations, which can be considered as the baseline. NFPs are almost unchanged in both Table 1 and Table 2, which is mainly because the classification of NP and AP images is a relatively simple task. Thus, we focus on the recall and FROC, while precision is briefly discussed since the absolute value is meaningless.

Firstly, we evaluate the influence of modulating factors μ_n and μ_c (varying from 1/2.0 to 2.0) with $\lambda = 0.9$ fixed. As shown in Table 1, singly setting $\mu_n = 2.0$ or $\mu_c = 0.5$ both boost the performance of recall and FROC compared to baseline. It quite makes sense since noisy example overfitting is expected to alleviated by down-weighting ($\mu_n > 1.0$) and hard example under-learning is expected to eased by up-weighting ($\mu_c < 1.0$). However, the improvement of the former is limited. That is mainly because that down-weighting is already achieved to some extent for this dataset via GD function even without further modulating ($\mu_n = 1.0$, $\mu_c = 1.0$). We further set $\mu_n > 1.0$ and $\mu_c < 1.0$ at the same time to relieve both problems, and the performance are further improved as expected. The opposite operation ($\mu_n < 1.0$ and $\mu_c > 1.0$) definitely deteriorates the performance. The decrease of precision is mainly because more unannotated SRCs are detected. Thus, $\mu_n > 1.0$ and $\mu_c < 1.0$ is recommended.

Secondly, we assess the effect of outlier threshold by varying λ from 0.7 to 0.9 with $\mu_n = 2.0$, $\mu_c = 0.5$ fixed. As shown in Table 2, recall and FROC increase significantly with λ decreasing. However, precision also decreases dramatically. When $\lambda = 0.7$, precision is even lower than 20%, which can be considered as a failure result. With low λ, the model tends to predict more anchors as positive in AP images. In other words, recall and FROC may be improved at the cost of precision. This is because that, extreme penalties are introduced even when the model is not very confident with the results. Intuitively, model stops from learning early, which causes underfitting. To avoid this situation, a relatively high λ value is selected.

Table 1
Performance comparison with varying μ_n and μ_c ($\lambda = 0.9$). The average performance and standard deviation over 5 different fold are provided.

μ_n	μ_c	Precision	NFPs	Recall	FROC
1.0	1.0	0.4822±0.0667	99.96±0.0233	0.7790±0.0253	0.7716±0.0265
2.0	1.0	0.4756±0.0633	99.98±0.0310	0.7816±0.0342	0.7814±0.0339
1.0	0.5	0.4224±0.0527	99.98±0.0221	0.8292±0.0228	0.8134±0.0274
0.5	2.0	**0.4950±0.0527**	**99.98±0.0172**	0.7590±0.0258	0.7538±0.0173
1/1.5	1.5	0.4828±0.0722	99.92±0.0946	0.7742±0.0371	0.7706±0.0410
1.5	1/1.5	0.4534±0.0698	99.95±0.0620	0.8004±0.0300	0.7962±0.0277
2.0	0.5	0.4229±0.0658	99.92±0.0459	**0.8294±0.0251**	**0.8189±0.0337**

Table 2
Performance comparison with varying λ ($\mu_n = 2.0$, $\mu_c = 0.5$). The performance on a random subset is presented.

λ	Precision	NFPs	Recall	FROC
0.7	0.0844	99.69	**0.9810**	**0.9796**
0.8	0.2888	**99.99**	0.9111	0.9045
0.9	0.3568	99.97	0.8732	0.8732

We take $\mu_n = 2.0$, $\mu_c = 0.5$ and $\lambda = 0.9$ as our default settings, which is the same as the challenge submission. It's worth noting that performance could be further improved by tuning hyper parameters. For example, the modulating factors for outliers in S_c and S_n are not necessarily to be reciprocal relationship. However, it's not our main concern. Under the current settings, $\mu_n > 1$ improves recall and FROC, and precision is guaranteed by higher λ, from which the recall-precision trade off can be easily obtained. Other hyper parameters like number of unit region are chosen to be 10 for low computational complexity, which has been discussed in [21].

4.5. Comparing with Other Losses

In this experiment, the same network under the same setting is trained with different losses for classification branch. We compare DGHM-C loss with CE loss, focal loss, GHM-C loss and SCE loss, where CE loss is considered as the baseline. As can be seen in Table 3, focal loss deteriorates the performance compared with baseline. This is mainly because noisy examples are mistaken as hard examples, resulting in noisy example overfitting. On the contrary, GHM-C loss improves recall and FROC by nearly 15%, which largely benefits from down-weighting the outliers. For SCE loss, two hyper parameters $\alpha_{SCE} = 0.01$ and $\beta_{SCE} = 1.0$ are used to ease overfitting as in [43]. Nonetheless, SCE loss brings no benefits compared with baseline even in NFPs. This is likely because that, good performance of SCE is not guaranteed under asymmetric noise in PAOD. However, when applying DGHM-C loss, the best performance of NFPs, recall and FROC are achieved, where The decline of precision is also acceptable as discussed in section 4.4. The substantial improvement of performance is mainly because DGHM-C loss facilitates the adaptive outlier operations for examples in data space of clean(S_c) and noisy(S_n). The problem of hard example under-learning is alleviated by up-weighting the outliers in S_c to encourage learning, while the problem of noisy example overfitting is addressed by down-weighting the outliers in S_n to enhance noise tolerance. This can also be proved by qualitative result in Figure 5. Furthermore, we have also tried applying different losses to noisy and clean examples respectively, such as combination of focal/CE loss with GHM-C loss. It is found that tuning weighting factor of different losses is quite trivial since their magnitude are distinct. The convergence of model is not guaranteed. In contrast, models trained with DGHM-C loss also yield better convergence.

4.6. Controlled Label Missing Rate

To investigate the effect of missing label rate η, we randomly discard the provided annotation by rate of [20%, 70%]. For simplicity, we assume the provided annotation is perfect and take η as the annotation drop rate. The performance of different methods is listed in Table 4. As can be seen, our method achieves outstanding recall from 0.6917 to 0.8666 and FROC from 0.6720 to 0.8630 under different η, which outperforms other methods by a significant margin. Furthermore, when changing η from 20% to 70%, for model trained with focal loss, CE loss and SCE loss, the FROC performance significantly drop by nearly 35%. Though GHM-C loss shows a better tolerance to noise, it still suffers a 25% drop in FROC performance. However, a 20% drop is obtained by DGHM-C loss, which is a much smaller gap compared with others. The same behavior is also observed for recall.

As shown in Table 4, R-recall of our DGHM-C loss surpasses other methods by over 15% under varying η. In the extreme case of dropping 70% of annotations, our method registers nearly 30% improvement (0.6872 vs. 0.3872) compared with the closest competitor (GHM-C), from which we can conclude that our model is more robust to label noise even when η is large. As for T-recall, though it is a relatively simple task, 3% to 10% enhancement are achieved by DGHM-C loss, which indicates our model can learn hard examples well. When considering R-recall and T-recall together, one can find CE loss, focal loss and SCE loss are all with relatively high T-recall but low R-recall. (The R-recall of 0.6590 for SCE under $\eta = 40\%$ is considered as an out-

Table 3
Quantitative results of RetinaNet with different loss functions. The average performance and standard deviation of 5 different fold is presented.

Method	Precision	NFPs	Recall	FROC
CE	0.6088±0.0741	99.37±0.2059	0.5960±0.0258	0.5931±0.0251
Focal	0.6556±0.0791	99.26±0.4685	0.5086±0.0197	0.5027±0.0214
GHM-C	0.4968±0.0678	99.72±0.1867	0.7565±0.0396	0.7429±0.0395
SCE	0.5926±0.0703	97.90±0.5922	0.5779±0.0191	0.5594±0.0178
DGHM-C	0.4229±0.0658	**99.92**±0.0459	**0.8294**±0.0251	**0.8188**±0.0337

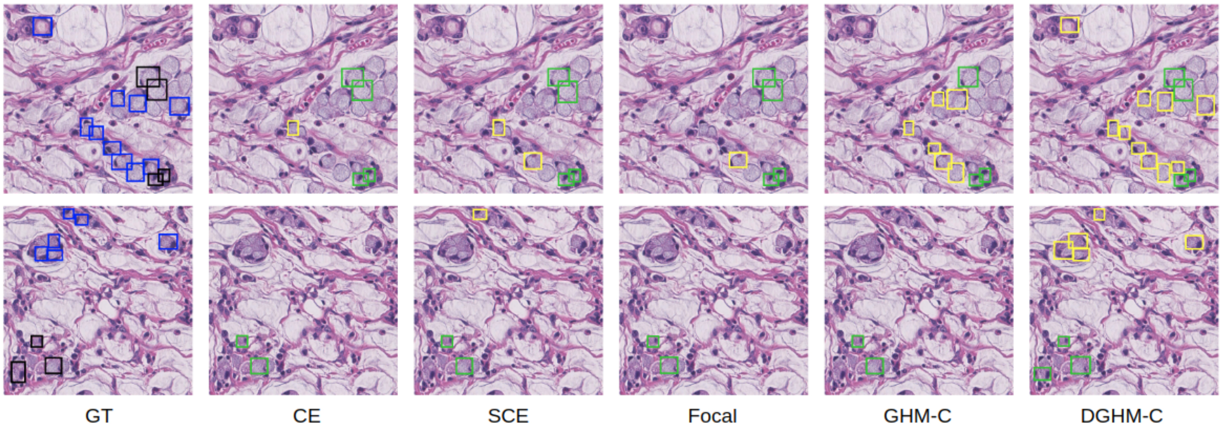


Figure 5: Qualitative comparison of our method and four references(GHM-C loss, focal loss, CE loss and SCE loss) under 70% missing label rates. The first column shows the ground truth, where black bounding boxes are training annotations and the blue ones are removed annotations. From second column, green bounding boxes indicate annotated SRC correctly detected by models, while the yellow ones correspond to unannotated SRC detected.

lier.) This is mainly because that they suffer from noisy example overfitting. Though relative high R-recall is acquired by GHM-C loss, the T-recall is unstable, e.g. T-recall is 0.6446 when $\eta = 70\%$. It seems that the noise robustness is achieved at cost of hard example under-learning. Only through DGHM-C loss, both good performance of T-recall and R-recall can be accomplished.

Moreover, we present a few qualitative results in Figure 5 for models trained with $\eta = 70\%$. As can be seen, the left bottom SRC in figures of the second row is missed by other methods, indicating it might be a hard example, while our method successfully detect it. Moreover, our method captures more removed SRCs (yellow bounding boxes), illustrating the robustness of our model to partial annotation.

5. Discussion

5.1. Decoupling More

Though great performance is achieved by decoupling noisy examples from clean ones, we believe that the optimal decoupling method is hard to define. As an additional research, we further decouple clean positive examples AP_P and clean negative examples NP_N, applying similar DGHM with $\mu_n = 2.0$ for both AP_P and NP_N. We denote it as DGHM-C*. The comparison result is shown in Table 5. We observe that DGHM-C* achieves a 3% improvement on recall and FROC compared to DGHM-C. To analyze the potential reason for the performance boost, we illustrate the according gradient norm distributions in Figure 6. As can be seen, a large parts of examples from NP_N show a very low gradient norm, which is consistent with the high NFPs in above experiments. It indicates that learning to distinguish NP images from abnormal ones is relatively easy for model. If the gradient density is shared between positive and negative examples, their loss contribution are the same. However, if the gradient density is calculated separately, the loss contribution of examples from AP_P are larger than that from NP_N, even the same whole gradient norm is unchanged. The model is forced to learn to correctly classify positive from negative. From empirical understanding, more delicate decoupling methods potentially boost DGHM, resulting in better performance. Still, computational complexity should be taken into consideration.

5.2. Recall/Precision Trade-off

A trade-off between recall and precision can be observed in DGHM-C loss. Actually, we can manually settle the recall/precision trade-off in the way of tuning the hyper parameters of the modulating factor μ_n, μ_c and outliers threshold λ. However, a model is hard to simultaneously achieves

Table 4
Performance comparison under different missing label rates on a single subset. Recall, precision, FROC and NFPs metrics are evaluated on full-annotated test set.

Method	Precision						NFPs					
	20%	30%	40%	50%	60%	70%	20%	30%	40%	50%	60%	70%
CE	0.5803	0.6096	0.6020	0.6211	0.6445	0.6117	99.54	99.77	99.73	99.59	99.81	99.67
Focal	0.6077	0.6167	0.6359	0.6074	0.6552	0.5878	99.53	99.64	99.63	99.46	99.62	99.67
GHM-C	0.4795	0.4992	0.5539	0.4949	0.6267	0.6867	99.91	99.97	**100.0**	99.92	100.0	**100.0**
SCE	0.5428	0.5835	0.5818	0.5893	0.5933	0.6006	97.97	98.95	98.51	98.67	99.45	99.47
DGHM-C	0.3785	0.3155	0.3702	0.4071	0.4609	0.4229	**99.91**	**99.98**	99.99	**99.96**	**100.0**	99.96
	Recall						FROC					
CE	0.5532	0.4519	0.387	0.3513	0.2682	0.1902	0.5255	0.4490	0.3870	0.3513	0.2682	0.1866
Focal	0.4585	0.3870	0.2762	0.2289	0.1939	0.1050	0.4541	0.3856	0.2719	0.2172	0.1837	0.1050
GHM-C	0.6642	0.6983	0.6480	0.6727	0.5138	0.4089	0.6545	0.6968	0.6480	0.6669	0.5138	0.4089
SCE	0.5175	0.4278	0.4249	0.3440	0.2340	0.1545	0.4928	0.4253	0.4147	0.3352	0.2340	0.1505
DGHM-C	**0.8666**	**0.8476**	**0.8105**	**0.7682**	**0.7128**	**0.6917**	**0.8630**	**0.8331**	**0.7922**	**0.7412**	**0.7128**	**0.6720**
	R-recall						T-recall					
CE	0.3612	0.2353	0.2131	0.1845	0.1477	0.1036	0.9105	0.9033	0.9078	0.9070	0.8989	0.9054
Focal	0.2268	0.1431	0.1029	0.0925	0.0597	0.0397	0.8932	0.8843	0.8821	0.8865	0.8844	0.8496
GHM-C	0.7556	0.7044	0.6547	0.6754	0.5233	0.3872	0.8341	0.7525	0.8616	0.7194	0.8509	0.6446
SCE	0.4607	0.3725	0.6590	0.2595	0.2002	0.1117	0.8960	0.8816	0.8899	0.8851	0.8736	0.8544
DGHM-C	**0.8560**	**0.8693**	**0.8067**	**0.7761**	**0.6836**	**0.6872**	**0.9340**	**0.9903**	**0.9935**	**0.9887**	**0.9876**	**0.9900**

Table 5
Performance difference between DGHM-C and DGHM-C*. The average performance and standard deviation of 5 different fold are reported. For both methods, we set hypermeters $\mu_n = 2$, $\mu_c = 0.5$ and $\lambda = 0.9$.

Method	Precision	NFPs	Recall	FROC
DGHM-C	0.4229±0.0658	99.92±0.0455	0.8294±0.0252	0.8188±0.0338
DGHM-C*	0.3722±0.0683	**99.96±0.0249**	**0.8533±0.0366**	**0.8494±0.0372**

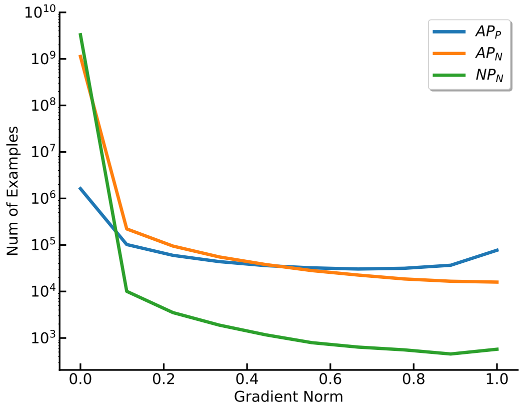


Figure 6: Gradient norm distribution of a converged model on AP_P, AP_N and NP_N. The y-axis is in log scale.

both high recall and good precision in above experiments. The main reason as mentioned before is that SRCs are partially annotated, some correctly detected SRCs are counted as false positives. Therefore, the measured precision is not accurate, which is lower than the real one. Besides, bounding boxes regression could be another reason. Though unannotated SRCs may be detected in classification by DGHM-C loss, their regression are never trained, which leads to localization failure during non-maximum suppression. It will be our future work to improve recall and precision at the same time.

6. Conclusion

In this paper, we formulate the partial annotation problem as an integrated task of noisy-supervised learning and full-supervised learning. To solve the problem, we propose a novel Decoupled Gradient Harmonizing Mechanism (DGHM) and embed it into classification loss (DGHM-C). Thus, both hard example under-learning and noisy example overfitting are addressed simultaneously. Experiments show that our method surpasses other baselines on MICCAI DigesPath2019 challenge. We further demonstrate the efficacy of our method by reporting an extensive controlled missing rates experiment. In the future we would like to explore the decoupling mechanism and regression branch of PAOD.

References

[1] Akiba, T., Kerola, T., Niitani, Y., Ogawa, T., Sano, S., Suzuki, S., 2018. Pfdet: 2nd place solution to open images challenge 2018 object detection track. arXiv:1809.00778.

[2] Bandos, A.I., Rockette, H.E., Song, T., Gur, D., 2009. Area under the free-response roc curve (froc) and a related summary index. Biometrics 65, 247–256.

[3] Belli, S., Aytac, H.O., Karagulle, E., Yabanoglu, H., Kayaselcuk, F., Yildirim, S., 2014. Outcomes of surgical treatment of primary signet ring cell carcinoma of the colon and rectum: 22 cases reviewed with literature. International surgery 99, 691–698.

[4] Bilen, H., Vedaldi, A., 2016. Weakly supervised deep detection networks, in: The IEEE Conference on Computer Vision and Pattern Recognition (CVPR).

[5] Dai, J., Li, Y., He, K., Sun, J., 2016. R-fcn: Object detection via region-based fully convolutional networks, in: Advances in neural information processing systems, pp. 379–387.

[6] Dgani, Y., Greenspan, H., Goldberger, J., 2018. Training a neural network based on unreliable human annotation of medical images, in: 2018 IEEE 15th International Symposium on Biomedical Imaging (ISBI 2018), IEEE. pp. 39–42.

[7] Durand, T., Mehrasa, N., Mori, G., 2019. Learning a deep convnet for multi-label classification with partial labels, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 647–657.

[8] Everingham, M., Van Gool, L., Williams, C.K., Winn, J., Zisserman, A., 2010. The pascal visual object classes (voc) challenge. International journal of computer vision 88, 303–338.

[9] Feng, L., Song, J.H., Kim, J., Jeong, S., Park, J.S., Kim, J., 2019. Robust nucleus detection with partially labeled exemplars. IEEE Access 7, 162169–162178. URL: http://dx.doi.org/10.1109/ACCESS.2019.2952098, doi:10.1109/access.2019.2952098.

[10] Girshick, R., 2015. Fast r-cnn, in: Proceedings of the IEEE international conference on computer vision, pp. 1440–1448.

[11] Girshick, R., Donahue, J., Darrell, T., Malik, J., 2014. Rich feature hierarchies for accurate object detection and semantic segmentation, in: Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 580–587.

[12] Guo, S., Huang, W., Zhang, H., Zhuang, C., Dong, D., Scott, M.R., Huang, D., 2018. Curriculumnet: Weakly supervised learning from large-scale web images, in: Proceedings of the European Conference on Computer Vision (ECCV), pp. 135–150.

[13] Hamilton, S.R., 2000. World health organization classification of tumours. Pathology and genetics of tumours of the digestive system .

[14] Hastie, T., 2008. Tibshirani, r. and friedman, j.(2009): The elements of statistical learning. data mining, inference, and prediction.

[15] He, K., Zhang, X., Ren, S., Sun, J., 2015. Deep Residual Learning for Image Recognition URL: http://arxiv.org/abs/1512.03385, arXiv:1512.03385.

[16] Hinton, G., Vinyals, O., Dean, J., 2015. Distilling the knowledge in a neural network. arXiv preprint arXiv:1503.02531 .

[17] Hoffman, J., Guadarrama, S., Tzeng, E.S., Hu, R., Donahue, J., Girshick, R., Darrell, T., Saenko, K., 2014. Lsda: Large scale detection through adaptation, in: Advances in Neural Information Processing Systems, pp. 3536–3544.

[18] Jiang, L., Zhou, Z., Leung, T., Li, L.J., Fei-Fei, L., 2017. Mentornet: Learning data-driven curriculum for very deep neural networks on corrupted labels. arXiv preprint arXiv:1712.05055 .

[19] Kantorov, V., Oquab, M., Cho, M., Laptev, I., 2016. Contextlocnet: Context-aware deep network models for weakly supervised localization, in: European Conference on Computer Vision, Springer. pp. 350–365.

[20] Lee, K.H., He, X., Zhang, L., Yang, L., 2018. CleanNet: Transfer Learning for Scalable Image Classifier Training with Label Noise. Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition , 5447–5456doi:10.1109/CVPR.2018.00571, arXiv:1711.07131.

[21] Li, B., Liu, Y., Wang, X., 2019a. Gradient harmonized single-stage detector, in: Proceedings of the AAAI Conference on Artificial Intelligence, pp. 8577–8584.

[22] Li, D., Huang, J.B., Li, Y., Wang, S., Yang, M.H., 2016. Weakly supervised object localization with progressive domain adaptation, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 3512–3520.

[23] Li, J., Yang, S., Huang, X., Da, Q., Yang, X., Hu, Z., Duan, Q., Wang, C., Li, H., 2019b. Signet ring cell detection with a semi-supervised learning framework. CoRR abs/1907.03954. URL: http://arxiv.org/abs/1907.03954, arXiv:1907.03954.

[24] Li, Y., Yang, J., Song, Y., Cao, L., Luo, J., Li, L.J., 2017. Learning from Noisy Labels with Distillation. Proceedings of the IEEE International Conference on Computer Vision 2017-Octob, 1928–1936. doi:10.1109/ICCV.2017.211, arXiv:1703.02391.

[25] Lin, T.Y., Dollár, P., Girshick, R., He, K., Hariharan, B., Belongie, S., 2017a. Feature pyramid networks for object detection, in: Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 2117–2125.

[26] Lin, T.Y., Goyal, P., Girshick, R., He, K., Dollár, P., 2017b. Focal loss for dense object detection, in: Proceedings of the IEEE international conference on computer vision, pp. 2980–2988.

[27] Luo, B., Feng, Y., Wang, Z., Zhu, Z., Huang, S., Yan, R., Zhao, D., 2017. Learning with noise: Enhance distantly supervised relation extraction with dynamic transition matrix. arXiv preprint arXiv:1705.03995 .

[28] Nettleton, D.F., Orriols-Puig, A., Fornells, A., 2010. A study of the effect of different types of noise on the precision of supervised learning techniques. Artificial intelligence review 33, 275–306.

[29] Pang, J., Chen, K., Shi, J., Feng, H., Ouyang, W., Lin, D., 2019. Libra r-cnn: Towards balanced learning for object detection, in: The IEEE Conference on Computer Vision and Pattern Recognition (CVPR).

[30] Patrini, G., Rozza, A., Krishna Menon, A., Nock, R., Qu, L., 2017. Making deep neural networks robust to label noise: A loss correction approach, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 1944–1952.

[31] Pereyra, G., Tucker, G., Chorowski, J., Kaiser, Ł., Hinton, G., 2017. Regularizing neural networks by penalizing confident output distributions. arXiv preprint arXiv:1701.06548 .

[32] Reed, S., Lee, H., Anguelov, D., Szegedy, C., Erhan, D., Rabinovich, A., 2014. Training deep neural networks on noisy labels with bootstrapping. arXiv preprint arXiv:1412.6596 .

[33] Ren, S., He, K., Girshick, R., Sun, J., 2015. Faster r-cnn: Towards real-time object detection with region proposal networks, in: Advances in neural information processing systems, pp. 91–99.

[34] Shrivastava, A., Gupta, A., Girshick, R., 2016. Training region-based object detectors with online hard example mining, in: Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 761–769.

[35] Szegedy, C., Vanhoucke, V., Ioffe, S., Shlens, J., Wojna, Z., 2016. Rethinking the inception architecture for computer vision, in: Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 2818–2826.

[36] Tanaka, D., Ikami, D., Yamasaki, T., Aizawa, K., 2018. Joint optimization framework for learning with noisy labels, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 5552–5560.

[37] Tang, P., Wang, X., Bai, X., Liu, W., 2017. Multiple instance detection network with online instance classifier refinement, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 2843–2851.

[38] Tang, Y., Wang, J., Gao, B., Dellandréa, E., Gaizauskas, R., Chen, L., 2016. Large scale semi-supervised object detection using visual and semantic knowledge transfer, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 2119–2128.

[39] Veit, A., Alldrin, N., Chechik, G., Krasin, I., Gupta, A., Belongie, S., 2017. Learning from noisy large-scale datasets with minimal supervision, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 839–847.

[40] Wan, F., Liu, C., Ke, W., Ji, X., Jiao, J., Ye, Q., 2019. C-mil: Continuation multiple instance learning for weakly supervised object detection, in: The IEEE Conference on Computer Vision and Pattern

Recognition (CVPR).

[41] Wan, F., Wei, P., Jiao, J., Han, Z., Ye, Q., 2018. Min-entropy latent model for weakly supervised object detection, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 1297–1306.

[42] Wang, J., Chen, K., Yang, S., Loy, C.C., Lin, D., 2019a. Region proposal by guided anchoring, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 2965–2974.

[43] Wang, Y., Ma, X., Chen, Z., Luo, Y., Yi, J., Bailey, J., 2019b. Symmetric cross entropy for robust learning with noisy labels, in: Proceedings of the IEEE International Conference on Computer Vision, pp. 322–330.

[44] Wu, Z., Bodla, N., Singh, B., Najibi, M., Chellappa, R., Davis, L.S., 2018. Soft sampling for robust object detection. arXiv preprint arXiv:1806.06986 .

[45] Xu, M., Bai, Y., Ghanem, B., Liu, B., Gao, Y., Guo, N., Ye, X., Wan, F., You, H., Fan, D., et al., 2019. Missing labels in object detection, in: The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops.

[46] Zhang, W., Wang, Y., Qiao, Y., 2019. Metacleaner: Learning to hallucinate clean representations for noisy-labeled visual recognition, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 7373–7382.

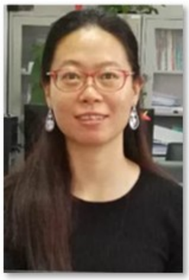
[47] Zhu, X., Wu, X., 2004. Class noise vs. attribute noise: A quantitative study. Artificial intelligence review 22, 177–210.

Tiancheng Lin received the B.S. degree from East China Normal University in 2016 and M.S. degree from Shanghai Jiao Tong University in 2019.He is working towards the Ph.D. degree of Artificial Intelligent Lab with Shanghai Jiao Tong University. His research interests are deep learning, with emphasis on medical image analysis.

Yuanfan Guo is working towards the B.S. degree of Shanghai Jiao Tong Univiersity. His research interests are deep learning, with emphasis on medical image analysis.

Canqian Yang is working towards the B.S. degree of Shanghai Jiao Tong University. His research interests are deep learning, with emphasis on medical image analysis and image super resolution.

Jiancheng Yang received the B.Eng. and M.Eng. degree in Automation from Shanghai Jiao Tong University, China, in 2015 and 2018, respectively. He also received engineer degree (double master degree) in Institute Mines-Telecom, France, in 2016. He is working towards the Ph.D. degree with Shanghai Jiao Tong University. His research interests are deep learning, with emphasis on medical image analysis, 3D computer vision, data efciency and interpretability in AI.



Yi Xu received the B.S. and M.S. degrees from the Nanjing University of Science and Technology, Nanjing, China, in 1996 and 1999, respectively, and the Ph.D. degree from Shanghai Jiao Tong University, Shanghai, China, in 2005. She is currently an Associate Professor with the Department of Electronic Engineering, Institute of Image Communication and Network Engineering, Shanghai Jiao Tong University. Her research interests include image processing, intelligent video analysis, and quaternion wavelet theory.