Tendentious Noise-rectifying Framework for Pathological HCC Grading

Xiaotian Yu*[1]
yuxiaotian@zju.edu.cn
Zunlei Feng*[1]
zunleifeng@zju.edu.cn
Mingli Song, Yuexuan Wang[1]
brooksong,amywang@zju.edu.cn
Xiuming Zhang[13]
1508056@zju.edu.cn
Thomas Li[2]
liktt@hku.hk

[1] Zhejiang University
Zhejiang, China
[2] University of Hong Kong,
Hong Kong, China
[3] The First Affiliated Hospital,
College of Medicine,
Zhejiang University,
Zhejiang, China

Abstract

Hepatocellular carcinoma (HCC) is one of the most common cancers with high mortality. In clinic, pathology grade assessment is laborious work with high divergence and misdiagnosis rate. Many computer-assisted grading methods were proposed to increase the objectivity of diagnosis and reduce workload. With massive accurate annotation, deep learning based methods have achieved promising results in tumor grading. However, annotating pathology images is very time-consuming and hard to be precise. The inevitable noises caused by manual annotation have been ignored by existing tumor grading methods, which leads to serious performance degradation. In this paper, we propose a Tendentious Noise-rectifying Framework (TNF) for HCC grading on pathology images with noisy annotations. A fundamental way to reduce the negative impact of those noisy data is finding and rectifying those noises. So, we devise a noise-rectifying loss to rectify those noisy labels with high confidence. The rectifying tendency is dynamically adjusted by the feature polymer that contains structural information of a large local area of pathology image. We collect 415 hepatocellular biopsy slides and crop 20,000 patches as the dataset. Exhaustive experiments on this dataset demonstrate that, with the noise-rectifying loss, TNF achieves state-of-the-art performance and finds out the carcinoma cells in the healthy area, which has significant meaning for the diagnosis of HCC.

1 Introduction

Hepatocellular carcinoma (HCC) is one of the most commonly diagnosed cancer with the second highest mortality rate among men [3]. Clinically, pathology slides are regarded as the 'gold standard' and can provide better evidence for diagnosing the disease than CT and X-ray samples. The pathological grading of HCC has significant implications for prognostic

*They are equally contributed and sorted in alphabetical order.

[19]. However, the grading results can be highly subjective and susceptible to misdiagnosis, and the process is also time-consuming. So, it's meaningful to develop an accurate and objective method for the HCC grading in pathology images.

Recently, deep learning approaches have achieved promising results in the medical image analysis area, such as pneumonia detection [11, 23], diabetic retinopathy detection [12], breast masses segmentation [7, 14], etc. Unlike these examples, the pathology slides of HCC have some unique characteristics. First of all, the size of the Whole-slide Images (WSIs) in the collected HCC dataset is up to 200,000 pixels wide, which is much larger than images of regular datasets. Since the diagnosis mainly bases on cellular level view and downsampling of the whole slides will lose critical information, the slides need to be cropped into small patches before inputting to the model. In this condition, a small deviation of the boundary will lead to many mislabeled patches. The ultra-large size makes the annotating very complicated and results in a particularly high noise rate. Secondly, although in most slides, cells of the same HCC slides have homogeneous grades, there still contains regions mixed of carcinoma cells and normal cells. Since it's impractical to check out every small area because of the large size of the WSIs, some components like blood vessels located in the tumor region or small clusters of carcinoma cells outside the tumor will be easily neglected. Those unique characteristics bring the inevitable noisy annotations, which leads to serious performance degradation and overfitting of existing methods [1, 17].

In order to have a clear understanding of this dataset, a sample of HCC slide is shown in Fig. 1. Here the tumor in this slide belongs to the grade 2, so all the patches located within the tumor region (blue circle) are annotated as grade 2 (cancerous) and the others are annotated as grade 0 (healthy). Besides, considering the potential errors, all patches have the reversed labels (healthy $\rightarrow$ cancerous, cancerous $\rightarrow$ healthy), which provides the opportunity for rectifying mislabeled samples.

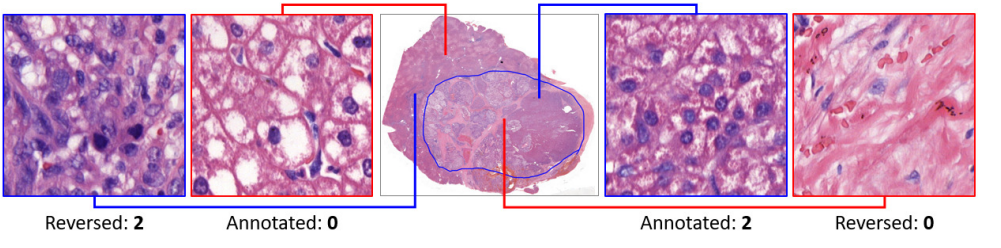

Figure 1: Illustration of 50x patches in a HCC slide. The blue circle in the middle figure is the annotation of the tumor. In this slide the tumor belongs to the grade 2. All patches cropped from tumor region are annotated as grade 2 and the reversed labels are grade 0. Labels of patches cropped outside this region are opposite. Some examples of the ground truth being consistent with the annotated or reversed labels are also shown.

In view of HCC samples' characteristics, the structural information can be utilized to rectify the mislabeled patches. We propose the Tendentious Noise-rectifying Framework (TNF) for HCC grading with noisy annotations. TNF takes two inputs: small-scaled patches and corresponding large-scaled feature polymers, representing the cellular information and the surrounding structure of each patch. Besides, noise-rectifying (NR) loss is proposed with two optimizing targets toward the annotated and reversed labels. The tendencies to these two labels in NR is determined by the surrounding structure extracted from feature polymer. The NR loss and feature polymer make the combined effect of rectifying the influence of the

noise. Moreover, an HCC dataset is collected, containing 20,000 patches from 415 WSIs with rough annotations by pathologists. Exhaustive experiments demonstrate the excellent performance of TNF on this dataset.

In summary, our main contribution is the first rectifying noise by integrating cellular and structural information for HCC grading. Also, The NR loss and feature polymer are designed according to the essential cause of the noise. With the combined effect of these two components, the model will be optimized selectively on the annotated or reversed labels, lessening the overfitting of incorrect labels.

2 Related Work

In this section, we survey two most related areas: computer-aided diagnosis on tumor and noisy label learning.

Computer-aided Tumor Diagnosis. For existing works aiming at tumor diagnosis, some researchers employed traditional machine learning methods such as random forest [16, 21] and support vector machine [9]. But the performance of these methods highly relies on the quality of hand-crafted features. Aiming at gigapixel histopathology images, Tellez et al. [28] applied neural image compression to map the whole-slide to low-dimensional embedding vectors for image-level classification. For patch-level prediction, Xu et al. [30] proposed CAMEL based on multiple-instance learning (MIL), which selected the most credible samples as clean data to retrain the model. Chikontwe et al. [4] proposed an end-to-end framework to combine instance-based MIL and embedding-based MIL. Diao et al. [8] added some artificial samples as negative ones for improving the model's generalization. In essence, MIL and artificial images merely reduce the weight of mislabeled patches, which cannot fundamentally solve the noise problem. So, the performance is still limited.

Learning with Noisy Labels. Referring to [1], researches on noisy label learning can be divided into model-based and model-free methods. For model-based methods, noise channel is one of the most wildly used techniques, which models an estimator to predict hidden true labels with the noisy labels. Most methods of this kind adopted confusing matrix as the noise channel [6, 20]. Take the work of Tanno et al. [27] as an example, the framework consisted of two connected models. The former predicted the hidden truth predictions, and the latter predicted the noisy labels by ground truth. Another kind of methods aims to model-freed denoising such as robust losses [22] and regularizers [17, 18]. These methods were designed to prevent overfitting without modeling the noise. For example, Wang et al. [29] researched the learning procedure and proposed Symmetric Cross Entropy (SCE). They mathematically proved the noise tolerance of the loss and achieved great results in the dataset with artificial noisy labels. These methods dealt with noise only by single images, and most of them evaluated on datasets with artificial noisy labels. Unlike that, there is a specific relationship between each mislabeled patch and the surrounding information in HCC samples. For instance, the blood vessel patch surrounded by cancer samples is preferred to be labeled as cancer, because such small regions in the tumor can be easily missed. Without large-scaled sight, these existing methods will lack crucial information for HCC pathology image grading.

3 Tendentious Noise-rectifying Framework

Aiming at the inevitable noises in pathology images, we propose the Tendentious Noise-rectifying Framework (TNF) for pathological grading of HCC. In TNF, noise-rectifying means the optimizing on NR loss makes the model rectify prediction on the reversed label when the predicted confidence on the original label is low, which is proved in Section 3.3. And different samples in this process has different tendencies, reflecting the confidence of the annotated labels derived from the structural information. The architecture is shown in Fig. 2. The TNF takes two components as input: HCC patches and corresponding feature polymers, representing the cellular detail feature and surrounding structural feature. In addition to assisting classification, the feature polymer also decides the tendency of the proposed Noise-Rectifying loss (NR), which is designed to rectify noises in pathology images. From the right figure it can be seen that NR loss will be minimized when either p or $\hat{p}$ is equal to 1, but the optimization tendency depends on α. In summary, TNF makes comprehensive utilization of both cellular and structural features, which effectively extracts information for noise rectification with the NR loss, and eventually improves the classification performance.

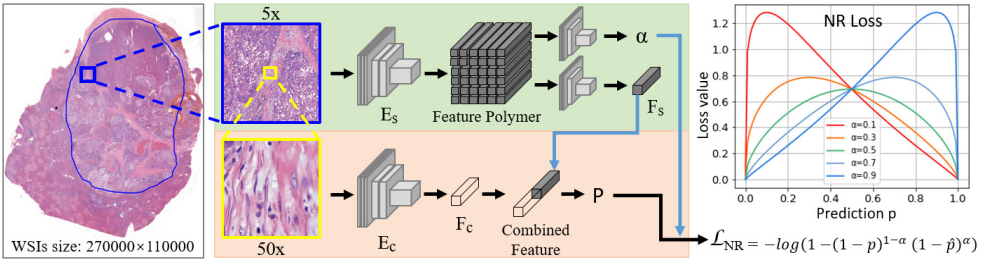

Figure 2: The flow diagram of Tendentious Noise-rectifying Framework (TNF), which is composed of two branches. The main branch (the bottom one) extracts cellular features F_c from central patches by encoder E_c. The top branch takes the surrounding areas from corresponding patches as input, and the features output by pre-trained encoder E_s are aggregated as feature polymer. After that, the feature polymer is embedded to the representations F_s, which is combined with F_c for final classification. What's more, the noise coefficient α is also predicted by feature polymer. The Noise-Rectifying (NR) loss is calculated according to the output results and α. The curves of NR loss reflects the effects of different α. In this figure it assumes that $p+\hat{p}=1$ to simplify the condition, where p and $\hat{p}$ denotes the predicted probability on the annotated label and reversed label, respectively.

3.1 Noise-Rectifying Loss

In deep learning, noisy labels has fatal impacts on the performance and robustness of models, which has been verified by [2, 17]. In regular classification tasks, the Cross-Entropy (CE) loss function is defined as follows:

$$\mathcal{L}_{CE} = -\sum_{k=0}^{K} Y_k log(P_k), \tag{1}$$

where K denotes the number of the classes, Y_k and P_k denote the GT and predicted probability of k-th class, respectively. During the optimization with CE, the predicted probability on

noisy labels will be forced to improve, resulting in the overfitting problem. It might aggravate in some hard tasks and lead to poor generalization performance.

A fundamental way to reduce the negative impact of those noisy labels is finding and rectifying the noise. So, we devise the Noise-Rectifying (NR) loss, which can dynamically rectify the high confidence samples in the training stage. In the collected HCC dataset, all patches from the same slide can be categorized as cancer of the same grade or benign one. However, the healthy and cancerous area of one slide both contains opposite patches. For a clear description, the reversed label is denoted as $\hat{Y}$ (original noisy label is Y). For example, for a patch in the cancerous area of grade 2 slides, the original label $Y_2 = 1$ (grade 2) and the reversed label $\hat{Y}_0 = 1$ (healthy). Due to the condition that only one cancer category occurs in a pathology image, there is no reverse between different cancer categories. An example is shown in Fig. 1.

Inspired by the fuzzy loss of [10], we devise a tendency hyper-parameter for rectifying noise labels in NR loss $\mathcal{L}_{NR}$, which is defined as follows:

$$\mathcal{L}_{NR} = -log(1-(1-p)^{1-\alpha}(1-\hat{p})^{\alpha}), \tag{2}$$

where α denotes the noise coefficient, p denotes the predicted probability on the annotated label and $\hat{p}$ denotes the predicted probability on the reversed label, which means $p = \sum_{k=1}^{K} Y_k P_k$ and $\hat{p} = \sum_{k=1}^{K} \hat{Y}_k P_k$. Low value of α represents low probability of mislabeling. Fig. 2 shows the curves of $\mathcal{L}_{NR}$ with the variable α. In the case that $\alpha = 0.1$, the minimization of NR leads to the growth of prediction on annotated label if $p > 0.1$. Otherwise, if $p < 0.1$, the prediction on annotated label will reduce and the prediction on the reversed label increases. The noise coefficient α in NR represents the tolerant threshold value. As will be proved in Section 3.3, a small constant of α is beneficial for rectifying target to the reversed label on samples with high confidence. What's more, for an adaptive tendency of different cases, the α is dynamically determined by the feature polymer, as described below.

Additional, extra regular term is introduced in TNF to restrain training process. Since the rectification ability of NR loss is highly dependent on the quality of α, $1-p$ is utilized as the supervised target of α in the early stage. In summary, the combined loss for each sample is derived as :

$$\mathcal{L}_{total} = -log(1-(1-p)^{1-\alpha}(1-\hat{p})^{\alpha}) + \beta|1-p-\alpha|, \tag{3}$$

where β denotes the weight of constraint of α. During training, the value of β decreases progressively, so that the model gradually learns the tendency from the feature polymer. In this paper, $\beta = 1 - e_c/e_m$, where e_c and e_m denote the current epoch and maximum epoch.

3.2 Feature Polymer

In HCC grading task with pathology slides, the cause of the noise is related to the rough annotation on ultra-large images. It indicates that the misclassified patches must locate in regions with specific characteristics, which inspires us that the surroundings may contain information for rectifying the noise label.

Our model takes feature polymer as input to represent the surrounding information. For dataset $\mathcal{D} = \{X^{(n)}, Y^{(n)}\}_{n=1}^{N}$, $X^{(n)} \in \mathbb{R}^{W\times H\times 3}$ denotes the patches cropped from HCC slides and $Y^{(n)} \in \mathbb{R}^{K}$ denotes the annotated labels. Among that N denotes the patch number, W, H denote the width and height of patches, and K denotes the category numbers. For each patch $X^{(n)}$, a group of surrounding patches are collected as $S(X^{(n)}) \in \mathbb{R}^{m\times W\times m\times H\times 3}$ (the patches

beyond the slide are replaced with 0), which will be input into the autoencoder for generating the feature polymer $P(S(X^{(n)})) \in \mathbb{R}^{d\times M\times M}$. The m denotes the patch number of surroundings, M denotes the size of polymerized feature map and d denotes the dimensionality. The autoencoder is pretrained on the training data, and the fully connected layer is removed to retain more location information.

In addition to combining original patches to get a joint feature vector, the polymers also dynamically adjust the noise coefficient α by the other branch of the framework. This coefficient represents the degree of the noise and will affect the training tendency of the NR loss. Generally, the feature polymer plays an auxiliary role in pathology image grading, mainly on providing peripheral structural information and determining the denoise tendency.

3.3 Theoretical Analysis

In NR, the loss minimizes when either p or $\hat{p}$ reaches 1, so the model prediction is relatively optional according to the extracted representation. A trainable exponent α that ranges from 0 to 1 is introduced to controls the tendency of the optimization. The effect will be clarified by the gradient of p and $\hat{p}$, which can be derived as:

$$\frac{\partial \mathcal{L}_{NR}}{\partial p} = \frac{-(1-\alpha)(\frac{1-\hat{p}}{1-p})^{\alpha}}{1-(1-p)^{1-\alpha}(1-\hat{p})^{\alpha}}, \qquad \frac{\partial \mathcal{L}_{NR}}{\partial \hat{p}} = \frac{-\alpha(\frac{1-p}{1-\hat{p}})^{1-\alpha}}{1-(1-p)^{1-\alpha}(1-\hat{p})^{\alpha}}. \tag{4}$$

It shows that both the gradients are negative, indicating the minimization of loss function increases p and $\hat{p}$ in varying degrees. To explore the relationship of two probabilities, we further calculate the gradients of the logits before the Softmax layer. For the Softmax function $p_i = \frac{e^{z_i}}{\sum_k e^{z_k}}$, the gradients varies under different conditions:

$$\frac{\partial p_i}{\partial z_j} = \begin{cases} p_i(1-p_i) & i = j, \\ -p_i p_j & i \neq j, \end{cases} \tag{5}$$

where p_i and z_i denote the probability and logit of the i-th category. According to the Eq. 4 and Eq. 5, the gradient of the logits can be derived as:

$$\frac{\partial \mathcal{L}_{NR}}{\partial z} = \frac{p(\frac{1-p}{1-\hat{p}})^{1-\alpha}(\hat{p}-(1-\alpha))}{1-(1-p)^{1-\alpha}(1-\hat{p})^{\alpha}}, \qquad \frac{\partial \mathcal{L}_{NR}}{\partial \hat{z}} = \frac{\hat{p}(\frac{1-\hat{p}}{1-p})^{\alpha}(p-\alpha)}{1-(1-p)^{1-\alpha}(1-\hat{p})^{\alpha}}, \tag{6}$$

where z and $\hat{z}$ denote the corresponding logits of p and $\hat{p}$. Excluding the terms that always positive, the gradient direction of z (or $\hat{z}$) depends on the difference between the exponent and the other probability (p against α or $\hat{p}$ against $1-\alpha$). For example, if the exponent of p is 0.9 (which means $1-\alpha = 0.9$), the logit z increases in most cases unless $\hat{p}$ is higher than 0.9, which implies the confidence is high enough to determine the reversed label as the target. In general, the exponents (α and $1-\alpha$) can be interpreted as the thresholds to tolerant noise. A low value of α makes the model have a preference for the annotated label in most conditions, but the model can also switch to the reversed labels on some samples that are convincing to be mislabeled. Specifically, the Cross-Entropy loss is a special condition with $\alpha = 0$, which has no tolerance for noise.

Simply setting the noise coefficient as constant such as 0.1, also empowers the model to denoise, but it will restrain the training of underfitting samples. So we make α adjustable

according to the surrounding information of each patch, which will be capable of distinguishing noisy samples and hard samples in pathology images. The gradient of α is derived as follows:

$$\frac{\partial \mathcal{L}_{NR}}{\partial \alpha} = \frac{(1-p)^{1-\alpha}(1-\hat{p})^{\alpha}}{1-(1-p)^{1-\alpha}(1-\hat{p})^{\alpha}} \ln(\frac{1-\hat{p}}{1-p}). \tag{7}$$

This gradient can be divided into two parts. The first item determined the gradient magnitude, which will be 0 when p or $\hat{p}$ reaches 1. The other item determined the direction of the gradient. It will be positive if p is larger than $\hat{p}$ and negative if it's the other way around. It conforms to the concept that the updating direction of α depends on the relative size of current p and $\hat{p}$. In the initial training phase, the α will be maximized for mislabeled samples that $\hat{p}$ is larger.

4 Experiment

4.1 Dataset

The collected HCC pathology dataset contains more than 20,000 patches cropped from 415 WSIs. The patch amplification and size are set to 50x and 448×448, respectively. For HCC diagnosis, the Edmondson-Steiner grading criterion is widely used to evaluate the disease degree on account of the robust prognostic implication in diagnosing HCC [15]. This grading system contains four patterns corresponding to various extents of cancerization. Higher grade represents a higher level of malignancy [24, 25]. With benign samples as grade 0, the slides have been labeled to five grades by pathologists, and the slides number of each grade are balanced. On each slide, the pathologists make rough annotations for the tumor region, so that normal and cancer patches are randomly cropped based on the annotation. Besides, in consideration of the similar feature among patches of the same slide, the partition of the training, validation, and test set is on slide level. All patches from the same slide only exist in one set to evaluate the generalization performance. For evaluating the model, the testing dataset is accurately annotated by pathologists. The pathologists spend more than an hour on each slide to ensure the testing labels are clean.

4.2 Experimental Settings

Network architecture. In the experiments, the ResNet18 [13] is adopted as the encoder of the main branch in TNF. It is pretrained on ImageNet [6]. In the branch of feature polymer, both the AE encoder and decoder consist 6 convolutional layers. As described in Section 3.2, the size of feature polymer is $d \times M \times M$. Here we set $d = 64, M = 63$, where M is the total feature dimension of 9 of surrounding patches (the size of each patch's feature is $64 \times 7 \times 7$). The polymer encoder contains 5 convolutional layers and an average pooling layer, and outputs the feature F_s with the same dimension as F_c (512×1). The network for outputting α contains 5 convolutional layers, followed by 2 fully connected layers.

Parameters. For a fair comparison, the SGD with the momentum of 0.9 and weight decay of 0.0001 is adopted in all the experiments. The batch size is set to 32 and the initial learning rate is set to 0.01. The learning rate is multiplied by 0.9 every two epochs. In TNF, the NR loss is divided into CNR (NR with a constant α) and ANR (NR with a adaptable α) for further exploration. The parameter α in CNR is set to 0.1 and that in ANR is determined by

feature polymer. The value of α for both methods are set to 0 in the first epoch, in order to make the model have the preliminary classification ability.

4.3 Quantitative Comparison

In our experiment, ResNet18 with Cross-Entropy (CE) loss is used as the baseline. The proposed method is evaluated on the collected HCC dataset against recent typical works, including model-based method (CM [20, 27]), model-free methods on robust losses (GCE [22], SCE [29]) and reguilarizers (LSR [18, 26], ELR and ELR+ [17]). The parameters of these methods are set according to the default setting in their experiments. For the proposed TNF, adaptable α is utilized.

Table 1: The classification results of different methods. The metrics contain accuracy (*Acc.*), sensitivity (*Sen.*), specificity (*Spe.*). Here TNF uses adaptable α determined by feature polymer. All these results averaged over ten experiments and the best results are marked in **bold** (All scores are in %).

	CE	CM [27]	GCE [22]	SCE [29]	LSR [18]	ELR [17]	ELR+ [17]	TNF
Acc.	64.97 ±0.021	66.51 ±0.027	66.34 ±0.018	67.27 ±0.020	65.78 ±0.022	63.26 ±0.031	65.95 ±0.021	**68.98 ±0.015**
Sen.	61.25 ±0.026	62.58 ±0.033	62.59 ±0.020	63.93 ±0.020	62.03 ±0.026	59.83 ±0.036	62.76 ±0.023	**65.77 ±0.022**
Spe.	78.16 ±0.032	**83.99 ±0.058**	80.02 ±0.028	79.70 ±0.042	80.73 ±0.051	76.69 ±0.048	78.04 ±0.033	81.37 ±0.057

Table 1 shows the classification results of different methods on HCC pathology dataset. We can see that TNF with self-adapting tendency achieves the best accuracy and sensitivity. Although CM [27] achieves the highest specificity, it is at the cost of low sensitivity. During confusion matrix training, it will optimize faster by transferring positive samples of multiple classes to negative ones. However, in tumor diagnosis, the low false negative is much more significant. Since the neglect of potential cancer regions will result in patients not being treated in time, sensitivity is important as well as grading accuracy. The HCC grading with noisy labels is a challenging task that even the latest methods are not effective. Nevertheless, TNF gets further improvement and outperforms the best existing methods with 1.71% accuracy increase.

For a fair comparison, we further incorporate feature polymer (FP) into these methods to provide sufficient information. The results are shown in Table 2. It's obvious to see that most of existing methods get limited improvement, and the performance of ELR+ is even worse. The possible reason is that simply combining two kinds of features can't fully utilize the structural information. Conversely the high dimension of feature will lead to overfitting. However, feature polymer is applied in TNF for better rectifying noise through structural information, which makes TNF outperform all these existing methods for HCC grading.

In order to evaluate the significance in clinical diagnosis, the cancerous patches in non-tumor areas predicted by baseline and TNF are confirmed again by pathologists. Among 500 test patches which are annotated as healthy (grade 0), the baseline model with CE detects 21 cancerous patches with 42.86% accuracy and TNF detects 60 cancerous patches with

Table 2: The classification results of different methods with the feature polymer (denoted by FP). Here TNF uses adaptable α determined by the FP. All these results averaged over ten experiments. The best results are marked in **bold** (All scores are in %).

Method(+FP)	CE	CM [21]	GCE [23]	SCE [29]	LSR [18]	ELR [17]	ELR+ [17]	TNF
Acc.	65.49 ±0.018	66.97 ±0.031	66.81 ±0.020	67.69 ±0.022	66.01 ±0.025	63.47 ±0.035	65.86 ±0.030	**68.98 ±0.015**

80.00% accuracy. It indicates that TNF has a powerful ability to find out the carcinoma cells in healthy areas, which has significant meaning for the diagnosis of cancer spreading.

4.4 Ablation Study

First of all, to evaluate the impact of different patch scales and select the most appropriate magnification, an ablation study is conducted on patches with magnifications of 10x, 30x, 50x, 70x by CE and TNF. The results are shown in Table 3. It can be seen that patches with 50x magnification are the most suitable for this dataset. From the samples in Fig. 1 we can also see that patches of 50x magnification have appropriate views for HCC grading. A smaller magnification results in the loss of cellular characteristics. And a larger magnification makes more patches that do not contain cancerous cells, which increases the noise ratio and leads to poor performance.

Table 3: Ablation results with different magnifications and surrounding patch numbers trained by CE and TNF. All scores are the average of ten experiments and the best result is marked in **bold** (All scores are in %).

Magnification	10x	30x	50x	70x	Surrounding patch number	m=5	m=7	m=9	m=11
CE	61.58 ±0.017	64.01 ±0.019	**64.97 ±0.021**	61.42 ±0.022	CE+FP	63.75 ±0.016	65.08 ±0.020	65.49 ±0.018	**65.52 ±0.016**
TNF	64.71 ±0.019	67.74 ±0.015	**68.98 ±0.015**	67.36 ±0.016	TNF	66.38 ±0.015	67.96 ±0.014	**68.98 ±0.015**	68.79 ±0.015

On the other hand, the patch number of the surrounding region (denoted as m) is also decided by the ablation study based on 50x magnification. In Table 3, there is little difference when m is set to 9 and 11. Considering the additional resource demand of large surrounding regions, $m = 9$ is more suitable for better extracting structural features.

Table 4: HCC grading results of NR with different α. Here CNR indicates NR with a constant α. Note that the NR with constant $\alpha = 0$ equals to CE. All scores are the average of ten experiments. The best results are marked in **bold** (All scores are in %).

	CNR					CNR+FP				
	$\alpha = 0$	$\alpha = 0.01$	$\alpha = 0.05$	$\alpha = 0.1$	$\alpha = 0.2$	$\alpha = 0$	$\alpha = 0.01$	$\alpha = 0.05$	$\alpha = 0.1$	$\alpha = 0.2$
Acc.	64.97 ±0.021	66.61 ±0.014	67.43 ±0.022	**67.46 ±0.015**	49.88 ±0.045	65.49 ±0.018	67.67 ±0.027	67.92 ±0.025	**68.26 ±0.026**	53.03 ±0.032

After deciding on the scale and surrounding patch number, additional experiments are conducted to verify the effectiveness of each component. Table 4 shows the performance of

methods with different constant value of α. We can find that method with $\alpha = 0.1$ achieves the highest score, which is consistent with our theory that a small α can help rectify the target to the reversed label. Here NR equals to the standard Cross-Entropy loss when $\alpha = 0$. It can be seen from experiments that a large value of α enhances the ability of NR to rectify the noise. But when the α is too large, the performance will decrease since the model tends to predict grade 0 for all the samples. So it's crucial to choose a suitable value. Actually, the NR with only a small value of α like 0.01 can exceed the CE. And an appropriate value helps to train the model to achieve better performance.

Table 5: Ablation results with different components of TNF. Here CNR indicates the NR with a constant α, and ANR indicates NR with an adaptable α. ANR+FP is TNF with all components. The value of α is set to 0.1 in CNR. All scores are the average of ten experiments and the best result is marked in **bold** (All scores are in %).

	CE	CNR	CNR+FP	ANR+FP
Acc.	64.97 ± 0.021	67.46 ± 0.015	68.26 ± 0.026	**68.98±0.015**

In order to improve the adaptability of CNR (NR with constant α), the feature polymer is utilized to determine the value of α in ANR (NR with adaptable α). The FP and ANR are in one system and can't be compared separately. So we conduct the ablation study on CNR, CNR+FP, and ANR+FP. From Table 5, it can be seen that CNR+FP exceeds CNR, which demonstrates the significance of the combining cellular and structural features for pathology image classification. Morever, ANF+FP (full components of TNF) further improves the performance on this basis. It shows the superiority of dynamic α determined by FP, which provides macro-view information. Since the tendency to annotated label of each sample is different, the dynamic α can better reflect the tendency than the constant value. As mentioned before, we believe that the surroundings of mislabeled patches have common characteristics and can be recognized by FP. The result verifies this assumption.

In addition, the complexity of the proposed TNF is analyzed in terms of parameters and training times. The ResNet18 for CE, LSR, etc. contains 11.24 million parameters, and TNF contains 15.81 million parameters. The average seconds of each batch in TNF and ResNet18 are 0.5982s and 0.5562s. On account of the additional branch for structural features, the model of TNF is more complex, but the difference in the training time is not significant.

5 Conclusion

In this paper, we present a noise-rectifying method TNF aiming at HCC grading on gigapixel pathology images. Considering the particular characteristics of HCC, we are the first to rectify the noise with surrounding structural information. Theoretical analysis proves that Noise-Rectifying loss with a small α indeed can rectify the noise samples with high confidence. And exhaustive experiments also verify the effectiveness of the proposed TNF with the adaptable α, which achieves SOTA performance. In the future, we will focus on the migration of our loss function to segmentation, and apply the noise-rectifying mechanism on other medical images. Since the lack of accurate annotations is a common problem in medical datasets, our method will make a breakthrough in computer-aided diagnosis.

6 Acknowledgement

This work is supported by National Natural Science Foundation of China (No.62002318), Key Research and Development Program of Zhejiang Province (2020C01023)Zhejiang Provincial Science and Technology Project for Public Welfare (LGF21F020020), the Major Scientific Research Project of Zhejiang Lab (No. 2019KD0AC01), and Ningbo Natural Science Foundation202003N4318).

References

[1] Görkem Algan and Ilkay Ulusoy. Image Classification with Deep Learning in the Presence of Noisy Labels: A Survey. *Knowledge-Based Systems*, 215:106771, 2021.

[2] Devansh Arpit, Stanisław Jastrzębski, Nicolas Ballas, David Krueger, Emmanuel Bengio, Maxinder S Kanwal, Tegan Maharaj, Asja Fischer, Aaron Courville, Yoshua Bengio, et al. A Closer Look at Memorization in Deep Networks. In *International Conference on Machine Learning*, pages 233–242. PMLR, 2017.

[3] Freddie Bray, Jacques Ferlay, Isabelle Soerjomataram, Rebecca L. Siegel, Lindsey A. Torre, and Ahmedin Jemal. Global Cancer Statistics 2018: GLOBOCAN Estimates of Incidence and Mortality Worldwide for 36 Cancers in 185 Countries. *Ca A Cancer Journal for Clinicians*, 2018.

[4] Philip Chikontwe, Meejeong Kim, Soo Jeong Nam, Heounjeong Go, and Sang Hyun Park. Multiple Instance Learning with Center Embeddings for Histopathology Classification. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 519–528. Springer, 2020.

[5] Jia Deng, Wei Dong, Richard Socher, Li Jia Li, and Fei Fei Li. ImageNet: A large-scale hierarchical image database. In *IEEE Conference on Computer Vision Pattern Recognition*, 2009.

[6] Yair Dgani, Hayit Greenspan, and Jacob Goldberger. Training a Neural Network Based on Unreliable Human Annotation of Medical Images. In *2018 IEEE 15th International Symposium on Biomedical Imaging (ISBI 2018)*, pages 39–42. IEEE, 2018.

[7] Neeraj Dhungel, Gustavo Carneiro, and Andrew P Bradley. Deep Learning and Structured Prediction for the Segmentation of Mass in Mammograms. In *International Conference on Medical image computing and computer-assisted intervention*, pages 605–612. Springer, 2015.

[8] Songhui Diao, Weiren Luo, Jiaxin Hou, Hong Yu, Yunqiang Chen, Jing Xiong, Yaoqin Xie, and Wenjian Qin. Computer Aided Cancer Regions Detection of Hepatocellular Carcinoma in Whole-slide Pathological Images based on Deep Learning. In *2019 International Conference on Medical Imaging Physics and Engineering (ICMIPE)*, pages 1–6. IEEE, 2019.

[9] Duc Fehr, Harini Veeraraghavan, Andreas Wibmer, Tatsuo Gondo, Kazuhiro Matsumoto, Herbert Alberto Vargas, Evis Sala, Hedvig Hricak, and Joseph O Deasy. Automatic Classification of Prostate Cancer Gleason Scores from Multiparametric Magnetic

Resonance Images. *Proceedings of the National Academy of Sciences*, 112(46):E6265–E6273, 2015.

[10] Zunlei Feng, Weixin Liang, Daocheng Tao, Li Sun, and Mingli Song. CU-Net: Component Unmixing Network for Textile Fiber Identification. *International Journal of Computer Vision*, 127(10):1443–1454, 2019.

[11] Tatiana Gabruseva, Dmytro Poplavskiy, and Alexandr Kalinin. Deep Learning for Automatic Pneumonia Detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pages 350–351, 2020.

[12] Varun Gulshan, Lily Peng, Marc Coram, Martin C Stumpe, Derek Wu, Arunachalam Narayanaswamy, Subhashini Venugopalan, Kasumi Widner, Tom Madams, Jorge Cuadros, et al. Development and Validation of a Deep Learning Algorithm for Detection of Diabetic Retinopathy in Retinal Fundus Photographs. *Jama*, 316(22):2402–2410, 2016.

[13] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[14] Viksit Kumar, Jeremy M Webb, Adriana Gregory, Max Denis, Duane D Meixner, Mahdi Bayat, Dana H Whaley, Mostafa Fatemi, and Azra Alizad. Automated and Real-time Segmentation of Suspicious Breast Masses using Convolutional Neural Network. *PloS one*, 13(5):e0195816, 2018.

[15] Li, Zhou, Jing-An, Rui, Wei-Xun, Zhou, Shao-Bin, Wang, Shu-Guang, and Chen and. Edmondson-Steiner Grade: A Crucial Predictor of Recurrence and Survival in Hepatocellular Carcinoma without Microvascular Invasio. *Pathology Research Practice*, 2017.

[16] Haotian Liao, Tianyuan Xiong, Jiajie Peng, Lin Xu, Mingheng Liao, Zhen Zhang, Zhenru Wu, Kefei Yuan, and Yong Zeng. Classification and Prognosis Prediction from Histopathological Images of Hepatocellular Carcinoma by a Fully Automated Pipeline Based on Machine Learning. *Annals of surgical oncology*, pages 1–11, 2020.

[17] Sheng Liu, Jonathan Niles-Weed, Narges Razavian, and Carlos Fernandez-Granda. Early-learning Regularization Prevents Memorization of Noisy Labels. *arXiv preprint arXiv:2007.00151*, 2020.

[18] Michal Lukasik, Srinadh Bhojanapalli, Aditya Menon, and Sanjiv Kumar. Does label smoothing mitigate label noise? In *International Conference on Machine Learning*, pages 6448–6458. PMLR, 2020.

[19] Sebastiao N. Martins-Filho, Paiva Caterina, Azevedo Raymundo Soares, and Alves Venancio Avancini Ferreira. Histological Grading of Hepatocellular Carcinoma—A Systematic Review of Literature. *Frontiers in Medicine*, 4:193–, 2017.

[20] Ishan Misra, C Lawrence Zitnick, Margaret Mitchell, and Ross Girshick. Seeing Through the Human Reporting Bias: Visual Classifiers from Noisy Human-centric Labels. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2930–2939, 2016.

[21] Tan Huu Nguyen, Shamira Sridharan, Virgilia Macias, Andre Kajdacsy-Balla, Jonathan Melamed, Minh N Do, and Gabriel Popescu. Automatic Gleason Grading of Prostate Cancer using Quantitative Phase Imaging and Machine Learning. *Journal of biomedical optics*, 22(3):036015, 2017.

[22] Giorgio Patrini, Alessandro Rozza, Aditya Krishna Menon, Richard Nock, and Lizhen Qu. Making Deep Neural Networks Robust to Label Noise: A Loss Correction Approach. In *IEEE Conference on Computer Vision Pattern Recognition*, 2017.

[23] Pranav Rajpurkar, Jeremy Irvin, Kaylie Zhu, Brandon Yang, Hershel Mehta, Tony Duan, Daisy Ding, Aarti Bagul, Curtis Langlotz, Katie Shpanskaya, et al. Chexnet: Radiologist-level pneumonia detection on chest x-rays with deep learning. *arXiv preprint arXiv:1711.05225*, 2017.

[24] Manuel Rodriguez-Peralvarez, Tu Vinh Luong, Lorenzo Andreana, Tim Meyer, Amar Paul Dhillon, and Andrew Kenneth Burroughs. A Systematic Review of Microvascular Invasion in Hepatocellular Carcinoma: Diagnostic and Prognostic Variability. *Annals of surgical oncology*, 20(1):325–339, 2013.

[25] Shuji Sumie, Ryoko Kuromatsu, Koji Okuda, Eiji Ando, Akio Takata, Nobuyoshi Fukushima, Yasutomo Watanabe, Masamichi Kojiro, and Michio Sata. Microvascular Invasion in Patients with Hepatocellular Carcinoma and its Predictable Clinicopathological Factors. *Annals of surgical oncology*, 15(5):1375–1382, 2008.

[26] Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jonathon Shlens, and Zbigniew Wojna. Rethinking the Inception Architecture for Computer Vision. 2015.

[27] Ryutaro Tanno, Ardavan Saeedi, Swami Sankaranarayanan, Daniel C. Alexander, and Nathan Silberman. Learning From Noisy Labels by Regularized Estimation of Annotator Confusion. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020.

[28] David Tellez, Geert Litjens, Jeroen van der Laak, and Francesco Ciompi. Neural Image Compression for Gigapixel Histopathology Image Analysis. *IEEE transactions on pattern analysis and machine intelligence*, 2019.

[29] Yisen Wang, Xingjun Ma, Zaiyi Chen, Yuan Luo, Jinfeng Yi, and James Bailey. Symmetric Cross Entropy for Robust Learning with Noisy Labels. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 322–330, 2019.

[30] Gang Xu, Zhigang Song, Zhuo Sun, Calvin Ku, Zhe Yang, Cancheng Liu, Shuhao Wang, Jianpeng Ma, and Wei Xu. Camel: A Weakly Supervised Learning Framework for Histopathology Image Segmentation. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 10682–10691, 2019.