



OPEN ACCESS

EDITED BY

Carlos S. Lima,
University of Minho, Portugal

REVIEWED BY

Guo Zhang,
Chongqing University of Posts and
Telecommunications, China
Yifan Gao,
University of Science and Technology of
China, China

*CORRESPONDENCE

Wei Lu
✉ luwei@wmu.edu.cn
Yi Wei
✉ drweiyi057@163.com

RECEIVED 25 June 2025

ACCEPTED 21 August 2025

PUBLISHED 10 September 2025

CITATION

Cheng X, Tian H, Zhou J, Xie T, Gong H, Liu M,
Wei Y and Lu W (2025) MPVT+: a noise-robust
training framework for automatic liver tumor
segmentation with noisy labels.
Front. Med. 12:1653865.
doi: 10.3389/fmed.2025.1653865

COPYRIGHT

© 2025 Cheng, Tian, Zhou, Xie, Gong, Liu, Wei
and Lu. This is an open-access article
distributed under the terms of the [Creative
Commons Attribution License \(CC BY\)](#). The
use, distribution or reproduction in other
forums is permitted, provided the original
author(s) and the copyright owner(s) are
credited and that the original publication in
this journal is cited, in accordance with
accepted academic practice. No use,
distribution or reproduction is permitted
which does not comply with these terms.

MPVT+: a noise-robust training framework for automatic liver tumor segmentation with noisy labels

Xuan Cheng¹, Haoxiang Tian², Jiajun Zhou², Tianshu Xie¹,
HaiGang Gong^{1,3}, Ming Liu³, Yi Wei^{4*} and Wei Lu^{3*}

¹University of Electronic Science and Technology of China, Chengdu, China, ²Yangtze Delta Region Institute (Quzhou), University of Electronic Science and Technology of China, Quzhou, China, ³The Quzhou Affiliated Hospital of Wenzhou Medical University, Quzhou People's Hospital, Quzhou, China, ⁴Department of Radiology, West China Hospital, Sichuan University, Chengdu, China

Introduction: Liver cancer is among the leading causes of cancer-related deaths worldwide. Accurate delineation of hepatic tumors is crucial for diagnosis, prognosis, and treatment planning, yet manual annotation is labor-intensive and subject to variability. Deep neural networks (DNNs) have shown promise in automating segmentation but require large amounts of high-quality labeled data, which is difficult to obtain. Incorporating noisy labels without proper handling can corrupt training and degrade performance.

Methods: We introduce MPVT+, a noise-robust training framework that integrates a pixel-wise noise-adaptation module with a multi-stage perturbation and variable-teacher (MPVT) consistency strategy. The noise adaptor infers corruption probabilities and re-weights unreliable supervision, while MPVT assembles an ensemble of stochastic teacher models that apply progressively stringent perturbations. This combination enables the network to exploit both clean and noisy labels without overfitting.

Results: Experiments conducted on 739 retrospectively collected liver-tumor CT datasets demonstrated that MPVT+ significantly outperformed baseline and traditional noise-handling approaches. Compared to a noise-free U-Net baseline (Dice Similarity Coefficient [DSC] 75.1%), MPVT+ improved segmentation accuracy to 80.3%. The framework consistently achieved superior results across multiple evaluation metrics, including DSC, JSC, SVD, and VOE.

Discussion: The MPVT+ framework demonstrates that principled noise modeling, coupled with consistency training, effectively unlocks the potential of imperfect medical datasets. This strategy reduces the dependency on perfectly labeled datasets and moves fully automated liver tumor delineation closer to clinical applicability.

KEYWORDS

liver tumor segmentation, noisy label, semi-supervised learning, deep learning, annotation

1 Introduction

Liver cancer is a prevalent malignant tumor and ranks as the third leading cause of cancer-related deaths globally in 2020 (1). An annual rise in liver cancer cases is noted, making it the only cancer among the top five deadliest (2). Diagnosing liver cancer typically involves various medical imaging techniques such as ultrasound, magnetic resonance imaging (MRI), and computer tomography (CT). Among these, CT is the most commonly used method because of its multiphasic contrast enhancement capabilities, providing more advanced imaging modalities (3). However, accurate diagnosis and subsequent treatment planning rely heavily on precise delineation of the tumor boundaries within the liver.

The physical characteristics of a tumor, such as its size, shape, and location, are important biomarkers that play a key role in accurate diagnosis and treatment (4).

Deep learning has shown great potential in medical segmentation tasks (5, 6). However, the quality of labels used for training is particularly concerning (7, 8) as it directly impacts model performance of correctly discerning and segmenting tumors and consequently influences clinician's decisions of diagnoses and treatments of tumors. Accurate, consistent, and clear labels not only enhance the performance and generalization ability of the model but also ensure that the segmentation results possess clinical interpretability and application reliability. Primarily, erroneous label data can impact the performance of deep learning models, potentially leading to the learning of incorrect features or segmentation boundaries. Second, noisy labels can affect the model's generalization ability, resulting in poor performance on unseen data (9). In addition, overfitting to noisy labels can cause a decline in the model's performance on test data as it fails to accurately capture the true distribution of the data. Finally, the presence of low-quality labels increases the difficulty of training and requires corresponding strategies to mitigate their impact, to enhance the model's performance and generalization ability. Therefore, ensuring high-quality label data is crucial for the success of tumor segmentation tasks.

To tackle the challenges posed by noisy labels, various strategies have been explored in this study. Some studies utilize annotations or consistent labels from numerous domain experts to improve the quality of data (10, 11). However, this method is impractical due to its significant financial and logistical resource consumption, which is not easily accessible. Moreover, inter-observer variability further compounds this issue as different radiologists may interpret ambiguous features differently based on their experience and training (12, 13). Other studies aim to reduce the workload of experts in medical image annotation by using computer assistance (14). However, variability among different observers and within the same observer cannot be disregarded when using this method, since a single radiologist may provide different annotations at different times, influenced by varying conditions such as fatigue or changes in perceptual criteria (15). Some studies utilize data automatically extracted from medical image databases such as hospital picture archiving and communication systems to create their training datasets (16, 17). Historical data may not always be available for every study and tend to be noisy. Some studies use crowd-sourcing to collect labeled data from non-experts (18, 19). However, they often do not produce labels of sufficient quality for general studies and typically have a high noise rate (19). Some researchers proposed to utilize noisy labels to enhance performance of models. These methods focus on creating a robust architecture to model latent parameters, enhance generalization, adjust loss functions, and select appropriate samples to reduce the impact of noisy labels on the model (9). For instance, some approaches leverage robust optimization techniques that specifically account for noise in the labels (20), some other proposed noise-robust loss functions aiming to lessen the influence of incorrectly labeled data during the training process (21). In addition, there is a growing interest in the implementation of transfer learning, where models pre-trained on large, clean datasets are fine-tuned on smaller,

noisier datasets, thereby capitalizing on the learned representations that are less susceptible to label noise (22). Techniques such as label smoothing (23) and uncertainty modeling (24, 25) can also help the model to not overfit to the noisy labels and instead focus on the underlying pattern in the data (26). Some methods completely ignore the noisy labels and train the networks using a semi-supervised approach. These evolving methodologies underscore the dynamic nature of deep learning in medical images, where ongoing research continually refines the balance between data quality and model reliability.

Probabilistic noise model is a fundamental concept in robust architecture. These methods are founded on the concept that noise can be represented by latent variables using neural networks (9, 27). Our method is inspired by both robust architecture and semi-supervised learning. To train deep neural network with noisy labels, it is crucial to identify noisy labels and eliminate the negative impact of them as much as possible. Based on the concept of noise adaptation layer (28), we introduced noise adaptor which alters convolutional neural networks by adding fully connected layers and softmax layers to predict hidden variables and simulate the generation of label noise. This module dynamically adjusts during training and refurbish the unreliable pixels in labels to produce unbiased supervisory signal. Furthermore, to prevent the model from overfitting due to potential changes in the data distribution, a semi-supervised framework called MPVT is implemented. This framework includes mean teachers with variable parameters updating inspired by dropout (29) to generate more robust pseudo-label guidance, which then allow us to introduce more perturbations into the training process to enhance the robustness and stability of the student model. This process effectively increases the model's resilience to overfitting by presenting it with a variety of training scenarios, ensuring that it learns to generalize from the core features of the data rather than its noisy aspects. This framework aims to help the model learn from reliable data and extract valuable information from noisy labels without being affected by the noise. Compared with previous methods, our method not only identifies and reduces the influence of noisy labels but also reinforces the robustness of the network by harnessing the inherent uncertainty within the data. This dual strategy effectively creates a noise-robust learning process, thereby improving the quality of the model's predictions and ensuring greater reliability in clinical applications.

2 Materials and methods

2.1 Datasets acquisition

This retrospective cross-sectional study received approval from the institutional review board at West China Hospital of Sichuan University. It primarily focused on patients diagnosed with hepatocellular carcinoma (HCC), intrahepatic cholangiocarcinoma (ICC), or liver metastasis (MET). Due to the retrospective nature of the study, patient consent was waived. The objective was to analyze imaging datasets derived from patients who underwent CT screening at our institution. To maintain scientific rigor, patients included in the study were required to meet specific inclusion

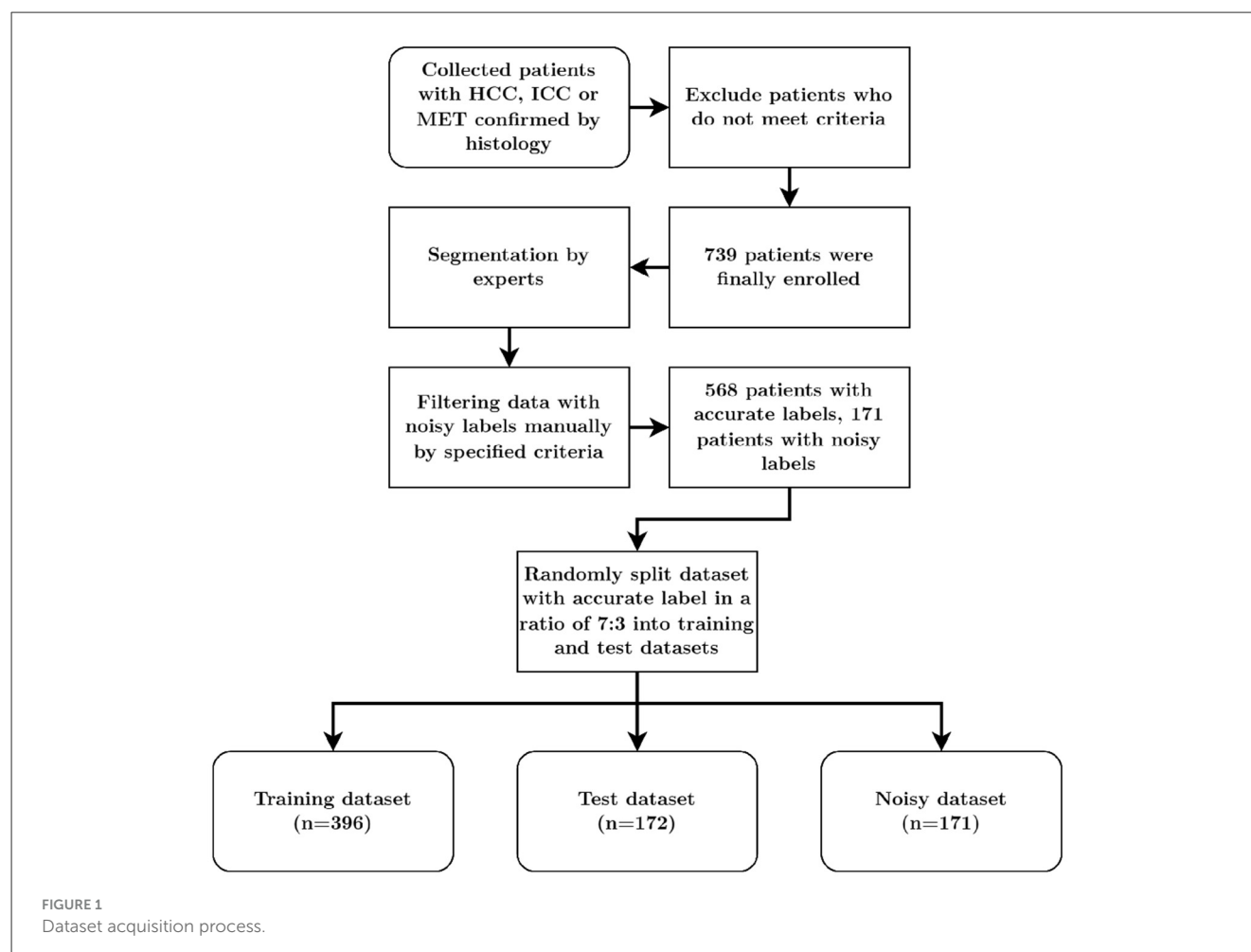
criteria: 1. Patients had to be at least 18 years old to ensure that the research subjects exhibited adult physiological characteristics, thereby enhancing the relevance of our findings to the adult population. 2. Patients with a history of hepatectomy, transarterial chemoembolization, or radiofrequency ablation prior to the CT examination were excluded to eliminate potential confounding factors that could affect the experimental outcomes. 3. All included patients had confirmed diagnoses of HCC, ICC, or MET, either through pathology reports or follow-up imaging studies lasting at least 6 months, verified by at least two independent imaging methods. This stringent criterion aimed to ensure the accuracy and reliability of the tumor diagnoses.

Furthermore, to identify data with noisy labels, several specific criteria were established for CT images of patients: 1. Cases with labels that are visibly misaligned with the anatomical structures they are supposed to represent should be marked. 2. Cases where labels bleed into adjacent structures due to segmentation errors should be marked. 3. Cases with labels that are inconsistent across slices or with conflicting boundaries should be marked. 4. Cases with labels that contain random noise or are not representative of the true anatomy should be marked. 5. Cases where the labeled regions are incomplete or cutoff should be marked. In this study, a total of 739 patients were included, with further details provided in Figure 1.

2.2 Overall structure

To address the detrimental impact of noisy labels on network training, we introduced a module termed the noise adaptor. This module is designed to mitigate the disturbances caused by noise. Furthermore, we proposed a consistency regularization scheme, Multi-Stage Perturbations and Variable Teachers (MPVT), to effectively harness the latent information present in the CT images. The noise adaptor module, akin to the decoder of U-Net (30), generates probabilities of mislabeling. These probabilities enable the prediction of the likelihood of mislabeling or the introduction of noise to good labels. This information, when used in conjunction with the noisy labels, aids the supervised learning process. The noise adaptor seeks to extract a clean supervisory signal from the noisy labels by simulating noise, allowing the backbone network to learn the true data distribution more effectively, thereby enhancing the model's generalization ability and robustness.

Considering the significant impact of semi-supervised learning on model performance in the absence of accurate labels, we proposed a model that employs consistency regularization with multi-stage perturbations and variable teachers. This combined approach enhances the model's ability to cope with noisy labels and optimizes the use of input images by introducing various perturbations using different teacher models ensemble from



different stages of training. We integrated the noise adaptor and MPVT scheme into the training process. Consequently, the overall training loss is:

$$\mathcal{L} = \mathcal{L}_{sup}(\mathcal{D}_c) + \beta_{ada}(\mathcal{L}_{ada}(\mathcal{D}_n) + \beta_{reg}\sigma) + \beta_{semi}(\mathcal{L}_{semi}(\mathcal{D}_c) + \mathcal{L}_{semi}(\mathcal{D}_n)), \quad (1)$$

where $\mathcal{D}_c$ and $\mathcal{D}_n$ represent datasets with accurate labels and noisy labels, respectively; $\mathcal{L}_{sup}$, $\mathcal{L}_{ada}$, and $\mathcal{L}_{semi}$ denote the supervision loss, adaptor loss, and semi-supervised loss, respectively; β_{ada} is the weight that determines the extent of the noise adaptor's supervision in training; σ represents the penalty term, and β_{reg} is the weight of the penalty term, which penalizes the noise adaptor to prevent it from finding shortcuts and not performing its supervisory role during training; β_{semi} is the weight of the semi-supervised learning component, and its magnitude is critical for the model's accuracy and generalization capability.

2.3 Noise adaptor module

To develop robust architecture that can adapt to label noise, it is essential to identify and model the types of label noise. Label corruption can be caused by the characteristics of images and labels themselves, which may increase inter- and intra-observer variation due to ambiguity that can prevent annotators from behaving consistently. In addition, random perturbation should be considered as the labeling process cannot be consistent and ideal.

Assuming that the corruption process is conditionally independent of label categories, label noise can be categorized as symmetrical and asymmetric noise. Symmetrical noise, or pure random noise, occurs when labels are corrupted by label transition probabilities p_{ij} , where i is the true label and j is the corrupted label. This type of noise is termed symmetrical because the probabilities of flipping a true label into any other labels are equal, which can be expressed as follows:

$$(\exists \rho \in [0, 1]) \wedge (\forall_{i=j} p_{ij} = 1 - \rho) \wedge \left(\forall_{i \neq j} p_{ij} = \frac{\rho}{1 - c} \right), \quad (2)$$

where ρ is the noise rate, and c is the number of classes. This equation reflects the equal likelihood of a label being incorrectly assigned to any other class in the presence of symmetrical noise;

$$(\exists \rho \in [0, 1]) \wedge (\forall_{i=j} p_{ij} = 1 - \rho) \wedge \left(\forall_{i \neq j} \sum_{j=0}^c p_{ij} = \rho \right), \quad (3)$$

where the probabilities of a true label being mislabeled as another one differ from one another. This implies that the true label is most likely to be mislabeled as a particular label. To implement this concept, we introduce a learnable layer at the end of the segmentation network output, inspired by Goldberger and Ben-Reuven (28).

In more realistic scenarios, label corruption is assumed to be related to both labels and features. This type of label noise

is referred to as instance-dependent label noise (27, 28). The probability with noise can be formulated as:

$$p_{ij}(x) = p(\tilde{y} = j | y = i, x), \quad (4)$$

where x represents the input to the networks. This implies that, in addition to the aforementioned noises, the features of the input data also have an impact on the annotations. For instance, in some images, the form of the portal vein may resemble a lesion, making it more likely to be mistaken as a tumor.

However, in the context of segmentation, there are differences compared to classification. Segmentation involves pixel-wise classification, where positional factors must be taken into account. Therefore, we modified the noise adaptation layer based on the concept that different pixels should have different probabilities. For instance, the edges around CT images are likely to have low signal intensity as patients are typically positioned in the middle of the images. Conversely, distinct areas with signal intensity differing from their surroundings are more likely to be lesions. This concept is referred to as positional label noise, and it can be expressed as:

$$p_{ij}(\omega) = p(\tilde{y} = j | y = i, \omega), \quad (5)$$

where ω represents a particular pixel in the images.

As depicted in Figure 2, we have implemented a noise adaptor based on the U-Net architecture (30). The adaptor is a decoder similar to the original U-Net decoder, receiving the extracted features from the encoder, which is part of the U-Net architecture. Then, the adaptor uses a single, direct convolutional layer (3×3 conv with 2 output channels) followed by a sigmoid layer to produce 2-channel probabilities. The 2-channel output represents the two probability endpoints that are used for linear interpolation with the main network's predictions. Through these two decoders, we obtain the predicted probabilities from the U-Net and the predicted corruption probabilities from the adaptor. The corruption probabilities consist of two parts: the probabilities of mislabeling background as foreground and the probabilities of correctly labeling foreground. We can then obtain the predicted probabilities with label noise according to the law of total probability, which can be formally expressed as:

$$\hat{y}_{ada} = \hat{y} \cdot p_{ada}(j = 1 | i = 1) + (1 - \hat{y}) \cdot p_{ada}(j = 1 | i = 0), \quad (6)$$

where $\hat{y}_{ada}$ represents the prediction after adaptation, $\hat{y}$ is the prediction of the U-Net, $p_{ada}(j = 1 | i = 0)$ is the probabilities of mislabeling background as foreground, and $p_{ada}(j = 1 | i = 1)$ is the probabilities of correctly labeling foreground. This process can be viewed as the corruption process from accurate labels to noisy labels. By employing this approach, we can train the model with data containing noisy labels in a normal supervised fashion.

To implement the aforementioned concept, we define the training loss for the network with the noise adaptor as:

$$\mathcal{L} = \mathcal{L}_{sup} + \beta_{ada}(\mathcal{L}_{ada} + \beta_{reg}\sigma). \quad (7)$$

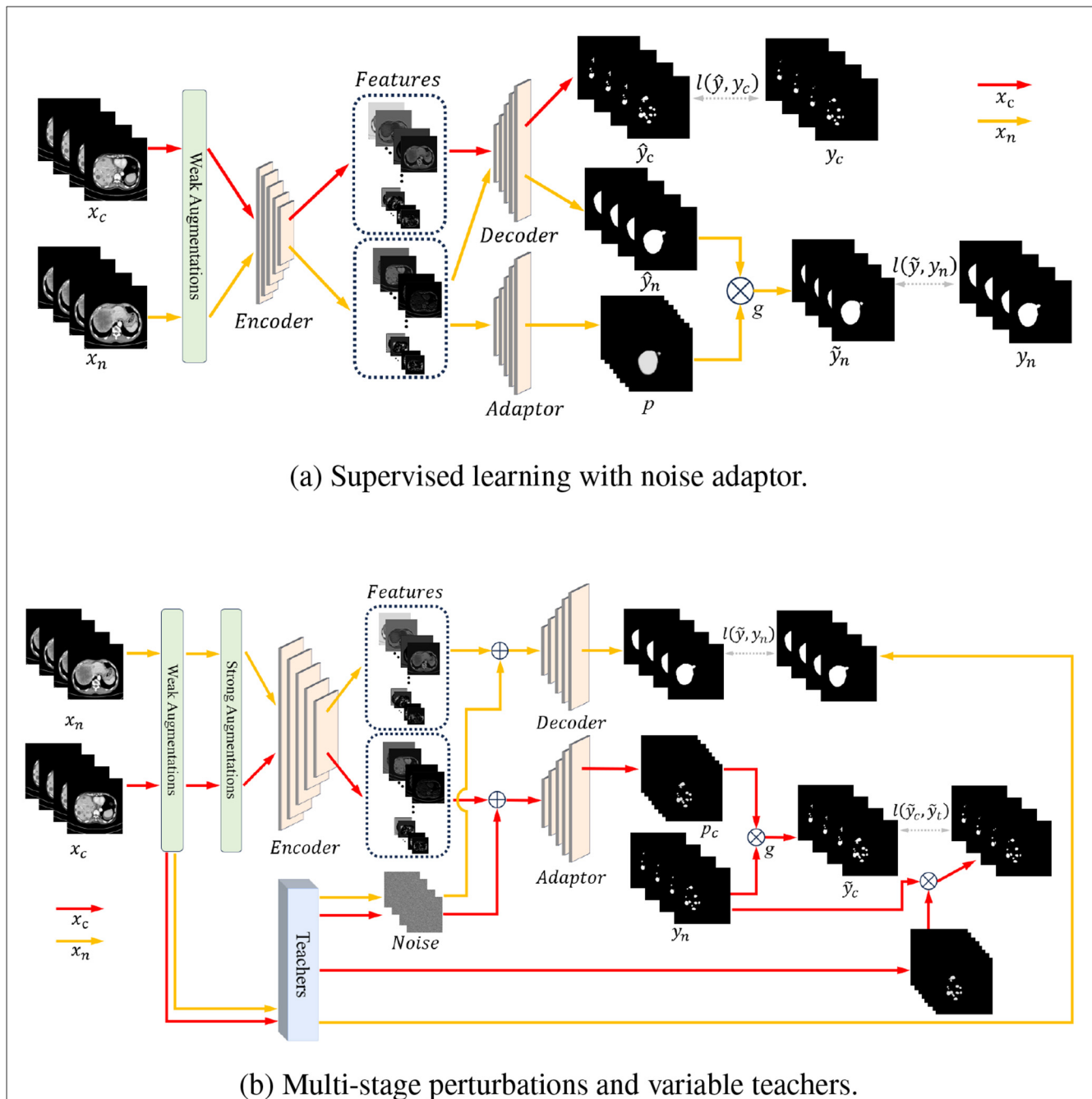


FIGURE 2

Overall structure diagram. (a) In the training process for supervised learning with the proposed noise adaptor, the data flow is represented by two distinct lines: The red line indicates the data flow of data with accurate labels, while the orange line represents the data flow of data with noisy labels. After the encoder processes the data, the features of data with noisy labels are fed into both the decoder and the adaptor. The adaptor then generates predicted noisy labels, which provide a supervisory signal for network training. (b) Consistency regularization is conducted with data that has accurate labels, while data with noisy labels are regarded as unlabeled data. This approach is used to enhance both the backbone networks and adaptor networks. By employing multi-stage perturbations and variable teachers, the consistency regularization process helps to improve the robustness and generalization capability of the model.

This loss consists of two terms: the supervised loss and the adaptor loss. The supervised loss is denoted as:

$$\mathcal{L}_{sup} = \frac{1}{|\mathcal{D}_n||\Omega|} \sum_{(\mathbf{x}, \mathbf{y}) \in \mathcal{D}_n} \sum_{\omega \in \Omega} l(f_{\theta}(\mathcal{T}_{weak}(\mathbf{x})), \mathcal{T}_{weak}(\mathbf{y})), \quad (8)$$

where $\mathcal{D}_n$ is the dataset with noisy labels, l is the loss function for supervised learning (we select cross-entropy loss and Dice loss), $\mathcal{T}_{weak}$ represents weak augmentation (both images and labels are augmented with the same settings), and f_{θ} is the U-Net parameterized with θ . Ω denotes the set of pixels in each training sample over which the pixel-wise loss is computed.

Since segmentation is formulated as pixel-wise classification, the summation over Ω ensures that the loss function accounts for every pixel in each training image. The adaptor loss is denoted as:

$$\mathcal{L}_{ada} = \frac{1}{|\mathcal{D}_n||\Omega|} \sum_{(\mathbf{x}, \mathbf{y}) \in \mathcal{D}_n} \sum_{\omega \in \Omega} l(g(f_{\theta}(\mathcal{T}_{weak}(\mathbf{x})), f_{\theta_{ada}}(\mathcal{T}_{weak}(\mathbf{x}))), \mathcal{T}_{weak}(\mathbf{y})), \quad (9)$$

where $f_{\theta_{ada}}$ is the adaptor and g is the transform function. According to the previous discussion, we define the transform function as:

$$g(\hat{y}, p_{ada}) = \hat{y}_{ada} = \hat{y} \cdot p_{ada} \quad (j = 1 | i = 1) \\ + (1 - \hat{y}) \cdot p_{ada} \quad (j = 1 | i = 0). \quad (10)$$

In addition, a regularization term σ is added to prevent the adaptor from taking shortcuts and preventing the U-Net from learning anything. σ can be defined as:

$$\sigma = \frac{1}{|\mathcal{D}_n||\Omega|} \sum_{(\mathbf{x}, \mathbf{y}) \in \mathcal{D}_n} \sum_{\omega \in \Omega} \|f_{\theta}(\mathcal{T}_{weak}(\mathbf{x})) - g(f_{\theta}(\mathcal{T}_{weak}(\mathbf{x}))), \mathcal{T}_{weak}(\mathbf{y}))\|_2^2. \quad (11)$$

2.4 Multi-stage perturbations and variable teachers

Semi-supervised learning techniques are valuable for extracting features from unlabeled data, aiding neural networks in learning from it to enhance their overall generalization. When training a model with noisy labels, it can eventually become biased and suffer from poor generalization. By discarding noisy labels, the training process transitions to semi-supervised learning and fully utilizes the data with corrupted labels. Common techniques include consistency regularization via pseudo-labeling and perturbations. To implement this concept, we propose the MPVT framework—a self-training and consistency regularization framework with multi-stage perturbations and variable teachers—with reference to (31) and (32).

In the training process, the loss is defined as:

$$\mathcal{L} = \mathcal{L}_{sup} + \beta_{semi} \mathcal{L}_{semi}, \quad (12)$$

where $\mathcal{L}_{sup}$ is the loss of supervised learning, and $\mathcal{L}_{semi}$ is the loss of semi-supervised learning, which is calculated as:

$$\mathcal{L}_{semi} = \frac{1}{|\mathcal{D}_n||\Omega|} \sum_{(\mathbf{x}, \mathbf{y}) \in \mathcal{D}_n} \sum_{\omega \in \Omega} l(f_{\theta_s}(\mathcal{T}_{strong}(\mathcal{T}_{weak}(\mathbf{x}))), \rho_{\tau_1, \tau_2}(\mathcal{F}_{\theta_t}(\mathcal{T}_{strong}(\mathcal{T}_{weak}(\mathbf{x}))), \mathbf{y})), \quad (13)$$

where $\mathcal{T}_{strong}$ represents the strong augmentations, f_{θ_s} denotes the student model parameterized by θ_s , $\mathcal{F}_{\theta_t}$ is the ensemble of teacher networks, and ρ_{τ} is the function used to obtain pseudo-labels from predicted probabilities with thresholds τ_1 and τ_2 . To address the

sample selection bias that can affect the training process, a tailored pseudo-label generation method is proposed, formulated as:

$$y_{pseudo}(c) = (\hat{y}(c) \geq \tau_1 \vee (\hat{y}(c) \geq \tau_2 \wedge y(c) = 1)). \quad (14)$$

In summary, we constructed a U-Net with a noise adaptor and trained it using the described training processes and losses. The main hyperparameters were as follows: we used Rectified Adam (RAdam) as the optimizer with an initial learning rate of 0.001 and a weight decay of 0.0001. Gradients were accumulated every 2 batches, with a batch size of 24 per worker. The learning rate was dynamically scaled down by a factor of 10 for epochs based on the current loss trend. We implemented the proposed model using PyTorch 1.2.0. The experiments were conducted on an Ubuntu 20.04.6 LTS server with two NVIDIA GeForce RTX 3090 cards.

2.5 Statistic indices

Here, we selected seven statistical indices to assess the performance of our method, including the Dice Similarity Coefficient (DSC), Jaccard Similarity Coefficient (JSC), Symmetric Volume Difference (SVD) and Volumetric Overlap Error (VOE)(33). These indices collectively provide a comprehensive evaluation of the method's accuracy, robustness, and efficiency in segmenting and quantifying the regions of interest.

DSC measures the overlap between two sets by calculating twice the intersection divided by the sum of the sizes of the two sets. JSC, also known as the Intersection over Union, is the ratio of the intersection of the sets to their union. SVD evaluates the symmetry of volume differences between two sets, providing insights into segmentation errors. VOE is the complement of the DSC and measures the volume not overlapping between two sets, often used to quantify segmentation accuracy.

To assess whether the proposed method achieved statistically significant improvements, we performed pairwise comparisons between the proposed method and other methods using a paired t-test on DSC, JSC, SVD, and VOE scores. The corresponding p-values were reported.

3 Data visualization and results

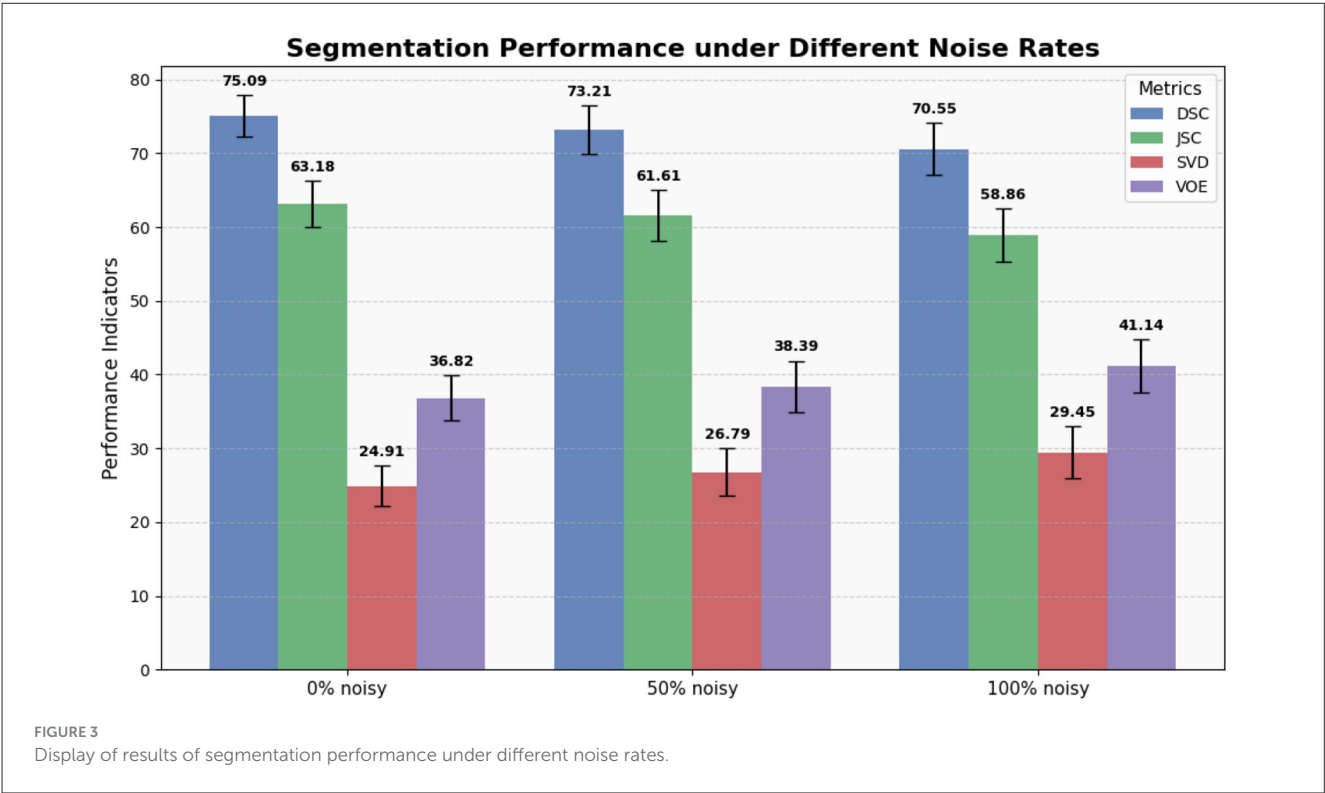
3.1 Negative impact of noisy label

Because introducing additional data with noisy labels may enhance the model's generalization ability, while noise also has the potential to impair generalization. Therefore, quantitatively analyzing the impact of noisy labels on the model becomes a pivotal aspect of our proposed noisy label handling framework. Hence, we conducted experiments with different noise rates (proportion of data with noisy labels in the training set) for evaluation. As shown in Table 1 and Figure 3, when the training data had a 0% noise rate, the U-Net with a ResNet34 (34) backbone achieved the best performance. The performance degraded with increasing noise rates, with the 50% and 100% noise rates resulting in the second and third-best performances, respectively. These results highlight that

TABLE 1 Comparison of segmentation performance of different noise rates.

Method	DSC (95% CI)	<i>p</i> -value	JSC (95% CI)	<i>p</i> -value	SVD (95% CI)	<i>p</i> -value	VOE (95% CI)	<i>p</i> -value
ResNet34 (0% noisy)	75.09 (72.30, 77.88)	-	63.18 (60.07, 66.28)	-	24.91 (22.12, 27.70)	-	36.82 (33.72, 39.93)	-
ResNet34 (50% noisy)	73.21 (69.97, 76.46)	0.068	61.61 (58.18, 65.04)	0.093	26.79 (23.54, 30.03)	0.068	38.39 (34.96, 41.82)	0.092
ResNet34 (100% noisy)	70.55 (67.01, 74.09)	0.000	58.86 (55.24, 62.48)	0.000	29.45 (25.91, 32.99)	0.000	41.14 (37.52, 44.76)	0.000

The best results are bolded.



data with noisy labels can significantly impair model performance, regardless of the quantity utilized. Therefore, it is crucial to develop robust techniques for handling noisy labels in training datasets to ensure the reliability and effectiveness of machine learning models.

3.2 Contrastive analysis

To validate the superior performance of our proposed method, we conducted a comparative analysis against several traditional methods designed for handling data with noisy labels (31, 34–37). The summarized results of these comparisons are presented in Table 2. We also plot column charts to provide a more intuitive visual comparison of performance indicators, as demonstrated in Figure 4.

Our experimental results clearly demonstrate that our method outperforms all other methods across various metrics, including DSC, JSC, SVD, and VOE. Particularly noteworthy

are the improvements in DSC and JSC, which are key indicators of segmentation performance, where our method exhibits enhancements of 1.79% and 1.93%, respectively, compared to the best-performing alternative methods. Moreover, to assess the impact of data containing noisy labels on network performance, we conducted experiments using varying amounts of data with noisy labels.

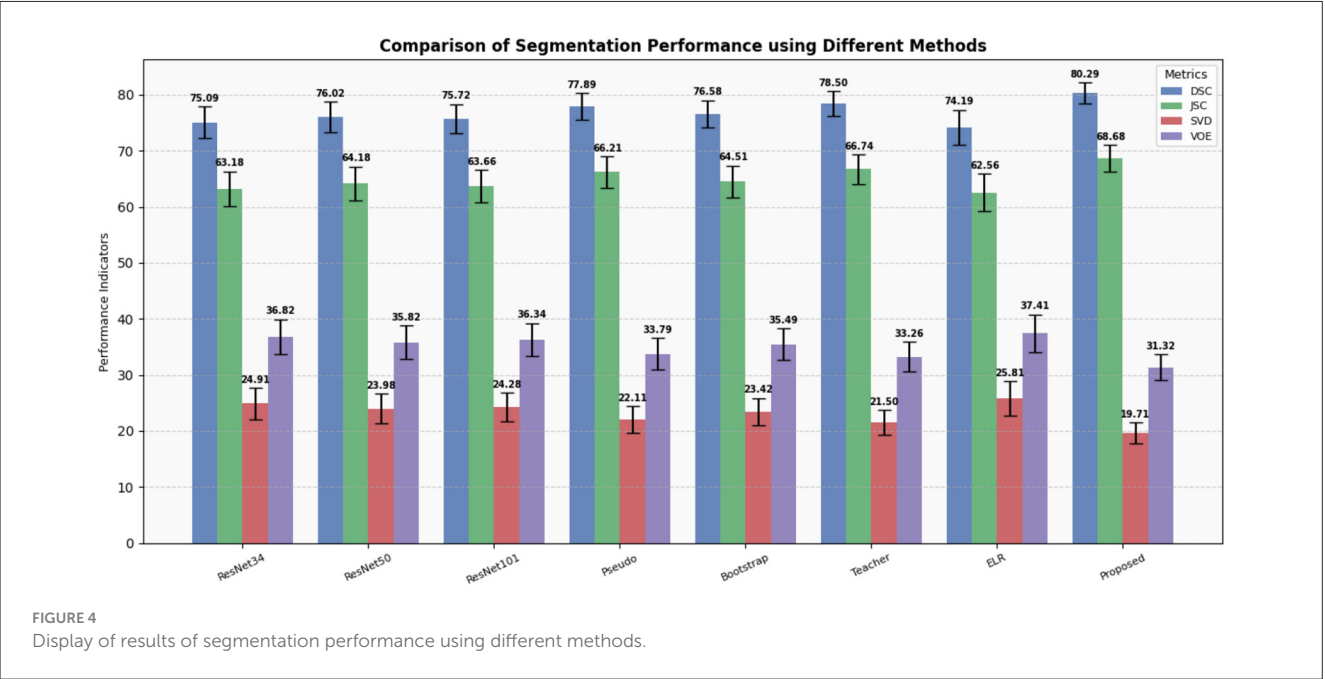
Furthermore, to investigate whether the observed improvements are solely due to the increased model parameters, we compared the performance of ResNet34 with ResNet50 and ResNet101 (34). The results suggest that the modest increase in parameters marginally enhances the model’s performance.

Therefore, our proposed method effectively leverages information from data with noisy labels, addressing the issue of deep learning models memorizing noisy information during training (38, 39), which can harm model performance. These findings suggest that our method can be a valuable tool in supporting the diagnosis of clinical liver tumors by providing auxiliary guidance.

TABLE 2 Comparison of segmentation performance of different methods for liver tumors.

Method	DSC (95% CI)	<i>p</i> -value	JSC (95% CI)	<i>p</i> -value	SVD (95% CI)	<i>p</i> -value	VOE (95% CI)	<i>p</i> -value
ResNet34 (34)	75.09 (72.30, 77.88)	0.000	63.18 (60.07, 66.28)	0.000	24.91 (22.12, 27.70)	0.000	36.82 (33.72, 39.93)	0.000
ResNet50 (34)	76.02 (73.32, 78.71)	0.000	64.18 (61.18, 67.18)	0.000	23.98 (21.29, 26.68)	0.000	35.82 (32.82, 38.82)	0.000
ResNet101 (34)	75.72 (73.11, 78.34)	0.000	63.66 (60.72, 66.60)	0.000	24.28 (21.66, 26.89)	0.000	36.34 (33.40, 39.28)	0.000
Pseudo (35)	77.89 (75.50, 80.28)	0.011	66.21 (63.41, 69.01)	0.019	22.11 (19.72, 24.50)	0.010	33.79 (30.99, 36.59)	0.020
Bootstrap (36)	76.58 (74.12, 79.03)	0.000	64.51 (61.69, 67.33)	0.000	23.42 (20.97, 25.88)	0.000	35.49 (32.67, 38.31)	0.000
Teacher (31)	78.50 (76.28, 80.72)	0.041	66.74 (64.10, 69.38)	0.023	21.50 (19.28, 23.72)	0.045	33.26 (30.62, 35.90)	0.022
ELR (37)	74.19 (71.10, 77.28)	0.000	62.56 (59.24, 65.88)	0.000	25.81 (22.72, 28.90)	0.000	37.44 (34.12, 40.76)	0.000
Proposed	80.29 (78.42, 82.17)	-	68.68 (66.35, 71.00)	-	19.71 (17.83, 21.58)	-	31.32 (29.00, 33.65)	-

The best results are bolded.

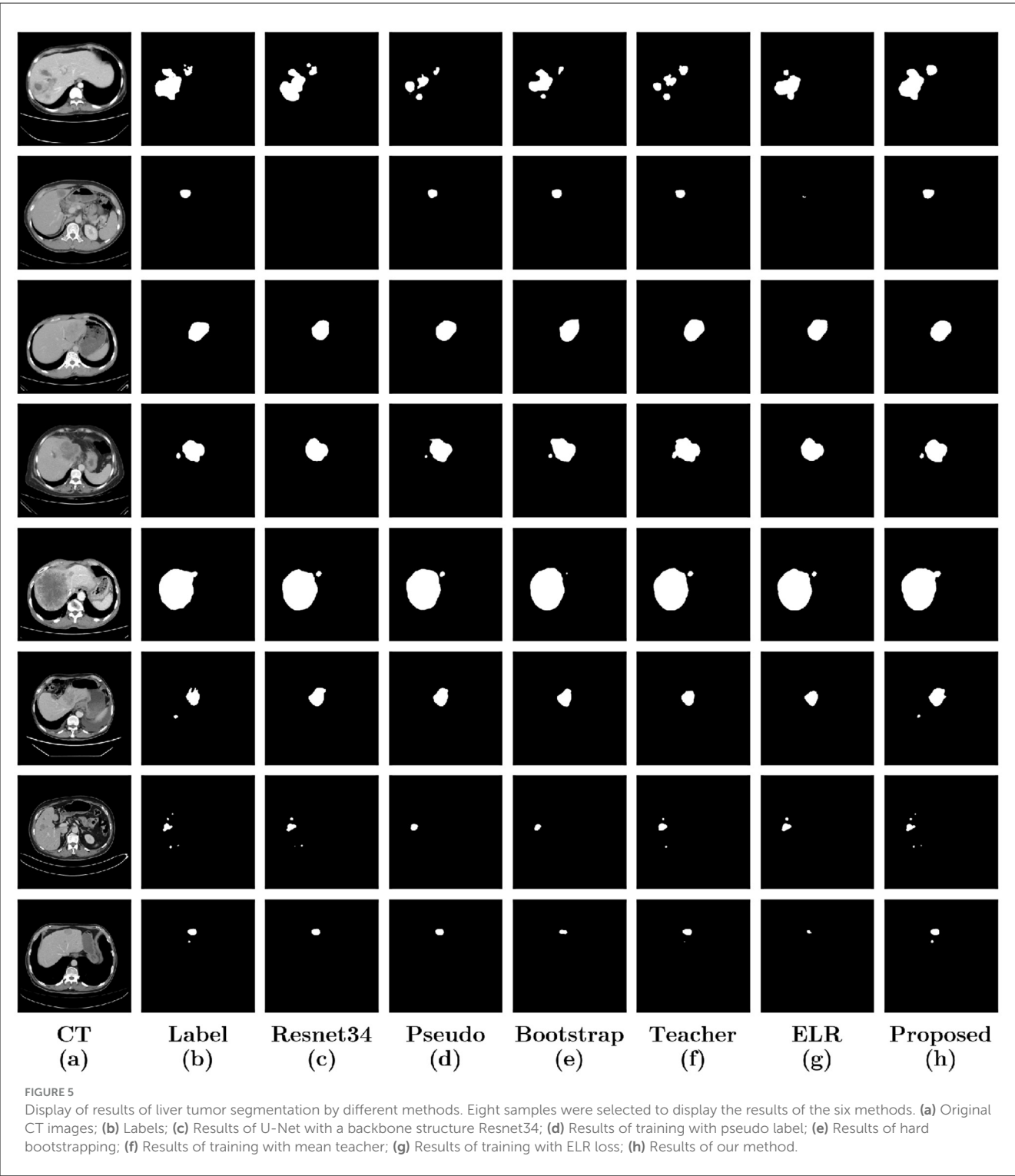


For a more intuitive qualitative assessment, Figure 5 illustrates eight liver tumor segmentation cases, each column representing a different stage of the segmentation process. The first column shows the original CT image, followed by manual annotations by a doctor in the second column. The third column presents the segmentation results using U-Net with the ResNet34 backbone. Subsequent columns demonstrate the impact of different techniques on the segmentation results, including pseudo-labeling, hard bootstrapping, mean teacher integration, ELR method for preventing noisy label memorization, and finally, our proposed method. A detailed examination of these columns reveals the effectiveness of our proposed method in addressing issues such as false positives, under-segmentation, and coarse boundaries across

various scenarios. Compared to alternative methods, our approach leads to notable improvements in liver segmentation accuracy.

3.3 Ablation analysis

To validate the effectiveness of our proposed method, we conducted an ablation study, the results of which are presented in Table 3. Our baseline model, depicted in the first row of Table 3, is the U-Net with backbone of ResNet34 trained using cross-entropy loss and dice loss. The ablation study focuses on evaluating the contributions of two key components: the noise adaptor



(NA) and the Multi-Stage Perturbations and Variable Teachers framework (MPVT).
The results of the ablation study indicate that each component contributes in a complementary manner to the overall performance. Specifically, the noise adaptor enhances accuracy by approximately 2.98% and 2.89% in terms of DSC and JSC, respectively, compared to the baseline model, indicating that the noise adaptor effectively learns label noise information,

thereby assisting the backbone network in learning without the perturbation caused by noise, which helps prevent the network from memorizing noisy data.
Furthermore, the implementation of the MPVT scheme resulted in improvements of approximately 4.18% and 4.07% in DSC and JSC, respectively, compared to the baseline model. Particularly noteworthy is the significant improvement in sensitivity, suggesting that the consistency regularization enhances

TABLE 3 Results of ablation analysis.

Method	DSC (95% CI)	JSC (95% CI)	SVD (95% CI)	VOE (95% CI)
ResNet34	75.09 (72.30, 77.88)	63.18 (60.07, 66.28)	24.91 (22.12, 27.70)	36.82 (33.72, 39.93)
Noise Adaptor	78.08 (75.92, 80.24)	66.08 (63.48, 68.68)	21.92 (19.76, 24.08)	33.92 (31.32, 36.52)
MPVT	79.27 (77.41, 81.13)	67.25 (64.92, 69.57)	20.73 (18.87, 22.59)	32.75 (30.43, 35.08)
w/o Input Perturbations	78.55 (76.28, 80.83)	66.85 (64.22, 69.49)	21.45 (19.17, 23.72)	33.15 (30.51, 35.78)
w/o Feature Perturbations	78.42 (76.30, 80.54)	66.44 (63.92, 68.97)	21.58 (19.46, 23.70)	33.56 (31.03, 36.08)
w/ One Teacher Model	77.92 (75.84, 80.00)	65.74 (63.20, 68.28)	22.08 (20.00, 24.16)	34.26 (31.72, 36.80)
Proposed Method	80.29 (78.42, 82.16)	68.68 (66.35, 71.00)	19.71 (17.83, 21.58)	31.32 (29.00, 33.65)

The best results are bolded.

the robustness of the backbone network and leverages latent information present in the CT images. In addition, to validate the improvement of model performance by multi-stage perturbations and variable teachers, we conducted ablation experiments on each perturbation and teacher model variability, respectively. The experimental results indicate that performance declines when any stage of perturbation is missing during training, compared to the complete MPVT. This demonstrates that the enhancement of model robustness from each stage of perturbation is indispensable. Specifically, when the training process lacks the variability of the teacher model, the model’s performance deteriorates the most. This underscores the necessity of using a variable teacher to enhance the model’s adaptability to strict multi-stage perturbations for training. These results indicate that training the network with the MPVT scheme enhances the network’s robustness.

In a comparative analysis, we have selected four cases depicted in Figure 6 to contrast seven distinct experiments. The results obtained from the baseline model exhibit a basic detection of lesion outlines and areas, with a tendency to mislabel regions resembling tumors and an inability to accurately identify all lesions. This limitation is attributed to the complex nature of liver CT images and the constraints of a limited training dataset.

In contrast, both the noise adaptor and MPVT approaches leverage additional data containing noisy labels, resulting in improved prediction quality. The noise adaptor tends toward conservative predictions, potentially overlooking some lesions, as seen in column 4 of Figure 6. On the other hand, MPVT leans toward bolder predictions, leading to over-expansion of the predicted area and mislabeling of other regions, as demonstrated in column 5 of Figure 6. Moreover, when multi-stage perturbations or variable teachers are removed, the model’s ability to segment details and recognize interference areas decreases. This reflects the impact

of lacking perturbations in the training process on the model’s generalization ability and robustness.

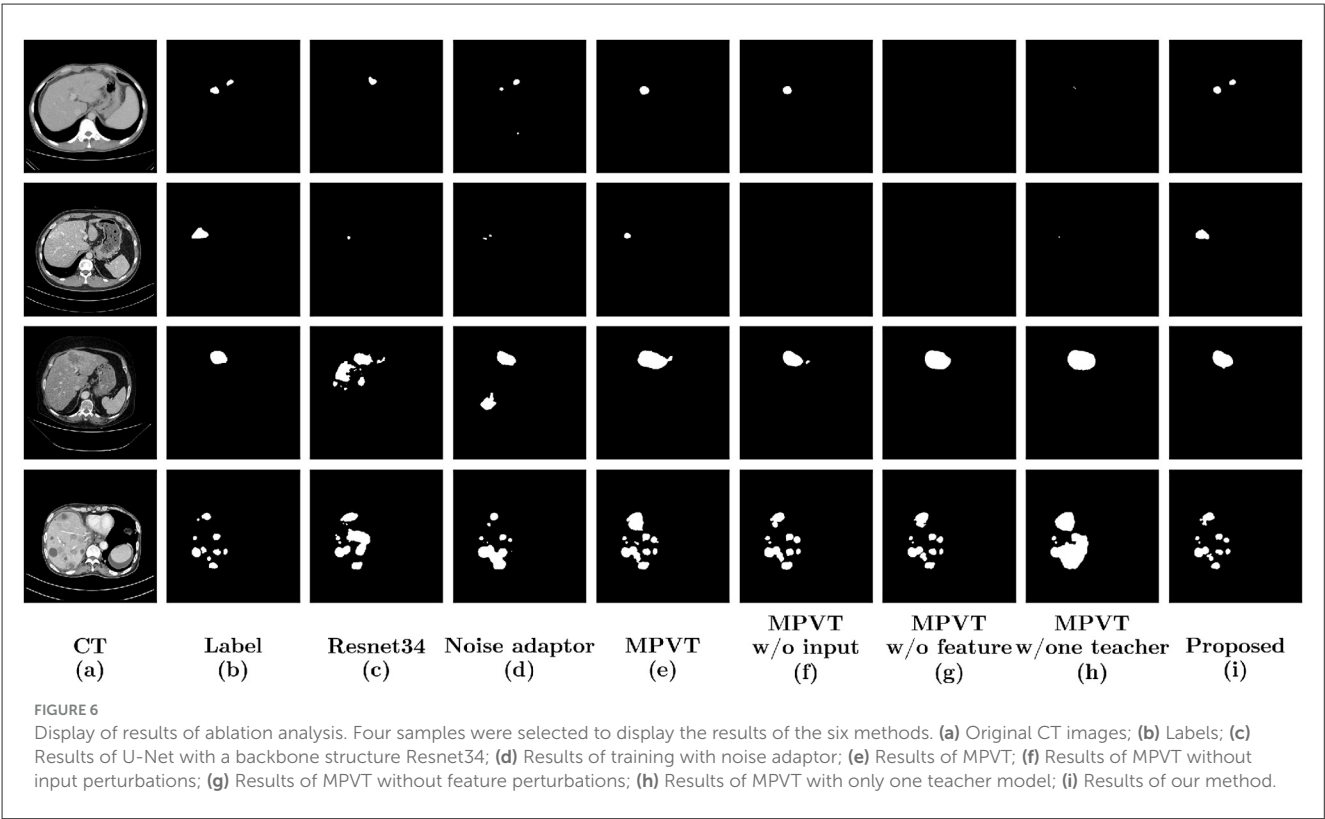
Our proposed method combines the advantages of both approaches, adopting a compromise strategy that yields the most favorable results among the four methods. This suggests that our approach, utilizing training with noise adaptor and MPVT schemes, enables these two distinct strategies to mitigate each other’s weaknesses and guide training outcomes toward a direction without bias.

4 Discussion

Automated liver-tumor segmentation remains hamstrung by the scarcity of perfectly annotated images and by the pernicious effect of label noise on deep neural networks. Here, we show that a multi-stage perturbation and variable-teacher (MPVT) consistency framework as well as a pixel-wise noise-adaptation module, termed MPVT+, enables convolutional networks to learn effectively from imperfect datasets, delivering state-of-the-art delineation accuracy. Across 739 retrospectively collected liver-tumor CT volumes, MPVT+ achieved the Dice similarity coefficient (DSC) from 75.1 % for a noise-free U-Net baseline to 80.3 %.

Various methods have been investigated to reduce the negative impact of noisy labels on medical image segmentation. For instance, obtaining annotations from multiple experts aims to establish a reliable ground truth but is limited by costs and remaining variability among annotators (10–13). Semi-automated tools can ease the annotation effort, although consistency issues among annotators still exist (14, 15). Utilizing historical hospital data provides scale but often suffers from inaccurate annotations (16, 17). Crowdsourcing involves more participants but typically introduces significant label noise (18, 19). To address noise during training, techniques such as designing robust models that account for latent noise parameters (9, 20), using custom loss functions to reduce the impact of incorrect labels (21), and applying transfer learning approaches that leverage knowledge from clean datasets (22) have been proposed. Additional regularization methods—such as label smoothing (23), help prevent the model from memorizing incorrect annotations (26). Building on these approaches, our MPVT+ framework integrates multiple components: a noise-adaptation module that explicitly models pixel-level corruption, multi-stage perturbations, and variable teachers that provide the network with self-supervised, noise-agnostic guidance. This combination advances liver-tumor segmentation beyond the performance of existing methods.

In addition, other recent studies have also emphasized the importance of robustness under imperfect or noisy annotations. Fang et al. (40) proposed a reliable mutual distillation framework, where two segmentation models collaborate to mitigate the impact of coarse or noisy labels through consistency constraints and sample selection strategies. Similarly, Qiu et al. (41) developed a hierarchical multimodal fusion framework that incorporates noisy label learning with attention mechanisms to enhance cancer classification performance from pathology and genomic data. Both works highlight complementary perspectives on leveraging imperfect supervision and noisy annotations, which aligns with



the design philosophy of MPVT+ that explicitly models label corruption while enforcing consistency. These directions together underscore a growing consensus that noise-robust strategies are crucial for advancing reliable medical image analysis in real-world clinical scenarios.

In the broader context of semi-supervised medical image segmentation, recent work has emphasized the integration of consistency learning with uncertainty modeling. Zhang et al. (25) proposed an uncertainty-guided mutual consistency framework that leverages intra-task and cross-task regularization while filtering unreliable predictions through uncertainty estimation, thereby enhancing the exploitation of unlabeled data. More recently, (24) introduced an uncertainty-aware consistency learning strategy that incorporates multi-level perturbations and perturbation-based uncertainty estimation to suppress noisy predictions and improve generalization across diverse segmentation tasks. Compared with these approaches, the present MPVT+ framework addresses robustness from a complementary perspective: Instead of relying solely on uncertainty-guided filtering, MPVT+ explicitly models label corruption through a noise adaptor while simultaneously enforcing multi-stage perturbations and variable-teacher consistency. This dual strategy allows the model to effectively utilize both reliable and noisy supervision, mitigating the adverse effects of corrupted labels. Taken together, these studies suggest that uncertainty-guided consistency learning and explicit noise-robust adaptation are not mutually exclusive but potentially synergistic directions. Future research may benefit from combining these paradigms to further advance the reliability and clinical applicability of automated tumor segmentation systems.

The devised framework presents a suite of significant enhancements to traditional deep learning models. Primarily, it utilizes a noise adaptor module that refurbishing noisy pixels. This reduces the detrimental impact of label noise on the training process without compromising the quantity of training samples, ensuring that the model can be trained on a vast and diverse dataset, crucial for the development of a robust liver tumor segmentation tool. In addition, the proposed module might be used for understanding mechanisms of noise generation and how to effectively eliminate the negative impact of noisy labels. Second, the introduction of variable teachers equips the model with the capability to generalize better by preventing it from focusing on any special subset of datasets and perturbations during training, thereby reducing overfitting to corner cases and enable the model to be trained with more strict perturbations. Third, the semi-supervised learning framework, along with strict perturbations, trains the network to be robust against data variability and capable of extracting more general and robust features. This addresses the weakness of the noise adaptor being susceptible to perturbations that could degrade its performance. Furthermore, by synthesizing the strengths of robust architecture with the flexibility of semi-supervised learning, our method demonstrates superior utilization of noisy-labeled data, significantly elevating the model's performance. Finally, the adaptability of the framework ensures its suitability across various extents of noise intensity and types, augmenting its potential application in diverse medical imaging contexts.

However, there are some limitations of our study. We did not consider discrepancies between annotators during the training process, as obtaining annotations from multiple radiologists and

pathologists is time-consuming. Future research will focus on incorporating factors such as the grading provided by each doctor to address label noise dependent on the annotator. In addition, while our implementation of the noise adaptor is effective, a more sophisticated architecture may be more adept at capturing intricate label noise patterns. Moreover, the use of multiple networks in the MPVT approach results in higher computational costs. Future research could explore training the network using a meta-learning framework to mitigate this issue. Finally, our study focused on scenarios where data with accurate labels could be differentiated from data with noisy labels, which can be a time-consuming process. Future research should investigate integrating mechanisms such as confidence curriculum learning to accelerate this data filtering process.

5 Conclusion

This study presents a novel deep learning framework that integrates a robust architecture with consistency regularization techniques to mitigate the adverse impact of noisy labels on DNNs training. By incorporating clean and noisy-labeled data, our approach enhances the model's robustness to noise and generalization capability, outperforming traditional methods and achieving better results. Our method shows promise for training segmentation CNNs tailored for detecting lesions in datasets affected by noisy labels. It effectively reduces the need for constructing a perfectly labeled dataset for DNNs training, serving as a valuable tool for aiding in clinical decision-making.

Data availability statement

The datasets presented in this article are not readily available because data from this study may contain potentially or sensitive patient information. Data are anonymized, nevertheless, due to relatively few severe cases, patients could be identified. Requests to access the datasets should be directed to Wei Yi, drweiyi057@163.com.

Ethics statement

The studies involving humans were approved by Institutional Review Board of West China Hospital of Sichuan University. The studies were conducted in accordance with the local legislation and institutional requirements. The Ethics Committee/Institutional Review Board waived the requirement of written informed consent for participation from the participants or the participants' legal guardians/next of kin because of the retrospective nature of the study.

Author contributions

XC: Conceptualization, Investigation, Methodology, Validation, Writing – original draft, Writing – review & editing. HT: Data curation, Methodology, Software, Visualization, Writing

– original draft. JZ: Data curation, Formal analysis, Software, Visualization, Writing – review & editing. TX: Investigation, Methodology, Visualization, Writing – review & editing. HG: Funding acquisition, Supervision, Validation, Writing – review & editing. ML: Conceptualization, Methodology, Project administration, Resources, Writing – review & editing. YW: Data curation, Investigation, Project administration, Resources, Supervision, Writing – review & editing. WL: Funding acquisition, Project administration, Supervision, Validation, Writing – review & editing.

Funding

The author(s) declare that financial support was received for the research and/or publication of this article. This work was supported by the National Key Research and Development Program under (Grant 2022YFB3104600), the Municipal Government of Quzhou (Grant 2023D007, 2023D014, 2023D033, 2023D034, 2023D035, 2024D058, and 2024D059), the China Postdoctoral Science Foundation Funded Project (2023M740519), and Guiding project of Quzhou Science and Technology Bureau (2022K50, 2023K013, and 2023K016).

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

The author(s) declared that they were an editorial board member of Frontiers, at the time of submission. This had no impact on the peer review process and the final decision.

Generative AI statement

The author(s) declare that Gen AI was used in the creation of this manuscript. During the preparation of this work, the author(s) used CHATGPT in order to improve language and readability. After using these tools/services, the author(s) reviewed and edited the content as needed and take(s) full responsibility for the content of the publication.

Any alternative text (alt text) provided alongside figures in this article has been generated by Frontiers with the support of artificial intelligence and reasonable efforts have been made to ensure accuracy, including review by the authors wherever possible. If you identify any issues, please contact us.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

References

- Sung H, Ferlay J, Siegel RL, Laversanne M, Soerjomataram I, Jemal A, et al. Global cancer statistics 2020: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. *CA Cancer J Clin.* (2021) 71:209–49. doi: 10.3322/caac.21660
- Siegel RL, Miller KD, Wagle NS, Jemal A. Cancer statistics, (2023). *CA Cancer J Clin.* (2023) 73:17–48. doi: 10.3322/caac.21763
- Bialecki ES, Bisceglie Di. Diagnosis of hepatocellular carcinoma. *HPB.* (2005) 7:26–34. doi: 10.1080/13651820410024049
- Christ PF, Elshaer MEA, Ettlinger F, Tatavarty S, Bickel M, Bilic P, et al. Automatic liver and lesion segmentation in CT using cascaded fully convolutional neural networks and 3D conditional random fields. In: *Medical Image Computing And Computer-Assisted Intervention-MICCAI* (2016). p. 415–423.
- Ma J, He Y, Li F, Han L, You C, Wang B. Segment anything in medical images. *Nat Commun.* (2024) 15:654. doi: 10.1038/s41467-024-44824-z
- Zhang L, Jindal B, Alaa A, Weinreb R, Wilson D, Segal E, et al. Generative AI enables medical image segmentation in ultra low-data regimes. *Nat Commun.* (2025) 16:6486. doi: 10.1038/s41467-025-61754-6
- Ashraf M, Robles WRQ, Kim M, Ko YS, Yi MY, A. loss-based patch label denoising method for improving whole-slide image analysis using a convolutional neural network. *Sci Rep.* (2022) 12:1392. doi: 10.1038/s41598-022-05001-8
- Shi J, Zhang K, Guo C, Yang Y, Xu Y, Wu J, et al. Survey of label-noise deep learning for medical image analysis. *Med Image Anal.* (2024) 95:103166. doi: 10.1016/j.media.2024.103166
- Song H, Kim M, Park D, Shin Y, Lee JG. Learning from noisy labels with deep neural networks: a survey. *IEEE Trans Neural Netw Learn Syst.* (2023) 34:8135–53. doi: 10.1109/TNNLS.2022.3152527
- Gulshan V, Peng L, Coram M, Stumpe MC, Wu D, Narayanaswamy A, et al. Development and validation of a deep learning algorithm for detection of diabetic retinopathy in retinal fundus photographs. *JAMA.* (2016) 316:2402. doi: 10.1001/jama.2016.17216
- Esteve A, Kuprel B, Novoa RA, Ko J, Swetter SM, Blau HM, et al. Dermatologist-level classification of skin cancer with deep neural networks. *Nature.* (2017) 542:115–8. doi: 10.1038/nature21056
- Warfield SK, Zou KH, Wells WM. Simultaneous truth and performance level estimation (STAPLE): an algorithm for the validation of image segmentation. *IEEE Trans Med Imaging.* (2004) 23:903–21. doi: 10.1109/TMI.2004.828354
- Rohlfing T, Russakoff DB, Maurer CR. Performance-based classifier combination in atlas-based image segmentation using expectation-maximization parameter estimation. *IEEE Trans Med Imaging.* (2004) 23:983–94. doi: 10.1109/TMI.2004.830803
- Deng Y, Sun Y, Zhu Y, Xu Y, Yang Q, Zhang S, et al. A new framework to reduce doctor workload for medical image annotation. *IEEE Access.* (2019) 7:107097–104. doi: 10.1109/ACCESS.2019.2917932
- Joskowicz L, Cohen D, Caplan N, Sosna J. Inter-observer variability of manual contour delineation of structures in CT. *Eur Radiol.* (2019) 29:1391–9. doi: 10.1007/s00330-018-5695-5
- Yan K, Wang X, Lu L, Summers RM. DeepLesion: automated mining of large-scale lesion annotations and universal lesion detection with deep learning. *J Med Imaging.* (2018) 5:1. doi: 10.1117/1.JMI.5.3.036501
- Irvine J, Rajpurkar P, Ko M, Yu Y, Ciurea-Ilcus S, Chute C, et al. Chexpert: a large chest radiograph dataset with uncertainty labels and expert comparison. In: *Proceedings of the AAAI Conference on Artificial Intelligence* (2019). p. 590–597.
- Kaur D, Kaur Y. Various image segmentation techniques: a review. *Int J Comp Sci Mobile Comp.* (2014) 3:809–14.
- Gurari D, Theriault D, Sameki M, Isenberg B, Pham TA, Purwada A, et al. How to collect segmentations for biomedical images? A benchmark evaluating the performance of experts, crowdsourced non-experts, and algorithms. In: *IEEE Winter Conference on Applications of Computer Vision.* (2015). p. 1.
- Frénay B, Verleysen M. Classification in the presence of label noise: a survey. *IEEE Trans Neural Netw Learn Syst.* (2013) 25:845–69. doi: 10.1109/TNNLS.2013.2292894
- Zhang Z, Sabuncu M. Generalized cross entropy loss for training deep neural networks with noisy labels. In: *Advances in Neural Information Processing Systems* (2018). p. 31.
- Shin HC, Roth HR, Gao M, Lu L, Xu Z, Nogueira I, et al. Deep convolutional neural networks for computer-aided detection: CNN architectures, dataset characteristics and transfer learning. *IEEE Trans Med Imaging.* (2016) 35:1285–98. doi: 10.1109/TMI.2016.2528162
- Karimi D, Rollins CK, Velasco-Annis C, Ouaalam A, Gholipour A. Learning to segment fetal brain tissue from noisy annotations. *Med Image Anal.* (2023) 85:102731. doi: 10.1016/j.media.2022.102731
- Dong M, Yang A, Wang Z, Li D, Yang J, Zhao R. Uncertainty-aware consistency learning for semi-supervised medical image segmentation. *Knowl-Based Syst.* (2025) 309:112890. doi: 10.1016/j.knsys.2024.112890
- Zhang Y, Jiao R, Liao Q, Li D, Zhang J. Uncertainty-guided mutual consistency learning for semi-supervised medical image segmentation. *Artif Intell Med.* (2023) 138:102476. doi: 10.1016/j.artmed.2022.102476
- Müller R, Kornblith S, Hinton GE. When does label smoothing help? In: *Advances in Neural Information Processing Systems* (2019). p. 32.
- Xiao T, Xia T, Yang Y, Huang C, Wang X. Learning from massive noisy labeled data for image classification. In: *IEEE Conference on Computer Vision and Pattern Recognition (CVPR).* (2015). p. 6.
- Goldberger J, Ben-Reuven E. Training deep neural-networks using a noise adaptation layer. In: *International Conference on Learning Representations* (2016).
- Srivastava N, Hinton G, Krizhevsky A, Sutskever I, Salakhutdinov R. Dropout: a simple way to prevent neural networks from overfitting. *J Mach Learn Res.* (2014) 15:1929–58.
- Ronneberger O, Fischer P, Brox T. U-net: Convolutional networks for biomedical image segmentation. In: *Medical Image Computing and Computer-Assisted Intervention-MICCAI 2015: 18th International Conference.* Munich: Springer (2015). p. 234–241.
- Tarvainen A, Valpola H. Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results. In: *Advances in Neural Information Processing Systems.* (2017). p. 30.
- French G, Laine S, Aila T, Mackiewicz M, Finlayson G. Semi-supervised semantic segmentation needs strong, varied perturbations. *arXiv [preprint] arXiv:1906.01916.* (2019). doi: 10.5244/C.34.154
- Yeghiazaryan V, Voiculescu I. Family of boundary overlap metrics for the evaluation of medical image segmentation. *J Med Imaging.* (2018) 5:015006–015006. doi: 10.1117/1.JMI.5.1.015006
- He K, Zhang X, Ren S, Sun J. Deep residual learning for image recognition. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (2016). p. 770–778.
- Lee DH, et al. Pseudo-label: The simple and efficient semi-supervised learning method for deep neural networks. In: *Workshop on Challenges in Representation Learning.* Atlanta: ICML (2013). p. 896.
- Reed S, Lee H, Angelov D, Szegedy C, Erhan D, Rabinovich A. Training deep neural networks on noisy labels with bootstrapping. *arXiv [preprint] arXiv:1412.6596.* (2014). doi: 10.48550/arXiv.1412.6596
- Liu S, Niles-Weed J, Razavian N, Fernandez-Granda C. Early-learning regularization prevents memorization of noisy labels. *Adv Neural Inf Process Syst.* (2020) 33:20331–42.
- Arpit D, Jastrzebski S, Ballas N, Krueger D, Bengio E, Kanwal MS, et al. A closer look at memorization in deep networks. In: *International Conference on Machine Learning.* New York: PMLR (2017). p. 233–242.
- Xia X, Liu T, Han B, Gong C, Wang N, Ge Z, et al. Robust early-learning: hindering the memorization of noisy labels. In: *International Conference on Learning Representations* (2020).
- Fang C, Wang Q, Cheng L, Gao Z, Pan C, Cao Z, et al. Reliable mutual distillation for medical image segmentation under imperfect annotations. *IEEE Trans Med Imaging.* (2023) 42:1720–34. doi: 10.1109/TMI.2023.3237183
- Qiu L, Zhao L, Hou R, Zhao W, Zhang S, Lin Z, et al. Hierarchical multimodal fusion framework based on noisy label learning and attention mechanism for cancer classification with pathology and genomic features. *Comput Med Imag Graph.* (2023) 104:102176. doi: 10.1016/j.compmedimag.2022.102176