Pick-and-Learn: Automatic Quality Evaluation for Noisy-Labeled Image Segmentation

2020.7.22

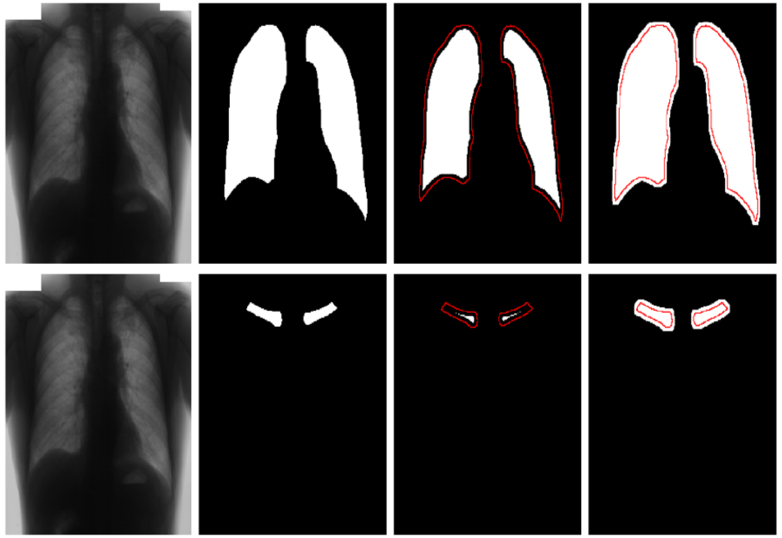
Haidong Zhu1, Jialin Shi1, and Ji Wu1,2

MICCAI 2019

Content

motivation

- The label of segmentation data set is very **easy to make mistakes**

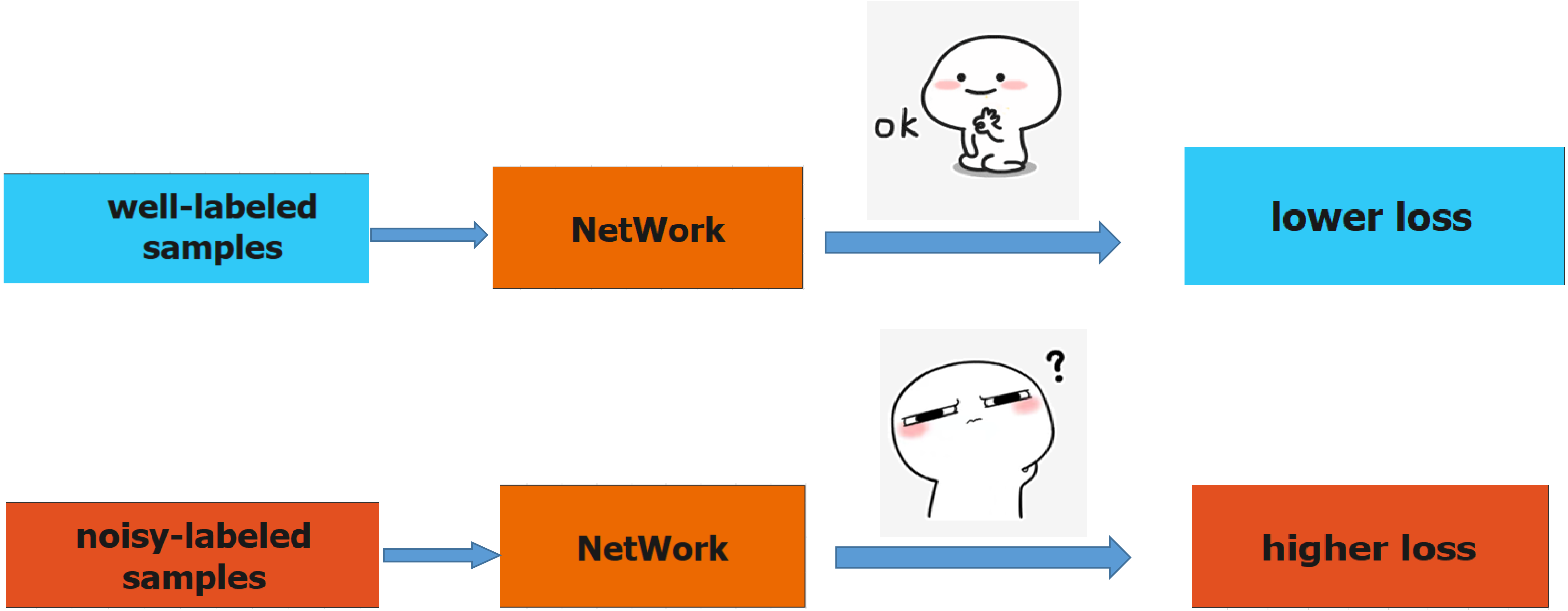


- **Most** of the **current solutions** for low quality annotations **are for classification tasks** and cannot be applied directly to segmentation tasks.

Method

Assumption

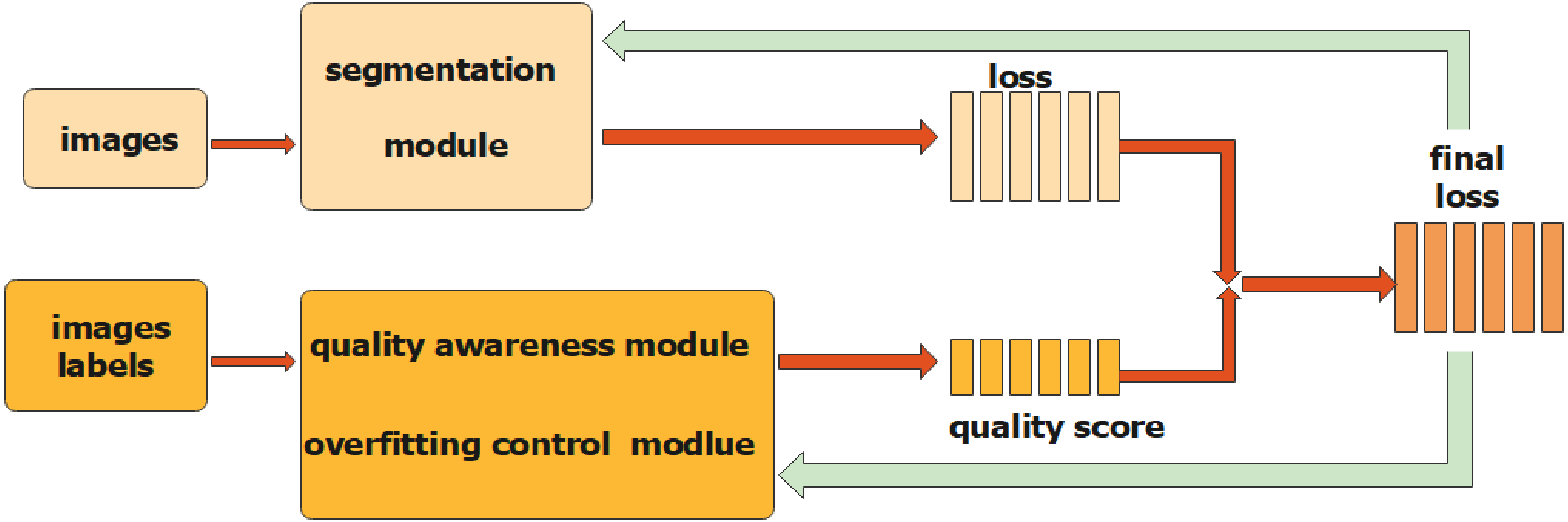
- during the training process, the predicted segmentation for the **noisy labeled samples** might have a **higher loss** compared with well-annotated ones.



Method

Framework

- Add a parallel branch to evaluate the quality score of each sample.
- Multiply the **loss of each sample** by **corresponding quality score** to obtain the final loss.
- Use the final loss to update the network.

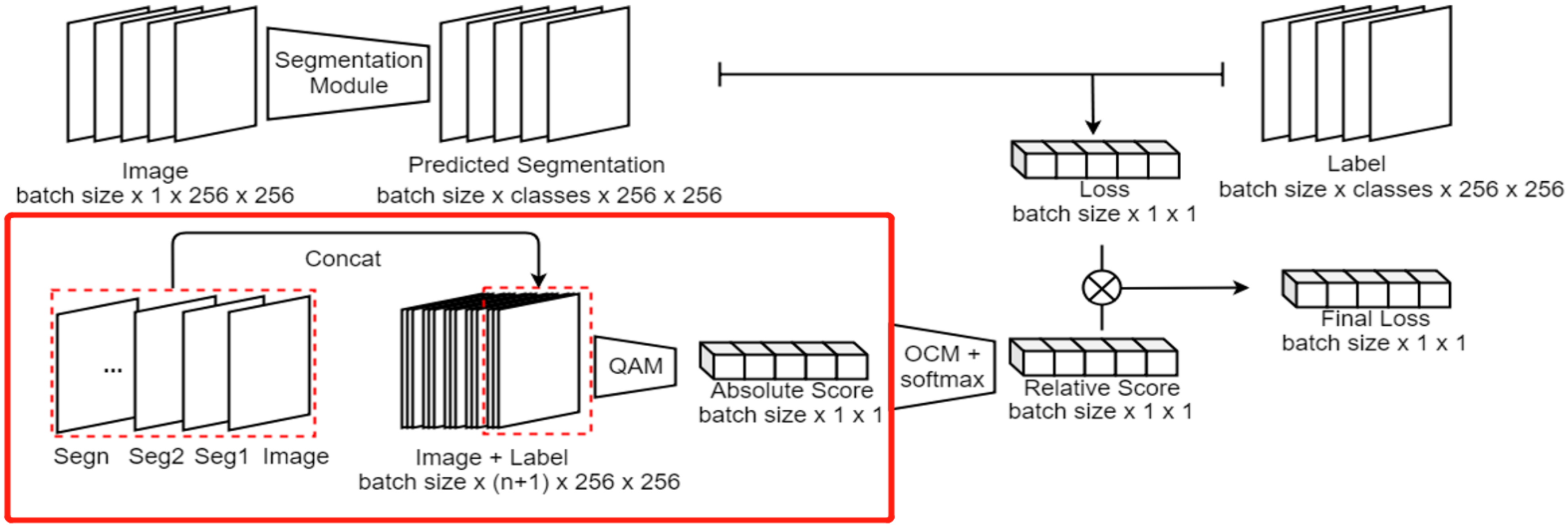


Method

Quality Awareness Module

- Use the VGG based network as the quality awareness module.

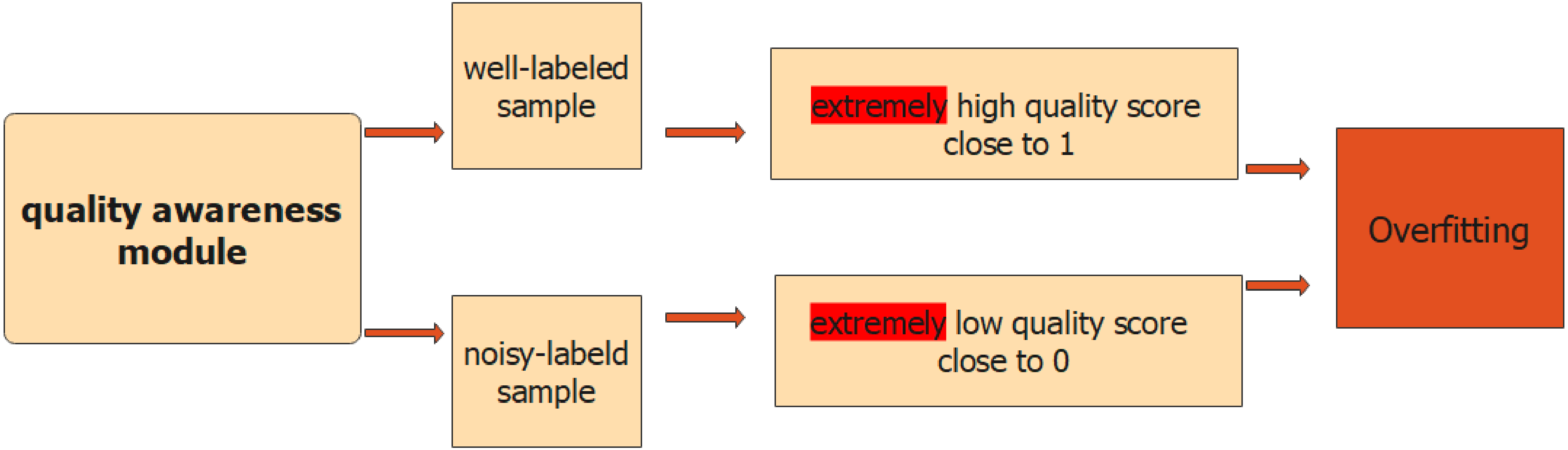
- Input QAM a batch of images and labels, the QMA output the quality score of each sample.



Method

The drawback of QAM

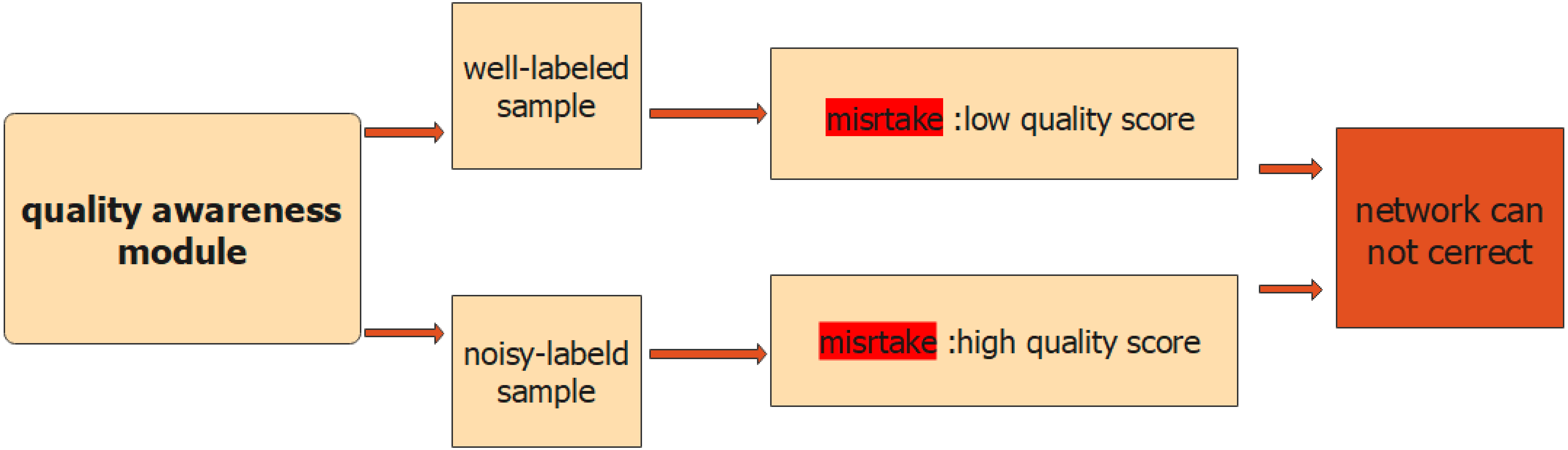
- On the one hand, the QAM might give the clear sample extremely high score (close to 1) and give the noisy sample extremely low score (close to 0), resulting in **overfitting.**



Method

The drawback of QAM

- On the other hand, the QAM might make mistakes and give the clear sample extremely low score and give the noisy sample extremely low score, resulting in the **network can not correct**.

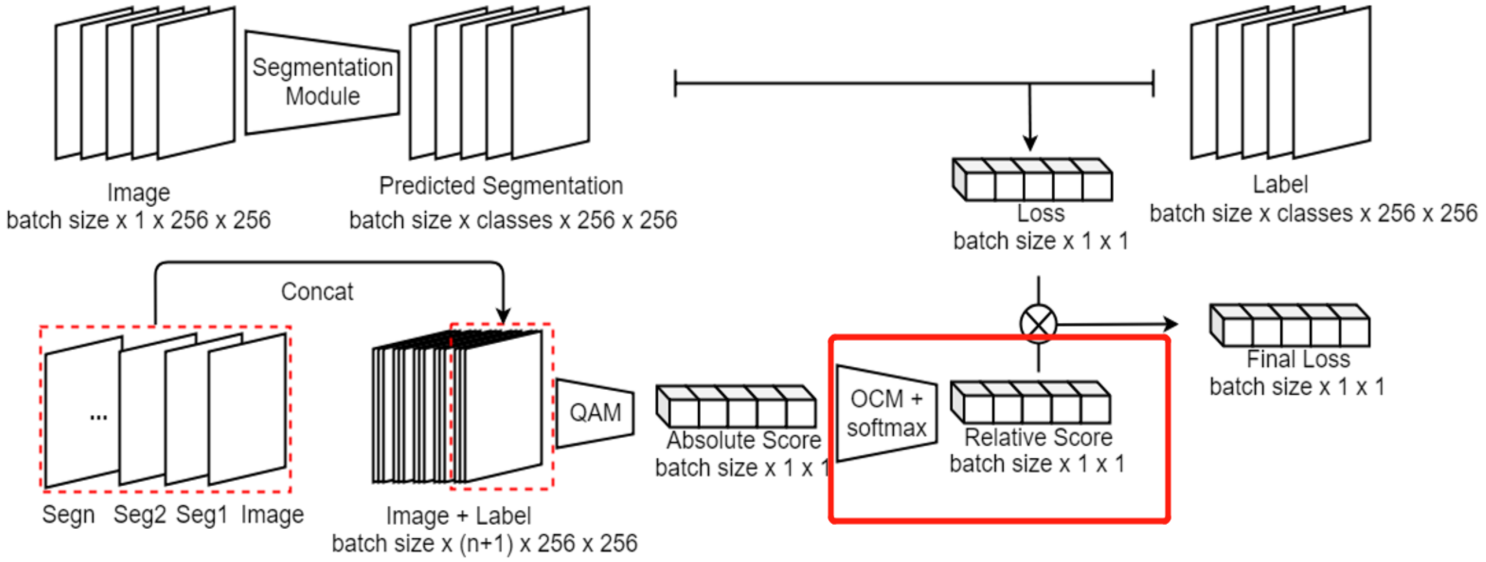


Method

Overfitting Control Module

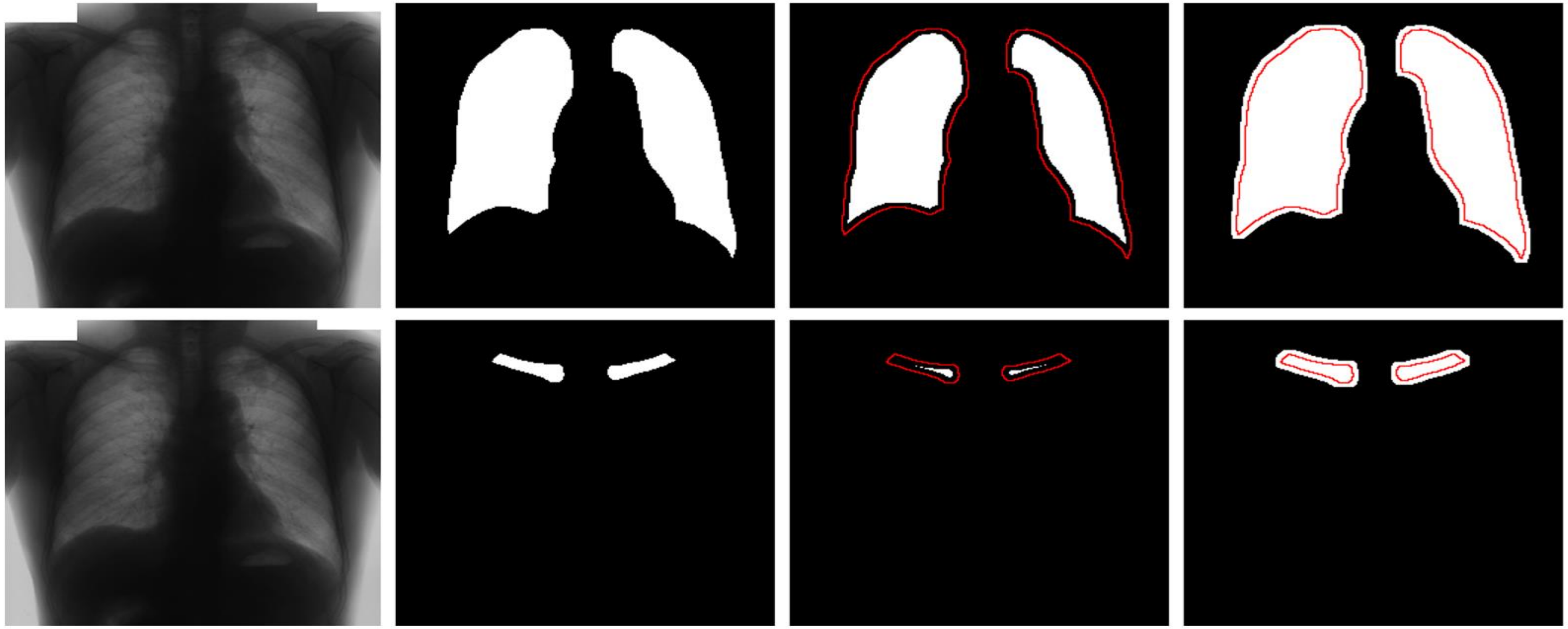
$$\Phi(t) = \lambda tanh(t)$$

This will rescale the quality score from $(-\infty, \infty)$ to $(-\lambda, \lambda)$.



Experiment

- Dataset: medical image dataset JSRT.
- Preprocessing: dilation and erosion.
  - we randomly selected 0%, 25%, 50% and 75% samples from the training set and further randomly eroded or dilated them with $1 \leqslant ni \leqslant 8$ and $5 \leqslant ni \leqslant 13$ pixels.



Experiment

Noise percentage: The percentage of the data set add noise

Noise level: The range of dilation and erosion

Table 1. Results on JSRT dataset

Noise percentage	Noise level	Strategy	Lungs	Heart	Clavicles	Average
No noise	-	baseline	0.943	0.941	0.862	0.915
No noise	-	QAM	0.939	0.923	0.831	0.898
No noise	-	QAM+OCM	0.941	0.940	0.852	0.911
25% noise	$1 \leq n_i \leq 8$	baseline	0.868	0.888	0.538	0.765
25% noise	$1 \leq n_i \leq 8$	QAM	0.925	0.926	0.748	0.866
25% noise	$1 \leq n_i \leq 8$	QAM+OCM	0.936	0.925	0.823	0.895
50% noise	$1 \leq n_i \leq 8$	baseline	0.873	0.884	0.539	0.765
50% noise	$1 \leq n_i \leq 8$	QAM	0.922	0.925	0.726	0.857
50% noise	$1 \leq n_i \leq 8$	QAM+OCM	0.936	0.929	0.828	0.898
75% noise	$1 \leq n_i \leq 8$	baseline	0.820	0.828	0.512	0.720
75% noise	$1 \leq n_i \leq 8$	QAM	0.898	0.825	0.536	0.753
75% noise	$1 \leq n_i \leq 8$	QAM+OCM	0.937	0.939	0.809	0.895
25% noise	$5 \leq n_i \leq 13$	baseline	0.865	0.857	0.422	0.715
25% noise	$5 \leq n_i \leq 13$	QAM	0.893	0.835	0.615	0.781
25% noise	$5 \leq n_i \leq 13$	QAM+OCM	0.935	0.935	0.801	0.890
50% noise	$5 \leq n_i \leq 13$	baseline	0.755	0.807	0.393	0.652
50% noise	$5 \leq n_i \leq 13$	QAM	0.828	0.853	0.491	0.714
50% noise	$5 \leq n_i \leq 13$	QAM+OCM	0.942	0.942	0.853	0.912
75% noise	$5 \leq n_i \leq 13$	baseline	0.745	0.738	0.381	0.621
75% noise	$5 \leq n_i \leq 13$	QAM	0.770	0.772	0.366	0.636
75% noise	$5 \leq n_i \leq 13$	QAM+OCM	0.938	0.937	0.801	0.892

Experiment

- As the number of training rounds increases, the **weight of clean-annotated** samples **increases** and the **weight of noisy-annotated** samples **decreases**.

- As the number of training rounds increases, **Variances** for clean and noisy-annotated samples are both **going down**.

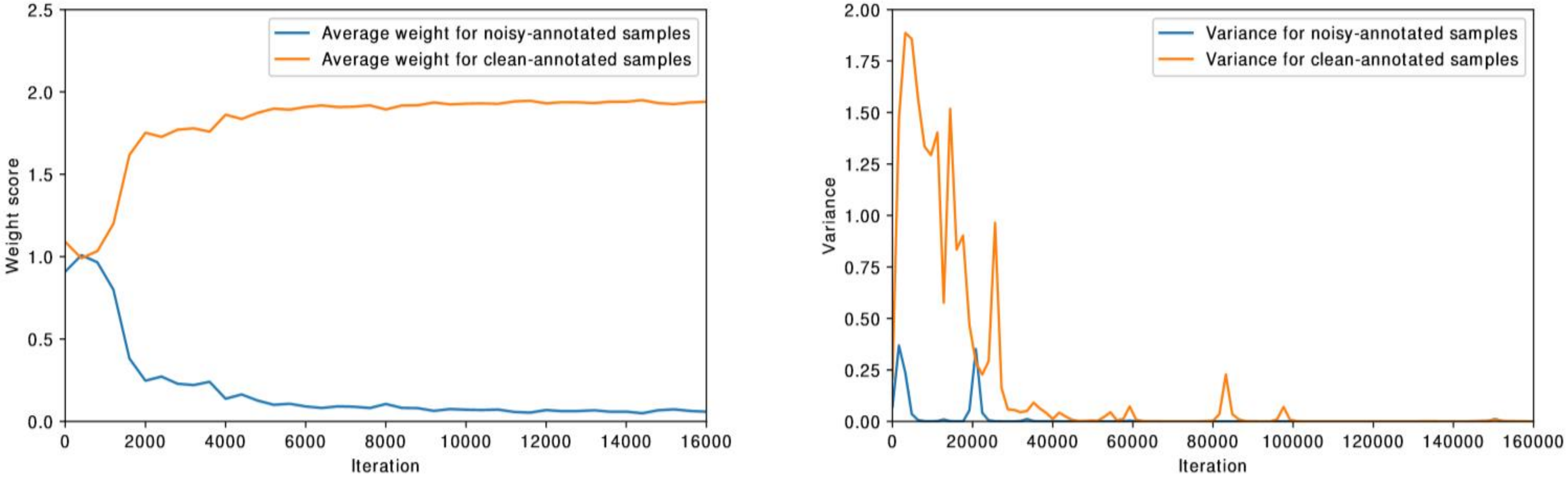


Fig. 4. Relative weights and variances for clean and noisy-labeled data.

Advantages:

- This work is novel for segmentation work.
- The experiment results are good.

Disadvantages:

- Only use one dataset and the size of dataset is too small.
- The added noise is relatively simple.

Thanks