CAMEL: A Weakly Supervised Learning Framework for Histopathology Image Segmentation

Gang Xu[1], Zhigang Song[2], Zhuo Sun[3], Calvin Ku[3], Zhe Yang[1], Cancheng Liu[3], Shuhao Wang[1,3], Jianpeng Ma[4*], Wei Xu[1*]

[1]Tsinghua University [2]The Chinese PLA General Hospital [3]Thorough Images [4]Fudan University

xug14@mails.tsinghua.edu.cn, songzhg301@139.com,
{zhuo.sun,calvin.j.ku,liucancheng,eric.wang}@thorough.ai, jpma@fudan.edu.cn,
{yangzhe2017,weixu}@tsinghua.edu.cn

Abstract

Histopathology image analysis plays a critical role in cancer diagnosis and treatment. To automatically segment the cancerous regions, fully supervised segmentation algorithms require labor-intensive and time-consuming labeling at the pixel level. In this research, we propose CAMEL, a weakly supervised learning framework for histopathology image segmentation using only image-level labels. Using multiple instance learning (MIL)-based label enrichment, CAMEL splits the image into latticed instances and automatically generates instance-level labels. After label enrichment, the instance-level labels are further assigned to the corresponding pixels, producing the approximate pixel-level labels and making fully supervised training of segmentation models possible. CAMEL achieves comparable performance with the fully supervised approaches in both instance-level classification and pixel-level segmentation on CAMELYON16 and a colorectal adenoma dataset. Moreover, the generality of the automatic labeling methodology may benefit future weakly supervised learning studies for histopathology image analysis.

1. Introduction

Histopathology image analysis is the gold standard for cancer detection and diagnosis. In recent years, the development of deep neural network has achieved many breakthroughs in automatic histopathology image classification and segmentation [15, 18, 19]. These methods highly depend on the availability of a large number of pixel-level labels, which are labor-intensive and time-consuming to obtain.

To relieve the demand for theses fine-grained labels, people have proposed many weakly supervised learning algorithms only requiring coarse-grained labels at the image-level [13, 25, 26]. However, due to the lack of sufficient supervision information, the accuracy is much lower than their fully supervised counterparts. One way to improve the performance of weakly supervised learning algorithms is to add more supervision constraints. For natural images, some studies [8, 14, 16] have proven the effectiveness of adding bounding boxes or scribble information artificially in their weakly supervised learning process. CDWS-MIL [13] has also shown the advantage of artificial area constraints for weakly supervised histopathological segmentation. However, it still takes much effort to obtain artificial constraints, especially in histopathology, where only well-trained pathologists can distinguish the cancerous regions from the normal ones. Therefore, automatically enriching labeling information instead of introducing artificial constraints before building the segmentation model is crucial for weakly supervised learning.

In this paper, we propose a weakly supervised learning framework, CAMEL, for histopathology image segmentation using only image-level labels. CAMEL consists of two steps: label enrichment and segmentation (Fig. 1). Instead of introducing more supervision constraints, CAMEL splits the image into latticed *instances* and automatically generates their instance-level labels in the label enrichment step, which can be regarded as a solution for a weakly supervised classification problem. In the label enrichment step, we use a combined multiple instance learning (cMIL) approach to construct a high-quality instance-level dataset with instance-level labels from the original image-level dataset. Then, we train a fully supervised classification

*Corresponding author.

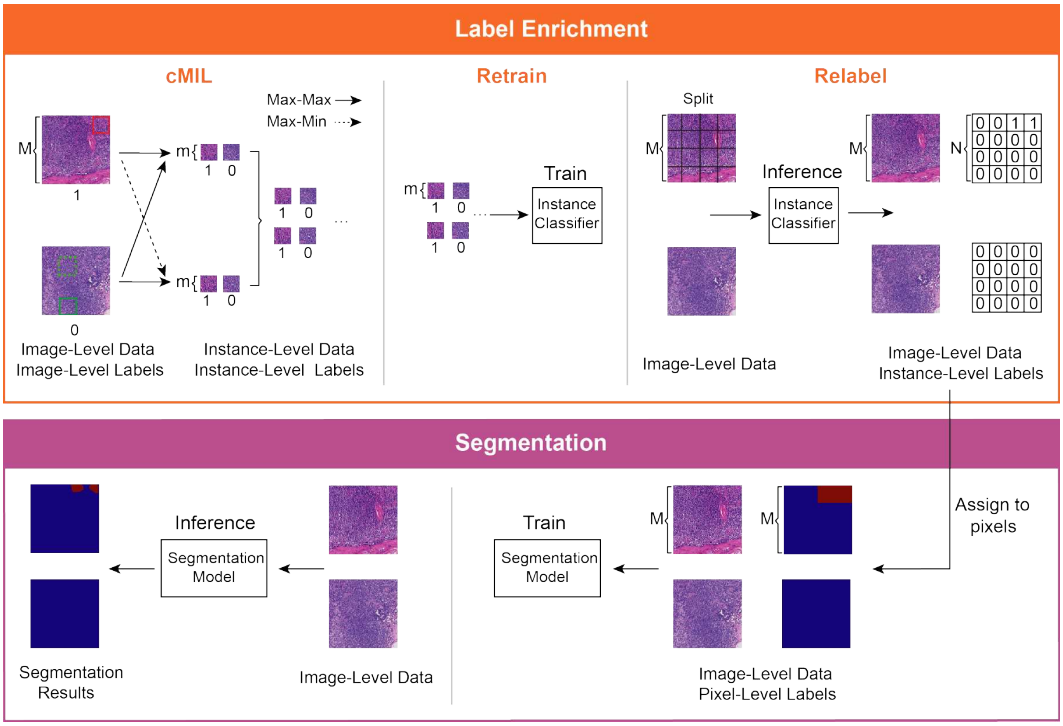

Figure 1. System architecture of CAMEL. CAMEL consists of two basic steps: label enrichment and segmentation. M and m represent the size of the image and the instance, respectively. N is the *scale factor* of cMIL where $N = \frac{M}{m}$.

model using this instance-level dataset. Once the model is trained, we split the images in the original image-level dataset into latticed instances and use this model to generate their labels. After label enrichment, the instance-level labels are directly assigned to their corresponding pixels, producing the approximate pixel-level labels and making fully supervised training of segmentation models possible. We conducted our experiments on CAMELYON16 [1, 5] and a colorectal adenoma dataset, the results of both instance-level classification and pixel-level segmentation were comparable with their fully supervised counterparts.

The contributions of this paper can be summarized as follows:

- We propose a weakly supervised learning framework, CAMEL, for histopathology image segmentation using only image-level labels. CAMEL automatically enriches supervision information of the image by generating the instance-level labels from the image-level ones and achieves comparable performance with the fully supervised baselines in both instance-level classification and pixel-level segmentation.

- To construct a high-quality instance-level dataset for fully supervised learning, we introduce a cMIL approach which combines two complementary instance selection criteria (Max-Max and Max-Min) in the data preparation process to balance the data distribution in the constructed dataset.

- To fully utilize the original image-level supervision information, we propose the cascade data enhancement method and add image-level constraints to boost the performance of CAMEL further.

- To facilitate the research in histopathology field, our colorectal adenoma dataset will be made publicly available at https://github.com/ThoroughImages/CAMEL.

2. Related Work

2.1. Weakly Supervision in Computer Vision

In computer vision, people have proposed many weakly supervised algorithms [3, 4, 9, 10, 12, 22, 23] for object detection and semantic segmentation. However, in histopathology image analysis scenarios, the difference of morphological appearance between foreground (cancerous region) and background (non-cancerous region) is less significant [17] compared to what is usually observed in natural images. Moreover, the cancerous regions are disconnected, and their morphologies are usually various. Therefore, the methods based on adversarial erasing [22] or seed growing [4] may not be suitable.

2.2. Weakly Supervision in Histopathology Image

2.2.1 Instance-Level Classification

MIL is widely applied in most weakly supervised methodologies [13, 25, 26]. However, despite the great success of MIL, many solutions need pre-specified features [21, 26], which require data specific prior knowledge and limit the general applications. Instead of using pre-specified features, Xu et al. [25] proposed to extract feature representations through a deep neural network automatically. However, the separation between feature engineering and MIL complicates the training process. In cMIL, the training procedure is end-to-end without deliberate feature extraction and feature learning, making the training process straightforward.

2.2.2 Pixel-Level Segmentation

Weakly supervised learning for histopathology image segmentation [13] has been proposed in recent years. The best performance was achieved by introducing artificial cancer area constraints. In CAMEL, the label enrichment step generates instance-level labels with more detailed supervision information and less labeling burden. In addition, compared to CDWS-MIL [13], the classifier in CAMEL does not need pre-training and thus increases the flexibility in choosing the network architecture.

3. Method

3.1. Label Enrichment

Due to the lack of sufficient supervision information, simply using the image-level labels is insufficient to train the segmentation model. Therefore, before building the segmentation model, we perform a label enrichment procedure by generating instance-level labels from the original image-level labels (see Fig. 1).

3.1.1 Combined Multiple Instance Learning

The effectiveness of CAMEL closely depends on the quality of our automatically enriched instance-level labels, which can also be regarded as a weakly supervised instance-level classification problem with only image-level labels. Here, we aim to transform this weakly supervised learning problem into a fully supervised instance-level classification one, and benefit from many existing well-developed fully supervised learning methods.

We introduce a new solution called *combined Multiple Instance Learning* (cMIL). The image is split into $N \times N$ latticed instances with equal size. Here, we consider the instances from the same image as in the same *bag*. In cMIL, two MIL-based classifiers with different instance selection criteria (Max-Max and Max-Min) are used to select

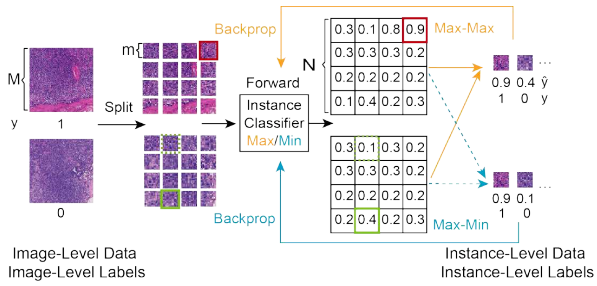

Figure 2. Training procedure of cMIL. M and m represent the size of the image and the instance, respectively. N is the *scale factor* of cMIL where $N = \frac{M}{m}$, here we require M to be divisible by m. We first split the image into $N \times N$ latticed instances with equal size. The selected instance can be considered as the representative of its corresponding image, therefore they own the same class label. We train two MIL models separately using two instance selection criteria (Max-Max and Max-Min).

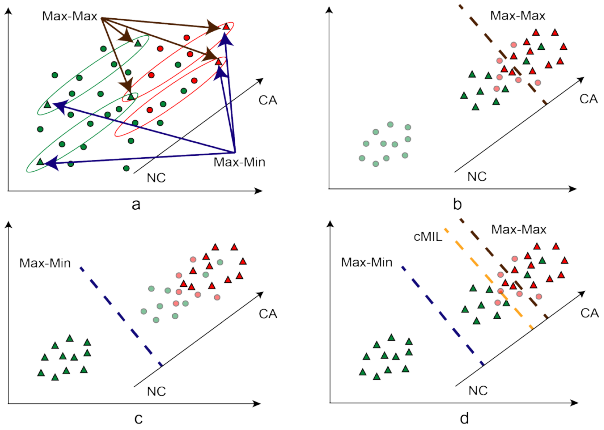

Figure 3. Intuition behind two instance selection criteria named Max-Max and Max-Min. Red and green circles represent the CA and NC instances, respectively. We use triangles to represent the selected instances, and circles with light colors to represent the instances that are not selected. Each dotted line represents the decision boundary of the classifier, which is trained with the selected instances. Each ellipse represents an image (or a bag in MIL). cMIL, which combines Max-Max and Max-Min, achieves a better decision boundary.

instances to construct the instance-level dataset (Fig. 2). The selected instance can be considered as the representative of its corresponding image, which determines the image class (similar to the attention mechanism [24]).

If the image has a cancerous (CA) region, we can reason that at least one instance is cancerous. On the other hand, if the label of the image is non-cancerous (NC), all the instances in it are non-cancerous. For both CA and NC images, Max-Max selects the instance with maximum CA response. As shown in Fig. 3(a) and (b), during the training

stage, in NC region, the Max-Max criterion will select the instance with maximum CA response, which has the highest similarity with CA, as the NC example. Therefore, the model trained with these data would give a decision boundary toward the CA direction, and this would lead to misclassification of CA instances with lower responses (as shown by light red circles). For example, CA instances with similar morphological appearances to NC may get misclassified. Max-Min acts as a countermeasure that selects the instances with the highest CA response for CA images and the instances with the lowest response for NC images. As shown in Fig. 3(c), Max-Min tends to have an opposite effect compared to Max-Max. Therefore, in cMIL we combine these two criteria to reduce the distribution deviation problem and obtain a more balanced instance-level dataset to be used in fully supervised learning (see Fig. 3(d)). It is worth noting that, for NC images, although each instance is NC, we only use the selected instances to avoid the data imbalance problem.

We choose ResNet-50 [11] as the classifier. The two MIL-based classifiers are trained separately under the same configuration (Fig. 2): in the forward pass, we use the Max-Max (or Max-Min for the other classifier) criterion to select one instance from each bag based on their predictions, and the prediction of the selected instance is regarded as the prediction of the image. In the backprop, we use the cross entropy loss between the image-level label and the prediction of the selected instance to update the classifier's parameters. The loss function for each classifier is defined as follows:

$$Loss = -\sum_{j}(y_j \log \hat{p_j} + (1 - y_j) \log(1 - \hat{p_j})), \quad (1)$$

where $\hat{p_j} = S_{criterion}(\{f(b_i)\})$, b_i is instances in image j, f is the classifier, $S_{criterion} \in \{\text{Max-Max, Max-Min}\}$. $S_{criterion}$ selects the target instance using the defined criterion, y_j is the image-level label.

For Max-Max criterion:

$$S_{Max-Max}(\{f(b_i)\}) = \max_i\{f(b_i)\}. \quad (2)$$

For Max-Min criterion:

$$S_{Max-Min}(\{f(b_i)\}) = \begin{cases} \max_i\{f(b_i)\} & if\ y = 1 \\ \min_i\{f(b_i)\} & if\ y = 0 \end{cases}. \quad (3)$$

After training, we again feed the same training data into the two trained classifiers and select the instances under the corresponding criterion, then the predictions are considered as their labels. We combine the instances selected by the two trained classifiers to construct the final fully supervised instance-level dataset. Noted that we discard those potentially confusing samples whose predicted labels are different from their corresponding image-level labels.

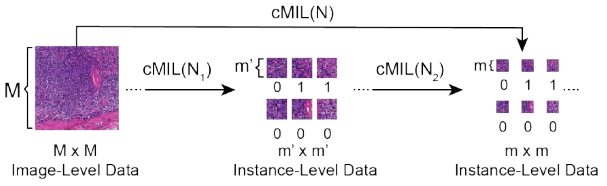

Figure 4. Cascade data enhancement. Beside constructing the $m \times m$ dataset using cMIL(N) directly, we can also first construct an intermediate $m' \times m'$ dataset using cMIL(N_1), then construct the final m×m dataset using cMIL(N_2) in a cascade manner ($N = N_1 \times N_2$).

3.1.2 Retrain and Relabel

Once the instance-level dataset is prepared, we are able to train an instance classifier in a fully supervised manner. The classifier we use in this step has the same architecture as the classifier in cMIL (ResNet-50), we name this step as *retrain*. Then, we split the original image into latticed instances and *relabel* them using the trained instance-level classification model (Fig. 1). For each image, we obtain N^2 high-quality instance labels from a single image-level label.

3.2. Segmentation

With enriched supervision information, the instance-level labels are directly assigned to the corresponding pixels, producing approximate pixel-level labels. Therefore, we can train segmentation models in a fully supervised way using well-developed architectures such as DeepLabv2 [6, 7] and U-Net [20]. To prevent the model from learning the checkboard-like artifacts in the approximate labels, in the training process, we perform data augmentation by feeding smaller images that are randomly cropped from the original training set and their corresponding masks into the segmentation model.

3.3. Further Improvement

The granularity of the enriched labels is determined by the scale factor N; larger scale factor results in finer labels. However, as a tradeoff, larger scale factor would lead to severe image information loss. To tackle this issue, we propose cascade data enhancement to recover the potential loss and add image-level constraints to make better use of the supervision information.

3.3.1 Cascade Data Enhancement

Each instance selection criterion only choose one instance from the image to construct the instance-level dataset, which only takes up a small portion of the image, resulting in losing a considerable amount of image information from

the original image-level dataset. In order to recover this information loss and increase data diversity in the instance-level dataset, we further introduce the cascade data enhancement method to generate the instance-level dataset by two concurrent routes (Fig. 4). Here, we use cMIL(N) to denote the cMIL with a scale factor of N. To derive labeled instances of a scale factor of N, we can either use cMIL(N) or cMIL(N_1) and cMIL(N_2) back-to-back where $N = N_1 \times N_2$. The two sources of data are combined before fed into the segmentation model.

3.3.2 Training with Image-Level Constraints

In order to maximize the utility of the original image-level supervision information, in the retrain step, we can further add the original image-level data as one additional input source going through the classifier. As shown in Fig. 5, the image-level constraint is imposed under Max-Max and Max-Min criteria to the instance level, the total loss is defined as the sum of the retrain loss and the constraint loss:

$$Loss = w_1 \cdot Loss_{constrain} + w_2 \cdot Loss_{retrain}, \quad (4)$$

where w_1 and w_2 are the weights of the two losses. We set $w_1 = w_2$ in our experiments.

$$Loss_{constrain} = -\sum_{S_{criterion}} (y \log \hat{p} + (1-y)\log(1-\hat{p})), \quad (5)$$

where $\hat{p} = S_{criterion}(\{f(b_i)\})$, b_i represents the selected instance, f is the image-level constrain route, $S_{criterion} \in \{\text{Max-Max, Max-Min}\}$, and y is the image-level label.

$$Loss_{retrain} = -\sum_{j}(y_j \log \hat{y_j} + (1-y_j)\log(1-\hat{y_j})), \quad (6)$$

where $\hat{y_j} = g(n_j)$, n_j represents the input instance, g is the retrain route, and y_j is the instance-level label. Since two routes share the same network, we have $f \equiv g$.

4. Experiments

4.1. Data Preparation

We conducted our experiments on CAMELYON16 [1, 5], a public dataset with 400 hematoxylin-eosin (H&E) stained whole-slide images (WSIs) of lymph node sections. In this research, same as CDWS-MIL [13], we regard the 1,280×1,280 patches at 20x magnification in the WSIs as image-level data. The training set of CAMELYON16 contains 240 WSIs (110 contain CA), which we split into 5,011 CA and 96,496 NC 1,280×1,280 patches, and we oversample the CA patches to match the number of NC ones.

Table 1. Instance-level classification performance of label enrichment on CAMELYON16 test set.

320×320 (%)	Sensitivity	Specificity	Accuracy
FSB320	90.0	97.4	94.5
Max-Max	56.9	98.1	81.9
Max-Min	82.0	82.6	82.3
Retrain (cMIL)	88.7	94.6	**92.3**
Retrain (constrained)	84.5	98.4	**92.9**
160×160 (%)	Sensitivity	Specificity	Accuracy
FSB160	89.0	95.0	92.8
Max-Max	44.9	99.3	79.3
Max-Min	87.7	86.5	86.9
Retrain (cMIL)	85.5	90.1	**88.4**
Retrain (constrained)	75.2	98.5	89.9
Cascade	87.7	92.0	90.4
Cascade (constrained)	83.6	96.4	**91.7**

Besides, we have also constructed two other fully supervised training sets containing 320×320 and 160×160 instances to help build the fully supervised baselines. The test set includes 160 WSIs (49 contain CA), and we split and select all the 3,392 1,280×1,280 CA patches, and then we randomly sample NC patches to match the number *. The 1,280×1,280 patches are further split into sizes of 320×320 and 160×160 to test the models with corresponding input sizes. The patches and the instances are labeled as CA if it contains any cancerous region. Otherwise, the label is NC.

4.2. Implementation

We applied rotation, mirroring, and scaling (between 1.0x and 1.2x) at random to augment the training data. All the models were implemented in TensorFlow [2] and trained on 4 NVIDIA GTX1080Ti GPUs. Both instance classifiers in cMIL and the retrain step were trained using Adam optimizer with a fixed learning rate of 0.0001. In cMIL, the batch size was set to 4 (one image-level patch on each GPU). In the retrain step, the batch size was set to 40 (ten instances on each GPU). During the segmentation stage, DeepLabv2 and U-Net were both trained using Adam optimizer with a fixed learning rate of 0.001 and the batch size of 24 (six images on each GPU). Due to the limitation of the GPU resources, we used 640×640 images that are randomly cropped from the original 1280×1280 training set and their corresponding masks to train the segmentation models.

4.3. Performance of Label Enrichment

As Table 1 and Fig. 6 show, in accordance with Fig. 3, models trained on data selected using Max-Max tends to have relatively low sensitivity and high specificity. On the contrary, Max-Min tends to help achieve relatively high sensitivity and low specificity. With the data selected with the

*We exclude Test_114 because of the duplicate labeling [15].

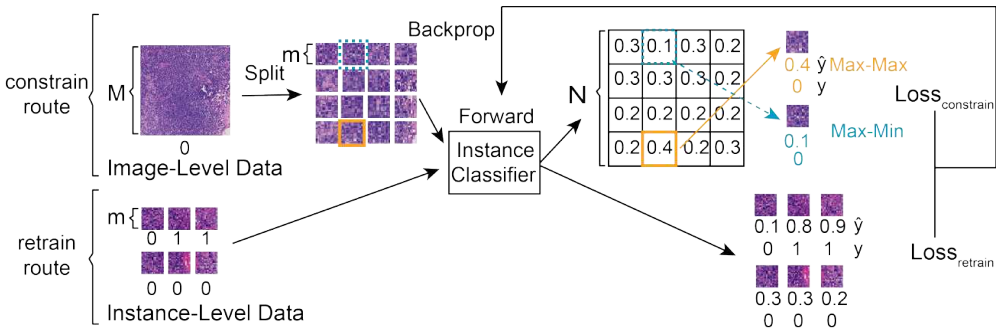

Figure 5. Illustration of model training under image-level constraints. The supervision information from the original image-level data is taken into consideration in the retrain step.

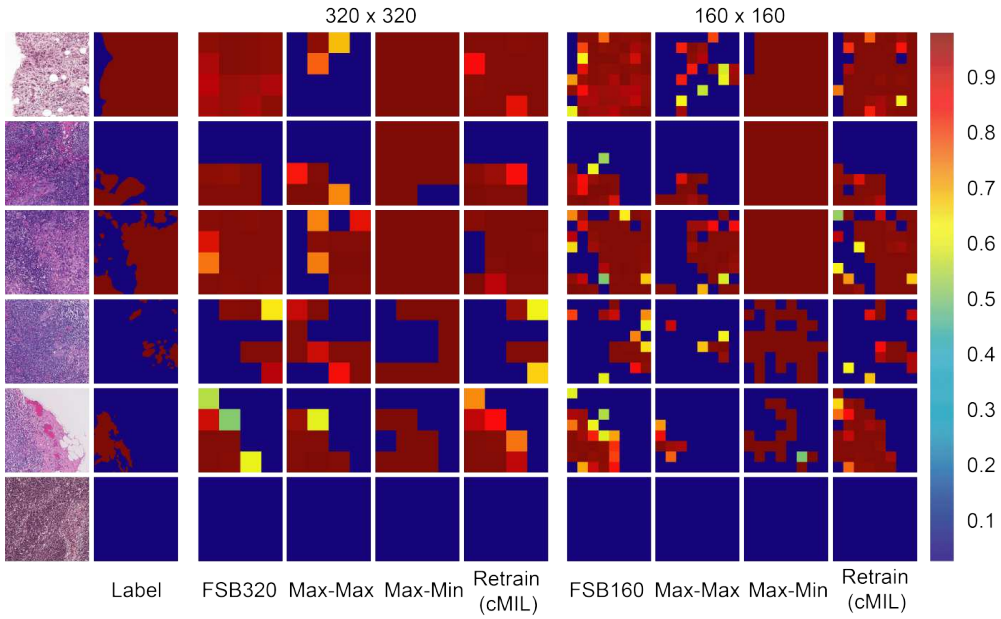

Figure 6. Instance-level classification results on CAMELYON16 test set. Compare to the ground truth, the model trained on the data selected using Max-Max tends to predict less CA, and more CA using Max-Min. Retrain (cMIL) achieves a more reasonable trade-off and better performance.

two criteria combined, the model can achieve a more reasonable trade-off and better performance. By using the cascade data enhancement method and adding the image-level constraints, we further improve the model's accuracy. To compare the performance between our model and the fully supervised baseline (FSB), we use the same classifier architecture (ResNet-50) for both models. On the 320×320 and the 160×160 test sets, the instance classification accuracy are comparable with the fully supervised baselines, which are only 1.6% and 1.1% lower, respectively.

The improvement from cascade data enhancement shows an effective way to recover from image information dilation in constructing the fully supervised instance-level dataset and suggests its potential for label enrichment on an even finer granularity. It also implicates the robustness of cMIL with different scale factors. The improvement from adding the image-level constraints shows the benefit of combining supervision information of image-level and instance-level.

We further verify the instance-level classification performance of our best models on the 320×320 and 160×160 training sets (Table 2), where they achieve 95.5% and 94.6% accuracies, respectively. After label enrichment, CAMEL successfully enriches the supervision information from single image-level label to N^2 instance-level granularity for the images in the original image-level dataset with high quality.

Table 2. Quality of automatically enriched instance-level labels for the original image-level dataset measured by the classification performance on CAMELYON16 training sets.

	N^2	Sensitivity	Specificity	Accuracy
160×160	64	89.9	94.7	94.6
320×320	16	91.4	95.7	95.5

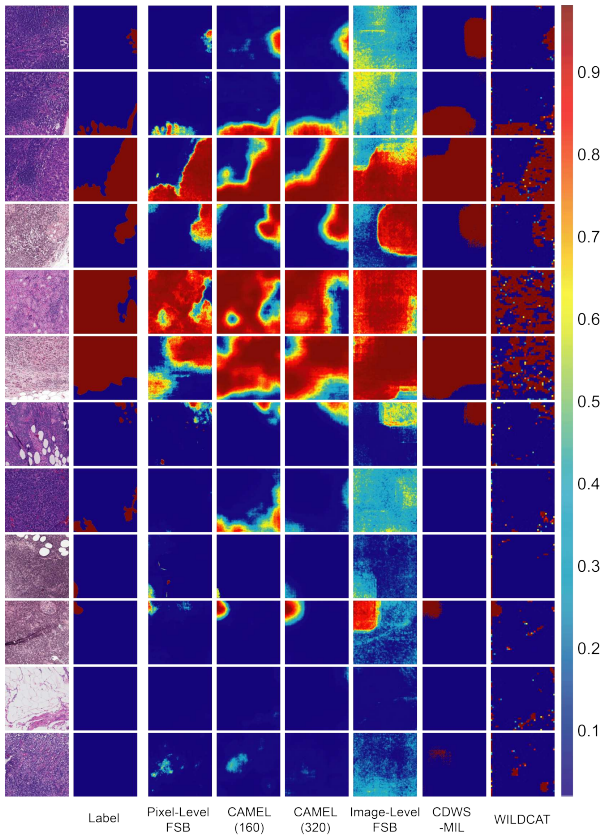

Figure 7. Pixel-level segmentation results (DeepLabv2) of CAMEL and other methods on CAMELYON16 test set.

4.4. Performance of Segmentation

After label enrichment, the instance-level labels of the training set are assigned to the corresponding pixels to produce approximate pixel-level labels. At this point, we can train the segmentation model in a fully supervised manner. We test the performance of DeepLabv2 with ResNet-34 [7] and U-Net [20].

As given in Table 3, we use sensitivity, specificity, accuracy, and intersection over union (IoU) to measure the pixel-level segmentation performance. For comparison, the performance of the fully supervised baseline pixel-level FSB and the performance of the weakly supervised methods WILDCAT [9], DWS-MIL, and CDWS-MIL [13] are also listed. WILDCAT is used for natural images in their paper [9], and DWS-MIL and CDWS-MIL [13] are used for histopathology image. Here, we add another baseline model (image-level FSB) to show the importance of label enrichment for segmentation performance. The image-level FSB is trained with the data whose label is generated by directly assigning the image-level labels to the pixels, while the pixel-level FSB is obtained using the original pixel-level ground truth. CAMEL outperforms the image-level FSB, WILDCAT, DWS-MIL, and CDWS-MIL, and is even comparable with the pixel-level FSB.

With the help of the efficient use of supervision information, finer granularity brings with better segmentation performance. Moreover, in the label enrichment step, the instance pixels are labeled as CA if it contains any cancerous region. This may lead to the over-labeling issue. As shown in Fig. 7, smaller instance size alleviates this issue by constructing finer pixel-level labels, demonstrating the effectiveness of finer labels and the potential of improvement for label enrichment on an even finer granularity.

We further evaluate our models on the WSIs of CAMELYON16 test set. Fig. 8 shows some examples.

4.5. Generality of CAMEL

To evaluate the generality of CAMEL, we test CAMEL on a colorectal adenoma dataset which contains 177 WSIs (156 contain adenoma) gathered and labeled by pathologists from the Department of Pathology, The Chinese PLA General Hospital. As Table 4 and Fig. 9 show, CAMEL consistently achieves comparable performance against the fully supervised baselines.

5. Conclusion

Computer-assisted diagnosis for histopathology image can improve the accuracy and relieve the burden for pathologists at the same time. In this research, we present a weakly supervised learning framework, CAMEL, for histopathology image segmentation using only image-level labels. CAMEL automatically enriches supervision information from image-level to instance-level with high quality and achieves comparable segmentation results with its fully supervised counterparts. More importantly, the automatic labeling methodology may generalize to other weakly supervised learning studies for histopathology image analysis.

In CAMEL, the obtained instance-level labels are directly assigned to the corresponding pixels and used as masks in the segmentation task, which may result in the over-labeling issue. We will tackle this challenge using mask boundary refinement [3, 4] in future work.

Acknowledgement. The authors would like to thank Xiang Gao, Lang Wang, Cunguang Wang, Lichao Pan, Fangjun Ding at Thorough Images for data processing and helpful discussions. This research is supported by National Natural Science Foundation of China (NSFC) (No.

Table 3. Pixel-level segmentation performance on CAMELYON16 test set.

DeepLabv2 (%)	Sensitivity	Specificity	Accuracy	F1-Score	IoU
Pixel-Level FSB	87.9	99.1	95.3	92.6	86.3
Image-Level FSB	89.2	88.7	88.9	84.4	72.9
CAMEL (160)	92.7	95.7	94.7	92.1	**85.4**
CAMEL (320)	94.7	93.8	94.1	91.5	**84.3**
U-Net (%)	Sensitivity	Specificity	Accuracy	F1-Score	IoU
Pixel-Level FSB	87.8	98.2	94.7	91.8	84.8
Image-Level FSB	95.5	82.1	86.6	82.8	70.6
CAMEL (160)	94.7	94.1	94.3	91.8	**84.8**
CAMEL (320)	94.7	94.0	94.2	91.7	**84.7**
Other Methods (%)	Sensitivity	Specificity	Accuracy	F1-Score	IoU
WILDCAT (w/ ResNet-50)	69.6	93.8	85.7	76.6	62.0
DWS-MIL (w/ ResNet-50)	86.0	93.4	90.9	86.4	76.0
CDWS-MIL (w/ ResNet-50)	87.2	93.8	91.5	87.4	77.6

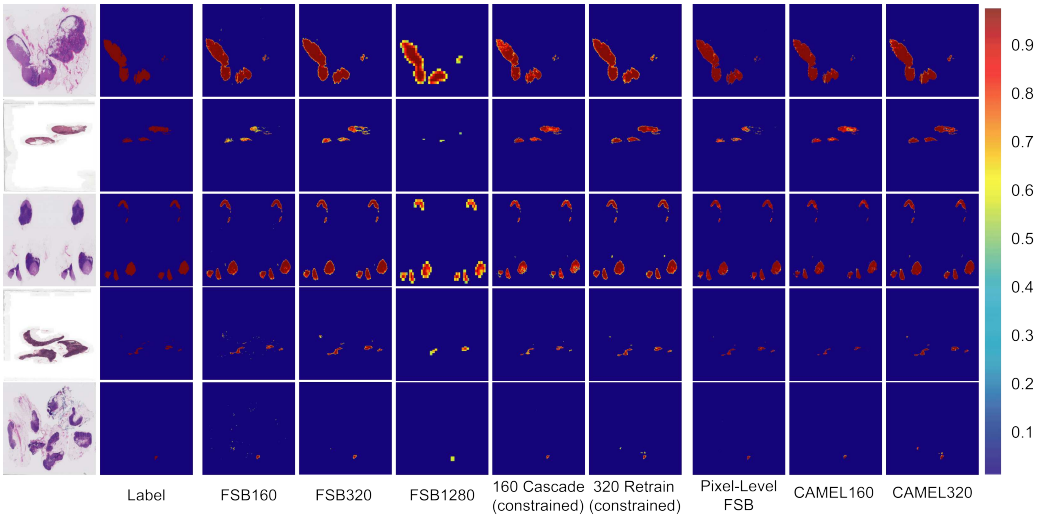

Figure 8. Some examples of instance-level classification and pixel-level segmentation (DeepLabv2) results on CAMELYON16 WSIs.

Table 4. Model performance on colorectal adenoma dataset.

Instance-level classification (%)	Recall	Precision	Accuracy
FSB320	81.1	90.0	87.1
Retrain (cMIL)	84.9	81.0	83.8
FSB160	80.7	87.6	87.0
Retrain (cMIL)	80.9	85.1	86.0
Pixel-level segmentation (%)	Recall	Precision	F1-Score
Pixel-Level FSB	86.1	89.0	87.5
CAMEL (160)	89.7	85.0	87.3
CAMEL (320)	95.4	78.5	86.1

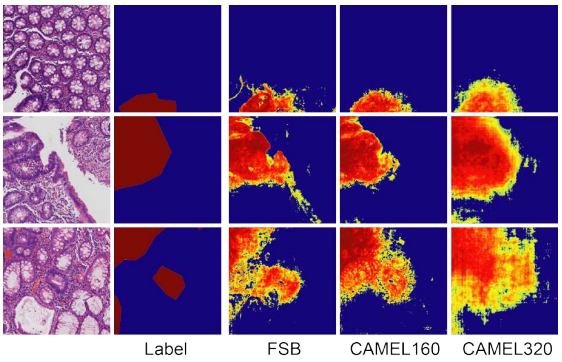

Figure 9. Pixel-level segmentation results (DeepLabv2) of CAMEL on colorectal adenoma dataset.

61532001), Tsinghua Initiative Research Program (No. 20151080475), Shanghai Municipal Science and Technology Major Project (No. 2018SHZDZX01) and ZJLab.

References

[1] CAMELYON 2016. https://camelyon16.grand-challenge.org, 2016.

[2] Martín Abadi, Paul Barham, Jianmin Chen, Zhifeng Chen, Andy Davis, Jeffrey Dean, Matthieu Devin, Sanjay Ghemawat, Geoffrey Irving, Michael Isard, et al. Tensorflow: A system for large-scale machine learning. In *OSDI*, volume 16, pages 265–283, 2016.

[3] Jiwoon Ahn, Sunghyun Cho, and Suha Kwak. Weakly supervised learning of instance segmentation with inter-pixel relations. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2209–2218, 2019.

[4] Jiwoon Ahn and Suha Kwak. Learning pixel-level semantic affinity with image-level supervision for weakly supervised semantic segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4981–4990, 2018.

[5] Babak Ehteshami Bejnordi, Mitko Veta, Paul Johannes Van Diest, Bram Van Ginneken, Nico Karssemeijer, Geert Litjens, Jeroen AWM Van Der Laak, and the CAMELYON16 Consortium. Diagnostic assessment of deep learning algorithms for detection of lymph node metastases in women with breast cancer. *JAMA*, 318(22):2199, 2017.

[6] Liang-Chieh Chen, George Papandreou, Iasonas Kokkinos, Kevin Murphy, and Alan L Yuille. Semantic image segmentation with deep convolutional nets and fully connected CRFs. *Computer Science*, (4):357–361, 2014.

[7] Liang-Chieh Chen, George Papandreou, Iasonas Kokkinos, Kevin Murphy, and Alan L Yuille. DeepLab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected CRFs. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 40(4):834–848, 2018.

[8] Jifeng Dai, Kaiming He, and Jian Sun. BoxSup: Exploiting bounding boxes to supervise convolutional networks for semantic segmentation. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1635–1643, 2015.

[9] Thibaut Durand, Taylor Mordan, Nicolas Thome, and Matthieu Cord. WILDCAT: Weakly supervised learning of deep convnets for image classification, pointwise localization and segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 642–651, 2017.

[10] Weifeng Ge, Sibei Yang, and Yizhou Yu. Multi-evidence filtering and fusion for multi-label classification, object detection and semantic segmentation based on weakly supervised learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1277–1286, 2018.

[11] 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.

[12] Zilong Huang, Xinggang Wang, Jiasi Wang, Wenyu Liu, and Jingdong Wang. Weakly-supervised semantic segmentation network with deep seeded region growing. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 7014–7023, 2018.

[13] Zhipeng Jia, Xingyi Huang, Eric I-Chang Chao, and Yan Xu. Constrained deep weak supervision for histopathology image segmentation. *IEEE Transactions on Medical Imaging*, 36(11):2376–2388, 2017.

[14] Anna Khoreva, Rodrigo Benenson, Jan Hendrik Hosang, Matthias Hein, and Bernt Schiele. Simple does it: Weakly supervised instance and semantic segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 876–885, 2017.

[15] Yi Li and Wei Ping. Cancer metastasis detection with neural conditional random field. *arXiv preprint arXiv:1806.07064*, 2018.

[16] Di Lin, Jifeng Dai, Jiaya Jia, Kaiming He, and Jian Sun. ScribbleSup: Scribble-supervised convolutional networks for semantic segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3159–3167, 2016.

[17] Huangjing Lin, Hao Chen, Simon Graham, Qi Dou, Nasir Rajpoot, and Pheng-Ann Heng. Fast Scannet: Fast and dense analysis of multi-gigapixel whole-slide images for cancer metastasis detection. *IEEE Transactions on Medical Imaging*, 38(8):1948–1958, 2019.

[18] Yun Liu, Krishna Gadepalli, Mohammad Norouzi, George E Dahl, Timo Kohlberger, Aleksey Boyko, Subhashini Venugopalan, Aleksei Timofeev, Philip Q Nelson, Greg S Corrado, et al. Detecting cancer metastases on gigapixel pathology images. *arXiv preprint arXiv:1703.02442*, 2017.

[19] Anant Madabhushi and George Lee. Image analysis and machine learning in digital pathology: Challenges and opportunities. *Medical Image Analysis*, 33:170–175, 2016.

[20] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-Net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 234–241. Springer, 2015.

[21] Paul Viola, John C. Platt, and Cha Zhang. Multiple instance boosting for object detection. In *International Conference on Neural Information Processing Systems*, pages 1417–1424, 2005.

[22] Yunchao Wei, Jiashi Feng, Xiaodan Liang, Ming-Ming Cheng, Yao Zhao, and Shuicheng Yan. Object region mining with adversarial erasing: A simple classification to semantic segmentation approach. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1568–1576, 2017.

[23] Yunchao Wei, Xiaodan Liang, Yunpeng Chen, Xiaohui Shen, Ming-Ming Cheng, Jiashi Feng, Yao Zhao, and Shuicheng Yan. STC: A simple to complex framework for weakly-supervised semantic segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 39(11):2314–2320, 2017.

[24] Tianjun Xiao, Yichong Xu, Kuiyuan Yang, Jiaxing Zhang, Yuxin Peng, and Zheng Zhang. The application of two-level attention models in deep convolutional neural network for fine-grained image classification. In *Proceedings of the*

IEEE Conference on Computer Vision and Pattern Recognition, pages 842–850, 2015.

[25] Yan Xu, Tao Mo, Qiwei Feng, Peilin Zhong, Maode Lai, and Eric I-Chang Chao. Deep learning of feature representation with multiple instance learning for medical image analysis. In *IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 1626–1630, 2014.

[26] Yan Xu, Jun-Yan Zhu, Eric I-Chang Chao, Maode Lai, and Zhuowen Tu. Weakly supervised histopathology cancer image segmentation and classification. *Medical Image Analysis*, 18(3):591–604, 2014.