Self- and Semi-supervised Learning for Gastroscopic Lesion Detection

Xuanye Zhang[1], Kaige Yin[2], Siqi Liu[1], Zhijie Feng[2], Xiaoguang Han[3,4], Guanbin Li[5(✉)], and Xiang Wan[1(✉)]

[1] Shenzhen Research Institute of Big Data, CUHK-Shenzhen, Shenzhen, China
wanxiang@sribd.cn
[2] Department of Gastroenterology, Hebei Key Laboratory of Gastroenterology, Hebei Institute of Gastroenterology, Hebei Clinical Research Center for Digestive Diseases, The Second Hospital of Hebei Medical University, Shijiazhuang, Hebei, China
[3] FNii, CUHK-Shenzhen, Shenzhen, China
[4] School of Science and Engineering, CUHK-Shenzhen, Shenzhen, China
[5] School of Computer Science and Engineering, Research Institute of Sun Yat-sen University in Shenzhen, Sun Yat-sen University, Guangzhou, China
liguanbin@mail.sysu.edu.cn

Abstract. Gastroscopic Lesion Detection (GLD) plays a key role in computer-assisted diagnostic procedures. However, this task is not well studied in the literature due to the lack of labeled data and the applicable methods. Generic detectors perform below expectations on GLD tasks for 2 reasons: 1) The scale of labeled data of GLD datasets is far smaller than that of natural-image object detection datasets. 2) Gastroscopic images exhibit distinct differences from natural images, which are usually of high similarity in global but high diversity in local. Such characteristic of gastroscopic images also degrades the performance of using generic self-supervised or semi-supervised methods to solve the labeled data shortage problem using massive unlabeled data. In this paper, we propose Self- and Semi-Supervised Learning (SSL) for GLD tailored for using massive unlabeled gastroscopic images to enhance GLD tasks performance, which consists of a Hybrid Self-Supervised Learning (HSL) method for backbone pre-training and a Prototype-based Pseudo-label Generation (PPG) method for semi-supervised detector training. The HSL combines patch reconstruction with dense contrastive learning to boost their advantages in feature learning from massive unlabeled data. The PGG generates pseudo-labels for unlabeled data based on similarity to the prototype feature vector to discover potential lesions and avoid introducing much noise. Moreover, we contribute the first Large-scale GLD Datasets (LGLDD), which contains 10,083 gastroscopic images with 12,292 well-annotated boxes for four-category lesions. Experiments on LGLDD demonstrate that SSL can bring significant improvement.

X. Zhang and K. Yin—Contribute equally to this paper.

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/978-3-031-43904-9_9.

H. Greenspan et al. (Eds.): MICCAI 2023, LNCS 14224, pp. 83–93, 2023.
https://doi.org/10.1007/978-3-031-43904-9_9

Keywords: Gastroscopic Lesion Detection · Self-Supervised Backbone Pre-training · Semi-supervised Detector Training

1 Introduction

Gastroscopic Lesion Detection (GLD) plays a key role in computer-assisted diagnostic procedures. Although deep neural network-based object detectors achieve tremendous success within the domain of natural images, directly training generic object detectors on GLD datasets performs below expectations for two reasons: 1) The scale of labeled data in GLD datasets is limited in comparison to natural images due to the annotation costs. Though gastroscopic images are abundant, those containing lesions are rare, which necessitates extensive image review for lesion annotation. 2) The characteristic of gastroscopic images exhibits distinct differences from the natural images [18,19,21] and is often of high similarity in global but high diversity in local. Specifically, each type of lesion may have diverse appearances though gastroscopic images look quite similar. Some appearances of lesions are quite rare and can only be observed in a few patients. Generic self-supervised backbone pre-training or semi-supervised detector training methods can solve the first challenge for natural images but its effectiveness is undermined for gastroscopic images due to the second challenge.

Self-Supervised Backbone Pre-training methods enhance object detection performance by learning high-quality feature representations from massive unlabelled data for the backbone. The mainstream self-supervised backbone pre-training methods adopt self-supervised contrast learning [3,4,7,9,10] or masked

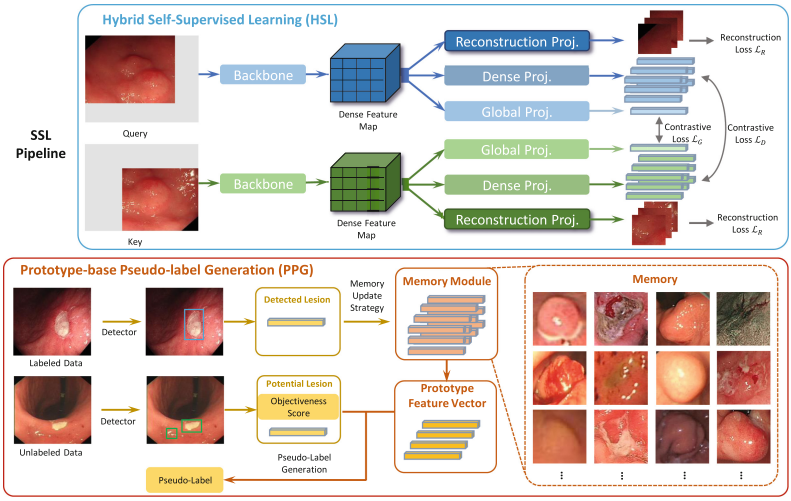

Fig. 1. Pipeline of Self- and Semi-Supervised Learning (SSL) for GLD. SSL consists of a Hybrid Self-Supervised Learning (HSL) method and a Prototype-based Pseudo-label Generation (PPG) method. HSL combines patch reconstruction with dense contrastive learning. PPG generates pseudo-labels for potential lesions based on the similarity to the prototype feature vectors.

image modeling [8,15]. Self-supervised contrastive learning methods [3,4,7,9] can learn discriminative global feature representations, and [10] can further learn discriminative local feature representations by extending contrastive learning to dense paradigm. However, these methods usually cannot grasp enough local detailed information. On the other hand, masked image modeling is expert in extracting local detailed information but is weak in preserving the discriminability of feature representation. Therefore, both types of methods have their own weakness for GLD tasks.

Semi-Supervised object detection methods [12,14,16,17,20,22,23] first use detectors trained with labeled data to generate pseudo-labels for unlabeled data and then enhance object detection performance by regarding these unlabeled data with pseudo-labels as labeled data to train the detector. Current pseudo-label generation methods rely on the objectiveness score threshold to generate pseudo-labels, which makes them perform below expectations on GLD, because the characteristic of gastroscopic lesions makes it difficult to set a suitable threshold to discover potential lesions meanwhile avoiding introducing much noise.

The motivation of this paper is to explore how to enhance GLD performance using massive unlabeled gastroscopic images to overcome the labeled data shortage problem. The main challenge for this goal is the characteristic of gastroscopic lesions. Intuitively, such a challenge requires local feature representations to contain enough detailed information, meanwhile preserving discriminability. Enlightened by this, we propose the **S**elf- and **S**emi-Supervised **L**earning (**SSL**) framework tailored to address challenges in daily clinical practice and use massive unlabeled data to enhance GLD performance. SSL overcomes the challenges of GLD by leveraging a large volume of unlabeled gastroscopic images using self-supervised learning for improved feature representations and semi-supervised learning to discover and utilize potential lesions to enhance performance. Specifically, it consists of a **H**ybrid **S**elf-Supervised **L**earning (**HSL**) method for self-supervised backbone pre-training and a **P**rototype-based **P**seudo-label **G**eneration (**PPG**) method for semi-supervised detector training. The HSL combines the dense contrastive learning [10] with the patch reconstruction to inherit the advantages of discriminative feature learning and grasp the detailed information that is important for GLD tasks. The PPG generates pseudo-labels based on the similarity to the prototype feature vectors (formulated from the feature vectors in its Memory Module) to discover potential lesions from unlabeled data, and avoid introducing much noise at the same time. Moreover, we propose the first Large-scale GLD Datasets (LGLDD), which contains 10,083 gastroscopic images with 12,292 well-annotated lesion bounding boxes of four categories of lesions (polyp, ulcer, cancer, and sub-mucosal tumor). We evaluate SSL with multiple detectors on LGLDD and SSL brings significant improvement compared with baseline methods (CenterNet [6]: +2.7AP, Faster RCNN [13]: +2.0AP). In summary, our contributions include:

- A Self- and Semi-supervise Learning (SSL) framework to leverage massive unlabeled data to enhance GLD performance.
- A Large-scale Gastroscopic Lesion Detection datasets (LGLDD)

– Experiments on LGLDD demonstrate that SSL can bring significant enhancement compared with baseline methods.

2 Methodology

In this section, we introduce the main ideas of the proposed **SSL** for GLD. The proposed approach includes 2 main components and is illustrated in Fig. 1.

2.1 Hybrid Self-supervised Learning

The motivation of Hybrid Self-Supervised Learning (HSL) is to learn the local feature representations of high discriminability meanwhile contain detailed information for the backbone from massive unlabeled gastroscopic images. Among existing backbone pre-training methods, dense contrastive learning can preserve local discriminability and masked image modeling can grasp local detailed information. Therefore, to leverage the advantages of both types of methods, we propose Hybrid Self-Supervised Learning (HSL), which combines patch reconstruction with dense contrastive learning to achieve the goal.

Structure. HSL heritages the structure of the DenseCL [10] but adds an extra reconstruction projection head to reconstruct patches. Specifically, HSL consists of a backbone network and 3 parallel sub-heads. The global projection head and the dense projection head heritages from the DenseCL [10], and the proposed reconstruction projection head is inspired by the Masked Image Modeling. Enlightened by the SimMIM [15], we adopt a lightweight design for the reconstruction projection head, which only contains 2 convolution layers.

Learning Pipeline. Like other self-supervised contrastive learning methods, HSL randomly generates 2 different "views" of the input image, uses the backbone to extract the dense feature maps $\mathbf{F}_1, \mathbf{F}_2 \in \mathbb{R}^{H \times W \times C}$, and then feeds them to the following projection heads. The global projection head of HSL uses $\mathbf{F}_1$, $\mathbf{F}_2$ to obtain the global feature vector $\mathbf{f}_{g1}$, $\mathbf{f}_{g2}$ like MoCo [9]. The dense projection head and the reconstruction projection head crop the dense feature maps $\mathbf{F}_1$, $\mathbf{F}_2$ into $\mathbf{S} \times \mathbf{S}$ patches and obtain the local feature vector sets $\mathbb{F}_1$ and $\mathbb{F}_2$ of each view ($\mathbb{F} = \{f_1, f_2, ..., f_{\mathbf{S}^2}\}$). The dense projection head use $\mathbb{F}_1$ and $\mathbb{F}_2$ to obtain local feature vector sets $\mathbb{F}_{l1}$ and $\mathbb{F}_{l2}$ ($\mathbb{F}_l = \{f_{l1}, f_{l2}, ..., f_{l\mathbf{S}^2}\}$) like DenseCL [10]. The reconstruction projection head uses each feature vector in $\mathbb{F}_1, \mathbb{F}_2$ to reconstruct corresponding patches and obtains the patch set $\mathbb{P}_1, \mathbb{P}_2 (\mathbb{P} = \{p_{i1}, p_{i2}, ..., p_{i\mathbf{S}^2}\}$.

Training Objective. The HSL formulates the two contrastive learning as dictionary look-up tasks like DenseCL [10] while the reconstruction learning as a regression task. The global contrastive learning uses the global feature vector $\mathbf{f}_g$ of an image as query $\mathbf{q}$ and feature vectors from the alternate view of the query image and the other images within the batch as keys $\mathbb{K} = \{k_1, k_2, ..., \}$. For each

query $\mathbf{q}$, the only positive key $\mathbf{k}_+$ is the different views of the same images and the others are all negative keys ($\mathbf{k}_-$) like MoCo [9]. We adopt the InfoNCE loss function for it:

$$\mathcal{L}_G = -log\frac{exp(q \cdot k_+/\tau)}{exp(q \cdot k_+) + \sum_{k_-} exp(q \cdot k_-/\tau)}$$

The dense contrastive learning uses the local feature vector in $\mathbb{F}_{li}$ as query r and keys $\mathbb{T}_l = \{t_1, t_2, ..., \}$. The negative keys t_- here are the feature vectors of different images while the positive key t_+ is the correspondence feature vector of r in another view of the images. Specifically, we adopt the correspondence methods in DenseCL [10] to obtain the positive key t_+, which first conducts the matching process based on vector-wise cosine similarity between r and feature vectors in $\mathbb{T}$ and then selects the t_j of highest similarity as the t_+. The loss function is also the InfoNCE loss but in a dense paradigm:

$$\mathcal{L}_L = \frac{1}{S^2}\sum -log\frac{exp(r^s \cdot t^s_+/\tau)}{exp(r^s \cdot t^s_+) + \sum_{t^s_-} exp(r^s \cdot t^s_-/\tau)}$$

The reconstruction task uses the feature vector in $\mathbb{F}$ to reconstruct each patch and obtain $\mathbb{P}_i$. The ground truth is the corresponding patches $\mathbb{V}_i = \{v_{i1}, v_{i2}, ..., v_{i\mathbf{S}^2}\}$ of the input view. We adopt the MSE loss function for it:

$$\mathcal{L}_R = \frac{1}{2S^2}\sum E(v_i - p_i)^2$$

The overall loss function is the weighted sum of these losses:

$$\mathcal{L}_H = \mathcal{L}_G + \lambda_D\mathcal{L}_D + \lambda_R\mathcal{L}_R$$

where λ_D and λ_R are the weights of $\mathcal{L}_D$ and $\mathcal{L}_R$ and are set to 1 and 2.

2.2 Prototype-Based Pseudo-label Generation Method

We propose the Prototype-based Pseudo-label Generation method (PPG) to discover potential lesions from unlabeled gastroscopic data meanwhile avoid introducing much noise to further enhance GLD performance. Specifically, PPG adopts a Memory Module to remember feature vectors of the representative lesions as memory and generates prototype feature vectors for each class based on the memories stored. To preserve the representativeness of the memory and further the prototype feature vectors, PPG designs a novel Memory Update Strategy. In semi-supervised learning, PPG generates pseudo-labels for unlabeled data relying on the similarity to the prototype feature vectors, which achieves a better balance between lesion discovery and noise avoidance.

Memory Module. Memory Module stores a set of lesion feature vectors as memory. For a C-class GLD task, the Memory Module stores $C \times N$ feature

vectors as memory. Specifically, for each lesion, we denote the feature vector used to classify the lesion in the detector as f_c. PPG stores N feature vectors for each class c to formulate the class memory $m_c = \{f_{c1}, f_{c2}, ..., f_{cN}\}$, and the memory $\mathbb{M}$ of PPG can be expressed as $\mathbb{M} = \{m_1, m_2, ..., m_C\}$. Then, PPG obtains the prototype feature vector p_c by calculating the center of each class memory m_c, and the prototype feature vector set can be expressed as $\mathbb{P}_t = \{p_1, p_2, ..., p_C\}$. Moreover, the prototype feature vectors further serve as supervision for detector training under a contrastive clustering formulation and adopt a contrastive loss:

$$\mathcal{L}_{cc} = \|f_c, p_c\| + \sum_{j \neq c}^{C} max(0, 1 - \|f_c, p_j\|)$$

If the detector training loss is $\mathcal{L}_{Det}$, the overall loss $\mathcal{L}$ can be expressed as:

$$\mathcal{L} = \mathcal{L}_{Det} + \lambda_{cc}\mathcal{L}_{cc}$$

where the λ_{cc} is the weight of the contrastive learning loss and is set to 0.5.

Memory Update Strategy. Memory Update Strategy directly influences the representativeness of the class memory m_c and further the prototype feature vector p_c. Therefore, PPG adopts a novel Memory Update Strategy, which follows the idea that "The Memory Module should preserve the more representative feature vector among similar feature vectors". The pipeline of the strategy is as follows: 1) Acquisition the lesion feature vector f'_c. 2) Identification of the most similar f_s to f_c from corresponding class memory m_c based on similarity:

$$f'_s = \underset{j}{max}\, sim(f_{cj}, f'_c)$$

3) Updating the memory by selecting more unique features f_s of $F' = \{f'_s, f'_c\}$ compared to the class prototype feature vector p_c based upon similarity:

$$f_s = \underset{f' \in F'}{argmin}\, sim(f', p_c)$$

The similarity function $sim(u, v)$ can be expressed as $sim(u, v) = u^T v / \|u\| \|v\|$. To initialize the memories, we empirically select 50 lesions randomly for each class. To maintain stability, we start updating the memory and calculating its loss after fixed epochs, and only the positive sample feature vector can be selected to update the memory.

Pseudo-label Generation. PPG proposes to generate pseudo-labels based on the similarity between the prototype feature vectors and the feature vector of potential lesions. To be specific, PPG first detects a large number of potential lesions with a low objectiveness score threshold τ_u and then matches them with all the prototype feature vectors $\mathbb{P}$ to find the most similar one:

$$c = \underset{p_c \in \mathbb{P}}{argmax}\, sim(p_c, f_u)$$

PPG assigns the pseudo-label c for similarity value $sim(p_c, f_u)$ greater than the similarity threshold τ_s otherwise omits it. We set $\tau_u = 0.5$ and $\tau_s = 0.5$

3 Datasets

We contribute the first **L**arge-scale **G**stroscopic **L**esion **D**etection **D**atasets (LGLDD) in the literature.

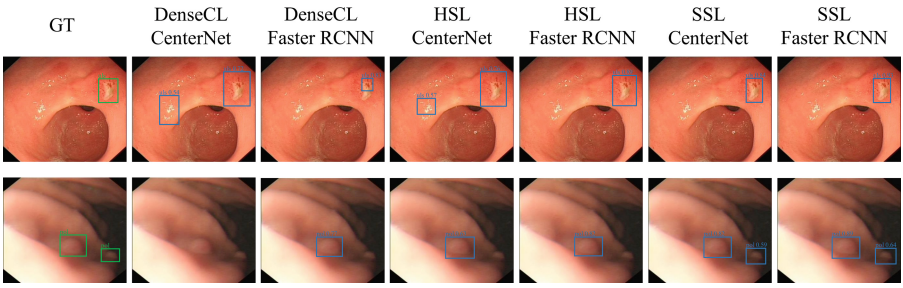

Fig. 2. Qualitative Results of SSL on LGLDD. SSL can actually enhance the GLD performance for some challenging cases.

Collection : LGMDD collects about 1M+ gastroscopic images from 2 hospitals of about 500 patients and their diagnosis reports. After consulting some senior doctors and surveying gastroscopic diagnosis papers [1], we select to annotate 4-category lesions: polyp(pol), ulcer(ulc), cancer(can) and sub-mucosal tumor(smt). We invite 10 senior doctors to annotate them from the unlabeled endoscopic images. To preserve the annotation quality, doctors can refer to the diagnosis reports, and each lesion is annotated by a doctor and checked by another. Finally, they annotates 12,292 lesion boxes in 10,083 images after going through about 120,000 images. The polyp, ulcer, cancer, and sub-mucosal tumor numbers are 7,779, 2,171, 1,164and 1,178, respectively. The train/val split of LGMDD is 8,076/2,007. The other data serves as unlabeled data.

Evaluation Metrics : We use standard object detection metrics to evaluate the GLD performance, which computes the average precision (AP) under multiple intersection-of-union (IoU) thresholds and then evaluate the performance using the mean of APs (mAP) and the AP of some specific IoU threshold. For mAP, we follow the popular object detection datasets COCO [11] and calculate the mean of 11 APs of IoU from 0.5 to 0.95 with stepsize 0.05 (mAP @[.5:.05:.95]). We also report AP under some specific IoU threshold (AP_{50} for .5, AP_{75} for .75) and AP of different scale lesions (AP_S, AP_M, AP_L) like COCO [11].

4 Experiments

Please kindly refer to the **Supplemental Materials** for implementation details and training setups.

Table 1. Quantitative Results of SSL on LGLDD. Both components of SSL (HSL & PPG) can bring significant performance enhancement for GLD tasks.

Detector	Pre-training	PPG	AP	AP_{50}	AP_{75}	AP_S	AP_M	AP_L	pol	stm	uls	can
CenterNet	Supervised	x	29.3	57.2	25.4	22.0	31.6	31.3	41.6	36.0	27.3	12.1
Faster RCNN	Supervised	x	34.1	70.6	28.1	28.2	29.2	35.6	44.0	44.4	24.0	24.2
CenterNet	DenseCL	x	31.9	60.7	29.5	22.6	32.1	34.7	43.2	43.1	28.2	13.2
Faster RCNN	DenseCL	x	35.3	71.9	29.4	29.9	31.8	37.1	44.9	46.7	25.4	24.3
CenterNet	HSL	x	33.7	64.2	30.6	23.1	33.7	35.9	42.5	45.3	28.8	18.0
Faster RCNN	HSL	x	36.4	74.0	31.4	27.9	31.8	38.3	43.7	48.0	26.1	**27.6**
CenterNet	HSL	✓	34.6	65.6	31.6	21.7	32.9	37.3	43.1	46.3	29.3	19.6
Faster RCNN	HSL	✓	**37.3**	**74.8**	**33.2**	**28.8**	**33.5**	**39.4**	**44.9**	**51.0**	**26.1**	27.3

Table 2. Parameters Analysis Experiment Results. (a) Reconstruction loss weight λ_R. (b) Objectiveness Score Threshold τ_u. (c) Memory update strategies. (d) Extension experiment on Endo21.

(a)

λ_R	AP	AP_{50}	AP_{75}
0.5	35.8	73.1	30.7
1	**36.4**	**74.0**	**31.4**
2	36.3	73.4	31.8
5	35.5	71.6	29.5

(b)

τ_u	AP	AP_{50}	AP_{75}
w/o	36.4	74.0	31.4
0.7	36.7	74.0	32.1
0.6	36.2	73.7	31.8
0.5	35.8	72.4	30.9
PPG	**37.3**	**74.8**	**33.2**

(c)

	AP	AP_{50}	AP_{75}
Q-like	37.0	74.2	31.4
PPG	**37.3**	**74.8**	**33.2**

(d)

	AP	AP_{50}	AP_{75}
YOLO v5	60.5	81.0	66.4
Faster RCNN	57.8	79.1	68.1
+DenseCL	59.0	80.9	66.0
+HSL	61.4	83.0	67.3
+PPG	**61.9**	**83.0**	**69.2**

Main Results. Table 1 shows the quantitative results of SSL on LGLDD. As is illustrated, when compared with the DenseCL [10] baseline, SSL can enhance 2.0AP and 2.7AP for Faster RCNN and CenterNet respectively. When compared with the supervised pre-training (ImageNet [5] weights) baseline, SSL can boost more AP enhancement (CenterNet: +5.3AP, FasterRCNN: +3.2AP). Qualitative Results are shown in Fig. 2. It can be noticed, SSL can actually enhance the GLD performance for both types of detectors, especially for some challenging cases.

Ablation Studies. We further analyze each component of SSL (HSL & PPG). HSL can bring 1.8 AP and 1.1 AP enhancement for CenterNet and FasterRCNN respectively compared with DenseCL. PPG can bring extra 0.9AP and 0.9AP enhancement for CenterNet and FasterRCNN respectively.

Parameter Analysis. We conduct extra experiments based on Faster RCNN to further analyze the effect of different parameter settings on LGLDD.
1) **Reconstruction Loss Weight** λ_R is designed to balance the losses of contrastive learning and the reconstruction, which is to balance the discriminability and the detailed information volume of local feature representations. As illustrated in Table 2.a, only suitable λ_R can fully boost the detection performance.
2) **Objectiveness score threshold** τ_u: We compare **PPG** with objectiveness score-based pseudo-label generation methods with different τ_u (Table 2.b). The Objectiveness score threshold controls the quality of pseudo-labels. a) A

low threshold generates noisy pseudo-labels, leading to reduced performance (-0.6/-0.2 AP at thresholds 0.5/0.6). b) A high threshold produces high-quality pseudo-labels but may miss potential lesions, resulting in only slight performance improvement (+0.3 AP at threshold 0.7). c) PPG approach uses a low threshold (0.5) to identify potential lesions, which are then filtered using prototype feature vectors, resulting in the most significant performance enhancement (+0.9 AP).
3) **Memory Update Strategy** influences the representativeness of memory and the prototype feature vectors. We compare our Memory Update Strategy with a queue-like ('Q-like') memory update strategy (first in & first out). Experiment results (Table 2.c) show our Memory Update Strategy performs better.
4) **Endo21**: To further evaluate the effectiveness of SSL, we conduct experiments on Endo21 [2] Sub-task 2 (Endo21 challenge consists of 4 sub-tasks and only the Sub-task 2 train/test split is available according to the [2]). Experimental results in Table 2.d show that SSL can bring significant improvements to publicly available datasets. Moreover, SSL overperforms current SOTA (YOLO v5 [2]).

5 Conclusion

In this work, we propose Self- and Semi-Supervised Learning (SSL) for GLD tailored for using massive unlabeled gastroscopic to enhance GLD performance. The key novelties of the proposed method include a Hybrid Contrastive Learning method for backbone pre-training and a Prototype-based Pseudo-Label Generation method for semi-supervised learning. Moreover, we contribute the first Large-scale GLD Datasets (LGLDD). Experiments on LGLDD prove that SSL can bring significant improvements to GLD performance. Since annotation cost always limits of datasets scale of such tasks, we hope SSL and LGLDD could fully realize its potential, as well as kindle further research in this direction.

Acknowledgement. This work is supported by Chinese Key-Area Research and Development Program of Guangdong Province (2020B0101350001), the National Natural Science Foundation of China (NO. 61976250), the Guangdong Basic and Applied Basic Research Foundation (NO. 2020B1515020048), the Shenzhen Science and Technology Program (JCYJ20220818103001002, JCYJ20220530141211024), the Shenzhen Sustainable Development Project (KCXFZ20201221173008022), the Guangdong Provincial Key Laboratory of Big Data Computing, and the Chinese University of Hong Kong, Shenzhen.

References

1. Ali, S., et al.: Endoscopy disease detection challenge 2020. arXiv preprint arXiv:2003.03376 (2020)
2. Ali, S., et al.: Assessing generalisability of deep learning-based polyp detection and segmentation methods through a computer vision challenge. arXiv preprint arXiv:2202.12031 (2022)

3. Chen, T., Kornblith, S., Norouzi, M., Hinton, G.: A simple framework for contrastive learning of visual representations. In: International Conference on Machine Learning, pp. 1597–1607. PMLR (2020)
4. Chen, X., He, K.: Exploring simple Siamese representation learning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 15750–15758 (2021)
5. Deng, J., Dong, W., Socher, R., Li, L.J., Li, K., Fei-Fei, L.: ImageNet: a large-scale hierarchical image database. In: 2009 IEEE Conference on Computer Vision and Pattern Recognition, pp. 248–255. IEEE (2009)
6. Duan, K., Bai, S., Xie, L., Qi, H., Huang, Q., Tian, Q.: CenterNet: keypoint triplets for object detection. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 6569–6578 (2019)
7. Grill, J.B., et al.: Bootstrap your own latent-a new approach to self-supervised learning. In: Advances in Neural Information Processing Systems, vol. 33, pp. 21271–21284 (2020)
8. He, K., Chen, X., Xie, S., Li, Y., Dollár, P., Girshick, R.: Masked autoencoders are scalable vision learners. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 16000–16009 (2022)
9. He, K., Fan, H., Wu, Y., Xie, S., Girshick, R.: Momentum contrast for unsupervised visual representation learning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 9729–9738 (2020)
10. Li, X., et al.: Dense semantic contrast for self-supervised visual representation learning. In: Proceedings of the 29th ACM International Conference on Multimedia, pp. 1368–1376 (2021)
11. Lin, T.-Y., et al.: Microsoft COCO: common objects in context. In: Fleet, D., Pajdla, T., Schiele, B., Tuytelaars, T. (eds.) ECCV 2014. LNCS, vol. 8693, pp. 740–755. Springer, Cham (2014). https://doi.org/10.1007/978-3-319-10602-1_48
12. Liu, Y.C., et al.: Unbiased teacher for semi-supervised object detection. arXiv preprint arXiv:2102.09480 (2021)
13. Ren, S., He, K., Girshick, R., Sun, J.: Faster R-CNN: towards real-time object detection with region proposal networks. In: Advances in Neural Information Processing Systems, vol. 28 (2015)
14. Sohn, K., Zhang, Z., Li, C.L., Zhang, H., Lee, C.Y., Pfister, T.: A simple semi-supervised learning framework for object detection. arXiv preprint arXiv:2005.04757 (2020)
15. Xie, Z., et al.: SimMIM: a simple framework for masked image modeling. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 9653–9663 (2022)
16. Xu, M., et al.: End-to-end semi-supervised object detection with soft teacher. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 3060–3069 (2021)
17. Yan, P., et al.: Semi-supervised video salient object detection using pseudo-labels. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 7284–7293 (2019)
18. Zhang, R., et al.: Lesion-aware dynamic kernel for polyp segmentation. In: Wang, L., Dou, Q., Fletcher, P.T., Speidel, S., Li, S. (eds.) International Conference on Medical Image Computing and Computer-Assisted Intervention, pp. 99–109. Springer, Cham (2022). https://doi.org/10.1007/978-3-031-16437-8_10
19. Zhang, R., Li, G., Li, Z., Cui, S., Qian, D., Yu, Y.: Adaptive context selection for polyp segmentation. In: Martel, A.L., et al. (eds.) MICCAI 2020. LNCS, vol.

12266, pp. 253–262. Springer, Cham (2020). https://doi.org/10.1007/978-3-030-59725-2_25
20. Zhang, R., Liu, S., Yu, Y., Li, G.: Self-supervised correction learning for semi-supervised biomedical image segmentation. In: de Bruijne, M., et al. (eds.) MICCAI 2021. LNCS, vol. 12902, pp. 134–144. Springer, Cham (2021). https://doi.org/10.1007/978-3-030-87196-3_13
21. Zhao, X., Fang, C., Fan, D.J., Lin, X., Gao, F., Li, G.: Cross-level contrastive learning and consistency constraint for semi-supervised medical image segmentation. In: 2022 IEEE 19th International Symposium on Biomedical Imaging (ISBI), pp. 1–5 (2022). https://doi.org/10.1109/ISBI52829.2022.9761710
22. Zhao, X., et al.: Semi-supervised spatial temporal attention network for video polyp segmentation. In: Wang, L., Dou, Q., Fletcher, P.T., Speidel, S., Li, S. (eds.) International Conference on Medical Image Computing and Computer-Assisted Intervention, pp. 456–466. Springer, Cham (2022). https://doi.org/10.1007/978-3-031-16440-8_44
23. Zhou, H., et al.: Dense teacher: dense pseudo-labels for semi-supervised object detection. In: Avidan, S., Brostow, G., Cissé, M., Farinella, G.M., Hassner, T. (eds.) Computer Vision-ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part IX, pp. 35–50. Springer, Cham (2022). https://doi.org/10.1007/978-3-031-20077-9_3