

Holistic White-light Polyp Classification via Alignment-free Dense Distillation of Auxiliary Optical Chromoendoscopy

Qiang Hu^{†1}, Qimei Wang^{†1}, Jia Chen², Xuanta Ji², Mei Liu³, Qiang Li (✉)¹,
and Zhiwei Wang (✉)¹

¹ Wuhan National Laboratory for Optoelectronics, Huazhong University of Science and Technology

² Changzhou United Imaging Surgical Co., Ltd

³ Tongji Medical College of Huazhong University of Science and Technology
{huqiang77, qimei_wang, liqiang8, zwwang}@hust.edu.cn

Abstract. White Light Imaging (WLI) and Narrow Band Imaging (NBI) are the two main colonoscopic modalities for polyp classification. While NBI, as optical chromoendoscopy, offers valuable vascular details, WLI remains the most common and often the only available modality in resource-limited settings. However, WLI-based methods typically underperform, limiting their clinical applicability. Existing approaches transfer knowledge from NBI to WLI through global feature alignment but often rely on cropped lesion regions, which are susceptible to detection errors and neglect contextual and subtle diagnostic cues. To address this, this paper proposes a novel holistic classification framework that leverages full-image diagnosis without requiring polyp localization. The key innovation lies in the Alignment-free Dense Distillation (ADD) module, which enables fine-grained cross-domain knowledge distillation regardless of misalignment between WLI and NBI images. Without resorting to explicit image alignment, ADD learns pixel-wise cross-domain affinities to establish correspondences between feature maps, guiding the distillation along the most relevant pixel connections. To further enhance distillation reliability, ADD incorporates Class Activation Mapping (CAM) to filter cross-domain affinities, ensuring the distillation path connects only those semantically consistent regions with equal contributions to polyp diagnosis. Extensive results on public and in-house datasets show that our method achieves state-of-the-art performance, relatively outperforming the other approaches by at least 2.5% and 16.2% in AUC, respectively. Code is available at: <https://github.com/Huster-Hq/ADD>.

Keywords: Colorectal polyps classification · Alignment-free distillation · Cross-domain learning.

1 Introduction

Colonoscopy is the gold standard for colorectal cancer (CRC) screening and diagnosis, where accurately distinguishing between *adenomatous* and *hyperplastic*

[†] Equal contribution; ✉ corresponding author.

polyps is essential [1]. Adenomatous polyps require removal due to their malignant potential, while hyperplastic polyps typically only necessitate periodic monitoring. However, variations in clinical experience can lead to inconsistencies in diagnosis. To enhance diagnostic objectivity, automated polyp classification algorithms for colonoscopy have been extensively studied [2,3,4].

Clinically, colonoscopy utilizes two primary imaging modes for polyp classification: white-light imaging (WLI), the standard modality, and narrow-band imaging (NBI), the optical chromoendoscopy modality. While NBI enhances vascular and morphological details by filtering specific wavelengths, WLI remains the predominant and only available option in resource-limited settings, such as capsule endoscopy. However, a significant performance gap exists between WLI- and NBI-based methods, with WLI generally exhibiting lower classification accuracy [5,6], limiting the broader adoption of polyp classification techniques.

To address the accuracy gap and enhance the diagnostic precision of WLI, existing methods [7,8] have sought to leverage domain alignment strategies, knowledge from NBI to WLI to improve its discriminative capabilities. However, these approaches fundamentally rely on cropped lesion patches for polyp classification, thereby necessitating additional detection steps [9,10] that may introduce systematic errors and hinder generalizability. Furthermore, by focusing solely on cropped lesion regions and employing global alignment techniques, these methods frequently overlook valuable contextual information (*e.g.*, growth position and color contrast with surrounding tissue) and fail to adequately capture subtle diagnostic vascular information in the NBI domain.

In this paper, we propose a novel holistic classification framework designed to eliminate geometric priors while effectively integrating both lesion characteristics and contextual information through cross-domain dense knowledge distillation. However, WLI and NBI images are often misaligned due to temporal discrepancies, variations in viewing angles, and motion artifacts, rendering conventional fine-grained guidance from NBI to WLI unreliable. To address this challenge, we introduce the innovative Alignment-free Dense Distillation (ADD) module, which rectifies dense knowledge distillation pathways from NBI to WLI along cross-domain pixel-wise affinities. These affinities identify the most relevant counterparts for each pixel between WLI and NBI feature maps, ensuring reliable knowledge transfer without requiring explicit spatial alignment overhead. To further enhance distillation reliability, we incorporate Class Activation Mapping (CAM) to filter cross-domain affinities, constraining distillation pathways to link regions across domains that contribute equally to polyp diagnosis.

In summary, our major contributions are as follows: (1) We propose a novel holistic WLI polyp classification framework. To the best of our knowledge, this is the first exploration of leveraging NBI knowledge to assist holistic WLI polyp classifier without requiring polyp location. (2) We propose the Alignment-free Dense Distillation (ADD) module, which establishes dense distillation pathways between misaligned cross-domain features guided by learned affinities. Additionally, we capture the semantic relations to ensure distillation is restricted to semantically consistent regions. (3) Extensive experiments demonstrate that

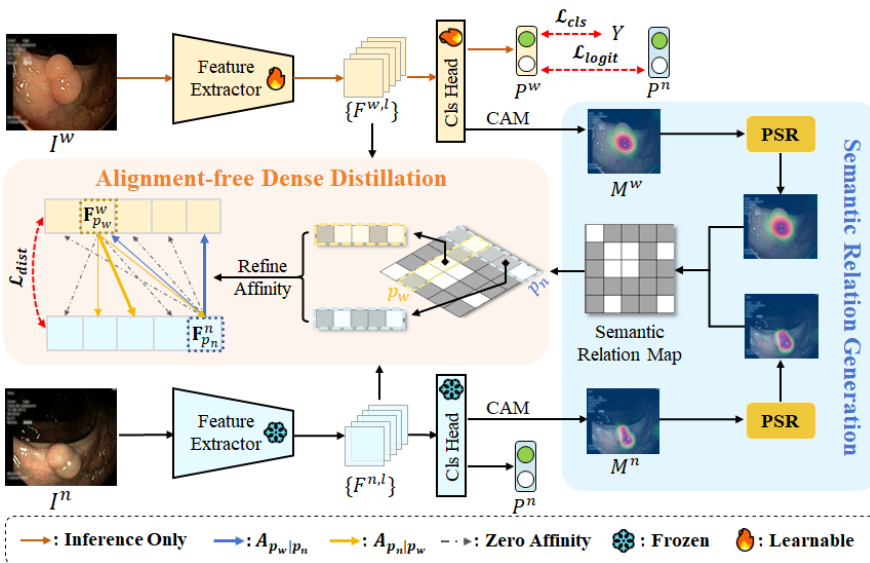


Fig. 1. Overview of our framework, which focuses on enabling knowledge transfer from a frozen NBI classifier to a learnable WLI classifier. Its core module, Alignment-free Dense Distillation (ADD), adaptively establishes distillation pathways between misaligned cross-domain features, guided by bidirectional affinities. Furthermore, we introduce the Semantic Relation Generation (SRG) module, which generates a semantic relation map between cross-domain features based on the CAM maps optimized by Pairwise Similarity Refinement (PSR). This semantic relation map ensures that distillation is restricted to cross-domain feature vectors sharing the consistent semantic.

our method achieves the state-of-the-art performance in WLI image classification on both the public CPC-Paired and our in-house datasets, with relative improvements of at least 2.5% and 16.2% in AUC, respectively.

2 Method

As shown in Fig. 1, our framework consists of two classifiers: a WLI classifier and a well-trained NBI classifier. During training, the NBI classifier remains frozen, while we introduce the Alignment-free Dense Distillation (ADD) module to facilitate efficient cross-domain knowledge transfer. ADD operates directly on the entire context-rich images, eliminating the need for explicit polyp location.

2.1 Alignment-free Dense Distillation (ADD)

Given paired but unaligned WLI-NBI images $(I^w, I^n) \in \mathbb{R}^{3 \times H \times W}$ and their shared ground truth (GT) category label $Y \in \{0, 1\}$, where 0 indicates hyperplastic and 1 indicates adenomatous, we first pass I^w and I^n through the

ResNet backbones of the learnable WLI classifier and the frozen NBI classifier to extract multi-scale feature maps, $\{\mathbf{F}^{w,l}, \mathbf{F}^{n,l}\} \in \mathbb{R}^{C_l \times H_l \times W_l}$, where l denotes the resolution levels of intermediate feature maps. These feature maps are processed independently at each scale, with no cross-scale operations involved, and therefore, we omit the scale notation in the following sections.

In typical clinical acquisitions, (I^w, I^n) exhibit geometric misalignment due to endoscopic movement between modality switches. This spatial inconsistency causes cross-domain feature vectors at the same spatial location to potentially represent divergent anatomical regions, making their distillation suboptimal. To address this issue, we propose the ADD module to establish dense distillation pathways guided by cross-domain affinities, which capture the relevance between cross-domain feature vectors.

For clarity, we introduce notations p and k as spatial position index, and use the subscripts w and n to indicate which modality being indexed (*e.g.*, p_w for WLI and p_n for NBI). Given a WLI feature vector $\mathbf{F}_{p_w}^w$, its normalized affinity to an NBI feature vector $\mathbf{F}_{p_n}^n$ (p_w can differ from p_n) can be formulated as follows:

$$A_{p_n|p_w} = \frac{\exp(S_{p_w,p_n}) \cdot R_{p_w,p_n}}{\sum_{k_n=0}^{H_l W_l - 1} (\exp(S_{p_w,k_n}) \cdot R_{p_w,k_n})}, \quad (1)$$

where S_{p_w,p_n} denotes the cosine similarity between $\mathbf{F}_{p_w}^w$ and $\mathbf{F}_{p_n}^n$. Note that R_{p_w,p_n} is a symmetric semantic relation between $\mathbf{F}_{p_w}^w$ and $\mathbf{F}_{p_n}^n$, which is used to filter out unreliable S_{p_w,p_n} . This will be explained in detail in Sec. 2.2.

To further handle domain discrepancies and better model the relationship between cross-domain feature vectors, we introduce a bidirectional affinity mechanism by simultaneously capturing the reverse affinity. That is, given $\mathbf{F}_{p_n}^n$, its affinity to $\mathbf{F}_{p_w}^w$ is defined as follows:

$$A_{p_w|p_n} = \frac{\exp(S_{p_n,p_w}) \cdot R_{p_n,p_w}}{\sum_{k_w=0}^{H_l W_l - 1} (\exp(S_{p_n,k_w}) \cdot R_{p_n,k_w})}. \quad (2)$$

Both $A_{p_n|p_w}$ and $A_{p_w|p_n}$ can be seen as the probability of a distillation path existing between $\mathbf{F}_{p_w}^w$ and $\mathbf{F}_{p_n}^n$, but they are asymmetric since they are computed from different image modality perspectives. We apply a simple symmetrization operation by taking the union of both affinities, establishing symmetric dense distillation pathways between the WLI and NBI feature maps. The dense distillation loss can then be expressed as follows:

$$\mathcal{L}_{dist} = \sum_{p_w=0}^{H_l W_l - 1} \sum_{p_n=0}^{H_l W_l - 1} (A_{p_n|p_w} + A_{p_w|p_n} - A_{p_n|p_w} \cdot A_{p_w|p_n}) \cdot \|\mathbf{F}_{p_w}^w - \mathbf{F}_{p_n}^n\|. \quad (3)$$

Here, every pixel between the unaligned WLI-NBI images can be connected, and knowledge is distilled from NBI to WLI according to the probability of path connectivity $(A_{p_n|p_w} + A_{p_w|p_n} - A_{p_n|p_w} \cdot A_{p_w|p_n})$.

2.2 Semantic Relations Generation (SRG) for Refining Affinities

In addition to densifying and rectifying the distillation pathways, we ensure that distillation occurs only between cross-domain feature vectors that share the same semantic meaning. Inspired by works [11, 12] using Class Activation Mapping (CAM) [13] to generate pseudo masks, we first generate CAM maps for the WLI and NBI images, denoted as M^w and M^n , respectively, to roughly indicate polyp regions. Next, we apply Pairwise Similarity Refinement (PSR) to iteratively refine these maps by aggregating pixel values within a window, guided by pairwise RGB-color similarities, for T iterations. Taking M^w as an example, at iteration t , the map is updated as follows:

$$M_{p_w}^{w,t} = \sum_{k_w \in \mathcal{N}(p_w)} \lambda_{k_w} \cdot M_{k_w}^{w,t-1}, \quad \lambda_{k_w} = \frac{1 - \|I_{p_w}^w - I_{k_w}^w\|}{\sum_{k_w \in \mathcal{N}(p_w)} (1 - \|I_{p_w}^w - I_{k_w}^w\|)}, \quad (4)$$

where $\mathcal{N}(\cdot)$ denotes the 3×3 neighboring pixel set.

Afterward, we resize the final refined CAM maps to the size of the feature maps. We then apply two thresholds, τ_1 and τ_2 ($0 < \tau_1 < \tau_2 < 1$), to binarize the maps, yielding $\mathbf{M}^w, \mathbf{M}^n \in \{0, \oslash, 1\}^{H_l \times W_l}$, where 0, $\oslash$, and 1 represent the background, unsure, and polyp, respectively. To determine whether the cross-domain pixels belong to the same class (excluding unsure pixels), the semantic relation between $\mathbf{F}_{p_w}^w$ and $\mathbf{F}_{p_n}^n$ is defined as follows:

$$R_{p_w, p_n} = \begin{cases} 1 & \text{if } \mathbf{M}_{p_w}^w = \mathbf{M}_{p_n}^n \text{ and } \mathbf{M}_{p_w}^w, \mathbf{M}_{p_n}^n \neq \oslash \\ 0 & \text{otherwise} \end{cases}. \quad (5)$$

R_{p_w, p_n} can be used to calculate $A_{p_n|p_w}$ and $A_{p_w|p_n}$ in Eq.(1) and Eq.(2), further pruning those unreliable WLI-NBI distillation paths connecting inconsistent CAM-derived semantics.

2.3 Training details

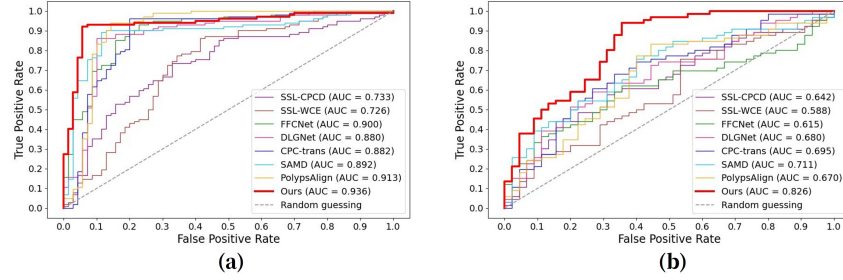
In addition to $\mathcal{L}_{dist}$, we compute additional loss terms: $\mathcal{L}_{logit}$ and $\mathcal{L}_{cls}$, to train the WLI classifier. Specifically, $\mathcal{L}_{logit}$ supervises the WLI classifier using the logit output of the NBI classifier as soft labels, while $\mathcal{L}_{cls}$ supervises the WLI classifier using the GT label Y . The overall training loss is formulated as:

$$\mathcal{L}_{total} = \mathcal{L}_{dist} + \mathcal{L}_{logit} + \mathcal{L}_{cls}, \quad (6)$$

where $\mathcal{L}_{cls} = \text{CE}(P^w, Y)$ and $\mathcal{L}_{logit} = \|P^w - P^n\|$. Here, $\text{CE}(\cdot, \cdot)$ denotes the cross-entropy loss, and P^w and P^n represent the predicted logits of the WLI and NBI classifiers, respectively. The implementation is by PyTorch [14] and the training is on a RTX 4090 GPU with 24GB memory. Weights of ResNet-50 [15] pre-trained on ImageNet [16] are loaded as initialization. The input size is set to 448×448 , and the batch size is set to 16. The optimizer is set to Adam [17] with $1e-8$ weight decay. The number of epochs is set to 200 and the initial learning rate is set to $1e-4$. We set T , τ_1 , and τ_2 as 10, 0.3, and 0.7, respectively, according to the parameter search experiment.

Table 1. Quantitative comparison on CPC-Paired dataset and our in-house dataset. The best performance is marked in bold and the second-best results are underlined.

Method	CPC-Paired						In-house					
	Acc	Pre	Rec	Spe	F1	AUC	Acc	Pre	Rec	Spe	F1	AUC
SSL-CPCD [18]	0.709	0.720	0.833	0.529	0.772	0.733	0.577	0.711	0.485	0.711	0.577	0.642
SSL-WCE [19]	0.750	0.752	0.863	0.586	0.804	0.726	0.523	0.614	0.530	0.511	0.569	0.588
FFCNet [20]	0.860	0.836	0.951	0.729	0.890	0.900	0.622	0.707	0.621	<u>0.622</u>	0.661	0.615
DLGNet [21]	0.872	0.917	0.863	<u>0.886</u>	0.889	0.880	0.649	0.690	0.742	0.511	0.715	0.680
CPC-trans [8]	0.895	0.875	0.961	0.800	0.916	0.882	0.667	0.644	0.985	0.200	<u>0.779</u>	0.695
SAMD [22]	0.884	0.902	0.902	0.857	0.902	0.892	0.694	0.700	0.848	0.467	0.767	<u>0.711</u>
PolypsAlign [7]	<u>0.907</u>	<u>0.906</u>	<u>0.941</u>	0.857	<u>0.923</u>	<u>0.913</u>	<u>0.712</u>	<u>0.724</u>	0.833	0.533	0.775	0.670
Ours	0.930	0.959	0.922	0.943	0.940	0.936	0.802	0.775	<u>0.939</u>	0.600	0.849	0.826

**Fig. 2.** Comparison of ROC curves on (a) CPC-Paired and (b) our in-house dataset.

3 Experiments

3.1 Datasets and Evaluation Metrics

We evaluate our method on two datasets: the CPC-Paired dataset [7] and our in-house dataset, both containing misaligned WLI-NBI polyp image pairs. The in-house dataset, collected from a local hospital, was annotated by two experienced colonoscopists. Both datasets consist of two polyp categories: hyperplastic and adenomatous. The CPC-Paired dataset includes 63 hyperplastic and 102 adenomatous pairs, while our in-house dataset contains 514 hyperplastic and 598 adenomatous pairs.

We perform 5-fold cross-validation on both datasets, ensuring patient-level separation between the train and test sets. To evaluate performance, we use six metrics: Accuracy (ACC), Precision (Pre), Sensitivity (Sen), Specificity (Spe), F1-score (F1), and Area Under the Curve (AUC). All evaluations are conducted on the WLI test set.

3.2 Comparisons with State-of-the-art Methods

We compare our method with seven state-of-the-art (SOTA) methods: four cross-domain independent classification (CIC) methods [18,19,20,21] and three cross-domain distillation classification (CDC) methods [7,8,22]. For CIC methods,

Table 2. Ablation study of two key components: Alignment-free Dense Distillation (ADD) and Semantic Relation Generation (SRG).

Components		CPC-Paired						In-house					
ADD	SRG	Acc	Pre	Rec	Spe	F1	AUC	Acc	Pre	Rec	Spe	F1	AUC
		0.721	0.776	0.745	0.686	0.760	0.801	0.568	0.650	0.591	0.533	0.619	0.603
✗	✗	0.826	0.891	0.804	0.857	0.845	0.857	0.676	0.800	0.606	0.778	0.690	0.683
✓	✗	0.913	0.922	0.931	0.886	0.926	0.925	0.766	0.885	0.697	0.867	0.780	0.775
✓	✓	0.930	0.959	0.922	0.943	0.940	0.936	0.802	0.775	0.939	0.600	0.849	0.826

Table 3. Ablation study of two sub-strategies: Bidirectional Affinities and Pairwise Similarity Refinement, denoted as Bi-A and PSR, respectively.

Methods	CPC-Paired						In-house					
	Acc	Pre	Rec	Spe	F1	AUC	Acc	Pre	Rec	Spe	F1	AUC
Ours	0.930	0.959	0.922	0.943	0.940	0.936	0.802	0.775	0.939	0.600	0.849	0.826
w/o Bi-A	0.913	0.931	0.922	0.900	0.926	0.918	0.757	0.783	0.818	0.667	0.800	0.762
w/o PSR	0.919	0.907	0.961	0.857	0.933	0.928	0.775	0.815	0.803	0.733	0.809	0.786

we train models without any cross-domain interaction. To ensure fairness, all methods use ResNet-50 [15] as the backbone, except CPC-trans, which uses ViT-S [23]. Notably, we implement CPC-trans [8] and PolypsAlign [7] to receive images cropped by GT polyp boxes during training and testing, following their original settings.

Table 1 shows the quantitative comparison on the CPC-Paired and in-house datasets. CDC methods generally outperform CIC methods, confirming that knowledge transfer from the NBI domain to the WLI domain improves WLI classification. Our method achieves the best performance on most metrics for both datasets. Surprisingly, our method, which receives holistic images, outperforms the two methods that receive GT polyp patches, which demonstrates the importance of global context for polyp classification. Additionally, the ROC curves in Fig. 2 show that our method achieves the highest True Positive Rates across most False Positive Rates. This improvement is due to our alignment-free dense distillation (ADD) module, which handles spatial misalignment and effectively reconstructs distillation pathways, unlike other CDC methods that rely on pixel-level alignment, missing complex spatial relationships across domains.

3.3 Ablation Study

Effectiveness of Key Components To evaluate the effectiveness of two key components, *i.e.*, ADD and SRG, we train variants of our method by disabling ADD and/or SRG, and compare them with a variant trained in CIC manner. Specifically, we disable ADD and SRG by replacing $\mathcal{L}_{dist}$ in Eq. (6) with a consistency loss between paired WLI-NBI features, and train the CIC variant similarly to the methods in Table 1. Note that SRG is designed to assist ADD, so it cannot operate independently. The comparison results in Table 2 show that even

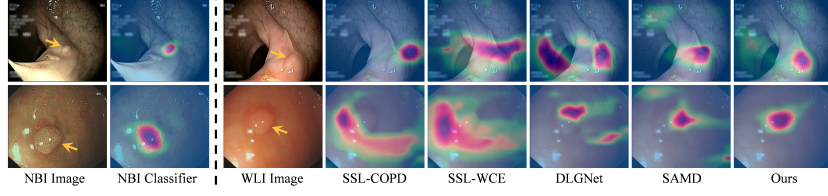


Fig. 3. Visualization of CAM maps of different methods, which reflect the ‘attention region’ while model classifying. Yellow arrows indicate polyps.

without optimal feature-level distillation, aligning output logits alone can effectively improve classification accuracy for WLI images (comparing the first two rows). Further analysis reveals that the ADD module significantly boosts performance, with AUC increasing by 0.068 and 0.092 on the CPC-Paired and in-house datasets, respectively (comparing the first three rows). This improvement stems from ADD’s ability to establish adaptive distillation paths, effectively addressing spatial misalignment. Finally, the SRG module further enhances performance by restricting distillation paths to features with semantic consistency, optimizing knowledge transfer.

Effectiveness of Sub-strategies To assess the impact of sub-strategies, *i.e.*, bidirectional affinities (Bi-A) and pairwise similarity refinement (PSR), we train variants with these components disabled. Specifically, we disable Bi-A by keeping only $A_{p_n|p_w}$ in Eq. (3). As shown in Table 3, removing Bi-A or PSR results in varying performance degradation. Undirectional affinities fail to capture complex cross-modal relationships, while the absence of PSR introduces noise into the semantic relations from coarse CAM maps.

3.4 Visualization of CAM

We visualize the CAM maps of different classifiers that receive holistic images in Fig. 3. CDC methods show more precise focus on polyp regions than CIC methods, confirming the importance of NBI knowledge distillation for WLI classification. Our method, in particular, produces attention regions that are most aligned with the NBI classifier, demonstrating the effectiveness of our cross-domain distillation and validating the use of CAM to guide relation map construction.

4 Conclusion

In this work, we propose a novel framework for WLI polyp classification, introducing the Alignment-Free Dense Distillation (ADD) module and Semantic Relation Generation (SRG) module. The ADD module enables effective knowledge distillation from the NBI domain to the WLI domain, while the SRG module

ensures distillation occurs only between semantically consistent regions, optimizing the knowledge transfer. By addressing spatial misalignment and ensuring semantic alignment, our method significantly enhances WLI classification accuracy. Experimental results on the CPC-Paired and in-house datasets show that our method outperforms state-of-the-art polyp classification methods, achieving superior performance. Visualization results further confirm that our method produces the most accurate attention regions, validating the alignment of cross-domain features and demonstrating the effectiveness of CAM-guided distillation.

Acknowledgments. This work was supported in part by National Natural Science Foundation of China (Grant No. 62202189), Key R&D Program of Hubei Province of China (No. 2023BCB003), and Changzhou United Imaging Surgical Co., Ltd. Thanks to Professor Mei Liu’s team for providing the in-house dataset used in this work at the Department of Gastroenterology, Tongji Medical College, Huazhong University of Science and Technology.

Disclosure of Interests. The authors have no competing interests to declare that are relevant to the content of this article.

References

1. Roger Fonollà, Quirine EW van der Zander, Ramon M Schreuder, Ad AM Masclee, Erik J Schoon, Fons van der Sommen, and Peter HN de With. A cnn cadx system for multimodal classification of colorectal polyps combining wl, bli, and lei modalities. *Applied Sciences*, 10(15):5040, 2020.
2. Roger Fonolla, Fons van der Sommen, Ramon M Schreuder, Erik J Schoon, and Peter HN de With. Multi-modal classification of polyp malignancy using cnn features with balanced class augmentation. In *2019 IEEE 16th International Symposium on Biomedical Imaging (ISBI 2019)*, pages 74–78. IEEE, 2019.
3. Yu Tian, Leonardo ZCT Pu, Rajvinder Singh, Alastair D Burt, and Gustavo Carneiro. One-stage five-class polyp detection and classification. In *2019 IEEE 16th international symposium on biomedical imaging (ISBI 2019)*, pages 70–73. IEEE, 2019.
4. Juan Francisco Ortega-Morán, Águeda Azpeitia, Luisa F Sánchez-Peralta, Luis Bote-Curiel, Blas Pagador, Virginia Cabezón, Cristina L Saratxaga, and Francisco M Sánchez-Margallo. Medical needs related to the endoscopic technology and colonoscopy for colorectal cancer diagnosis. *BMC cancer*, 21(1):467, 2021.
5. Hiroyasu Usami, Yuji Iwahori, Yoshinori Adachi, Manas Kamal Bhuyan, Aili Wang, Satoshi Inoue, Masahide Ebi, Naotaka Ogasawara, and Kunio Kasugai. Colorectal polyp classification based on latent sharing features domain from multiple endoscopy images. *Procedia Computer Science*, 176:2507–2514, 2020.
6. Franklin Sierra-Jerez and Fabio Martínez. A deep representation to fully characterize hyperplastic, adenoma, and serrated polyps on narrow band imaging sequences. *Health and Technology*, 12(2):401–413, 2022.
7. Qin Wang, Hui Che, Weizhen Ding, Li Xiang, Guanbin Li, Zhen Li, and Shuguang Cui. Colorectal polyp classification from white-light colonoscopy images via domain alignment. In *Medical Image Computing and Computer Assisted Intervention—MICCAI 2021: 24th International Conference, Strasbourg, France, September 27–October 1, 2021, Proceedings, Part VII 24*, pages 24–32. Springer, 2021.

8. Weijie Ma, Ye Zhu, Ruimao Zhang, Jie Yang, Yiwen Hu, Zhen Li, and Li Xiang. Toward clinically assisted colorectal polyp recognition via structured cross-modal representation consistency. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 141–150. Springer, 2022.
9. Qiang Hu, Zhenyu Yi, Ying Zhou, Fang Peng, Mei Liu, Qiang Li, and Zhiwei Wang. Sali: Short-term alignment and long-term interaction network for colonoscopy video polyp segmentation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 531–541. Springer, 2024.
10. Qiang Hu, Zhenyu Yi, Ying Zhou, Fan Huang, Mei Liu, Qiang Li, and Zhiwei Wang. Monobox: Tightness-free box-supervised polyp segmentation using monotonicity constraint. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pages 3572–3580, 2025.
11. Zhaozheng Chen, Tan Wang, Xiongwei Wu, Xian-Sheng Hua, Hanwang Zhang, and Qianru Sun. Class re-activation maps for weakly-supervised semantic segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 969–978, 2022.
12. Lixiang Ru, Heliang Zheng, Yibing Zhan, and Bo Du. Token contrast for weakly-supervised semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3093–3102, 2023.
13. Bolei Zhou, Aditya Khosla, Agata Lapedriza, Aude Oliva, and Antonio Torralba. Learning deep features for discriminative localization. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2921–2929, 2016.
14. Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. Pytorch: An imperative style, high-performance deep learning library. *Advances in neural information processing systems*, 32, 2019.
15. 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.
16. Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009.
17. Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.
18. Ziang Xu, Jens Rittscher, and Sharib Ali. Ssl-cpcd: Self-supervised learning with composite pretext-class discrimination for improved generalisability in endoscopic image analysis. *IEEE Transactions on Medical Imaging*, 2024.
19. Xiaoqing Guo and Yixuan Yuan. Semi-supervised wce image classification with adaptive aggregated attention. *Medical Image Analysis*, 64:101733, 2020.
20. Kai-Ni Wang, Yuting He, Shuaishuai Zhuang, Juzheng Miao, Xiaopu He, Ping Zhou, Guanyu Yang, Guang-Quan Zhou, and Shuo Li. Ffcnnet: Fourier transform-based frequency learning and complex convolutional network for colon disease classification. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 78–87. Springer, 2022.
21. Kai-Ni Wang, Shuaishuai Zhuang, Qi-Yong Ran, Ping Zhou, Jie Hua, Guang-Quan Zhou, and Xiaopu He. Dlgnet: A dual-branch lesion-aware network with the supervised gaussian mixture model for colon lesions classification in colonoscopy images. *Medical Image Analysis*, 87:102832, 2023.
22. Yu Shen, Xijun Wang, Peng Gao, and Ming Lin. Auxiliary modality learning with generalized curriculum distillation. In *International Conference on Machine Learning*, pages 31057–31076. PMLR, 2023.

23. Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.