

Multimodal Medical Image Classification via Synergistic Learning Pre-training

Qinghua Lin, *Student Member, IEEE*, Guang-Hai Liu, Zuoyong Li, *Senior Member, IEEE*,
Yang Li, *Senior Member, IEEE*, Yuting Jiang and Xiang Wu

Abstract—Multimodal pathological images are usually in clinical diagnosis, but computer vision-based multimodal image-assisted diagnosis faces challenges with modality fusion, especially in the absence of expert-annotated data. To achieve the modality fusion in multimodal images with label scarcity, we propose a novel “pretraining + fine-tuning” framework for multimodal semi-supervised medical image classification. Specifically, we propose a synergistic learning pretraining framework of consistency, reconstructive, and aligned learning. By treating one modality as an augmented sample of another modality, we implement a self-supervised learning pre-train, enhancing the baseline model’s feature representation capability. Then, we design a fine-tuning method for multimodal fusion. During the fine-tuning stage, we set different encoders to extract features from the original modalities and provide a multimodal fusion encoder for fusion modality. In addition, we propose a distribution shift method for multimodal fusion features, which alleviates the prediction uncertainty and overfitting risks caused by the lack of labeled samples. We conduct extensive experiments on the publicly available gastroscopy image datasets Kvasir and Kvasirv2. Quantitative and qualitative results demonstrate that the proposed method outperforms the current state-of-the-art classification methods. The code will be released at: <https://github.com/LQH89757/MICS>.

Index Terms—Multimodal, Medical image classification, Semi-supervised learning, Distribution shift, Deep learning.

This work was supported in part by National Natural Science Foundation of China (62471207), Natural Science Foundation of Fujian Province (2024J02029), Joint Funds for The Innovation of Science and Technology in Fujian province (2024Y9028, 2023Y9280), Open Project of Fujian Key Laboratory of Medical Big Data Engineering (KLKF202301). (Corresponding authors: Zuoyong Li; Xiang Wu.)

Qinghua Lin is with the College of Biomedical Engineering, Fudan University, Shanghai 200433, China (e-mail: akametriz@163.com).

Guang-Hai Liu is with the College of Computer Science and Engineering, Guangxi Normal University, Guilin 541004, China (e-mail: liuguanghai009@163.com). Zuoyong Li is with the Fujian Provincial Key Laboratory of Information Processing and Intelligent Control, School of Computer and Big Data, Minjiang University, Fuzhou 350121, China (e-mail: fzulzytdq@126.com).

Yang Li is with the Department of Automation Science and Electrical Engineering, Beihang University, Beijing 100191, China (e-mail: liyang@buaa.edu.cn).

Yuting Jiang is with the Department of Digestive Endoscopy, Fuzhou University Affiliated Provincial Hospital, Provincial Clinical Medical College of Fujian Medical University, 350001, Fuzhou, China (email: yutingjiang@fjmu.edu.cn). Xiang Wu is with the Department of Urology, Fuzhou University Affiliated Provincial Hospital, Provincial Clinical Medical College of Fujian Medical University, 350001, Fuzhou, China (email: tianyang0909@pku.org.cn).

This work has been submitted to the IEEE for possible publication. Copyright may be transferred without notice, after which this version may no longer be accessible.

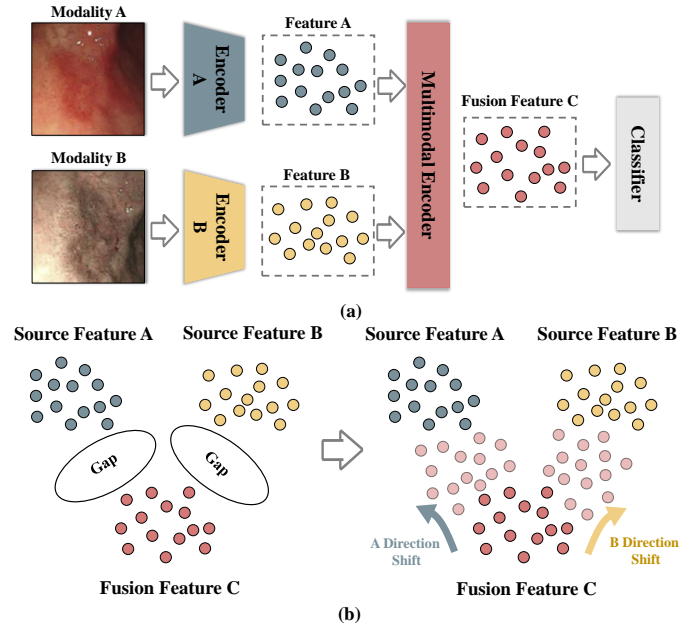


Fig. 1. The proposed distribution shift method. (a) Multimodal features fusion for classification during the fine-tuning stage. (b) The generation of shift directions based on feature distribution guides the augmentation of fused features.

I. INTRODUCTION

WITH the rapid development of medical imaging technology, medical images play an increasingly important role in clinical diagnosis and treatment [1]–[4]. However, single-modal medical image information is limited and often fails to fully reflect the characteristics of lesions, which in turn affects the accuracy of diagnosis. By integrating different imaging techniques, multimodal medical images provide more diverse lesion information and offer a more reliable basis for clinical diagnosis and disease analysis [5]. For example, physicians usually use white light to inspect the patient’s gastrointestinal tract and roughly identify suspicious areas. White light imaging uses the entire visible spectrum to produce red, green, and blue images, making it more sensitive in examining morphology, mucosa, and blood vessel distribution. This helps doctors detect gastric ulcers and tumors and is the current standard mode of endoscopic imaging. When a

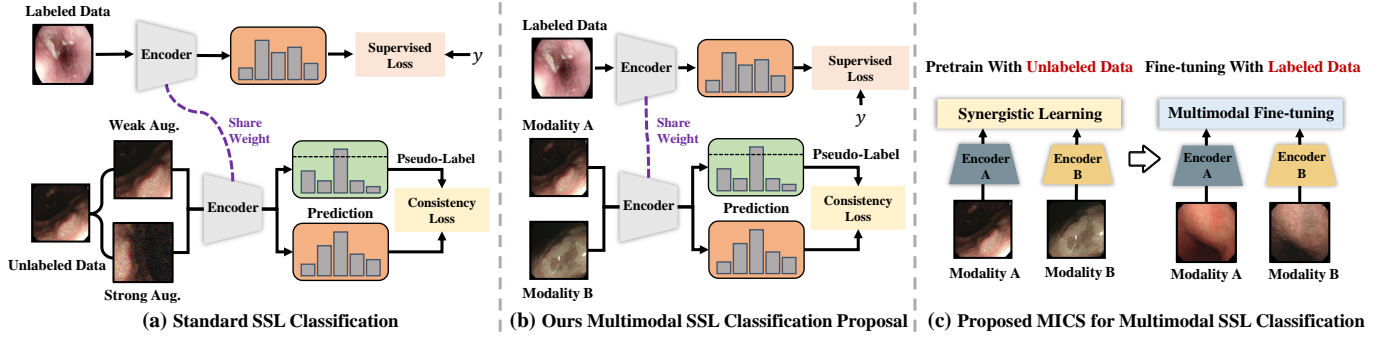


Fig. 2. Comparison of Semi-Supervised Learning (SSL).

suspicious area is identified, the doctor switches to narrow-band imaging to further assess the possible type of condition. Narrow-band imaging uses specific filters to produce light in the blue and green wavelengths, enhancing the mucosal features and vascular structure of suspicious areas, which aids in the early detection of gastric cancer [6].

However, due to the differences and redundancies between modalities, effectively fusing multimodal image information has become a pressing challenge [7]. Deep learning has made significant progress in medical image analysis in recent years, particularly demonstrating superior feature extraction and pattern recognition capabilities. As data-driven methods, deep models require considerable labeled data for training. The acquisition of data labeled by professional physicians is challenging and inefficient. Especially in multimodal tasks, doctors need to repeatedly switch between different modalities to ensure the accuracy of data annotation. Therefore, multimodal medical image tasks face the following two challenges:

(1) Multimodal medical images provide diverse lesion information, but effectively fusing features from different modalities remains challenging.

(2) Data annotation is more time-consuming and labor-intensive in multimodal image tasks than single-modal data.

To address these limitations, we rethink semi-supervised image classification methods and propose a **Multimodal Medical Image Classification via Synergistic Learning Pre-training (MICS)**. As shown in Fig. 2, current semi-supervised classification methods rely on consistency predictions between strong and weak augmentations of unlabeled images. Strong augmentation techniques [8], [9] have successfully perturbed natural images for consistency regularization (e.g., cats, airplanes). However, these standard augmentation techniques can easily destroy lesion features of medical images, and it is hard to find an appropriate augmentation combination [10]. Inspired by these semi-supervised classification methods, we treat one modality of the same disease type as the optimal augmented sample for another modality for consistency prediction. Moreover, the scale of multimodal datasets is usually tiny, and the mutual influence between the training of labeled and unlabeled samples limits the model's optimization [11]. In view of these issues, we separate the training of unlabeled and labeled samples and propose a self-supervised pretraining method based on synergistic learning with consistency, reconstructive, and aligned learning. To achieve multimodal feature fusion

on labeled samples, we propose a fusion fine-tuning method based on feature distribution shift.

Specifically, we use self-supervised learning for the pre-train stage. Consistency learning treats different modalities as distinct augmented samples and predicts their consistency to learn feature representations of unlabeled samples. Reconstructive learning, after randomly masking the original image, extracts features from different encoders to a unified decoder for image reconstruction, focusing on the local detail representations of original modalities. Aligned learning forces the two modalities to approach each other in the high-dimensional feature space through instance-level contrastive learning, further enhancing the similarity of paired image representations.

During the fine-tuning stage, we follow a multimodal fusion framework to fuse feature representations from different modalities. To alleviate overfitting risks caused by a small number of labeled samples, we propose an implicit augmentation method based on a shift vector dictionary (SVD). We use pre-trained weights to construct an SVD for modality clustering. SVD treats the output of the feature by the multimodal encoder as prototypes and obtains shift vectors. Then, according to shift vectors, the fusion features generate perturbed features with a feature distribution close to the initial modality. As shown in Fig. 1, the augmented samples fill the gap between the initial and fused modalities in the latent space, compensating for the model's perceptual differences between the fusion and original modalities.

Our contributions are summarized as follows:

- To the best of our knowledge, we propose the first multimodal-based semi-supervised medical image classification method by rethinking the implementation of multimodal in semi-supervised learning.
- We propose a synergistic learning pretraining framework that combines consistency, reconstructive, and aligned learning, effectively enhancing the model's multimodal representation capability in a self-supervised way.
- We propose a multimodal fusion fine-tuning method based on feature distribution shift, effectively alleviating the overfitting risk caused by the lack of labeled samples and mitigating the model's perceptual differences between the fusion and initial modalities.
- We conduct extensive experiments using two public gastroscopy image datasets, where the paired modality images are generated from WtNGAN [12] to verify the

feasibility of using multimodal images as augmented samples. The experimental results show that the proposed method achieves promising results in multimodal medical image classification.

II. RELATED WORK

In this section, we first review several advanced semi-supervised classification methods, then introduce the current state of the multimodal domain, and finally focus on some works related to multimodal fusion.

A. Semi-Supervised Classification

Semi-supervised learning aims to achieve performance comparable to supervised learning by leveraging a small amount of labeled data and a large amount of unlabeled data. Mainstream semi-supervised classification methods primarily integrate pseudo-labeling and consistency regularization to learn feature representations from unlabeled samples. For example, FixMatch [13] sets a fixed threshold to generate pseudo-labels from weakly augmented unlabeled samples and enforces consistency by predicting these pseudo-labels on their strongly augmented versions. FlexMatch [14] introduces curriculum pseudo-labeling to flexibly adjust thresholds for different classes. FreeMatch [15] dynamically adjusts the pseudo-labeling threshold based on the model's learning state and introduces class fairness regularization to reduce sample bias. SimMatch [11] combines contrastive self-supervised pretraining with consistency regularization, narrowing the gap between supervised and semi-supervised learning during fine-tuning. PEFAT [16] enhances semi-supervised classification performance for medical images through adversarial training from the perspective of loss distribution.

B. Multimodal Task

Multimodal visual tasks have garnered significant attention from researchers in recent years, as multimodal representations effectively leverage information from different modalities to improve downstream task performance. For example, ALBEF [17] aligns feature representations of images and text through cross-modal attention. DeepGuide [18] utilizes knowledge learned from a superior modality to guide the use of an inferior modality, effectively improving diagnostic performance for the latter. FusionM4Net [19] employs a two-stage multimodal learning approach to effectively fuse clinical and dermoscopic image representations at the feature level. GiMP [20] incorporates high-order correlation modeling, introducing a group multi-head self-attention gene encoder to capture the global structure of gene expression. FusAtNet [21] utilizes multi-spectral and hyperspectral images for land cover classification. MSAN [22] focuses on modality-specific features, using two attention modules to guide eye disease classification based on fundus and OCT images.

C. Multimodal Image Fusion

Image fusion combines information from heterogeneous images to generate an image with rich details [23]. MATR [5]

introduces a globally complementary context-adaptive modulation convolution to fuse multimodal medical images. ITFuse [24] employs interactive attention to merge complementary information from infrared and visible light images, effectively leveraging the shared attributes of heterogeneous images. TG-Fuse [25] integrates features extracted by Transformers with shallow features from convolutional neural networks, refining cross-channel interactions within the spatial domain.

Although semi-supervised classification has achieved performance close to supervised classification, it heavily relies on large-scale data and sufficient iterations. However, compared to natural images, acquiring medical imaging data is often more challenging. As a result, semi-supervised models face limitations in learning feature representations from unlabeled samples with small-scale medical imaging datasets. Considering clinicians typically analyze diseased areas using different imaging modalities and the advancements in multimodal tasks within computer vision, we propose MICS to enhance the benchmark for semi-supervised multimodal medical image classification.

III. METHOD

The proposed MICS consists of three stages: the pre-train stage, the dictionary construction stage, and the multimodal fusion fine-tuning stage. The pre-train stage includes consistency learning, reconstructive learning, and aligned learning. In the dictionary construction stage, a Shift Vector Dictionary (SVD) is built based on the feature representations learned from the pre-train stage. During the multimodal fusion fine-tuning stage, the model is initialized with pre-trained weights, while additional training samples are generated using the distribution shifts produced by the SVD.

A. Synergistic Learning Pretrain

The proposed synergistic learning pretrain framework is illustrated in Fig. 3. At the pretrain stage, we extract features from different modalities using single-modal encoders. For the input paired modality images $X = \{w, n\}$, the collaborative learning framework employs encoders f_θ and f_φ to extract the white-light image features z^w and narrow-band image features z^n , respectively. These latent features are first aligned in a low-dimensional feature space through consistency learning and calculate the consistency loss \mathcal{L}_{cons} . Then, we randomly mask $X = \{w, n\}$ as $X_m = \{w_m, w_n\}$. The masked images are then processed by the encoders f_θ and f_φ to extract features z_{mask}^w and z_{mask}^n , which are subsequently reconstructed by a shared decoder g and compute the reconstruction loss \mathcal{L}_{res} . Finally, to align the high-dimensional feature representations of different modalities, z^w and z^n are projected into a high-dimensional latent space using a global embedding GloE(.) for an alignment loss \mathcal{L}_a .

1) *Consistency Learning*: Inspired by image-text contrastive learning [17], we propose a consistency Learning method designed for use prior to the fusion of different modalities. Similar to MoCo [26], the features of the most recently selected K image pairs $T = \{t^w, t^n\}$ are stored in two separate queues as negative samples. During training, these

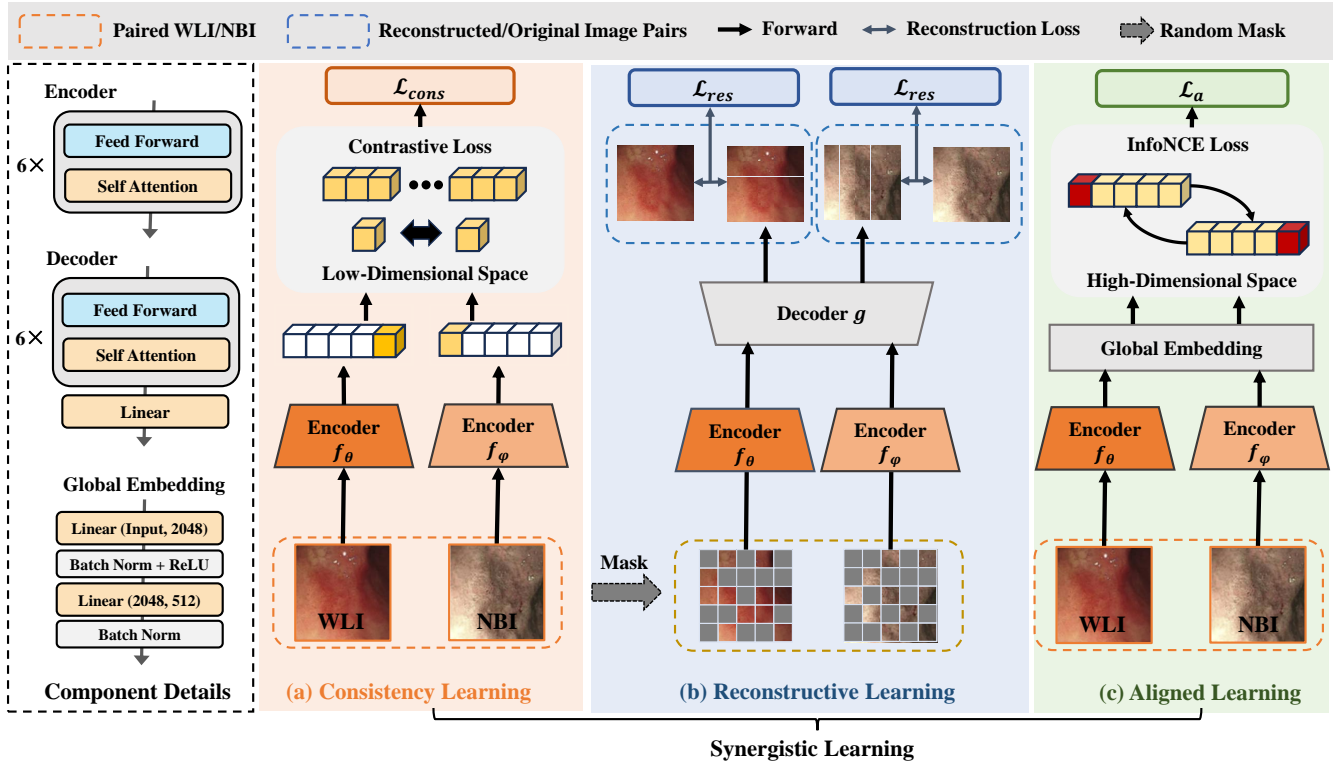


Fig. 3. The overall framework of synergistic learning pre-train: (a) Consistency learning constrains the latent representations of paired images in low-dimensional space to ensure Consistency. (b) Reconstructive learning randomly masks the original images and reconstructs them using a unified decoder based on the output features of single-modal encoders. (c) Aligned learning aims to align more complex and detailed single-modal representations in high-dimensional space. Synergistic learning provides robust single-modal representation capabilities for multimodal feature fusion models.

K image feature pairs are passed through a projection head p , which maps them into a low-dimensional space, and their similarity is computed as follows:

$$\begin{aligned} sim_{wn}^m &= \frac{z^w \times t^{n'}}{\tau}, sim_{nw}^m = \frac{z^n \times t^{w'}}{\tau}, \\ sim_{wn} &= \frac{z^w \times t^{n'}}{\tau}, sim_{nw} = \frac{z^n \times t^{w'}}{\tau}, \end{aligned} \quad (1)$$

where z^w and z^n represent the image features extracted by the momentum model, $t^{n'} = cat((z^w)^T, t^n)$, $t^{w'} = cat((z^n)^T, t^w)$, and τ denotes the temperature parameter. Subsequently, we construct a target similarity matrix \mathbb{I} and compute the target similarity as follows:

$$\begin{aligned} sim_{t_{wn}} &= \alpha \cdot softmax(sim_{wn}^m) + (1 - \alpha) \cdot \mathbb{I}, \\ sim_{t_{nw}} &= \alpha \cdot softmax(sim_{nw}^m) + (1 - \alpha) \cdot \mathbb{I}. \end{aligned} \quad (2)$$

Therefore, the similarity between the features of the current image pair and the features in the momentum queue is defined as:

$$\begin{aligned} \mathcal{L}_{wn} &= -\frac{1}{N} \sum_i \sum_j \log softmax(sim_{wn}) \cdot sim_{t_{wn}}, \\ \mathcal{L}_{nw} &= -\frac{1}{N} \sum_i \sum_j \log softmax(sim_{nw}) \cdot sim_{t_{nw}}. \end{aligned} \quad (3)$$

The final consistency loss \mathcal{L}_{cons} is defined as:

$$\mathcal{L}_{dis} = \frac{\mathcal{L}_{wn} + \mathcal{L}_{nw}}{2}. \quad (4)$$

2) Reconstructive Learning: The masked autoencoder [27] is considered a practical self-supervised learning approach. Therefore, we introduce a generative learning strategy to capture local detail representations across different modalities in the synergistic learning pre-train stage. Different from [27], we employ a unified decoder g for multimodal tasks to facilitate interaction between modal features, which means that regardless of whether the image features extracted by encoder f_θ and f_ϕ , the decoder can reconstruct the original images. The motivation of reconstructive learning is analogous to adversarial learning, where we aim for different modalities to align only in critical regions within the feature space rather than achieving complete consistency. The unified decoder g helps retain modality-specific feature representation capabilities for each encoder.

Specifically, we reconstruct the masked features z_{mask}^w and z_{mask}^n into the original images w^r and n^r :

$$z_{mask}^w = f_\theta(\text{Mask}(w, \sigma)), z_{mask}^n = f_\phi(\text{Mask}(w, \sigma)), \quad (5)$$

where $\text{Mask}(\cdot)$ denotes random masking, and $\sigma = 0.75$ represents the masking ratio. Then, through the decoder g , the reconstruction loss \mathcal{L}_{res} is defined as:

$$\mathcal{L}_{res} = \|w, g(z_{mask}^w)\|_2 + \|n, g(z_{mask}^n)\|_2. \quad (6)$$

3) Aligned Learning: The research [28] reveals the role of high-dimensional features in domain adaptation. For medical images, high-dimensional vectors facilitate the model to enrich

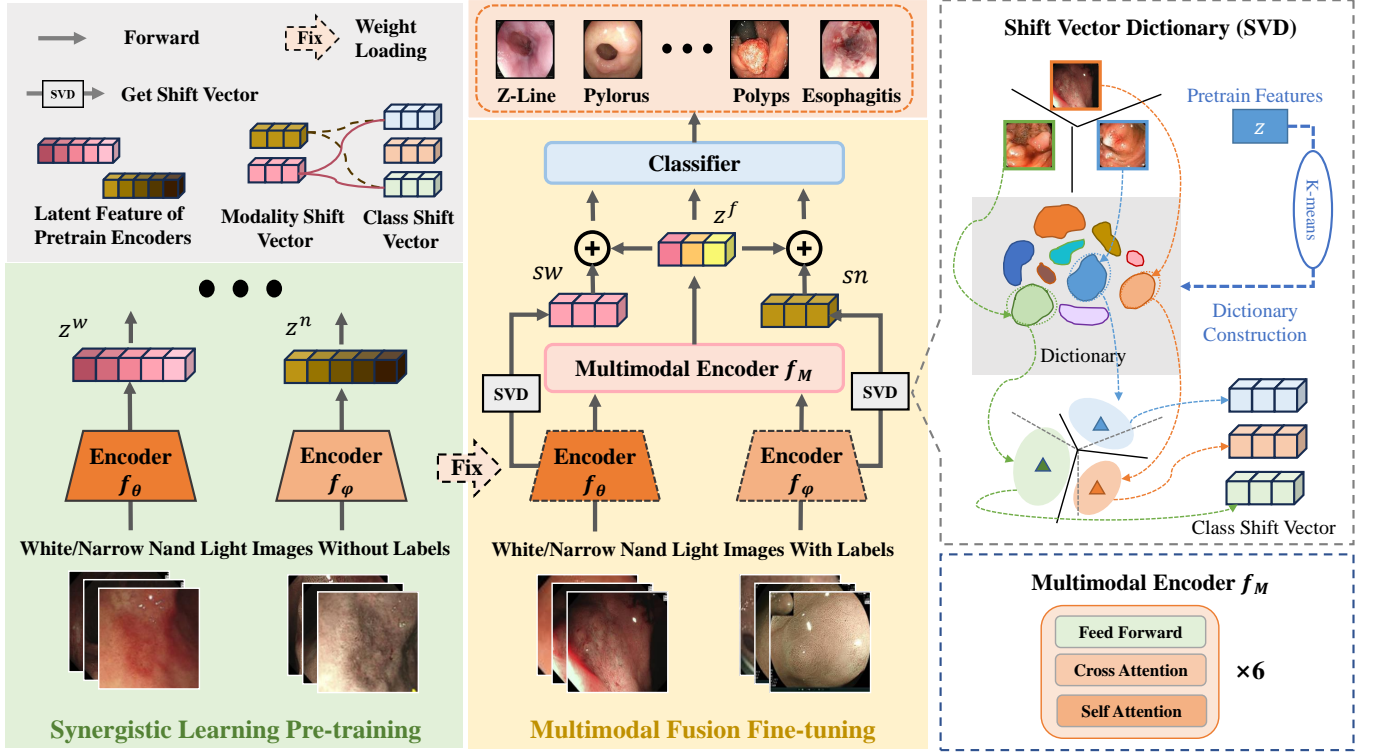


Fig. 4. The overview of proposed multimodal fusion fine-tuning framework. At the fine-tuning stage, the encoders load the weights from pre-training. A small number of labeled multimodal samples are encoded by f_θ and f_ϕ , and then fed into the multimodal encoder f_M to learn the fused representation. Subsequently, the multimodal features z^f are perturbed based on the shift vectors generated from the shift vector dictionary.

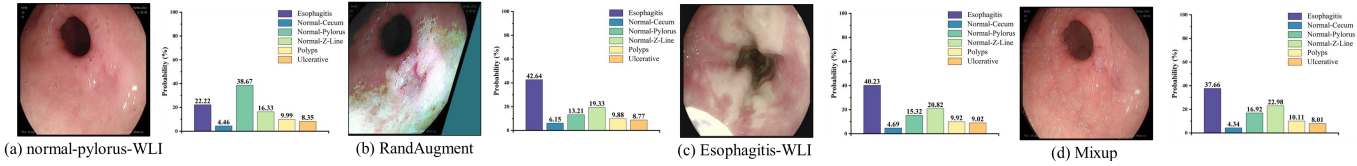


Fig. 5. The sample of white light image. (a) Normal-pylorus sample. (b) The sample (a) after randaugment. (c) Esophagitis sample. (d) The sample (a) mixes with Polyps using a mixup.

feature representations and recognize subtle lesions. Specifically, inspired by the research [29], we set a global embedding to project z^w and z^n into high-dimensional latent feature space vectors Z^w and Z^n . Aligned learning optimizes the InfoNCE [30] loss as the objective through instance-level contrastive learning to align representations. Similar to Eq. 1, we compute the similarity between the embedding vectors of WLI and NBI as follows:

$$S(w_i, n_j) = \frac{\exp(\cosine(Z_i^w, Z_j^n)/\tau)}{\sum_{b=1}^B \exp(\cosine(Z_i^w, \tilde{Z}_b^n)/\tau)}, \quad (7)$$

$$S(n_i, w_j) = \frac{\exp(\cosine(Z_j^n, Z_i^w)/\tau)}{\sum_{b=1}^B \exp(\cosine(Z_i^n, \tilde{Z}_b^w)/\tau)},$$

where $\cosine(\cdot)$ denotes cosine similarity, τ is the temperature as in Eq. 1, (Z_i^w, Z_j^n) represents positive instance pairs, \tilde{Z}_b^w and \tilde{Z}_b^n represent negative instances. Thus, the optimization objective of aligned learning is defined as follows:

$$\mathcal{L}_a = \frac{1}{B} \sum_{i=1}^B (\mathcal{H}(y_i, S(w_i, n)) + \mathcal{H}(y_i, S(n_i, w))), \quad (8)$$

where $\mathcal{H}(\cdot)$ denotes the cross-entropy, and $y_i \in \mathbb{R}^B$ represents the one-hot labels of the instance-level samples.

4) **Loss Function in Pretrain Stage:** The overall optimization objective of MICS during the pre-training stage is summarized as follows:

$$L_{pre} = \alpha^{dis} L_{dis} + \alpha^{res} L_{res} + \alpha^a L_a, \quad (9)$$

where α^{dis} , α^{res} and α^a represent the weighting coefficients for the corresponding losses.

B. Shift Vector Dictionary

Previous explicit augmentation methods (e.g., mixup [32] and randaugment [8]) are prone to generate unreliable augmented samples for medical images. As shown in Fig. 5, we provide an example of a standard white-light image. In semi-supervised classification methods, the prediction of weakly augmented (a) is typically used as a pseudo-label, forcing the model to predict (b) as the same class. The ambiguous consistency between (a) and (b) affects the model's prediction for (c). The augmented samples with randaugment intuitively

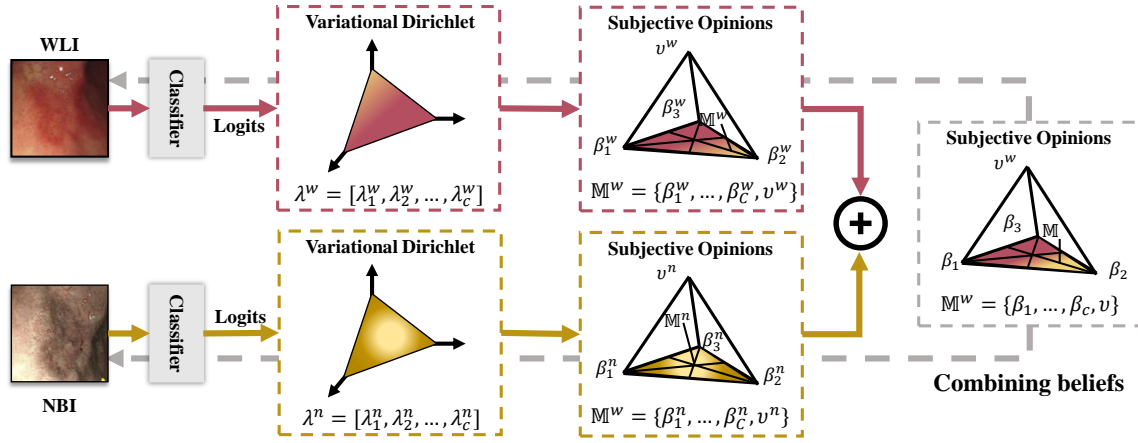


Fig. 6. Overview of Dynamic Evidential Fusion [31].

resemble real Esophagitis more closely, even if a well-trained ViT-B/16 model misclassifies them. Besides, the augmented sample of mixup from Normal-Pylorus and Polyps is incorrectly identified as Esophagitis.

Therefore, we consider designing an implicit augmentation scheme that requires no prior knowledge. On the one hand, it avoids introducing potentially ambiguous semantic information. On the other hand, it transfers transferable knowledge from existing data. Inspired by [33], we propose an implicit feature augmentation scheme based on the distribution shift of the base class dictionary. The difference between the proposed method and [33] is that the former is suitable for multimodal tasks, where the sample distribution of the base class dictionary generates shift components to guide the fusion features toward the initial modality, while the latter seeks the nearest set of samples within a single-modal dictionary's clustering to join the training process.

Specifically, we fully leverage the pre-trained encoders f_θ and f_φ . Since the intermediate layers of the encoders operate independently during the pretraining stage, f_θ and f_φ retain the ability to map the original modality distribution. First, the encoders extract features from unlabeled samples:

$$\begin{aligned} z_w^* &= f_\theta(w_q), q = \{1, 2, \dots, Q\}, \\ z_n^* &= f_\varphi(n_q), q = \{1, 2, \dots, Q\}, \end{aligned} \quad (10)$$

where Q represents the total number of samples in the training set. Then, the extracted features are clustered into C clusters by K-means:

$$\min \sum_{i=1}^Q \sum_{j=1}^C \mathbf{1}\{c_i = j\} \|z_i^* - \mu_j\|^2, \quad (11)$$

where $c = \{c_1, c_2, \dots, c_Q\}$ represents the cluster labels and $c_i \in \{1, 2, \dots, C\}$. C is the number of classification categories, μ_j is the prototype vector for each cluster, and $\mathbf{1}\{c_i = j\}$ is the indicator function. Then, for each cluster $j \in \{1, 2, \dots, C\}$, the sample mean is calculated to update the prototype μ_j as:

$$\mu_j = \frac{1}{Q_j} \sum_{i:c_i=j} z_i^*, \quad (12)$$

where Q_j is the number of samples in cluster j . Therefore, the sample covariance matrix for cluster j can be represented as:

$$\Sigma_j = \frac{1}{Q_j - 1} \sum_{i:c_i=j} (z_i^* - \mu_j)(z_i^* - \mu_j)^T. \quad (13)$$

Finally, we compute the shift vector dictionary based on the covariance matrix Σ_j . Sampling P shift vectors $\{s_1^j, s_2^j, \dots, s_P^j\}$ from a multivariate Gaussian distribution with mean μ_j and covariance Σ_j . The shift vector s_p^j is defined as:

$$s_p^j \sim \mathcal{N}(\mu_j, \Sigma_j), p = 1, 2, \dots, P, \quad (14)$$

The shift vector dictionary for all clusters is $SVD \in \mathbb{R}^{C \times P \times D} = \{\{s_p^j\}_{p=1}^P\}_{j=1}^C$, where D represents the dimensionality of the original input features.

C. Multimodal Fusion Fine-tuning

The multimodal fusion fine-tuning process is shown in Fig. 4. The fine-tuning stage includes the initial modality encoders f_θ and f_φ , provided by synergistic learning pretrain, and the multimodal encoder f_M . We fine-tune the model using a small amount of labeled data $L_X = \{l_w, l_n\}$. Specifically, f_θ and f_φ extract the corresponding modality features z^{l_w} and z^{l_n} , features are fused by f_M to obtain the multimodal fused feature z^f . Then, based on the initial modality Shift Vector Dictionary constructed in Section III.B, we randomly select the prototypes $\mu = \{\mu^w, \mu^n\}$ and their corresponding shift vectors s_w and s_n . Finally, z^f is element-wise added to the shift vectors s_w and s_n to obtain the knowledge transfer-based distribution-shifted features z^{wf} and z^{nf} . Therefore, the loss function for the proposed method during the multimodal fusion fine-tuning stage is defined as follows:

$$\begin{aligned} \mathcal{L}_f &= \mathcal{H}(y, cls(z^f)) + \mathcal{H}(y, cls(z^f + s_w)) \\ &\quad + \mathcal{H}(y, cls(z^f + s_n)), \end{aligned} \quad (15)$$

where y represents the class label, and cls denotes the classification head.

We aim to achieve information interaction between modalities and remove redundant features during the fine-tuning stage through the cross-attention mechanism of the multimodal

fusion encoder. Additionally, the proposed distribution-shift-based implicit data augmentation method mitigates the impact of ambiguous semantic information on model training and avoids introducing additional parameter computations. It expands the training data and alleviates the overfitting risk caused by the lack of labeled samples.

D. Uncertainty-Based Evidential Fusion

Inspired by multi-view classification, uncertainty-based evidential fusion can enhance the reliability of multimodal classification in addition to feature fusion. As shown in Fig. 6, we introduce a dynamic evidential fusion method for multimodal medical image classification called Trusted Multi-View Classification (TMC) [31]. We did not modify the theory of TMC but extended it to the gastroscopy multimodal classification domain, providing new insights for subsequent multimodal classification tasks.

Specifically, we obtain the C concentration parameters of the Dirichlet distribution for each modality sample in the fine-tuned model through TMC:

$$\lambda = [\lambda_1, \lambda_2, \dots, \lambda_c]. \quad (16)$$

Therefore, Dirichlet distribution is defined as $Dir(\sigma|\lambda)$, λ is the mean of the Dirichlet distribution, Dirichlet strength can be defined as $DS = \sum_{c=1}^C \lambda_c$. Then, belief mass β and uncertainty ν are defined as:

$$\begin{aligned} \beta_c &= \frac{\lambda_c - 1}{DS}, \\ \nu &= \frac{C}{DS}, \end{aligned} \quad (17)$$

where C is the number of classification categories. For samples from two modalities, TMC generates β_c^w and ν_c^w for WLI, and β_c^n and ν_c^n for NBI. Therefore, the probability mass assignments for different modalities are defined as:

$$\begin{aligned} \mathbb{M}^w &= \{\{\beta_c^w\}_{c=1}^C, \nu^w\}, \\ \mathbb{M}^n &= \{\{\beta_c^n\}_{c=1}^C, \nu^n\}. \end{aligned} \quad (18)$$

Given the quality of the fine-tuned model for each modality, these beliefs and uncertainties from different modalities are fused as:

$$\mathbb{M} = \mathbb{M}^w \oplus \mathbb{M}^n. \quad (19)$$

Although \mathbb{M} provides a decision-level fusion for the fine-tuned model, the external modality complementary optimization objectives at the representation layer can promote information interaction between modalities. Therefore, the final fine-tuning loss is defined as follows:

$$\mathcal{L} = \mathcal{L}_f + \mathcal{L}_{wn} + \mathcal{L}_{fuse}, \quad (20)$$

where \mathcal{L}_f is shown in Eq. 15. \mathcal{L}_{wn} is the fusion loss between white light samples and narrow-band light samples, defined as follows:

$$\begin{aligned} \mathcal{L}_{wn} &= \log p^w(y|\sigma^w) + \log p^n(y|\sigma^n) \\ &- \theta \cdot KL[Dir(\sigma^w, \lambda^w) || Dir(\sigma^w, [1, \dots, 1])] \\ &- \theta \cdot KL[Dir(\sigma^n, \lambda^n) || Dir(\sigma^n, [1, \dots, 1])], \end{aligned} \quad (21)$$

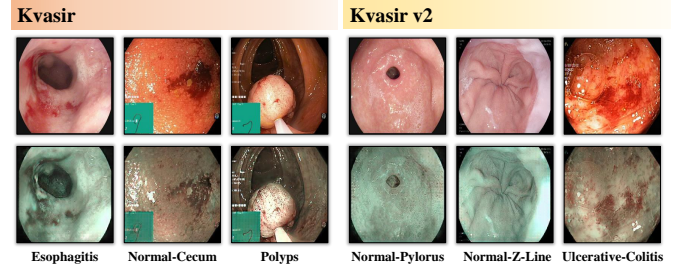


Fig. 7. Sample examples from the Kvasir and Kvasirv2 datasets. The top row shows the real white light images, and the bottom row shows the paired narrow-band light images generated by WtNGAN.

where p^w and p^n represent the predicted probabilities for white light and narrow-band light images, respectively. σ is the mean value of the Dirichlet Distribution for the corresponding modality, and KL denotes the Kullback-Leibler Divergence. Additionally, the fusion sample loss \mathcal{L}_{fuse} is defined as:

$$\begin{aligned} \mathcal{L}_{fuse} &= \log p(y|\sigma) \\ &- \theta \cdot KL[Dir(\sigma, \lambda) || Dir(\sigma, [1, \dots, 1])]. \end{aligned} \quad (22)$$

IV. EXPERIMENT

A. Implementation Details

At the pre-training stage, the initial learning rate is set to $1 \times e^{-4}$ and gradually decays to $1 \times e^{-5}$ with a cosine schedule. We use the AdamW [34] optimizer with a weight decay coefficient of 0.02 and a momentum parameter of 0.995, training for 100 epochs. At the fine-tuning stage, the parameter settings remain similar to the pre-training, except that the minimum learning rate $1 \times e^{-6}$ and the decay rate 0.999. We implemented all methods in Python 3.9.13 and PyTorch 1.12.1. A computer equipped with an Intel Xeon Silver 4208 CPU and seven NVIDIA GeForce RTX 3090 GPUs for train and testing.

TABLE I
THE DETAILED COMPOSITION OF KVASIR AND KVASIRV2.

	Categories	Kvasir	Kvasirv2
Normal	Z-line	500	1,000
	Pylorus	500	1,000
	Cecum	500	1,000
Abnormal	Esophagitis	500	1,000
	Polyps	500	1,000
	Ulcerative-Colitis	500	1,000
	Total	3,000	6,000

B. Datasets

On the one hand, to verify the practical significance of the generated paired modality images. On the other hand, it is limited by publicly available paired medical image multimodal datasets. We use the public gastroscopy white light image datasets Kvasir and Kvasirv2. WtNGAN generates the narrow-band light image paired with each image, and all comparison methods use the same multimodal data. The dataset example is shown in Fig. 7. The Kvasir dataset contains three normal

TABLE II

RESULTS OF COMPARATIVE EXPERIMENTS. SEMI-SUPERVISED METHODS ARE TRAINED IN A STANDARD WAY AND A PROPOSED MULTI-MODALITY WAY. **GREEN** INDICATES THAT MULTI-MODAL TRAINING IS IMPROVED COMPARED TO SINGLE-MODALITY, AND **RED** INDICATES A DECREASE. **BOLD** INDICATES THE BEST RESULT.

Method	Year	Modality		Kvasir				Kvasir v2			
		WLI	NBI	50% labels	30% labels	10% labels	5% labels	50% labels	30% labels	10% labels	5% labels
FixMatch	2020	✓		71.67	51.67	46.67	44.67	73.83	71.75	60.42	49.75
FlexMatch	2021	✓		73.00	67.33	52.33	54.67	73.00	74.67	63.50	56.41
Dash	2021	✓		70.33	55.33	44.50	40.17	75.17	71.75	62.42	46.08
FreeMatch	2023	✓		73.83	69.17	52.33	45.67	75.17	73.42	69.25	60.83
SoftMatch	2023	✓		73.33	69.83	62.67	54.33	73.00	71.50	66.33	60.41
HABIT	2023	✓		73.17	69.83	61.00	54.83	75.50	73.08	68.91	65.00
SIABC	2025	✓		73.33	69.83	62.50	56.50	78.00	75.58	70.25	61.83
FixMatch	2020	✓	✓	70.50 (-1.17)	66.83 (+15.16)	58.17 (+11.50)	57.00 (+12.33)	74.75 (+0.92)	71.50 (-0.25)	64.33 (+3.91)	69.50 (+19.75)
FlexMatch	2021	✓	✓	72.00 (+1.00)	70.83 (+3.50)	66.67 (+14.34)	57.50 (+2.83)	75.50 (+2.50)	74.08 (-0.59)	69.67 (+6.17)	64.33 (+7.92)
Dash	2021	✓	✓	71.67 (+1.34)	71.33 (+16.00)	63.83 (+19.33)	58.67 (+18.50)	72.58 (-2.59)	70.33 (-1.42)	69.67 (+7.25)	62.75 (+16.67)
FreeMatch	2023	✓	✓	75.50 (+1.67)	72.83 (+3.66)	64.83 (+12.50)	58.67 (+13.00)	77.17 (+2.00)	73.08 (-0.34)	68.58 (-0.67)	63.75 (+2.92)
SoftMatch	2023	✓	✓	71.17 (-2.16)	69.00 (-0.83)	63.83 (+1.16)	51.33 (-3.00)	75.92 (+2.92)	72.83 (+1.33)	68.75 (+2.42)	63.58 (+3.17)
HABIT	2023	✓	✓	74.33(+ 1.16)	71.33 (+1.50)	61.17 (+0.17)	52.33 (-2.50)	76.75 (+1.25)	72.67 (-0.41)	69.25 (+0.34)	64.83 (-0.17)
SIABC	2025	✓	✓	72.17 (-1.16)	66.83 (-3.00)	64.00 (+1.50)	54.67 (-1.83)	74.42 (-3.58)	69.75 (-5.83)	69.67 (-0.58)	64.42 (+2.59)
MICS	Ours	✓	✓	76.67	75.33	67.17	60.88	78.67	77.92	74.33	70.25

gastric anatomical structure types and three common gastric disease images. We unify the image resolution to 256×256 and divide the training set, validation set, and test set into a ratio of 6:2:2. The detailed composition of the dataset used is shown in Table I.

C. Baseline Methods

We compare with the state-of-the-art semi-supervised classification methods trained with white light images only and with images from both modalities. The comparative methods are as follows:

- FixMatch [13] applies consistency regularization to pseudo-labeling methods and sets a fixed confidence threshold to predict the consistency of different augmented samples.
- FlexMatch [14] dynamically adjusts the pseudo-label confidence threshold of each class of samples according to the model's learning state.
- Dash [35] selects a subset of training samples to filter out erroneous pseudo-label samples based on the dynamic adjustment of the threshold.
- FreeMatch [15] sets a lower pseudo-label threshold in the early stage of training to accelerate model convergence and increases the threshold in the later stage of training to eliminate erroneous pseudo-label predictions.
- SoftMatch [36] focuses on the quality-quantity trade-off of pseudo-label samples, maintaining numerous high-quality pseudo-labels during training to effectively utilize unlabeled samples.
- HABIT [37] proposes a consistency-aware momentum heredity to alleviate bias in pseudo-label sample selection.
- SIABC [38] proposes a super augmentation block and a cross-set augmentation module so that model optimization benefits from samples that have been well-learned.

D. Quantitative and Qualitative Experimental Results

Quantitative results are shown in Table II. We conduct two sets of experiments under four label ratios. The results demonstrate that replacing strongly augmented images with weakly augmented ones from another modality effectively enhances semi-supervised classification performance, validating our hypothesis in Section III.B. Notably, FixMatch with multimodal training significantly improves accuracy with only 5% labeled data, likely because it relies on abundant unlabeled samples for optimization, avoids erroneous feature representations from excessive strong augmentations, and filters out incorrect pseudo-labels with high thresholds. Other threshold-based methods show accuracy improvements ranging from 0.17% to 18.50%, while composite augmentation methods experience accuracy drops of 0.17% to 5.83% at specific label ratios. Among all results, the proposed MICS leverages multimodal data, achieving the best classification outcomes, with an accuracy of 70.25% on the Kvasir v2 dataset using only 5% labeled data.

The qualitative experimental results are shown in Fig. 8. We provide the UMAP feature visualization results of comparative methods trained with only 5% labeled data on the Kvasir v2 dataset. The experimental results indicate that almost all classification methods exhibit confusion between Normal-Cecum, Ulcerative, and Polyps. On the one hand, the cecum is part of the colon, and ulcers and polyps are more likely to appear in the colon region. On the other hand, with only a small amount of labeled data provided, these methods fail to focus on the lesions' characteristics and instead memorize the scenarios where the lesions appear, which is easier than the former. Notably, in the feature visualization results of the proposed method, normal cecum samples are clearly separated from ulcerative or polyps in the feature space, indicating that MICS distinguishes well between lesion locations and the lesions. MICS, SIABC, and Dash also differentiate between the normal pyloric structure and the Z-Line. The structural

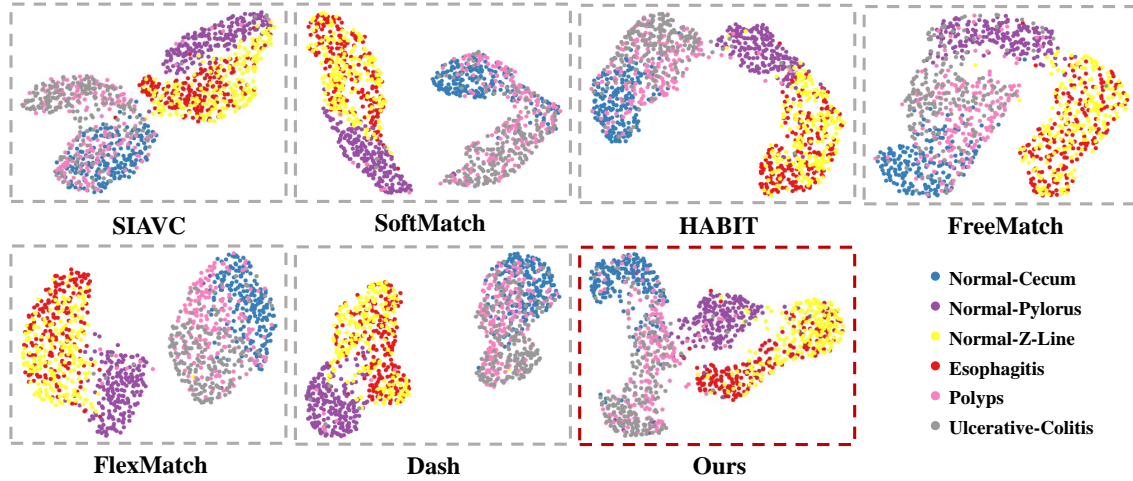


Fig. 8. The UMAP (n_components=2, n_neighbors=30, min_dist=1.0) feature visualization of each method on the Kvasir v2 dataset with only 5% labeled data provided.

TABLE III

ABLATION EXPERIMENT OF THE FINAL FINE-TUNED MODEL WITH WEIGHTS OF EACH PRE-TRAINED COMPONENT PROVIDED.

Train Way	Components			Kvasir	Kvasirv2
	Consistency Learning	Reconstructive Learning	Aligned Learning	Top-1 Acc	Top-1 Acc
Pretrain	✓			71.33%	72.50%
	✓	✓		73.33%	75.17%
	✓	✓	✓	77.04%	78.67%
Vanilla				70.33%	70.92%

TABLE IV

ABLATION STUDIES OF VARIOUS FINE-TUNED MODEL COMPONENTS USING THE COMPLETE PRE-TRAINED WEIGHTS.

Train Way	Components				Kvasir	Kvasir v2
	Encoder WLI	Encoder NBI	SVD	TMC	Top-1 Acc	Top-1 Acc
Fine-tuning	✓				70.33%	74.08%
	✓	✓			73.33%	76.91%
	✓	✓	✓		76.17%	78.00%
	✓	✓	✓	✓	77.04%	78.67%

similarity between the pylorus and the z-line is high, and sensitivity to the hole-like black structure may cause the model to confuse the two, resulting in low differentiation in the feature space. Finally, when labeled data is scarce, all methods fail to distinguish between Normal-Z-Line and Esophagitis. Anatomically, the Z-Line is located at the junction of the stomach and esophagus, and Esophagitis is one of the most common pathological manifestations of the Z-Line. Therefore, when Esophagitis images show low-grade inflammation or low contrast, models trained with limited labels struggle to distinguish between normal esophagus and mild esophageal inflammation.

E. Ablation Study

To demonstrate the contribution of each component to multimodal classification, we conduct ablation studies on the proposed method with 50% of the labels. The ablation experiments for MICS are shown in Table III and Table IV. Table III demonstrates the classification results of the final fine-tuned model with weights from different pre-trained components. The experimental results indicate that the synergistic learning pretraining (Pretrain) improves the classification accuracy by 6.71% and 7.75% on two gastroscopy datasets, respectively, compared to the model with randomly initialized weights (Vanilla). Table IV shows the ablation results of each component during the fine-tuning phase. Compared to the single-modality WLI, introducing the NBI encoder for multimodal fusion improves the performance by 3.00% and 2.08% on the two gastroscopy datasets, respectively. SVD provides more generalized training samples based on the distributions of different modalities and contributes to a 2.84% accuracy improvement on the relatively more minor Kvasir dataset. Finally, the TMC based on Dynamic Evidential fusion further enhances the classification capability of MICS by leveraging uncertainty estimation.

V. CONCLUSION

In this paper, we rethink the role of multimodal images in semi-supervised learning based on consistency regularization and propose MICS, the first semi-supervised medical image classification network to our knowledge that exploits multimodal consistency. Specifically, we use real gastroscopic WLI images and paired NBI images generated by the algorithm to classify gastroscopic diseases and anatomical structures. MICS separates the learning process of labeled data from unlabeled data, avoiding the high-confidence misprediction of unlabeled samples affecting the learning of labeled samples. MICS implements adversarial alignment pre-training of unlabeled samples and distribution shift with uncertainty fusion of a small number of labeled samples in the framework of "pre-training + fine-tuning". The proposed MICS shows promising

recognition ability on gastroscopic multimodal classification datasets, especially when only small labels are provided. In the future, we will verify the superiority of MICS on more types of medical image tasks and generate more reliable multimodal data to further improve the classification accuracy.

REFERENCES

- [1] S. Xu, W. Li, Z. Li, T. Zhao, and B. Zhang, "Facing differences of similarity: Intra-and inter-correlation unsupervised learning for chest x-ray anomaly detection," *IEEE Transactions on Medical Imaging*, 2024.
- [2] Y. Ma, W. Cui, J. Liu, Y. Guo, H. Chen, and Y. Li, "A multi-graph cross-attention-based region-aware feature fusion network using multi-template for brain disorder diagnosis," *IEEE Transactions on Medical Imaging*, vol. 43, no. 3, pp. 1045–1059, 2023.
- [3] J. Liu, W. Cui, Y. Chen, Y. Ma, Q. Dong, R. Cai, Y. Li, and B. Hu, "Deep fusion of multi-template using spatio-temporal weighted multi-hypergraph convolutional networks for brain disease analysis," *IEEE Transactions on Medical Imaging*, vol. 43, no. 2, pp. 860–873, 2023.
- [4] Y. Li, Y. Zhang, W. Cui, B. Lei, X. Kuang, and T. Zhang, "Dual encoder-based dynamic-channel graph convolutional network with edge enhancement for retinal vessel segmentation," *IEEE Transactions on Medical Imaging*, vol. 41, no. 8, pp. 1975–1989, 2022.
- [5] W. Tang, F. He, Y. Liu, and Y. Duan, "Matr: Multimodal medical image fusion via multiscale adaptive transformer," *IEEE Transactions on Image Processing*, vol. 31, pp. 5134–5149, 2022.
- [6] S. H. Yun and S. J. Kwok, "Light in diagnosis, therapy and surgery," *Nature Biomedical Engineering*, vol. 1, no. 1, p. 0008, 2017.
- [7] D. Lahat, T. Adali, and C. Jutten, "Multimodal data fusion: an overview of methods, challenges, and prospects," *Proceedings of the IEEE*, vol. 103, no. 9, pp. 1449–1477, 2015.
- [8] E. D. Cubuk, B. Zoph, J. Shlens, and Q. V. Le, "RandAugment: Practical automated data augmentation with a reduced search space," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, 2020, pp. 702–703.
- [9] D. Berthelot, N. Carlini, E. D. Cubuk, A. Kurakin, K. Sohn, H. Zhang, and C. Raffel, "Remixmatch: Semi-supervised learning with distribution matching and augmentation anchoring," in *International Conference on Learning Representations*, 2020. [Online]. Available: <https://openreview.net/forum?id=HkLkeR4KPB>
- [10] Z. Li, Q. Lin, J. Wu, T. Lai, R. Wu, and D. Zhang, "Leukocyte classification using relative-relationship-guided contrastive learning," *Expert Systems with Applications*, vol. 260, p. 125390, 2025.
- [11] M. Zheng, S. You, L. Huang, F. Wang, C. Qian, and C. Xu, "Simmatch: Semi-supervised learning with similarity matching," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 14471–14481.
- [12] Q. Lin, Z. Li, K. Zeng, J. Wen, Y. Jiang, and J. Chen, "Wtngan: Unpaired image translation from white light images to narrow-band images," *Pattern Recognition*, p. 111431, 2025.
- [13] K. Sohn, D. Berthelot, N. Carlini, Z. Zhang, H. Zhang, C. A. Raffel, E. D. Cubuk, A. Kurakin, and C.-L. Li, "Fixmatch: Simplifying semi-supervised learning with consistency and confidence," *Advances in Neural Information Processing Systems*, vol. 33, pp. 596–608, 2020.
- [14] B. Zhang, Y. Wang, W. Hou, H. Wu, J. Wang, M. Okumura, and T. Shinozaki, "Flexmatch: Boosting semi-supervised learning with curriculum pseudo labeling," *Advances in Neural Information Processing Systems*, vol. 34, pp. 18408–18419, 2021.
- [15] Y. Wang, H. Chen, Q. Heng, W. Hou, Y. Fan, Z. Wu, J. Wang, M. Savvides, T. Shinozaki, B. Raj, B. Schiele, and X. Xie, "Freematch: Self-adaptive thresholding for semi-supervised learning," in *The Eleventh International Conference on Learning Representations*, 2023. [Online]. Available: <https://openreview.net/forum?id=PDRUPTXJLA>
- [16] Q. Zeng, Y. Xie, Z. Lu, and Y. Xia, "Pefat: Boosting semi-supervised medical image classification via pseudo-loss estimation and feature adversarial training," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 15 671–15 680.
- [17] J. Li, R. Selvaraju, A. Gormare, S. Joty, C. Xiong, and S. C. H. Hoi, "Align before fuse: Vision and language representation learning with momentum distillation," *Advances in Neural Information Processing Systems*, vol. 34, pp. 9694–9705, 2021.
- [18] M. Mallya and G. Hamarneh, "Deep multimodal guidance for medical image classification," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer, 2022, pp. 298–308.
- [19] P. Tang, X. Yan, Y. Nan, S. Xiang, S. Krammer, and T. Lasser, "Fusionm4net: A multi-stage multi-modal learning algorithm for multi-label skin lesion classification," *Medical Image Analysis*, vol. 76, p. 102307, 2022.
- [20] T. Jin, X. Xie, R. Wan, Q. Li, and Y. Wang, "Gene-induced multimodal pre-training for image-omic classification," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer, 2023, pp. 508–517.
- [21] S. Mohla, S. Pande, B. Banerjee, and S. Chaudhuri, "Fusatnet: Dual attention based spectrospatial multimodal fusion network for hyperspectral and lidar classification," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, 2020, pp. 92–93.
- [22] X. He, Y. Deng, L. Fang, and Q. Peng, "Multi-modal retinal image classification with modality-specific attention network," *IEEE Transactions on Medical Imaging*, vol. 40, no. 6, pp. 1591–1602, 2021.
- [23] R. Liu, J. Liu, Z. Jiang, X. Fan, and Z. Luo, "A bilevel integrated model with data-driven layer ensemble for multi-modality image fusion," *IEEE Transactions on Image Processing*, vol. 30, pp. 1261–1274, 2020.
- [24] W. Tang, F. He, and Y. Liu, "Itfuse: An interactive transformer for infrared and visible image fusion," *Pattern Recognition*, vol. 156, p. 110822, 2024.
- [25] D. Rao, T. Xu, and X.-J. Wu, "Tgfuse: An infrared and visible image fusion approach based on transformer and generative adversarial network," *IEEE Transactions on Image Processing*, 2023.
- [26] K. He, H. Fan, Y. Wu, S. Xie, and R. Girshick, "Momentum contrast for unsupervised visual representation learning," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 9729–9738.
- [27] K. He, X. Chen, S. Xie, Y. Li, P. Dollár, and R. Girshick, "Masked autoencoders are scalable vision learners," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 16 000–16 009.
- [28] N. Vesdapunt, K. K. Fu, Y. Wu, X. Zhang, and P. Natarajan, "Hvclip: High-dimensional vector in clip for unsupervised domain adaptation," in *European Conference on Computer Vision*, 2024, pp. 1–18.
- [29] Q. Li, X. Yan, J. Xu, R. Yuan, Y. Zhang, R. Feng, Q. Shen, X. Zhang, and S. Wang, "Anatomical structure-guided medical vision-language pre-training," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer, 2024, pp. 80–90.
- [30] A. v. d. Oord, Y. Li, and O. Vinyals, "Representation learning with contrastive predictive coding," *arXiv preprint arXiv:1807.03748*, 2018.
- [31] Z. Han, C. Zhang, H. Fu, and J. T. Zhou, "Trusted multi-view classification with dynamic evidential fusion," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 45, no. 2, pp. 2551–2566, 2022.
- [32] H. Zhang, M. Cisse, Y. N. Dauphin, and D. Lopez-Paz, "mixup: Beyond empirical risk minimization," in *International Conference on Learning Representations*, 2018. [Online]. Available: <https://openreview.net/forum?id=r1Ddp1-Rb>
- [33] J. Yang, H. Chen, J. Yan, X. Chen, and J. Yao, "Towards better understanding and better generalization of low-shot classification in histology images with contrastive learning," in *International Conference on Learning Representations*, 2022. [Online]. Available: <https://openreview.net/forum?id=kQ2SOfHIOVC>
- [34] I. Loshchilov and F. Hutter, "Decoupled weight decay regularization," in *International Conference on Learning Representations*, 2019. [Online]. Available: <https://openreview.net/forum?id=Bkg6RiCqY7>
- [35] Y. Xu, L. Shang, J. Ye, Q. Qian, Y.-F. Li, B. Sun, H. Li, and R. Jin, "Dash: Semi-supervised learning with dynamic thresholding," in *International Conference on Machine Learning*, 2021, pp. 11 525–11 536.
- [36] H. Chen, R. Tao, Y. Fan, Y. Wang, J. Wang, B. Schiele, X. Xie, B. Raj, and M. Savvides, "Softmatch: Addressing the quantity-quality tradeoff in semi-supervised learning," in *The Eleventh International Conference on Learning Representations*, 2023. [Online]. Available: <https://openreview.net/forum?id=ymlzQXBDIF>
- [37] Q. Yang, Z. Chen, and Y. Yuan, "Hierarchical bias mitigation for semi-supervised medical image classification," *IEEE Transactions on Medical Imaging*, vol. 42, no. 8, pp. 2200–2210, 2023.
- [38] Z. Li, Q. Lin, H. Fan, T. Zhao, and D. Zhang, "Siavc: Semi-supervised framework for industrial accident video classification," *IEEE Transactions on Circuits and Systems for Video Technology*, pp. 1–1, 2024.