Information Fusion 113 (2025) 102631

Contents lists available at ScienceDirect

Information Fusion

journal homepage: www.elsevier.com/locate/inffus

Adapting the segment anything model for multi-modal retinal anomaly detection and localization

Jingtao Li [a,c,1], Ting Chen [a,c,1], Xinyu Wang [b,c,*], Yanfei Zhong [a,c], Xuan Xiao [a,c,*]

[a] State Key Laboratory of Information Engineering in Surveying, Mapping and Remote Sensing & Department of Ophthalmology, Renmin Hospital of Wuhan University, Wuhan University, PR China
[b] School of Remote Sensing and Information Engineering, Wuhan University, PR China
[c] Cross Research Institute of Medical Remote Sensing, Wuhan University, PR China

ARTICLE INFO

Keywords:
Anomaly detection
Anomaly localization
Retinal artery occlusion
Visual foundation model
Multi-modal medical diagnosis
Segment anything model

ABSTRACT

The fusion of optical coherence tomography (OCT) and fundus modality information can provide a comprehensive diagnosis for retinal artery occlusion (RAO) disease, where OCT provides the cross-sectional examination of the fundus image. Given multi-modal retinal images, an anomaly diagnosis model can discriminate RAO without the need for real diseased samples. Despite this, previous studies have only focused on single-modal diagnosis, because of: 1) the lack of paired modality samples; and 2) the significant imaging differences, which make the fusion difficult with small-scale medical data. In this paper, we describe how we first built a multi-modal RAO dataset including both OCT and fundus modalities, which supports both the anomaly detection and localization tasks with pixel-level annotation. Motivated by the powerful generalization ability of the recent visual foundation model known as the Segment Anything Model (SAM), we adapted it for our task considering the small-scale property of retinal samples. Specifically, a modality-shared decoder with task-specific tokens is introduced to make SAM support the multi-modal image setting, which includes a mask token for the anomaly localization task at the pixel level and a fusion token for the anomaly detection task at the case level. Since SAM has little medical knowledge and lacks the learning of the "normal" concept, it is infeasible to localize RAO anomalies in the zero-shot manner. To integrate expert retinal knowledge while keeping the general segmentation knowledge, general anomaly simulation for both modalities and a low-level prompt-tuning strategy are introduced. The experiments conducted in this study show that the adapted model can surpass the state-of-the-art model by a large margin. This study sets the first benchmark for the multi-modal anomaly detection and localization tasks in the medical community. The code is available at https://github.com/Jingtao-Li-CVer/MMRAD.

1. Introduction

Retinal disease is not only an indicator of a visual health problem, but also has a correlation with hypertension and heart and brain diseases [1]. Detecting the three most important causes of blindness (i.e., macular degeneration, glaucoma, and diabetic retinopathy (DR) [2]) at an early stage can effectively avoid the loss of vision in many cases [3,4]. Especially with an aging population, the number of patients suffering from chorioretinal diseases such as age-related macular degeneration (AMD) can be expected to increase, and more precise retinal diagnosis is required [5]. Furthermore, certain cardiovascular diseases (e.g., coronary heart disease) [6] and brain diseases (e.g., Alzheimer's disease) [7] can also be observed through an eye examination with a high diagnosis accuracy.

Retinal fundus and optical coherence tomography (OCT) images are the most important modalities, and their fusion has proven useful when evaluating retinal pathology in supervised communities [8,9]. A fundus image is captured with contrast filters (red, blue, and green filters) in a noninvasive manner, providing a two-dimensional (2-D) representation of the retinal surface. In contrast, OCT provides the three-dimensional (3-D) structural information, showing the cross-sectional lesion of the fundus image. Due to the different imaging mechanisms, the use of only a single modality is insufficient to spot all the potential diseases. For example, dry AMD at the early stage can only be observed in the OCT

* Corresponding authors.
E-mail addresses: wangxinyu@whu.edu.cn (X. Wang), xiaoxuan1111@whu.edu.cn (X. Xiao).
[1] The authors contributed equally.

https://doi.org/10.1016/j.inffus.2024.102631

Available online 8 August 2024

modality, while some typical DR alterations are more observable in the fundus modality. To date, the multi-modal diagnosis works have mostly concentrated on supervised learning. For example, Yoo et al. [8] combined OCT and fundus images for improving the accuracy of AMD diagnosis. To reduce the requirement for strict paired data, Wang et.al. introduced the loose pair training and designed a two-stream multi--modal CNN (MM-CNN) to classify three classes (normal, dryAMD and wetAMD) [10]. Multi-Modal Multiple Instance Learning (MM-MIL) improved the MM-CNN by designing instance learning module and increased the model interpretation [11].

In contrast with the supervised methods, medical anomaly detection and localization (MADL) can provide an initial diagnosis, and can thus assist the related experts in an unsupervised manner [10]. A MADL model is always trained with only healthy samples and gives each test sample a continuous anomaly score [11]. Considering the high cost of retinal samples, the MADL model is more practical than the supervised models, without the need for expensive expert annotation [12,13]. Furthermore, the MADL model is not limited to preset disease categories, and is thus more flexible for unknown retinal diseases [14,15].

Despite all of this, the current MADL methods only focus on single-modal retinal data because of the lack of paired modality samples, and the significant imaging differences make the fusion difficult with small-scale medical data. Researchers on MADL have considered many individual modalities [16], including retinal fundus images [13], OCT images [17], chest X-ray images [18], computed tomography (CT) images of the brain [19], and brain functional magnetic resonance imaging (fMRI) [20]. To the best of our knowledge, no prior research has attempted multi-modal MADL, although the effect of medical multi-modal fusion has been validated in supervised communities (where the used datasets are not publicly available) [8,21]. To this end, we first established a cooperation with Renmin Hospital of Wuhan University in China, and then built the first multi-modal dataset for MADL research. Focusing on retinal artery occlusion (RAO) diagnosis, we collected 370 paired samples, made up of fundus image and OCT image modalities. RAO is a rare but visually threatening ophthalmic emergency and associated with cardiovascular diseases closely. The final constructed **m**ulti-**m**odal **r**etinal (MMR) dataset contains 266 paired images of healthy people and 104 paired images of patients with RAO disease. Each patient image has been labeled at the pixel level (nine symptom classes in the fundus image modality and eight symptom classes in the OCT image modality). The constructed MMR dataset supports both multi-modal anomaly detection at the case level and the localization task at the pixel level.

To make the MADL model suitable for small-scale multi-modal data, we chose to fine-tune the recent Segment Anything Model (SAM) [22], considering its powerful zero-shot generalization ability [23]. Pre-trained MADL models [24,16] have shown promising performances, compared to models trained from scratch, such as the reconstruction-based models [25,26]. As a visual foundation model, SAM has been trained on the 1 billion SA-1B dataset and has a powerful segmentation ability under the open-set setting [22]. However, since SAM has little medical knowledge and lacks the learning of the "normal" concept, it is infeasible to complete the retinal anomaly diagnosis task in a zero-shot manner. To integrate expert retinal knowledge, we decided to fine-tune SAM for the multi-modal MADL task.

There are two barriers when fine-tuning SAM for the MADL task. Firstly, SAM does not support the multi-modal anomaly detection task, and the fine-tuning process requires labeled samples. Labeled samples are difficult to access and contradictory with the anomaly-free training setting of the anomaly detection task [27]. To adapt SAM for supporting both the anomaly detection and localization tasks, task-specific tokens are introduced for the decoder, i.e., a mask token for the anomaly localization task and a fusion token for the anomaly detection task. The task-specific tokens are input into the modality-shared decoder with image embeddings and trained to contain the task-specific information. To fine-tune the model in a pseudo-supervised manner while keeping the SAM encoder frozen, we propose a general sample simulation strategy for both the fundus and OCT anomalies, using only healthy images and the data augmentation technique. The transformed data act as the anomalies during the training stage and have shown a promising generalization ability for real anomalies at the test stage. With the simulated samples, a prompt-tuning paradigm is further introduced, which keeps the heavy backbone weights frozen and generates the learned low-level prompts for each encoder layer. The prompt-tuning method reduces the training burden with limited samples, and the low-level prompts are consistent with the semantically independent property of the MADL task [28]. We name the adapted model the **m**ulti-**m**odal model for **R**AO **a**nomaly **d**iagnosis (MMRAD), which accepts the fundus image and OCT image as input, and outputs two single-modal anomaly localization maps and the multi-modal anomaly detection score for each case.

The main contributions of this paper can be summarized as follows.

1) We built the first multi-modal dataset for MADL research, which focuses on RAO diagnosis, including both the fundus image modality and OCT image modality. The dataset is labeled at the pixel level, supporting both the multi-modal anomaly detection and localization tasks.
2) Based on the recent visual foundation model, i.e., SAM, the multi-modal model for RAO anomaly diagnosis (MMRAD) is proposed. A general anomaly simulation strategy for both modalities and a low-level prompt-tuning strategy are introduced to fine-tune the model while keeping the SAM encoder frozen.
3) A modality-shared decoder with task-specific tokens is proposed, which includes a mask token for the anomaly localization task and a fusion token for the anomaly detection task. With these tokens, MMRAD can output two single-modal anomaly localization maps and the multi-modal anomaly detection score for each case.

The rest of this paper is organized as follows. Section 2 gives a brief review of the related medical anomaly diagnosis work and the visual foundation model. Section 3 introduces the constructed MMR dataset for RAO diagnosis. Section 4 presents the proposed MMRAD model and the fine-tuning strategies. The experimental results and analysis are given in Section 5. Finally, the paper is concluded in Section 6.

2. Related work

2.1. Medical anomaly detection and localization for a single modality

Motivated by reducing the need for labeled data, the retinal anomaly detection and localization tasks have been actively researched, which are aimed at finding the anomaly patterns deviating from the healthy patterns at the image level or pixel level [14]. The current related methods can be roughly divided into autoencoder (AE)-based methods, generative adversarial network (GAN)-based methods, and pretraining-based methods. The AE-based methods assume that the normal samples can be more easily reconstructed than anomalies, and the reconstruction error indicates the anomaly score [16,25,29]. The traditional AE-based methods apply convolutional AEs directly [29]. Some of the more recent AE-based methods focus on adding more learning constraints [25] or using an advanced architecture (e.g., transformer model) [27] for better normal representations. Similarly, the GAN-based methods assume that the normal samples can be more easily regenerated, and the generation error is treated as the anomaly score [10,17]. AnoGAN is a classical method which maps the test image from image space back to latent space and assumes that the regenerated image will only be constructed well for normal OCT images [10]. To speed up the inference process, f-AnoGAN was proposed by training an encoder to map the image to latent space automatically [17]. The introduction of a discriminator can make the generated image more realistic than only using the mean squared error (MSE) loss. Despite this,

the assumption of both the AE and GAN methods is not met in general, and the anomaly regions may be also reconstructed [27].

In contrast to learning the normal representations from scratch, the pretraining-based methods use a pretext task to pre-train the model [24, 16] or use an ImageNet-pretrained model to extract the features [13]. A significant improvement can be observed, compared to training from scratch [24]. To recognize the subclasses of diseased images, the multi-class strong augmentation via contrastive learning (MSACL) method was further proposed, which extends the contrastive learning into the four-class setting with pretext training [16]. However, these pretraining-based methods need relatively large-scale medical images, which is not feasible in our multi-modal case. Without the additional pretraining stage, some methods use the ImageNet-pretrained weights directly to extract the features and combine this with a one-class classifier [30]. For example, Ouardini et al. [13] concatenated the pretrained features and used them to train an isolation forest based anomaly detection model. However, the ImageNet-pretrained features may not be suitable for medical modalities [30,31], and the current medical anomaly detection research only focuses on a single modality.

2.2. Visual foundation model

A visual foundation model refers to an emerging paradigm where the model is trained on large-scale data and can be adapted to a wide range of downstream tasks [32]. Inspired by the success in the natural language processing (NLP) community, such as the bidirectional encoder representations from transformers (BERT) model [33] and the generative pretraining (GPT) series of models [34], many visual foundation models have been developed based on an advanced transformer architecture, which is compatible with the language and visual signals and can learn the dependency relationship over a long range. Representative models include contrastive language-image pretraining (CLIP) [35], grounded language-image pretraining (GLIP) [36], Grounding DINO [37], and SAM [22], where SAM is the only visual foundation model at the pixel level, and the other models are mostly at the image level or instance level. For example, CLIP is trained on large-scale network image-text pairs, focusing on the image-level recognition task, without explicit modeling of spatial relationships, and has a larger gap than SAM with the dense anomaly localization task. Differing from the classical ImageNet-pretrained model, SAM is powered by 11 million images and has an excellent zero-shot generalization ability for unseen scenes [22, 38]. After fine-tuning the visual foundation model on the small-scale downstream images, an obvious improvement can be observed across a wide range of tasks [32,39]. Currently, although SAM has already been successfully applied to certain supervised medical segmentation tasks [40], no prior study has tried SAM in the medical anomaly detection or localization tasks, due to the need for supervised fine-tuning.

3. Multi-modal retinal (MMR) dataset

Although the fusion of the fundus image modality and the OCT image modality is beneficial for judging retinal disease, there are no open-source multi-modal data (i.e., fundus image and OCT image) available, due to the high acquisition cost. To this end, we established a cooperation with Renmin Hospital of Wuhan University in China, and aimed to build a multi-modal dataset for anomaly detection and localization research on RAO disease. The paired images were acquired by a Zeiss VISUCAM 200 Fundus camera for the fundus images and spectral-domain optical coherence tomography (SD-OCT) for the OCT images. The fundus image size is 2124×2056, and the OCT image size is 868×1536 or 596×1264. The scan line information in the fundus image is also included in each OCT image. All the OCT images were acquired with single-line scanning in the macular area since the main symptoms of RAO are in the macular area and single-line scanning is sufficient for the anomaly diagnosis task. The final constructed dataset contains 266 paired images of healthy people and 104 paired images of patients with retinal disease. The healthy samples are considered as normal samples and the patient samples as anomaly samples.

To support both the anomaly detection and localization tasks, the patient samples have been further labeled at the pixel level with LabelMe software by retinal specialists. A medical image was treated as an anomaly at the case level for whatever symptom had been recognized in either the OCT or fundus images. To ensure the label consistency, the retinal specialists were trained in the software use and the set standard operating procedure (SOP).

The training set includes 199 normal samples only, and the test set includes 67 normal samples and 104 anomaly samples. Each sample is a pair made up of a fundus image and an OCT image. To make the anomaly symptoms clear, we gave each RAO symptom class a unique ID (as detailed in Table 1) and labeled them separately. Retinal whitening (17.74%) is the main symptom in the fundus anomalies. Middle limiting membrane (0.65%) and atrophy of the retinal inner layers (0.22%) are the most common OCT anomalies. All the symptom classes were treated as RAO anomalies when evaluating the model in the study. We have named the dataset the **m**ulti-**m**odal **r**etinal (MMR) dataset. The MMR dataset is a challenging dataset because it supports both multi-modal anomaly detection at the case level and the localization task at the pixel level, and only the small-scale training samples can be used for learning the normal fundus representations.

4. Methods

To diagnose disease with both color fundus images and OCT images, we propose the unsupervised MMRAD model. Considering the limited medical samples, MMRAD fine-tunes the SAM visual foundation model for both the single-modal anomaly localization task at the pixel level and the modal fusion anomaly detection task at the case level (as shown in Fig. 1). To adapt SAM for the multi-modal setting, task-specific tokens are introduced for anomaly localization and detection with the attention decoder. Specifically, the fusion token interacts with both modality embeddings successively to give a comprehensive anomaly score, while the mask token processes each modality separately to output the anomaly map. To reduce the training burden, unsupervised anomaly sample simulation and low-level prompt-tuning strategies are proposed to inject the medical knowledge into SAM while keeping the heavy backbone frozen.

4.1. Modality-shared SAM encoder

Due to the high cost of obtaining multi-modal medical samples, the training samples in the MMR dataset are small in scale, and it is

Table 1
The specific RAO symptom classes in the anomaly images of the MMR dataset. Each symptom class is given an ID and the pixel percentage.

Modality	Symptom class (class id, class name, pixel percentage)		
Fundus image modality	**1** Retinal whitening – 17.74%	**2** Cherry-red spot – 0.13%	**3** Segmental changes of retina vessel – 0.96%
	4 Retinal hemorrhage – 0.67%	**5** Cotton-wool spots – 1.80%	**6** Retinal embolus – 0.02%
	7 Retinal arteriovenous cross – 0.52%	**8** Neovascularization – 0.35%	**9** Retinal hard exudates – 0.01%
OCT image modality	**1** Prominent middle limiting membrane – 0.65%	**2** Atrophy of retinal inner layers – 0.22%	**3** Subretinal fluid – 0.04%
	4 Outer retinal and choroidal hyperreflectivity in the macular – 0.23%		
	5 Outer retinal and choroidal hyperreflectivity in the parafoveal – 0.31%		
	6 Middle retinal hyperreflectivity – 0.05%		**7** Intraretinal fluid – 0.02%
	8 Posterior vitreous detachment/epiretinal membrane – 0.26%		

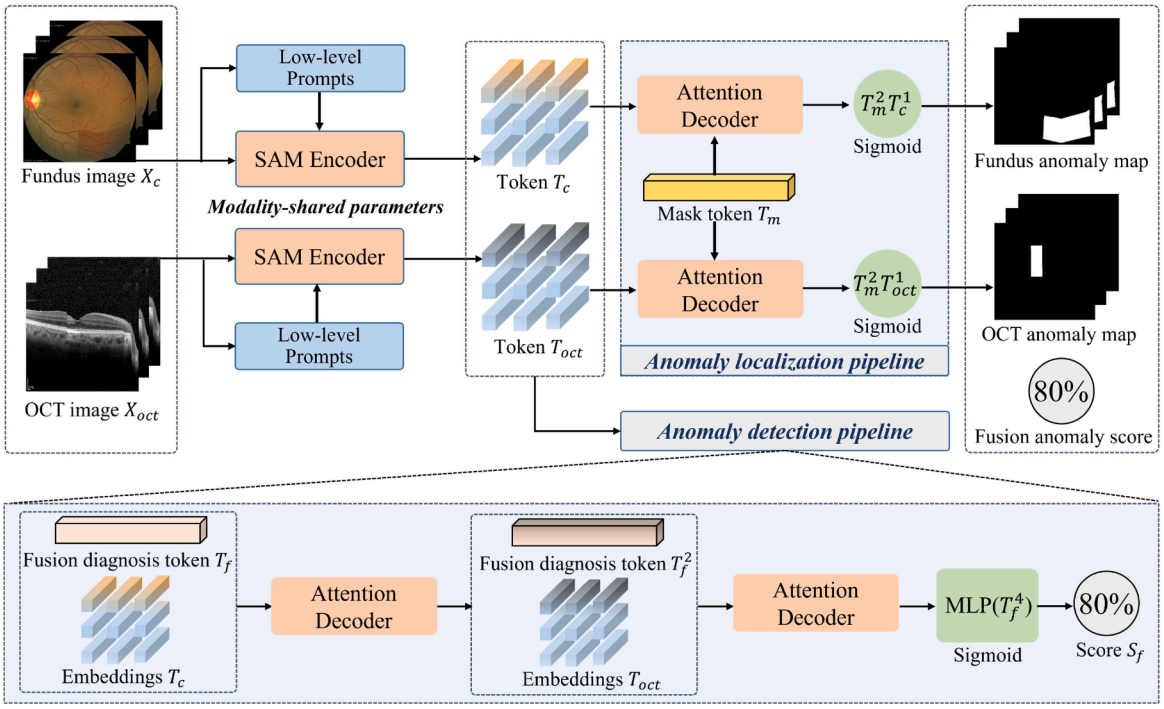

Fig. 1. The framework of the proposed **m**ulti-**m**odal model for **R**AO **a**nomaly **d**iagnosis (MMRAD). The task-specific tokens are designed to adapt SAM for supporting both the single-modal anomaly localization task at the pixel level and the modal fusion anomaly detection task at the case level with the modality-shared decoder. MMRAD fine-tunes SAM with simulated samples and low-level prompt tuning, keeping the encoder frozen to reduce the training burden.

infeasible to train an anomaly detection model from scratch. Inspired by the success of the large visual foundation models, which are trained on large-scale data and can be transferred to downstream tasks with zero or few shots, we chose to fine-tune the recent visual foundation model known as SAM [22] for the anomaly localization and detection tasks.

Before diving into the MMRAD model, it is necessary to first introduce the SAM encoder, which transforms the input image from RGB space into the embedding space, and plays the key role for the zero-shot segmentation ability. The encoder of SAM is based on the vision transformer (ViT) model [41] with cascaded transformer blocks. Similar to the NLP models [33,34], the SAM encoder processes the features in a patch as a unit, where each patch of the original image is first transformed into an embedding vector. The transformer block refines the embedding space with cross attention to exchange the mutual information between each pair of embedding vectors. Benefiting from the long-distance modeling of transformer blocks [42] and the large-scale datasets, the pretrained SAM encoder can extract high-quality embeddings for a variety of domains.

4.2. Attention decoder with task-specific tokens

The decoder of the MMRAD model is intended to support both the anomaly localization and detection tasks, given multi-modal embeddings from the SAM encoder. To achieve this, learnable task-specific tokens are introduced for the attention decoder, i.e., the mask token for the anomaly localization task and the fusion token for the anomaly detection task. The task-specific tokens and the image embeddings from the encoder constitute the query, key, and value elements in the attention mechanism [43], which injects the anomaly information into the mask token and fusion token by attention computing. The attention processing architecture in the decoder (Fig. 2) is motivated by the original SAM decoder, which alternately uses self-attention and cross-attention modules to refine the varying input query. Compared to the other decoder designs, the architecture depicted in Fig. 2 has the advantage of being compatible with the two different task-specific tokens and can use the pretraining parameters of the original SAM decoder. Since the anomaly localization and detection tasks diagnose the same patient in different levels, the decoder is made weight-shared between the two task pipelines for consistent results and to reduce the training burden. With the shared attention decoder, the mask token deals with the two modalities in a parallel manner, and the fusion token works in a serial manner for an overall examination, as shown in Fig. 1.

4.2.1. Anomaly localization with the mask token

The mask token is expected to be aware of the anomaly localization information in each modality image and is used to segment the anomaly pixels. The initial mask token is set as a learnable vector with the same feature dimension as the image embeddings. The anomaly information is gradually injected into the mask token by interacting the image embedding and the mask token with the attention decoder. By treating

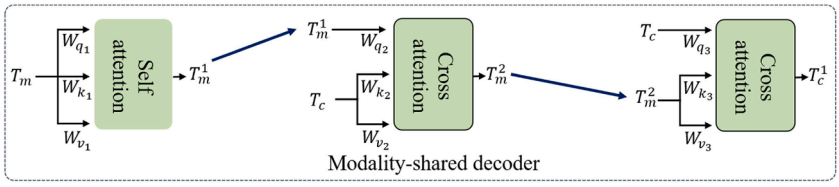

Fig. 2. The detailed architecture of the modality-shared decoder, consisting of the self-attention and cross-attention modules to exchange the information of the task specific tokens (e.g., T_m) and the image embeddings (e.g., T_c). The decoder is shared with the same weights for the RAO anomaly localization and detection pipelines.

some of the input (embeddings or the mask token) as the query value, the attention mechanism can refine this by modeling the relationship of all the query and key values. During the training stage, the mask token is optimized to have high correlation with the embeddings corresponding to the anomaly regions.

The mask token interacts with the two modality embeddings in a parallel manner since the anomaly localization is conducted at the pixel level in each modality. The localization process is the same for both the fundus and OCT images with the shared attention decoder to reduce the training burden. For simplicity, we take the fundus embeddings $T_c \in R^{L\times C}$ as an example to show the anomaly localization process in Fig. 1. L is the embedding number and C is the dimension of the embedding vector. Given the mask tokens $T_m \in R^C$ and T_c, the modality-shared decoder updates their representation with several cascaded self-attention and cross-attention modules. Eqs. (1)–(3) show the inner process of the decoder. T_m is first updated with the self-attention computation to obtain the refined T_m^1, where W_{q_1}, W_{k_1}, and W_{v_1} are the linear weights used for projecting T_m into the query, key, and value features, as in Eq. (1). The two cross-attention modules are then executed alternately, where the first module updates T_m^1 with T_c, as in Eq. (2), and the second one updates T_c with T_m^2, as in Eq. (3).

$$T_m^1 = softmax\left(\frac{T_m W_{q_1}(T_m W_{k_1})^T}{\sqrt{d}}\right)(T_m W_{v_1}) \tag{1}$$

$$T_m^2 = softmax\left(\frac{T_m^1 W_{q_2}(T_c W_{k_2})^T}{\sqrt{d}}\right)(T_c W_{v_2}) \tag{2}$$

$$T_c^1 = softmax\left(\frac{T_c W_{q_3}(T_m^2 W_{k_3})^T}{\sqrt{d}}\right)(T_m^2 W_{v_3}) \tag{3}$$

The modality-shared decoder builds a connection between the mask token and the image token and injects the anomaly localization information into T_m. The final obtained T_c^1 and T_m^2 are multiplied directly and activated with the sigmoid function to output the anomaly map M_c, as in Eq. (4).

$$M_c = Sigmoid(T_m^2 T_c^1) \tag{4}$$

Differing from most of the traditional anomaly localization methods, MMRAD outputs the normalized anomaly scores directly, without relying on any distance measurement. This end-to-end localization has been named the "one-step paradigm" in the remote sensing community [44], and MMRAD is the first method to implement this in the medical community. Without adopting any proxy task (e.g., reconstruction or generation) [10,17], the proposed model can be optimized directly with a variety of classification losses.

4.2.2. Anomaly detection with the fusion token

Differing from the anomaly localization task, which is a dense prediction task, where the mask token processes the two modalities in a parallel manner, the anomaly detection task only outputs a comprehensive score by processing the two modalities in series. The bottom part of Fig. 1 shows the serial workflow. The processed fusion token is expected to be aware of whether there is an anomaly in the given multi-modal embeddings. Compared to the mask token, the fusion token does not need to locate the anomalies but needs to consider both the fundus and OCT modalities together. This process is similar to the work of a medical professional, who will always look at multi-modal data to give a comprehensive diagnosis. To fuse the anomaly information in both modalities, the fusion token interacts with both modality embeddings successively by the use of the attention modules in the decoder. When conducting the cross-attention computation, the fusion token is treated as the query, and the two modality embeddings act as the key and value in turn, where the fusion token is updated gradually to achieve a comprehensive diagnosis.

The interaction order for the fundus and OCT modalities is flexible, and Fig. 1 shows the detailed workflow when processing T_c first. The influence of the interaction order is studied in Section 5.4.7. The interaction process for the fusion token with image embeddings in the attention decoder is identical to that for the mask token in the localization pipeline. Differently, the attention decoder is used twice to fuse the multi-modal information, and the final T_f^4 after four updates is expected to contain the comprehensive diagnostic information. To obtain the final anomaly detection score S, a multilayer perceptron (MLP) is applied to T_f^4 with the sigmoid function, as in Eq. (5).

$$S = Sigmoid\left(MLP\left(T_f^4\right)\right) \tag{5}$$

4.3. Fine-tuning strategies

To make full use of the pretrained SAM encoder weights, MMRAD is fine-tuned with simulated samples and generates low-level prompts while keeping the SAM encoder frozen. Since labeled retinal samples are difficult to access, we propose a general sample simulation strategy to fine-tune MMRAD in a pseudo-supervised manner with only healthy images. With the simulated samples, we follow the prompt-tuning paradigm for efficient fine-tuning and emphasize generating low-level prompts that are suitable for the MADL task.

4.3.1. Medical anomaly sample simulation

Real medical anomalies are difficult to acquire, and the label cost is expensive, so we chose to simulate the medical anomaly samples for end-to-end fine-tuning. From the perspective of the data distribution, the anomalies are the out-of-distribution samples, given a normal distribution [45]. The anomaly diagnosis model does not discriminate the special categories of the out-of-distribution samples but only discriminates whether the given sample is of the healthy distribution. Inspired by this observation, we simulate the anomalies by shifting the distribution of the normal samples violently with data augmentation techniques. With different kinds of augmentation combinations, various out-of-distribution samples can be generated.

Fig. 3 shows the overall simulation workflow. The workflow is universal for both the fundus and OCT modalities, and we take a fundus image X_c as an example. Given a healthy image X_c, the workflow chooses the anomaly generation area (i.e., the π operation) first and then transforms its distribution with the data augmentation techniques (i.e., the φ operation) to create the out-of-distribution anomalies. We found that a single kind of augmentation technique performs worse than the combined use of various augmentation techniques, which provide more diverse anomalies and benefit the diagnosis results on real anomalies (as discussed in Section 5.4.1). Since a real anomaly area may have irregular shapes, the workflow finally applies affine transformation to twist the anomaly area. It is worth noting that the proposed workflow has great randomness in the selection of anomaly location, data augmentation of the anomaly generation, and affine transform, which can make the generated samples diverse and prevent the model learning specific kinds of anomalies.

4.3.2. Low-level prompt tuning

The simulated paired samples make it possible to fine-tune SAM in a supervised manner. However, there is still one remaining problem, i.e., which fine-tuning paradigm to choose. The popular paradigm is partial tuning (such as the decoder), which is always inferior to full tuning with the high storage cost. Recently, the prompt-tuning paradigm has demonstrated a strong generalization ability in high-level recognition tasks [39,46]. Unfortunately, the MADL task is a low-level task, and a gap exists when directly using recognition-specific prompts.

To deal with this problem, we developed **l**ow-level **p**rompt **t**uning (LPT) to complete the fine-tuning stage. According to Fig. 1, the LPT

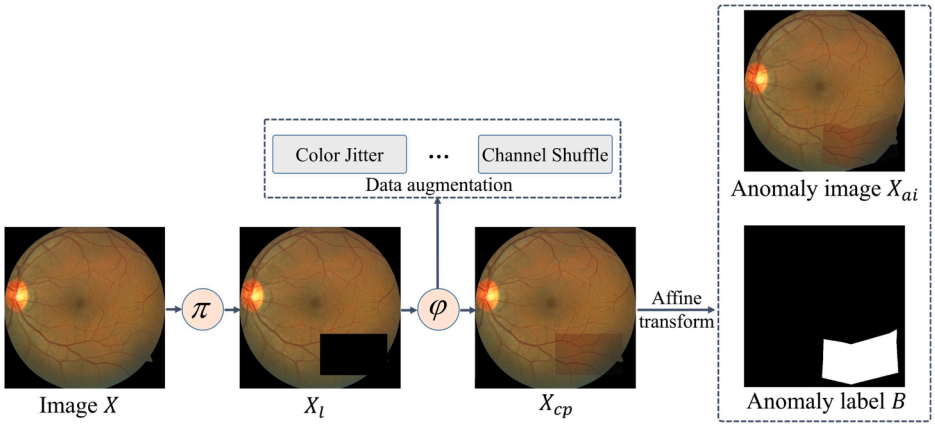

Fig. 3. The workflow of the proposed anomaly sample simulation strategy, which is general for both the OCT and fundus images.

freezes the heavy backbone weight (i.e., the SAM encoder) and changes the embedding space for each encoder transformer layer with the generated prompts. The prompts are learnable embeddings [47], which are trained to contain medical-specific knowledge to assist the original SAM encoder. To make the prompts aware of the low-level features, we use both the shallow features from the large-scale pretrained convolutional blocks (i.e., features $F_c \in R^{64\times64\times h}$) and the pretrained patch embeddings (i.e., features $F_p \in R^{64\times64\times h}$) from the original SAM encoder to generate the prompts, where h represents the channel dimension of the obtained feature cubes. The pretrained features from the two different architecture families (i.e., CNN and transformer) are expected to complement each other [42,48]. As in Fig. 4(a), F_c is extracted mainly by the m individual pretrained convolutional blocks $C_1 - C_m$. The process can be formally expressed as shown in Eq. (6), where [ ] represents the concatenation operation.

$$F_c = LN([C_1(X), ..., C_m(X)]) \tag{6}$$

Both F_c and F_p have smaller feature dimensions than the encoder embedding space, to reduce the learning burden with limited medical samples. F_c and F_p are fused by element-wise addition, and act as the base to generate the n prompts for the n transformer encoder layers (as in Fig. 4(b)). Since the embedding space varies for each layer, N unshared linear layers $LN_1 - LN_n$ are set to customize the common low-level features for each encoder layer, and one shared layer LN_s increases the dimension of $LN_1 - LN_n$ to be the same as the layer input. In total, each prompt Pr_i for the i-th transformer layer can be represented as shown in Eq. (7).

$$Pr_i = LN_s(LN_i(F_c + F_p)) \tag{7}$$

4.4. Multi-modal training and testing

With the redesigned decoder and the fine-tuning strategies, the MMRAD model can process both the fundus and the OCT modalities simultaneously, without the need for real anomalies. At the training stage, if both modality images are available, as in Fig. 5(a), MMRAD is optimized jointly for the anomaly localization task at the pixel level and the anomaly detection task at the case level. The anomaly localization loss L_{al} is composed of cross-entropy (CE) loss on both modalities, as in Eq. (8), where B_c and B_{oct} are the simulated binary labels corresponding to the output anomaly maps M_c and M_{OCT}, respectively.

$$L_{al} = CE(M_c, B_c) + CE(M_{oct}, B_{oct}) \tag{8}$$

Since the anomaly detection result for a pair of fundus and OCT images is considered an anomaly once any modality shows disease, the anomaly detection loss is computed as shown in Eq. (9), with $\max(B_c, B_{oct})$ as the ground truth. The $\max(B_c, B_{oct})$ value is a scalar value of 0 or 1. L_{al} and L_{ad} work together (as in Eq. (10)) to optimize MMRAD for both the anomaly localization and anomaly detection tasks, given both modalities.

$$L_{ad} = CE(S, \max(B_c, B_{oct})) \tag{9}$$

$$L = L_{al} + L_{ad} \tag{10}$$

If only single-modal images are available at the training stage, as in Fig. 5(b), the model only outputs the anomaly localization map and is optimized only with $CE(M_c, B_c)$ for the fundus modality (or $CE(M_{oct}, B_{oct})$ for the OCT modality). Since there is no fusion requirement to output the anomaly detection results, we found that the traditional manner of taking the maximum or mean value as the image-level detection results

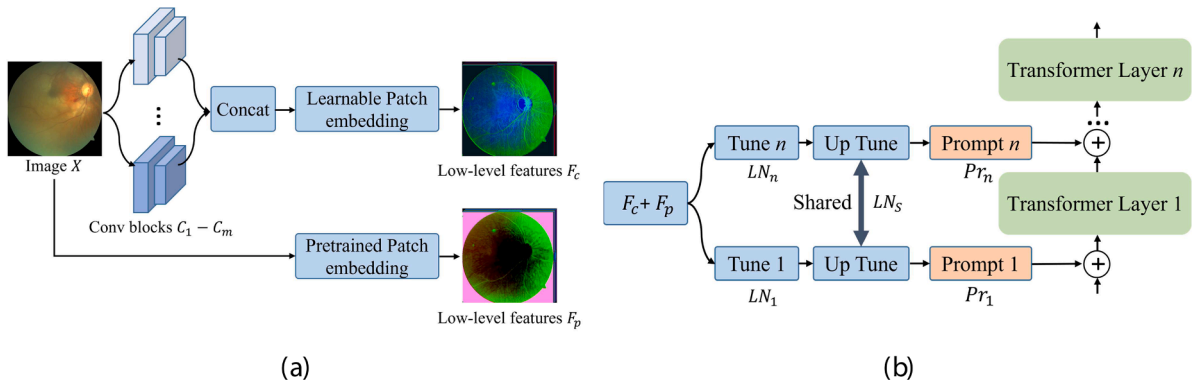

Fig. 4. (a) Low-level extraction component. (b) Prompt generation component. (a) and (b) belong to the low-level prompt tuning, which generates prompts in the transformer input space to fine-tune MMRAD.

Fig. 5. The training process for the proposed model with both modalities (a) and a single modality (b). When both modalities are available, the model is optimized jointly for the anomaly localization task at the pixel level and the anomaly detection task at the case level. When only a single modality is available, the model outputs and optimizes the anomaly localization task only. At the test stage, the trained model from (a) can also process a single modality to output the anomaly localization map, as shown in (b).

is enough for the single-modal input. CE loss is chosen in Eqs. (8)–(10) since it is the classical baseline of many advanced classification or segmentation loss functions. Certain advanced loss functions could improve the performance further, but that was not the focus in this study (as discussed in Section 5.4.5).

At the test stage, the model trained with multi-modal images from Fig. 5(a) can infer both a multi-modal image pair and a single-modal image. Given a pair of fundus and OCT images, the model can output two anomaly localization maps and one anomaly detection score simultaneously. Given a single-modal image, the model outputs the anomaly localization map only. In contrast, the model trained with single-modal images from Fig. 5(b) can only process the same modality images and generate the anomaly localization map, which is exactly how we used the MMRAD model on the Indian Diabetic Retinopathy Image Dataset (IDRiD) dataset with the fundus modality only.

5. Results

5.1. Experimental settings

5.1.1. Datasets

Due to the sparsity of retinal images with pixel-level annotation, two accessible datasets were used to evaluate the proposed MMRAD model. The MMR dataset built in this study contains multi-modal OCT and fundus images, which supports both the anomaly detection task at the case level and the anomaly localization task at the pixel level for RAO diagnosis. Detailed information about the MMR dataset is provided in Section 2. The IDRiD dataset is publicly available [49], but supports only the fundus anomaly localization task. IDRiD focuses on DR detection, and the lesions include microaneurysms, hemorrhages, hard exudates, and soft exudates. To train the anomaly localization model, we chose the healthy images of the IDRiD dataset as the training set (168 images) and the lesion images as the test set (81 images).

5.1.2. Comparison methods and evaluation metrics

The proposed MMRAD model was compared with two classical medical anomaly diagnosis methods and four state-of-the-art methods. The former methods were f-AnoGAN [17] and the convolutional autoencoder (CAE) [29]. The latter methods were a transfer learning (TL)-based model [13], PatchCore [50], the patch distribution modeling method (PaDiM) [51], and UniAD [52]. The comparison models were trained separately on each modality to output the anomaly location map, where the maximum value of both modalities was taken to act as the anomaly detection score at the case level.

We take the area under the receiver operating characteristic (ROC) curve (AUC) as the main quantitative metric for both tasks, which considers the trade-off between the true positive rate (TPR) and the false positive rate (FPR) at different thresholds. The varying thresholds determine whether a pixel is an anomaly for the anomaly localization task and whether an image contains the anomaly for the anomaly detection task. However, AUC is always biased toward the large anomalies and has limitations in evaluating small anomalies for the anomaly localization task [53]. To address this problem, we also introduce the per-region overlap (PRO) metric to evaluate the localization task, following [53,54], which treats the anomaly regions of different size equally.

5.1.3. Implementation details

The f-AnoGAN model used the ResNet-based Wasserstein GAN (WGAN) with gradient penalty (WGAN-GP) [55] for stable GAN training. To balance the rich spatial detail of medical images and the intractable GAN training, the generated image size was set to 256×256 and the latent space dimension was set to 128, following Thomas et al. [17]. We implemented CAE with ResNet-50 as the backbone and the same DeConvNet architecture as Rashmi et al. [29]. To strengthen the detection ability of the TL-based method, we used the ImageNet-pretrained EfficientNet v2 [56] to extract the features and implemented the isolation forest algorithm using the Scikit-learn library [57]. The memory bank in PatchCore was constructed from the pre-trained ResNet-50 model, and the sampling ratio of PatchCore set was set to 0.1. PaDiM selected the same pretrained features as PatchCore, and the randomly selected dimension was set to 100. All the settings of UniAD followed the original paper [52].

The proposed MMRAD model was built based on SAM with the adapted ViT-B encoder. The π operation was executed with a probability of 0.7, which means that pure healthy samples could also be generated. For the LPT, we used the first three layers of two ImageNet-pretrained ResNet-50 models with different parameters to extract the low-level features. The mask token T_m and fusion token T_f had the same token length of 256. In practice, T_m and T_f were concatenated together for efficient inference. The attention decoder and the mask token are weight-sharing modules for both modalities. T_f interacts with the fundus modality first for the anomaly detection at the case level (as discussed in Section 5.4.7). The augmentation techniques used for the anomaly simulation are a composition of pixel-level and spatial-level transformations (as discussed in Section 5.4.1). MMRAD was optimized using the AdamW optimizer [58] (learning rate $= 0.0002, \beta_1 = 0.9$. $\beta_2 = 0.999$) and the pretrained SAM parameter was fine-tuned for 1 epoch.

5.2. Results obtained on the MMR dataset

5.2.1. Anomaly localization results obtained on the single-modal images

The results obtained on the constructed MMR dataset are made up of the single-modal anomaly localization results and the modal fusion anomaly detection results. Fig. 6 shows the qualitative anomaly localization results for the fundus images, where we masked the black background for better visualization. The f-AnoGAN and CAE models have the highest anomaly degree on the optic disc area, and the real RAO anomaly area has a lower response. This is because the used reconstruction error in image space is sensitive to extreme values. In contrast, the remaining methods compute the anomaly score in feature space and show a higher response in the anomaly area. The PaDiM and PatchCore models are the only methods that locate the anomaly vessels in the first example, but fail for the remaining anomaly areas. The proposed MMRAD model found most of the anomaly areas, with fewer false alarms than the comparison methods. Table 2 lists the corresponding quantitative results. Except for the proposed MMRAD, PaDiM and PatchCore both obtain PRO values that are higher than 0.40. Under the AUC metric, MMRAD achieves the highest score of 0.6997 and is around 3 points ahead of the second-place PaDiM model.

Fig. 7 shows the qualitative anomaly localization results for the OCT images. Without expert knowledge, readers could identify the RAO anomalies in the samples of Fig. 6, but they would struggle to process the sample OCT images in Fig. 7. Identifying the anomalies in the OCT modality is more difficult than in the fundus images. Almost all the comparison models fail to localize the anomaly region in Fig. 7. The f-AnoGAN model tends to recognize the brightest region as an anomaly region due to the almost fully black background. Although MMRAD shows a green color rather than the ideal blue color in the imaging region, the real anomaly regions are identified with the highest anomaly response. Table 2 lists the corresponding quantitative results. The AUC values of the f-AnoGAN and CAE models are around 0.5, showing the great challenge for the OCT modality. MMRAD obtains the best performance under both the AUC and PRO metrics, surpassing the second-place PaDiM model by around 3 points in AUC score and around 10 points in the PRO metric.

5.2.2. Anomaly detection results obtained on the multi-modal images

Table 2 also lists the multi-modal anomaly detection results. The multi-modal anomaly score of the comparison models was obtained by using some of the statistical values of the single-modal anomaly maps, following Zavrtanik et al. [59]. Compared to the anomaly localization task at the pixel level, the anomaly detection task at the case level only needs to diagnose whether there is an RAO anomaly in the images, without localizing it. The relatively simpler demand brings a higher detection AUC score when compared to the localization scores. Most models obtain an AUC score of higher than 0.70. MMRAD surpasses the comparison models by a large margin and achieves the highest AUC score of 0.7664, which shows the superiority of the designed fusion token, compared to the use of the statistical values.

5.3. Results for the IDRiD dataset

The IDRiD dataset is a valuable open-source retinal dataset with pixel-level annotation, for which we report the anomaly localization results in Fig. 8 and Table 3. The IDRiD dataset was annotated in a semi-automatic mode and has many sporadic diseased regions. Similar to the results for the MMR dataset, the f-AnoGAN and CAE models can recognize some of the anomaly areas with a low response. The remaining comparison models can locate the anomaly region more accurately, but their resolutions are low since their detection process is conducted in feature space. Many tiny diseased areas are ignored in the anomaly maps of PatchCore, PaDiM, and UniAD. Since the black background in the IDRiD dataset is not all zero pixels, the TL-based method also has a high response to the background in the third example, caused by some tiny value changes. MMRAD finds most of the large and small diseased areas, and surpasses the second-place model by around 2 points in AUC score and 4 points in the PRO metric.

5.4. Model analysis

5.4.1. Anomaly transformation

The anomaly transformation generates the pseudo-anomaly regions in the proposed simulation strategy, and the simulation quality has a great effect on the diagnosis results. Table 4 lists the related results when changing the transformation settings. The tested settings included both the pixel-level transformations and the spatial-level transformations. All the transformations were implemented using the Albumentations library [60]. Color jitter, channel shuffle, and RGBShift transformation change the pixel value while keeping the spatial location unchanged. Among these transformations, RGBShift achieves the best overall performance with the localization AUC in the fundus image modality (0.6839) and the AUC in the OCT image modality (0.8594). The channel shuffle transformation results in an obvious accuracy drop for the anomaly localization task in the OCT modality. We speculate that this may have been caused by the gray style of the OCT image, and shuffling the channels may not change the distribution distinctly. Compared to the pixel-level transformation, the spatial-level transformation (i.e., rotation) results in a higher localization AUC in the fundus image modality (0.7037), but with an obvious drop in the multi-modal anomaly detection task. By combining the transformations, MMRAD obtains the best

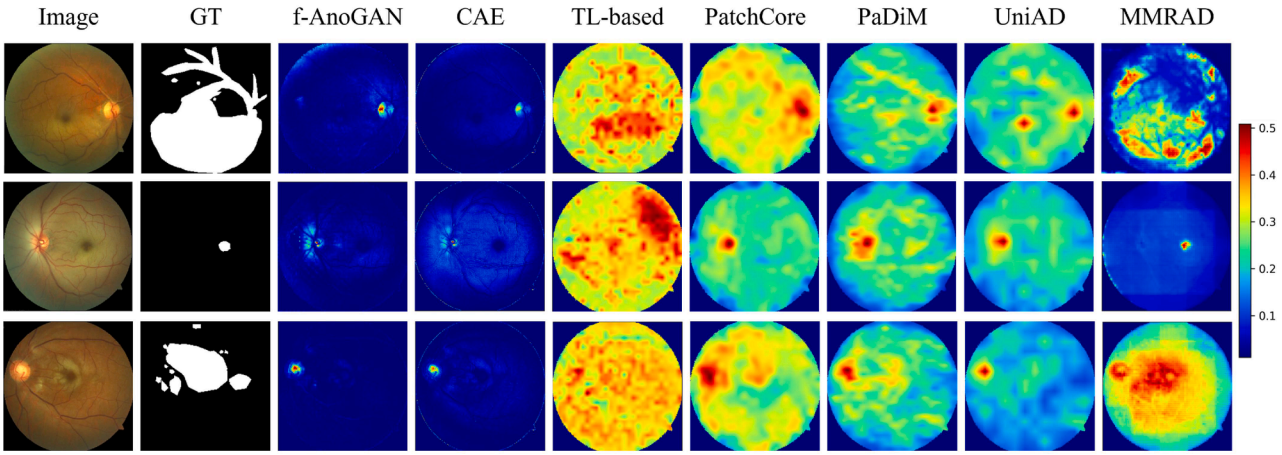

Fig. 6. The qualitative detection results for the RAO anomaly localization task on the fundus image modality.

Table 2
The quantitative comparison results for the anomaly localization and detection tasks on the constructed MMR dataset.

Method	Anomaly localization (fundus modality)		Anomaly localization (OCT modality)		Anomaly detection (with both modalities)
	AUC	PRO	AUC	PRO	AUC
f-AnoGAN	0.5958	0.2577	0.5360	0.3052	0.7232
CAE	0.5751	0.2197	0.4225	0.2710	0.6447
TL-based	0.5303	0.1714	0.6496	0.2510	0.5979
PatchCore	0.6531	0.4446	0.7809	0.4311	0.7062
PaDiM	0.6720	**0.4639**	0.8042	0.4417	0.7196
UniAD	0.6676	0.3551	0.7809	0.3526	0.7014
MMRAD	**0.6997**	0.4303	**0.8295**	**0.5339**	**0.7664**

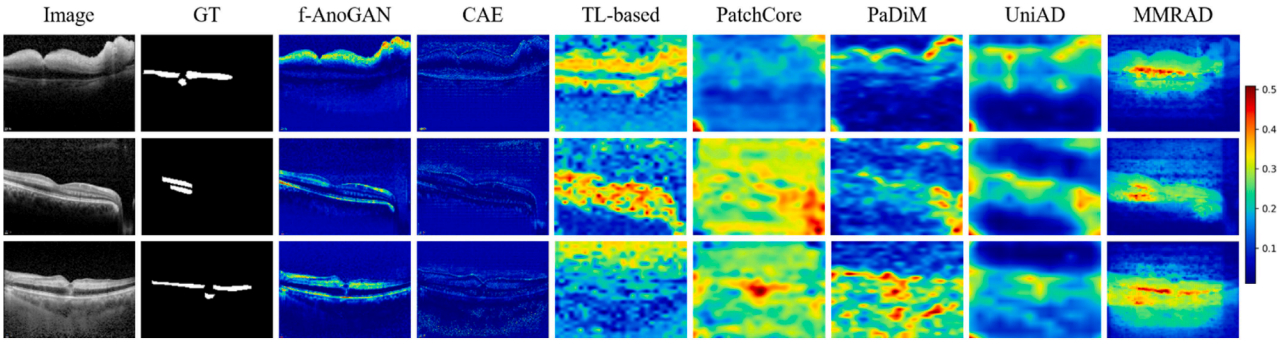

Fig. 7. The qualitative detection results for the RAO anomaly localization task on the OCT image modality (MMR dataset).

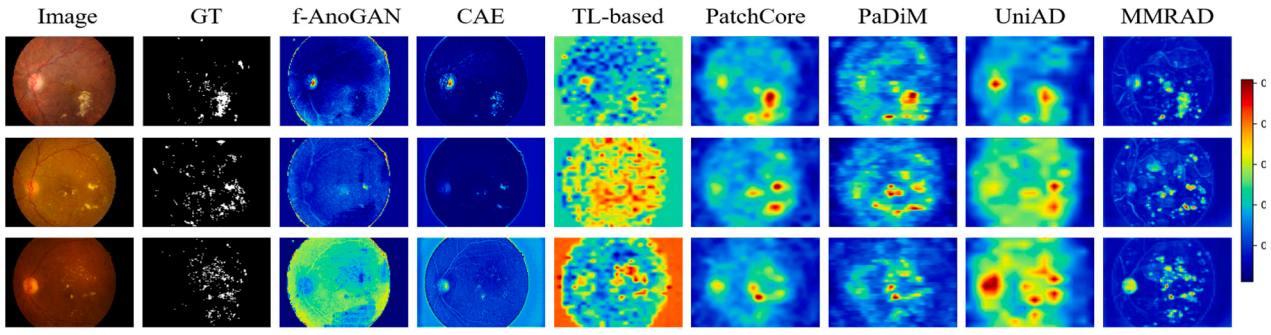

Fig. 8. The qualitative detection results for the anomaly localization task on the fundus images (IDRiD dataset).

Table 3
The quantitative comparison results for the anomaly localization task on the IDRiD dataset.

Method	Anomaly localization	
	AUC	PRO
f-AnoGAN	0.6649	0.2794
CAE	0.4809	0.1049
TL-based	0.4680	0.1716
PatchCore	0.8126	0.3036
PaDiM	0.7892	0.3147
UniAD	0.7641	0.2394
MMRAD	**0.8375**	**0.3544**

Table 4
Ablation comparison results for the chosen anomaly transform on the MMR dataset.

Anomaly transformation setting	Anomaly localization in fundus		Anomaly localization in OCT		Multi-modal anomaly detection
	AUC	PRO	AUC	PRO	AUC
Color jitter	0.6456	0.3412	0.8241	0.4999	0.6815
Channel shuffle	0.6736	0.3688	0.6314	0.4364	0.6518
RGBShift	0.6839	0.4251	0.8594	**0.5607**	0.6708
Rotate	**0.7037**	**0.4504**	0.8262	0.5232	0.5679
Composition	0.6997	0.4303	**0.8295**	0.5339	**0.7664**

overall performance (ranking first or second place), and especially the highest detection AUC score of 0.7664. The combination of transformations can increase the anomaly diversity and generate anomalies that are suitable for both modalities.

5.4.2. *Multi-modal vs single-modal images*

OCT and fundus images are the two most important modalities for retinal diagnosis, where the complementary information has been proven useful in many previous works [8,9]. To validate the superiority of multi-modal images over single-modal images in our task, we report the anomaly detection results of the different models for single-modal images in Table 5. Most models achieve a better RAO detection performance with both modalities, compared to a single fundus or OCT modality, with a stable promotion. The f-AnoGAN, CAE, and MMRAD models obtain an increment of around 3 points over the single-modal results. There are exceptions for PatchCore and UniAD, where the single modality performs better than the modality fusion. Note that both PatchCore and UniAD achieve extremely imbalanced detection

Table 5
The RAO anomaly detection results with a single modality and both modalities on the MMR dataset.

Method	Fundus modality AUC	OCT modality AUC	With both modalities AUC
f-AnoGAN	0.6985	0.6288	0.7232
CAE	0.5530	0.6058	0.6447
TL-based	0.5746	0.5919	0.5979
PatchCore	**0.7399**	0.6287	0.7062
PaDiM	0.6756	0.7065	0.7196
UniAD	0.7271	0.6524	0.7014
MMRAD	0.7350	**0.7121**	**0.7664**

performances on the two modalities, which may increase the fusion barrier and thus decrease the performance.

For the anomaly localization task, the promotion of multi-modal training can also be observed in Table 6. Since the comparison models were trained for each modality to output the anomaly localization map in turn rather than together, the comparison experiment was conducted with the proposed MMRAD. MMRAD uses both modalities in a parallel manner with the modality-shared decoder, where the multi-modal training improves the decoder with more images, and a better decoder can benefit both modalities. Tables 5 and 6 show that the combined use of fundus and OCT modalities can benefit both the anomaly localization and detection tasks.

5.4.3. Analysis of the prompt-tuning method

The LPT method is designed to tune the heavy SAM encoder by injecting the learnable low-level prompts into each encoder layer, which consists of the low-level feature extraction module and the prompt generation module. Since the existing prompt-tuning methods cannot be strictly divided into these two modules, the related analyses were conducted by varying the design choices. For the analysis of the low-level feature extraction module, we compared the extraction strategies of CNN blocks without pretraining, Fourier transform, as in [39], and the pretrained CNN used in the proposed method. The results are reported in Table 7. Since there is no low-level prior in CNN blocks without pretraining, this approach performs around 3–6 points lower than the other two strategies. Benefiting from the large-scale pretraining dataset and the advantages of the data-driven approach, the pretrained CNN blocks can extract better low-level features than the hand-crafted Fourier transform approach, especially under the PRO metric for both the fundus and OCT modalities.

The main unique design of the prompt generation module lies in the LN_s with shared weights and $LN_1 - LN_n$ without shared weights for each transform layer, so we varied this design to validate its effectiveness. In the four different settings listed in Table 7, sharing all the layer weights results in the worst performance, which can be understood since the generated prompts would be the same for all the layers with different levels of features. If the setting goes to the other extreme, without any shared weights, the performance is better but there is still a great gap with the partly shared setting. The large number of trained parameters makes the fine-tuning process with limited medical samples difficult. Comparing the partly shared setting of sharing $LN_1 - LN_n$ or LN_s, we found that there is only a stable promotion when sharing $LN_1 - LN_n$, and thus we chose this setting for the prompt generation module.

We also compared the proposed LPT method with the famous visual prompt tuning (VPT) [47] and AdaptFormer [61] methods, the recent explicit visual prompting (EVP) method [39] for structural segmentation, and also the baseline with trained decoder only. The results are reported in Table 8. Unsurprisingly, all the prompt-tuning methods surpass the decoder-only baseline by injecting the downstream knowledge into the backbone layers. VPT and AdaptFormer are designed as general tuning methods without special consideration for the low-level tasks such as anomaly diagnosis, and perform worse than EVP and the proposed LPT. Compared to EVP, the proposed method extracts the low-level features from the shallow layers in large-scale pretrained CNN blocks, and has a better overall performance.

Table 6
The anomaly localization results of the proposed model trained with a single modality and both modalities on the MMR dataset.

Metric	Single-modal training		Multi-modal training	
	Fundus modality	OCT modality	Fundus modality	OCT modality
AUC	0.6873	0.8095	0.6997	0.8295
PRO	0.4384	0.5156	0.4303	0.5339

5.4.4. Modality-shared mask token and decoder

Considering the limited medical samples, we made the mask token and the decoder shared by both modalities to reduce the training burden. To validate the effectiveness of this approach, we compared the performances of different weight-sharing choices and report the results in Table 9. Setting either the mask token and the decoder weight unshared leads to an obvious performance drop in both the localization and detection tasks. Compared to the choice of shared decoder only, the choice of shared mask token only uses two separate decoders for the fundus and OCT modalities, where the attention mechanisms increase the number of learned parameters, and this choice leads to a worse performance. Overall, the decrease of learned parameters benefits the RAO anomaly diagnosis performance with limited medical samples.

5.4.5. Comparison of different training losses

The choice of different training losses has an obvious effect on the RAO anomaly diagnosis results. To make the effect clear, we compared the CE loss with the advanced focal loss [62], dice loss [63], and Lovasz loss [64] on the MMR dataset, as shown in Table 10. Both the dice loss and focal loss train the model effectively, but the focal loss surpasses the CE loss in many metrics by a large margin. The results show that the performance of the proposed MMRAD method can be enhanced further with a more advanced loss function. Despite this, the CE loss was chosen to train the proposed model for the comparison in Table 2 and Table 3, since it is the classical baseline of many advanced classification or segmentation losses, and the main diagnosis promotion is expected to come from the SAM pretraining, new multi-modal decoder, and the fine-tuning strategies, and not an advanced loss function.

5.4.6. Model generalization

The model generalization ability is of great significance for practical diagnosis. To validate this, since there are no publicly available multi-modal retinal datasets, we tested the MMRAD model trained on the constructed MMR dataset directly on the single-modal IDRiD [49] and (Retinal Edema Segmentation Challenge) RESC datasets [65]. The IDRiD and RESC datasets have 81 test fundus images and 15 test OCT slices, respectively. Since each OCT sample in the RESC dataset has 128 slices, the first layer of the SAM encoder was also made trainable to adapt to this situation. We report the results in Table 11. Table 11 also reports the results obtained with the same training and test datasets, which can be regarded as an evaluation reference for the transfer diagnosis results. Despite the different datasets being collected from different hospitals and the great domain differences existing, the proposed MMRAD model achieves an AUC score of higher than 0.75. Especially for the IDRiD dataset, the transfer diagnosis results of MMRAD are still be better than those of the many comparison models, which were both trained and tested on the IDRiD dataset, as in Table 2. The quantitative results fully prove the model generalization ability for data from different hospitals.

5.4.7. Effect of the fusion token

The fusion token is designed to conduct the multi-modal anomaly detection task. To prove its superiority, we compared the fusion token with some anomaly detection strategies based on statistical values, and also explored the modality processing sequence for the fusion token.

Table 7
The comparison results with different design choices in the low-level extraction and prompt generation modules on the MMR dataset.

Module	Design choice	Anomaly localization (fundus modality)		Anomaly localization (OCT modality)		Anomaly detection (with both modalities)
		AUC	PRO	AUC	PRO	AUC
Low-level extraction	Without pre-training	0.6628	0.3037	0.7647	0.4333	0.7257
	Fourier transform	**0.7197**	0.3554	0.8241	0.5068	0.7524
	Pretrained CNN (proposed)	0.6997	**0.4303**	**0.8295**	**0.5339**	**0.7664**
Prompt generation	Without shared weights	0.6464	0.3386	0.8154	0.4573	0.7089
	Shared $LN_1 - LN_n$ only	0.6782	0.3972	**0.8405**	0.4880	0.7407
	Shared $LN_s$and $LN_1 - LN_n$	0.6253	0.2932	0.7706	0.3674	0.6693
	Shared LN_sonly (proposed)	**0.6997**	**0.4303**	0.8295	**0.5339**	**0.7664**

Table 8
The comparison results for different prompt-tuning methods on the MMR dataset.

Method	Anomaly localization (fundus modality)		Anomaly localization (OCT modality)		Anomaly detection (with both modalities)
	AUC	PRO	AUC	PRO	AUC
Only decoder	0.6206	0.2261	0.7713	0.3008	0.6872
VPT	0.6305	0.3996	0.7895	0.4109	0.7166
AdaptFormer	0.6472	0.3176	0.8142	0.4616	0.7225
EVP	0.6979	**0.4502**	0.8144	0.4671	**0.7758**
LPT (proposed)	**0.6997**	0.4303	**0.8295**	**0.5339**	0.7664

Table 9
The comparison results with different weight-sharing choices in the decoder and mask token on the MMR dataset.

Design choice	Anomaly localization (fundus modality)		Anomaly localization (OCT modality)		Anomaly detection (with both modalities)
	AUC	PRO	AUC	PRO	AUC
Shared decoder only	0.6846	0.3842	0.8188	0.4827	0.7624
Shared mask token only	0.6609	0.3088	0.7573	0.4584	0.7237
Shared mask token and decoder (proposed)	**0.6997**	**0.4303**	**0.8295**	**0.5339**	**0.7664**

Table 10
Comparison results of different training loss functions on the MMR dataset.

Loss	Anomaly localization (fundus modality)		Anomaly localization (OCT modality)		Anomaly detection (with both modalities)
	AUC	PRO	AUC	PRO	AUC
Focal	**0.7210**	0.3477	**0.8491**	**0.5936**	**0.7773**
Dice	0.6928	0.3882	0.8356	0.5187	0.7137
Lovasz	0.6451	0.3836	0.7677	0.3029	0.7086
Cross-entropy	0.6997	**0.4303**	0.8295	0.5339	0.7664

Table 11
The anomaly localization transfer results of the proposed MMRAD model.

Test dataset	Training dataset	AUC	PRO
IDRiD	IDRiD	0.8375	0.3544
	MMR	0.7836	0.2601
RESC	RESC	0.8171	0.4371
	MMR	0.7549	0.3974

Table 12
Comparison results for the fusion token and other fusion strategies on the MMR dataset.

Anomaly detection setting	Multi-modal anomaly detection AUC
Maximum	0.7338
Mean	0.4597
Fusion token ($T_{oct}-T_c$)	0.7456
Fusion token (T_c-T_{oct})	**0.7664**

Table 12 reports the related results. With the maximum value in both modalities, the obtained results show a small drop when compared to the fusion token. In contrast, the mean value fails to detect the case-level anomalies, with an AUC score of 0.4597. We speculate that two different cases can have similar mean values, leading to the low representation with the mean value. For the fusion token, changing the modality processing sequence from OCT image to fundus image, or the opposite, results in higher AUC scores than the single-modal detection results in Table 5. This shows that once the fusion token has seen both modalities, it can obtain better results, regardless of the processing sequence.

5.4.8. Results for each symptom

The MMR dataset was constructed with unique labels for the different RAO symptoms (nine in the fundus modality and eight in the OCT modality). To explore the diagnosis ability for each single symptom, we evaluated the anomaly localization performance for each symptom, and the results are given in Table 13. For each symptom, we consider its separability with the normal patterns while ignoring the remaining symptoms. We found that the proposed MMRAD model shows a satisfactory performance in most symptoms, such as the cherry-red spot (ID 2) and neovascularization (ID 8) symptoms in the fundus modality, and the subretinal fluid (ID 3) and intraretinal fluid (ID 7) symptoms in the OCT modality. In the fundus modality, there are two symptoms with an AUC score of lower than 0.60, i.e., retinal hemorrhage (ID 4) and retinal embolus (ID 6). In the OCT modality, the symptom of posterior vitreous detachment/epiretinal membrane (ID 8) is the only class with an AUC score of lower than 0.75.

5.4.9. Token initialization

To explore the effect of the token initialization method on the performance, we tested three different initialization methods and report the results in Table 14. We applied the same initialization method for both the mask token and fusion token in each test. Normal initialization makes the initial weights follow a normal distribution with mean 0 and standard deviation 0.01 [66]. Kaiming initialization goes a step further by taking into account the impact of the token dimensionality [67]. These two methods had not undergone any data training process. In contrast, the third method (pretrained token) means using the pretrained mask token and intersection over union (IoU) token from SAM to initialize the corresponding mask token and fusion token, respectively. Similar to the fusion token, IoU also gives an image-level evaluation for the IoU score. In comparison, there are no significant differences

Table 13
The anomaly localization performance for each single RAO symptom on the MMR dataset.

Modality	Results for each symptom											
	ID	AUC	PRO	ID	AUC	PRO	ID	AUC	PRO	ID	AUC	PRO
Fundus image modality	1	0.7078	0.2104	2	0.8575	0.5786	3	0.6802	0.2498	4	0.5580	0.4801
	5	0.6603	0.5171	6	0.5904	0.2965	7	0.6545	0.5533	8	0.7929	0.6511
	9	0.6697	0.5837									
OCT image modality	1	0.8855	0.3870	2	0.7574	0.4279	3	0.8758	0.5456	4	0.8910	0.5378
	5	0.8539	0.3789	6	0.8399	0.3290	7	0.8651	0.7227	8	0.6413	0.4633

Table 14
Comparison effects of different token initialization methods on the MMR dataset.

Method	Anomaly localization (fundus modality)		Anomaly localization (OCT modality)		Anomaly detection (with both modalities)
	AUC	PRO	AUC	PRO	AUC
Normal	0.6844	0.4280	0.8229	0.5326	0.7493
Kaiming	**0.7097**	0.4274	**0.8302**	0.5258	0.7459
Pretrained	0.6997	**0.4303**	0.8295	**0.5339**	**0.7664**

between normal initialization and Kaiming initialization in Table 14. The initialization method of pretrained token shows a similar performance in the anomaly localization task, without obvious differences, but the anomaly detection score shows an improvement of around 2 points. Therefore, we finally chose the pretrained tokens from the original SAM to initialize the mask token and fusion token in MMRAD.

5.4.10. Distribution of simulated samples

Simulated anomalies can be expected to be out-of-distribution with healthy samples, and we verify this with the t-distributed stochastic neighbor embedding (t-SNE) [68] visualization technique in Fig. 9. Since the original image space has only three channels (RGB) and may be insufficient to represent the retinal pattern, we converted the images into the high-level embeddings space from MMRAD and visualized the embeddings with the t-SNE technique. Fig. 9 visualizes real anomalies, the simulated anomalies, and normal samples in both the fundus and OCT modalities. Although there is some overlap, the distributions of the three categories exist independently, overall, with relatively weak correlations. The OCT modality exhibits greater separability, which is consistent with the higher AUC score in Table 2. Furthermore, a wider gap exists between the simulated anomalies and normal samples, compared to the real anomalies, since the simulation process described in Section 4.3.1 tends to alter the distribution more aggressively, which proves that the simulated samples can meet the out-of-distribution condition.

6. Conclusion

Limited by the lack of paired modality samples and the significant imaging differences results in image fusion being difficult with small-scale medical data, and the previous anomaly detection and localization studies have only considered single-modal settings. To this end, in this paper, we focused on the multi-modal MADL task for retinal RAO diseases with the fundus image modality and the OCT modality. The MMR dataset was first constructed, which supports both anomaly localization at the pixel level and anomaly detection at the case level. Furthermore, we embraced the recent SAM visual foundation model and adapted SAM for the MADL task, due to its powerful generalization ability and the small-scale medical data. With the designed modality-shared decoder and task-specific tokens, SAM has been given the ability to process multi-modal images. The developed anomaly simulation and low-level prompt-tuning strategies achieve the fine-tuning process in a pseudo-supervised manner and keep the heavy backbone weights frozen. The adapted model (MMRAD) showed an outstanding anomaly diagnosis ability, with a large margin over the previous state-of-the-art model (at least 3 points in AUC score).

However, there are still some limitations. Although MMRAD performs at a similar level on most specific symptoms, each modality has its own challenging symptom (e.g., retinal hemorrhage in the fundus modality and retinal pigment epithelial detachment in the OCT modality). Incorporating more expert prior knowledge into the anomaly simulation process may be a possible solution for this. Currently, medical anomaly detection, especially in the multi-modal case, is largely underexplored, and we hope that this study will assist the community in achieving a practical diagnosis for retinal disease.

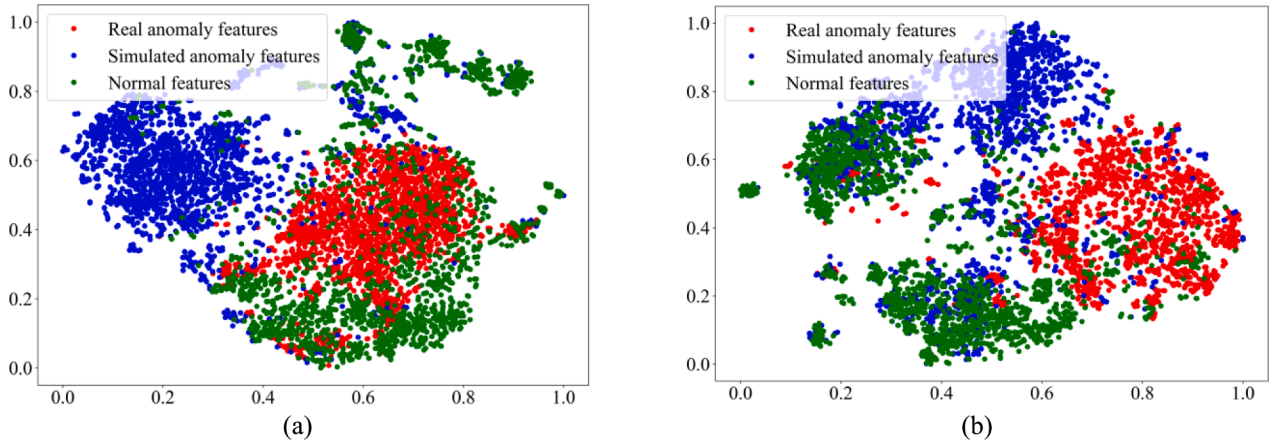

Fig. 9. t-SNE visualization of real anomalies, simulated anomalies, and normal samples for the fundus modality (a) and OCT modality (b). The corresponding embeddings were extracted from the SAM latent space.

Funding

This work was supported by the National Natural Science Foundation of China under Grant No. 42325105, and in part by the Interdisciplinary Innovative Talents Foundation from Renmin Hospital of Wuhan University (JCRCYG-2022-010).

CRediT authorship contribution statement

Jingtao Li: Conceptualization, Methodology, Software, Investigation, Writing – original draft. **Ting Chen:** Conceptualization, Data curation, Writing – original draft. **Xinyu Wang:** Methodology, Validation, Writing – review & editing. **Yanfei Zhong:** Supervision, Funding acquisition, Data curation, Formal analysis. **Xuan Xiao:** Resources, Project administration, Writing – review & editing.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

Data will be made available on request.

Acknowledgments

We would like to thank Yang Yang and Han Cai (Renmin Hospital of Wuhan University, Wuhan University, Wuhan, China) for the assistance in labeling the databases at the pixel level. When constructing the MMR dataset, any disagreements were assessed by kappa text and resolved by discussion and then obtaining consensus from two authors (T.C. and X. X.). The annotation used for the dataset construction came from the best labeled annotation chosen or revised by the two authors (T. C. and X.X.), after comparing all the annotations.

References

[1] Y. Zhou, et al., A foundation model for generalizable disease detection from retinal images, Nature 622 (7981) (2023) 156–163.
[2] W. He, et al., Incremental learning for exudate and hemorrhage segmentation on fundus images, Inform. Fusion 73 (2021) 157–164.
[3] M. Badar, M. Haris, A. Fatima, Application of deep learning for retinal image analysis: A review, Comput. Sci. Rev. 35 (2020) 100203.
[4] J. De Fauw, et al., Clinically applicable deep learning for diagnosis and referral in retinal disease, Nat. Med. 24 (9) (2018) 1342–1350.
[5] V. Emilsson, et al., A proteogenomic signature of age-related macular degeneration in blood, Nat. Commun. 13 (1) (2022) 3401.
[6] C.Y. Cheung, et al., A deep-learning system for the assessment of cardiovascular disease risk via the measurement of retinal-vessel calibre, Nat. Biomed. Eng. 5 (6) (2021) 498–508.
[7] V.B. Gupta, et al., Retinal changes in Alzheimer's disease—integrated prospects of imaging, functional and molecular advances, Prog. Retin. Eye Res. 82 (2021) 100899.
[8] T.K. Yoo, J.Y. Choi, J.G. Seo, B. Ramasubramanian, S. Selvaperumal, D.W. Kim, The possibility of the combination of OCT and fundus images for improving the diagnostic accuracy of deep learning for age-related macular degeneration: a preliminary experiment, Med. Biol. Eng. Comput. 57 (2019) 677–687.
[9] M. Mokhtari, et al., Local comparison of cup to disc ratio in right and left eyes based on fusion of color fundus images and OCT B-scans, Inform. Fusion 51 (2019) 30–41.
[10] T. Schlegl, P. Seeböck, S.M. Waldstein, U. Schmidt-Erfurth, G. Langs, Unsupervised anomaly detection with generative adversarial networks to guide marker discovery, in: International conference on information processing in medical imaging, 2017, pp. 146–157.
[11] T. Fernando, H. Gammulle, S. Denman, S. Sridharan, C. Fookes, Deep learning for medical anomaly detection–a survey, ACM Comput. Surv. (CSUR) 54 (7) (2021) 1–37.
[12] Z. Yuan, H. Chen, C. Luo, D. Peng, MFGAD: Multi-fuzzy granules anomaly detection, Inform. Fusion 95 (2023) 17–25.
[13] K. Ouardini, et al., Towards Practical Unsupervised Anomaly Detection on Retinal Images. MICCAI Workshop on Domain Adaptation and Representation Transfer, 2019, pp. 225–234.
[14] K. Zhou, et al., Proxy-bridged image reconstruction network for anomaly detection in medical images, IEEe Trans. Med. ImAging 41 (3) (2021) 582–594.
[15] Y. Li, et al., Self-supervised anomaly detection, staging and segmentation for retinal images, Med. Image Anal. 87 (2023) 102805.
[16] Y. Tian, et al., Self-supervised pseudo multi-class pre-training for unsupervised anomaly detection and segmentation in medical images, Med. Image Anal. 90 (2023) 102930.
[17] T. Schlegl, P. Seeböck, S.M. Waldstein, G. Langs, U. Schmidt-Erfurth, f-AnoGAN: Fast unsupervised anomaly detection with generative adversarial networks, Med. Image Anal. 54 (2019) 30–44.
[18] J. Zhang, et al., Viral pneumonia screening on chest X-rays using confidence-aware anomaly detection, IEEe Trans. Med. ImAging 40 (3) (2020) 879–890.
[19] S. Lee, et al., Emergency triage of brain computed tomography via anomaly detection with a deep generative model, Nat. Commun. 13 (1) (2022) 4251.
[20] J. Su, H. Shen, L. Peng, D. Hu, Few-shot domain-adaptive anomaly detection for cross-site brain images, IEEe Trans. Pattern. Anal. Mach. Intell. 46 (3) (2021) 1819–1835.
[21] W. Wang, et al., Two-stream CNN with loose pair training for multi-modal AMD categorization, in: Medical Image Computing and Computer Assisted Intervention–MICCAI2019: 22nd International Conference, Shenzhen, China, October 13–17, 2019, Proceedings, Part I 22, Springer, 2019, pp. 156–164.
[22] A. Kirillov, et al., Segment anything, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2023, pp. 4015–4026.
[23] M.A. Mazurowski, H. Dong, H. Gu, J. Yang, N. Konz, Y. Zhang, Segment anything model for medical image analysis: an experimental study, Med. Image Anal. 89 (2023) 102918.
[24] Y. Tian, et al., Constrained contrastive distribution learning for unsupervised anomaly detection and localisation in medical images, in: Medical Image Computing and Computer Assisted Intervention–MICCAI 2021: 24th International Conference, Strasbourg, France, September 27–October 1, 2021, Proceedings, Part V 24, Springer, 2021, pp. 128–140.
[25] S.D. Das, S. Dutta, N.A. Shah, D. Mahapatra, Z. Ge, Anomaly detection in retinal images using multi-scale deep feature sparse coding, in: 2022 IEEE 19th International Symposium on Biomedical Imaging (ISBI), IEEE, 2022, pp. 1–5.
[26] L. Chen, Z. You, N. Zhang, J. Xi, X. Le, Utrad: Anomaly detection and localization with u-transformer, Neural Networks 147 (2022) 53–62.
[27] N. Madan, et al., Self-supervised masked convolutional transformer block for anomaly detection, IEEe Trans. Pattern. Anal. Mach. Intell. (2023).
[28] Y. Fei, C. Huang, C. Jinkun, M. Li, Y. Zhang, C. Lu, Attribute restoration framework for anomaly detection, IEEe Trans. Multimedia (2020).
[29] R. Siddalingappa, S. Kanagaraj, Anomaly detection on medical images using autoencoder and convolutional neural network, Int. J. Adv. Comput. Sci. Appl. (7) (2021).
[30] G. Pang, C. Shen, L. Cao, A. Van Den Hengel, Deep learning for anomaly detection: A review, ACM Comput. Surv. (CSUR) 54 (2) (2021) 1–38.
[31] M.R. Hosseinzadeh Taher, F. Haghighi, R. Feng, M.B. Gotway, J. Liang, A systematic benchmarking analysis of transfer learning for medical image analysis, in: Domain Adaptation and Representation Transfer, and Affordable Healthcare and AI for Resource Diverse Global Health: Third MICCAI Workshop, DART 2021, and First MICCAI Workshop, FAIR 2021, Held in Conjunction with MICCAI 2021, Strasbourg, France, September, Springer, 2021, pp. 3–13.
[32] R. Bommasani et al., "On the opportunities and risks of foundation models," *arXiv preprint arXiv:2108.07258*, 2021.
[33] J.D.M.-W.C. Kenton, L.K. Toutanova, BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding, in: Proceedings of NAACL-HLT, 2019, pp. 4171–4186.
[34] A. Radford, K. Narasimhan, T. Salimans, and I. Sutskever, "Improving language understanding by generative pre-training," 2018.
[35] A. Radford, et al., Learning transferable visual models from natural language supervision, in: International conference on machine learning, 2021, pp. 8748–8763.
[36] L.H. Li, et al., Grounded language-image pre-training, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2022, pp. 10965–10975.
[37] S. Liu et al., "Grounding dino: Marrying dino with grounded pre-training for open-set object detection," *arXiv preprint arXiv:2303.05499*, 2023.
[38] T. Chen, et al., Sam-adapter: Adapting segment anything in underperformed scenes, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2023, pp. 3367–3375.
[39] W. Liu, X. Shen, C.-M. Pun, X. Cun, Explicit visual prompting for low-level structure segmentations, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2023, pp. 19434–19445.
[40] C. Hu and X. Li, "When sam meets medical images: An investigation of segment anything model (sam) on multi-phase liver tumor segmentation," *arXiv preprint arXiv:2304.08506*, 2023.
[41] A. Dosovitskiy, et al., An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale, in: International Conference on Learning Representations, 2020.
[42] P. Gu, Y. Zhang, C. Wang, D.Z. Chen, ConvFormer: Combining CNN and Transformer for Medical Image Segmentation, in: 2023 IEEE 20th International Symposium on Biomedical Imaging (ISBI), IEEE, 2023, pp. 1–5.
[43] A. Vaswani, et al., Attention is all you need, Adv. Neural Inf. Process. Syst. 30 (2017).

[44] J. Li, X. Wang, S. Wang, H. Zhao, L. Zhang, and Y. Zhong, "One-Step Detection Paradigm for Hyperspectral Anomaly Detection via Spectral Deviation Relationship Learning," *arXiv preprint arXiv:2303.12342*, 2023.
[45] M. Salehi, H. Mirzaei, D. Hendrycks, Y. Li, M.H. Rohban, M. Sabokrou, A unified survey on anomaly, novelty, open-set, and out of-distribution detection: solutions and future challenges, Trans. Mach. Learn. Res. (234) (2022).
[46] A. Bulat, R. Guerrero, B. Martinez, G. Tzimiropoulos, FS-DETR: Few-Shot DEtection TRansformer with prompting and without re-training, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2023, pp. 11793–11802.
[47] X. Nie, et al., Pro-tuning: Unified prompt tuning for vision tasks, IEEE Trans. Circuits Syst. Video Technol. (2023).
[48] F. Yuan, Z. Zhang, Z. Fang, An effective CNN and Transformer complementary network for medical image segmentation, Pattern. Recognit. 136 (2023) 109228.
[49] P. Porwal, et al., Idrid: Diabetic retinopathy–segmentation and grading challenge, Med. Image Anal. 59 (2020) 101561.
[50] K. Roth, L. Pemula, J. Zepeda, B. Schölkopf, T. Brox, P. Gehler, Towards total recall in industrial anomaly detection, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2022, pp. 14318–14328.
[51] T. Defard, A. Setkov, A. Loesch, R. Audigier, Padim: a patch distribution modeling framework for anomaly detection and localization, in: International Conference on Pattern Recognition, Springer, 2021, pp. 475–489.
[52] Z. You, et al., A Unified Model for Multi-class Anomaly Detection, Adv. Neural Inf. Process. Syst. 35 (2022) 4571–4584.
[53] J. Bao, H. Sun, H. Deng, Y. He, Z. Zhang, and X. Li, "BMAD: benchmarks for medical anomaly detection," *arXiv preprint arXiv:2306.11876*, 2023.
[54] P. Bergmann, M. Fauser, D. Sattlegger, C. Steger, Uninformed students: Student-teacher anomaly detection with discriminative latent embeddings, in: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 2020, pp. 4183–4192.
[55] I. Gulrajani, F. Ahmed, M. Arjovsky, V. Dumoulin, A.C. Courville, Improved training of wasserstein gans, Adv. Neural Inf. Process. Syst. 30 (2017).
[56] M. Tan, Q. Le, Efficientnetv2: Smaller models and faster training, in: International conference on machine learning, PMLR, 2021, pp. 10096–10106.
[57] F. Pedregosa, et al., Scikit-learn: Machine learning in Python, J. Mach. Learn. Res. 12 (2011) 2825–2830.
[58] I. Loshchilov, F. Hutter, Decoupled Weight Decay Regularization, in: International Conference on Learning Representations, 2018.
[59] V. Zavrtanik, M. Kristan, D. Skočaj, Reconstruction by inpainting for visual anomaly detection, Pattern. Recognit. 112 (2021) 107706.
[60] A. Buslaev, V.I. Iglovikov, E. Khvedchenya, A. Parinov, M. Druzhinin, A.A. Kalinin, Albumentations: Fast and Flexible Image Augmentations, Information 11 (2) (2020).
[61] S. Chen, et al., Adaptformer: Adapting vision transformers for scalable visual recognition, Adv. Neural Inf. Process. Syst. 35 (2022) 16664–16678.
[62] M. Yeung, E. Sala, C.-B. Schönlieb, L. Rundo, Unified focal loss: Generalising dice and cross entropy-based losses to handle class imbalanced medical image segmentation, Comput. Med. Imaging Gr. 95 (2022) 102026.
[63] C.H. Sudre, W. Li, T. Vercauteren, S. Ourselin, M.J. Cardoso, Generalised dice overlap as a deep learning loss function for highly unbalanced segmentations, in: Deep learning in medical image analysis and multimodal learning for clinical decision support: Third International Workshop, 2017, pp. 240–248.
[64] M. Berman, A.R. Triki, M.B. Blaschko, The lovász-softmax loss: A tractable surrogate for the optimization of the intersection-over-union measure in neural networks, in: Proceedings of the IEEE conference on computer vision and pattern recognition, 2018, pp. 4413–4421.
[65] J. Hu, Y. Chen, Z. Yi, Automated segmentation of macular edema in OCT using deep neural networks, Med. Image Anal. 55 (2019) 216–227.
[66] Y. LeCun, L. Bottou, G.B. Orr, K.-R. Müller, Efficient backprop. Neural networks: Tricks of the trade, Springer, 2002, pp. 9–50.
[67] K. He, X. Zhang, S. Ren, J. Sun, Delving deep into rectifiers: Surpassing human-level performance on imagenet classification, in: Proceedings of the IEEE international conference on computer vision, 2015, pp. 1026–1034.
[68] L. Van der Maaten, G. Hinton, Visualizing data using t-SNE, J. Mach. Learn. Res. 9 (11) (2008).