Medical Image Analysis 87 (2023) 102805

Contents lists available at ScienceDirect

Medical Image Analysis

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

Self-supervised anomaly detection, staging and segmentation for retinal images

Yiyue Li [a,c], Qicheng Lao [b,e,*], Qingbo Kang [c,e], Zekun Jiang [c], Shiyi Du [c], Shaoting Zhang [e], Kang Li [c,d,e,**]

[a] Department of Ophthalmology and West China Biomedical Big Data Center, West China Hospital, Sichuan University, Chengdu, Sichuan, 610041, China
[b] School of Artificial Intelligence, Beijing University of Posts and Telecommunications, Beijing, 100876, China
[c] West China Biomedical Big Data Center, Med-X Center for Informatics, Sichuan University, Chengdu, Sichuan, 610041, China
[d] Sichuan University Pittsburgh Institute, Chengdu, Sichuan, 610065, China
[e] Shanghai Artificial Intelligence Laboratory, Shanghai, 200030, China

ARTICLE INFO

Keywords:
Anomaly detection
Anomaly staging
Anomaly segmentation
Retinal images

ABSTRACT

Unsupervised anomaly detection (UAD) is to detect anomalies through learning the distribution of normal data without labels and therefore has a wide application in medical images by alleviating the burden of collecting annotated medical data. Current UAD methods mostly learn the normal data by the reconstruction of the original input, but often lack the consideration of any prior information that has semantic meanings. In this paper, we first propose a universal unsupervised anomaly detection framework *SSL-AnoVAE*, which utilizes a self-supervised learning (SSL) module for providing more fine-grained semantics depending on the to-be detected anomalies in the retinal images. We also explore the relationship between the data transformation adopted in the SSL module and the quality of anomaly detection for retinal images. Moreover, to take full advantage of the proposed SSL-AnoVAE and apply towards clinical usages for computer-aided diagnosis of retinal-related diseases, we further propose to stage and segment the anomalies in retinal images detected by SSL-AnoVAE in an unsupervised manner. Experimental results demonstrate the effectiveness of our proposed method for unsupervised anomaly detection, staging and segmentation on both retinal optical coherence tomography images and color fundus photograph images.

1. Introduction

Eye diseases are the most common cause of vision impairment and loss, with more than 300 million people worldwide suffering from various eye diseases, such as diabetic retinopathy (DR), age-related macular degeneration (AMD) and choroidal Neovascularization (CNV) (Alqudah, 2020). Optical coherence tomography (OCT) B-scan images and color fundus photograph (CFP) are the two most widely used methods for eye examination due to their numerous advantages such as non-invasive, fast, and highly reproducible (Trichonas and Kaiser, 2014). The OCT imaging can outline multiple layers of the retina and visualize the structural changes of the retina in a large range while the CFP images can display en face information of color fundus (Li et al., 2021). However, it still remains a big challenge for anomaly detection algorithms to accurately detect the abnormal images from a large amount of normal OCT and CFP images accumulated through routine clinical examinations.

Unsupervised anomaly detection (UAD) is typically based on the reconstruction of normal data to identify abnormal, novel, or invisible data (Zimmerer et al., 2018; Chen and Konukoglu, 2018; Schlegl et al., 2019). It aims to learn the normal data distribution, and therefore only normal images are used for training. In the testing phase, the test images that have high reconstruction error are treated as detected anomalies. UAD can be well suited for applications in medical images where normal data is relatively easy to collect and no annotation is required, which can solve the lack of label problem in many diseases and avoid the time-consuming and labor-intensive labeling processing (Mahapatra et al., 2021). Furthermore, normal (healthy) medical images often have identical anatomical structures, texture features, and color distributions, which are beneficial for learning the same shared distribution on the normal data. For example, in the retinal-related images, the CFP retinal images share similar blood vessel texture and color distribution, while the OCT retinal images have similar layer

* Corresponding author at: School of Artificial Intelligence, Beijing University of Posts and Telecommunications, Beijing, 100876, China.
** Corresponding author at: West China Biomedical Big Data Center, Med-X Center for Informatics, Sichuan University, Chengdu, Sichuan, 610041, China.
E-mail addresses: qicheng.lao@bupt.edu.cn (Q. Lao), likang@wchscu.cn (K. Li).

https://doi.org/10.1016/j.media.2023.102805
Received 5 May 2022; Received in revised form 14 November 2022; Accepted 30 March 2023
Available online 11 April 2023

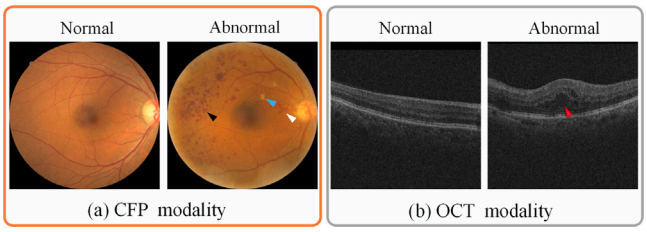

Fig. 1. The two common retinal imaging modalities: CFP and OCT retinal images. In the abnormal CFP image, the black arrowhead represents hemorrhage; the white arrowhead denotes neovascularization; and the blue arrowhead is soft exudates. On the right, the anomaly of the OCT image is the retinal edema region (red arrowhead).

structure (Zhou et al., 2020). Conversely, the abnormal images typically follow a significantly different distribution in the lesion regions, and as a result, they cannot be fitted into the normal data distribution. For instance, in Fig. 1, the edema regions can destroy the layer structure of the OCT images; and for the CFP images, the color distribution (e.g., soft exudate) and blood vessel structure (e.g., neovascularization and hemorrhage) of certain areas can be disturbed by the corresponding anomalies.

Most of the current existing UAD methods directly feed the original normal images to the image reconstruction model without considering any prior information that can provide semantic meanings. In practice, however, when a professional doctor diagnoses the retinal lesions, all kinds of information such as texture, structural features, and changes in the color regions are used in the diagnostic criteria. To specifically overcome such limitation, P-Net (Zhou et al., 2020) has been proposed where the structures in the retinal images are pre-extracted as prior information to improve the subsequent anomaly detection performance. However, we argue that there are still several limitations. First, the previous method cannot be universally applied to all different types of images, e.g., the OCT and CFP images, since the structure information can be different among different modalities and new structure extraction networks are required before the method can be used for the new problem. Moreover, the previous method only considers the structural information as prior information that may be suitable for only certain types of anomalies while ignoring other likewise important prior information, such as the texture, color-related information, etc.

In this work, we first propose a universal unsupervised anomaly detection framework *SSL-AnoVAE* where we introduce a self-supervised learning (SSL) module to participate in anomaly detection. The SSL module, trained with 'free' labels from the transformations of the raw images without any manual annotations, can provide more useful semantic features (e.g., texture, structure, and color-related features) as prior information for better image reconstruction, since the 'free' labels can represent various colors, structures, and contextual information of the images (Koohbanani et al., 2021). Moreover, the proposed framework is also flexible and can be applied to any anomaly detection in different modalities, since the SSL module can utilize different 'free' labels to extract their feature information with different semantic meanings depending on the to-be detected anomalies. Furthermore, we also explore the strategy of how the SSL module is integrated into anomaly detection, where we find that the concatenation of both sampled representations through variational auto-encoders (VAE) from the SSL module and the original image encoder gives the best anomaly detection performance for retinal images.

Finally, to take full advantage of the proposed SSL-AnoVAE and apply towards clinical usages for computer-aided diagnosis of retinal-related diseases, we further propose anomaly staging and segmentation methods based on the anomalies detected by SSL-AnoVAE, i.e., to classify the abnormal retinal CFP images into different severe stages based on their clusters and anomaly scores, and to segment the edema regions of the OCT images with a proposed layer-wise grey scale comparison method. We evaluate the proposed anomaly detection, staging and segmentation methods on two public available datasets for retinal images, and experimental results demonstrate that the proposed methods achieve the state-of-the-art performance. In sum, our main contributions are as follows:

- We propose a novel anomaly detection framework SSL-AnoVAE where we employ a self-supervised learning module to obtain more prior semantic features, which are then concatenated with the representation from the original encoder for better image reconstruction. The proposed SSL-AnoVAE achieves state-of-the-art performance on unsupervised anomaly detection, significantly exceeding the best existing methods.
- We explore the relationship between the data transformations used in the SSL module and the anomaly detection, which can be flexibly adjusted depending on the to-be detected anomalies in different modalities, e.g., the OCT and CFP images. To the best of our knowledge, this is the first study showing how data transformation can help unsupervised anomaly detection.
- An unsupervised anomaly staging method is introduced based on the residual clustering and anomaly score from the trained SSL-AnoVAE model, which is of great help to understand the severity or progression of the retinal disease.
- A label-free anomaly segmentation method is proposed based on layer-wise grey scale comparisons between the original image and the reconstructed image from the trained SSL-AnoVAE model, which greatly improves the performance of traditional unsupervised anomaly segmentation.

2. Related work

2.1. Unsupervised Anomaly Detection (UAD)

Unsupervised anomaly detection (UAD) is to find outliers that are not in the normal data distribution. Many UAD methods have been proposed, including density estimation-based methods (Yang et al., 2009; Kim and Scott, 2012), clustering-based methods (He et al., 2003), and one-class support vector machine methods (Shyu et al., 2003), etc. Recently, due to the excellent performance of deep neural networks (DNNs), more works based on DNNs have been introduced to the UAD fields. Among them, reconstruction-based methods (Zhou and Paffenroth, 2017; Zimmerer et al., 2018; Schlegl et al., 2019) detect abnormal images by reconstruction errors, and treat images that have high reconstruction error as detected anomalies. Zimmerer et al. (2018) proposed a context-encoding variational auto-encoder, which combines context encoding with VAE, adding density-based anomaly scoring to anomaly detection in medical images. Zong et al. (2018) proposed DAGMM using a compression network as auto-encoder to obtain latent representation and reconstruction error, and then feed them into Gaussian mixture models to detect anomalies in the estimation network. Besides, Schlegl et al. (2017) proposed AnoGAN using the GAN framework (Goodfellow et al., 2014) to learn latent representation distribution of normal data while the unfitted latent representations are then distinguished as the anomalies in test stage. Chen et al. (2020c) proposed an auto-encoder based anomaly detection network called MAMA (multi-scale attention memory with hash addressing), combining pixel patch attention and channel attention layer which is easy to participate in any network.

However, all the above-mentioned approaches have one common drawback, i.e., only the original images are utilized during reconstruction while the prior information that is helpful for anomaly detection is not well exploited. Zhou et al. (2020) proposed to use the prior information of image structure but ignored other important prior information such as image texture, color-related, etc., and thus the method is only effective for detecting structure-related diseases, and cannot cover other feature-related diseases. Note that although Zhou et al. (2020) mentioned texture information, it is immersed in the image reconstruction branch but not highlighted as prior information.

2.2. Self-supervised learning

The self-supervised learning methods of learning deep features can be divided into two categories (Wang et al., 2021). The first one is to learn the deep features by aligning to a target task between inputs and self-defined signals (Schlegl et al., 2017; Komodakis and Gidaris, 2018). For example, in medical imaging field, Li et al. (2020) presented a self-supervised learning method by effectively exploiting multi-modal data for retinal disease diagnosis, and learning deep features by aligning the feature-based softmax embedding objective. The other category is contrastive leaning (Komodakis and Gidaris, 2018; Chen et al., 2020a; Grill et al., 2020), i.e., learning deep features by minimizing the distance between transformed images of same image with two contrastive networks. For example, Azizi et al. (2021) first applied contrastive learning on natural images, and then adopted multi-instance contrastive learning on medical images. Nevertheless, their deep feature learning stage is unsupervised but the label supervision is still required in the final classification stage.

For the anomaly detection task, Wang et al. (2021) used the SSL method to constrain a DNN in training stage, and then treated the output of the DNN network as anomaly scores to detect anomalies. However, only the deep feature information from the self-supervised learning network is used as anomaly scores. Zhao et al. (2021) proposed SALAD, which is based on extracting translation-consistent features both from the original image and latent self-supervised spaces. SALAD helps to improve the robustness of the model, but only considering the consistent features, which may lose some other crucial features from image and the SSL feature space.

Many contrastive learning SSL methods utilize data transformations to learn deep features (He et al., 2020; Chen et al., 2020a; Chen and He, 2021). They train the SSL network by making transformed images from the same group closer and those from different groups farther. But for these methods, many abnormal images (Arora et al., 2019) and large batches (Chen et al., 2020a; Tian et al., 2020) are required. Grill et al. (2020) proposed a contrastive learning SSL method BYOL that only uses normal images, which does not rely on abnormal pairs and is more robust. In other works, Tack et al. (2020) and Yoa et al. (2021) used data transformations with a self-supervised learning model to tackle the image-level anomaly detection task. Nevertheless, they have not considered the relationship between data transformation and lesion detection.

2.3. Anomaly staging and segmentation

Most anomaly detection works generally either classify anomalies (e.g., SLA^2P (Wang et al., 2021), $E^3Outlier$ (Wang et al., 2019), etc.) or segment anomalies (Hansen et al., 2022), for example, Seeböck et al. (2019) first trained a Bayesian U-Net network with layer labels in a weakly supervised manner, and then added epistemic uncertainty in the anomaly detection stage to obtain OCT segmentation results. But less attention has been paid to concatenating the processes of detection and segmentation into a UAD model. Fortunately, in recent years, some of the reconstruction-based UAD tasks in medical images are not only simple detection, but also segmentation of abnormal regions (Yao et al., 2021; Chen et al., 2021), which are widely used for brain MR images (Chen et al., 2020b). For instance, Baur et al. (2018) proposed AnoVAEGAN, which adds a discriminator to the output of the VAE framework to train the decoder, and finally uses the residual between the reconstructed image and the original image as the segmented abnormal regions. Moreover, Baur et al. (2021) also compared the segmentation performance of various anomaly detection methods for brain MR images by using residuals. However, this reconstruction-based segmentation method utilizing pixel-level and patch-level residuals cannot achieve satisfactory results when applied to retinal OCT images, especially for the segmentation of retinal edema. Moreover, the above methods only use anomaly detection for anomaly segmentation, but few explore the staging of abnormal images or to distinguish the disease severity.

3. Methodology

In order to obtain more semantic prior information (e.g., texture, color-related, etc.) to improve the performance of anomaly detection and allow the network to be flexibly applied to different modalities, this paper proposes a universal anomaly detection method *SSL-AnoVAE*, which introduces a self-supervised learning module into the anomaly detection module (Fig. 2(A)). The SSL module can act as a prior and provide more fine-grained semantics depending on the to-be detected anomalies, such as structure anomalies in optical coherence tomography (OCT) images, or color anomalies in color fundus photograph (CFP) images.

Meanwhile, to take full advantage of the proposed SSL-AnoVAE and apply towards clinical usages for computer-aided diagnosis of retinal-related diseases, we further propose to classify and segment the anomalies in an unsupervised manner, i.e., classify the abnormal CFP images into different stages based on their clusters and anomaly scores (Fig. 2(B)), and segment the edema regions of OCT images with layer-wise grey scale comparisons (Fig. 2(C)).

3.1. The proposed SSL-AnoVAE

3.1.1. The self-Supervised Learning module of SSL-AnoVAE

To obtain the prior information from normal images with self-supervised learning, we adopt a contrastive network, which can learn specific information by choosing selected image transformations as inputs to the network. Inspired by Grill et al. (2020), we design the contrastive network in our SSL module with two unbalanced branches: online branch and target branch. As such, the network can learn similar distribution information of normal images by maximizing the similarity between outputs from these two branches. Note that the inputs of the SSL module in this work are all positive pairs (i.e., normal retinal image), which is beneficial for the network to learn the shared features of normal retinal images, consistent with learning the distribution of normal data for UAD. Moreover, it has been shown that the self-supervised learning network with only inputs of positive pairs is more robust to perturbation than that with negative pairs for the image transformations (Grill et al., 2020).

As shown in Fig. 2, given an input image x, we obtain two transformed images $T_1(x)$ and $T_2(x)$ by using selected transformations T_1 and T_2 (will be detailed later in Section 3.1.2). Then $T_1(x)$ and $T_2(x)$ are passed through encoders E_{S_θ}, E_{S_ξ}, and projectors P_{j_θ}, P_{j_ξ}, respectively to get the outputs h_θ and h_ξ. Different to the target branch, additional predictor P_{d_θ} is used in the online branch for the unbalanced output h'_θ. Following Grill et al. (2020), we also use mean squared error to measure the difference between the two ℓ_2-normalized outputs $\overline{h'_\theta}$ and $\overline{h_\xi}$:

$$\mathcal{L}_{\theta,\xi} = \|\overline{h'_\theta} - \overline{h_\xi}\|_2^2 = 2 - 2 \cdot \frac{\langle h'_\theta, h_\xi \rangle}{\|h'_\theta\|_2 \cdot \|h_\xi\|_2}. \tag{1}$$

For the simplicity of notation, we denote $F(\cdot) = P_{d_\theta}(P_{j_\theta}(E_{S_\theta}(\cdot)))$ for the online branch and $G(\cdot) = P_{j_\xi}(E_{S_\xi}(\cdot))$ for the target branch. Then the loss $\mathcal{L}_{\theta,\xi}$ in Eq. (1) can also be written as:

$$\mathcal{L}_{\theta,\xi} = \|\overline{F}(T_1(x)) - \overline{G}(T_2(x))\|_2^2, \tag{2}$$

where the overline operation denotes the ℓ_2-normalization operation. By symmetrizing $\mathcal{L}_{\theta,\xi}$, the total loss for the self-supervised learning module $\mathcal{L}_{ssl}$ is given by:

$$\mathcal{L}_{ssl} = \|\overline{F}(T_1(x)) - \overline{G}(T_2(x))\|_2^2 + \|\overline{F}(T_2(x)) - \overline{G}(T_1(x))\|_2^2, \tag{3}$$

where the first term is $\mathcal{L}_{\theta,\xi}$ in Eq. (2) and the second term is its symmetric loss by separately feeding $T_2(x)$ to the online branch and $T_1(x)$ to the target branch. Note that we only optimize θ in the online branch, and apply a stop-gradient to ξ in the target branch.

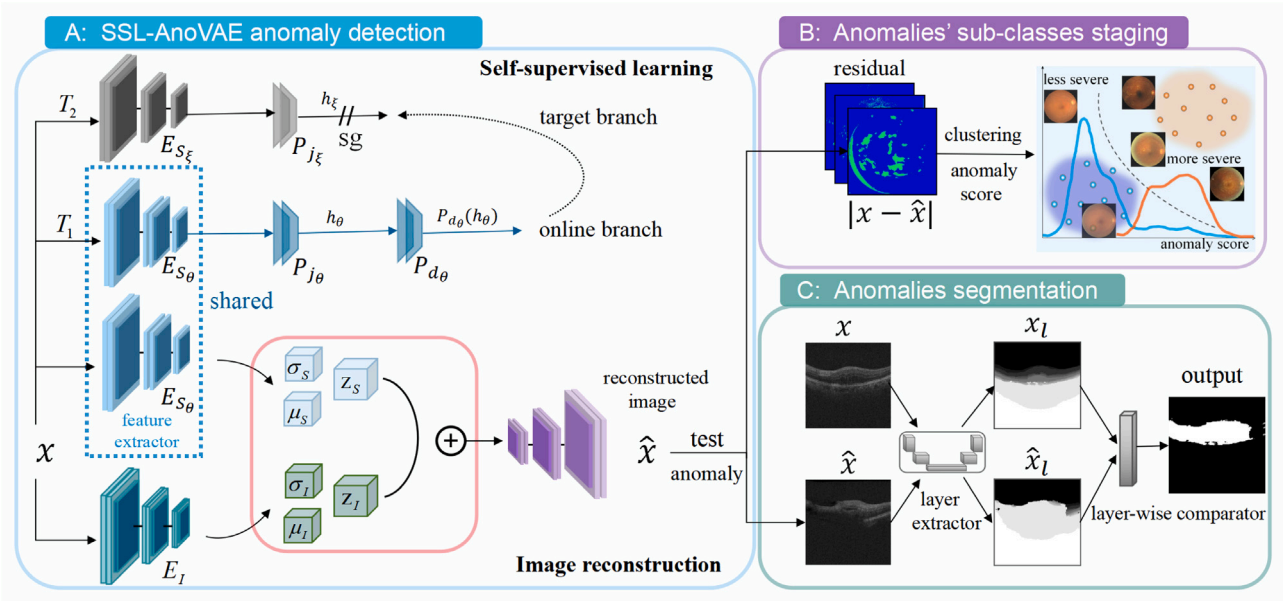

Fig. 2. Schematic diagram of our proposed method consisting of three parts: (A) SSL-AnoVAE anomaly detection, (B) anomalies' sub-classes staging, (C) anomalies segmentation. The proposed SSL-AnoVAE consists of a self-supervised learning (SSL) module and an image reconstruction module. In the training of SSL-AnoVAE, the SSL module trains E_{S_θ} to extract features, and shares online encoder E_{S_θ} with the image reconstruction module. Then we concatenate the two latent representations z_I and z_S, from the shared online encoder E_{S_θ} and the original image encoder E_I for image reconstruction. In the anomalies' sub-classes staging part, the CFP abnormal image is fed into the trained SSL-AnoVAE, and the output residuals $|x - \hat{x}|$ are clustered into K clusters, e.g., $K = 2$ for less severe and more severe. In the anomaly segmentation part, we first input the reconstructed image $\hat{x}$ and original image x to the layer extractor and then introduce a layer-wise grey scale comparator, which outputs the segmentation map of retinal edema regions in OCT image (more details are provided in Fig. 3).

3.1.2. Choices of transformations

For learning good representations via self-supervised learning, the choices of data transformation are critical (Chen et al., 2020a). Common data transformations can be mainly divided into two types: spatial/geometric transformations such as cropping, flipping and rotation (Komodakis and Gidaris, 2018), and appearance transformations, such as color-related changes (e.g., color jitter, color drop and Canny filtering), Gaussian blur, etc. The contrastive network can extract various features of interest by minimizing the distance between images with different selected transformations. For example, it has been shown that with color-irrelevant appearance transformations, the contrastive network mostly learns the color representation (Chen et al., 2020a). Similarly, when we only use color-related appearance transformations, the contrastive network will pay more attention to the invariant structural information (e,g., blood vessels, retinal layers) while paying less attention to the varying color information.

As shown in Table 1, we adopt different transformations on the retinal images for the SSL module in this work. Specifically, the data transformations of the OCT and CFP datasets are selected depending on the to-be detected anomalies. The OCT retinal images are mainly composed of layers, whose structure can be destroyed by abnormal regions (lesions). Thus, we apply appearance transformations on the OCT images to provide more structural and textural information. In addition, since some diseases (e.g., retinal edema) may also cause intensity changes around the lesions, we add spatial/geometric transformations as supplement. For the CFP dataset, the abnormal regions of CFP are mostly changes in the color (e.g., soft exudate) and structure (e.g., neovascularization and hemorrhage) of certain areas. Therefore the chosen data transformations for the CFP dataset aim to provide more color and structural information. We will show later in our experiments that the selected transformations are optimal choices.

3.1.3. The image reconstruction module of SSL-AnoVAE

For better reconstruction of normal images, we combine two latent representations for the image decoding: one from the feature extractor E_{S_θ} in the SSL module and the other from the original encoder E_I. By doing so, both the original image information and the prior information from the SSL module are integrated. Formally, we first encode the original image x into two latent representations z_I, z_S through original encoder E_I and the shared encoder E_{S_θ} from the SSL module. Then the VAE (Larsen et al., 2016) is introduced to regularize the E_I and E_{S_θ} by matching the latent distributions $p(z_I)$, $p(z_S)$ to a prior distribution:

$$\mathcal{L}_{prior} = \lambda_I D_{KL}(p(z_I)\|N(0,1)) + \lambda_S D_{KL}(p(z_S)\|N(0,1)), \tag{4}$$

where D_{KL} is the Kullback–Leibler divergence.

Finally, we concatenate these two latent representations z_I and z_S as input to a decoder for the reconstructed image $\hat{x}$. Following Isola et al. (2017), Akcay et al. (2018), Zhou et al. (2020), the ℓ_1 norm is used for the reconstruction loss:

$$\mathcal{L}_{rec} = \|x - \hat{x}\|_1. \tag{5}$$

Moreover, to further match the distributions of the original images and the reconstructed images, we use PatchGAN (Isola et al., 2017) to constrain the difference between the reconstructed image $\hat{x}$ and the original image x. Concretely, we introduce an adversarial D to compete with the whole image reconstruction module, and the adversarial loss L_{adv} can be expressed as:

$$\mathcal{L}_{adv} = \mathbb{E}[log(1 - D(\hat{x}))] + \mathbb{E}[log D(x)]. \tag{6}$$

Table 1
Choices of data transformations on the Davis (CFP) and RESC (OCT) datasets.

	Data transformation				
	Flip	Guassian blur	Color jitter	Color drop	Canny filtering
Davis (CFP)	✓	✓	✓	✓	
RESC (OCT)	✓	✓			✓

3.1.4. The overall loss of SSL-anovae

For the overall framework of SSL-AnoVAE, the SSL module and the image reconstruction module are trained simultaneously in an end-to-end manner. More precisely, after the two transformations of $T1$

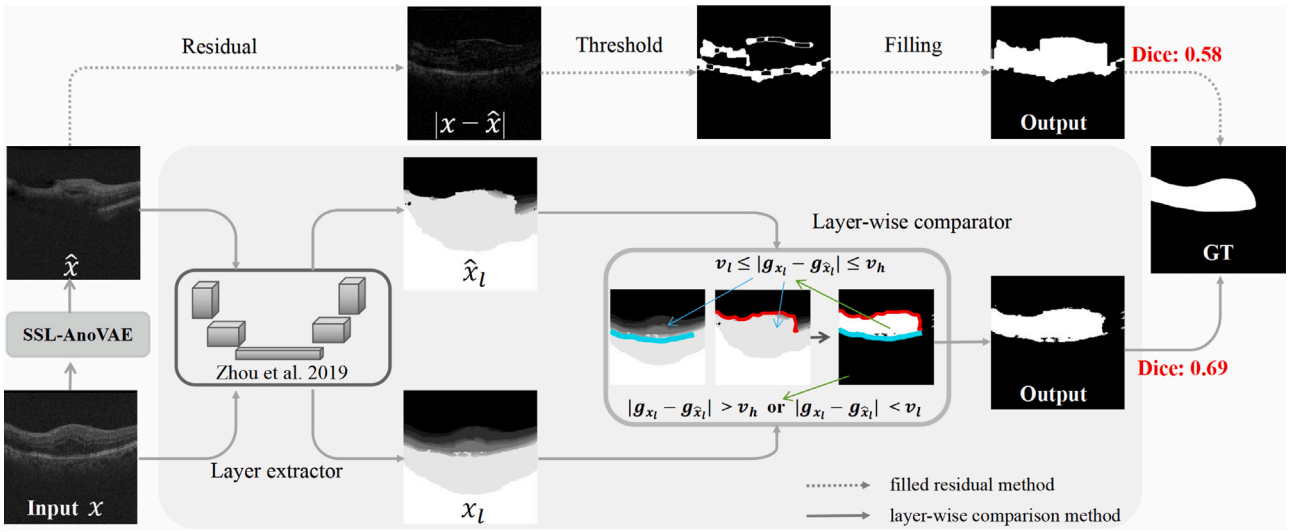

Fig. 3. The proposed unsupervised anomaly segmentation method. Two different methods for the OCT retinal edema segmentation are presented here. The first baseline method (top) uses thresholding, edge detection, and filling operations on the residual $|x - \hat{x}|$, which only yields Dice score of 0.58. Our proposed method (bottom) is through layer-wise grey scale comparison, where we first feed the original image $\boldsymbol{x}$ and the reconstructed image $\hat{\boldsymbol{x}}$ to a layer extractor (Zhou et al., 2020), separately, which then outputs the layer-wise images $\boldsymbol{x}_l$ and $\hat{\boldsymbol{x}}_l$. By layer-wisely comparing their grey scales, the regions whose grey scale difference $(\boldsymbol{g}_{\boldsymbol{x}_l} - \boldsymbol{g}_{\hat{\boldsymbol{x}}_l})$ between $\boldsymbol{x}_l$ and $\hat{\boldsymbol{x}}_l$ is greater than $\boldsymbol{v}_l$ and less than $\boldsymbol{v}_h$ is regarded as the abnormal regions. The proposed method improves the Dice score to 0.69.

and $T2$, the transformed images are input into E_{S_θ} and E_{S_ξ} in the SSL module. At the same time, the original image x is also input into E_{S_θ} and E_I, respectively in the image reconstruction module. Note that the SSL module shares the online encoder E_{S_θ} with the image reconstruction module. Therefore, the similarity loss between the two transformed images can participate in the training of the reconstruction module while constraining the self-supervised learning. As such, the overall loss for the whole training process is:

$$\mathcal{L} = \lambda_1 \mathcal{L}_{ssl} + \lambda_2 \mathcal{L}_{rec} + \lambda_3 \mathcal{L}_{prior} + \lambda_4 \mathcal{L}_{adv}, \tag{7}$$

where λ_1, λ_2, λ_3, λ_4 are the hyper-parameters. At each training step, we perform a stochastic optimization to minimize the loss $\mathcal{L}$ in Eq. (7) with respect to θ only, without updating ξ due to the stop-gradient strategy as shown in Fig. 2. After that, ξ is then updated iteratively based on θ. The dynamics of training the proposed SSL-AnoVAE is shown in the following:

$$\theta \leftarrow optimizer(\theta, \nabla_\theta \mathcal{L}, \eta),$$

$$\xi \leftarrow \tau\xi + (1-\tau)\theta,$$

where *optimizer* denotes an optimizer, η is the learning rate, and τ is the weight hyper-parameter. For the testing, we use reconstruction error as the measurement score of anomaly detection, i.e.anomaly score, which can be expressed as:

$$\mathcal{A}(x) = \|x - \hat{x}\|_1. \tag{8}$$

For the choices of whether to use a pretrained model for the SSL module as opposed to end-to-end training, we also perform comparative experiments, where we first pre-train the online encoder via self-supervised learning and then integrate it in the image reconstruction module for the anomaly detection task. However, as will be shown in the experiments section, our result suggests that the end-to-end training is preferable compared to the pre-training strategy.

3.2. Unsupervised anomaly staging

The abnormal residual from an anomaly detection model, i.e., the differences between the reconstructed and original image, represents the detected lesion regions of the abnormal image (Schlegl et al., 2019; Baur et al., 2021). Generally, the more severe the abnormal image is, the larger lesions it detects. To better assist patients in diagnosis and treatment by identifying disease severity or progression, i.e., the sub-classes of the abnormal images, in this work, we also tackle the problem of anomaly staging in an unsupervised manner, i.e., to classify the residuals extracted from the abnormal images without labels.

One of the most severe lesions in retinal diseases is proliferative diabetic retinopathy (PDR). It is generally recommended that eyes with PDR together with vitreous hemorrhage or fibrovascular proliferative membranes require a vitrectomy surgery (Takahashi et al., 2017). Therefore, to better diagnose and treat retinal disease, we aim to identify PDR from the other retinopathy (i.e., Non-PDR), including simple diabetic retinopathy (SDR) and pre-proliferative retinopathy (PPDR). We show the overall process of anomaly staging for retinal disease in Fig. 2(B). Specifically, we first obtain the abnormal residuals $|x - \hat{x}|$ of the CFP images with the proposed SSL-AnoVAE, and then cluster these residuals of abnormal images into K clusters using K-means clustering, e.g., $K = 2$ in this case for PDR and Non-PDR. To determine the corresponding categories of the two clusters, we assign the anomaly score $\mathcal{A}(x)$ in Eq. (8) to each sample in the clusters, and the centroid of each cluster is calculated for its anomaly score. The cluster with higher anomaly score corresponds to a more severe lesion group (e.g., PDR), and vice versa for the Non-PDR lesions.

3.3. Unsupervised segmentation via layer-wise gray scale comparison

Next, to further locate and delineate the lesion regions towards clinical applications and take full advantage of the proposed SSL-AnoVAE, we also propose an unsupervised anomaly segmentation method via layer-wise grey scale comparison. Previous anomaly segmentation methods based on reconstruction (Baur et al., 2018, 2021) rely on pixel-wise comparisons between the original images and their reconstructions. However, some diseases in the OCT retinal images (e.g., retinal edema) cannot achieve satisfactory segmentation by directly using the pixel-wise segmentation because the tissues in the edema region are normal. For example, as shown in the top of Fig. 3, if we transform the residuals by simply using threshold and filling to segment the abnormal regions of retinal edema, these pixel-wise comparison methods only improve the segmentation results to some extent (Dice: 0.58) while the shape of retinal edema is still not well segmented.

Given the fact that the OCT retinal image is composed of layers and the abnormal regions (lesions) can destroy the layer structure, the abnormal regions (e.g., retinal edema) between the layer structures of the original image and the reconstructed image can be quite different, namely, the abnormal regions can be amplified in layer-wise comparisons. Motivated by this, in our proposed method (shown in the bottom of Fig. 3), a layer extractor (Zhou et al., 2020) is introduced for extracting layers for both the original image x and the reconstructed image $\hat{x}$ from the SSL-AnoVAE. The introduced layer extractor has good quality in layer segmentation since it uses domain adaptation to mitigate the domain differences between our dataset (i.e., RESC dataset) and source segmentation datasets (i.e., Toncon dataset used in Zhou et al. (2020)). Therefore, the layer extractor can segment the layer structure of the original image with the highest fidelity and minimize the layer segmentation error, so that it can be better used for the abnormal regions segmentation task to obtain a good segmentation performance. It can be seen from the figure that the layer structures x_l and $\hat{x}_l$ are more different in the lesion regions than x and $\hat{x}$ because the lesion is enlarged after the layer extractor. Then we propose a grey scale comparator, where the regions whose grey scale difference between x_l and $\hat{x}_l$ greater than v_l and less than v_h are considered as the abnormal regions (i.e., retinal edema), and the rest are the background. The upper limit (i.e., v_h) is set based on the calculation that the grey scale difference among retinal layers typically falls within a certain range, which is due to the grey scale definition of each layer in the layer extractor. Furthermore, the non-retinal layer, e.g., the background of the white area in the layer structure, has a larger grey scale difference compared with other layers. Thus we set an upper limit v_h to exclude the error resulting from the background regions. The final segmentation map obtained by comparing the grey value differences of x_l and $\hat{x}_l$ from the layer extractor, i.e., through the layer-wise grey scale comparison, can be written as:

$$x_{seg} = \begin{cases} 255, & v_l \le |g_{x_l} - g_{\hat{x}_l}| \le v_h \\ 0, & others \end{cases}$$

where g denotes the grey scale value.

As shown in Fig. 3 and will be detailed in Section 5.3, the proposed layer-wise comparison method significantly improves the performance of unsupervised anomaly segmentation. To the best of our knowledge, this is the first time using layer-wise comparison instead of the pixel-wise comparison between the original and reconstructed images for unsupervised anomaly segmentation.

4. Experiments

We first give a description of the CFP and OCT datasets we used in this work. Then we introduce our experimental settings and several UAD baseline methods.

4.1. Datasets

Davis Datasets (Takahashi, 2017) The Davis dataset has 9939 posterior pole photographs from 2740 diabetic patients, which divides diabetic retinopathy into four stages: no diabetic retinopathy (NDR), simple diabetic retinopathy (SDR), pre-proliferative retinopathy (PPDR), and proliferative diabetic retinopathy (PDR). Note that the NDR images are normal images while the rest are abnormal images. Due to the clinical importance of PDR, we consider SDR and PPDR as the Non-PDR stage in this work for anomaly staging. To alleviate the computational burden, we set the size of the CFP images from the original 848 × 848 to 224 × 224. For the training, we use 3888 normal CFP images from 1500 subjects in the NDR stage. For the testing, we randomly selected 1578 images, including 780 NDR images, 276 SDR images, 248 PPDR images, and 274 PDR images, from the rest 1240 subjects (i.e., 776 NDR, 256 SDR, 102 PPDR, and 106 PDR).

Retinal Edema Segmentation Challenge Dataset (RESC)[1] (Hu et al., 2019) The RESC dataset is an extensive OCT dataset provided by the "AI Challenger" competition platform. The dataset is well-annotated at the pixel level and has a resolution of 512 × 1024. It consists of training, validation, and test set, but only the annotations for the training and validation sets are provided. We use the 4296 normal OCT images in the training set for training, and the original validation set for testing which includes a total of 1920 images, of which 1086 images are normal images, and 834 images are abnormal images with retinal edema. In order to alleviate the computational burden, we set the size of the images to 224 × 224. For the unsupervised anomaly segmentation, we validate our method using annotations with 834 retinal edema images from in the validation set.

4.2. Experimental settings

Implementation details. We implement our proposed SSL-AnoVAE with PyTorch. For the encoders in the SSL-AnoVAE, i.e., E_{S_θ}, E_{S_ξ} and E_I, we use the same architecture, and the dimension of the output feature is 512. The projectors P_{j_θ}, P_{j_ξ} and predictor P_{d_θ} also have the same architecture, i.e., multi-layer perceptron (MLP). The model is updated using the ADAM optimizer with weight decay 1e−4, and the batch size is set to 48. We train our model for 600 epochs. The learning rate is 1e−3 for the first 300 epochs, and decreased to 1e−4 for the last 300 epochs, until convergence. Both the self-supervised learning module and the reconstruction module are trained from scratch, which is the same as all baseline methods. For a fair comparison, we perform three independent experiments (with different random seeds) for all methods including the baselines and the average is reported as the final results. For the OCT anomaly segmentation, we use the layer extractor provided by Zhou et al. (2020). We set $v_l = 25$ and $v_h = 205$.

Baselines. In our experiments, several baselines including the state-of-the-art anomaly detection methods are compared to evaluate the effectiveness of the proposed method:

- **Auto-Encoder based**, which employs an encoder for the latent representation encoding and an decoder for reconstructing the original normal images, and the reconstruction error is considered as the detected anomalies.
- **CycleGAN** (Zhu et al., 2017), which involves the mapping between two domains with a cycle consistency regularization. For anomaly detection, the normal images can be translated into a predefined domain (e.g., the extracted structures in P-NET (Zhou et al., 2020)) and then reconstructed back, and the anomaly score is obtained by using reconstruction error between the original and reconstructed images.
- **AnoGAN** (Schlegl et al., 2017), which proposes to use the GAN framework to fit the normal data distribution while the unfitted data is then distinguished as the anomalies.
- **VAE-GAN** (Baur et al., 2018), which combines the VAE framework with GAN, where the latter uses a discriminator to distinguish between the original and reconstructed images. Anomalies are detected by comparing the original and reconstructed images.
- **GANomaly** (Akcay et al., 2018), employs an encoder–decoder-encoder network for learning the image and latent spaces jointly. During inference, the anomaly score is the distance between the latent representation of the input image and the encoded features of the generated image.

[1] The AI Challenger is at the time of writing not publicly available.

Table 2
Performance comparison of different anomaly detection methods.

Method	RESC (OCT)				Davis (CFP)			
	AUC(%)	ACC(%)	SPE(%)	SEN(%)	AUC(%)	ACC(%)	SPE(%)	SEN(%)
Auto-Encoder (Zhou and Paffenroth, 2017)	82.52 ± 1.86	77.92 ± 1.94	80.48 ± 1.21	74.58 ± 1.66	76.49 ± 1.81	71.58 ± 1.04	74.92 ± 1.38	66.98 ± 1.12
CycleGAN (Zhu et al., 2017)	86.37 ± 1.09	82.15 ± 1.08	83.69 ± 0.98	78.30 ± 1.24	83.46 ± 2.07	78.56 ± 1.87	80.98 ± 2.06	75.21 ± 1.59
AnoGAN (Schlegl et al., 2017)	85.81 ± 0.87	81.34 ± 0.69	82.94 ± 0.72	78.01 ± 0.56	82.68 ± 0.56	77.42 ± 0.45	80.02 ± 0.72	74.26 ± 0.54
VAE-GAN (Baur et al., 2018)	90.64 ± 1.13	84.96 ± 1.51	86.13 ± 1.26	82.01 ± 0.93	88.20 ± 1.22	82.60 ± 1.04	84.62 ± 1.07	79.98 ± 0.98
GANomaly (Akcay et al., 2018)	89.96 ± 2.07	83.42 ± 1.86	85.92 ± 2.01	81.69 ± 1.41	86.17 ± 0.82	81.15 ± 1.17	83.33 ± 1.47	78.30 ± 1.23
f-AnoGAN (Schlegl et al., 2019)	90.78 ± 0.86	85.12 ± 0.98	86.34 ± 1.16	82.29 ± 1.09	88.54 ± 0.91	82.34 ± 0.87	84.46 ± 0.91	80.13 ± 0.94
P-NET (Zhou et al., 2020)	93.10 ± 1.22	87.60 ± 1.04	89.04 ± 1.23	85.73 ± 1.14	90.31 ± 0.92	84.31 ± 0.66	85.92 ± 0.89	81.64 ± 0.71
SSL-AnoVAE	**98.04 ± 0.93**	**93.34 ± 0.86**	**94.01 ± 0.92**	**92.30 ± 0.86**	**94.41 ± 0.87**	**89.61 ± 0.76**	**91.02 ± 0.78**	**87.23 ± 0.62**

- **f-AnoGAN** (Schlegl et al., 2019), has two steps in the training phase, i.e., first train a Wasserstein GAN with normal data, and followed by an encoder training which can fast map the image to the latent space of GAN. The calculation of the anomaly score is jointly guided by the discriminator feature residual error and the image reconstruction error.

- **P-NET** (Zhou et al., 2020), which proposes to pre-extract the image structures as prior information and combine it with the original images for better reconstruction of the normal images.

For anomaly segmentation and staging parts, we compared the three baselines (Auto-Encoder, VAE-GAN, P-Net), which are all reconstruction-based anomaly detection methods similar to our method. Therefore, for these baselines, we can get a residual of the original and reconstructed images, and then use the obtained residuals and reconstructed images as input for the segmentation and staging networks for further calculations.

Evaluation indicators. We use a variety of metrics to measure the performance of the proposed anomaly detection method, including AUC (Area Under the Receiver Operating Characteristics), ACC (accuracy), SPE (specificity), SEN (sensitivity). We continuously adjust the threshold of $\mathcal{A}(x)$ to classify normal and abnormal images (Zhou et al., 2020).

5. Results

In this section, we present experimental results and ablation studies on self-supervised anomaly detection, staging and segmentation for retinal images. Firstly, we show quantitative and qualitative results of anomaly detection to validate the effectiveness of our proposed SSL-AnoVAE on the Davis (CFP) and RESC (OCT) datasets. Moreover, we evaluate the proposed anomaly staging and layer-wise anomaly segmentation methods on the two datasets. Finally, we carry out comprehensive ablation studies to justify the rationality for the design of the proposed SSL-AnoVAE.

5.1. The anomaly detection results

5.1.1. The proposed SSL-AnoVAE accurately detects abnormal images, and achieves state-of-the-art detection performance

Table 2 presents in-depth comparisons of the image-level anomaly detection accuracy between the proposed SSL-AnoVAE and other anomaly detection methods on the Davis (CFP) and RESC (OCT) datasets. As explained earlier in the experimental setting, we report the average AUC (%), ACC(%), SPE(%), and SEN(%) of three independent experiments using different random seeds, and seven baselines are included in the table: Auto-Encoder (Zhou and Paffenroth, 2017), CycleGAN (Zhu et al., 2017), AnoGAN (Schlegl et al., 2017), f-AnoGAN (Schlegl et al., 2019), VAE-GAN (Baur et al., 2018), GANomaly (Akcay et al., 2018), and P-Net (Zhou et al., 2020). The table shows that our method achieves the best performance on both datasets and greatly outperforms the current anomaly detection methods by a big margin. For example, on the Davis dataset, it increases the AUC, ACC, SPE, and SEN by 4.1%, 5.3%, 5.1%, and 5.59% respectively, compared to the second-best baselines; on the RESC dataset, it is 4.94%, 5.74%, 4.97%, and 6.57% higher than the second-highest baselines. These results confirm that our proposed method, by providing more semantic information on color and structure, is effective for retinal images of different modalities (i.e., CFP and OCT images).

Furthermore, we compare the performance of pixel-level anomaly localization between SSL-AnoVAE and three other state-of-the-art anomaly detection methods in Fig. 4. As shown in the columns 2 and 3 of the figure, the Auto-Encoder method fails to reconstruct the colors of normal images in some regions (the first row), whereas some normal colored regions are often mistakenly detected as abnormal in the abnormal image (the last two rows). In the columns for the VAE-GAN method, the normal blood vessels and a few normal colored regions (e.g., optical disc) cannot be well reconstructed in the abnormal images, and therefore are incorrectly identified as abnormal, leading to inaccurate detection results. The results from the P-NET method also have issues in the reconstruction of normal blood vessels, and thus the pixel-level localization of abnormal regions is also far from satisfactory. Conversely, our proposed SSL-AnoVAE can not only better reconstruct normal images including the blood vessels and colored regions, but also precisely locate most anomalies in the abnormal image while the normal regions are rarely mistakenly detected as abnormal.

5.1.2. The proposed SSL-AnoVAE improves the anomaly detection performance in all stages

Next, we investigate whether the proposed SSL-AnoVAE can detect all stages of anomalies or only the most severe anomalies. To do so, we analyze the anomaly detection results on the Davis (CFP) dataset where the labels of different anomaly stages/sub-classes are annotated (i.e., SDR, PPDR, and PDR). In Fig. 5, we plot the histograms of the anomaly score $\mathcal{A}(x)$ given by different methods according to Eq. (8) for both normal (NDR) and abnormal images (SDR, PPDR and PDR) from the test set. The corresponding quantitative results of AUC (%) are shown in Table 3. Both results demonstrate that our proposed SSL-AnoVAE significantly improves the anomaly detection performance in all stages. Remarkably, as shown in the table, SSL-AnoVAE improves the AUC of the SDR detection from 55.93% (with the Auto-Encoder method) to 85.44%, and the PPDR detection from 65.92% (with the Auto-Encoder method) to 92.09%. Compared to the second-best baseline, SSL-AnoVAE increases the AUC of the SDR, PPDR, and PDR detections by 4.89%, 9.5%, 1% respectively. In brief, the improvement of the

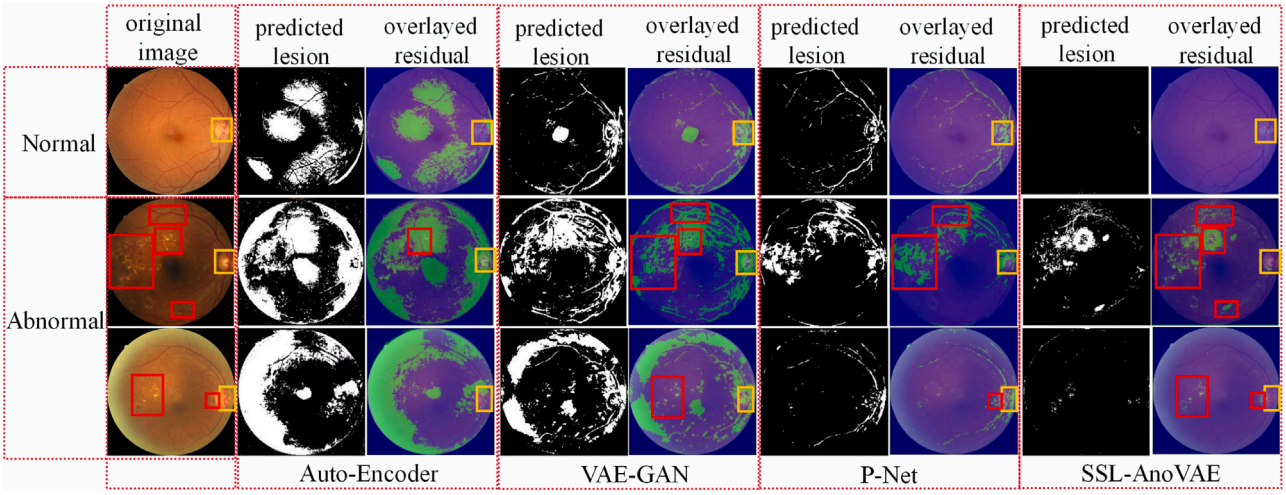

Fig. 4. The comparisons of four state-of-the-art anomaly detection methods for pixel-level localization of abnormal regions. The first row is a representative normal image from healthy cases (NDR) in the test set, and compares the reconstruction ability of different methods. The last two rows are two representative abnormal images in the test set. Notably, to facilitate comparison, the red boxes are used to highlight most lesion regions and do not participate in any calculations. The yellow box is the optical disc, and the predicted lesion is the binary map obtained from the residual after thresholding.

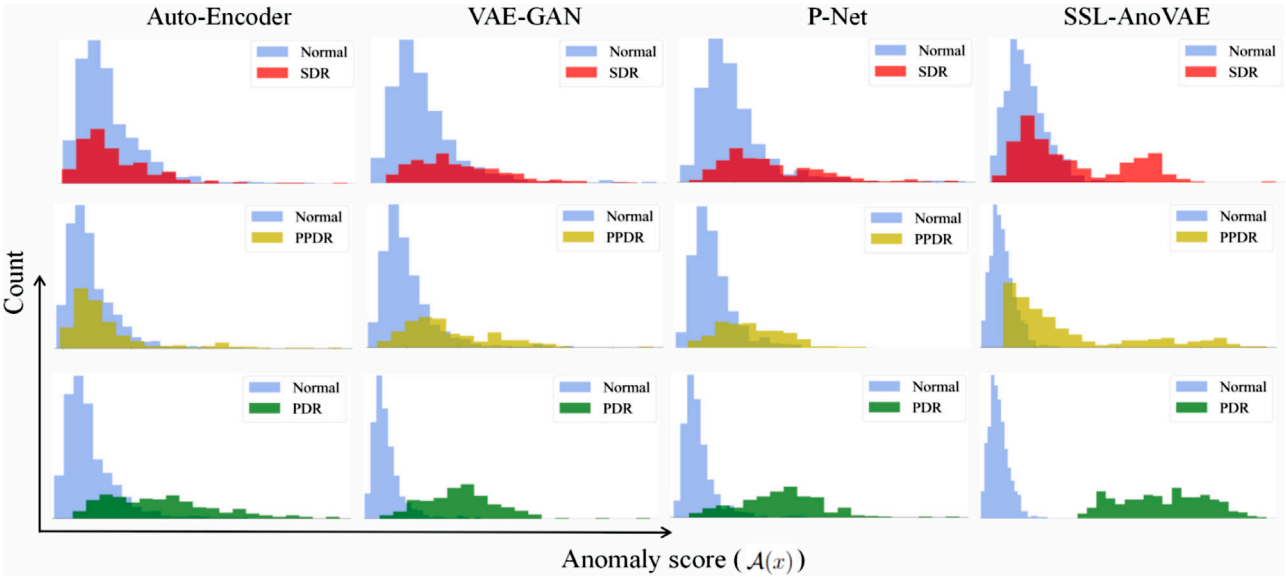

Fig. 5. Histograms of anomaly scores plotted for normal (NDR) images and three sub-classes of abnormal images in the Davis (CFP) test set. The first row compares the quantity histograms of anomaly scores for normal (NDR) with SDR between the four methods. The second row compares normal (NDR) with PPDR, and the last row compares PDR with normal (NDR).

Table 3
The anomaly detection performance (AUC) of different stages in the Davis (CFP) dataset.

Methods	SDR	PPDR	PDR
Auto-Encoder (Zhou and Paffenroth, 2017)	55.93	65.92	93.22
VAE-GAN (Baur et al., 2018)	75.68	82.59	97.77
P-Net (Zhou et al., 2020)	80.55	81.79	98.98
SSL-AnoVAE	**85.44**	**92.09**	**99.98**

anomaly detection performance by our proposed method occurs in all anomaly stages, especially in the SDR and PPDR stages, which is of great importance to patients for diagnosing the disease at earlier stages.

5.1.3. The proposed SSL-AnoVAE improves the pixel-level anomaly localization via different data transformations

In Fig. 6, we qualitatively compare the performance of pixel-level anomaly localization by changing different types of data transformations used in the self-supervised learning module of SSL-AnoVAE. The first group in the figure shows that when the data transformations are not color-related where the color remains the same as the original image after the transformation, i.e., with only flip, Gaussian blurring, etc., the SSL module pays more attention to the shared color information, which is more conducive to SSL-AnoVAE to detect color-related anomalies in the image at the pixel level. Similarly, the second group shows that when the data transformations are color-related while the structure remains the same as the original image after transformation, such as color jitter, color drop, etc., the SSL module can extract more structure-related features to participate in the anomaly detection, which helps SSL-AnoVAE to detect more structure-related anomalies. Finally, the last group contains the data transformations we proposed in the SSL module, which can detect both color-related and structure-related anomalies in the images. The results suggest that the proposed SSL-AnoVAE improves the pixel-level anomaly localization via different data transformations, and also confirm the validity of the proposed data

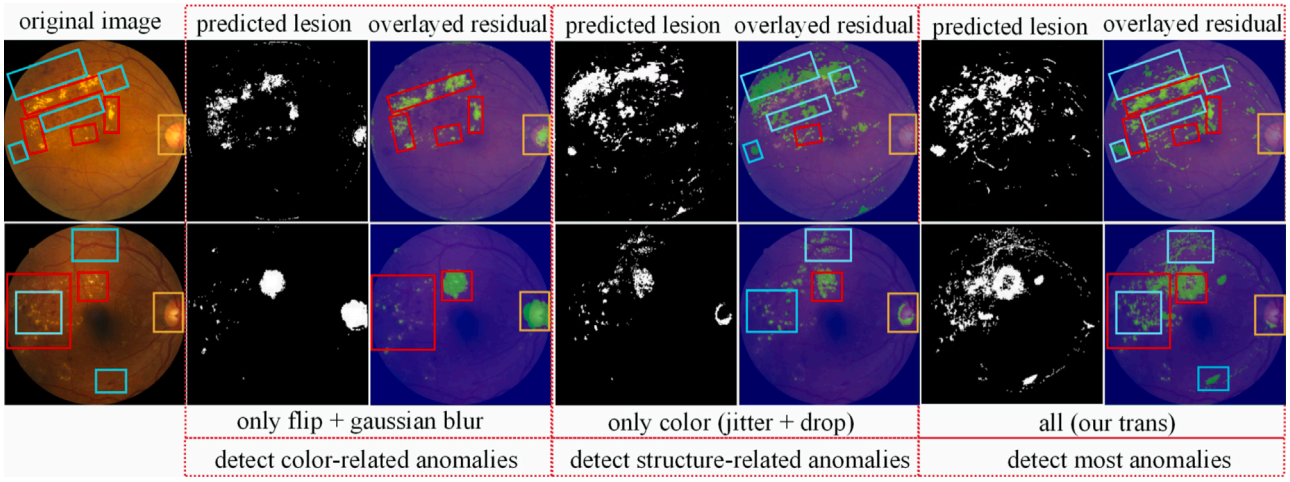

Fig. 6. The effect of data transformation on pixel-level localization of abnormal regions. The first column shows the original input of abnormal images. Red boxes indicate color-related anomalies (e.g., soft exudate), and blue boxes indicate structure-related anomalies (e.g., neovascularization, hemorrhage, etc.). The 2nd and 3rd columns: the predicted lesion (binary residual) and overlayed residual of abnormal images with data transformations of flip and Gaussian blurring used in SSL-AnoVAE. The 4th and 5th columns: the predicted lesion (binary residual) and overlayed residual of abnormal images with corlor-related data transformations (jittering and grey scale). The last group uses all selected data transformations to detect almost all abnormal regions.

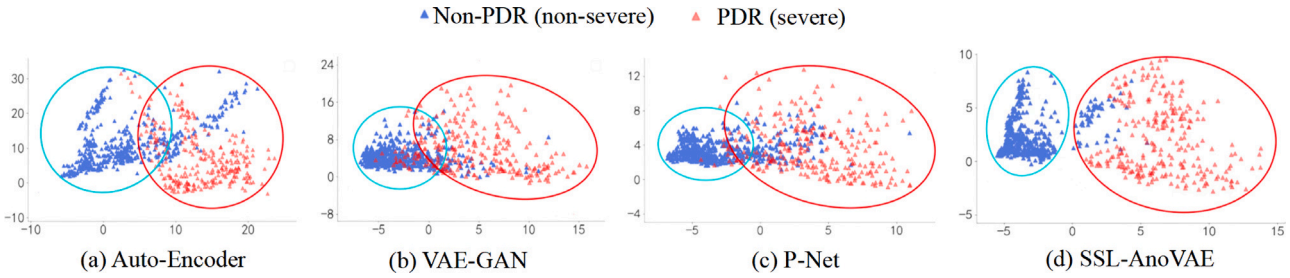

Fig. 7. Clustering results obtained from residuals of abnormal images on the Davis (CFP) dataset, i.e., clusters of Non-PDR and PDR (severe).

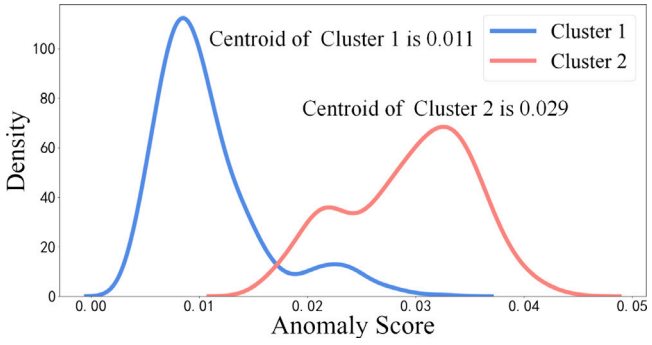

Fig. 8. The density distribution of the anomaly score of each cluster from SSL-AnoVAE. The cluster of centroid with a smaller value is considered as non-PDR and vice versa for PDR.

transformations for SSL-AnoVAE (Table 1). Moreover, the SSL module can be flexibly adjusted depending on the to-be detected anomalies, facilitating the detection of different modalities of disease, which also conform that SSL-AnoVAE is effective for retinal images of different modalities.

5.2. The anomaly staging results

Here, we qualitatively and quantitatively compare the performance of anomaly severity staging (i.e., to distinguish PDR from Non-PDR anomalies given the clinical importance of identifying PDR), based on the anomalies detected by the proposed SSL-AnoVAE and the other three state-of-the-art methods (Auto-Encoder, VAE-GAN, P-Net). As shown in Table 4, SSL-AnoVAE achieves the best ACC (92.86%) and F1-score (90.52%) among the listed methods, which are 7.27% and 6.94% higher than the second-highest baseline, respectively. These results indicate that our proposed SSL-AnoVAE outputs the abnormal residuals in better quality that can facilitate the distinguishing between Non-PDR and PDR. Moreover, we find that during the staging process, the abnormal residuals generated by other anomaly detection methods cannot be clearly separated into different clusters through K-means clustering (Fig. 7(a), (b) and (c)), which suggests that the subtle anomaly differences between PDR and Non-PDR are not well captured in these methods. However, as shown in Fig. 7(d) and the anomaly score curves in Fig. 8, SSL-AnoVAE can distinguish severe disease (PDR) and non-severe disease (Non-PDR) from residuals of abnormal images, which also supports the results presented in Table 4. These findings may be explained by the introduction of the SSL module in SSL-AnoVAE as prior information, which can detect subtle abnormal structures and color changes more accurately.

5.3. The anomaly segmentation results

In this subsection, we compare the unsupervised anomaly segmentation performance of our proposed layer-wise method with typically reconstruction-based unsupervised anomaly segmentation methods on the RESC (OCT) dataset. The typically reconstruction-based method where the threshold processes abnormal regions from potentially abnormal samples in the residual images to obtain the binary segmentation maps (Baur et al., 2018, 2021). Due to the large area of the

Table 4
The unsupervised anomaly staging performance of different methods on the Davis (CFP) dataset (Non-PDR and PDR).

Method	ACC (%)	F1-score (%)
Auto-Encoder (Zhou and Paffenroth, 2017)	82.96	78.95
VAE-GAN (Baur et al., 2018)	85.59	80.63
P-Net (Zhou et al., 2020)	84.96	83.58
SSL-AnoVAE	**92.86**	**90.52**

Table 5
The unsupervised anomaly segmentation performance of different methods on the RESC (OCT) dataset.

Method	Dice
U-Net (supervised)	0.8081
Auto-Encoder (residual) (Baur et al., 2018, 2021)	0.0638
Auto-Encoder (filled residual)	0.1631
Auto-Encoder (layer-wise)	0.3876
VAE-GAN (residual) (Baur et al., 2018, 2021)	0.1047
VAE-GAN (filled residual)	0.4760
VAE-GAN (layer-wise)	0.5941
P-Net (residual) (Baur et al., 2018, 2021)	0.0969
P-Net (filled residual)	0.4297
P-Net (layer-wise)	0.5146
WeakAnD (Seeböck et al., 2019)	0.6312
SSL-AnoVAE (residual)	0.1267
SSL-AnoVAE (filled residual)	0.5842
SSL-AnoVAE (layer-wise)	**0.6889**

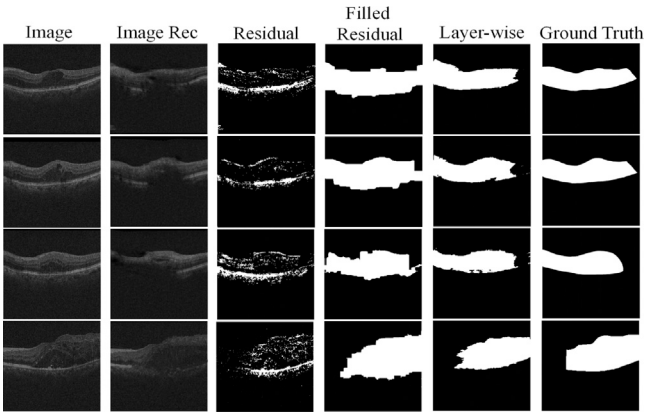

Fig. 9. The segmentation results of three unsupervised anomaly segmentation methods (i.e., residual, filled residual, and layer-wise comparison). The results are all based on the anomalies obtained from the SSL-AnoVAE model. The ground truth is abnormal region.

retinal edema lesion region, one can further improve the segmentation maps by morphological filling the abnormal regions in the residual images. We term the above two baselines as *residual* and *filled residual*, respectively. Furthermore, to test the performance of the supervised segmentation model when inferring unseen data, we also train a supervised U-Net as comparison (top row in Table 5).

Table 5 quantitatively compares the segmentation performance of the three unsupervised anomaly segmentation methods, i.e., residual, filled residual, and our proposed layer-wise grey scale comparison method, on four reconstruction-based anomaly detection methods (Auto-Encoder, VAE-GAN, P-Net, and SSL-AnoVAE). To further demonstrate the effectiveness of our model, we also compare it with WeakAnD (Seeböck et al., 2019), which is the anomaly segmentation method based on epistemic uncertainty. For each reconstruction-based anomaly detection method, the proposed layer-wise segmentation method significantly improves the Dice over the other two baselines, for example, SSL-AnoVAE (layer-wise) tremendously increases the Dice from 0.1267 in SSL-AnoVAE (residual) and 0.5842 in SSL-AnoVAE (filled residual) to 0.6889. Overall, the combination of the proposed SSL-AnoVAE and layer-wise comparison method gives the best Dice score of 0.6889, which is also approaching the supervised segmentation performance (Dice: 0.8081). Fig. 9 visually presents some representative examples of the segmentation results, showing that our proposed layer-wise method gives more accurate segmentation maps, compared to the other baselines. Together, the results demonstrate the effectiveness of our proposed SSL-AnoVAE for anomaly detection in retinal images, followed by the layer-wise comparison method for the subsequent unsupervised anomaly segmentation.

5.4. Ablation studies

5.4.1. Choices of image reconstruction strategies

In Fig. 2, our proposed SSL-AnoVAE concatenates two sampled latent representations from feature extractor E_{S_θ} in the SSL module and encoder E_I as input for the image reconstruction. We perform ablation experiments to investigate the effectiveness of this design. The results are summarized in Table 6. The first four rows are all using a single feature input for the image reconstruction, i.e., either from the SSL module or from the image module. For instance, the first two rows are anomaly detection using latent features only from the encoder E_{S_θ} of the SSL module without sampling (row 1) or with sampling (row 2). Similarly, rows 3 and 4 represent the input for image reconstruction by only using latent representations from original encoder E_I without sampling or with sampling. It is observed that for the image reconstruction with single input, the sampled representation strategy performs better than that without sampling. Next, the rest rows are using both representations from the encoders E_I and E_{S_θ} as input for the image reconstruction. For example, row 5 indicates that we sample the concatenated representations, i.e., first concatenate the two latent representations and then sample the concatenated representation, which slightly improves the detection performance compared to the single input strategy. The 6th row concatenates the representation from encoder E_I without sampling and the sampled representation from the SSL module. Notably, the last row is the strategy we used in our SSL-AnoVAE method, i.e., concatenating the two sampled latent representations from both the SSL module and the image module as input for the image reconstruction. It can be concluded from the table that the current design for the proposed SSL-AnoVAE is optimal, and gives the best performance by using representations from both the SSL and image modules with the sampling strategy.

5.4.2. Choices of data transformations

The self-supervised learning module in SSL-AnoVAE is responsible for exploiting useful information from the raw normal data that can be used for anomaly detection tasks. Different image transformations of the SSL module can guide anomaly detection to focus on different feature related anomalies in the image. In this work, we choose the data transformations (Table 1) by taking into consideration that the SSL module needs to provide more prior information such as color-related and structure-related information according to the characteristics of the data. Nevertheless, it is not true that with more data transformations, more useful information can be obtained. In fact, certain data transformations, if not matched with the characteristics of retinal images, may cause the SSL module to extract misleading information to the network, thereby reducing the performance. Therefore, we use ablation experiments to enumerate the impact of the chosen data transformations on our model.

Fig. 10 presents the ablation results of the chosen data transformations in our method. On right of the figure, we try seven types of data transformations which may be suitable for acquiring CFP image features. The results on the Davis (CFP) dataset show that when the data transformations are the same as listed in Table 1, i.e., color drop, flip, Gaussian blur and color jitter, the SSL-AnoVAE model gives the

Table 6
Ablation study on different strategies of image reconstruction on both RESC (OCT) and Davis (CFP) datasets.

Index	Input of generator					AUC (%)	
	SSL		Image		$\mathcal{L}_{prior_{I+S}}$	RESC (OCT)	Davis (CFP)
	No sampling	Sampling($\mathcal{L}_{prior_S}$)	No sampling	Sampling($\mathcal{L}_{prior_I}$)			
1	✓					87.68	86.83
2		✓				91.62	90.06
3			✓			82.07	84.20
4				✓		89.04	88.20
5					✓	94.02	92.18
6	✓		✓			95.42	92.68
7	✓			✓		95.57	92.05
8		✓	✓			96.63	93.47
9		✓		✓		**98.04**	**94.41**

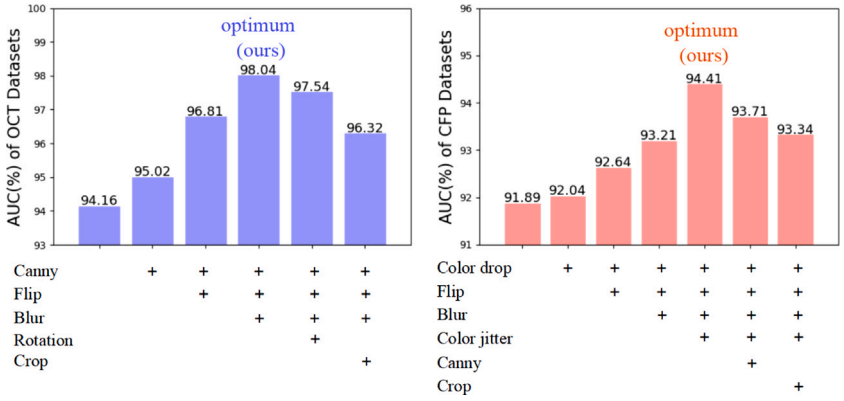

Fig. 10. The ablation study of different data transformations for SSL-AnoVAE on both Davis (CFP) and RESC (OCT) datasets. For instance, on the x-axis (left figure), the first bar (94.16%) is the AUC of the proposed method without any data transformation. The third bar (96.81%) is with data transformations of Canny and Flip.

Table 7
Ablation study on different manners of training the SSL module for anomaly detection (AUC %).

Methods	Davis (CFP)	RESC (OCT)
Pre-train (fix weights)	88.69	90.54
Pre-train	91.87	94.12
End-to-end	94.41	98.04

best anomaly detection performance. However, when Canny or crop transformations are added, the model's performance is impaired due to the noisy edge information by the Canny extraction, or the missing information during random cropping. Similarly, for the OCT images, six data transformations according to OCT image characteristics are investigated on the left of Fig. 10. As shown in the figure, when we choose data transformations with Canny, flip, and Gaussian blur, the model achieves the best performance for anomaly detection. Notably, Canny transformation is effective for OCT images but not for CFP images, which can be explained by the importance of the layer structure extracted by the Canny transformation for reconstructing the original OCT images.

5.4.3. Choices of pre-train or end-to-end training strategy

For the choices of different strategies of the SSL module participating in the anomaly detection, we perform comparative experiments to show the effectiveness of end-to-end training strategy over using pretrained models. As shown in Table 7, the first row is the result where we first pre-train the SSL module, and then the SSL module with fixed weights is integrated into the image reconstruction module for the training of anomaly detection network, resulting in the AUC of 88.69%, 90.54% on the Davis (CFP) and RESC (OCT) datasets, which is 3.18% and 3.58% lower than that of using the SSL module without fixed weights (the second row). In contrast, our proposed SSL-AnoVAE using the end-to-end training strategy achieves a significant improvement of 2.54%, 3.92% for AUC over the above two pre-train based strategies on the Davis (CFP) and RESC (OCT) datasets.

6. Discussion and conclusion

We propose a novel general unsupervised anomaly detection framework SSL-AnoVAE, which utilizes a self-supervised learning module with two unbalanced branches to obtain more semantic feature information depending on the to-be detected anomalies from retinal images. Then SSL-AnoVAE concatenates sampled feature information from both the SSL module and the original encoder for image reconstruction, which is proved to be effective for anomaly detection. Since our model can accurately detect the lesions, we also apply it towards clinical usages for computer-aided diagnosis of retinal-related diseases, where we utilize residuals of abnormal images and anomaly score output by SSL-AnoVAE to determine the disease stage, and further propose a layer-wise segmentation method to segment the anomalies.

We demonstrate the effectiveness of our proposed methods through extensive experiments and ablation studies on the Davis (CFP) and RESC (OCT) datasets. The experimental results reveal four critical points. Firstly, our proposed anomaly detection method SSL-AnoVAE, which utilizes the SSL module to obtain semantic prior information and concatenates two sampled representations for image reconstruction, is effective by producing both plausible and accurate anomaly detection results in retinal images for different modalities. Secondly, our SSL-AnoVAE presents the relationship between data transformation and different types of lesion detection, which can also be extended to future research, exploring to treat other complex medical disease with different data transformations. Moreover, we use the Davis (CFP) data to validate our method's ability to identify disease stage, which helps to understand the severity or progression of the retinal disease. Finally, the proposed layer-wise segmentation method based on SSL-AnoVAE yields significantly better results than the traditional UAD segmentation

methods. To the best of our knowledge, this is the first time using layer-wise comparison between the reconstructed image and the original image for unsupervised anomaly segmentation.

We demonstrate the effectiveness of our proposed methods through extensive experiments and ablation studies on the Davis (CFP) and RESC (OCT) datasets. The experimental results reveal four critical points. (1) Sections 5.1.1 and 5.1.2 proved that SSL-AnoVAE outperforms state-of-the-art methods at both image and pixel levels performance and improves diagnosis of all stage diseases, which helps patients for diagnosing the disease at earlier stages. It also confirms that our method utilizes the SSL module to obtain prior semantic information and concatenates two sampled representations for image reconstruction, which is effective by producing both plausible and accurate anomaly detection results in retinal images for different modalities. (2) Section 5.1.3 compared the one and multi transformed inputs and presented the relationship between data transformation and different types of lesion detection, which can also be extended to future research to explore the treatment of other complex medical diseases with different data transformations. (3) In Section 5.2 we use the Davis (CFP) data to validate our method's ability to identify disease stages. The SSL-AnoVAE can distinguish severe disease (PDR) and non-severe disease (Non-PDR) from residuals of abnormal images, which helps to understand the severity or progression of the retinal disease. (4) In Section 5.3, the proposed layer-wise segmentation method based on SSL-AnoVAE yields significantly better results than the traditional UAD segmentation methods. To the best of our knowledge, this is the first time using the layer-wise comparison between the reconstructed image and the original image for unsupervised anomaly segmentation.

Although our method achieves excellent performance, it still comes with some unresolved limitations. (1) Since SSL-AnoVAE takes two sampled latent representations from the SSL module and the original image as input for image reconstruction, although being proved to be effective in the ablation experiments, the mathematical interpretability of combining two sampled latent representations has not been explored. (2) In retinal image applications, for CFP images, some neovascularizations whose color, shape, and location distributions are very similar to healthy blood vessels, which will mistakenly cause our model to recognize neovascularizations as healthy tissue. For OCT images, some diseases (e.g., pigment epithelium detachment) may not completely destroy the layer structure and the intensity of abnormal regions does not change significantly, which will cause the model to output small anomaly scores of these regions. However, we think that inputting reconstructed abnormal images into the layer extractor can alleviate such problems to a certain extent. Although this is only slightly destroyed in the original image, the reconstructed abnormal region may cause the slightly damaged region to be emphasized, especially the layer-wise reconstructed image processed by the layer extractor (see Section 3.3). Therefore, inputting the reconstructed image into the layer extractor for the segmentation task can alleviate this problem to a certain extent. (3) In general, our model is conducive to generalization to other medical datasets because the prior information it extracts comes from its features and does not require additional extraction. Therefore, for many medical tasks, the SSL module can be designed depending on the to-be-detected anomalies. But for some tasks where the structure and color information is not obvious (e.g., ultrasound images), our method may not have an excellent strategy to obtain good prior information through the SSL model.

Our future work will explore the mathematical interpretability of combining two sampled latent representations in the image reconstruction module. Moreover, we will extend the application of SSL-AnoVAE to other complex medical problems.

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

The datasets used in this work are public.

Acknowledgments

This study was supported by National Key Research and Development Program of China (2020YFB1711500, 2020YFB1711503), the 1.3.5 project for disciplines of excellence, West China Hospital, Sichuan University (ZYYC21004, ZYJC18010).

References

Akcay, S., Atapour-Abarghouei, A., Breckon, T.P., 2018. Ganomaly: Semi-supervised anomaly detection via adversarial training. In: Asian Conference on Computer Vision. Springer, pp. 622–637.

Alqudah, A.M., 2020. AOCT-NET: A convolutional network automated classification of multiclass retinal diseases using spectral-domain optical coherence tomography images. Med. Biol. Eng. Comput. 58 (1), 41–53.

Arora, S., Khandeparkar, H., Khodak, M., Plevrakis, O., Saunshi, N., 2019. A theoretical analysis of contrastive unsupervised representation learning. arXiv preprint arXiv:1902.09229.

Azizi, S., Mustafa, B., Ryan, F., Beaver, Z., Freyberg, J., Deaton, J., Loh, A., Karthikesalingam, A., Kornblith, S., Chen, T., et al., 2021. Big self-supervised models advance medical image classification. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 3478–3488.

Baur, C., Denner, S., Wiestler, B., Navab, N., Albarqouni, S., 2021. Autoencoders for unsupervised anomaly segmentation in brain MR images: A comparative study. Med. Image Anal. 69, 101952.

Baur, C., Wiestler, B., Albarqouni, S., Navab, N., 2018. Deep autoencoding models for unsupervised anomaly segmentation in brain MR images. In: International MICCAI Brainlesion Workshop. Springer, pp. 161–169.

Chen, X., He, K., 2021. Exploring simple siamese representation learning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 15750–15758.

Chen, X., Konukoglu, E., 2018. Unsupervised detection of lesions in brain MRI using constrained adversarial auto-encoders. arXiv preprint arXiv:1806.04972.

Chen, T., Kornblith, S., Norouzi, M., Hinton, G., 2020a. A simple framework for contrastive learning of visual representations. In: International Conference on Machine Learning. PMLR, pp. 1597–1607.

Chen, X., Pawlowski, N., Glocker, B., Konukoglu, E., 2021. Normative ascent with local gaussians for unsupervised lesion detection. Med. Image Anal. 74, 102208.

Chen, X., You, S., Tezcan, K.C., Konukoglu, E., 2020b. Unsupervised lesion detection via image restoration with a normative prior. Med. Image Anal. 64, 101713.

Chen, Y., Zhang, H., Wang, Y., Yang, Y., Zhou, X., Wu, Q.J., 2020c. MAMA net: Multi-scale attention memory autoencoder network for anomaly detection. IEEE Trans. Med. Imaging 40 (3), 1032–1041.

Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., Bengio, Y., 2014. Generative adversarial nets. Adv. Neural Inf. Process. Syst. 27.

Grill, J.-B., Strub, F., Altché, F., Tallec, C., Richemond, P.H., Buchatskaya, E., Doersch, C., Pires, B.A., Guo, Z.D., Azar, M.G., et al., 2020. Bootstrap your own latent: A new approach to self-supervised learning. arXiv preprint arXiv:2006.07733.

Hansen, S., Gautam, S., Jenssen, R., Kampffmeyer, M., 2022. Anomaly detection-inspired few-shot medical image segmentation through self-supervision with supervoxels. Med. Image Anal. 78, 102385.

He, K., Fan, H., Wu, Y., Xie, S., Girshick, R., 2020. Momentum contrast for unsupervised visual representation learning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 9729–9738.

He, Z., Xu, X., Deng, S., 2003. Discovering cluster-based local outliers. Pattern Recognit. Lett. 24 (9–10), 1641–1650.

Hu, J., Chen, Y., Yi, Z., 2019. Automated segmentation of macular edema in OCT using deep neural networks. Med. Image Anal. 55, 216–227.

Isola, P., Zhu, J.-Y., Zhou, T., Efros, A.A., 2017. Image-to-image translation with conditional adversarial networks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 1125–1134.

Kim, J., Scott, C.D., 2012. Robust kernel density estimation. J. Mach. Learn. Res. 13 (1), 2529–2565.

Komodakis, N., Gidaris, S., 2018. Unsupervised representation learning by predicting image rotations. In: International Conference on Learning Representations. ICLR.

Koohbanani, N.A., Unnikrishnan, B., Khurram, S.A., Krishnaswamy, P., Rajpoot, N., 2021. Self-Path: Self-supervision for classification of pathology images with limited annotations. IEEE Trans. Med. Imaging.

Larsen, A.B.L., Sønderby, S.K., Larochelle, H., Winther, O., 2016. Autoencoding beyond pixels using a learned similarity metric. In: International Conference on Machine Learning. PMLR, pp. 1558–1566.

Li, X., Jia, M., Islam, M.T., Yu, L., Xing, L., 2020. Self-supervised feature learning via exploiting multi-modal data for retinal disease diagnosis. IEEE Trans. Med. Imaging 39 (12), 4023–4033.

Li, X., Zhou, Y., Wang, J., Lin, H., Zhao, J., Ding, D., Yu, W., Chen, Y., 2021. Multi-modal multi-instance learning for retinal disease recognition. arXiv preprint arXiv:2109.12307.

Mahapatra, D., Poellinger, A., Shao, L., Reyes, M., 2021. Interpretability-driven sample selection using self supervised learning for disease classification and segmentation. IEEE Trans. Med. Imaging 40 (10), 2548–2562.

Schlegl, T., Seeböck, P., Waldstein, S.M., Langs, G., Schmidt-Erfurth, U., 2019. f-AnoGAN: Fast unsupervised anomaly detection with generative adversarial networks. Med. Image Anal. 54, 30–44.

Schlegl, T., Seeböck, P., Waldstein, S.M., Schmidt-Erfurth, U., Langs, G., 2017. Unsupervised anomaly detection with generative adversarial networks to guide marker discovery. In: International Conference on Information Processing in Medical Imaging. Springer, pp. 146–157.

Seeböck, P., Orlando, J.I., Schlegl, T., Waldstein, S.M., Bogunović, H., Klimscha, S., Langs, G., Schmidt-Erfurth, U., 2019. Exploiting epistemic uncertainty of anatomy segmentation for anomaly detection in retinal OCT. IEEE Trans. Med. Imaging 39 (1), 87–98.

Shyu, M.-L., Chen, S.-C., Sarinnapakorn, K., Chang, L., 2003. A Novel Anomaly Detection Scheme Based on Principal Component Classifier. Technical Report, Miami Univ Coral Gables Fl Dept of Electrical and Computer Engineering.

Tack, J., Mo, S., Jeong, J., Shin, J., 2020. Csi: Novelty detection via contrastive learning on distributionally shifted instances. Adv. Neural Inf. Process. Syst. 33, 11839–11852.

Takahashi, H., 2017. Davis Grading of One and Concatenated Figures. figshare.

Takahashi, H., Tampo, H., Arai, Y., Inoue, Y., Kawashima, H., 2017. Applying artificial intelligence to disease staging: Deep learning for improved staging of diabetic retinopathy. PLoS One 12 (6), e0179790.

Tian, Y., Sun, C., Poole, B., Krishnan, D., Schmid, C., Isola, P., 2020. What makes for good views for contrastive learning? Adv. Neural Inf. Process. Syst. 33, 6827–6839.

Trichonas, G., Kaiser, P.K., 2014. Optical coherence tomography imaging of macular oedema. Br. J. Ophthalmol. 98 (Suppl 2), ii24–ii29.

Wang, Y., Qin, C., Wei, R., Xu, Y., Bai, Y., Fu, Y., 2021. SLA^{2}P: Self-supervised anomaly detection with adversarial perturbation. arXiv e-prints, arXiv–2111.

Wang, S., Zeng, Y., Liu, X., Zhu, E., Yin, J., Xu, C., Kloft, M., 2019. Effective end-to-end unsupervised outlier detection via inlier priority of discriminative network. Adv. Neural Inf. Process. Syst. 32.

Yang, X., Latecki, L.J., Pokrajac, D., 2009. Outlier detection with globally optimal exemplar-based GMM. In: Proceedings of the 2009 SIAM International Conference on Data Mining. SIAM, pp. 145–154.

Yao, Q., Xiao, L., Liu, P., Zhou, S.K., 2021. Label-free segmentation of COVID-19 lesions in lung CT. IEEE Trans. Med. Imaging 40 (10), 2808–2819.

Yoa, S., Lee, S., Kim, C., Kim, H.J., 2021. Self-supervised learning for anomaly detection with dynamic local augmentation. IEEE Access 9, 147201–147211.

Zhao, H., Li, Y., He, N., Ma, K., Fang, L., Li, H., Zheng, Y., 2021. Anomaly detection for medical images using self-supervised and translation-consistent features. IEEE Trans. Med. Imaging 40 (12), 3641–3651.

Zhou, C., Paffenroth, R.C., 2017. Anomaly detection with robust deep autoencoders. In: Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. pp. 665–674.

Zhou, K., Xiao, Y., Yang, J., Cheng, J., Liu, W., Luo, W., Gu, Z., Liu, J., Gao, S., 2020. Encoding structure-texture relation with P-Net for anomaly detection in retinal images. In: Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XX 16. Springer, pp. 360–377.

Zhu, J.-Y., Park, T., Isola, P., Efros, A.A., 2017. Unpaired image-to-image translation using cycle-consistent adversarial networks. In: Proceedings of the IEEE International Conference on Computer Vision. pp. 2223–2232.

Zimmerer, D., Kohl, S.A., Petersen, J., Isensee, F., Maier-Hein, K.H., 2018. Context-encoding variational autoencoder for unsupervised anomaly detection. arXiv preprint arXiv:1812.05941.

Zong, B., Song, Q., Min, M.R., Cheng, W., Lumezanu, C., Cho, D., Chen, H., 2018. Deep autoencoding Gaussian mixture model for unsupervised anomaly detection. In: International Conference on Learning Representations.