Anomaly Detection for Medical Images Using Self-Supervised and Translation-Consistent Features

He Zhao, Yuexiang Li, Nanjun He, Kai Ma, Leyuan Fang, *Senior Member, IEEE*,
Huiqi Li, *Senior Member, IEEE*, and Yefeng Zheng

***Abstract*—As the labeled anomalous medical images are usually difficult to acquire, especially for rare diseases, the deep learning based methods, which heavily rely on the large amount of labeled data, cannot yield a satisfactory performance. Compared to the anomalous data, the normal images without the need of lesion annotation are much easier to collect. In this paper, we propose an anomaly detection framework, namely SALAD, extracting Self-supervised and trAnsLation-consistent features for Anomaly Detection. The proposed SALAD is a reconstruction-based method, which learns the manifold of normal data through an encode-and-reconstruct translation between image and latent spaces. In particular, two constraints (*i.e.*, structure similarity loss and center constraint loss) are proposed to regulate the cross-space (*i.e.*, image and feature) translation, which enforce the model to learn translation-consistent and representative features from the normal data. Furthermore, a self-supervised learning module is engaged into our framework to further boost the anomaly detection accuracy by deeply exploiting useful information from the raw normal data. An anomaly score, as a measure to separate the anomalous data from the healthy ones, is constructed based on the learned self-supervised-and-translation-consistent features. Extensive experiments are conducted on optical coherence tomography (OCT) and chest X-ray datasets. The experimental results demonstrate the effectiveness of our approach.**

***Index Terms*—Medical image analysis, anomaly detection, feature space constraint, generative adversarial networks.**

I. Introduction

DEEP learning methods have achieved state-of-the-art performance in many computer vision tasks. Its superior performance mainly relies on large amounts of labeled data [1]–[3]. However, annotated medical images are usually difficult to acquire, especially for rare diseases, which limits the application of deep learning models for medical diagnosis of anomalous physiological readings. Compared to anomalous cases, the annotation of normal images from healthy subjects is much easier. In clinical practice, human experts are able to identify unexpected cases that differ from healthy references. To imitate the process and loose the annotation requirement of anomalous data, researchers make their efforts to develop anomaly detection approaches that select data showing solely normal appearance to train their models.

Extensive studies have been proposed to explore the effectiveness of anomaly detection for computer vision tasks [4], [5] and medical image analysis [6]–[8], respectively. Most of the existing frameworks are based on the encoder-decoder structure, which use the image reconstruction error as a metric to detect outliers. Nevertheless, the accuracy of those anomaly detection algorithms is still unsatisfactory, due to limited attention paid on latent feature space. The performance can still be improved if the feature constraint is concerned. In this paper, we propose a novel anomaly detection approach that takes both image and feature spaces into consideration. Particularly, the image space and feature space are treated as two domains. As shown in Fig. 1, an encode-and-reconstruct translation is performed between the two spaces. Different from approaches exploiting the image space only, our approach offers a robust feature representation by deeply exploiting the intrinsic information from normal data, thereby results in a higher sensitivity to the anomalies. In other words, our model can reconstruct normal images and corresponding features with a small error, making the error of the anomalies more obvious. For this purpose, two constraints (*i.e.*, structure similarity loss and center constraint) are proposed to enforce the network to learn translation-consistent features from the image and latent spaces, respectively.

The main contribution of our approach can be summarized into four-fold:

Manuscript received April 20, 2021; revised May 25, 2021; accepted June 24, 2021. Date of publication July 1, 2021; date of current version November 30, 2021. This work was supported in part by the Key-Area Research and Development Program of Guangdong Province, China under Grant 2018B010111001, in part by the National Key Research and Development Program of China under Grant 2018YFC2000702, in part by the Scientific and Technical Innovation 2030—New Generation Artificial Intelligence Project under Grant 2020AAA0104100, and in part by the China Postdoctoral Science Foundation under Grant 2020M680387. *(Corresponding authors: Yuexiang Li; Huiqi Li.)*

He Zhao and Huiqi Li are with the Beijing Institute of Technology, Beijing 100811, China (e-mail: huiqili@bit.edu.cn).

Yuexiang Li, Nanjun He, Kai Ma, and Yefeng Zheng are with the Tencent Jarvis Laboratory, Shenzhen 518057, China (e-mail: vicyxli@tencent.com).

Leyuan Fang is with the College of Electrical and Information Engineering, Hunan University, Changsha 410082, China.

Digital Object Identifier 10.1109/TMI.2021.3093883

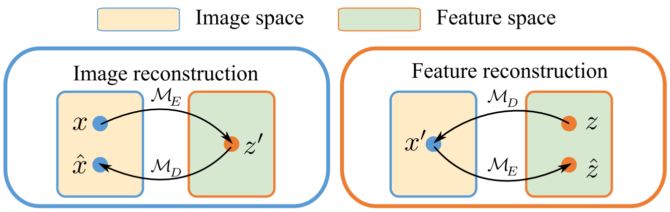


Fig. 1. Overview of the proposed anomaly detection approach. There are two translations (*i.e.*, image reconstruction and feature reconstruction) between image and latent spaces in our framework to train the same encoder ($\mathcal{M}_E$) and decoder ($\mathcal{M}_D$), which enforce them to deeply exploit useful information from normal data for anomaly detection.

- We propose a framework, namely **SALAD**,[1] extracting from **S**elf-supervised and tr**A**nsL**a**tion-consistent features for **A**nomaly **D**etection. The proposed SALAD is a reconstruction-based anomaly detection framework, which performs image-feature translations to simultaneously extract useful information from both image and latent feature spaces.
- Two additional constraints, *i.e.*, structured similarity loss and center constraint loss, are proposed to enforce the network to maintain the consistency in image space and feature space, respectively, during translation. Such a translation-consistent feature representation extracts more meaningful information of normal images and has a higher sensitivity to anomalous data compared to the existing approaches.
- A self-supervised learning module is combined in our framework to deeply exploit useful information from the raw normal data, which further boosts the classification performance.
- We extensively evaluate the proposed SALAD model on two publicly available datasets. The experimental results demonstrate the effectiveness of our self-supervised and translation-consistent features for anomaly detection on different medical imaging applications.

The rest of our paper is organized as follows. In Section II, we review the relevant studies on anomaly detection and self-supervised learning. In Section III, we describe the proposed SALAD model in details. The experiments and results are presented and discussed in Section IV. Finally, Section V concludes this study.

II. Related Work

A. Anomaly Detection

Anomaly detection, closely related to outlier detection or novelty detection, is the identification of rare events or observations which deviate from the distribution of normal data. It has been applicable in a variety of domains, such as fraud detection [9], network intrusion detection [10], medical imaging lesion detection [6] and vehicle detection in crowed scenes [11]. Based on the definition, it is natural to learn a decision boundary to separate anomalous data from normal data. One-class support vector machine [12] is an algorithm which learns a discriminative hyperplane by modeling the training distribution where a small region contains most of the training examples and anomalies are away from this region. Deep SVDD [13] utilizes convolutional networks to extract the common factors of variation by minimizing the volume of hypersphere that encloses the representations of normal data. Gaussian mixture model (GMM) can also be used for anomaly detection [14]–[17], which tends to model the distribution of normal data and detects anomalies based on probability. In [16], the authors presented a deep auto-encoding method to extract features and further fed the features to GMM to identify anomalous cases. These methods usually obtain a less desirable performance when dealing with high-dimensional data.

The reconstruction-based methods tend to learn a mapping function to reconstruct normal data with the assumption that anomalous ones will not be well recovered and lead to higher residuals. Sparse coding and dictionary learning are traditional ways to encode the normal patterns [18], [19], where the normal data can be easily represented by the linear combination of the bases which encode normal patterns in the training set, while the anomalous data is not. Bergmann *et al.* [20] utilized a student-teacher network for anomaly detection, where the anomaly score is based on how much the output of student network differs from that of the teacher network. AnoGAN [6] was proposed with a generative adversarial network. The query image is identified as abnormal if there is a large difference between the query image and its reconstruction which is generated by a latent code determined by iterative progress. Further, a fast version of AnoGAN was proposed in [8] termed as f-AnoGAN, where an encoder is trained to replace the iterative process for determining the latent code. Following the study of [6], Zenati *et al.* [21] used a bi-directional generative adversarial network to map an image to latent space, which reduces the computational complexity. Approaches based on auto-encoders [22] and variational auto-encoders [23] are very popular to learn the reconstruction function. Feature matching metric [24] and structure similarity metric [25] are also considered in the loss terms for better reconstruction performance on normal data. Akcay *et al.* [4] proposed a Ganomaly model which learns the image and latent spaces jointly. Their encoder-decoder structure is followed by another encoder to generate reconstructed latent vector to capture features in the latent space. Similarly, Zhou *et al.* [26] further introduced a sparsity regularization net to restrict the bottleneck features in the latent space to enhance the model ability. In [27], two discriminators were utilized on both image and latent spaces together with a classifier proposed to determine how well the given image resembles the content of the given class. As a result, their model can force the latent representation of any samples to reconstruct images of the given class. Most of existing methods utilize reconstruction error in image space as guideline, but feature representations are not well considered in the above methods, which motivates us to design constraints on the feature space to achieve better performance.

[1] Like salad mixing various ingredients, our approach fuses the information extracted from the image and feature spaces together for the robust anomaly detection.

B. Self-Supervised Learning

Self-supervised learning is a new paradigm to learn the underlying information from raw unlabeled data. For 2D natural images, various pretext tasks have been proposed. Doersch *et al.* [28] proposed a framework, learning the visual features by predicting the relative positions of two patches from the same image. Another representative approach of relative position prediction is the Jigsaw puzzles proposed by Noroozi and Favaro [29]. This work requires deep learning networks to rearrange the positions of nine patches cropped from the same image. Based on the Jigsaw puzzles, some variants are developed and colorization can also be formulated as a pretext task to pre-train neural networks [30]–[32]. More recently, studies have proposed to adopt rotation prediction [33], transformation estimation [34], and optical flow prediction [35] as additional pretext tasks. For the applications with medical data, self-supervised learning is also exploited and researchers take medical domain knowledge into account when formulating the pretext task [36]–[38]. For example, Zhou *et al.* [38] proposed multiple proxy tasks to train the model from multiple perspectives, including appearance, texture and context, which leads to a more robust model across all target tasks and the model achieves a superior performance to initialize parameters for other applications.

III. Method

This work aims to train a model to identify anomalies using only the normal data. The main difference between the proposed approach and existing ones is that our framework employs the feature space constraints together with image reconstruction to construct a comprehensive metric for anomaly detection. In this section, we introduce the proposed anomaly detection method in details.

A. Pipeline

Here, we give the definition of the anomaly detection task. Given a large training dataset $\mathcal{D}$ with N normal training samples only (*i.e.*, $\mathcal{D} = \{x_1, \ldots, x_N\}$) and a test set $\mathcal{D}_t$ with M normal and anomalous images (*i.e.*, $\mathcal{D}_t = \{(x_{t_1}, y_1), \ldots, (x_{t_M}, y_M)\}$), where $y_i \in \{0, 1\}$ is the image label (0 for normal image and 1 for anomalous image), our goal is to train a model that captures the distribution of training dataset $\mathcal{D}$ and detect the anomalies in the test set $\mathcal{D}_t$ as outliers during inference.

To achieve this, our SALAD approach encourages a model to fully exploit the useful information contained in normal data via the translation between image and feature spaces. There are two adversarial reconstruction processes for the image space and feature space, which are presented in the blue and orange boxes in Fig. 2(a), respectively. Specifically, the encoder translates the image $x \in \mathbb{R}^{W\times H}$ to a latent representation $z' \in \mathbb{R}^{n\times 1}$ ($\mathcal{M}_E : \mathcal{I} \to \mathcal{F}$), while the decoder reconstructs the latent representation back to an image ($\mathcal{M}_D : \mathcal{F} \to \mathcal{I}$). Similar to CycleGAN for image domain transfer [39], we also introduce two adversarial discriminators $D_\mathcal{I}$ and $D_\mathcal{F}$ for the image space and feature space respectively. Furthermore, a self-supervised learning module with proxy restoration tasks is proposed to encourage $\mathcal{M}_E$ to deeply exploit useful and robust representations from raw normal images for feature embedding.

Intuitively, if the feature representations of normal data generated by $\mathcal{M}_E$ are tightly clustered in the latent feature manifold, the encoded feature of an anomaly image lying far from the normal cluster is easy to identify. Hence, we propose a center constraint to compact the cluster of feature representations extracted from the normal images in the latent feature space. The regularization on the feature space can boost the robustness of the features learned by $\mathcal{M}_E$, which is the main difference between our approach to the existing frameworks.

The inference process is presented in Fig. 2 (b). For a test image x_t, an anomaly score is constructed by measuring the difference between the feature representation z_t and the reconstructed feature $\hat{z}_t$ of reconstructed image $\hat{x}_t$ to identify the anomalies from normal cases.

B. Adversarial Reconstruction in Image and Feature Spaces

There are two adversarial reconstruction processes for the image space and feature space in our SALAD framework as shown in Fig. 2(a). On one hand, similar to the image-space-only approaches, a reconstruction error is adopted to supervise the image space reconstruction process. On the other hand, the sampled feature vector z in the feature space is fed to the decoder-encoder architecture for the feature space reconstruction. A reconstruction error between the sampled feature vector z and reconstructed features $\hat{z}$ is calculated to supervise the reconstruction process. To ensure the encoded feature representation z' (a.k.a. $\mathcal{M}_E(x)$) and synthetic image x' (a.k.a. $\mathcal{M}_D(z)$) have the same distributions to the normal class, two discriminators $D_\mathcal{F}$ and $D_\mathcal{I}$ are implemented. In specific, $D_\mathcal{F}$ is utilized to distinguish z' from the sampled feature vector z which is generated by a multivariate Gaussian distribution. $D_\mathcal{I}$ is trained to identify x' from the original image x. Therefore, the whole framework can be adversarially trained. With the reconstructions in the two spaces, the encoder and decoder can excellently construct the manifold of normal data by fully exploiting the useful information from both image and feature spaces.

C. Self-Supervised Learning Module

Detecting anomalies only by learning the normal data is a challenging task, thus we propose to leverage self-supervised learning technique to assist the reconstruction process and thereby improve the model capability. The self-supervised learning module is shown in the dashed red box of Fig. 2 (a), where restoration-based proxy tasks are constructed by perturbing the images with three strategies. The intuition underlying this module is that the model is expected to learn useful information from the raw data via self-supervised learning, which can thereby assist the following data reconstruction stage for anomaly detection. The same encoder-decoder architecture is adopted for the restoration proxy task. In the experiments, the encoder shares weights with the one for image-wise adversarial reconstruction to improve the robustness of embedded feature z'. We present the detailed information of each

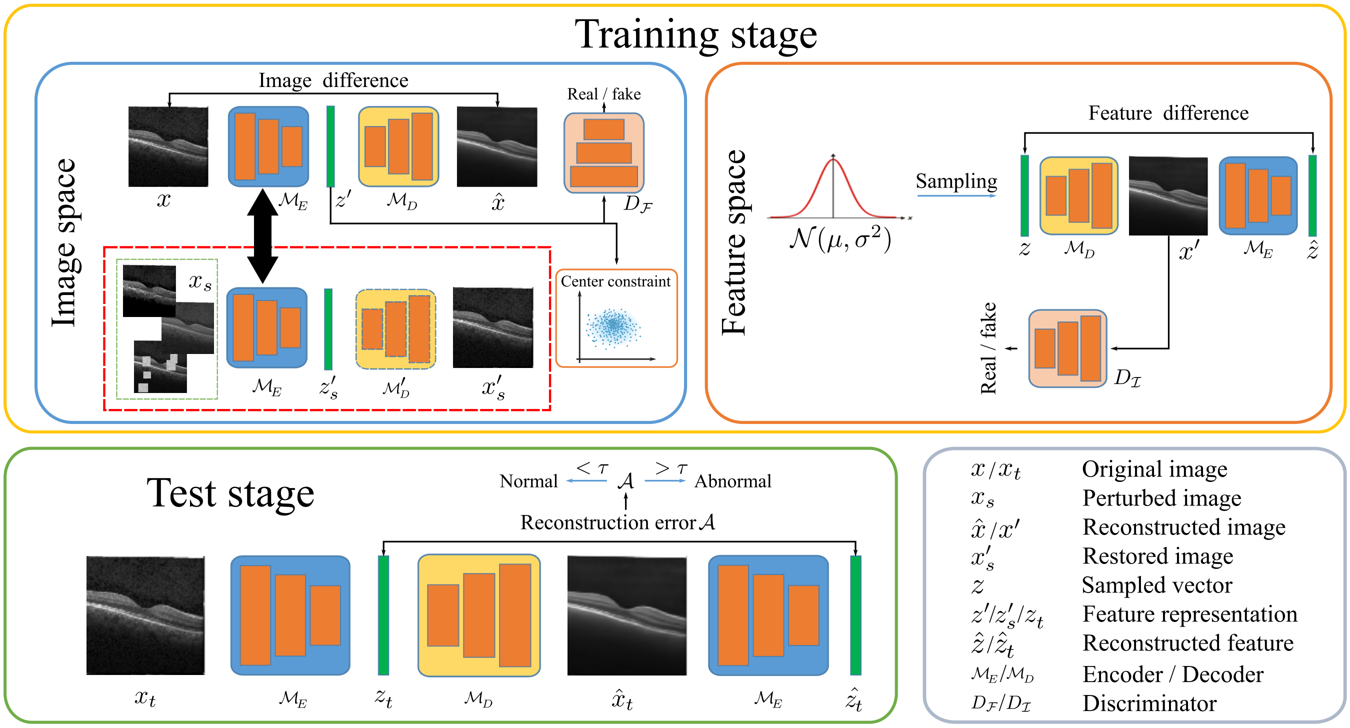


Fig. 2. Flowchart of the proposed SALAD approach. The yellow box shows the training stage of our approach where the reconstruction is applied on both image and feature spaces; The test stage of our approach is illustrated in green box; The legend of symbols are presented in gray box.

kind of perturbations adopted in our self-supervised learning module in the following.

1) In-Painting: The in-painting perturbation is performed by removing some regions in the image. The restoration of the partially blocked image regions drives the model to explicitly learn the anatomic information of normal data. In our experiments, we randomly crop five tiles with random sizes ranging in $[\frac{WH}{64}, \frac{WH}{25}]$ from the images, and fill the cropped areas with random noises drawn from uniform distribution.

2) Local Pixel Shuffling: The conventional pixel shuffling permutation randomly switches the pixel values of different positions in a given image. In our framework, we adopt a so-called local pixel shuffling method to perturb the pixel values. Specifically, we first sample a set of tiles (300 patches are selected with random size ranging from 1 to $\frac{WH}{100}$ in our experiment), and then perform the pixel shuffling in each tile for content permutation. Different from the in-painting perturbation, local pixel shuffling encourages the model to learn local boundaries and texture information.

3) Non-Linear Intensity Transformation: The pixel intensity in medical images contains rich information related to the shape, texture, etc. Thus, a non-linear transformation is adopted to perturb the pixel intensities. Reconstructing from such a perturbed image can improve the robustness of the model to intensity variations among samples. In the experiments, the Bezier curve function [40] with three randomly selected control points is used to alter the pixel intensity.

The examples of images perturbed by different proxy tasks are shown in Fig. 3. From left to right are original image, transformation results of in-painting, local pixel shuffling and non-linear intensity transformation, respectively. One of the perturbations is randomly selected to disarrange an input image. The encoder-decoder is trained to restore the input image from the perturbed one. Therefore, the original input image is used as the supervision signal for the restoration proxy task. We follow the same training procedure of [38]. The encoder is expected to exploit more useful information from the raw image data for anomaly detection via the self-supervised training. Compared with denoising auto-encoder, the proxy tasks utilized in this work are designed from various aspects, such as texture, context and anatomical structure, which results in a robust feature representation for anomaly detection.

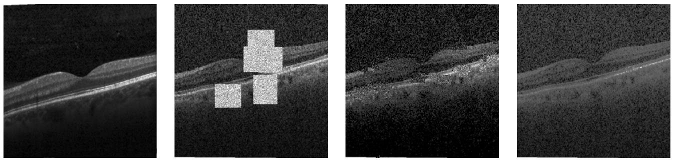


Fig. 3. Three proxy tasks utilized in our self-supervised learning module. From left to right are original image, image after in-painting, image after local pixel shuffling and image after non-linear intensity transformation, respectively.

D. Loss Functions

During training, the encoder $\mathcal{M}_E$, decoder $\mathcal{M}_D$ and discriminators ($D_{\mathcal{F}}$ and $D_{\mathcal{I}}$) are alternatively optimized. The whole framework is supervised by an objective function consisting of four losses (*i.e.*, adversarial loss, reconstruction loss, feature center loss and restoration loss), which are presented in details in the following.

1) Adversarial Loss: As shown in Fig. 2(a), there are two discriminators for image and feature spaces, respectively. The discriminator in the image space ($D_{\mathcal{I}}$) aims to distinguish between the original normal image and the one (*i.e.*, $\mathcal{M}_D(z)$ or x') reconstructed from a vector sampled from the Gaussian distribution. The discriminator implemented in the feature space ($D_{\mathcal{F}}$) shares the similar idea, which is trained to differentiate between the feature encoded from a normal image (*i.e.*, $\mathcal{M}_E(x)$ or z') and the one sampled from the Gaussian distribution. The least squared adversarial loss is used to optimize the discriminators:

$$\begin{aligned}\min_{D_{\mathcal{I}}} \mathcal{L}_{D_{\mathcal{I}}} = \; & \frac{1}{2}\mathbb{E}_{z\sim p_z}|D_{\mathcal{I}}(\mathcal{M}_D(z)) - a|^2 \\ & + \frac{1}{2}\mathbb{E}_{x\sim\mathcal{D}}|D_{\mathcal{I}}(x) - b|^2, \end{aligned} \tag{1}$$

and

$$\begin{aligned}\min_{D_{\mathcal{F}}} \mathcal{L}_{D_{\mathcal{F}}} = \; & \frac{1}{2}\mathbb{E}_{I\sim\mathcal{D}}|D_{\mathcal{F}}(\mathcal{M}_E(x)) - a|^2 \\ & + \frac{1}{2}\mathbb{E}_{z\sim p_z}|D_{\mathcal{F}}(z) - b|^2, \end{aligned} \tag{2}$$

where x and z are the original image and sampled vector, respectively; a and b are set to 0 and 1, respectively. Similar to [41], a standard multivariate Gaussian distribution $\mathcal{N}_n(\vec{\mu}, \vec{\Sigma})$ is chosen for distribution p_z where each component of z is a zero-mean unit-variance normally distributed random variable, *i.e.*, $z_n \sim \mathcal{N}(0, 1)$ for all n.

Through the adversarial training with the discriminators, the capacity of encoder-decoder on feature extraction and image reconstruction can be improved accordingly. Ideally, the encoded feature $\mathcal{M}_E(x)$ and translated image $\mathcal{M}_D(z)$ can gradually fool the corresponding discriminator by approaching the similar distributions as the real ones. The adversarial loss supervising this minimax game between encoder-decoder and discriminators is formulated as following:

$$\begin{aligned}\min_{\mathcal{M}_E}\min_{\mathcal{M}_D} \mathcal{L}_{adv} = \; & \frac{1}{2}\mathbb{E}_{x\sim\mathcal{D}}|D_{\mathcal{F}}(\mathcal{M}_E(x)) - c|^2 \\ & + \frac{1}{2}\mathbb{E}_{z\sim p_z}|D_{\mathcal{I}}(\mathcal{M}_D(z)) - c|^2, \end{aligned} \tag{3}$$

where c is set to 1.

2) Reconstruction Loss: Apart from the adversarial loss, which aligns the distribution of encoded feature and synthesized image to the real ones, we also calculate the reconstruction errors to encourage the encoder-decoder to exploit useful information from data for plausible reconstruction results. For the image space reconstruction, we employ the structure similarity loss for the encoder-decoder. Different from the widely-used pixel-wise $\mathcal{L}_1$ loss, the structure similarity loss takes the luminance, contrast and anatomical information into account [42], which is less sensitive to the location-shift between the original image and its reconstruction and thereby enables the network easier to converge. Hence, the model trained with the structure similarity loss tends to focus on the global information (*e.g.*, anatomical structures) rather than the local features (*e.g.*, pixel intensity) during the image reconstruction. The structure similarity loss optimizing the encoder-decoder can be formulated as:

$$\begin{aligned}\min \mathcal{L}_{str} &= -SSIM(x, \hat{x}) \\ &= -SSIM(x, \mathcal{M}_D(\mathcal{M}_E(x))). \end{aligned} \tag{4}$$

Here, the $SSIM(\cdot)$ is defined as:

$$SSIM(x, \hat{x}) = \frac{(2\mu_x\mu_{\hat{x}} + c_1)(2\sigma_{x\hat{x}} + c_2)}{(\mu_x^2 + \mu_{\hat{x}}^2 + c_1)(\sigma_x^2 + \sigma_{\hat{x}}^2 + c_2)}, \tag{5}$$

where x and $\hat{x}$ are the input and reconstructed images, respectively; μ is the mean intensity of image, σ denotes the standard deviation of image and $\sigma_{x\hat{x}}$ stands for the covariance of two images. The constants c_1 and c_2 are set to 0.01 and 0.03, respectively.

For the adversarial reconstruction in the feature space, the element-wise $\mathcal{L}_1$ loss is utilized to minimize the distance between the reconstructed feature $\hat{z}$ and the sampled vector z, which can be defined as:

$$\min \mathcal{L}_{fea} = |\hat{z} - z| = |\mathcal{M}_E(\mathcal{M}_D(z)) - z|. \tag{6}$$

3) Center Constraint Loss: A compact cluster of features encoded from images belonging to normal class in the latent feature space makes the model easier to identify the outliers as anomalies. To this end, we propose to adopt a center constraint on the latent space to push the encoded features towards the center of the cluster, which reduces the intra-class dissimilarity.

The feature center loss can be optimized with:

$$\min \mathcal{L}_{ct} = \frac{1}{2}||z' - C||_2^2 = \frac{1}{2}||\mathcal{M}_E(x) - C||_2^2, \tag{7}$$

where $C \in \mathbb{R}^n$ denotes the normal class center on the latent feature manifold and is updated based on the feature distances following [43].

4) Restoration Loss: As aforementioned, to further improve the model ability, we propose a self-supervised learning module for the framework. The input images are perturbed by a transformation randomly selected from three candidates (*i.e.*, in-painting, local pixel shuffling and non-linear intensity transformation). The encoder $\mathcal{M}_E$ and decoder $\mathcal{M}'_D$ are required to restore the original input image from the perturbed one. Here, the encoder $\mathcal{M}_E$ shares weights with the one for image and feature reconstruction, while the decoder $\mathcal{M}'_D$ has the same structure to $\mathcal{M}_D$ but with different network parameters. The original input image x is used as the supervision signal for the self-supervised learning module.

The pixel-wise $\mathcal{L}_1$ loss is adopted to supervise the proxy tasks to learn a better representation, which can be written as:

$$\min \mathcal{L}_{self} = |x'_s - x| = |\mathcal{M}'_D(\mathcal{M}_E(x_s)) - x|, \tag{8}$$

where x_s is the perturbed input image and x'_s is the restored result.

Finally, with the previously defined loss functions, the overall objective to optimize the encoder $\mathcal{M}_E$ and decoder $\mathcal{M}_D$ can be formulated as:

$$\mathcal{L} = \alpha\mathcal{L}_{adv} + \beta\mathcal{L}_{str} + \gamma\mathcal{L}_{fea} + \delta\mathcal{L}_{ct} + \eta\mathcal{L}_{self}, \tag{9}$$

where α, β, γ, δ, η are the loss weights.

E. Test Stage

The test phase is illustrated in Fig. 2 (b). A test image x_t is fed to the encoder $\mathcal{M}_E$ and decoder $\mathcal{M}_D$, which yields a reconstructed image $\hat{x}_t$ and an encoded feature representation z_t. The reconstructed image $\hat{x}_t$ is then passed through the $\mathcal{M}_E$ again to generate the feature $\hat{z}_t$. The reconstruction error between z_t and $\hat{z}_t$ is adopted as a metric, namely anomaly score $\mathcal{A}$, for anomaly detection, which is similar to most existing studies [4], [8]. Compared with the image-based reconstruction approaches, the feature $\hat{z}_t$ obtained with three translations ($x_t \rightarrow z_t$, $z_t \rightarrow \hat{x}_t$ and $\hat{x}_t \rightarrow \hat{z}_t$) can amplify the reconstruction error of anomalous data, which leads to a higher sensitivity for anomalies. The feature $\hat{z}_t$ with low reconstruction error ($\mathcal{A} < \tau$) to the encoded one z_t after three translations, *i.e.*, translation-consistent, is treated as normal data and otherwise anomaly.

IV. Experiments

In this section, we validate the proposed SALAD framework on detecting anomalies from two modalities of medical data, *i.e.*, optical coherence tomography (OCT) and chest X-ray images. First, the detail of each dataset is introduced, and then the experimental results comparing with state-of-the-art methods are given. Finally, an ablation study of our model is conducted to verify the contribution made by each of the proposed loss functions.

A. Datasets

Two publicly available datasets, *i.e.*, OCT dataset and chest X-ray dataset [44], are adopted to evaluate the proposed SALAD approach.

1) OCT Dataset: The dataset is categorized into four classes, including normal, drusen, diabetic macular edema (DME) and choroidal neovascularization (CNV), which are already separated to the training and test sets for a fair comparison. We use the 17,922 normal images in the training set to train the SALAD framework and evaluate its performance on the public test set, consisting of 769 images from four classes. Since the images from the dataset are of different sizes, we uniform them to the same size of 256 × 256 pixels.

2) Chest X-ray Dataset: The chest X-ray dataset consists of normal and pneumonia images collected from 6,480 patients. The image-level annotations are provided for each subject. There are 1,349 normal images in the training set, while 234 normal images and 390 pneumonia images are contained in the test set. Similar to the OCT dataset, we resize the chest X-ray images to a standard size of 256 × 256 pixels. Compared to OCT images, the chest X-ray images, containing multiple anatomical structures such as lung and bone, are more complicated. Such images increase the difficulty for accurate anomaly detection, since the lesions may appear at different positions in chest X-ray images.

B. Experimental Settings

Our approach is implemented using the PyTorch toolbox. For the two datasets, we use the bicubic interpolation for image size standardization and uniform the pixel values of an image to the range $[-1, 1]$. The Adam optimizer [45] with a learning rate of 0.0002 is adopted for network optimization. The model is trained for 200 epochs with mini-batch size of 64. We keep the same learning rate for the first 100 epochs and linearly decay the rate to zero over the next 100 epochs. Following the hyperparameter settings in [39], the loss weights (β, γ, δ, and η) in our experiments are all set to 10 except α is set to 1. The encoder and decoder consist of eight convolutional or deconvolutional layers, respectively. Each convolutional or deconvolutional layer is with the kernel size of 4 and stride of 2 to downsample or upsample the feature maps and it finally leads to a one-dimensional vector or an image. The InstanceNorm and LeakyReLU are used after each convolutional/deconvolutional layer. The PatchGAN [46] is used as the backbone for image space discriminator. The feature discriminator is a multi-layer perceptron (MLP) network.

In our experiments, apart from the original auto-encoder and variational auto-encoder (VAE) [41] using the reconstruction error for anomaly detection, several state-of-the-art anomaly detection methods are involved for comparison:

- **Auto-encoder** consists of two sub-networks, *i.e.* encoder and decoder, to learn the reconstruction of an image by mapping the input to latent space and remapping back to image space.
- **VAE** [41] is a variant of auto-encoder which not only penalizes the reconstruction difference but also regularizes the latent encoding distribution. This property makes the model capable of learning the data distribution instead of just remember the training data.
- **f-AnoGAN** [8] is a generative adversarial framework, which learns a mapping between image and latent spaces. The f-AnoGAN adopts the image and feature reconstruction error as the metric to identify anomalous images.
- **Ganomaly** [4] learns the image and latent spaces jointly. It has an encoder-decoder-encoder structure, which generates reconstructed latent vector to capture features in the latent space for anomaly detection.

The area under the receiver operating characteristic (ROC) curve (AUC), F1-score, average classification accuracy (ACC), sensitivity (SEN) and specificity (SPE) are adopted as the evaluation metrics. AUC and F1-score are the evaluation of overall performance, while SEN and SPE focus on the positive samples and negative samples, respectively. Our results are the average score of three runs. The threshold used for evaluation is determined based on the best value of F1-score.

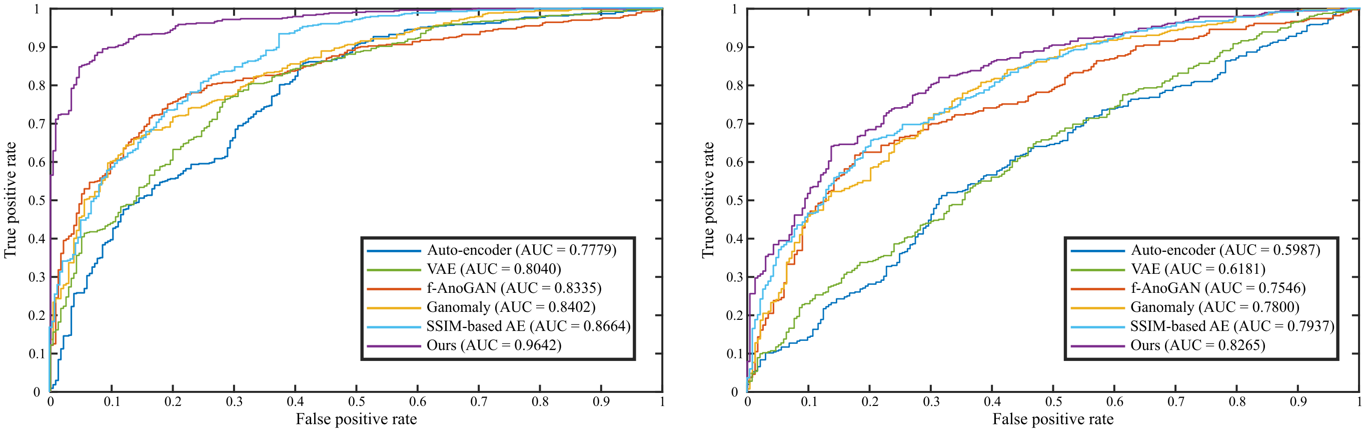
C. Comparison With State of the art

In this section, we compare the performance of our SALAD framework with the state-of-the-art methods on OCT and chest X-ray datasets.

1) Results on OCT Dataset: We first evaluate the performance of our SALAD and state-of-the-art approaches on the OCT dataset. The experimental results are presented in Table I. It can be observed that the image-reconstruction-based approaches (*i.e.*, the original auto-encoder, VAE and

TABLE I
CLASSIFICATION PERFORMANCE (AUC, F1 (%), ACC (%), SEN (%) AND SPE (%)) YIELDED BY DIFFERENT ALGORITHMS ON OCT AND CHEST X-RAY DATASETS

	OCT					**Chest X-Ray**				
	AUC	F1	ACC	SEN	SPE	AUC	F1	ACC	SEN	SPE
Auto-encoder	0.7779	85.79	78.28	94.38	41.70	0.5987	77.20	63.40	**98.97**	3.86
VAE [41]	0.8040	85.55	77.63	95.32	37.45	0.6181	77.37	64.04	98.21	6.87
f-AnoGAN [8]	0.8335	84.73	77.50	89.89	49.36	0.7546	81.00	74.00	88.97	36.48
Ganomaly [4]	0.8402	85.76	77.24	**98.69**	28.51	0.7800	78.97	69.98	90.00	48.93
SSIM-based AE	0.8664	89.10	83.88	94.94	58.72	0.7937	81.16	72.87	93.33	38.62
SALAD (*ours*)	**0.9642**	**93.42**	**90.64**	95.69	**79.15**	**0.8265**	**82.14**	**75.92**	88.46	**54.94**



(a) ROC curve on OCT dataset

(b) ROC curve on chest X-ray dataset

Fig. 4. Receiver operating characteristic curve (ROC) of the comparison methods on OCT and chest X-ray datasets.

f-AnoGAN) achieve relatively lower classification accuracy than the ones using information extracted from both image and latent spaces (*i.e.*, Ganomaly and our SALAD). The experimental results demonstrate that the image reconstruction error may be insufficient for the robust anomaly detection, which is also revealed by the existing study [4].

With the aid of information extracted from the latent feature space, the Ganomaly model yields a better anomaly detection performance—an AUC of 0.8402 is achieved. Compared with Ganomaly, our SALAD framework formulates a dual-track translation (*i.e.*, image-feature-image and feature-image-feature) and implements manifold constraints to enforce the model to deeply exploit useful information for anomaly detection from the image and latent spaces, resulting in a remarkable improvement of AUC (+0.1240). The center loss diminishes the distances among the representations of normal data, which amplifies the feature reconstruction error of anomalous data from the other side. A thorough analysis of the contribution made by each loss function is presented in the following section.

Furthermore, apart from AUC, our SALAD approach achieves the best F1-score (93.42%) and ACC (90.64%) among the listed algorithms, which are 7.63% and 12.36% higher than the runner-up SSIM-based auto-encoder in F1-score and ACC, respectively. Although Ganomaly yields the best sensitivity of 98.69% with a significant sacrifice in specificity, the result of our SALAD approach (*i.e.*, 95.69%) is still comparable. We also draw the ROC curve in Fig. 4 (a) for performance evaluation. It can be observed that our SALAD approach outperforms the state-of-the-art methods by a large margin.

2) Results on Chest X-ray Dataset: We further test our SALAD approach on the chest X-ray dataset. The experimental results are presented in Table I. The corresponding ROC curve is also drawn in Fig. 4 (b). As shown in Table I, similar trend of accuracy variation to the OCT dataset is observed—our approach using both image and feature information together with center constraint surpasses the image-reconstruction-based methods. Compared with the results on the OCT dataset, the anomaly detection accuracy of the listed models consistently decreases. The underlying reason for the performance degradation is that the content of chest X-ray image is more complicated than the OCT image, which increases the difficulty for feature embedding and thereby decreases the performance not only for image-reconstruction-based approaches (*e.g.*, auto-encoder and f-AnoGAN) but also the ones utilizing the information from both spaces (Ganomaly and SALAD). Nevertheless, the proposed SALAD approach obtains the highest AUC (0.8265), F1-score (82.14%) and ACC (75.92%) on the chest X-ray dataset, validating the robustness of the proposed SALAD approach. We also display the standard deviation

TABLE II
The Standard Deviation of Three-Run Experimental Results on the OCT and Chest X-ray Datasets With Evaluation Metrics of AUC, F1 (%), ACC (%), SEN (%), SPE (%)

	AUC	F1	ACC	SEN	SPE
OCT	0.0111	1.08	1.58	1.14	0.05
Chest X-ray	0.0083	0.32	0.89	1.35	0.05

TABLE III
Anomaly Detection Performance of Our SALAD Framework With Different Loss Functions in the Image and Feature Spaces on the OCT Dataset With AUC, F1 (%), ACC (%), SEN (%) and SPE (%)[2]

	AUC	F1	ACC	SEN	SPE
Image space					
$\mathcal{L}_1$	0.7447	84.10	74.77	96.07	26.38
$\mathcal{L}_{str}$	0.8886	89.03	83.75	94.94	58.30
Feature space					
$\mathcal{L}_{str}$ + $\mathcal{L}_{KL}$	0.9086	89.10	84.92	88.76	**76.17**
$\mathcal{L}_{str}$ + $\mathcal{L}_{ct}$	**0.9328**	**91.50**	**87.52**	**96.82**	66.38

of three-run experimental results on the OCT and Chest X-Ray datasets, respectively, with different evaluation metrics in Table. II.

D. Ablation Study

We conduct a thorough ablation study on the OCT dataset to evaluate the contribution made by each component of the proposed SALAD framework and present the evaluation results in this section.

1) Losses in Image and Feature Spaces: We evaluate the performance of SALAD framework with different loss functions used for image and feature spaces, while the self-supervised learning module is excluded for all the experiments (*i.e.*, without $\mathcal{L}_{self}$). Besides, the feature reconstruction loss $\mathcal{L}_{fea}$ and adversarial loss $\mathcal{L}_{adv}$ are used as default settings. The evaluation results are shown in Table III.

a) Image space reconstruction: It can be observed that the $\mathcal{L}_1$-only model yields a relatively low AUC (0.7447), which is increased to 0.8886 by using the $\mathcal{L}_{str}$ instead. The large difference between the $\mathcal{L}_1$ model and the $\mathcal{L}_{str}$ model validates the benefit of our $\mathcal{L}_{str}$ loss. The same trend can also be found in auto-encoder, which increase the performance from 0.7779 to 0.8664 in Table I. The results demonstrate that the $\mathcal{L}_{str}$ enables the network to learn the richer information (*e.g.*, texture and anatomical information) from the image and extract more robust feature representation for anomaly detection than the pixel-wise $\mathcal{L}_1$ loss.

b) Feature space constraint: The center constraint loss $\mathcal{L}_{ct}$ is proposed to engage the prior information in the feature space, which enforces the model to compact the cluster of feature representation extracted from normal images. As shown in Table III, it boosts the AUC to 0.9328, which is about 0.05 higher than the $\mathcal{L}_{str}$-only model. To further demonstrate the superior performance of the proposed center loss, we select the widely-used KL-divergence loss $\mathcal{L}_{KL}$ for comparison. Although the model trained with KL-divergence constraint (*i.e.* $\mathcal{L}_{str} + \mathcal{L}_{KL}$) achieves an AUC of 0.9086, the performance of our model (*i.e.* $\mathcal{L}_{str} + \mathcal{L}_{ct}$) is higher with an improvement of 0.03.

TABLE IV
Performance (AUC, F1 (%), ACC (%), SEN (%) and SPE (%)) of Models Trained With Different Settings of Self-Supervised Learning Module on the OCT Dataset, Where Structure Similarity in Image Space and Center Constraint in Feature Space Are Utilized (SSL: Self-Supervised Learning Module; E.: Encoder $\mathcal{M}_E$; D.: Decoder $\mathcal{M}_D$)

		AUC	F1	ACC	SEN	SPE
w/o SSL		0.9328	91.5	87.52	**96.82**	66.38
w/ SSL	E. + D.	0.9309	91.27	87.26	95.88	67.66
	E.	**0.9642**	**93.42**	**90.64**	95.69	**79.15**

The experimental results thereby validate the effectiveness of the proposed structure similarity loss and center constraint, which successfully assist the network to learn a robust feature representation from normal images for anomaly detection.

2) Self-Supervised Learning Module: The self-supervised learning module is implemented to fully exploit useful information from the raw normal data for networks to accurately detect anomalous images. As aforementioned, the self-supervised learning module consists of an encoder sharing weights with the image reconstruction branch and a specific decoder, due to the different optimization directions between the image reconstruction and proxy task. To validate the previous statement, we evaluate the performance of SALAD framework using weight-sharing encoder-decoder and weight-sharing encoder-only, respectively. As shown in Table IV, the anomaly detection performance of SALAD framework degrades to an AUC of 0.9309 using weight-sharing encoder-decoder, which is just comparable to the SALAD without the self-supervised learning module. In contrast, our SALAD only sharing the encoder achieves a significant improvement (+0.0314) for AUC. To further validate the sharing-encoder-only design, we conduct an extra experiment—the self-supervised learning module shares the parameters of the encoder and the last few layers of the decoder with the image reconstruction branch. Such a framework yields an AUC of 0.9436—higher than the sharing-weight encoder-decoder SALAD but still lower than the encoder-only one, which demonstrates the negative effect caused by sharing the parameter of decoder between the two tasks.

3) Metric for Anomaly Detection: Since our SALAD involves different errors, such as image-reconstruction and feature-reconstruction errors, both of which can be used as the metric for anomaly detection, we conduct an ablation study to evaluate the performance of SALAD using different metrics. Three kinds of errors are involved for comparison:

a) Error in the image space: ($\mathcal{A}_{\mathcal{I}}$) calculates the reconstruction error of images $\mathcal{A}_{\mathcal{I}} = |x_t - \hat{x}_t|$.

[2] Note that all the models compared in the table are trained together with feature reconstruction loss $\mathcal{L}_{fea}$ and adversarial loss $\mathcal{L}_{adv}$ while not using self-supervised learning module (*i.e.*, without $\mathcal{L}_{self}$).

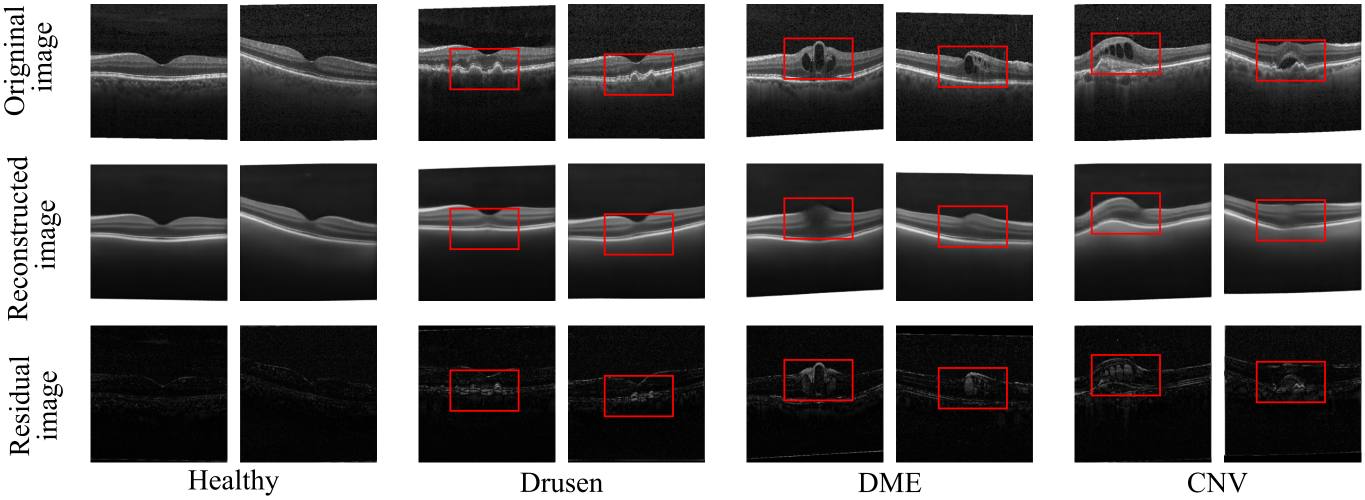


Fig. 5. Reconstruction results of our model on the OCT dataset including two exemplar images for normal, drusen, DME, CNV cases, respectively.

TABLE V
ABLATION STUDY ON DIFFERENT METRICS USED FOR ANOMALY SCORE $\mathcal{A}$ ON THE OCT DATASET WITH AUC, F1 (%), ACC (%), SEN (%) AND SPE (%)

	AUC	F1	ACC	SEN	SPE
$\mathcal{A}_{\mathcal{I}}$	0.8696	88.50	82.96	94.38	57.02
$\mathcal{A}_{\mathcal{C}}$	0.8299	85.54	78.67	90.82	51.06
$\mathcal{A}_{\mathcal{F}}$	**0.9642**	**93.42**	**90.64**	**95.69**	**79.15**

b) Error in the center distance: ($\mathcal{A}_{\mathcal{C}}$) calculates the distance between feature representation z_t generated by $\mathcal{M}_E$ and the center C obtained from normal training data, *i.e.*, $\mathcal{A}_{\mathcal{C}} = |z_t - C|$.

c) Error in the feature space: ($\mathcal{A}_{\mathcal{F}}$) calculates the reconstruction error of feature representations $\mathcal{A}_{\mathcal{F}} = |z_t - \hat{z}_t|$.

The experimental results are shown in Table V. Compared with the center distance, the reconstruction-based metrics provide a more intuitive measurement for anomaly detection, which consistently surpass the center distance metric. Compared with image space reconstruction error ($\mathcal{A}_{\mathcal{I}}$), the feature space one ($\mathcal{A}_{\mathcal{F}}$) alleviates the influence caused by local information, such as position shifting and local pixel difference, and therefore achieves a more robust anomaly detection performance. Hence, the proposed SALAD uses the feature space reconstruction error as the metric.

E. Visualization

In this section, we visualize the reconstructed images and learned features to further demonstrate the effectiveness of our SALAD framework.

1) Visualization of Reconstructed Images: The OCT images reconstructed by our SALAD approach are shown in Fig. 5, including two exemplars of each category (*i.e.*, normal, drusen, DME and CNV). The original images x_t, reconstructed images $\hat{x}_t$ and residual images $(x_t - \hat{x}_t)$ are presented in the first, second and third rows of the figure, respectively. We can observe that the SALAD highlights the lesion areas with a large image-based reconstruction error, as illustrated in the residual images. Hence, the following image-to-feature translation can amplify the error and accordingly increase the model sensitivity to anomalies. Conversely, the normal data is well reconstructed with negligible reconstruction error. The reconstructed chest X-ray images are presented in Fig. 6. We have similar observations to the OCT reconstruction results—our SALAD framework can distinguish pneumonia images with large reconstruction difference.

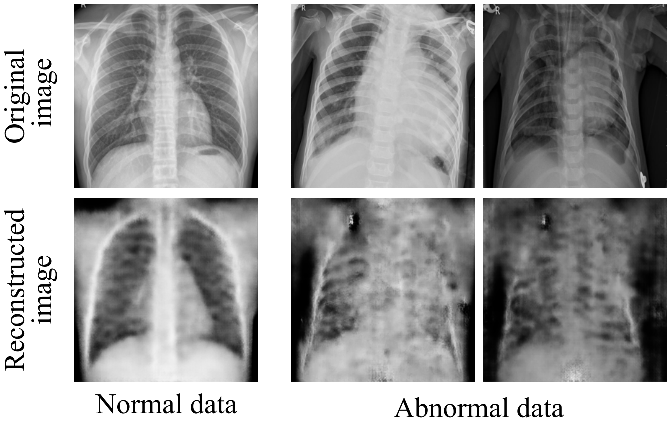


Fig. 6. Reconstruction results of our model on the chest X-ray dataset where samples of each class (normal and anomalous data) are displayed.

We also quantitatively analyze the reconstruction errors ($\mathcal{A}$) for normal and anomalous data in latent space, respectively, as presented in Table VI, where the mean error and standard deviation for different classes are calculated. For the OCT dataset, the reconstruction error of normal images is 0.0112 ± 0.0018, which is significantly smaller than those of the anomalous classes (*i.e.*, CNV, DME and drusen). The same trend is observed on the chest X-ray dataset. The error difference between normal images and anomalous images illustrates that our approach can well distinguish the anomalies from normal data with the feature space reconstruction error.

2) Feature Visualization Using t-SNE: The center loss compacts the normal cluster in the latent space, which leads to

TABLE VI
COMPARISON OF THE RECONSTRUCTION ERRORS IN THE FEATURE SPACE BETWEEN NORMAL DATA AND ANOMALOUS DATA WITH MEAN AND STANDARD DEVIATION ON OCT AND CHEST X-RAY DATASETS (PNEUM.: PNEUMONIA)

OCT		**Chest X-Ray**	
Normal	0.0112± 0.0018	Normal	0.0250 ± 0.0067
CNV	0.0229 ± 0.0051	Pneum.	0.0449 ± 0.0278
DME	0.0200 ± 0.0045		
Drusen	0.0162 ± 0.0028		

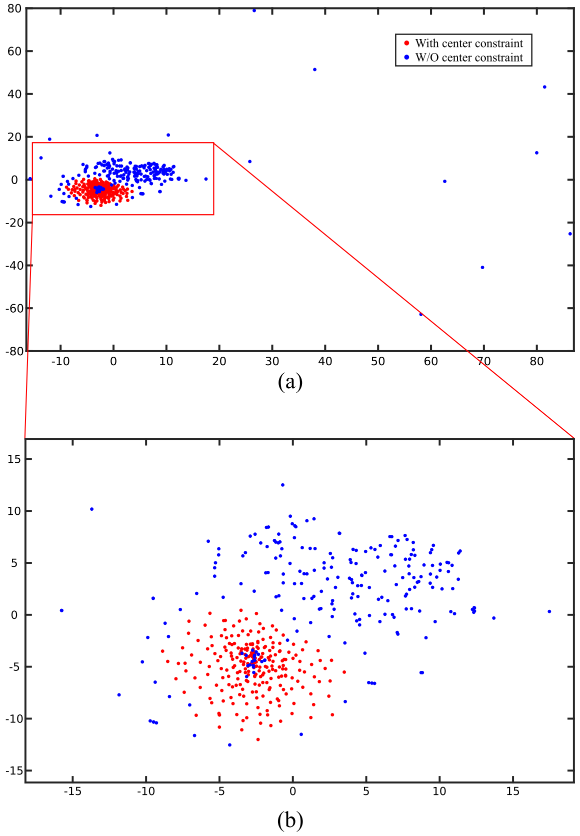


Fig. 7. Visualization of compressed feature distribution generated by model trained with center constraint (red points) and without center constraint (blue points) via t-SNE. (a) The overall feature distribution. (b) Zoomed-in view of feature cloud.

an easier identification of anomalies. To validate this claim, we use the t-SNE method [47] to visualize the normal cluster with and without the proposed center loss for illustration purpose. The visualization result is shown in Fig. 7—the features embedded by the model trained with and without the proposed center constraint $\mathcal{L}_{ct}$ are drawn with *red* and *blue* points, respectively. It can be observed that the cluster generated by the model trained without the proposed center loss has a sparse distribution with lots of outliers. Different from that, the features generated by the SALAD using the center constraint locate closer to each other and lead to a compact cluster, which therefore decreases the possibility of false-positive and increases model sensitivity to anomalies.

V. CONCLUSION

In this paper, we proposed a framework, namely **SALAD**, extracting **S**elf-supervised and tr **A**ns**L**ation-consistent features for **A**nomaly **D**etection, which is only trained with the normal data. We considered the reconstruction tasks on both the image and feature spaces, which helps our model to learn a meaningful representation. The reconstruction error in the latent feature space is utilized for distinguishing the anomalies in the test stage. In order to pay attention to the structure but not detailed difference on the reconstructed map, we introduced a structure similarity loss instead of pixel-wise $\mathcal{L}_1$ loss. By engaging center constraint on the feature space and self-supervised learning module, our approach enforces the encoder to learn a desirable feature representation of normal data, which further improves the classification performance. Experiments on different medical image datasets demonstrated the effectiveness of our approach. In the future work, we will explore the model ability on more complex image content and investigate on more medical applications.

REFERENCES

[1] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "ImageNet: A large-scale hierarchical image database," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2009, pp. 248–255.

[2] V. Gulshan *et al.*, "Development and validation of a deep learning algorithm for detection of diabetic retinopathy in retinal fundus photographs," *J. Amer. Med. Assoc.*, vol. 316, no. 22, pp. 2402–2410, 2016.

[3] A. Esteva *et al.*, "Dermatologist-level classification of skin cancer with deep neural networks," *Nature*, vol. 542, no. 7639, pp. 115–118, Feb. 2017.

[4] S. Akcay, A. Atapour-Abarghouei, and T. P. Breckon, "Ganomaly: Semi-supervised anomaly detection via adversarial training," in *Proc. Asian Conf. Comput. Vis.* Cham, Switzerland: Springer, 2018, pp. 622–637.

[5] M. Sabokrou, M. Khalooei, M. Fathy, and E. Adeli, "Adversarially learned one-class classifier for novelty detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 3379–3388.

[6] T. Schlegl, P. Seeböck, S. M. Waldstein, U. Schmidt-Erfurth, and G. Langs, "Unsupervised anomaly detection with generative adversarial networks to guide marker discovery," in *Proc. Int. Conf. Inf. Process. Med. Imag.* Cham, Switzerland: Springer, 2017, pp. 146–157.

[7] D. Zimmerer, F. Isensee, J. Petersen, S. Kohl, and K. Maier-Hein, "Unsupervised anomaly localization using variational auto-encoders," in *Int. Conf. Med. Image Comput. Comput. Assist. Intervent.* Cham, Switzerland: Springer, 2019, pp. 289–297.

[8] T. Schlegl, P. Seeböck, S. M. Waldstein, G. Langs, and U. Schmidt-Erfurth, "F-AnoGAN: Fast unsupervised anomaly detection with generative adversarial networks," *Med. Image Anal.*, vol. 54, pp. 30–44, May 2019.

[9] J. O. Awoyemi, A. O. Adetunmbi, and S. A. Oluwadare, "Credit card fraud detection using machine learning techniques: A comparative analysis," in *Proc. Int. Conf. Comput. Netw. Informat.*, 2017, pp. 1–9.

[10] A. Javaid, Q. Niyaz, W. Sun, and M. Alam, "A deep learning approach for network intrusion detection system," in *Proc. 9th EAI Int. Conf. Bio-inspired Inf. Commun. Technol. (formerly BIONETICS)*, 2016, pp. 21–26.

[11] W. Li, V. Mahadevan, and N. Vasconcelos, "Anomaly detection and localization in crowded scenes," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 36, no. 1, pp. 18–32, Jan. 2014.

[12] B. Schölkopf, R. C. Williamson, A. J. Smola, J. Shawe-Taylor, and J. C. Platt, "Support vector method for novelty detection," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 12, 1999, pp. 582–588.

[13] L. Ruff *et al.*, "Deep one-class classification," in *Proc. Int. Conf. Mach. Learn.*, 2018, pp. 4393–4402.

[14] L. Xiong, B. Póczos, and J. G. Schneider, "Group anomaly detection using flexible genre models," in *Proc. Adv. Neural Inf. Process. Syst.*, 2011, pp. 1071–1079.

[15] J. Wen, Z. Lai, Z. Ming, W. K. Wong, and Z. Zhong, "Directional Gaussian model for automatic speeding event detection," *IEEE Trans. Inf. Forensics Security*, vol. 12, no. 10, pp. 2292–2307, Oct. 2017.

[16] B. Zong *et al.*, "Deep autoencoding Gaussian mixture model for unsupervised anomaly detection," in *Proc. Int. Conf. Learn. Represent.*, 2018, pp. 1–19.

[17] D. Sidibé *et al.*, "An anomaly detection approach for the identification of DME patients using spectral domain optical coherence tomography images," *Comput. Methods Programs Biomed.*, vol. 139, pp. 109–117, Feb. 2017.

[18] B. Zhao, L. Fei-Fei, and E. P. Xing, "Online detection of unusual events in videos via dynamic sparse coding," in *Proc. CVPR*, Jun. 2011, pp. 3313–3320.

[19] Y. Cong, J. Yuan, and J. Liu, "Sparse reconstruction cost for abnormal event detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2011, pp. 3449–3456.

[20] P. Bergmann, M. Fauser, D. Sattlegger, and C. Steger, "Uninformed students: Student-teacher anomaly detection with discriminative latent embeddings," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 4183–4192.

[21] H. Zenati, M. Romain, C.-S. Foo, B. Lecouat, and V. Chandrasekhar, "Adversarially learned anomaly detection," in *Proc. IEEE Int. Conf. Data Mining (ICDM)*, Nov. 2018, pp. 727–736.

[22] C. Zhou and R. C. Paffenroth, "Anomaly detection with robust deep autoencoders," in *Proc. 23rd ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, Aug. 2017, pp. 665–674.

[23] J. An and S. Cho, "Variational autoencoder based anomaly detection using reconstruction probability," SNU Data Mining Center, Seoul, South Korea, Tech. Rep. 1, 2015.

[24] A. Dosovitskiy and T. Brox, "Generating images with perceptual similarity metrics based on deep networks," in *Proc. Adv. Neural Inf. Process. Syst.*, 2016, pp. 658–666.

[25] P. Bergmann, S. Löwe, M. Fauser, D. Sattlegger, and C. Steger, "Improving unsupervised defect segmentation by applying structural similarity to autoencoders," 2018, *arXiv:1807.02011*. [Online]. Available: http://arxiv.org/abs/1807.02011

[26] K. Zhou *et al.*, "Sparse-GAN: Sparsity-constrained generative adversarial network for anomaly detection in retinal OCT image," in *Proc. IEEE 17th Int. Symp. Biomed. Imag.*, Apr. 2020, pp. 1227–1231.

[27] P. Perera, R. Nallapati, and B. Xiang, "OCGAN: One-class novelty detection using GANs with constrained latent representations," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 2898–2906.

[28] C. Doersch, A. Gupta, and A. A. Efros, "Unsupervised visual representation learning by context prediction," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Dec. 2015, pp. 1422–1430.

[29] M. Noroozi and P. Favaro, "Unsupervised learning of visual representations by solving Jigsaw puzzles," in *Proc. Eur. Conf. Comput. Vis.* Cham, Switzerland: Springer, 2016, pp. 69–84.

[30] G. Larsson, M. Maire, and G. Shakhnarovich, "Colorization as a proxy task for visual understanding," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 6874–6883.

[31] D. Kim, D. Cho, D. Yoo, and I. S. Kweon, "Learning image representations by completing damaged jigsaw puzzles," in *Proc. IEEE Winter Conf. Appl. Comput. Vis. (WACV)*, Mar. 2018, pp. 793–802.

[32] C. Wei *et al.*, "Iterative reorganization with weak spatial constraints: Solving arbitrary Jigsaw puzzles for unsupervised representation learning," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2019, pp. 1910–1919.

[33] H. Wang, G. Pang, C. Shen, and C. Ma, "Unsupervised representation learning by predicting random distances," in *Proc. 29th Int. Joint Conf. Artif. Intell.*, Jul. 2020, pp. 1–4.

[34] L. Zhang, G.-J. Qi, L. Wang, and J. Luo, "AET vs. AED: Unsupervised representation learning by auto-encoding transformations rather than data," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Dec. 2019, pp. 2547–2555.

[35] X. Zhan, X. Pan, Z. Liu, D. Lin, and C. C. Loy, "Self-supervised learning via conditional motion propagation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 1881–1889.

[36] P. Zhang, F. Wang, and Y. Zheng, "Self supervised deep representation learning for fine-grained body part recognition," in *Proc. IEEE 14th Int. Symp. Biomed. Imag. (ISBI)*, Apr. 2017, pp. 578–582.

[37] H. Spitzer, K. Kiwitz, K. Amunts, S. Harmeling, and T. Dickscheid, "Improving cytoarchitectonic segmentation of human brain areas with self-supervised Siamese networks," in *Proc. Int. Conf. Med. Image Comput. Comput. Assist. Intervent.* Cham, Switzerland: Springer, 2018, pp. 663–671.

[38] Z. Zhou *et al.*, "Models genesis: Generic autodidactic models for 3D medical image analysis," in *Proc. Int. Conf. Med. Image Comput. Comput. Assist. Intervent.* Cham, Switzerland: Springer, 2019, pp. 384–393.

[39] J.-Y. Zhu, T. Park, P. Isola, and A. A. Efros, "Unpaired image-to-image translation using cycle-consistent adversarial networks," in *Proc. IEEE Int. Conf. Comput. Vis.*, Jun. 2017, pp. 2223–2232.

[40] M. E. Mortenson, *Mathematics for Computer Graphics Applications*. New York, NY, USA: Industrial Press, 1999.

[41] D. P. Kingma and M. Welling, "Auto-encoding variational Bayes," in *Proc. Int. Conf. Learn. Represent.*, 2014, pp. 1–14.

[42] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli, "Image quality assessment: From error visibility to structural similarity," *IEEE Trans. Image Process.*, vol. 13, no. 4, pp. 600–612, Apr. 2004.

[43] Y. Wen, K. Zhang, Z. Li, and Y. Qiao, "A discriminative feature learning approach for deep face recognition," in *Proc. Eur. Conf. Comput. Vis.* Cham, Switzerland: Springer, 2016, pp. 499–515.

[44] D. S. Kermany *et al.*, "Identifying medical diagnoses and treatable diseases by image-based deep learning," *Cell*, vol. 172, no. 5, pp. 1122–1131, 2018.

[45] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," 2014, *arXiv:1412.6980*. [Online]. Available: http://arxiv.org/abs/1412.6980

[46] C. Li and M. Wand, "Precomputed real-time texture synthesis with Markovian generative adversarial networks," in *Proc. Eur. Conf. Comput. Vis.* Cham, Switzerland: Springer, 2016, pp. 702–716.

[47] L. van der Maaten and G. Hinton, "Visualizing data using t-SNE," *J. Mach. Learn. Res.*, vol. 9, pp. 2579–2605, Nov. 2008.