Unsupervised Local Discrimination for Medical Images

Huai Chen, Renzhen Wang, Xiuying Wang, Jieyu Li, Qu Fang, Hui Li, Jianhao Bai, Qing Peng, Deyu Meng, Member, IEEE, and Lisheng Wang

Abstract—Contrastive learning, which aims to capture general representation from unlabeled images to initialize the medical analysis models, has been proven effective in alleviating the high demand for expensive annotations. Current methods mainly focus on instance-wise comparisons to learn the global discriminative features, however, precluding the local details to distinguish tiny anatomical structures, lesions, and tissues. To address this challenge, in this paper, we propose a general unsupervised representation learning framework, named local discrimination (LD), to learn local discriminative features for medical images by closely embedding semantically similar pixels and identifying regions of similar structures across different images. Specifically, this model is equipped with an embedding module for pixel-wise embedding and a clustering module for generating segmentation. And these two modules are unified by optimizing our novel region discrimination loss function in a mutually beneficial mechanism, which enables our model to reflect structure information as well as measure pixel-wise and region-wise similarity. Furthermore, based on LD, we propose a center-sensitive one-shot landmark localization algorithm and a shape-guided cross-modality segmentation model to foster the generalizability of our model. When transferred to downstream tasks, the learned representation by our method shows a better generalization, outperforming representation from 18 state-of-the-art (SOTA) methods and winning 9 out of all 12 downstream tasks. Especially for the challenging lesion segmentation tasks, the proposed method achieves significantly better performance.

Index Terms—Contrastive learning, local discrimination, one-shot landmark localization, shape-guided segmentation.

I. INTRODUCTION

A RTIFICIAL intelligence-based computer-aided diagnosis has recently become the mainstream of various fields of precision medicine [1], such as radiotherapy and disease screening [2]. Intelligent medical image analysis plays a crucial role in the more efficient and effective diagnosis and is the basis for computer-aided systems [3]. However, the heavy reliance on vast amounts of annotations for establishing a generalized medical image analysis model hinders its broad applications. To address this issue and alleviate the high annotation demand, contrastive learning has been a hot research topic and achieved huge breakthroughs [4], [5].

In contrastive learning methods, the contrastive rule for constructing negative and positive pairs is of vital importance. Recently, ingenious rules have been proposed for various medical modalities images, including color fundus images [4], [5], chest X-ray images [6], [7], and MRI [8]. For example, based on the similarity between fundus fluorescein angiography (FFA) images and color fundus images, Li et al. [4] treated the color fundus image and the FFA image of the same patient as a positive pair and otherwise as negative pairs, by which modality-invariant and patient-similarity features were captured to enhance retinal disease diagnosis. Based on the hypothesis that aligned MRI images share similar contents in the same anatomical region, Krishna et al. [8] managed volumes of similar positions as positive pairs to learn features to improve segmentation tasks. These methods have thoroughly investigated the characteristics of medical images and the demands of tasks, and the proposed models are proven effective for solving the target tasks.

Most of the recent work focuses on constructing positive/negative pairs from a global perspective to learn discriminative features [5], [6], [9]; however, the local discrimination, which is essential for depicting lesions, minor tissues, and tiny structures, has not been sufficiently investigated and exploited.

Aiming to learn local features from aligned corresponding positions in the registered images, [8] relies on image registration and is ineffective for images with severe anatomical deformation. Furthermore, most current techniques are specifically designed for specific imaging modalities or tasks. Correspondingly, particular training data is needed for the modality-specific or task-driven models, which in turn hinders them from broad clinical applications. For example, corresponding color fundus images

Manuscript received 21 July 2022; revised 17 May 2023; accepted 21 July 2023. Date of publication 26 July 2023; date of current version 3 November 2023.

This work was supported in part by the NSFC of China under Grant 12090024, in part by the SJTU Trans-med Awards Research under Grant 20220101, and in part by the Shanghai Municipal Science and Technology Major Project under Grant 2021SHZDZX1012. Recommended for acceptance by S. K. Zhou.

(Corresponding authors: Qing Peng; Lisheng Wang.)

Huai Chen, Jieyu Li, and Lisheng Wang are with the Institute of Image Processing and Pattern Recognition, Department of Automation, Shanghai Jiao Tong University, Shanghai 200240, China (e-mail: chenhua@sjtu.edu.cn; lijieyu@sjtu.edu.cn; lwang@sjtu.edu.cn).

Renzhen Wang and Deyu Meng are with the School of Mathematics and Statistics and Ministry of Education Key Lab of Intelligent Networks and Network Security, Xi’an Jiaotong University, Xi’an, Shaanxi 710049, China (e-mail: wrzhen@stu.sjtu.edu.cn; dymeng@mail.xjtu.edu.cn).

Xiuying Wang is with the School of Computer Science, The University of Sydney, Sydney, NSW 2006, Australia (e-mail: xi.wang@sydney.edu.au).

Qu Fang is with Changchun GeneScience Pharmaceutical Company Ltd., Shanghai 202150, China (e-mail: patrickfang@qq.com).

Hui Li is with Cooperative Medianet Innovation Center, Shanghai Jiao Tong University, Shanghai 200240, China (e-mail: evelynli@sjtu.edu.cn).

Jianhao Bai and Qing Peng are with the Department of Ophthalmology, Shanghai Tenth People’s Hospital, Tongji University, Shanghai 200240, China (e-mail: dbai2020@tongji.edu.cn; pengqing@tongji.edu.cn).

The source codes are publicly available at https://github.com/HuaiChen-1994/LDlearning.

Digital Object Identifier 10.1109/TPAMI.2023.3299038
and FFA images are needed in [4], and image-text pairs are needed in [7].

To address the challenge of learning local discriminative representation for medical image analysis, alleviate the high demand for expensive annotations, and ensure that the representation learning framework is generalized for most medical modalities and tasks, we design a universal contrastive rule and propose a general local discrimination learning framework. Concretely, as shown in Fig. 1(a), almost all medical images possess the property of intra-modality structure similarity due to the consistency of anatomical structures and the stable focusing region of the imaging modality, which means that images of the same medical modality share structure regions with similar patterns [10]. For instance, each color fundus image has the optic disk (OD), vessels, and fovea. Moreover, the pixels within each structure region are semantically consistent. Therefore, for medical images of the same modality, clustering semantically similar pixels in an image to form structure regions and recognizing regions with similar patterns in different images can be a universal contrastive rule. Based on this rule, we propose region discrimination, which attempts to closely embed semantically consistent pixels and distinguish regions of different semantic categories. By optimizing region discrimination model, local discriminative features that measure both pixel-wise and region-wise similarity can be captured.

Furthermore, based on the ability of region discrimination to cluster semantically similar pixels, we make a new perspective to combine region discrimination with the cross-modality shape prior to realize segmentation. As shown in Fig. 1(b), although there are obvious gaps in the content and visual characteristics of color fundus images, optical coherence tomography angiography (OCTA) images, and digital subtraction angiography (DSA) images, their vessels have similar shapes. This property is named inter-modality shape similarity and is widely shared among anatomical structures, especially vessels [10]. Based on this property, we can re-utilize the existing manual segmentation masks of similar structures as the reference for the clustering of region discrimination, ensuring that the segmentation region maintains the interior semantic consistency and possess the expected shape, and thereby obtaining the segmentation of the target structure.

Finally, to take full advantage of the learned representation that is capable of measuring regional similarity, we propose a one-shot center-sensitive landmark localization model. It introduces the center-sensitive property into the discriminative representation, by which landmarks can be accurately localized with the guidance of one labeled sample. Since the learned features are augmentation-invariant and orientation-invariant, this model is generalized and robust for multiple medical tasks.

In summary, our major contributions are three folds:

- A universal representation learning framework is proposed to generate representation with the ability to measure pixel-wise and region-wise similarity. The learned representation significantly enhances the medical image models focusing on analyzing tiny structures/tissues and alleviates their demand for expensive annotations. More specifically, our model improves the mean Dice Similarity Coefficient (DSC) by 5.79% for 8 color fundus tasks and 3.68% for 4 chest X-ray tasks, respectively.

- A new schema to transfer shape knowledge across medical modalities is presented, which can be fused with our representation learning model to realize shape-guided segmentation. This model demonstrates the feasibility of identifying anatomical structures by harnessing the shape knowledge from similar structures and the interior semantic consistency, and opens up a new mechanism to effectively re-utilize the existing annotations.

- Based on the learned representation with the ability to measure regional similarity, a center-sensitive one-shot landmark localization method is proposed, which makes landmark localization more economical and convenient. It is one of the pioneering self-supervised one-shot landmark localization frameworks and is generalized for various medical tasks.

This paper is an extended version of our conference paper [11]. Besides improving the solid theoretical basis and providing sufficient experiments and discussions, the substantial extensions are as follows: 1) Most of all, this study has innovated the optimization model to enhance our model’s generalization, robustness and practicability for real-world medical datasets, including proposing region discrimination loss function based on the comprehensive structure information to enhance the model’s robustness against noisy data, and presenting an evolutionary optimization strategy to enhance the stability of simultaneously training the clustering and the embedding. 2) To further demonstrate our model’s generalization for broad clinical applications, this work has presented the enhanced shape-guided segmentation and the center-sensitive one-shot landmark localization, both of which are of beneficial clinical and practical value to alleviate the clinician’s burden and accelerate model construction.

The following paper is organized as follows: Section II overviews the related work. The preliminary knowledge is elucidated in Section III. Section IV presents the details of our
representation learning model. Section V deeply investigates the clinical application of our method for segmentation and localization. In Section VI, extensive experiments are conducted to analyze our method quantitatively and qualitatively. Finally, conclusions are summarized in Section VII.

II. RELATED WORK

A. Contrastive Learning

The core idea of contrastive learning is to construct positive/negative pairs based on some similarity measuring rules, and learn a semantic embedding space to map positive pairs closely and negative pairs separately [12], [13], [14]. By minimizing contrastive loss, such as InfoNCE [15], the models are encouraged to learn discriminative features that can be transferred to enhance downstream tasks. To construct positive and negative pairs, the early work [16] was based on datasets with prior knowledge, or manual annotations, and treated images with neighborhood relationships as positive pairs and otherwise as negative pairs. More recently, contrastive learning methods based on unlabeled images have been the mainstay, and can be divided into instance-wise contrastive learning [12], [13], [14] and cluster-wise contrastive learning [17], [18], [19].

Instance-wise contrastive learning can be regarded as the extension of exemplar-CNN [20], in which, the images of the same instance but with varied augmentations are treated as the same category and the discriminative features are learned through the classification procedure. Although this framework is capable of capturing semantic features, it is impractical when training with large-scale datasets due to the high computational demand of adopting a parametric paradigm. To address this issue, more recent instance discrimination methods [13], [14], [21], [22] adopt a non-parametric classifier and a noise-contrastive estimation to recognize instances with a fixed computational cost. MoCo [12] and MoCov2 [23] are two typical instance-wise discriminative methods, in which, an on-the-fly dictionary composed of an embedding encoder and a momentum encoder is built to match queries and keys. They have achieved milestones in closing the performance gap between unsupervised and supervised representation learning.

In our framework, the preparatory stage of patch discrimination is mostly related to instance-wise contrastive learning. However, to satisfy the demand of medical image processing for describing local details, patch discrimination maps an image to dense spatial feature vectors instead of the global feature vector. Each of these feature vectors represents a patch and reflects its comprehensive semantic information by integrating the patch’s inner content and nearby context. In patch discrimination, a patch instance is treated as an individual category, and the feature measuring patch-wise similarity can be obtained by forcing the model to recognize different views of the same patch instance. Furthermore, hypersphere mixup is presented for patch discrimination to improve the smoothness of embedding space and enhance the generalization of the representation.

Cluster-wise contrastive learning [17], [18], [19] attempts to harness clustering and contrastive learning to investigate the inter-image similarity and treat two similar images as a positive pair. For example, to enhance the ability on depicting the semantic structure, Li et al. [18] assigned image instances to several prototypes of different granularity, and then the contrastive loss was constructed to enforce an image instance and its assigned prototype to have similar embeddings.

Our region discrimination also aims to introduce clustering to enhance contrastive learning. However, rather than clustering images into several cluster assignments in a global perspective, our method focuses on clustering semantically consistent regions based on the intra-modality structure similarity, which contributes to much finer features with the ability to measure both pixel-wise and region-wise similarities. Furthermore, due to the mutually beneficial training mechanism, the clustered results and the embeddings are directly generated by our model without extra clustering stages as K-means in [18] and FINCH in [24], which is efficient and economic.

B. Contrastive Learning in Medical Images

Contrastive learning has been a hot topic in medical image analysis [6], [7], [8]. According to the design mechanisms, recent work can be mainly classified into specific-properties-based methods [4], [25], [26] and application-driven methods [5], [6]. Specific-properties-based methods aim to fuse the modality-specific characteristics into contrastive learning. For instance, considering the knowledge that different dyes can enhance different types of tissue components, Yang et al. [26] separated the stain results of hematoxylin and eosin, and fused them into the design of contrastive learning.

Application-driven methods take account of the properties of the processing tasks when designing models. For instance, based on the observation that the structure information in retinal images is sensitive to the orientations and is beneficial for retinal disease recognition, Li et al. [5] fused the rotation prediction task with contrastive learning to improve the classification of pathologic myopia.

Most of the recent medical contrastive learning methods focus on using instance discrimination to learn global features, while neglecting local details [5], [6], [7], [9]. And further, the special requirements for the training data hinder the current contrastive learning from broad medical applications [4], [6], [7]. Comparatively, our method learns local discriminative representation with the ability to measure pixel-wise and region-wise similarity. And by incorporating the common property of intra-modality structure similarity, our model is capable of processing and analyzing images for multiple medical modalities and downstream tasks.

C. Cross-Domain Knowledge Transfer

The purpose of cross-domain knowledge transfer is to improve the performance of tasks in the target domain by utilizing the knowledge from related domains. According to differences in the working mechanisms, it can be divided into domain adaptation and synthetic segmentation. Domain adaptation aims to adjust the trained model in the source domain to be generalized in the target domain [27], [28], which effectively alleviates the degeneration of the model caused by the changing imaging conditions.

Authorized licensed use limited to the terms of the applicable license agreement with IEEE. Restrictions apply.
**Synthetic segmentation** is a unique cross-domain knowledge transfer framework for medical images to share knowledge across modalities [29], [30], [31]. It is based on the hypothesis that images of medical modalities focusing on the same body part share similar organs and tissues. Based on this, synthetic segmentation usually utilizes image translation to get synthetic target modality images from source labeled images. After that, the synthetic images are accompanied by original real labels to train a segmentation network and this model can be directly applied to the target modality. Synthetic segmentation has been widely investigated in cross-modality pairs, such as CT-MR [29], and CBCT-MR [31]. Recently, the unsupervised framework of CycleGAN-based synthetic segmentation has obtained more attention than supervised ones [29], [30], [32].

Synthetic segmentation is a cost-effective method to transfer knowledge across modalities. It demonstrates that the deep network owns a similar ability as humans to memory and transfer knowledge from familiar medical scenarios to unknown ones [33]. However, it can only share knowledge among modalities with almost the same content, i.e., modalities focusing on the same body part. For radiologists, the ability to transfer knowledge is more powerful, by which, the knowledge can still be shared across modalities with big content differences. For example, the knowledge of vessel shape in OCTA or DSA is enough for a radiologist to recognize vessels in color fundus images. The inherent mechanism is that people can recognize a target based on shape knowledge and its interior pattern similarity. Therefore, we attempt to demonstrate the feasibility of transferring shape knowledge among modalities with big content gaps to realize segmentation, and successfully propose shape-guided cross-modality segmentation.

### D. Self-Supervised One-Shot Landmark Localization

Anatomical landmark localization is essential for medical therapy planning and intervention, and has been a hot research topic [34], [35]. The supervised localization models often heavily rely on expensive annotations on the images. To alleviate the burden of manual annotating, self-supervised one-shot landmark localization frameworks [36], [37] have been designed, which train models in easily accessible unlabeled images and only require one labeled sample as supporting information. For example, RPR-Loc [36] was based on the idea that tissues and organs from different human bodies own similar relative positions and contexts. And thus, this method predicted the relative positions between two patches to locate the target organ. Cascade Comparing to Detect (CC2D) [37] was a coarse-to-fine localization framework, where coarse-grained corresponding areas were first identified and then the finer-grained areas were matched in the selected coarse area. For CC2D, the target point is randomly selected in the original image, and the training target is to detect the best matching point in the cropped image.

Our center-sensitive one-shot landmark localization method is based on our patch discrimination, and utilizes the learned patch discriminative features to search for the best matching patch. To ensure that the detected results are precisely aligned with target landmarks, we further propose the center-sensitive averaging pooling, making the representation of a patch pay more attention to the semantic information in the patch’s central regions. Compared with recent self-supervised methods, where CC2D requires that most pixels in the training images have structural information and RPR-Loc needs the structures in training images should have stable relative positions, our method is more generalized for fitting different medical modalities since the trained patch-wise embedding is augmentation-invariant and orientation-invariant.

### III. PRELIMINARIES

#### A. Base Contrastive Formulation

Traditional contrastive learning commonly formulates an embedding function that projects images into a K-d embedding space to make positive pairs closely clustered and negative pairs dispersedly distributed. We denote \( v_i \in \mathbb{R}^K \) as the embedded vector of the \( i \)-th sample, \( v^+ \in V^+_i \) and \( v^- \in V^-_i \) as its positive and negative samples, where \( V^+_i \) and \( V^-_i \) are respectively the sets of positive and negative embedded vectors for \( v_i \). The contrastive loss is defined as:

\[
\mathcal{L}_i = \frac{1}{|V^-_i|} \sum_{v^+ \in V^+_i} - \log \frac{e^{\text{sim}(v^+, v_i) / \tau}}{\sum_{v^- \in V^-_i} e^{\text{sim}(v^-, v_i) / \tau}},
\]

where \( \text{sim}(a, b) \) means the similarity between two vectors and is usually set as the cosine similarity, i.e., \( \text{sim}(a, b) = a^\top b / \|a\|_2 \|b\|_2 \). And \( \tau \) is a temperature parameter to control the concentration level of the distribution [38].

#### B. From Instance-Wise Discriminative Features to Local Discriminative Features

**Instance-wise discriminative features:** The common rule of constructing positive and negative pairs is to create multiple views of images with varied augmentations, and a positive/negative pair refers to two views from the same/different images [12], [13], [21]. Following this rule, instance-wise comparisons are conducted to learn the global representation that can distinguish different image instances. The flowchart is shown in Fig. 2(a), where image sample \( x_n \) is fed into an encoder \( f(\cdot) \) with a global average pooling layer, by which the discriminative information for the whole image can be captured, i.e., \( v_n = f(x_n) \).
Local discriminative features: Instance-wise discriminative features are highly concentrated for semantic context; however, medical image tasks, especially segmentation of lesions, tissues, and tiny structures, need finer features describing local details. To address this issue, we attempt to learn local discriminative features, which evolve from patch-wise discriminative features to region/pixel-wise discriminative features. To learn patch-wise discriminative features, dense spatial features are generated to represent patch instances, and each patch is treated as an individual class (shown in Fig. 2(b)). As shown in Fig. 2(c), region-wise discriminative features are captured by identifying the structure region composed of semantically consistent pixels in an image and clustering regions of the same semantic class among different images. Compared with patch-wise discriminative features, region-wise discriminative features are capable of reflecting comprehensive structure information as well as measuring pixel-wise and region-wise similarity, which better satisfies the demand of medical image analysis for distinguishing tiny structures/tissues.

The pipeline of the overall framework: As shown in Fig. 3, our proposed framework is composed of three major parts, including 1) a backbone to generate embeddings and segmentation masks, 2) a systematic local discrimination model, containing patch discrimination and region discrimination, to learn local discriminative features, and 3) the clinical applications of one-shot landmark localization and the shape-guided segmentation.

IV. LOCAL DISCRIMINATION LEARNING

The illustration of our local discriminative representation learning framework is shown in Fig. 4. It is designed as an evolutionary framework, which first utilizes patch discrimination to warm up the embedding module with the initial patch-wise discriminative ability and then further optimizes the total model by region discrimination to learn region-wise and pixel-wise discriminative features. In Section IV-A, we present the definition and notation of the backbone. Then, patch discrimination and region discrimination are respectively articulated in Sections IV-B and IV-C. Finally, hypersphere mixup to enhance the generalization is elucidated in Section IV-D.

A. The Definition and Notation of the Backbone

The backbone aims to generate pixel-wise embedding and segmentation masks, and therefore should support point-to-point mapping. Its design is based on U-Net [39] which is one of the most commonly utilized segmentation models with the ability to fuse low-level and high-level features. Specifically, as shown in Fig. 4(a), the backbone is composed of an encoder \( f(\cdot) \) and a decoder \( g(\cdot) \). And the final block is composed of a clustering module and an embedding module to realize segmenting and embedding.

We denote the unlabeled image set as \( X = \{x_1, x_2, \ldots, x_N\} \) and the image instance as \( x_n \in \mathbb{R}^{H \times W} \), where \( H \) and \( W \) are respectively the height and width of this image. After feeding \( x_n \) into this model, we can get the segmentation mask \( r_n \in \mathbb{R}^{H \times W \times C} \) and the pixel-wise embedding \( v_n \in \mathbb{R}^{H \times W \times K} \), i.e.:

\[
v_n, r_n = g(f(x_n)).
\]

\( v_n(h, w) \in \mathbb{R}^K \) is the embedded vector for pixel \( x_n(h, w) \) and \( r_{nc}(h, w)(0 \leq c < C) \) is the probability of classifying this pixel into \( c \)-th semantic class. \( v_n(h, w) \) is further processed by \( l_2 \)-normalization to project the embedded vector into a unit space, i.e., \( ||v_n(h, w)||_2 = 1 \). Meanwhile, \( r_n(h, w) \) is processed by \( l_1 \)-normalization to make the sum of classification probabilities of a single pixel to be 1, i.e., \( ||r_n(h, w)||_1 = 1 \).

B. Patch Discrimination

Patch discrimination is the first step to empowering our model with the initial discriminative ability. Its core idea is to treat each patch instance as an individual class and force the model to cluster different views of the same patch together. Such a contrastive learning scheme enables the model to measure the semantic similarity between patch instances and the resulting model can capture patch-wise discriminative representation.

As shown in Fig. 4(b), both image sample \( x_n \) and its augmented sample \( \tilde{x}_n \) are fed into the model to get the pixel-wise embedding \( v_n \) and \( \tilde{v}_n \). Then, the embedded vector for a patch can be denoted as the sum of pixel-wise embedded vectors within the patch. To simplify the process of constructing pairs and improve the practicability, we evenly divide an image sample into \( H_p \times W_p \) patches. After that, the embedded feature of the \( i \)-th patch \( p_{ni} \) (\( 0 \leq i < H_p W_p \)) is denoted as \( s_{ni} \), and can be formulated as:

\[
s_{ni} = \frac{\sum_{h \in [h_b, h_e), w \in [w_b, w_e)} v_n(h, w)}{\sum_{h \in [h_b, h_e), w \in [w_b, w_e)} v_n(h, w)}
\]

where \((h_b, w_b)\) and \((h_e, w_e)\) are respectively the top-left point and the bottom-right point of this patch, and their detailed
definitions are as follows:

\[ h_b = \frac{H}{H_p} \times \left[ i/W_p \right], \quad h_e = \frac{H}{H_p} + h_b, \]
\[ w_b = \frac{W}{W_p} \times \left( i - W_p \left[ i/W_p \right] \right), \quad w_e = \frac{W}{W_p} + w_b, \quad (4) \]

where \([ \cdot ]\) means rounding numbers down to the nearest integer.

Following the similar analysis as above, we can capture the embedded vector for the corresponding augmented patch in \( \hat{s}_{ni} \) and denote it as \( \hat{s}_{ni} \). Therefore, \( S_{ni}^- = \{ s_{ni} \} \) is the positive embedding set for \( \hat{s}_{ni} \), and the negative embedding set can be denoted as \( S_{ni}^+ = \{ s_{ni'} | n' \neq n \text{ or } i' \neq i \} \). Then, the probability of recognizing \( s_{ni} \) as \( s_{ni} \) can be defined as:

\[
P(n_i|\hat{s}_{ni}) = \frac{e^{\alpha(s_{ni}, \hat{s}_{ni})/\tau}}{\sum_{s \in S_{ni}^- \cup S_{ni}^+} e^{\alpha(s, \hat{s}_{ni})/\tau}}, \quad (5)\]

On the other hand, the probability of incorrectly recognizing \( s_{ni} \) as \( s_{ni'} \) (\( n' \neq n \text{ or } i' \neq i \)) can be formulated as:

\[
P(n_i'|\hat{s}_{ni}) = \frac{e^{\alpha(s_{ni'}, \hat{s}_{ni})/\tau}}{\sum_{s \in S_{ni}^- \cup S_{ni}^+} e^{\alpha(s, \hat{s}_{ni})/\tau}}. \quad (6)\]

Assuming each patch being recognized as \( p_{ni} \) is independent, then the joint probability of making the right classifications can be formulated as:

\[
P_{ni} = P(n_i|\hat{s}_{ni}) \prod_{n' \neq n \text{ or } i' \neq i} (1 - P(n_i'|\hat{s}_{ni})), \quad (7)\]

Patch discrimination attempts to optimize the model to make correct patch classification, i.e., to optimize \( f(\cdot) \) and \( g(\cdot) \) to maximize \( P_{ni} \). We can apply minimizing the corresponding negative log-likelihood as the alternative training target, and the loss function for patch discrimination (\( L_{Pd} \)) can be formulated as:

\[
L_{Pd} = \sum_{n,i} M_{ni}, \quad (8)\]
\[
M_{ni} = -\log(P_{ni}), \quad (9)\]

By optimizing the model by minimizing \( L_{Pd} \), patch instances can be defensively projected over the learned hypersphere space and the cosine distance between embedded features can be used to measure the similarity between the correlated patches.

C. Region Discrimination

While patch discrimination is capable of extracting finer patch-level discriminative features, the mechanism of dividing an image instance into patches may impede the integrity of anatomical structures, and treating each patch as an individual class neglects the similarity among patches. Such a representation learning scheme has a limited ability to capture the comprehensive structure information, and its training is easily disturbed by the patches with similar contents across different images since they are forcibly classified into different classes.

To address the aforementioned issues and satisfy the demand of medical image analysis for representation that is capable of reflecting comprehensive structure information and measuring pixel/region-wise similarity, we further propose region discrimination, which creatively incorporates the intra-modality structure similarity into the model design. More specifically, besides the embedding module for pixel-wise embedding, region discrimination is further equipped with the clustering module (shown in Section IV-A) for generating segmentation to discriminate structure regions in an image and assimilate semantically consistent regions among different images. These two modules are jointly optimized by the novel region discrimination loss function, which enables the model to closely embed semantically similar pixels and distinguish regions of different semantic classes.

As shown in Fig. 4(c) and (2), the segmentation mask \( r_n \) generated by the clustering module indicates the probability of classifying pixels into \( C \) classes. A high value of \( r_{nc}(h, w) \) means the pixel \( x_n(h, w) \) is categorized into \( c \)-th semantic class with high confidence. Therefore, when representing the \( c \)-th semantic region in \( x_n \), the semantic information of the pixel with a high \( r_{nc} \) is reasonable to play a more important role. Fusing this idea into the model design, we denote the embedded vector for \( c \)-th structure region in \( x_n \) as \( t_{nc} \) and it can be deduced by combining the pixel-wise embedding \( v_n \) and the segmentation mask \( r_{nc} \) as follows:

\[
t_{nc} = \frac{\sum_{h,w} r_{nc}(h, w)v_n(h, w)}{\left\| \sum_{h,w} r_{nc}(h, w)v_n(h, w) \right\|_2}. \quad (10)\]

Since images of the same modality commonly share similar structure regions, we can obtain the prototype (cluster centroid) vector \( T_c \) for the \( c \)-th semantic class as follows:

\[
T_c = \frac{\sum_{n} t_{nc}}{\left\| \sum_{n} t_{nc} \right\|_2}. \quad (11)\]

Then, the optimization objective of region discrimination is to cluster regions of the same semantic class together, i.e., to assign \( t_{nc} \) to the prototype of \( T_c \).

The probability of assigning \( t_{nc} \) to \( T_c \) can be defined as:

\[
P_{nc} = P(c|t_{nc}) = \frac{e^{\alpha(T_c, t_{nc})/\tau}}{\sum_{0 \leq c < C} e^{\alpha(T_c', t_{nc})/\tau}}. \quad (12)\]

Since the negative log-likelihood of the joint probability \( \prod_{n,c} P_{nc} \) can be set as the loss function for region discrimination, we denote the loss as \( L_{Rd} \) and the detailed definition can be formulated as follows:

\[
L_{Rd} = -\sum_{n,c} \log(P_{nc}), \quad (13)\]

Formulation analysis: Minimizing \( L_{Rd} \) ((13)) equals to maximizing \( P_{nc} \) (12). Eq. (12) can be re-written as:

\[
P_{nc} = P(c|t_{nc}) = \frac{1}{1 + \sum_{c' \neq c} e^{\alpha(T_{c'}, t_{nc})/\tau}}. \quad (14)\]

Then, the training objective can be replaced by maximizing \( \text{sim}(T_c, t_{nc}) \) and minimizing \( \text{sim}(T_{c'}, t_{nc}) \), i.e., pulling \( t_{nc} \) toward \( T_c \) and pushing \( t_{nc} \) away from \( T_c \). To realize this target, according to the definition of \( t_{nc} \) (10), the pixel with the embedding \( v_n(h, w) \) similar to \( T_c \) should have a high probability value \( r_{nc}(h, w) \), and vice versa. This mechanism enables the clustering module and the embedding module to be optimized in
a mutually beneficial mechanism, where the clustering module generates \( r_{nc} \) to identify which pixels should be closely embedded for the embedding module, and the embedding module generates \( v_{n} \) to decide which pixels should be clustered together for the clustering module.

The evolutionary optimization strategy: As analyzed above, the optimization of the embedding module and the clustering module in region discrimination rely on each other. As a consequence, if directly optimizing these two modules from scratch, they will provide random information to disturb each other, which may lead to meaningless clustering or a training collapse, i.e., the clustering module maps all pixels into a single category. To address this issue, our representation learning framework is designed as a gradual evolution model, by which the embedding is first trained by patch discrimination to with the initial patchwise discriminative ability, and then the whole model is further optimized by region discrimination to possess the region-wise discriminative ability. This mechanism significantly enhances the training stability and, in particular, improves the mutually beneficial optimization of the clustering and the embedding.

D. Hypersphere Mixup

Patch discrimination learning described in Section IV-B mainly applies consistency regularization to the hypersphere, specifically, it gathers multiple views of the same patch together in the embedding space. This constraint encourages the proposed model to measure similarity in semantic context and makes patch instances sparsely distributed over the hyperspace. However, since the number of training patch instances is limited, a large part of the embedding space is blank, that is, no patches are projected into it, which limits the generalization of the learned representation [40], [41]. To tackle this issue and constrain the embedding space with smoothness, we propose hypersphere mixup to fill the void between embedded vectors of patch instances. Its main idea is to construct virtual training patches as the linear interpolations of real patch instances and these virtual patches should be mapped to the linear interpolations of the associated embedded vectors. By conducting hypersphere mixup, the hyperspace is smoother and the learned presentation is more generalized and robust.

In hypersphere mixup, we first construct a virtual training sample \( \tilde{x}_{n_1n_2} \) as the linear interpolation of two real image samples \( x_{n_1} \) and \( x_{n_2} \), i.e.:

\[
\tilde{x}_{n_1n_2} = \lambda x_{n_1} + (1 - \lambda)x_{n_2}, \quad \lambda \in (0, 1).
\]  

After that, \( \tilde{x}_{n_1n_2} \) is fed into the model to get pixel-wise embeddings \( \tilde{v}_{n_1n_2} \), i.e., \( \tilde{v}_{n_1n_2} = g(\tilde{x}_{n_1n_2}) \). Then, similar to (3), the embedded vector for \( i \)-th virtual patch \( \tilde{p}_{n_1n_2i} \) is denoted as \( \tilde{s}_{n_1n_2i} \) and can be formulated as:

\[
\tilde{s}_{n_1n_2i} = \frac{\sum_{h \in [i_h, h_u], w \in [w_l, w_u]} \tilde{v}_{n_1n_2i}(h, w)}{\left| \sum_{h \in [i_h, h_u], w \in [w_l, w_u]} \tilde{v}_{n_1n_2i}(h, w) \right|_2}.
\]  

On the other hand, we construct a virtual embedded vector \( \tilde{s}_{1n_2} \), which is the linear interpolation of \( s_{n_1i} \) and \( s_{n_2i} \) with the same \( \lambda \) as (15), where \( s_{n_1i} \) and \( s_{n_2i} \) are respectively the embeddings for \( i \)-th patch in \( x_{n_1} \) and \( i \)-th patch in \( x_{n_2} \). The formulation is defined as follows:

\[
\tilde{s}_{n_1n_2i} = \frac{\lambda s_{n_1i} + (1 - \lambda)s_{n_2i}}{\left| \lambda s_{n_1i} + (1 - \lambda)s_{n_2i} \right|_2}.
\]  

The training target for hypersphere mixup is to optimize \( f(\cdot) \) and \( g(\cdot) \) to make \( \tilde{s}_{n_1n_2i} \) similar to \( \tilde{s}_{n_1n_2i} \). To realize this, we can also maximize the probability of recognizing \( \tilde{s}_{n_1n_2i} \) as \( \tilde{s}_{n_1n_2i} \). Similar to the analysis of \( L_{pd} \) (Section IV-B), we set \( S_{n_1n_2} = \{ \tilde{s}_{n_1n_2i} \} \) and \( \tilde{S}_{n_1n_2} = \{ \tilde{s}_{1n_2i} \mid i \neq n_1 \} \) as the positive embedding set and the negative embedding set for \( \tilde{s}_{n_1n_2} \). Accordingly, the probability of making the right classification and the probability of making the incorrect classification are respectively denoted as \( P(n_1n_2 | \tilde{s}_{n_1n_2}) \) and \( P(n'_1n'_2 | \tilde{s}_{n_1n_2}) \), and their detailed formulations are as follows:

\[
P(n_1n_2 | \tilde{s}_{n_1n_2}) = \frac{e^{s_{sim}(\tilde{s}_{n_1n_2}, \tilde{s}_{n_1n_2})/\tau}}{\sum_{n'_1} e^{s_{sim}(\tilde{s}_{n_1n_2}, \tilde{s}_{n_1n_2})/\tau}},
\]

\[
P(n'_1n'_2 | \tilde{s}_{n_1n_2}) = \frac{e^{s_{sim}(\tilde{s}_{n_1n_2}, \tilde{s}_{n_1n_2})/\tau}}{\sum_{n_1, 2} e^{s_{sim}(\tilde{s}_{n_1n_2}, \tilde{s}_{n_1n_2})/\tau}}.
\]

Then, the joint probability of making the correct classification can be defined as follows:

\[
P_{n_1n_2} = P(n_1n_2 | \tilde{s}_{n_1n_2}) \prod_{n'_1n'_2 \neq n_1n_2} (1 - P(n'_1n'_2 | \tilde{s}_{n_1n_2})).
\]  

Finally, we can deduce the loss function for hypersphere mixup (\( L_{Hm} \)) as follows:

\[
L_{Hm} = - \sum_{n_1, n_2, i} \log(P_{n_1n_2i}).
\]

V. SHAPE-GUIDED SEGMENTATION AND ONE-SHOT LOCALIZATION

The previous section proposes a local discriminative representation learning framework composed of patch discrimination and region discrimination. Patch discrimination enables the trained model to measure similarity between patches, and region discrimination empowers the model to cluster semantically similar pixels and distinguish regions of different semantic classes. Based on these abilities, we attempt to investigate its potential clinical applications. Specifically, in Section V-A, the shape prior of similar structures is first introduced to be fused with region discrimination to realize segmentation of the target structure with the expected shape distribution. Then, in Section V-B, we introduce the center-sensitive property into patch discrimination and the improved model is suitable for detecting landmarks with only one labeled sample.

A. Shape-Guided Cross-Modality Segmentation

As inter-modality shape similarity indicated, some structures or tissues in different medical modalities, such as vessels in color fundus images, OCTA images, and DSA images, have similar shapes. If one of these structures has manual segmentation
trained by the constraint of region discrimination and

\[ r_{nm}^*(i, j) = \frac{e^{\text{sim}(t_{nm}, v_{n}(h, w))/\tau}}{\sum_{0 \leq e < C} e^{\text{sim}(t_{nc}, v_{n}(h, w))/\tau}}. \]  

(21)

The utilized prototype vector is specifically captured for the test sample, which better fits with the test instance and can improve the accuracy of segmentation.

Uncertainty estimation refinement attempts to address the model’s fragility caused by the absence of definite annotations. Its working mechanism is to utilize uncertainty estimation [43, 44] to identify reliable predictions as pseudo labels, and then re-train a segmentation model based on these labels in a supervised mechanism.

As shown in Fig. 5(b), pseudo labels and uncertainty maps are first generated by the label generator, which is composed of \( f(\cdot) \) and \( g(\cdot) \) trained by the constraint of region discrimination and shape prior. Specifically, we first utilize varied augmentations to process \( x_n \) to generate multiple image samples, denote \( x_n' \) \((0 \leq e < E)\) as the \( e \)-th augmented sample. Then \( x_n' \) is fed into \( f(\cdot) \) and \( g(\cdot) \) to generate \( r_{nm}^e \), according to (21). Based on the above processings, the pseudo label \( r_{nm}^* \) and the uncertainty map \( u_{nm} \) can be calculated as follows:

\[
\begin{align*}
    r_{nm}^*(i, j) &= \begin{cases} 
    1, & \text{if } \frac{1}{E} \sum_{e} r_{nm}^e(i, j) > 0.5 \\
    0, & \text{otherwise}
    \end{cases} \tag{22}
\end{align*}
\]

\[
\begin{align*}
    u_{nm}(i, j) &= -\frac{1}{E} \sum_{e} r_{nm}(i, j) \log(r_{nm}^e(i, j)) - \frac{1}{E} \sum_{e} (1 - r_{nm}^e(i, j)) \log(1 - r_{nm}^e(i, j)). \tag{23}
\end{align*}
\]

Finally, \((x_{nm}, r_{nm}^*, u_{nm})\) are utilized to re-train a segmentation network. We define the final output as \( r' \) and the loss function set as a binary cross-entropy loss weighted by the uncertainty map, the detailed definition is as follows:

\[
\begin{align*}
    \mathcal{L}_{wce} &= -\frac{1}{N} \sum_{n, i, j} u_{nm}(i, j) r_{nm}^*(i, j) \log r_{nm}(i, j) - \frac{1}{N} \sum_{n, i, j} u_{nm}(i, j) (1 - r_{nm}^*(i, j)) \log(1 - r_{nm}(i, j)). \tag{24}
\end{align*}
\]

B. Center-Sensitive One-Shot Localization

The ability to measure similarity is the mainstay of metric-learning-based one-shot methods [45], and the trained model by patch discrimination is capable of measuring patch-wise similarity. Therefore, we attempt to take full advantage of patch discrimination to realize unsupervised one-shot landmark localization to alleviate the burden of manual annotating.

One-shot landmark localization aims to localize landmarks in unlabeled images (query samples) based on one labeled sample
(support sample). A natural idea based on patch discrimination is that we can first construct a patch with the labeled point as the center and capture its embedded vector as the support embedded vector. Then, we can measure the cosine similarity between the support embedded vector and the vectors of patches in the query sample. After that, the detected point is identified as the center point of the best matching patch. However, when deducing the embedded vector for a patch instance, original patch discrimination gives equal importance to pixel-wise embeddings within it (shown in 3), which makes the identification of the target point easily disturbed by its nearby pixels.

To address this challenge, rather than utilizing averaging pooling to get patch embedded vectors, we present the center-sensitive averaging pooling, ensuring the embedding of a patch pays more attention to its central regions and contributes to better-aligned results. The center-sensitive averaging pooling can be seen as a convolutional operation, and its kernel values get bigger when closer to the center point. The kernel weight \( W_c \in \mathbb{R}^{|H| \times |W|} \) is defined as follows:

\[
W_c(h, w) = e^{- \frac{\|h - v_c\|^2 + \|w - w_c\|^2}{2\sigma^2}},
\]

\[
\text{dis}_h(h, w) = \frac{h}{h_{cen}} - 1, \quad \text{dis}_w(h, w) = \frac{w}{w_{cen}} - 1,
\]

\[
h \in [0, |H|/H_p], \quad w \in [0, |W|/W_p], \quad (25)
\]

where \( W_c(h, w) \) is a Gaussian-liked formulation and \( \sigma \) is a hyper-parameter to control the level of focusing on the center region. \((h_{cen}, w_{cen})\) is the center point of the kernel, i.e., \( h_{cen} = 0.5 |H|/H_p \) and \( w_{cen} = 0.5 |W|/W_p \).

**Training stage:** To make the trained model have the center-sensitive ability, we first fuse the center-sensitive averaging pooling into the formulation of patch-wise embedding \( s_n \) (3) and \( \pi_{n_1,n_2} \) (17) as:

\[
\hat{s}_{ni} = \frac{\sum v_n(h, w)W_c(h - h_b, w - w_b)}{\|\sum v_n(h, w)W_c(h - h_b, w - w_b)\|_2},
\]

\[
\hat{s}_{n_1,n_2} = \frac{\sum \hat{v}_{n_1,n_2}(h, w)W_c(h - h_b, w - w_b)}{\|\sum \hat{v}_{n_1,n_2}(h, w)W_c(h - h_b, w - w_b)\|_2},
\]

where \( h \in [h_b, h_c) \) and \( w \in [w_b, w_c) \). Then, \( \hat{s}_{ni} \) and \( \hat{s}_{n_1,n_2} \) are utilized to replace the original \( s_{ni} \) and \( \pi_{n_1,n_2} \) to compute \( L_{pd} \) (8) and \( L_{Hm} \) (20). Finally, the model is trained based on the modified loss function, which ensures that the patch comparison focuses on the center region.

**Test stage:** As shown in Fig. 6, the center-sensitive embedding for the test sample can be deduced by utilizing \( W_c \) to filter original pixels-wise embedded vectors \( v_n \), i.e.:

\[
\hat{v}_n = \text{conv}(v_n, W_c).
\]

Meanwhile, the support embedding \( \hat{s}_t \) is set as the center-sensitive embedding of the patch with the labeled point as the center. Then, the similarity map is captured as the cosine distance between \( \hat{s}_t \) and \( \hat{v}_n \), i.e., \( \text{sim}(\hat{s}_t, \hat{v}_n(h, w)) = \frac{\langle \hat{s}_t, \hat{v}_n(h, w) \rangle}{\|\hat{s}_t\|_2 \|\hat{v}_n(h, w)\|_2} \). After that, the point with the max similarity value can be identified:

\[
(h_{max}, w_{max}) = \arg \max_{h,w} \text{sim}(\hat{s}_t, \hat{v}_n(h, w)).
\]

And we further adopt threshold segmentation to get regions with similarity values greater than 95% of the max similarity value. Finally, we choose the connective region, which contains the point of \((h_{max}, w_{max})\), and get its centroid point \((h_d, w_d)\) as the final result.

### VI. EXPERIMENTAL RESULTS

This section contains three parts: (1) The generalization and robustness of our local discrimination model are analyzed in Section VI-A. (2) The comparative results of the shape-guided segmentation are presented in Section VI-B. (3) Section VI-C shows the results of the center-sensitive one-shot localization.

#### A. Local Discriminative Representation Learning

1) **Datasets and Preprocessing:** Color fundus datasets: The color fundus images from Kaggle diabetic retinopathy (DR) detection dataset are first utilized to train the representation. Then, the trained representation is transferred to 8 downstream tasks.

**Chest X-ray datasets:** The dataset of ChestX-ray14 with about 112k training data is utilized to train the representation. 4 downstream segmentation tasks are initialized by the learned representation.

The basic information of datasets is shown in Table I. Both the Kaggle DR dataset and ChestX-ray14 are real-world datasets containing images with different imaging conditions and from varied centers. To validate the robustness of our representation model to real-world datasets, we apply very limited preprocessing. Concretely, we only use the simple threshold segmentation to filter out the black background part of color fundus images and resize all images to 512 × 512.

2) **Implementation Details:** Network architecture: To make the training friendly in consumption time and GPU memory, we utilize a tiny version of VGG16 [57] as the encoder \( f(\cdot) \) and the decoder \( g(\cdot) \) is also designed as a light-weighted network. Specifically, the setting of \( f(\cdot) \) is almost the same as VGG16 without fully connected (FC) layers, but the number of channels
is reduced to a quarter of VGG16. For $g(\cdot)$, each of the first four base blocks is composed of an up-pooling layer and two convolutional layers. In the final part of the decoder, the clustering module and the embedding module, each of which is made up of two convolutional layers, are set to obtain the segmentation mask $r_n$ and the pixel-wise embedding $v_n$, respectively. It is worth noting that the features of each encoder block are connected with the corresponding decoder block by the skip connection layer.

Training process: Learning local discriminative features is an evolutionary process from patch discrimination to region discrimination. Patch discrimination accompanied by hypersphere mixup is first utilized to train the model to get initial discriminative ability, the loss is formulated as:

$$L = L_{pd} + L_{hm}.$$  

(29)

$f(\cdot)$ and $g(\cdot)$ are optimized by minimizing this loss in the first 20 epochs and updated $1k$ iterations in each epoch.

After pre-training model based on patch-wise loss, region discrimination is further fused into the training process, and the total loss ($L$) is defined as follows:

$$L = L_{pd} + L_{hm} + 10L_{rm} + 0.1L_{entropy},$$  

(30)

$$L_{entropy} = -\frac{1}{NCHW} \sum_{n,c,h,w} r_{nc}(h,w) \log(r_{nc}(h,w)) - \frac{1}{NCHW} \sum_{n,c,h,w} (1-r_{nc}(h,w)) \log(1-r_{nc}(h,w)),$$  

(31)

where $L_{entropy}$ is a constraint on entropy to make $r_{nc}(h,w)$ with high confidence. The model is optimized by minimizing the new loss in the following 80 epochs.

In each training iteration, four data groups are fed into the model. Concretely, two images are first randomly sampled and mixed to generate the mixup image followed (15); then, each sampled image is processed by different augmentation settings to generate two augmented versions; and the final group contains these four augmented images and the mixup image. The optimizer is Adam with the initial learning rate ($lr$) of 0.001, and $lr$ will decline to half after every 10 epochs.

Transfer to downstream tasks: The segmentation model for downstream tasks is a U-Net, its encoder is the same as $f(\cdot)$ and its decoder is similar to $g(\cdot)$ but without the embedding module, and a final layer is activated by the Sigmoid. We first utilize the learned feature extractor as the initial encoder for models of downstream tasks and only update the decoder in the first 100 epochs. Then, all of the model’s parameters are updated in the following 100 epochs. The Adam with $lr = 0.001$ is utilized as the optimizer, and 20% of the training data is split as the validation dataset to alleviate the over-fitting. The loss function is set as dice loss as follows:

$$L_{dice} = 1 - 2 \sum_{n} y_n^t y_n + \epsilon \sum_{n}(y_n^t + y_n + \epsilon),$$  

(32)

where $y_n^t \in Y^t$, $y_n \in Y$, $Y$ is the ground truth and $Y^t$ is the prediction. The smoothness term $\epsilon$ is set as 1.

Since the number of the image in chest X-ray datasets (SCR and SIIM-ACR) is relatively abundant, we make comparative experiments to analyze the improvements brought by our method for different data numbers in Section VI-A4. Specifically, we respectively use 20%, 40%, 60%, 80%, and 100% of the training data to train the network. Accordingly, The evaluation metric for chest X-ray tasks in this section is the average value of results based on different ratios of the training data.

Experiment settings: To construct augmentation image pairs, RandomResizedCrop, RandomGrayscale, ColorJitter, RandomHorizontalFlip, and Rotation90 in PyTorch are adopted. $H_p$ and $W_p$ are set as 8 to divide an image into $8 \times 8$ patches, and $C$ is set as 16 to cluster 16 regions for each image.

Experiments are performed on a workstation platform with Intel(R) Xeon(R) CPU E5-2640 v4 @ 2.40 GHz, 128 GB RAM and 4 $\times$ NVIDIA GeForce GTX 1080 GPU. The code is implemented in Ubuntu 18.04LTS with PyTorch 1.8.0.

3) Comparative Experiments: We compare our representation learning framework with recent SOTA methods. The comparative methods contain: Scratch is the network trained from scratch. Supervised is the supervised representation learning framework based on classification labels. Instance-wise contrastive learning: InstDisc [21], EmbInva [22], MoCo [12], MoCoV2 [23], SimCLR [13], DenseCL [14]. Clustering-based methods: DC [58], JC [59], SwAV [17], CC [19], and PCL [18]. SwAV, CC, and PCL are cluster-wise contrastive learning methods. Self-supervised medical representation learning methods: RotOri [5], CR [60], CGM [61], C2L [62], SimTriple [25], and EMMD [4]. PJM121 [11], Pd, Pd+mix, and Rd are respectively the model in the conference version, patch discrimination, patch discrimination accompanied by hypersphere mixup, and region discrimination.
TABLE II
COMPARISON OF THE RESULTS OF DOWNSTREAM SEGMENTING TASKS (DSC%)
By comparing the results of Pd, Pd+mix, and Rd, we can see that our proposed evolutionary framework works well to obtain continuous improvements. Meanwhile, the effectiveness of hyperspace mixup (Pd+mix) is demonstrated by improving the mean DSC of patch discrimination by 0.46% for color fundus tasks and 1.83% for chest X-ray tasks.

Ablation Experiments: Patch-out size: The model is trained by patch discrimination accompanied by hypersphere mixup, and the patch-out size \((H_p, W_p)\) is respectively set as (1,1), (2,2), (4,4), (8,8) and (16,16). The mean DSC and the detailed DSC for different patch-out sizes are shown in Fig. 8. We can see that:

1) The performance is rapidly improved when the patch-out size changes from (1,1) to (8,8), which shows the finer representation learned by comparing smaller patches is more generalized for downstream tasks. And the mean DSC changes slowly after the patch-out size is larger than (8,8). Thus, we set \(H_p = W_p = 8\) in the following experiments.

2) When the patch-out size changes from (8,8) to (16,16), the image is divided into more patches with smaller sizes. For the color fundus image with a large background region, the number of background patches lacking informative structure information significantly increases; consequently, the corresponding models are disturbed by these patches, and the performance degenerates (shown in Fig. 8(a)). For the chest X-ray image, it has complex structures spreading over the image and its background region is relatively small, thus, the influence of background patches is less significant. Meanwhile, the comparisons among smaller patches contribute to features with a better ability to distinguish tiny regions, satisfying the demand of lesion segmentation. Therefore, the result of pneumothorax segmentation is significantly improved, ensuring the total results go up (shown in Fig. 8(d) and (b)).

Network complexity: Encoders with different complexities are implemented as the feature extractor of our model, containing tiny VGG16 (similar to VGG16 but with a quarter of the channels), small VGG16 (similar to VGG16 but with half the channels), VGG16, ResNet-18, ResNet-34, and ResNet-50. And it is worth noting that the first layer of ResNet, which is originally a convolutional layer with the kernel of \(7 \times 7\) and the stride of \(2 \times 2\), is replaced by two \(3 \times 3\) convolutional layers and a \(2 \times 2\) max-pooling to provide finer features for segmentation. According to the results shown in Fig. 9 and the detailed parameter information shown in Table III, we have the following conclusions:

1) Compared with models trained from scratch, models pre-trained by our method can get significant improvements (as the red font values shown in Fig. 9).

2) When the number of parameters in the randomly initialized decoder is relatively small, the performance of the model is improved with the increase of the model complexity. Specifically, all of tiny VGG16, small VGG16, ResNet-18, VGG16, and ResNet-34 have the corresponding decoder with no more than 8 M (million) parameters. When the encoder changes from tiny VGG16 to small VGG16 to ResNet-18 to VGG16 to ResNet-34, the corresponding performance is roughly improved with the increase of model parameters. VGG16 and ResNet-34 have similar numbers of parameters, and they respectively gain the best results in color fundus tasks and chest X-ray tasks.

3) If the randomly initialized decoder has extremely large parameters, it is challenging to improve the performance with the increase of the model complexity. Specifically, the corresponding decoder for ResNet-50 is extremely complex with 40.37 M parameters. Consequently, while...
TABLE IV

| Cluster num | 0 | 4  | 8  | 12 | 16 |
|-------------|---|----|----|----|----|
| Mean DSC for color fundus | 84.65 | 84.51 | 84.92 | 84.86 | 85.12 |
| Mean DSC for Chest X-ray | 81.81 | 82.33 | 82.12 | 82.15 | 82.11 |

![Fig. 10. The results with different numbers of data (DSC%).](image)

ResNet-50 is the most complex model, its performance is not satisfactory.

**Cluster number**: We respectively set $C$ as 4, 8, 12, 16 and the comparative results are shown in Table IV. We have the following conclusions:

1) The models further trained by region discrimination have better results than models without it (i.e., $C = 0$). The reason is that the model trained by region discrimination, focusing on clustering pixels to form regions with interior semantic consistency and distinguishing regions of different structures/tissues, can measure pixel-wise/region-wise similarity, satisfying the demand of medical image analysis for processing tiny targets.

2) For targets with small sizes, the model trained by region discrimination with a bigger $C$ can contribute to better performance. Target structures in color fundus images, such as vessels and lesions, are of small size, making results with $C = 16$ the best. However, since targets in chest X-ray tasks have relatively big sizes, the small clustering number of $C = 4$ is enough to reach the best results.

**Annotation demand**: Based on the chest X-ray downstream tasks, we investigate the ability of our method to reduce the demand for labeled datasets. Specifically, 20%, 40%, 60%, 80%, and 100% of the annotated data in the downstream tasks are set as the training data for models. According to the results shown in Fig. 10, we can see that:

1) Models based on the trained model of our method (Rd) have the best results in most situations (15 out of 20).

2) In the situation that the training data is extremely scarce (20% of training data are available), Rd can significantly improve the performance by 8.81%, 7.1%, and 12.49% for the segmentation of heart, clavicle, and pneumothorax.

3) Models based on Rd can achieve better results with fewer data. Specifically, utilizing 40%, 40%, 60%, and 60% of the training data can ensure that our models have better results than segmentation models trained from scratch with full data of heart, lung, clavicle, and pneumothorax.

**B. Shape-Guided Segmentation**

1) **Implementation Details**: **Datasets**: About 11-k color fundus images from Kaggle DR [46] are set as the training data; 10 OCTA vessel masks and 10 DSA vessel masks are utilized as the reference masks. The trained model is evaluated in the test data of CHASEDB [47], HRF [48], RITE [49] and STARE [50], [51].

**Training details for clustering targets**: Similar to Section VI-A2, we first warm up models by patch discrimination with hypersphere mixup in the first 20 epochs. Then, in the following 80 epochs, we aim to generate segmentation masks ($R_{nm} = \{r_{nm}\}_{n=1}^N$) to cluster target regions that are composed of semantically consistent pixels and possess a similar shape distribution to reference masks ($R_e = \{r_{en}\}_{n=1}^N$). We utilize adversarial learning [42] to align the shape distribution of predictions and reference masks. Specifically, besides the original $f(\cdot)$ and $g(\cdot)$, we build a discriminator network $d(\cdot)$ composed of 7 convolutional layers and 2 FC layers. The channel number for layers is respectively 16, 32, 32, 32, 32, 32, and 1. After each convolutional layer, a LeakyReLU and a (2,2) max-pooling layer are set. Features of CNN are further processed by a global averaging pooling and fed into FC layers. The final output is activated by the Sigmoid.

The training can be divided into two parts. In part one, the discriminator is updated to discriminate $R_e$ and $R_m$. The loss function $L_d$ is defined as:

$$L_d = L_{bce}(d(R_e), 1) + L_{bce}(d(R_m), 0),$$

$$L_{bce}(y', y) = -\frac{1}{N} \sum_{n} (y_n \log y_n' + (1 - y_n) \log(1 - y_n')).$$

In part two, $f(\cdot)$ and $g(\cdot)$ are updated to simultaneously cluster semantically similar pixels and cheat $d(\cdot)$. Accordingly, the loss function can be defined as follows:

$$L = L_{pd} + L_{hm} + 10L_{rd} + 0.1L_{entropy} + L_{adv},$$

$$L_{adv} = L_{bce}(d(R_m), 1).$$

The optimizers for both parts are Adams. And to ensure the training stability, we set $lr$ to 0.00005 for the first part and 0.001 for the second part.

**Training details for refinement segmentation**: We first capture 30 re-clustered results with different augmentations to identify pseudo labels and uncertainty maps according to (22) and (23). Then a new network is built, which has similar architectures as $f(\cdot)$ and $g(\cdot)$, but the embedding module is removed from $g(\cdot)$ and the final layer of $g(\cdot)$ is a convolutional layer with only one output channel. The loss function is $L_{bce}$ defined in (24). The optimizer is an Adam with $lr = 0.001$ and the max training epoch is 100.

**Comparative methods**: CycleGAN is the basic model for synthetic segmentation [65], [66]. SynSeg [30] is an improved
framework based on cycleGAN, IPMI21 shows the results of the previous conference version. OutOri, OutReclu and OutUncer respectively show the results of the clustering module, the re-clustering refinement and the uncertainty estimation refinement in this study.

2) Experimental Results: The results based on 10 DSA masks and 10 OCTA masks are shown in Table V, and Fig. 11 shows some synthetic samples and predictions. The results demonstrate that:

1) Overall, when transferring knowledge across modalities with a big content gap, our method significantly outperforms CycleGAN and SynSeg. Specifically, compared with SynSeg, our method gains a 27.22% improvement in DSA-guided experiments, and a 44.02% improvement in OCTA-guided experiments. As shown in Fig. 11, although CycleGAN-based methods can get synthetic samples with similar styles as the color fundus image, the significant difference in contents makes the model trained based on the synthetic images failed in the real color fundus image.

2) Compared with our conference version (IPMI21), the enhanced model in this study (OutUncer) has significantly better results with 20.11% DSC gains in DSA-guided segmentation and 9.32% DSC gains in OCTA-guided segmentation, which shows its effectiveness and superiority.

3) In comparison with the results of OutOri, the results of OutReclu are better, demonstrating utilizing individual cluster prototypes to make pixel re-clustering for each image is effective. As shown in Fig. 11, OutReclu can recognize more thin vessels compared with OutOri.

4) The results of OutUncer show that it is an effective strategy to further improve segmentation by utilizing reliable pseudo labels identified by the uncertainty estimation to re-train the model. Compared with OutClu, OutUncer can further remove false positives as shown in Fig. 11.

C. One-Shot Localization

1) Implementation Details: Datasets: 1) OD localization (HRF) [48]: The center of OD is manually labeled in 45 color fundus images. One data is set as the support data and the rest is set as the query data. 2) Heart localization (SCR) [56]: The centroid points are identified by the manual heart masks. Similarly, one annotated sample is set as the support data.

Training details: The basic model is trained based on patch discrimination accompanied by hypersphere mixup to obtain the ability to measure patches’ similarity, specifically, the loss is defined as $L = L_{Pd} + L_{Hm}$. Then, we conduct the following two improvements to make the learned representation more suitable for landmark localization:

1) Multi-size pooling: Rather than utilizing the adaptive average pooling with a fixed output size to gain the patch-wise representation, we use the average pooling layer with a random pooling size ranging from 28 to 112 to ensure that our model is adaptive for targets with varied sizes.

2) Center-sensitive averaging: As described in Section V-B, the averaging pooling layer is replaced by the center-sensitive averaging pooling to obtain well-aligned predictions. The $\sigma$ in (25) is set as 0.5.

The total training epoch is 100, the optimizer is an Adam with $lr = 0.001$, and four groups of images form a batch as described in Section VI-A2.

Evaluation metric: To evaluate the performance, we first compute the euclidean distance between the real landmark point $(h_t, w_t)$ and the detected point $(h_d, w_d)$ as follows:

$$\text{dis}((h_d, w_d), (h_t, w_t)) = \sqrt{(h_d - h_t)^2 + (w_d - w_t)^2}.$$  

(37)

Then, we define various threshold distances to identify the correct detection; specifically, the landmark is correctly localized if the distance between it and the predicted point is less than the threshold distance. The threshold distance is respectively set as 0.05, 0.1, 0.15, 0.2, 0.25 of the pooling size.

Competitive methods: Besides 2 recent self-supervised one-shot landmark localization methods (CC2D [37], RPR-Loc [36]), we compare our method with 9 template matching methods, including 5 template matching methods based on hand-crafted features (CCOEFF-N [67], CCORR [67], CCORR-N [67], SQDIFF [67], SQDIFF-N [67]), and 4 template matching methods based on the representation learned by SOTA.
TABLE VI
ACCURATE RATES WITH DIFFERENT THRESHOLD DISTANCES FOR ONE-SHOT LOCALIZATION (%)

| Methods   | 0.05 | 0.1  | 0.15 | 0.2  | 0.25 | Mean |
|-----------|------|------|------|------|------|------|
| OD localization |      |      |      |      |      |      |
| CCOEFF-N  | 22.72| 40.90| 40.90| 43.18| 43.18| 38.18|
| CCORR     | 4.54 | 18.18| 31.81| 36.36| 45.45| 27.27|
| CCORR-N   | 13.63| 31.81| 31.81| 31.81| 31.81| 28.18|
| SQDIFF    | 15.90| 34.09| 34.09| 34.09| 34.09| 30.45|
| SQDIFF-N  | 18.18| 34.09| 34.09| 36.36| 36.36| 31.81|
| MoCo      | 4.34 | 15.90| 20.45| 27.27| 27.27| 19.09|
| C2L       | 4.54 | 6.81 | 13.63| 20.45| 22.72| 13.63|
| SimTriplet| 9.09 | 25   | 31.81| 43.18| 54.54| 32.72|
| DenseCL   | 6.81 | 18.18| 38.63| 54.54| 63.63| 36.36|
| CC2D      | 0.00 | 6.82 | 22.73| 29.55| 29.55| 17.73|
| RPR-Loc   | 0.00 | 2.27 | 2.27 | 11.36| 18.18| 6.82 |
| Ours      | 25   | 61.36| 88.63| 97.72| 97.72| 74.09|

| Heart localization |      |      |      |      |      |      |
|--------------------|------|------|------|------|------|------|
| CCOEFF-N           | 2.44 | 14.23| 26.02| 39.80| 46.34| 25.37|
| CCORR              | 0.41 | 0.41 | 0.41 | 0.41 | 0.41 | 0.41 |
| CCORR-N            | 2.44 | 12.60| 28.46| 38.62| 54.88| 27.40|
| SQDIFF             | 1.63 | 6.10 | 14.23| 20.33| 27.24| 13.90|
| SQDIFF-N           | 1.63 | 6.10 | 13.82| 20.33| 27.64| 13.90|
| MoCo               | 2.44 | 8.13 | 13.41| 19.51| 27.24| 14.15|
| C2L                | 3.66 | 10.57| 18.29| 29.27| 37.80| 19.92|
| SimTriplet         | 2.44 | 4.47 | 9.35 | 15.04| 19.92| 10.24|
| DenseCL            | 5.28 | 19.11| 34.15| 52.03| 65.85| 35.28|
| CC2D               | 8.54 | 26.02| 49.19| 67.48| 80.49| 46.34|
| RPR-Loc            | 6.10 | 22.36| 42.68| 62.20| 80.89| 42.85|
| Ours               | 18.29| 44.31| 55.28| 63.82| 67.07| 49.76|

TABLE VII
ABLATION EXPERIMENTS FOR ONE-SHOT LOCALIZATION (%)

| Methods   | 0.05 | 0.1  | 0.15 | 0.2  | 0.25 | Mean |
|-----------|------|------|------|------|------|------|
| OD localization |      |      |      |      |      |      |
| Baseline  | 6.81 | 22.72| 36.36| 54.54| 65.90| 37.27|
| Baseline+ms | 6.81 | 29.54| 59.09| 72.72| 88.63| 51.36|
| Baseline+ms+cs | 25   | 61.36| 88.63| 97.72| 97.72| 74.09|
| Heart localization |      |      |      |      |      |      |
| Baseline  | 2.03 | 7.32 | 17.89| 32.93| 46.75| 21.38|
| Baseline+ms | 1.22 | 10.98| 23.58| 36.18| 52.65| 24.96|
| Baseline+ms+cs | 18.29| 44.31| 55.28| 63.82| 67.07| 49.76|

2) Experimental Results: The comparative results are shown in Table VI and the results of the ablation experiments are shown in Table VII, where Baseline is the original patch discrimination accompanied by hypersphere mixup, Baseline+ms is the enhanced model with multi-size pooling, and Baseline+ms+cs is the improved model with multi-size center-sensitive averaging pooling. Furthermore, some localization results and similarity maps of Baseline+ms+cs are shown in Fig. 12. Accordingly, we come to the following conclusions:

1) Overall, our proposed center-sensitive one-shot landmark localization method is more generalized and gains the best results in both tasks.

2) For template matching methods, reliable features that are capable of measuring regional similarity are critically important to achieving stable performance in landmark localization tasks. Hand-crafted features are sensitive to image characteristics, such as color and contrast, leading to unstable performance in different tasks (as shown by the results of CCORR, SQDIFF, and SQDIFF-N). Since most of the SOTA contrastive learning methods, such as MoCo and SimTriplet, aim to learn global discriminative features, neglecting regional similarity, the representation learned by these methods may not be suitable for the localization tasks. Comparatively, the focus on measuring patch-wise similarity and the design of center-sensitive averaging pooling, secure our model the robust performance in achieving accurate landmark localization.

3) While the self-supervised one-shot localization methods (CC2D, RPR-Loc) achieve relatively competitive results in the heart localization task for chest X-ray images, the performance of both methods declines dramatically in the OD localization task for color fundus images. The major reason for the unsatisfactory results of CC2D is that the CC2D is based on a pixel-matching mechanism, which is fragile for the color fundus image with a large number of meaningless background pixels. Since RPR-Loc requires the training dataset with stable relative positions, the differences between the left and the right eyes, the influence of different photography centers, and the varied fields of view, all together account for the less satisfactory results of RPR-Loc on color fundus images. In comparison, the augmentation-invariant and orientation-invariant patch discriminative features as learned and utilized in our method contribute to better generalization and robust performance in both medical tasks.

4) Further, as shown in Fig. 12, the center-sensitive averaging pooling empowers the model to learn features focusing on the patch’s central region. Thus, the matching point has a peak value in the heat-map and the detected results align well with the target.

5) As shown by the results in Table VII, the effectiveness of the multi-size pooling and the center-sensitive averaging pooling is evidenced by the continuous improvements.

VII. CONCLUSION

In this work, we proposed an unsupervised local discriminative representation learning framework to address the challenging task of distinguishing tiny structures and tissues from different medical modality images and for various medical...
The generalization and robustness of the representation learned by our method are fully validated on 12 downstream tasks, and our method achieves the best results on 9 tasks when compared to 18 SOTA methods. Based on the thorough analyses of the clinical applications, our framework attains a new perspective on knowledge transfer among modalities with enormous content gaps exemplar with vessel segmentation in color fundus images taking only shape prior from OCTA and DSA. Moreover, patch discrimination is further enhanced by the center-sensitive averaging pooling for one-shot landmark localization, which facilitates the detection of OD center and heart center.

Discriminative representation learning for medical image analysis remains an open question, and there are three particular challenges to be addressed in our future work. First, while our method has comprehensively explored the intra-modality structure similarity and the inter-modality shape similarity to learn discriminative features and transfer shape knowledge to realize segmentation, further intra-modality/inter-modality properties, such as the common property across images with various field of views, could be taken into account to learn more generalized discriminative features. Second, an adaptive algorithm, which may be designed as a meta-learning model to identify the hyperparameters, could further improve our model. Finally, we would like to translate the open-source framework to potentially support more clinical applications, such as one-shot medical image segmentation and anomaly detection.

REFERENCES

[1] Z. Ahmed, K. Mohamed, S. Zeeshan, and X. Dong, “Artificial intelligence with multi-functional machine learning platform development for better healthcare and precision medicine,” Database, vol. 2020, 2020, Art. no. baa010.
[2] P. Meyer, V. Noblet, C. Mazzara, and A. Lallemont, “Survey on deep learning for radiotherapy,” Comput. Biol. Med., vol. 98, pp. 126–146, 2018.
[3] G. Litjens et al., “A survey on deep learning in medical image analysis,” Med. Image Anal., vol. 42, pp. 60–88, 2017.
[4] X. Li, M. Jia, M. T. Islam, L. Yu, and L. Xing, “Self-supervised feature learning via exploiting multi-modal data for retinal disease diagnosis,” IEEE Trans. Med. Imaging, vol. 39, no. 12, pp. 4023–4033, Dec. 2020.
[5] X. Li et al., “Rotation-oriented collaborative self-supervised learning for retinal disease diagnosis,” IEEE Trans. Med. Imaging., vol. 40, no. 9, pp. 2284–2294, Sep. 2021.
[6] Y. Han, C. Chen, A. Tewfik, Y. Ding, and Y. Peng, “Pneumonia detection on chest X-ray using radiomic features and contrastive learning,” in Proc. IEEE/ACM Comput. Vis. Pattern Recognit., 2021, pp. 247–251.
[7] Y. Zhang, H. Jiang, Y. Miura, C. D. Manning, and C. P. Langlotz, “Contrastive learning of medical visual representations from paired images and text,” 2020, arXiv: 2010.00747.
[33] E. Akagunduz, A. G. Bors, and K. K. Evans, “Defining image memorability using the visual memory schema,” IEEE Trans. Pattern Anal. Mach. Intell., vol. 42, no. 9, pp. 1916–1925, Sep. 2020.

[34] X. Xu, F. Zhou, B. Liu, D. Fu, and X. Bai, “Efficient multiple organ localization in CT image using 3D region proposal network,” IEEE Trans. Med. Imag., vol. 38, no. 8, pp. 1885–1898, Aug. 2019.

[35] S. K. Zhou et al., “A review of deep learning in medical imaging: Imaging traits, technology trends, case studies with progress highlights, and future promises,” in Proc. IEEE, vol. 109, no. 5, pp. 820–838, May 2021.

[36] W. Lei et al., “Contrastive learning of relative position regression for one-shot object localization in 3D medical images,” in Proc. Int. Conf. Comp. Comput.- Assist. Intervention, 2021, pp. 155–165.

[37] Q. Yao, Q. Quan, L. Xiao, and S. Kevin Zhou, “One-shot medical landmark detection,” in Proc. Med. J. Comp. Comput. Comput.- Assist. Intervention, 2021, pp. 177–188.

[38] G. Hinton, O. Vinyals, and J. Dean, “Distilling the knowledge in a neural network,” 2015, arXiv:1503.02531.

[39] X. Wang, Y. Peng, L. Lu, Z. Lu, M. Bagheri, and R. M. Summers, “ChestX-ray8: Hospital-scale chest X-ray database and benchmarks on weakly-supervised classification and localization of common thorax diseases,” in Proc. IEEE Trans. Med. Imag., vol. 38, no. 8, pp. 1885–1898, Aug. 2019.

[40] H. Zhang, M. Cisse, Y. N. Dauphin, and D. Lopez-Paz, “mixup: Beyond empirical risk minimization,” in Proc. Int. Conf. Learn. Representations, 2018, pp. 1–9.

[41] V. Ververa et al., “Manifold mixup: Better representations by interpolating hidden states,” in Proc. Int. Conf. Mach. Learn., 2019, pp. 6434–6447.

[42] I. Goodfellow et al., “Generative adversarial nets,” in Proc. Adv. Neural Inf. Process. Syst., 2014, pp. 2672–2680.

[43] A. Kendall and Y. Gal, “What uncertainties do we need in Bayesian deep learning for computer vision?” in Proc. Adv. Neural Inf. Process. Syst., 2017, pp. 5574–5584.

[44] C. Bian et al., “Uncertainty-aware domain alignment for anatomical structure segmentation,” Med. Image Anal., vol. 64, 2020, Art. no. 101732.

[45] J. Snell, K. Swersky, and R. Zemel, “Prototypical networks for few-shot learning,” in Proc. 31st Int. Conf. Neural Inf. Process. Syst., 2017, pp. 4080–4090.

[46] J. Cudnov and G. Bresnick, “EyePACS: An adaptable telemedicine system for diabetic retinopathy screening,” J. diabetes Sci. Technol., vol. 2, no. 1, 2008, Art. no. 1004.

[47] M. M. Fraz et al., “An ensemble classification-based approach applied to retinal blood vessel segmentation,” IEEE Trans. Biomed. Eng., vol. 59, no. 9, pp. 2538–2548, 2012.

[48] A. Buadi, R. Bock, A. Maier, J. Hornegger, and G. Michelson, “Robust vessel segmentation in fundus images,” Int. J. Biomed. Imag., vol. 2013, Art. no. 154860.

[49] Q. Hu, M. D. Abràmoff, and M. K. Garvin, “Automated separation of blood vessels using the fuzzy convergence of the blood vessels,” Med. Image Anal., vol. 141, pp. 192–199, 2019.

[50] J. Cuadros and G. Bresnick, “EyePACS: An adaptable telemedicine system for diabetic retinopathy screening,” J. diabetes Sci. Technol., vol. 2, no. 1, 2008, Art. no. 1004.

[51] H.-Y. Zhou, S. Yu, C. Bian, Y. Hu, K. Ma, and Y. Zheng, “Comparing to learn: Surpassing imagenet pretraining on radiographs by comparing image representations,” in Proc. Int. Conf. Med. Imag. Comput. Comput.- Assist. Intervention, 2020, pp. 398–407.

[52] M. Long, Y. Cao, Z. Cao, J. Wang, and M. I. Jordan, “Transferable representation learning with deep adaptation networks,” IEEE Trans. Pattern Anal. Mach. Intell., vol. 41, no. 12, pp. 3071–3085, Dec. 2019.

[53] K. He, X. Zhang, S. Ren, and J. Sun, “Deep residual learning for image recognition,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit., 2016, pp. 770–778.

[54] X. Dong et al., “Synthetic MRI-aided multi-organ segmentation on male pelvic CT using cycle consistent deep attention network,” Radiotherapy Oncol., vol. 141, pp. 192–199, 2019.

[55] V. Sandfort, K. Yan, P. J. Pickhardt, and R. M. Summers, “Data augmentation using generative adversarial networks (CycleGAN) to improve generalizability in CT segmentation tasks,” Sci. Rep., vol. 9, no. 1, pp. 1–9, 2019.

[56] G. Bradski and A. Kaehler, Learning OpenCV: Computer Vision With the OpenCV Library. Sebastopol, CA, USA: O’Reilly Media, Inc., 2008.

Hui Chen received the BSc degree in electronic science and technology from Beijing Normal University, Beijing, China, in 2017. He is currently working toward the PhD degree in control science and engineering with Shanghai Jiao Tong University. His primary research interests include medical image processing, self-supervised learning, and contrastive learning.

Renzhen Wang received the BSc degree from the Dalian University of Technology, Dalian, China, in 2016, and the PhD degree from Xi’an Jiaotong University, Xi’an, China, in 2022. He is currently an Assistant Professor with the School of Mathematics and Statistics, Xi’an Jiaotong University. His current research interests include semi-supervised learning, continual learning, and medical image analysis.

Xiying Wang received the PhD degree in computer science from the University of Sydney, in 2006. She is currently an associate professor and Associate Director with the Multimedia Lab, School of Computer Science, University of Sydney. Her research interests include quantitative data analysis and fusion as relates to multimodal imaging data computing, such as clustering and segmentation, image registration for biomedical and multimedia applications, saliency detection, identification, and visual analytics of medical and health data.
Jieyu Li received the bachelor’s degree from Zhejiang University, Hangzhou, China, in 2014, and PhD degree from Shanghai Jiao Tong University, Shanghai, China, in 2021. She was a visiting PhD student in Medical Image Processing Group, University of Pennsylvania, Philadelphia, Pennsylvania, from 2017 to 2020. She is currently a postdoc researcher with Tsinghua University, Beijing, China. Her research interest includes focuses on biomedical image analysis.

Qu Fang received the BSc degree from Shanghai Tongji University, Shanghai, China, in 2007. He is currently working with Changchun GeneScience Pharmaceutical Company, Ltd as chief technology officer of Medical Device Center. His current research interest includes the application of computer vision in the field of medical device.

Hui Li received the bachelor’s degree in 2018 and master’s degree from Shanghai Jiao Tong University, Shanghai, China, in 2021. She is currently working toward the PhD degree in information and communication engineering with Shanghai Jiao Tong University. Her research interests include focuses on biomedical image analysis, vision, and language navigation.

Jianhao Bai received the bachelor’s degree in medicine from Shanxi Medical University, in 2020. He is currently working toward the master’s degree with Tongji University. Meanwhile, he is also a resident of the Shanghai Tenth people’s Hospital. His current research direction interests include the basic and clinical diagnosis and treatment of vitreoretinal diseases and the development and application of multimodal imaging.

Qing Peng received the PhD degree from Tongji Medical University, WuHan, China, in 2006. She is currently a chief ophthalmologist with the Shanghai Tenth People’s Hospital of Tongji University. Her current basic research and clinical work focus on the diagnosis and treatment of vitreoretinal diseases, especially the multimodal images of fundus diseases.

Deyu Meng (Member, IEEE) received the BSc, MSc, and PhD degrees from Xi’an Jiaotong University, Xi’an, China, in 2001, 2004, and 2008, respectively. He is currently a professor with the School of Mathematics and Statistics, Xi’an Jiaotong University, and adjunct professor with the Faculty of Information Technology, The Macau University of Science and Technology. He currently serves as an associate editor of the IEEE Transactions on Pattern Recognition and Machine Intelligence, Science China-Information Sciences, and Frontiers of Computer Science. His current research interests include model-based deep learning, variational networks, and meta-learning.

Lisheng Wang received the MS degree in mathematics and the PhD degree in electronic and information engineering from Xi’an Jiaotong University, China, in 1993 and 1999, respectively. In 2003, he joined the Department of Automation, Shanghai Jiao Tong University, China, and now is a professor. His research interests include AI analysis and visualization of 3D medical images, computer-aided diagnosis, and surgery planning.