





Unsupervised Learning of Local Discriminative Representation for Medical Images

Huai Chen¹ , Jieyu Li¹, Renzhen Wang², Yijie Huang¹, Fanrui Meng¹,
Deyu Meng², Qing Peng³, and Lisheng Wang¹ (✉) 

¹ Institute of Image Processing and Pattern Recognition, Department of Automation, Shanghai Jiao Tong University, Shanghai 200240, People's Republic of China

lswang@sjtu.edu.cn

² Xi'an Jiaotong University, Xi'an, People's Republic of China

³ Department of Ophthalmology, Shanghai Tenth People's Hospital, Tongji University, Shanghai, People's Republic of China

Abstract. Local discriminative representation is needed in many medical image analysis tasks such as identifying sub-types of lesion or segmenting detailed components of anatomical structures. However, the commonly applied supervised representation learning methods require a large amount of annotated data, and unsupervised discriminative representation learning distinguishes different images by learning a global feature, both of which are not suitable for localized medical image analysis tasks. In order to avoid the limitations of these two methods, we introduce local discrimination into unsupervised representation learning in this work. The model contains two branches: one is an embedding branch which learns an embedding function to disperse dissimilar pixels over a low-dimensional hypersphere; and the other is a clustering branch which learns a clustering function to classify similar pixels into the same cluster. These two branches are trained simultaneously in a mutually beneficial pattern, and the learnt local discriminative representations are able to well measure the similarity of local image regions. These representations can be transferred to enhance various downstream tasks. Meanwhile, they can also be applied to cluster anatomical structures from unlabeled medical images under the guidance of topological priors from simulation or other structures with similar topological characteristics. The effectiveness and usefulness of the proposed method are demonstrated by enhancing various downstream tasks and clustering anatomical structures in retinal images and chest X-ray images.

Keywords: Unsupervised representation learning · Local discrimination · Topological priors

1 Introduction

In medical image analysis, transferring pre-trained encoders as initial models is an effective practice, and supervised representation learning is widely applied,

while it usually depends on a large amount of annotated data and the learnt features might be less efficient for new tasks differing from original training task [4]. Thus, some researchers turn to study unsupervised representation learning [9, 17], and particularly unsupervised discriminative representation learning was proposed to measure similarity of different images [7, 16, 18]. However, these methods mainly learn the instance-wise discrimination based on global semantics, and cannot characterize the similarities of local regions in image. Hence, they are less efficient for many medical image analysis tasks, such as lesion detection, structure segmentation, identifying distinctions between different structures, in which local discriminative features are needed to be captured. In order to make unsupervised representation learning suitable for these tasks, we introduce local discrimination into unsupervised representation learning in this work.

It is known that medical images of humans contain similar anatomical structures, and thus pixels can be classifying into several clusters based on their context. Based on such observations, a local discriminative embedding space can be learnt, in which pixels with similar context will distribute closely and dissimilar pixels can be dispersed. In this work, a model containing two branches is constructed following a backbone network, in which an embedding branch is used to generate pixel-wise embedding features and a clustering branch is used to generate pseudo segmentations. Through jointly updating these two branches, pixels belonging to the same cluster will have similar embedding features and different clusters will have dissimilar ones. In this way, local discriminative features can be learnt in an unsupervised way, which can be used for evaluating similarity of local image regions.

The proposed method is further applied to several typical medical image analysis tasks respectively in fundus images and chest X-ray images: (1) The learnt features are utilized in 9 different downstream tasks via transfer learning, including segmentations of retinal vessel, optic disk (OD) and lungs, detection of haemorrhages and hard exudates, etc., to enhance the performances of these tasks. (2) Inspired by specialists' ability of recognizing anatomical structures based on prior knowledge, we utilize the learnt features to cluster local regions of the same anatomical structure under the guidance of topological priors, which are generated by simulation or from other structures with similar topology.

2 Related Work

Instance discrimination learning method [1, 7, 16, 18] is an unsupervised representation learning framework providing a good initialization for downstream tasks and it can be considered as an extension of exemplar convolution neural network (CNN) [4]. The main conception of instance discrimination is to build an encoder to dispersedly embed training samples over a hypersphere [16]. Specifically speaking, a CNN is trained to project each image onto a low-dimensional unit hypersphere, in which the similarity between images can be evaluated by cosine similarity. In this embedding space, dissimilar images are forced to be

separately distributed and similar images are forced to be closely distributed. Thus, the encoder can make instance-level discrimination. Wu et al. [16] introduce a memory bank to store historical feature vectors for each image. Then the probability of image being recognized as i -th example can be expressed by inner product of the embedding vector and vectors stored in the memory bank. And the discrimination ability of encoder is obtained by learning to correctly classify image instance into the corresponding record in the memory bank. However, the vectors stored in the memory bank are usually outdated caused by discontinuous updating. To address this problem, Ye et al. [18] propose a framework with siamese network which introduces augmentation invariant into the embedding space to cluster similar images to realize real-time comparison.

The ingenious design enables instance discrimination effectively utilize unlabeled images to train a generalized feature extractor for downstream tasks and shrink the gap between unsupervised and supervised representation learning [7]. However, summarizing a global feature for image instance miss local details, which are crucial for medical image tasks, and the high similarity of global semantics between images of same body part makes instance-wise discrimination less practical. Therefore, it is more convinced to focus on local discrimination of medical images. Meanwhile, medical images of the same body part can be divided into several clusters due to the similar anatomical structures, which inspires us to propose a framework to cluster similar pixels to learn local discrimination.

3 Methods

The illustration of our unsupervised framework is shown in Fig. 1. This model has two main components. The first is learning a local discriminative representation, which aims to project pixels into a l_2 -normalized low-dimensional space, i.e. a K -D unit hypersphere, and pixels with similar context should be closely distributed and dissimilar pixels should be far away from each other on this embedding space. The learnt local discriminative representation can be taken as a good feature extractor for downstream tasks. The second is introducing prior

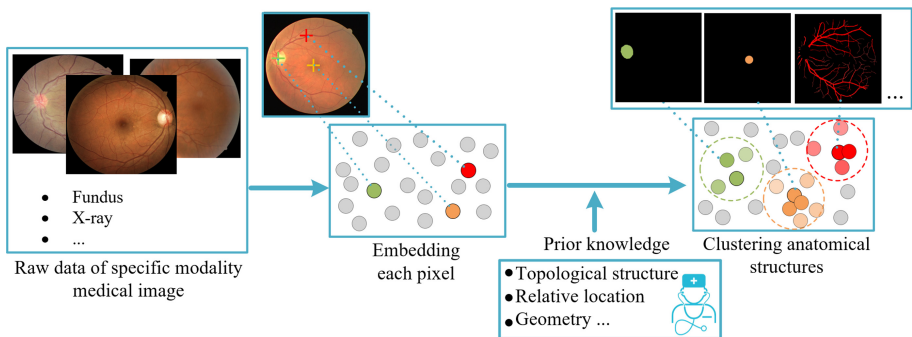


Fig. 1. Illustration of our proposed learning model.

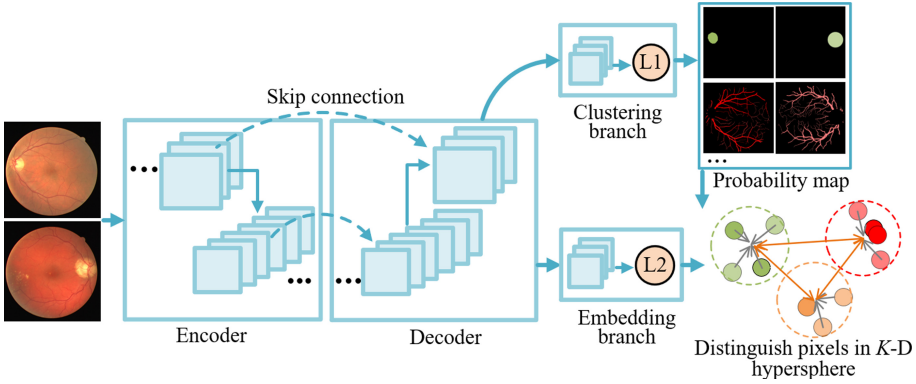


Fig. 2. Illustration of local discrimination learning.

knowledge of topological structure and relative location into local discrimination, where the prior knowledge will be fused into the model to make the distribution of pseudo segmentations closer to the distribution of priors. By combining priors of structures with local discrimination, regions of the expected anatomical structure can be clustered.

3.1 Local Discrimination Learning

As medical images of the same body region contain same anatomical structures, image pixels can be classified into several clusters, each of which corresponds to a specific kind of structure. Therefore, local discrimination learning is proposed to train representations to embed each pixel onto a hypersphere, on which pixels with similar context will be encoded closely. To achieve this, two branches, including an embedding branch to encode each pixel and a clustering branch to generate pseudo segmentations to cluster pixels, are built following a backbone network and trained in a mutually beneficial manner (Fig. 2).

Notation: We denote f_θ as the deep neural network, where θ is the parameters of network. The unlabeled image examples are denoted as $X = \{x_1, \dots, x_N\}$ where $x_i \in \mathbb{R}^{H,W}$. After feeding x_i into the network, we can get embedding features v_i and probability map r_i , i.e., $v_i, r_i = f_\theta(x_i)$, where $v_i(h, w) \in \mathbb{R}^K$ is the K -dimensional encoded vector for position (h, w) of image x_i and $r_i(h, w) \in \mathbb{R}^M$ is a vector representing the probability of classifying pixel $x_i(h, w)$ into M clusters. And $r_{mi}(h, w)$ denotes the probability of classifying pixel $x_i(h, w)$ into the m -th cluster. We force $\|r_i(h, w)\|_1 = 1$ and $\|v_i(h, w)\|_2 = 1$ by respectively setting l_1 and l_2 normalization in clustering branch and embedding branch.

Jointly Train Clustering Branch and Embedding Branch: After getting embedding features and pseudo segmentations, the center embedding feature c_m

of m -th cluster can be formulated as followed:

$$c_m = \frac{\sum_{i,h,w} r_{mi}(h,w)v_i(h,w)}{\|\sum_{i,h,w} r_{mi}(h,w)v_i(h,w)\|_2} \quad (1)$$

Where l_2 normalization is used to make c_m on the hypersphere. Thus, the similarity between c_m and $v_i(w,h)$ can be evaluated by cosine similarity as followed:

$$t(c_m, v_i(w,h)) = c_m^T v_i(w,h) \quad (2)$$

To make pixels of same cluster closely distributed and pixels of different clusters dispersedly distributed, there should be high similarity between $v_i(w,h)$ and corresponding center embedding features c_m , and low similarity between c_m and $c_n(m \neq n)$ as well. Thus, the loss function can be formulated as followed:

$$loss_{ld} = -\frac{1}{MNHW} \sum_{m,i,h,w} r_{mi}(h,w)t(c_m, v_i(w,h)) + \frac{1}{M(M-1)} \sum_{m,n \neq m} c_m^T c_n \quad (3)$$

More Constraints: We also add entropy loss and area loss to make high confidence of predictions and avoid blank outputs for some clusters. The losses are as followed:

$$loss_{entropy} = -\frac{1}{MNHW} \sum_{m,i,h,w} r_{mi}(h,w) \log r_{mi}(h,w) \quad (4)$$

$$area_{mi} = \sum_{h,w} r_{mi}(h,w) \quad (5)$$

$$loss_{area} = \frac{1}{NM} \sum_{m,i} \text{relu}(\frac{1}{4M}HW - area_{mi}) \quad (6)$$

where relu is rectified linear units [5], $loss_{area}$ will impose punishment if the area of pseudo segmentation is smaller than $\frac{1}{4M}HW$.

3.2 Prior-Guided Anatomical Structure Clustering

Commonly, specialists can easily identify anatomical structures based on corresponding prior knowledge, including relative location, topological structure, and even based on knowledge of similar structures. Therefore, DNN's ability of recognizing structures based on local discrimination and topological priors is studied in this part. Reference images, which are binary masks of similar structures, real data or simulation and show knowledge of location and topological structure, is introduced to the network to force the clustering branch to obtain corresponding structures as shown in Fig. 3.

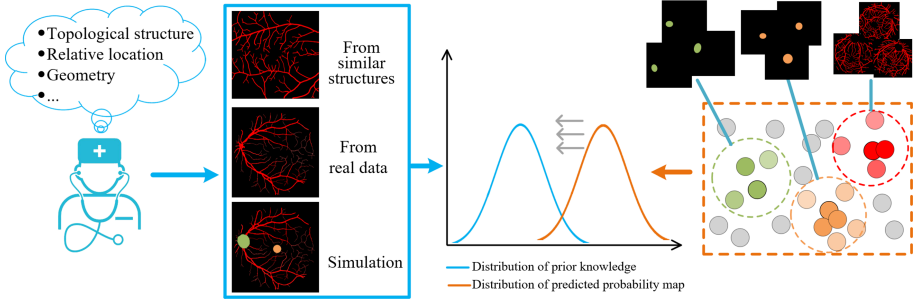


Fig. 3. Based on prior knowledge and local discrimination to recognize structures.

We denote the distribution of m -th cluster as P_m and the distribution of corresponding references as Q_m . The goal of optimization is to minimize Kullback-Leibler (KL) divergence between them, and it can be formulated as followed:

$$\min_{f_\theta} KL(P_m || Q_m) = \sum_i P_m(r_{mi}) \log \frac{P_m(r_{mi})}{Q_m(r_{mi})} \quad (7)$$

To minimize the KL divergence between P_m and Q_m , adversarial learning [6] is utilized to encourage the produced pseudo segmentation to be similar as the reference mask. During training, a discriminator D is set to discriminate pseudo segmentation r_m and reference mask s_m , while f_θ aims to cheat D . The loss function for D and adversarial loss for f_θ are defined as followed:

$$loss_D = loss_{bce}(D(s_m), 1) + loss_{bce}(D(r_m), 0) \quad (8)$$

$$loss_{bce}(\hat{y}, y) = -\frac{1}{N} \sum_i (y_i \log \hat{y}_i + (1 - y_i) \log (1 - \hat{y}_i)) \quad (9)$$

$$loss_{adv} = loss_{bce}(D(r_m), 1) \quad (10)$$

Reference Masks: (1) From similar structures: Similar structures share similar geometry and topology. Therefore, we can utilize segmentation annotations from similar structures to guide the segmentation of target, e.g., annotations of vessel in OCTA can be utilized for the clustering of retinal vessel in fundus images. (2) From real data: Corresponding annotations of target structure can be directly set as the prior knowledge. (3) Simulation: Based on the comprehension, experts can draw the pseudo masks to show the information of relative location, topology, etc. For example, based on retinal vessel mask, the approximate location of OD and fovea can be identified. Then, ellipses can be placed at these positions to represent OD and fovea based on their geometry priors.

4 Experiments and Discussion

The experiments can be divided into two parts to show the effectiveness of our proposed unsupervised local discrimination learning and prove the feasibility of combining local discrimination and topological priors to cluster target structures.

4.1 Network Architectures and Initialization

The backbone is a U-net consisted with a VGG-liked encoder and a decoder. The encoder is a tiny version of VGG-16 without fully connection layers (FCs), whose channel number is quarter of VGG-16. The decoder is composed with 4 convolution blocks, each of which is made up of two convolution layers. The final features of decoder will be concatenated with the features generated by the first convolution block of encoder for the further processing of the clustering branch and the embedding branch. Embedding branch is formed of 2 convolution layers with 32 channels and a l_2 normalization layer to project each pixel onto a 32-D hypersphere. Clustering branch is consisted with 2 convolution layers with 8 channels followed by a l_1 normalization layer.

To minimize the KL divergence between the pseudo segmentation distribution and the references distribution, a discriminator is created. The discriminator is a simple classifier with 7 convolution layers and 2 FCs. The channel numbers of convolution layers are 16, 32, 32, 32, 32, 64, 64 and the first 5 layers are followed by a max-pooling layer to halve the image size. FCs' channels are 32 and 1, and the final FC is followed by a Sigmoid layer.

Patch Discrimination to Initialize the Network: It is hard to simultaneously train the clustering branch and the embedding branch from scratch. Thus, we firstly jointly pre-train the backbone and the embedding branch by patch discrimination, which is an improvement of instance discrimination [18]. The main idea is that the embedding branch should project similar patches (patches under various augmentations) onto close positions on the hypersphere. The embedding features v_i will be firstly processed by an adaptive average pooling layer (APP) to generate spatial features, each of which represents feature of corresponding patches of image x_i . We denote $s_i(j)$ as the embedding vector for $x_i(j)$ (j -th patch of x_i), where $\|s_i(j)\|_2 = 1$ by applying a l_2 normalization. $\hat{s}_i(j)$ denotes the embedding vector of corresponding augmentation patch $\hat{x}_i(j)$. The probability of region $\hat{x}_i(j)$ being recognized as region $x_i(j)$ can be defined as followed:

$$P(ij|\hat{x}_i(j)) = \frac{\exp(s_i^T(j)\hat{s}_i(j)/\tau)}{\sum_{k,l} \exp(s_k^T(l)\hat{s}_i(j)/\tau)}, \quad (11)$$

Assuming all patches being recognized as $x_i(j)$ is independent, then the joint probability of $\hat{x}_i(j)$ being recognized as $x_i(j)$ and $x_k(l)$ ($k \neq i$ or $l \neq j$) not being recognized as $x_i(j)$ is as followed:

$$P_{ij} = P(ij|\hat{x}_i(j)) \prod_{k \neq i \text{ or } l \neq j} (1 - P(ij|\hat{x}_k(l))). \quad (12)$$

The negative log likelihood and loss function are formulated as followed:

$$J_{ij} = -\log P(ij|\hat{x}_i(j)) - \sum_{k \neq i \text{ or } l \neq j} \log(1 - P(ij|\hat{x}_k(l))). \quad (13)$$

$$loss_{pd} = \sum_{i,j} J_{ij} \quad (14)$$

We also introduce mixup [19] to make the representations more robust. Based on mixup, virtual sample $\tilde{x}_i = \lambda x_a + (1 - \lambda)x_b$ is firstly generated by linear interpolation of x_a and x_b , where $\lambda \in (0, 1)$. The embedded representation for patch $\tilde{x}_i(j)$ is $\tilde{s}_i(j)$, and we expect it is similar to the mixup feature $z_i(j)$. The loss is defined as followed:

$$z_i(j) = \frac{\lambda s_a(j) + (1 - \lambda)s_b(j)}{\|\lambda s_a(j) + (1 - \lambda)s_b(j)\|_2}. \quad (15)$$

$$\tilde{P}(ij|\tilde{s}_i(j)) = \frac{\exp(z_i^T(j)\tilde{s}_i(j)/\tau)}{\sum_{k,l} \exp(z_k^T(l)\tilde{s}_i(j)/\tau)}. \quad (16)$$

$$\tilde{J}_{ij} = -\log \tilde{P}(ij|\tilde{s}_i(j)) - \sum_{k \neq i \text{ or } l \neq j} \log(1 - \tilde{P}(ij|\tilde{s}_k(l))), \quad (17)$$

$$loss_{mixup} = \sum_{i,j} \tilde{J}_{ij} \quad (18)$$

When pre-training this model, we set the training loss as $loss_{pd} + loss_{mixup}$. The output size of APP is set as 4×4 to split each image into 16 patches. And each batch contains 16 groups of images and 8 corresponding mixup images, and each of group contains 2 augmentations of one image. The augmentation methods contain *RandomResizedCrop*, *RandomGrayscale*, *ColorJitter*, *RandomHorizontalFlip*, *Rotation90* in pytorch. The optimizer is Adam with initial learning rate (lr) of 0.001, which will be half if the validation loss does not decrease over 3 epochs. The maximum training epoch is 20.

4.2 Experiments for Learning Local Discrimination

Datasets and Preprocessing: Our method is evaluated in two medical scenes.

Fundus images: The model will be firstly trained on diabetic retinopathy (DR) detection dataset of kaggle [3]¹ (30k for training, 5k for validation). Then, the pre-trained encoder is transferred to 8 segmentation tasks: (1) Retinal vessel: DRIVE [13] (20 for training, 20 for testing), STARE [8] (10 for training, 10 for testing) and CHASEDB1 [10] (20 training, 8 testing). (2) OD and cup: Drishti-GS [12] (50 for training, 50 for testing). ID (OD) [11] (54 for training, 27 for testing). (3) Lesions: Haemorrhages dataset (Hae) and hard exudates dataset (HE) from IDRID [11]. **Chest X-ray:** The encoder is pre-trained on ChestX-ray8 [15] (100k for training and 12k for validation) and transferred to lung segmentation [2] (69 for training, 69 for testing). All images of above datasets are resized to 512×512 .

¹ <https://www.kaggle.com/c/diabetic-retinopathy-detection/data>.

Table 1. Comparison of results of downstream segmentation tasks.

Encoder	Retinal vessel			Optic disc and cup			Lesions		X-ray
	DRIVE	STARE	CHASE	GS(cup)	GS(OD)	ID(OD)	Hae	HE	Lung
Random	80.76	76.26	78.30	77.76	95.41	89.11	37.76	57.44	96.34
Supervised	81.06	80.59	78.56	86.94	96.40	93.56	51.11	61.34	—
Wu et al.	74.98	66.15	68.31	84.59	94.58	88.70	26.34	48.67	96.27
Ye et al.	80.87	81.22	79.85	87.30	97.40	94.68	46.79	59.40	96.63
LD	82.15	83.42	80.35	89.30	96.53	95.59	46.72	65.77	97.51

Implementation Details:

- (1) Local discriminative representation learning: The model is firstly initialized by pre-trained model of patch discrimination. Then the training loss is set as $loss_{pd} + loss_{mixup} + 10loss_{id} + loss_{entropy} + 5loss_{area}$. Each batch has 6 groups of images, each of which contains 2 augmentations of one image, and 3 mixup images. The maximum training epoch is 80 and the optimizer is Adam with $lr = 0.001$.
- (2) Transferring: The encoder of downstream tasks is initialized by the learnt feature extractor of local discrimination. The decoder is composed with 5 convolution blocks, each of which contains 2 convolution layers and is followed by a up-pooling layer. The loss is set as $loss_{dsc} = \frac{2|p \times g|}{|g| + |p|}$. This model will be firstly trained in 100 epochs in frozen pattern with Adam with $lr = 0.001$, and then be trained in fine-tune pattern with $lr = 0.0001$ in the following 100 epochs.
- (3) Comparative methods: **Random:** The network is trained from scratch. **Supervised:** Supervised by the manual score of DR, the encoder will be firstly trained by making classification. **Wu et al. [16]** and **Ye et al. [18]:** Instance discrimination methods proposed in [16] and [18]. **LD:** The proposed method.

Results: The evaluation metric is mean Dice-Sørensen coefficient (DSC): $DSC = \frac{2|P \times G|}{|P| + |G|}$, where P is the binary results of predictions and G is the ground truth. Quantitative evaluations for downstream tasks are shown in Table 1, and we can have following observations:

- 1) The generalization ability of the trained local discriminative representation is demonstrated by the leading performance in the 6 fundus tasks and lung segmentation. Compared with models trained from scratch, models initialized by our pre-trained encoder can respectively gain improvements of 1.39%, 7.16%, 2.05%, 11.54%, 1.12%, 6.48%, 8.96%, 8.33% and 1.17% in DSC for all 9 tasks.
- 2) Compared with instance discrimination methods by Wu et al. [16] and Ye et al. [18], the proposed local discrimination is capable to learn finer features

and is more suitable for unsupervised representation learning of medical images.

- 3) The proposed unsupervised method is free from labeled images and the learnt representation is more generalized, while supervised representation learning relies on expensive manual annotations and learns specialized representations. As shown in Table 1, our method shows better performance than supervised representation learning, whose target is to classification DR, and the only exception is on segmenting haemorrhages which is the key evidence for DR.

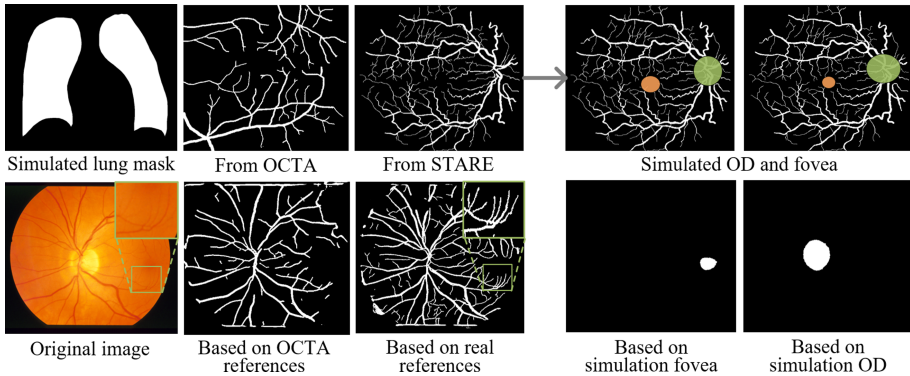


Fig. 4. Some examples of reference masks and predicted results: The first row show some reference images and the second row show the predictions.

4.3 Experiments for Clustering Structures Based on Prior Knowledge

Implementation Details: In this part, we respectively fuse reference images from real data, similar structures and simulations into local discrimination to investigate the ability of clustering anatomical structures. A dataset with 3110 high-quality fundus images from [14] and 1482 frontal X-rays from [15] are utilized as the training data. The reference images can be constructed in 3 ways: (1) From real references: ALL 40 retinal vessel masks of DRIVE are utilized as the references for clustering pixels of vessel. (2) From similar structures: Similar structures share similar priors, thus, 10 OCTA vessel masks are utilized as the references for retinal vessel of fundus. (3) Simulation: We directly draw 20 simulated lung masks to guide lung segmentation. Meanwhile, based on vessel masks of DRIVE, we place ellipses at approximate center location of OD and fovea to generate pseudo masks. Some reference masks are shown in Fig. 4.

f_θ needs to jointly learn local discrimination and cheat D , thus, it will be updated by minimizing the following loss:

$$loss_{f_\theta} = loss_{pd} + loss_{mixup} + 10loss_{ld} + loss_{entropy} + 5loss_{area} + 2loss_{adv} \quad (19)$$

The optimizer for f_θ is Adam with $lr = 0.001$. The discriminator is optimized by minimizing $loss_D$ and the optimizer is Adam with $lr = 0.0005$. It is worth noting that during the clustering training of OD and fovea, all masks of real vessel, fovea and OD are concatenated and fed into D to provide enough information and f_θ is firstly pre-trained to cluster retinal vessel. The maximum training epoch is 80.

Results: Visualization examples are shown in Fig. 4. Quantitative evaluations are as followed: (1) Retinal vessel segmentation is evaluated in the test data of STARE. And the DSC are respectively 66.25% and 57.35% for models based on real references and based on OCTA annotations. (2) The segmentation of OD is evaluated in the test data of Drishti-GS and gains DSC of 83.60%. (3) The segmentation of fovea is evaluated in the test data of STARE. Because the region of fovea is fuzzy, we measure the mean distance between the real center of fovea and the predicted center. The mean distance is 7.63pixels. (4) The segmentation of lung is evaluated in NLM [2] and the DSC is 81.20%.

Based on above results, we can have following observations:

- 1) In general, topological priors generated from simulation or similar structures in a different modality is effective to guide the clustering of target regions.
- 2) However, real masks contain more detailed information and are able to provide more precise guidance. For example, compared with vessel segmentations based on OCTA annotations, which missing the thin blood vessels due to the great thickness of OCTA mask, segmentations based on real masks can recognize thin vessels due to the details provided and the constraint of clustering pixels with similar context.
- 3) For anatomical structures with fuzzy intensity pattern, such as fovea, combining local similarity and structure priors is able to guide precise recognition.

5 Conclusion

In this paper, we propose an unsupervised framework to learn local discriminative representation for medical images. By transferring the learnt feature extractor, downstream tasks can be improved to decrease the demand for expensive annotations. Furthermore, similar structures can be clustered by fusing prior knowledge into the learning framework. The experimental results show that our methods have best performance on 7 out of 9 tasks in fundus and chest X-ray images, demonstrating the great generalization of the learnt representation. Meanwhile, the feasibility of clustering structures based on prior knowledge and unlabeled images is demonstrated by combining local discrimination and topological priors from real data, similar structures or even simulations to segment anatomical structures including retinal vessel, OD, fovea and lung.

References

1. Bachman, P., Hjelm, R.D., Buchwalter, W.: Learning representations by maximizing mutual information across views. In: *Advances in Neural Information Processing Systems*, pp. 15535–15545 (2019)
2. Candemir, S., et al.: Lung segmentation in chest radiographs using anatomical atlases with nonrigid registration. *IEEE Trans. Med. Imaging* **33**(2), 577–590 (2013)
3. Cuadros, J., Bresnick, G.: Eyepacs: an adaptable telemedicine system for diabetic retinopathy screening. *J. Diabetes Sci. Technol.* **3**(3), 509–516 (2009)
4. Dosovitskiy, A., Fischer, P., Springenberg, J.T., Riedmiller, M., Brox, T.: Discriminative unsupervised feature learning with exemplar convolutional neural networks. *IEEE Trans. Pattern Anal. Mach. Intell.* **38**(9), 1734–1747 (2015)
5. Glorot, X., Bordes, A., Bengio, Y.: Deep sparse rectifier neural networks. In: *Proceedings of the Fourteenth International Conference on Artificial Intelligence and Statistics*. pp. 315–323 (2011)
6. Goodfellow, I., et al.: Generative adversarial nets. In: *Advances in Neural Information Processing Systems*, pp. 2672–2680 (2014)
7. He, K., Fan, H., Wu, Y., Xie, S., Girshick, R.: Momentum contrast for unsupervised visual representation learning. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 9729–9738 (2020)
8. Hoover, A., Kouznetsova, V., Goldbaum, M.: Locating blood vessels in retinal images by piecewise threshold probing of a matched filter response. *IEEE Trans. Med. Imaging* **19**(3), 203–210 (2000)
9. Mahmood, U., et al.: Whole MILC: generalizing learned dynamics across tasks, datasets, and populations. In: Martel, A.L., et al. (eds.) *MICCAI 2020*. LNCS, vol. 12267, pp. 407–417. Springer, Cham (2020). https://doi.org/10.1007/978-3-030-59728-3_40
10. Owen, C.G., et al.: Measuring retinal vessel tortuosity in 10-year-old children: validation of the computer-assisted image analysis of the retina (caiar) program. *Invest. Ophthalmol. Vis. Sci.* **50**(5), 2004–2010 (2009)
11. Porwal, P., et al.: Indian diabetic retinopathy image dataset (idrid): a database for diabetic retinopathy screening research. *Data* **3**(3), 25 (2018)
12. Sivaswamy, J., Krishnadas, S., Joshi, G.D., Jain, M., Tabish, A.U.S.: Drishti-gs: Retinal image dataset for optic nerve head (onh) segmentation. In: *2014 IEEE 11th International Symposium on Biomedical Imaging (ISBI)*, pp. 53–56. IEEE (2014)
13. Staal, J., Abràmoff, M.D., Niemeijer, M., Viergever, M.A., Van Ginneken, B.: Ridge-based vessel segmentation in color images of the retina. *IEEE Trans. Med. Imaging* **23**(4), 501–509 (2004)
14. Wang, R., Chen, B., Meng, D., Wang, L.: Weakly-supervised lesion detection from fundus images. *IEEE Trans. Med. Imaging* **38**, 1501–1512 (2018)
15. Wang, X., Peng, Y., Lu, L., Lu, Z., Bagheri, M., Summers, R.M.: Chestx-ray8: Hospital-scale chest x-ray database and benchmarks on weakly-supervised classification and localization of common thorax diseases. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 2097–2106 (2017)
16. Wu, Z., Xiong, Y., Yu, X.S., Lin, D.: Unsupervised feature learning via non-parametric instance discrimination. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 3733–3742 (2018)

17. Xie, X., Chen, J., Li, Y., Shen, L., Ma, K., Zheng, Y.: Instance-aware self-supervised learning for nuclei segmentation. In: Martel, A.L., et al. (eds.) MICCAI 2020. LNCS, vol. 12265, pp. 341–350. Springer, Cham (2020). https://doi.org/10.1007/978-3-030-59722-1_33
18. Ye, M., Zhang, X., Yuen, P.C., Chang, S.F.: Unsupervised embedding learning via invariant and spreading instance feature. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 6210–6219 (2019)
19. Zhang, H., Cisse, M., Dauphin, Y.N., Lopez-Paz, D.: Mixup: beyond empirical risk minimization. In: International Conference on Learning Representations (2018)