# Combating Ambiguity for Hash-code Learning in Medical Instance Retrieval

Jiansheng Fang, Huazhu Fu, Dan Zeng, Xiao Yan, Yuguang Yan, and Jiang Liu

***Abstract*— When encountering a dubious diagnostic case, medical instance retrieval can help radiologists make evidence-based diagnoses by finding images containing instances similar to a query case from a large image database. The similarity between the query case and retrieved similar cases is determined by visual features extracted from pathologically abnormal regions. However, the manifestation of these regions often lacks specificity, i.e., different diseases can have the same manifestation, and different manifestations may occur at different stages of the same disease. To combat the manifestation ambiguity in medical instance retrieval, we propose a novel deep framework called Y-Net, encoding images into compact hash-codes generated from convolutional features by feature aggregation. Y-Net can learn highly discriminative convolutional features by unifying the pixel-wise segmentation loss and classification loss. The segmentation loss allows exploring subtle spatial differences for good spatial-discriminability while the classification loss utilizes class-aware semantic information for good semantic-separability. As a result, Y-Net can enhance the visual features in pathologically abnormal regions and suppress the disturbing of the background during model training, which could effectively embed discriminative features into the hash-codes in the retrieval stage. Extensive experiments on two medical image datasets demonstrate that Y-Net can alleviate the ambiguity of pathologically abnormal regions and its retrieval performance outperforms the state-of-the-art method by an average of $9.27\%$ on the returned list of $10$.**

***Index Terms*— Medical Instance Retrieval, Convolutional Features, Deep Hashing Methods, Content-based Image Retrieval**

## I. Introduction

Content-based image retrieval (CBIR) has been mostly tackled as the problem of instance-level image retrieval [1] and has been a long-standing research topic in the computer vision society [2]. When encountering a dubious diagnostic case, CBIR systems can help radiologists search for similar cases in their decision-making process. Instance-level image retrieval is to hunt for images with the same instance as a query image in a large image database [3]. The benefit of instance-level retrieval for medical image screening and diagnosing can be witnessed in an observer study [4]. Five participating radiologists were given the task of querying nodules, for which they were required to infer the likelihood of malignancy. The task was performed twice: once with the aid of the search engine and once not. The search engine returned 3 instances of the most similar malignant images and 3 instances of the most similar benign images to help these radiologists making the inference. The average performance of the five radiologists was shown to increase from $0.56$ to $0.63$ with the aid of similar nodules.

The relevancy of instance-level retrieval is mainly grounded on the visual similarity of instances rather than the whole image [5], so the features of a region-wise instance residing in a retrieved image should be explored effectively. Recently, many existing works on instance-level retrieval typically extracted visual features by using convolutional neural networks (CNN) to prevent the visual features unique to an instance from drowning in the global image. Early works [6], [7] focused on replacing traditional hand-crafted descriptors with features from fully-connected layers. The second generation of works [8], [9] achieved significant gains by encoding the activations of convolutional layers as region-wise feature descriptors. Among CNN-based approaches of instance-level retrieval, deep hashing methods [10], [11] have arisen as a promising solution because of their efficient data storage and fast searching.

Deep hashing methods can preserve the information of high-dimensional images by jointly learning image descriptors and hash-codes in an end-to-end framework [12]–[14]. The image descriptors from fully-connected layers or convolutional layers are mapped into compact hash-codes for similarity comparison. Existing deep hash methods for instance-level retrieval have been shown to be effective and efficient [15], [16]. However, generating hash-codes in medical instance retrieval is challenging due to the manifestation ambiguity of pathologically abnormal regions. Such an issue plagues radiologists in the clinically routine screening and largely affects medical instance retrieval performance. It can be varied in two kinds: different diseases can have the same pathological abnormalities (**SPDD**), while different pathological manifestations may occur at different stages of the same disease (**DPSD**). As Fig. 1 shows, 1) **SPDD** problem: it is difficult to interpret chest X-ray images and recognize the subtle difference between malignant and benign nodules, the lesion region of both images is on the left lung's upper lobe and has similar manifestations. However, the malignant image is diagnosed as lung cancer, and the benign image is pulmonary hematoma. Only professional radiologists can find the difference between benign and malignant nodules. 2) **DPSD** problem: cup to disk ratio (CDR), which is the ratio of cup diameter to disc diameter and often be employed as the main clue of glaucoma diagnose, varies at different stages.

The ambiguity of pathologically abnormal regions may prevent the assimilation of medical instance retrieval into an assistant tool for medico-decision [17]. One solution is to provide ground-truth fine-grained labels to combat the ambiguity of pathologically abnormal regions. However, medical annotations remain highly dependant on manual expert feedback with high inter-observer variability [18]. Generally, medical image datasets can provide class labels for classification and pixel-wise masks for segmentation. Hence, a feasible solution is to effectively exploit the visual contents of pathologically abnormal regions based on class labels and pixel-wise masks [19]. Following this way, we present a novel deep framework, called Y-

This work was supported in part by the Science and Technology Innovation Committee of Shenzhen City (20200925174052004 and JCYJ20200109140820699).

J.Fang is with the School of Computer Science and Technology, Harbin Institute of Technology, Harbin, China, and also with the Department of Computer Science and Engineering, Southern University of Science and Technology, Shenzhen, China, and also with CVTE Research, Guangzhou, China (e-mail: 11949039@mail.sustech.edu.cn).

H.Fu is with the Inception Institute of Artificial Intelligence, Abu Dhabi, United Arab Emirates (e-mail: hzfu@ieee.org).

D.Zeng and X.Yan are with the Department of Computer Science and Engineering, Southern University of Science and Technology, Shenzhen, China.

Y.Yan is with the Department of Mathematics, University of Hong Kong, Hong Kong, China.

J.Liu is with the Department of Computer Science and Engineering, Southern University of Science and Technology, Shenzhen, China, and also with the Cixi Institute of Biomedical Engineering, Chinese Academy of Sciences, Ningbo, China (e-mail: liuj@sustech.edu.cn).

Corresponding author: Jiang Liu.

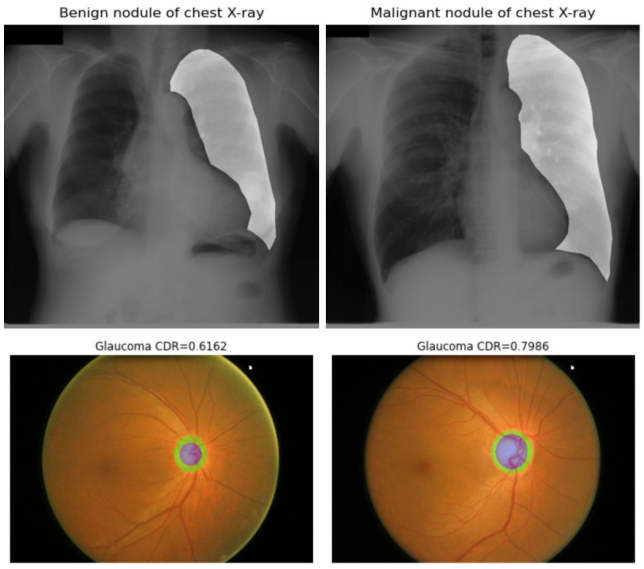


Fig. 1. Illustration of ambiguity in medical image diagnosis. The upper row is two chest X-ray images labeled benign and malignant nodule. The down row is two glaucoma fundus images with a cup to disk ratio (CDR).

Net, to learn deep representations from image spaces by unifying segmentation and classification losses. During the training stage, the spatially subtle differences and class-aware semantic information of pathological regions are simultaneously learned into convolutional features. In the test stage, the learned convolutional features are aggregated into the hash-codes to preserve visual features unique to pathologically abnormal regions.

There are two main motivations to present the Y-Net framework for alleviating the specificity shortage in medical instance retrieval. First, traditional deep hashing networks are to learn the global descriptor in an end-to-end way. They are prone to make the discriminative regions drown in the global descriptor. On the contrary, our Y-Net aims to explore the pixel-wise discriminative information by segmentation guidance, which pays more attention on the pathologically abnormal regions. Second, existing instance retrieval methods using local aggregation usually locate local regions in an unsupervised or weakly-supervised manner, which ignores the label information, while our Y-Net exploits class labels to locate the discriminative regions. The main contributions of this work are summarized as follows:

1) To combat ambiguity of pathologically abnormal regions in medical instance retrieval, we present a novel Y shape deep network, named Y-Net, encoding images into compact hash-codes. Our Y-Net can improve the differentiating ability of the hash-codes by exploiting the visual features unique to pathologically abnormal regions.
2) Y-Net unifies classification and pixel-wise segmentation training to learn good semantic-separability and spatial-discriminability convolutional features. The segmentation branch learns subtle spatial differences to avoid the **SPDD** problem while the classification branch locates the discriminative regions by class-aware semantic information to overcome the **DPSD** problem.
3) Extensive experiments on two public medical datasets demonstrate that our proposed Y-Net can further improve the retrieval performance compared to the state-of-the-art instance retrieval methods. Our code and model have been released in `https://github.com/fjssharpsword/YNet`.

The rest of this work is organized as follows: Section II discusses related works. Section III describes our methodology in detail. Section IV extensively evaluates the proposed method on two medical images datasets. Section V gives concluding remarks.

## II. Related Works

This section gives some related works that have contributed to instance-level retrieval and discuss current research progress of medical instance retrieval.

### A. Instance-level Retrieval

Hashing methods can be divided into data-independent methods and data-dependent methods. The data-independent methods [20], [21] learn hashing functions in a two-stage manner from hand-crafted features such, and the hash-codes learning procedure is independent of the image features, which may lead to sub-optimal performance. The data-dependent methods, also called learning-based hashing methods, can be further categorized into [22]: (1) shallow learning-based hashing methods, like metric hashing forests [23], and kernel sensitive hashing [24]; (2) deep learning-based hashing methods, like image inpainting-based compact hash-code learning [25], and deep hashing network [26]. In contrast to the data-independent methods, they extract global features for hashing in an end-to-end manner. Early works [27], [28] for instance-level retrieval rely on hand-crafted local descriptors such as SIFT [29] and SURF [30]. Prior to deep learning, these works based on local features extraction, then aggregated into a global vector [31], [32]. The instances relevant to a query are discovered in the candidate images for similarity search by matching local descriptors. However, hand-crafted local features are vulnerable to non-rigid deformations and heavy viewpoint changes. Due to the promising performance in computer vision, CNN-based approaches have been introduced to instance-level retrieval. Instead, the global vector is extracted by a single forward-pass through a CNN, in which the extraction and aggregation steps are not separated. Existing deep hashing methods [33], [34] can be grouped into this category using feature embedding tailor features from fully-connected layers for hash-codes generating. The representative methods include deep pairwise-supervised hashing (DPSH) [7], deep supervised hashing (DSH) [35], and deep residual hashing (DRH) [36]. Since convolutional features have been found to be reasonably discriminative [37], recent CNN-based approaches have shifted to concentrate on feature aggregating rather than feature embedding. CNN-based approaches aggregating convolutional feature maps as global image representation can be roughly divided into two categories.

The first category is the works encoding the activations of a convolutional layer by weighted aggregation. These works' key idea is to assign different weights to different regions' activations in feature maps after global convolutional layers generate. SPoC [8] showed that a simple spatial pooling on the convolutional layer outperformed fully-connected layers, and the power of this representation could be enhanced by applying the Gaussian center prior scheme to weight the contribution of the activations before aggregation. Following a similar idea, CroW [38] proposed a non-parametric spatial- and channel-wise weighting method for focusing on salient regions. Unlike the spatial weighting scheme, class activation maps (CAMs) [39] are employed for calculating semantic-aware weights of a convolutional feature map. Based on the bags of local convolutional features (BLCF) [40], BLCF-SALGAN [41] build an efficient image representation by saliency weighting.

The second category is the works performing region analysis using convolutional features. Unlike the first category, this category first generates regions' convolutional features after region proposal, then

aggregates them into global features. The representative work is regional maximum activation of convolutions (R-MAC) [42], which generated a set of regional vectors by performing spatial max-pooling within a particular region and aggregates features from several local regions into a single compact feature. Gordo *et al.* [43] improved over the original R-MAC encoding by explicitly learning a region proposal network [44] and training in an end-to-end framework with a triplet loss. Laskar *et al.* [45] used a saliency measure directly derived from the convolutional features to weigh the contribution of the regions of R-MAC before aggregation. Similar to R-MAC, Cao *et al.* [46] proposed a method to derive a set of base regions directly from the convolutional layer, followed by a query adaptive re-ranking strategy. DeepVision [47] extracted region-level features from the bounding boxes generated by the object detection framework. Regional attention [37] proposed a context-aware regional attention network that weighs an attentive score of a region considering global attentiveness.

### B. Medical Instance Retrieval

Recently, deep hashing methods using feature embedding on the fully-connected layer have also been widely proposed for medical instance retrieval, such as deep multiple instances hashing for tumor assessment [11], deep residual hashing for chest X-ray images [36], order-sensitive deep hashing method for multi-morbidity medical image retrieval [22], deep disentangled momentum hashing for Neuroimage Search [16], etc. Although the prior works have facilitated medical instance retrieval's prosperity, pathologically abnormal regions' manifestation ambiguity is challenging for current deep hashing methods. Recent studies [37], [39], [46] have shown that using feature aggregating on the convolutional features achieves promising performance in instance-level retrieval. Following this direction, our work improves the current deep hashing method to combat pathologically abnormal regions' ambiguity in medical instance retrieval.

In this work, the improvement for the current deep hashing methods includes:

- Unlike the current deep hashing methods jointly learning image descriptors and hash-codes, our work first learns convolutional features from image spaces by supervised training, then aggregates them as hash-codes. The learned convolutional features and following generated hash-codes can effectively preserve the differentiating information of pathologically abnormal regions.
- Inspired to CAMs [39] and R-MAC [42], we endow the class-aware information to the R-MAC descriptors by classification training. The R-MAC descriptors related to classes can enhance their differentiating ability and help to avoid the **SPDD** problem.
- Motivated by regional attention [37], we adopt feature pyramid networks (FPN) [48] to exploit multi-scale pathologically abnormal regions by pixel-wise segmentation training. The subtle differences are encoded into the convolutional features to overcome the **DPSD** problem.

We detect if pathologically abnormal regions are presented in each image with classification training, and we locate pathologically abnormal regions using activations with the help of segmentation training. In the end, the convolutional features, having learned class-aware information and subtle spatial differences, are mapped into the hash-codes. Based on the class labels and pixel-wise masks, we argue that our work is a beneficial exploration to combat pathologically abnormal regions' ambiguity in medical instance retrieval.

## III. Methodology

Our Y-Net aims to generate highly distinctive hash-codes from the learned convolutional features. The hash-codes should meet three requirements: $(a)$ the query image should be encoded close to positive images with the same instance and far from negative images without the same instance in the hashing space; $(b)$ the class-aware semantic information and subtle differences of pathologically abnormal regions should be effectively encoded in convolutional features; $(c)$ The convolutional features should be effectively aggregated to the compact hash-codes to preserve the learned visual cues. This section will elaborate on our Y-Net, including the main branch, R-MAC branch, FPN branch, and the coupled loss function.

### A. Framework Overview

Each image is represented by an instance-invariant feature vector, i.e., hash-code. As shown in Fig.2, we present a deep learning framework, called Y-net, to generate distinctive hash-codes from convolutional features. Our Y-Net contains three parts, main branch, R-MAC branch (right), and FPN branch (left). In the training stage (double arrow), we input an image into the main branch and feed-forward to the core node (red rectangle). The core node is a convolutional layer, followed by the R-MAC branch and the FPN branch. In the R-MAC branch, the classification loss minimizes intra-class distance and maximize inter-class distance. The inter-class separation can help avoid **SPDD** problem. But, to overcome the **DPSD** problem, the intra-class distance needs to be preserved but not minimized. The FPN branch can locate intra-class differences by pixel-wise segmentation training to balance the reduction of intra-class distance in the R-MAC branch. The core node learns the class-aware semantic information of pathological regions from the R-MAC branch for differentiating the same manifestation of different diseases. Simultaneously, the spatially subtle differences of pathological regions from the FPN branch are encoded into the core node to locate the same disease's subtle differences at different stages. After the core node absorbing the visual cues from the R-MAC branch and the FPN branch in the training stage, we can generate hash-codes from the learned core node by feature aggregation in the test stage (single arrow).

### B. Main Branch

The image encoding pipeline of the main branch is depicted as follows:

- **Training Stage.** The input image $\boldsymbol{I}$ with a resolution $3 \times 256 \times 256$ is feed-forwarded into the main branch. The main branch computes a feature hierarchy consisting of a bottom-up block at three scales with a scaling step of 2. At each scale, we use the feature activations output of bottom-up block [49] to get a receptive field. The three bottom-up blocks are merged into the FPN branch by addition. In the core node of the main branch, the convolutional feature maps $\boldsymbol{X} \in \mathbb{R}^{C \times H \times W}$ can be arranged in a tensor of the size $C \times H \times W$, where $H$ and $W$ denote the height and width of each feature map, and $C$ denotes the number of feature maps (or channels) in the convolutional layer. In a convolutional layer, the activations at the same spatial location across all feature maps can be composed into a $C$-dimensional local descriptor for a certain image region. Compared to the activations of the fully-connected layer, the convolutional features retain the spatial information of local image descriptors and are essentially similar to the traditional hand-crafted local features [50], [51]. The convolutional feature maps $\boldsymbol{X}$ are further feed-forwarded into the R-MAC branch and the FPN branch. Based on the classification and segmentation training, the semantic and spatial information of pathological regions is encoded into the convolutional feature maps in the feedback process.

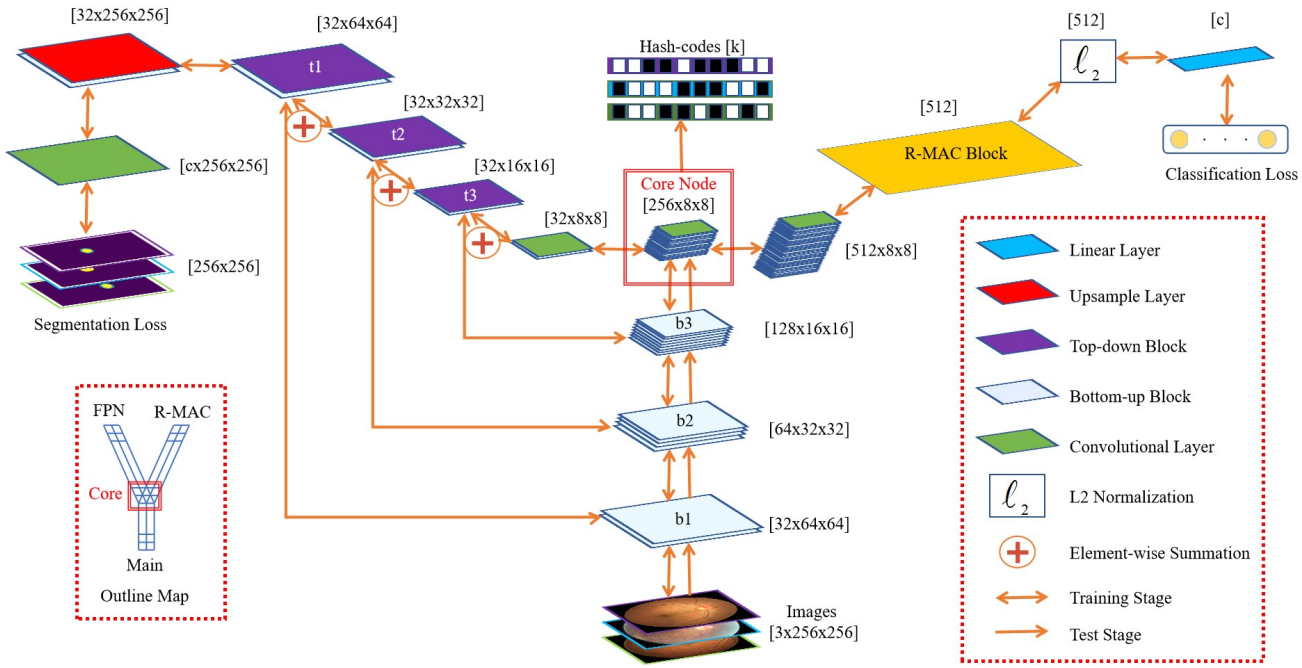


Fig. 2. A novel deep framework for medical instance retrieval. We feed an image to the main branch, followed by the R-MAC branch (right) for classification training and the FPN branch (left) for segmentation training. The class-aware semantic information of pathological regions from the R-MAC branch and the spatially subtle differences of pathological regions from the FPN branch is effectively learned in the convolutional feature maps in the core node (red rectangle). The convolutional feature maps are mapped into the hash-codes via feature aggregating in the test stage. Our framework's shape is similar to a Y shape, including the main branch, the R-MAC branch, and the FPN branch, called Y-Net.

- **Test Stage.** Based on the pre-trained Y-Net, an image with a resolution $3 \times 236 \times 256$ is feed-forwarded into the main branch and terminated in the core node. The core node is the conjunct point of a Y shape and is the core component in the framework of Y-Net. We apply feature aggregation to generate a $k$-bits hash-code from the learned convolutional feature maps $\boldsymbol{X}$ with $C \times H \times W$ in the core node. Based on the existing works, the feature aggregation does not take part in the training and has been found to be more capable of preserving the discriminative information than the feature embedding. Considering the convolutional feature maps $\boldsymbol{X}$ with $C \times H \times W$ have learned the visual cues of pathological regions effectively, we convolute the size of $C \times H \times W$ into the size of $c \times h \times w$ without any weighting strategy. Then the three dimensions vectors further are squeezed into one dimension; its size equals the hash-code size of $k$-bits. Such a convolution process can aggregate feature maps of various sizes in three dimensions, such as $1 \times 8 \times 8$ and $128 \times 1 \times 1$. Lastly, we apply the hyperbolic tangent function to generate the value between $-1$ and 1, following by signed as binary hash-code. At this step, we do not introduce any weighting strategy on feature aggregation because the convolutional feature maps have learned the visual cues of pathological regions effectively.

### C. R-MAC Branch

The R-Mac branch contains a convolutional layer using $3 \times 3$ filters and followed by batch normalization [52], then an R-MAC block generating a feature vector of length 512. The feature vector is mapped into a linear layer after the $L_2$ normalization. The length of the linear layer is the number of classes. The R-MAC block generates a compact representation by aggregating multiple regions at different scales. By classification training, the highly activated regions can correspondingly respond to the semantic information of the belonging class. The pipeline of the R-MAC block is summarized as follows:

- Based on a convolutional layer with $512 \times 8 \times 8$, we sample square regions with a region size, $R_s$, of a specific scale $s$ in a sliding window manner of 0.4 overlap between neighbor windows, for all $s = 0, ..., S$. The region size at a specific scale can be calculated as:

$$R_s = 2 \times \min(W_r, H_r)/(s+2), \tag{1}$$

  where $W_r$ and $H_r$ are width and height of the feature map in the convolutional layer. In our Y-Net, with $W_r = 8$ and $H_r = 8$, we set $S = 3$, then we totally get sample region of 14.
- After sampling the regional feature maps, we perform a max-pooling for all regional feature maps of 14. Each regional feature maps generate a feature vector with 512, the same as the channel's size. Last, we aggregate all feature vector of sample regions in the whole image as a global feature vector with 512 dimensions, named R-MAC descriptor used as a discriminative image representation.

In the pipeline of R-MAC, the local features from a certain convolutional layer are max-pooled across several multi-scale overlapping regions, obtained from a rigid grid covering the whole image, similar in spirit to spatial pyramids, producing a single feature vector per region. Then these region descriptors are sum-aggregated and $L_2-$ normalized into a global image representation. The discriminative global image representation is a compact vector whose size is independent of the size of the image and the number of regions. The region pooling is different from a spatial pyramid. The latter concatenates the region descriptors, while the former sum-aggregates them. Comparing the R-MAC descriptors of two images with a dot-product can then be interpreted as a many-to-many region matching.

R-MAC has been known for effective and efficient performance in image retrieval. Nonetheless, the main issue of R-MAC is that all sampled regions are equally treated without considering their

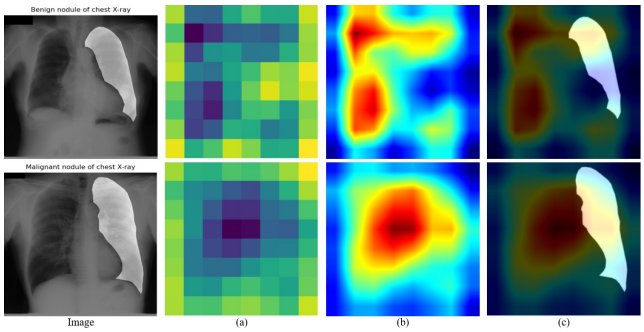


Fig. 3. Feature maps of the R-MAC branch. We calculate the mean value along the channel axis of the convolutional features with $\mathbf{512} \times \mathbf{8} \times \mathbf{8}$, then visualize the mean value: (a) the feature map with $\mathbf{8} \times \mathbf{8}$, (b) the color map resized to the size of the mask image with $\mathbf{256} \times \mathbf{256}$, and (c) the overlay map combined the color map with the mask image.

varying importance. When aggregating their regional feature vectors, all regions construct their equal attentiveness to the last R-MAC descriptor. To overcome this problem, we integrate R-MAC in our Y-Net for supervised training to avoid the class-agnostic problem of the descriptors in R-MAC. We argue that the convolutional layer activations before the R-MAC block can respond to the semantic information during classification training. The class-based semantic information is conveyed to the sample regions in the R-MAC block. Thus the different regions responding to the classification devote their varying contribution to generating the R-MAC descriptor. The learned R-MAC descriptor containing class-based semantic information can help address the **SPDD** problem by differentiating regions with a similar texture. Put an example of a chest X-ray, although two chest nodules with different sizes are the same texture, they are labeled benign and malignant and diagnosed with different diseases, respectively. As shown in Fig. 3, the two chest nodules that belonged to different classes vary differently in the learned feature maps of the convolutional layer. By training the R-MAC branch, the R-MAC descriptor can exploit the class-based semantic information of chest nodules and feedback to the convolutional layer in the R-MAC block, then the core node of the main branch.

### D. FPN Branch

Pixel-wise segmentation help extract features that emphasize the pathological abnormal regions. Beneficial from the segmentation training, the FPN branch explores the multi-scale subtle differences of pathological regions at different stages and then give feedback to the core node in the main branch. FPN leverages the convolutional features from low to high levels to extract multi-scale spatial information by building a pyramidal feature hierarchy. FPN has been a criteria component in the network of object detection and shows its powerful feature extraction capability to achieve higher accuracy [53], [54]. The multi-scale spatial information of pathological regions helps generate differentiating features in medical instance retrieval. In our Y-Net, we leverage the FPN components to extract multi-scale spatial information from medical images for semantic segmentation by setting the label of mask images semantically. The structure of the FPN branch is introduced as follows:

- Following the core node in the main branch, the FPN branch provides two convolutional layers and three top-down blocks. Last, the FPN branch generates a predicted mask image for pixel-wise segmentation training. The segmentation loss is feedback to the FPN branch and the main branch to help exploit the multi-scale subtle differences of pathological regions at different stages.
- In the feed-forwarding process, corresponding to the three bottom-up blocks $\{32 \times 64 \times 64$ as b1, $64 \times 32 \times 32$ as b2, $128 \times 16 \times 16$ as b3$\}$ in the main branch, three top-down blocks $\{32 \times 16 \times 16$ as t3, $32 \times 32 \times 32$ as t2, $32 \times 64 \times 64$ as t1$\}$ in the FPN branch merge them by element-wise addition. The outputs of two bottom-up blocks $\{$b2, b3$\}$ convolute into 32-dimensions channel. The output of the convolutional layer before block $\{$t3$\}$ and two top-down blocks $\{$t2, t3$\}$ are resized into twice times their width and height by bi-linear up-sampling. Then, the convolutional layer before top-down block $\{$t3$\}$ and the bottom-up block $\{$b3$\}$, the top-down block $\{$t3$\}$ and the bottom-up block $\{$b2$\}$, the top-down block $\{$t2$\}$ and the bottom-up block $\{$b1$\}$, these pairs with the same spatial size are merged by element-wise summation. The addition operation of these three pairs generate the top-down blocks $\{$t3, t2, t1$\}$ successively. Last, the top-down block $\{$t3$\}$ convolutes into the size of the mask image.

In the main branch, the features of bottom-up blocks with lower-level information are more accurately localized by sub-sampling. In the FPN branch, the features of top-down blocks with higher-level information have a stronger spatial resolution by up-sampling. The features of top-down blocks can be enhanced by merging the features from bottom-up blocks. Based on the segmentation training, the learned pyramidal features can learn multi-scale spatial information and are feedback to the main branch's core node. As shown in Fig. 4, the subtle CDR differences between two glaucoma images can be observed on the learned feature maps. The minor difference reflects the different information on different CDRs of same glaucoma, and the difference is encoded into the core node. Based on the pixel-wise segmentation training, we argue that the multi-scale subtle differences of pathological regions at different stages can be learned by the FPN branch to tackle the **DPSD** problem.

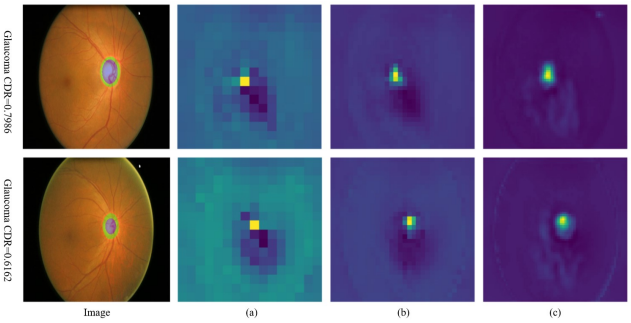


Fig. 4. Feature maps of the FPN branch. We calculate the mean value along the channel axis of the features of the three top-down blocks $\{$t3, t2,t1$\}$, then visualize the mean value: (a) the feature map of the top-down block $\{$t3$\}$ with $\mathbf{16} \times \mathbf{16}$, (b) the feature map of the top-down block $\{$t2$\}$ with $\mathbf{32} \times \mathbf{32}$, and (c) the feature map of the top-down block $\{$t1$\}$ with $\mathbf{64} \times \mathbf{64}$.

### E. Coupled Loss

In Y-Net, we integrate the classification task in the R-MAC branch and the segmentation task in the FPN branch to learn the semantic and spatial information of pathological regions simultaneously. To balance the two tasks' loss, we design a coupled loss to unify the classification and segmentation learning. In general, the gradient size is different in the convergence process of different tasks, and the sensitivity to different learning rates is also different. Unifying the scale of different loss functions can prevent the loss items with small gradients from being covered by the loss items with large gradients. Unifying the losses to the same order of magnitude can help improve

the generalization of the learned features [55]. The coupled loss function is defined as:

$$\mathcal{L} = \omega\mathcal{L}_l + (1-\omega)\mathcal{L}_r, \tag{2}$$

where $\mathcal{L}_r$ is the circle loss [56] for the classification training, $\mathcal{L}_l$ denotes the cross-entropy (CE) loss for the pixel-wise segmentation training, and $\omega$ is the weight factor. In the circle loss, each similarity score is given different penalties according to its distance to the optimal effect. In the R-MAC branch, instead of the CE loss, we adopt the circle loss to preserve the class-aware similarity of pathological regions and help prevent minimizing intra-class distance.

Based on the coupled loss unifying the classification loss and segmentation loss, the main branch's core node can effectively retain the multi-scale spatial information from segmentation training and the class-aware semantic information from classification training simultaneously. The convolutional feature maps from a certain convolutional layer can be viewed as an array of local features sampled from a dense sampling grid. In Fig. 5, the pathological region is the cup and disk of glaucoma. By observing the core node's feature maps, the FPN branch focuses on exploring the pathological region (cup and disk), and the R-MAC branch concerns the highly activated region of the whole image (glaucoma). With the help of the coupled loss balancing the two losses, the Y-Net row's feature maps confirm the effectiveness of preserving the information from the R-MAC branch and the FPN branch. Hence, the learned convolutional features of the core node can be used to generate hash-codes to combat pathological regions' ambiguous manifestations.

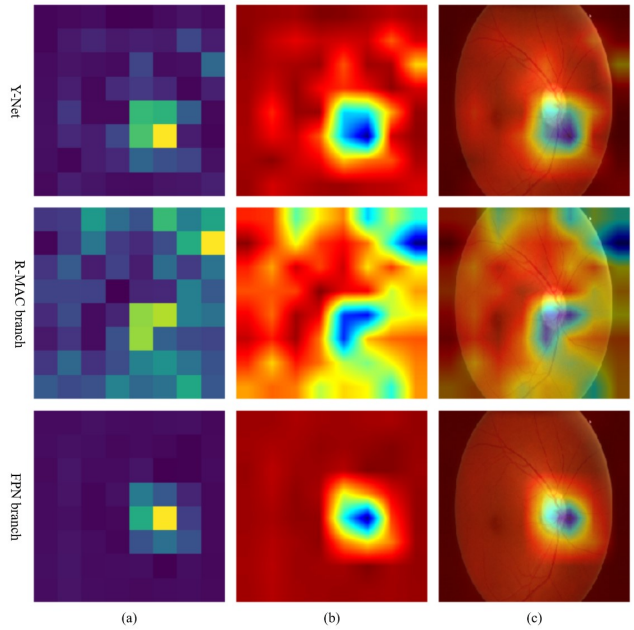


Fig. 5. Feature maps of the core node in the main branch. We calculate the mean value along the channel axis of the convolutional features with $\mathbf{256 \times 8 \times 8}$, then visualize the mean value: (a) the feature map with $\mathbf{8 \times 8}$, (b) the color map resized to the size of the input image with $\mathbf{256 \times 256}$, and (c) the overlay map combined the color map with the input image.

## IV. Experiments and Analysis

To evaluate the performance of our proposed Y-Net, we conduct extensive experiments on two public medical image datasets to verify our method's effectiveness in combating the ambiguous manifestation of pathological regions. In this section, we will introduce the experimental details and analyze the experimental results.

### A. Datasets

**Fundus** [57] contains 650 annotated retina images. Each image is tagged with classification information and manually segmented the result of optic disc and cup. This dataset is obtained from a population-based study and is therefore suitable for evaluating glaucoma screening performance. In this dataset, 168 images from glaucomatous eyes and 482 images from normal eyes are classified. Manual CDR computed from the manually segmented disc, and cup boundaries are necessary for segmentation training. Based on the classification and segmentation labels, we split this dataset into the train set and the test set by ratio $9:1$. The test set of 65 consists of 16 glaucoma images and 49 normal images, and the train set of 585 images covers 152 glaucoma images and 433 normal images.

**JSRT** [58] provides 154 nodule and 93 non-nodule chest X-ray images. Each nodule case contains a nodule only, which is rated as benign or malignant by 20 different radiologists. A detailed delineation of the segmentation's nodule is publicly available to train a lung segmentation [59]. This dataset annotates the lesion position and responding diagnosis. For example, the lesion region of a malignant image is located on the left lung's upper lobe and diagnosed as lung cancer. The annotation images for segmentation tasks are binary images in which pixels are either 255 for the foreground or 0 for the background. We sample 138 images containing 89 malignant nodules and 49 benign nodules to form a train set and 16 images containing 11 malignant nodules and 5 benign nodules to form a test set. The ratio of the train set and the test set is $9:1$.

### B. Experimental Setups

We mainly use mean average precision (**mAP**) for quantitative evaluation. In the returned list, mAP averages the ranks of images similar to the query image to measure the rank quality. The mAP is usually adopted for evaluating the retrieval performance [1], [2], and is calculated as follows:

$$AP = \frac{\sum_{k=1}^{n} P(k) \cdot rel(k)}{R}, \tag{3}$$

where $R$ denotes the number of similar results for the current query image, $P(k)$ denotes the precision of top-$k$ retrieval results, $rel_k$ is a binary indicator function equaling 1 when the $k$-th retrieved results is similar to the current query image and 0 otherwise, and $n$ denotes the total number of retrieved results. Based on the class labels and the aim of instance retrieval assisting the clinician's own decision-making by reviewing similar cases, the success criteria of similar images are defined as that the two images have similar pathological patterns.

Y-Net is compared against several representative approaches of instance-level retrieval. The comparative approaches are categorized as: weight feature aggregating on convolutional features, regional feature aggregating on convolutional features, and feature embedding from a full-connected layer.

- **Weight feature aggregating. CroW** [38] estimates a spatial weighting of the features as a combination of convolutional feature maps across all channels of the layer. Features at locations with salient visual content are boosted while weights in non-salient locations are decreased. To explicitly leverage semantic information, **CAM** [39] obtains semantic-aware weights for convolutional features by exploiting the predicted classes. CAM generates a set of spatial maps highlighting the contribution of the regions within an image. Each map is used to weigh the convolutional features and generate a set of class vectors that are aggregated as the region vectors over the fixed region strategy of R-MAC. CAM inspires our R-MAC branch of Y-Net. **BLCF**

[41] builds an efficient image representation by combining saliency weighting over convolutional features aggregated by using a large vocabulary with a bag of words (BoW) model [60]. **SOLAR-Local** [61] focuses on second-order spatial information to learn local patch descriptors without extra supervision. Based on the feature weighting strategy [62], it combines the second-order spatial attention and the second-order descriptor loss to improve image features for retrieval and matching.

- **Regional feature aggregating. R-MAC** [42] is an aggregation method for convolutional features to generate a set of regional vectors by performing spatial max-pooling within a particular region. Building on the R-MAC descriptor, **R-MAC + RPN** [43], [62] can enhance the ability to focus on relevant regions in the image by replacing the rigid grid with a region proposal network (RPN) trained to localize regions of interest in images. **Regional Attention** [37] presents a context-aware regional attention network for tackling the problem of region-based feature aggregation suffering from the background clutter and varying importance of regions, especially in R-MAC, by weighting an attentive score of a region. **Deep Vision + SOLO** [47] is trained for instance-level retrieval of image- and region-wise representations pooled from an object detection CNN. In this experiment, we take advantage of the object proposals learned by SOLO [63] and their associated convolutional features to build an instance search pipeline.
- **Feature embedding.** Three deep hashing methods using feature embedding are used to build benchmarks for our Y-Net, including **DPSH** [7], **DSH** [35], **DRH** [36], **DDMH** [16]. DPSH performs simultaneous feature learning and hash-code learning with deep neural networks by maximizing pairwise similarities. Inspired by DPSH, DSH proposes a triplet label-based deep hashing method to maximize the given triplet labels' likelihood. DRH offers good separability of classes in hashing space while preserving semantic similarities in local embedding neighborhoods for supervised hashing of medical images through residual learning. DPSH and DSH use AlexNet [64] as the backbone. Recently, the residual block [49] has been used popularly as the backbone in deep hashing methods such as DRH and shows the advantage of feature extraction. In our Y-Net, the main branch also uses the residual block as the backbone. DDMH proposes a unique disentangled triplet loss to effectively push positive and negative sample pairs by desired Hamming distance discrepancies for hash-codes with different lengths.

Our Y-Net is implemented under the PyTorch framework, and experiments are run on Geforce RTX 2080 Ti. In our work, the indexing and similarity calculation for evaluation uses Faiss [65], a library for efficient similarity search and clustering of dense vectors. We use the mini-batch stochastic gradient descent with $0.9$ momentum. The mini-batch size of images is fixed as $32$, and the weight decay parameter is $0.001$. All deep models are trained from scratch with $500$ epochs. It spends approximately $3$ hours for training our Y-Net. The pixel-wise cross-entropy loss is used in the segmentation task. The circle loss [56] is used for classification training by using cosine similarity and setting a scale of $32$, a margin of $0.25$. The weight factor $\omega$ in the coupled loss is initially set as $0.5$. We use the $5$-fold cross-validation to select the best classification and segmentation model. The parameters of comparative methods are set according to their implementation details in the corresponding papers, and the best performance is reported. Based on top-10 retrieval results, we investigate our Y-Net's performance over hash-code with lengths of $36$, $64$, $128$, $256$, respectively. According to Table I, with the hash-code lengthen, the performance can correspondingly improve at the cost of storage and search efficiency. As a trade-off between performance and search cost, we report all the performances over $64$-bits hash-code for our Y-Net.

TABLE I
MAP OF Y-NET OVER THE VARYING LENGTH OF HASH-CODES ON THE FUNDUS AND JSRT DATASETS.

| Datasets | mAP@36 | mAP@64 | mAP@128 | mAP@256 |
|---|---|---|---|---|
| **Fundus** | 0.5903 | 0.6102 | 0.6266 | 0.6308 |
| **JSRT** | 0.5361 | 0.5518 | 0.5732 | 0.5809 |

### C. Experimental Results

The following research questions will be answered by analyzing experimental results:

- **RQ1** Does our proposed Y-Net outperform the state-of-the-art methods on retrieval performance in medical instance retrieval?
- **RQ2** Can our proposed Y-Net help to combat the ambiguity of pathological regions in medical instance retrieval?
- **RQ3** What are the effectiveness of the R-MAC branch, the FPN branch, and the coupled loss in our proposed Y-Net framework?
- **RQ4** How is the retrieval efficiency of our proposed Y-Net?

*1) Quantitative Analysis (RQ1):* The performance of the mAP over the returned list of $5$, $10$, $20$, and $50$ on Fundus and JSRT datasets are reported in Table II, respectively. On the whole, when the returned list lengthens, all methods' performance declines to some extent. Our Y-Net all achieves significant gains of mAP over the varying returned list on the two datasets. Experimental results on the Fundus dataset show that Y-Net outperforms the second-highest methods (underline) by $7.60\%$, $11.18\%$, $9.35\%$, $7.26\%$ correspond to the different number of the returned list. Y-Net also achieves the best performance on the JSRT dataset compared to the other methods. For the methods obtaining the second-highest performance, CAM is a weighing feature method aggregating on convolutional features, and DRH is a method of feature embedding. This demonstrates that the methods of regional feature aggregating on convolutional features may lose related information between regions after region proposals. This loss prevents them from obtaining better performance. Among methods of weight feature aggregating, SOLAR-Local yields good performance by exploiting the second-order spatial information. CAM can achieve better performance than SOLAR-Local by exploiting class semantic information. The retrieval performance on the Fundus dataset is higher than that on the JSRT dataset by $10.58\%$ on the returned list of $10$. The reason for this gap has two points. The shortage of specificity is the main challenge for chest X-ray image analysis tasks. The JSRT dataset only provides lung masks but not lesion masks; those non-lesion regions in the lung mask may affect the discriminative information learning.

Compared to DRH, CAM acquires a better performance over the returned list of $5$ and $10$. This demonstrates that it effectively explore pathological regions and weigh their activations by exploiting the correlation between class labels and pathological regions. Inspired to CAM, the R-MAC branch in our Y-Net contributes to increasing the retrieval performance by focusing on the pathological regions and weights these regions with class activations. Benefiting from adopting the residual block as the backbone, DRH is superior to CAM over the returned list of $20$ and $50$. To further improve the performance over the longer returned list, we need to exploit spatially subtle differences of pathologically abnormal regions with

TABLE II
MAP OVER THE VARYING NUMBER OF THE RETURNED LIST ON THE FUNDUS AND JSRT DATASETS.

| Methods | Dim | Fundus | | | | JSRT | | | |
|---|---|---|---|---|---|---|---|---|---|
| | | top-5 | top-10 | top-20 | top-50 | top-5 | top-10 | top-20 | top-50 |
| **CroW [38]** | 512 | 0.5223 | 0.4681 | 0.4471 | 0.4366 | 0.4993 | 0.4705 | 0.4396 | 0.4189 |
| **CAM [39]** | 2048 | 0.5917 | 0.5488 | 0.4982 | 0.4609 | 0.5611 | 0.5124 | 0.4497 | 0.4187 |
| **BLCF [41]** | 1000 | 0.4890 | 0.4793 | 0.4463 | 0.4216 | 0.4701 | 0.4356 | 0.4096 | 0.3903 |
| **SOLAR-Local [61]** | 1024 | 0.5701 | 0.5274 | 0.4766 | 0.4482 | 0.5443 | 0.4987 | 0.4264 | 0.4051 |
| **R-MAC [42]** | 512 | 0.5016 | 0.4884 | 0.4585 | 0.4528 | 0.4682 | 0.4191 | 0.3965 | 0.3812 |
| **R-MAC + RPN [62]** | 3072 | 0.5483 | 0.5024 | 0.4685 | 0.4446 | 0.4805 | 0.4461 | 0.4098 | 0.3951 |
| **Regional Attention [37]** | 2048 | 0.5674 | 0.5279 | 0.5070 | 0.4854 | 0.4984 | 0.4621 | 0.4289 | 0.4069 |
| **Deep Vision + SOLO [47]** | 3072 | 0.5486 | 0.5001 | 0.4889 | 0.4815 | 0.5123 | 0.4756 | 0.4358 | 0.4123 |
| **DPSH [7]** | 64 | 0.5044 | 0.4693 | 0.4451 | 0.4270 | 0.4581 | 0.4203 | 0.3891 | 0.3677 |
| **DSH [35]** | 64 | 0.5052 | 0.4882 | 0.4788 | 0.4734 | 0.5487 | 0.4921 | 0.4578 | 0.4332 |
| **DRH [36]** | 64 | 0.5712 | 0.5435 | 0.5322 | 0.5203 | 0.5306 | 0.4912 | 0.4651 | 0.4498 |
| **DDMH [16]** | 32 | 0.5231 | 0.5051 | 0.4962 | 0.4802 | 0.5396 | 0.4869 | 0.4421 | 0.4284 |
| **Y-Net (ours)** | 64 | **0.6367** | **0.6102** | **0.5820** | **0.5581** | **0.6013** | **0.5518** | **0.5284** | **0.4976** |

the help of pixel-wise segmentation training in the FPN branch. Due to the differentiating ability of the subtle differences in pathological regions, our Y-Net surpasses DRH over the returned list of 20 and 50 compared to CAM. In summary, three points contribute to the performance of our Y-Net. (1) The R-MAC branch learns the class-aware semantic information of pathological regions. (2) The FPN branch explores the multi-scale subtle spatial information of pathological regions. (3) The main branch uses the residual block as the backbone.

*2) Qualitative Analysis (RQ2):* Lung nodules are small masses of tissue in the lung and quite common. They appear as round, white shadows on a chest X-ray. Lung nodules are usually about 0.2 inches (5 millimeters) to 1.2 inches (30 millimeters) in size. A larger lung nodule, such as 30 millimeters or larger, is more likely to be cancerous than a smaller lung nodule. The regions of chest nodules in X-ray images are hard to differentiate malignant or benign according to the spatial information, including texture and size. So this is a typical **SPDD** problem. As shown in Fig. 6, our Y-Net returns more malignant images and ranking ahead than DRH by querying a malignant image. Based on the FPN branch exploiting spatially subtle differences of nodule regions, the R-MAC branch cooperatively encodes the class-aware semantic information of pathological regions into the hash-codes. By exploiting the correlation between class labels and pathological regions, the R-MAC branch can address the **SPDD** problem in medical instance retrieval. In fact, the R-MAC branch weighs the regional of maximum activation by conveying the class-based semantic information to the R-MAC descriptor and the convolutional features. The class-weighted regional of maximum activation can differentiate the same performance of different diseases of medical images.

The size of CDR computation from color fundus images is the main clue for glaucoma diagnosis [66]. The different size of CDR denotes different grading of glaucoma. It is useful for clinicians to find the most similar images with closer CDR sizes to make a medico-decision. As shown in Fig. 7, compared to the DRH, Y-Net returned more glaucoma images with closer CDR sizes and ranked ahead by querying a glaucoma image. According to this experimental result, we argue that the FPN branch can effectively encode the subtle differences of pathological regions into the hash-codes to address the **DPSD** problem by mining the multi-scale spatial information. The FPN branch can locate pathological regions' subtle differences at different stages of the same disease based on the pixel-wise segmentation training. In essence, the FPN branch weights the pathological regions by segmentation training. The weighted pathological regions can be encoded as the most discriminative parts of the hash-codes to differentiate the same disease's different manifestations at different stages.

Our Y-Net's R-MAC branch exploits the class semantic information to weigh regions of maximum activation to tackle the SPDD problem. Apart from the same pathological criteria evaluation (benign and malignant), we also apply the disease label to evaluate the performance to embody the effectiveness of tackling the SPDD problem. The large disease label consists of lung cancer, granuloma, cryptococcosis, inflammatory mass, etc. The fine disease label for lung cancer includes adenocarcinoma, large cell carcinoma, small cell carcinoma, etc. On the returned list of 10, our method outperforms CAM by 8.12% average precision on diagnosing disease. This demonstrates that our method can effectively differentiate the similar manifestation of different diseases. Our Y-Net's FPN branch explores the spatially subtle differences of the lesion region to overcome the DPSD problem. Regarding the DPSD problem, we apply average CDR to evaluate the performance on differentiating the different manifestations of the same disease in different stages. Our Y-Net yields the average CDR gap of 0.2157 between the query image and the retrieved images, while CAM obtains 0.3521. The convolutional features in the core node of the main branch learn the information from both branches to promote hash-codes' discriminative ability.

*3) Ablation Study (RQ3):* To further research the R-MAC branch and FPN branch's contribution, we conduct an ablation study by cropping the corresponding branch of Y-Net. As shown in Table III, Y-Net without the FPN branch can achieve better performance than Y-Net without the R-MAC branch, and Y-Net achieves convincing performance by unifying the FPN branch and R-MAC branch. Without the FPN branch, Y-net can achieve competitive performance compared to CAM and DRH. Upon the R-MAC branch, Y-Net can obtain a significant gain by adding the FPN branch. This demonstrates that the R-MAC branch can differentiate pathological regions' similar manifestations by weighing the regional of maximum activation based on the classification training. The added gain benefits from the FPN branch, which exploits the subtle differences of pathological regions by mining the multi-scale spatial information based on the segmentation training. As shown in Fig. 7, the glaucoma images ranked ahead are closer to the query image in CDR size. This also confirms the FPN branch's effectiveness in preventing the R-MAC branch from minimizing the intra-class distance. Based on this joint learning scheme, the core node in the main branch absorbs the class-aware semantic information from the R-MAC branch and spatially subtle differences from the FPN branch, then are mapped into the hash-codes. The learned hash-codes can be used to combat the ambiguous manifestations of pathological regions.

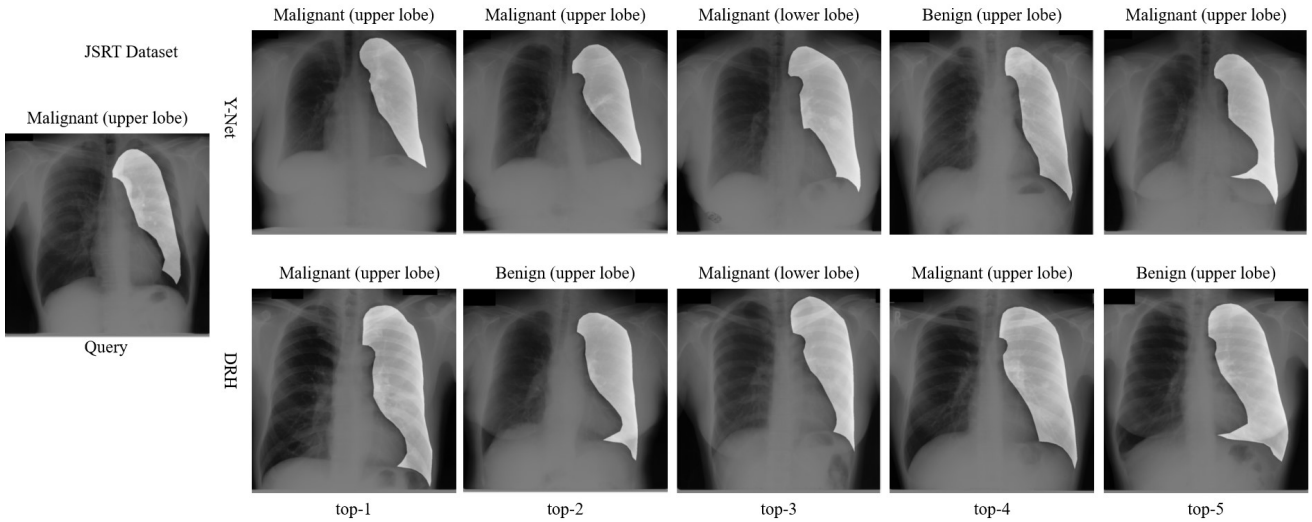


Fig. 6. Ranking of the top-5 returned list on the JSRT dataset. We query a malignant image and obtain the ranking of the top-5 returned list for Y-Net and DRH, respectively. Each image shows the position of the chest nodule labeled manually.

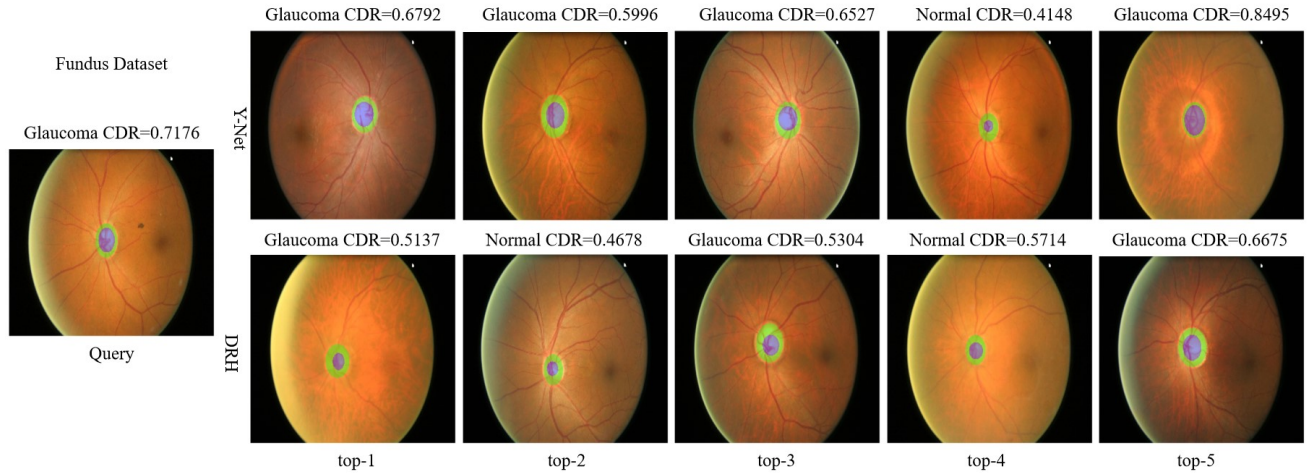


Fig. 7. Ranking of the top-5 returned list on the Fundus dataset. We query a glaucoma image and obtain the ranking of the top-5 returned list for Y-Net and DRH, respectively. Each image shows its Cup to Disk Ratio (CDR) size labeled manually.

TABLE III

MAP OF BRANCHES OF Y-NET OVER THE VARYING NUMBER OF THE RETURNED LIST ON THE FUNDUS AND JSRT DATASETS.

| Branches | Fundus | | | | JSRT | | | |
|---|---|---|---|---|---|---|---|---|
| | top-5 | top-10 | top-20 | top-50 | top-5 | top-10 | top-20 | top-50 |
| **Y-Net w/o FPN and R-MAC branch** | 0.5001 | 0.4871 | 0.4679 | 0.4575 | 0.5324 | 0.4856 | 0.4509 | 0.4297 |
| **Y-Net w/o FPN branch** | 0.5881 | 0.5656 | 0.5443 | 0.5033 | 0.5325 | 0.5114 | 0.4831 | 0.4501 |
| **Y-Net w/o R-MAC branch** | 0.5561 | 0.5179 | 0.4854 | 0.4536 | 0.5210 | 0.4914 | 0.4597 | 0.4285 |
| **Y-Net w/o Circle loss** | 0.6061 | 0.5879 | 0.5554 | 0.5136 | 0.5684 | 0.5291 | 0.4976 | 0.4703 |
| **Y-Net** | **0.6367** | **0.6102** | **0.5820** | **0.5581** | **0.6013** | **0.5518** | **0.5284** | **0.4976** |

Based on the above experimental analysis, we confirm the effect of unifying classification and segmentation. Next, we would like to discuss the coupled loss function's effect in both branches' unified training. First, as shown in Table III, the R-MAC branch with the circle loss achieves better performance than the R-MAC branch with the CE loss. The circle loss can help the R-MAC branch maximize the intra-class similarity and minimize inter-class similarity by pair similarity optimization. Second, Compared to the sum of the two losses, the coupled loss function can improve retrieval performance averagely by 2% on mAP over the varying number of the returned list on the Fundus and JSRT datasets. This demonstrates that the coupled loss can help facilitate the generalization of the learned convolutional features by unifying the losses to the same order of magnitude. As shown in Fig 8, compared to the sum of the two losses (blue), the coupled loss (red) unifies the scale of the circle loss (yellow) and the cross-entropy loss (green) to prevent the loss unbalance in the convergence process of different tasks. In the process of screening and diagnosis, the ambiguous manifestation of pathological regions may be varied. Hence, the two tasks can be mutually beneficial to enhance Y-Net's generalization by the joint learning scheme.

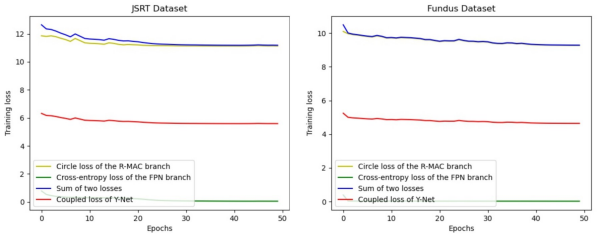


Fig. 8. Qualitative results of the coupled loss. The different loss of top-50 iterations during training on the Fundus and JSRT datasets are compared.

*4) Retrieval Efficiency Analysis (RQ4):* In this section, we discuss the efficiency of the proposed Y-Net from three-folds by putting the Fundus dataset as an example.

1) **Feature computation time.** Based on the pre-trained Y-Net model, we inference the hash-codes of 64-bits from the core node in the main branch. Hence, the feature computation of the main branch occupies the most time cost in the test stage. We can complete the hash-codes generating for the training set of 585 images in 4 seconds on GPU. The feature computation time of our Y-Net is fair to the most comparative methods.
2) **Retrieval time.** After hash-codes generating, we build the index in 1 second for the training set by using Faiss. By querying the test set of 65 images, returning top-10 most similar images can be done in $34ms$. The time-consuming processes of the search engine are the indexing search and similarity calculation. The time cost of both lengthens when the size of feature vectors used for similarity calculation extends. As Table II shows (column: Dim), Y-Net's hash-code length is equal to the methods using feature embedding.
3) **Memory cost.** The memory-consuming is about $2,000$ Mbps during model training by setting the batch size at 32. The online search for the index also consumes about $2,000$ Mbps. The memory cost depends on the model complexity where our Y-Net is fair to the methods aggregating regional features.

According to the above analysis of efficiency, our Y-Net can provide fair real-time responses with significantly improving the performance by comparing to the state-of-the-art methods.

## V. Conclusions

To combat the manifestation ambiguity in medical instance retrieval, we propose a novel framework called Y-Net, encoding images into compact hash-codes aggregating from convolutional features. The proposed Y-Net contains the main branch, the R-MAC branch, the FPN branch. Based on the classification loss, the R-MAC branch encodes the class-aware semantic information of pathological regions into the convolutional features to avoid **SPDD** problem. And based on the pixel-wise segmentation loss, the FPN branch encodes the spatially subtle differences of pathological regions into the convolutional features to overcome the **DPSD** problem. After unifying the classification and segmentation training, the learned convolutional features in the main branch are directly aggregated to generate the hash-codes for similarity measure. The extensive experiments on the two medical image datasets with class and pixel-wise mask labels show that our Y-Net can alleviate pathologically abnormal regions' ambiguity.

There also exist two limitations of this work. First, it is hard to acquire medical image datasets with pixel-wise segmentation annotations, while detecting the subtle differences with the bounding box of pathological regions is challenging. This restricts our Y-Net's availability and universality. Second, the multi-instances and multi-labels of medical images significantly lift the difficulty of combating pathologically abnormal regions' ambiguity. In the future, we would like to explore the solutions to address such issues.

## References

[1] L. Zheng, Y. Yang, and Q. Tian, "Sift meets cnn: A decade survey of instance retrieval," *IEEE transactions on pattern analysis and machine intelligence*, vol. 40, no. 5, pp. 1224–1244, 2017.
[2] W. Zhou, H. Li, and Q. Tian, "Recent advance in content-based image retrieval: A literature survey," *arXiv preprint arXiv:1706.06064*, 2017.
[3] H.-C. Xiao and W.-L. Zhao, "Deeply activated salient region for instance search," *arXiv preprint arXiv:2002.00185*, 2020.
[4] Q. Li, F. Li, J. Shiraishi, S. Katsuragawa, S. Sone, and K. Doi, "Investigation of new psychophysical measures for evaluation of similar images on thoracic computed tomography for distinction between benign and malignant nodules," *Medical Physics*, vol. 30, no. 10, pp. 2584–2593, 2003.
[5] Y. Zhan and W.-L. Zhao, "Instance search via instance level segmentation and feature representation," *arXiv preprint arXiv:1806.03576*, 2018.
[6] A. Babenko, A. Slesarev, A. Chigorin, and V. Lempitsky, "Neural codes for image retrieval," in *European conference on computer vision*, pp. 584–599, Springer, 2014.
[7] W.-J. Li, S. Wang, and W.-C. Kang, "Feature learning based deep supervised hashing with pairwise labels," *arXiv preprint arXiv:1511.03855*, 2015.
[8] A. Babenko and V. Lempitsky, "Aggregating local deep features for image retrieval," in *Proceedings of the IEEE international conference on computer vision*, pp. 1269–1277, 2015.
[9] A. S. Razavian, J. Sullivan, S. Carlsson, and A. Maki, "Visual instance retrieval with deep convolutional networks," *ITE Transactions on Media Technology and Applications*, vol. 4, no. 3, pp. 251–258, 2016.
[10] B. Zhuang, G. Lin, C. Shen, and I. Reid, "Fast training of triplet-based deep binary embedding networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 5955–5964, 2016.
[11] S. Conjeti, M. Paschali, A. Katouzian, and N. Navab, "Deep multiple instance hashing for scalable medical image retrieval," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pp. 550–558, Springer, 2017.
[12] T.-T. Do, T. Hoang, D.-K. Le Tan, A.-D. Doan, and N.-M. Cheung, "Compact hash code learning with binary deep neural network," *IEEE Transactions on Multimedia*, vol. 22, no. 4, pp. 992–1004, 2019.
[13] J. Fang, Y. Xu, X. Zhang, Y. Hu, and J. Liu, "Attention-based saliency hashing for ophthalmic image retrieval," in *2020 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, pp. 990–995, IEEE, 2020.
[14] J. Fang, H. Fu, and J. Liu, "Deep triplet hashing network for case-based medical image retrieval," *Medical Image Analysis*, vol. 69, p. 101981, 2021.
[15] D. Wu, Q. Dai, J. Liu, B. Li, and W. Wang, "Deep incremental hashing network for efficient image retrieval," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 9069–9077, 2019.
[16] E. Yang, D. Yao, B. Cao, H. Guan, P.-T. Yap, D. Shen, and M. Liu, "Deep disentangled hashing with momentum triplets for neuroimage search," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pp. 191–201, Springer, 2020.
[17] M. Loyman and H. Greenspan, "Semi-supervised lung nodule retrieval," *arXiv preprint arXiv:2005.01805*, 2020.
[18] J. Faruque, C. F. Beaulieu, J. Rosenberg, D. Rubin, D. Yao, and S. Napel, "Content-based image retrieval in radiology: analysis of variability in human perception of similarity," *Journal of Medical Imaging*, vol. 2, no. 2, p. 025501, 2015.
[19] Z. Li, X. Zhang, H. Müller, and S. Zhang, "Large-scale retrieval for medical image analytics: A comprehensive review," *Medical image analysis*, vol. 43, pp. 66–84, 2018.
[20] M. Slaney and M. Casey, "Locality-sensitive hashing for finding nearest neighbors [lecture notes]," *IEEE Signal processing magazine*, vol. 25, no. 2, pp. 128–131, 2008.
[21] M. Raginsky and S. Lazebnik, "Locality-sensitive binary codes from shift-invariant kernels," in *Advances in neural information processing systems*, pp. 1509–1517, 2009.

[22] Z. Chen, R. Cai, J. Lu, J. Feng, and J. Zhou, "Order-sensitive deep hashing for multimorbidity medical image retrieval," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pp. 620–628, Springer, 2018.

[23] S. Conjeti, A. Katouzian, A. Kazi, S. Mesbah, D. Beymer, T. F. Syeda-Mahmood, and N. Navab, "Metric hashing forests," *Medical image analysis*, vol. 34, pp. 13–29, 2016.

[24] H. Liu, R. Wang, S. Shan, and X. Chen, "Deep supervised hashing for fast image retrieval," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 2064–2072, 2016.

[25] Ş. Öztürk, "Image inpainting based compact hash code learning using modified u-net," in *2020 4th International Symposium on Multidisciplinary Studies and Innovative Technologies (ISMSIT)*, pp. 1–5, IEEE, 2020.

[26] H. Zhu, M. Long, J. Wang, and Y. Cao, "Deep hashing network for efficient similarity retrieval," in *Thirtieth AAAI Conference on Artificial Intelligence*, 2016.

[27] H. Jegou, M. Douze, and C. Schmid, "Hamming embedding and weak geometric consistency for large scale image search," in *European conference on computer vision*, pp. 304–317, Springer, 2008.

[28] R. Arandjelović and A. Zisserman, "Smooth object retrieval using a bag of boundaries," in *2011 International Conference on Computer Vision*, pp. 375–382, IEEE, 2011.

[29] D. G. Lowe, "Distinctive image features from scale-invariant keypoints," *International journal of computer vision*, vol. 60, no. 2, pp. 91–110, 2004.

[30] H. Bay, T. Tuytelaars, and L. Van Gool, "Surf: Speeded up robust features," in *European conference on computer vision*, pp. 404–417, Springer, 2006.

[31] G. Tolias, Y. Avrithis, and H. Jégou, "To aggregate or not to aggregate: Selective match kernels for image search," in *Proceedings of the IEEE International Conference on Computer Vision*, pp. 1401–1408, 2013.

[32] G. Tolias, T. Furon, and H. Jégou, "Orientation covariant aggregation of local descriptors with embeddings," in *European Conference on Computer Vision*, pp. 382–397, Springer, 2014.

[33] Ş. Öztürk, "Stacked auto-encoder based tagging with deep features for content-based medical image retrieval," *Expert Systems with Applications*, vol. 161, p. 113693, 2020.

[34] Ş. ÖZTÜRK, "Two-stage sequential losses based automatic hash code generation using siamese network," *Avrupa Bilim ve Teknoloji Dergisi*, pp. 39–46.

[35] X. Wang, Y. Shi, and K. M. Kitani, "Deep supervised hashing with triplet labels," in *Asian conference on computer vision*, pp. 70–84, Springer, 2016.

[36] S. Conjeti, A. G. Roy, A. Katouzian, and N. Navab, "Hashing with residual networks for image retrieval," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pp. 541–549, Springer, 2017.

[37] J. Kim and S.-E. Yoon, "Regional attention based deep feature for image retrieval.," in *BMVC*, p. 209, 2018.

[38] Y. Kalantidis, C. Mellina, and S. Osindero, "Cross-dimensional weighting for aggregated deep convolutional features," in *European conference on computer vision*, pp. 685–701, Springer, 2016.

[39] A. Jimenez, J. M. Alvarez, and X. Giro-i Nieto, "Class-weighted convolutional features for visual instance search," *arXiv preprint arXiv:1707.02581*, 2017.

[40] E. Mohedano, K. McGuinness, N. E. O'Connor, A. Salvador, F. Marques, and X. Giro-i Nieto, "Bags of local convolutional features for scalable instance search," in *Proceedings of the 2016 ACM on International Conference on Multimedia Retrieval*, pp. 327–331, 2016.

[41] E. Mohedano, K. McGuinness, X. Giró-i Nieto, and N. E. O'Connor, "Saliency weighted convolutional features for instance search," in *2018 international conference on content-based multimedia indexing (CBMI)*, pp. 1–6, IEEE, 2018.

[42] G. Tolias, R. Sicre, and H. Jégou, "Particular object retrieval with integral max-pooling of cnn activations," *arXiv preprint arXiv:1511.05879*, 2015.

[43] A. Gordo, J. Almazan, J. Revaud, and D. Larlus, "End-to-end learning of deep visual representations for image retrieval," *International Journal of Computer Vision*, vol. 124, no. 2, pp. 237–254, 2017.

[44] S. Ren, K. He, R. Girshick, and J. Sun, "Faster r-cnn: Towards real-time object detection with region proposal networks," in *Advances in neural information processing systems*, pp. 91–99, 2015.

[45] Z. Laskar and J. Kannala, "Context aware query image representation for particular object retrieval," in *Scandinavian Conference on Image Analysis*, pp. 88–99, Springer, 2017.

[46] J. Cao, L. Liu, P. Wang, Z. Huang, C. Shen, and H. T. Shen, "Where to focus: Query adaptive matching for instance retrieval using convolutional feature maps," *arXiv preprint arXiv:1606.06811*, 2016.

[47] A. Salvador, X. Giró-i Nieto, F. Marqués, and S. Satoh, "Faster r-cnn features for instance search," in *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*, pp. 9–16, 2016.

[48] T.-Y. Lin, P. Dollár, R. Girshick, K. He, B. Hariharan, and S. Belongie, "Feature pyramid networks for object detection," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 2117–2125, 2017.

[49] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 770–778, 2016.

[50] J. L. Long, N. Zhang, and T. Darrell, "Do convnets learn correspondence?," in *Advances in neural information processing systems*, pp. 1601–1609, 2014.

[51] L. Liu, C. Shen, and A. van den Hengel, "The treasure beneath convolutional layers: Cross-convolutional-layer pooling for image classification," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 4749–4757, 2015.

[52] S. Ioffe and C. Szegedy, "Batch normalization: Accelerating deep network training by reducing internal covariate shift," *arXiv preprint arXiv:1502.03167*, 2015.

[53] W. Liu, D. Anguelov, D. Erhan, C. Szegedy, S. Reed, C.-Y. Fu, and A. C. Berg, "Ssd: Single shot multibox detector," in *European conference on computer vision*, pp. 21–37, Springer, 2016.

[54] K. He, G. Gkioxari, P. Dollár, and R. Girshick, "Mask r-cnn," in *Proceedings of the IEEE international conference on computer vision*, pp. 2961–2969, 2017.

[55] A. Kendall, Y. Gal, and R. Cipolla, "Multi-task learning using uncertainty to weigh losses for scene geometry and semantics," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 7482–7491, 2018.

[56] Y. Sun, C. Cheng, Y. Zhang, C. Zhang, L. Zheng, Z. Wang, and Y. Wei, "Circle loss: A unified perspective of pair similarity optimization," *arXiv preprint arXiv:2002.10857*, 2020.

[57] J. Cheng, Z. Zhang, D. Tao, D. W. K. Wong, J. Liu, M. Baskaran, T. Aung, and T. Y. Wong, "Similarity regularized sparse group lasso for cup to disc ratio computation," *Biomedical optics express*, vol. 8, no. 8, pp. 3763–3777, 2017.

[58] J. Shiraishi, S. Katsuragawa, J. Ikezoe, T. Matsumoto, T. Kobayashi, K.-i. Komatsu, M. Matsui, H. Fujita, Y. Kodera, and K. Doi, "Development of a digital image database for chest radiographs with and without a lung nodule: receiver operating characteristic analysis of radiologists' detection of pulmonary nodules," *American Journal of Roentgenology*, vol. 174, no. 1, pp. 71–74, 2000.

[59] B. Van Ginneken, M. B. Stegmann, and M. Loog, "Segmentation of anatomical structures in chest radiographs using supervised methods: a comparative study on a public database," *Medical image analysis*, vol. 10, no. 1, pp. 19–40, 2006.

[60] J. Brownlee, "A gentle introduction to the bag-of-words model," *Machine Learning Mastery*, vol. 21, 2017.

[61] T. Ng, V. Balntas, Y. Tian, and K. Mikolajczyk, "Solar: Second-order loss and attention for image retrieval," *arXiv preprint arXiv:2001.08972*, 2020.

[62] J. Revaud, J. Almazán, R. S. Rezende, and C. R. d. Souza, "Learning with average precision: Training image retrieval with a listwise loss," in *Proceedings of the IEEE International Conference on Computer Vision*, pp. 5107–5116, 2019.

[63] X. Wang, T. Kong, C. Shen, Y. Jiang, and L. Li, "Solo: Segmenting objects by locations," *arXiv preprint arXiv:1912.04488*, 2019.

[64] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *Advances in neural information processing systems*, pp. 1097–1105, 2012.

[65] J. Johnson, M. Douze, and H. Jégou, "Billion-scale similarity search with gpus," *IEEE Transactions on Big Data*, 2019.

[66] T. Li, W. Bo, C. Hu, H. Kang, H. Liu, K. Wang, and H. Fu, "Applications of deep learning in fundus images: A review," *Medical Image Analysis*, p. 101971, 2021.