Combating Ambiguity for Hash-code Learning in Medical Instance Retrieval

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

Dear Editors and Reviewers,

On behalf of all co-authors, we thank you very much for giving us an opportunity to revise our manuscript submitted to JBHI (ID: JBHI-01721-2020; Title: Combating Ambiguity for Hash-code Learning in Medical Instance Retrieval), which obtained the decision of “major revisions needed” in the first review round. We appreciate the high quality of the peer review team and the high standard of the JBHI submission. Benefiting from the significant inputs and recommendations from the review team, we have been able to prepare a much better manuscript with great effort on addressing all concerns proposed by the reviewers. The major changes of the revised version include:

1. We conduct the experiments on a chest X-ray dataset to replace the ultrasound dataset. The results show that our method achieves the best performance compared to other methods.

2. We add two state-of-the-art methods as baselines published in 2020. The comparison result demonstrates that our method has a competitive advantage over these latest methods.

3. We provide more results and analysis to demonstrate the effectiveness of our method in combating the ambiguity for medical instance retrieval.

Please find the detailed point-by-point responses to all the raised concerns in the next pages.

Yours sincerely,

Jiansheng Fang
On behalf of all co-authors

Response to Reviewer #1:

[Q1] Please add a strong motivation section before presenting the contributions.

[R1] In response to the reviewer's comment, we have refined two motivations of our Y-Net framework in Section 1, and as follows:

"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."

[Q2] Contributions should be presented more simply and clearly. Please review and change them.

[R2] According to your useful suggestions, we simplify our contributions in Section 1 and rewrite 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. We will release our code and model after paper acceptance."

[Q3] Problems are mentioned, but the way to go for a solution is not specified in introduction.

[R3] Thank you for your kind suggestions. After the problem definition, we have separated the solution and clarify the motivation more clearly the introduction in Section 1. The solution is introduced as follows:

"The ambiguity of pathologically abnormal regions may prevent the assimilation of medical instance retrieval into an assistant tool for medico-decision[1]. One solution is to provide fine-grained labels to combat the ambiguity of pathologically abnormal regions, but medical annotations remain highly dependent on manual feedback with high inter-observer variability[2]. Generally, medical image datasets can provide 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[3]. Following this way, we present a novel end-to-end framework, called Y-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. ”

[Q4] Retrieval is a very technical subject and the main idea can slip when the related works section is kept this short. Please develop a related work including feature extraction classes.

[R4] We thank the reviewers for providing the thoughtful comments. We have interspersed related works of feature extraction for image retrieval in Section 2. Such as:

“Hashing methods can be divided into data-independent methods and data-dependent methods. The data-independent methods[6,7] 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[8]: (1) shallow learning-based hashing methods, like metric hashing forests[9], and kernel sensitive hashing[10]; (2) deep learning-based hashing methods, like image inpainting-based compact hash code learning[11], and deep hashing network[12]. In contrast to the data-independent methods, they extract global features for hashing in an end-to-end manner.”

“Prior to deep learning, these works based on local features extraction, then aggregated into a global vector [4, 5].”

“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[13,14] can be grouped into this category using feature embedding tailor features from fully-connected layers for hash-codes generating. ”

[Q5] Please add current state-of-the-art studies about medical image retrieval. Stacked auto-encoder based tagging with deep features for content-based medical image retrieval, Image Inpainting based Compact Hash Code Learning using Modified U-Net, Two-Stage Sequential Losses based Automatic Hash Code Generation using Siamese Network.

[R5] Thank you for your kind sharing. We have cited these three works in Section 2 (References: [11, 13, 14]).

[Q6] I think find another place for your own method contributions that you added under the 'Medical Instance Retrieval' section.

[R6] Thanks for your detailed review. In section 2, we highlighted the difference compared with the existing works.

[Q7] How does the proposed approach deal with SPDD and DPSD problems? thanks to what features?

[R7] Thanks for your professional review. This problem has be answered by analyzing experimental results in Section 4.3.3 [Ablation Study(RQ3)]. Here, we briefly clarify our viewpoint. In the R-MAC branch, the classification loss minimizes intra-class distance and maximizes 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.

[Q8] How are the two tasks combined in the core node part? What are the loss functions?

[R8] Nice question. On the one hand, 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. On the other hand, 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. We apply the circle loss to train the R-MAC branch and the cross-entropy loss to train the FPN branch. To balance the loss of two tasks, 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[15].

[Q9] One of the two missions can be dominant, how did you deal with that?

[R9] This is a professional question. In fact, we have encountered this problem in the experimental process. We apply the coupled loss to replace the sum loss and tune the weights of these two tasks in the coupled loss.

[Q10]- Provide details of the end-to-end training.

[R10] Thank you for your kind suggestions. We have emphasized the difference compared to the current deep hashing methods in Section 2. 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 pipeline of our method has been clarified in Section 3.

[Q11] Please compare state-of-the-art methods with our method.

[R11] Based on the prior state-of-the-art methods, we have added experiments on the two latest methods published in 2020 to demonstrate the supervised effectiveness and pixel-segmentation. The two latest methods are introduced in Section 4.2, and the experimental results are shown in Table 2 of Section 4.3, as follows. The results show that our proposed method can achieve better performance than them.

"SOLAR-Local[16] focuses on second-order spatial information to learn local patch descriptors without extra supervision. Based on the feature weighting strategy, it combines the second-order spatial attention and the second-order descriptor loss to improve image features for retrieval and matching."

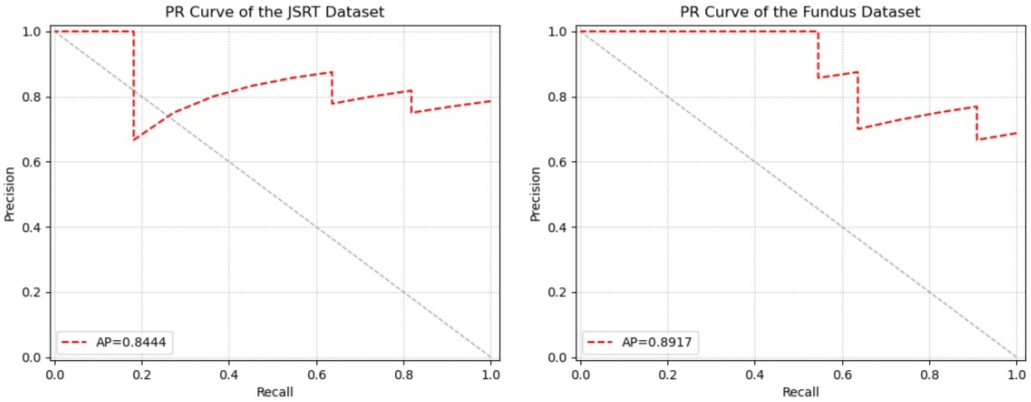
"DDMH[17] proposes a unique disentangled triplet loss to effectively push positive and negative sample pairs by desired Hamming distance discrepancies for hash-codes with different lengths."

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 [42]	512	0.5223	0.4681	0.4471	0.4366	0.4993	0.4705	0.4396	0.4189
CAM [43]	2048	0.5917	0.5488	0.4982	0.4609	0.5611	0.5124	0.4497	0.4187
BLCF [45]	1000	0.4890	0.4793	0.4463	0.4216	0.4701	0.4356	0.4096	0.3903
SOLAR-Local [65]	1024	0.5701	0.5274	0.4766	0.4482	0.5443	0.4987	0.4264	0.4051
R-MAC [46]	512	0.5016	0.4884	0.4585	0.4528	0.4682	0.4191	0.3965	0.3812
R-MAC + RPN [66]	3072	0.5483	0.5024	0.4685	0.4446	0.4805	0.4461	0.4098	0.3951
Regional Attention [41]	2048	0.5674	0.5279	0.5070	0.4854	0.4984	0.4621	0.4289	0.4069
Deep Vision + SOLO [51]	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 [39]	64	0.5052	0.4882	0.4788	0.4734	0.5487	0.4921	0.4578	0.4332
DRH [40]	64	0.5712	0.5435	0.5322	0.5203	0.5306	0.4912	0.4651	0.4498
DDMH [14]	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**

[Q12] Please share the average precision curves. On the other hand, evaluate the retrieval performance.

[R12] The success criteria of similar images are defined as that the two images have similar pathological regions. In our work, we apply the mean average precision (mAP) to evaluate the retrieval performance. In the R-MAC branch, we extract feature vectors to compute the circle loss for similarity training. Above this, our Y-Net achieves the best performance to search for more similar images containing pathological regions. If we add a linear layer to embed the feature vector to the softmax layer for class prediction in the R-MAC branch, we obtain the average precision score of 0.8444 for the JSRT dataset and 0.8917 for the Fundus dataset, and as follows.


PR Curve of the JSRT Dataset
Precision
Recall
AP=0.8444
PR Curve of the Fundus Dataset
Precision
Recall
AP=0.8917

Response to Reviewer #2:

Instance based image retrieval is an important task for radiology and medical diagnostics. This paper presents a CNN based approach that learns features and generate hash-codes to help in instance based image retrieval (combat manifestation ambiguity).

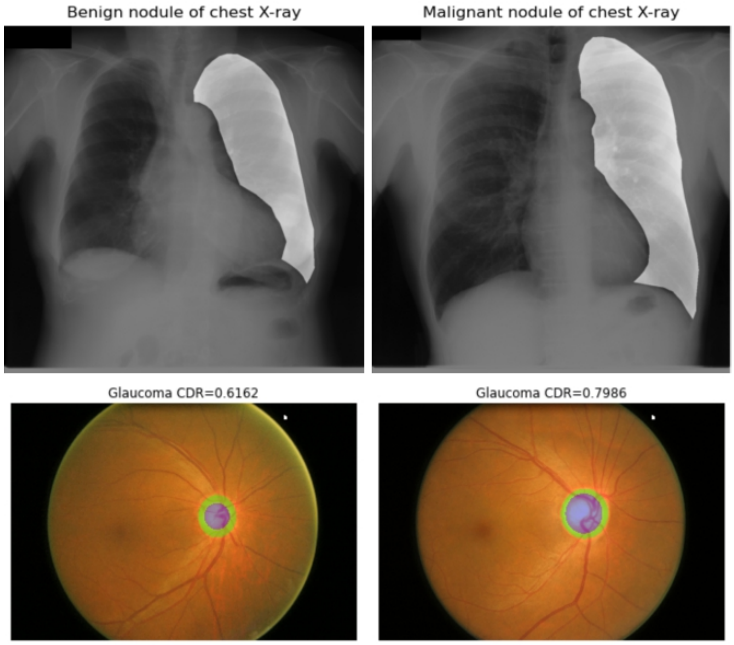
[Q1] The cup to disk ratio (CDR) in glaucoma is not explained (not clear at this point); for example what is CDR=0.616 vs CDR=0.798

[R1] Thanks for your kind mentions. CDR is the ratio of cup diameter to disc diameter, which is an important clinical factor for glaucoma diagnosis. We have added this introduction in Section 1.

[Q2] Figure 1 is not clear (not helpful). I think if it is explained well it will be informative.

[R2] Thanks for your king suggestions. We have replaced the ultrasound dataset with the chest X-ray dataset and rewrite the explanation for Figure 1. Both chest X-ray and ultrasound image have specificity shortage. The figure and the corresponding explanation are shown as follows:

"As below figure 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."



[Q3] There are quite a few repetition and redundancy in the paper (e.g. "….novel end-to-end framework called Y-net…." many times).

[R3] We thank the reviewer for careful polishing. We have revised the paper elaborately to improve the expression.

[Q4] How did you randomly select the training portion and the testing images (65 images) from the Fundus dataset? And why it is not repeated? why you did not repeat it 10 times (10f-cv)?

[R4] Thanks for your useful comment. We have revised the introduction of splitting datasets and added the more details of our experiments, as follows. We applied the 5-fold cross-validation to select the best classification and segmentation model.

"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."

"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 is used for classification training by using cosine similarity and setting a scale of 32, a margin of 0.25. The weight factor 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. "

[Q5] Table 2 needs more explanation.

[R5] Thanks for your kind suggestion. Based on the prior state-of-the-art methods, we have added experiments on the two latest methods published in 2020 to demonstrate the supervised effectiveness and pixel-segmentation. Then we have given more analysis for Table 2.

1) Two methods are introduced in Section 4.2, and the experimental results are shown in Table 2 of Section 4.3, as follows:

"SOLAR-Local[16] focuses on second-order spatial information to learn local patch descriptors without extra supervision. Based on the feature weighting strategy, it combines the second-order spatial attention and the second-order descriptor loss to improve image features for retrieval and matching."

"DDMH[17] 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."

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 [42]	512	0.5223	0.4681	0.4471	0.4366	0.4993	0.4705	0.4396	0.4189
CAM [43]	2048	0.5917	0.5488	0.4982	0.4609	0.5611	0.5124	0.4497	0.4187
BLCF [45]	1000	0.4890	0.4793	0.4463	0.4216	0.4701	0.4356	0.4096	0.3903
SOLAR-Local [65]	1024	0.5701	0.5274	0.4766	0.4482	0.5443	0.4987	0.4264	0.4051
R-MAC [46]	512	0.5016	0.4884	0.4585	0.4528	0.4682	0.4191	0.3965	0.3812
R-MAC + RPN [66]	3072	0.5483	0.5024	0.4685	0.4446	0.4805	0.4461	0.4098	0.3951
Regional Attention [41]	2048	0.5674	0.5279	0.5070	0.4854	0.4984	0.4621	0.4289	0.4069
Deep Vision + SOLO [51]	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 [39]	64	0.5052	0.4882	0.4788	0.4734	0.5487	0.4921	0.4578	0.4332
DRH [40]	64	0.5712	0.5435	0.5322	0.5203	0.5306	0.4912	0.4651	0.4498
DDMH [14]	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**

2) The added explanation for Table 2 is as follow:

"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 Fundus dataset's retrieval performance outperforms the JSRT dataset by 10.58% on the returned list of 10. There are two reasons for this gap. 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."

[Q6] The two example cases in Figure 1 may not be enough to explain SPDD and DPSD (more examples will be very helpful)

[R6] Thanks for the useful comment. We have strengthened the explanation for Figure 1. The improvement can be found in [R2].

Response to Reviewer #3:

The authors propose a deep learning architecture to train a medical image hash-coding pipeline which is to be used to retrieve similar medical images for evidence based diagnostics. The main goal of the proposed training scheme is to resolve the SPDD (Similar Pathology Different Disease) and DPSD (Different Pathology Similar Disease) "problems".

The proposed training architecture is novel to the best to my knowledge and makes sense. In short, they train a rather simple CNN based network (main branch) using 2 types of losses (classification and segmentation losses) simultaneously. The testing is performed on the main branch, which is concatenated with a static hash-code generation block. As said, this is a reasonable and novel approach to the best of my knowledge.

We thank the reviewer for the supportive comments and appreciate the highly refined summary. In the revised version, we have addressed all the concerns as follows.

[Q1] Before getting to some specific comments about the paper, I would like to express my major criticism which is not related to the techniques used but to the paradigm here. I believe a medical CBIR system that is to be used as part of an MD-in-the-loop Interpretability-Guided Content-Based Medical Image Retrieval should not be class sensitive. Designing such a CBIR positions the whole concept closer to a CAD system away from the MD-in-the-loop evidence based diagnosis paradigm. The sole aim of such systems must be to provide MDs with objective (image based) evidence and the MDs shall be giving the diagnosis. Presenting similar images from the "same" class is biasing the MD towards a CAD diagnostic result. If that is very reliable, there is no need for the MD in the loop. I, personally, would like to see such aspects discussed in the paper. Nevertheless, authors may not be sharing my point of view.

[R1] Personally, we agree with this opinion, which points to the essence of assisting evidence-based diagnosis for the CBIR system. In our Y-Net retrieval stage, images are retrieved according to their similar hash-codes but not the same class. The hash-codes are generated from the convolutional features which have learned the pathological information from the R-MAC branch and the FPN branch. In the R-MAC branch, we rely on the class label to enhance the visual features in pathologically abnormal regions and suppress the disturbing of the background during model training. Hence, in our work, class-aware information is used to overcome the DPSD problem by enhancing the response of pathological regions in the training stage but not for presenting similar images in the retrieval stage. Our Y-Net follows the motivation of CBIR systems that provide MDs with objective evidence by hash-code learning and comparison.

As for the paper itself:
[Q2] Please provide some quantitative results in the abstract

[R2] Thanks for the useful suggestion. Because the TNSCUI dataset can not get the authority to publish, we use a public chest x-ray dataset (JSRT) to test our method. The overall performance has been provided in the abstract, as follows.

"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."

[Q3] Why would CNNs be "smoothing out" visual features unique to an instance? I find such statements hard to fully understand what is meant.

[R3] Thanks for the kind comment. We argue that an instance (pathological region) in a medical image usually occurs in a small area, which may be affected by the large normal area during model training[19]. We have revised the related description in this work to avoid hard understand, as follows.

"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."

[Q4] Good overview of SoA

[R4] We thank the reviewer for the supportive comments.

[Q5] Good experimental evaluation, yet aren't the results still far from being clinically applicable level? How one can truly measure the impact of such CBIR systems? Without a clinical field experiment that assesses the improvement in clinical performance (i.t.o. both time and accuracy), such system proposals may be left in papers only. Hence a clinical experiment would be much appreciated.

[R5] We highly appreciate the reviewer for caring about the clinic application. When we explore CBIR methods, we also synchronously start to build the CBIR system. We have completed the data collection for the fundus image and will apply our method in Shenzhen People Hospital, China. This work was supported in part by The Science and Technology Innovation Committee of Shenzhen City (20200925174052004 and JCYJ20200109140820699). Besides, we have bulid the CBIR system for assisting chest X-ray screening in CVTE medical center (http://www.yibicom.com/). Now, our model is tested on the CVTE chest X-ray to deploy. The clinic feedback will help improve our model.

Response to Reviewer #4:

This paper proposes a novel architecture, Y-Net, to combat the manifestation ambiguity in medical instance retrieval. Y-Net consists of the main branch, R-MAC branch, FPN branch, and the coupled loss function, aiming to learn discriminative convolutional features by unifying the pixel-wise segmentation loss and classification loss. Regarding the novelty and quality of this paper, I have the following major concerns.

We thank the reviewer for the supportive comments. In the revised version, we have addressed all the concerns as follows.

[Q1] The TNSCUI dataset was used for this study. However, according to the challenge guidelines, "the publish right of this dataset is limited to the purpose of this challenge ONLY due to the ethical approval was obtained." Have you got the privilege to use this dataset for this publication?

[R1] Thanks for your reminder. We review the challenge policy again and have replaced the TNSCUI dataset with the JSRT dataset. There are two reasons for us to select the JSRT dataset. First, the X-ray images have lower specificity same as the ultrasound images. Second, this dataset provides pixel-level lung masks. The introduction of the JSRT dataset and its experimental results follow:

"JSRT[20] 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 [21]. 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."

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 [42]	512	0.5223	0.4681	0.4471	0.4366	0.4993	0.4705	0.4396	0.4189
CAM [43]	2048	0.5917	0.5488	0.4982	0.4609	0.5611	0.5124	0.4497	0.4187
BLCF [45]	1000	0.4890	0.4793	0.4463	0.4216	0.4701	0.4356	0.4096	0.3903
SOLAR-Local [65]	1024	0.5701	0.5274	0.4766	0.4482	0.5443	0.4987	0.4264	0.4051
R-MAC [46]	512	0.5016	0.4884	0.4585	0.4528	0.4682	0.4191	0.3965	0.3812
R-MAC + RPN [66]	3072	0.5483	0.5024	0.4685	0.4446	0.4805	0.4461	0.4098	0.3951
Regional Attention [41]	2048	0.5674	0.5279	0.5070	0.4854	0.4984	0.4621	0.4289	0.4069
Deep Vision + SOLO [51]	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 [39]	64	0.5052	0.4882	0.4788	0.4734	0.5487	0.4921	0.4578	0.4332
DRH [40]	64	0.5712	0.5435	0.5322	0.5203	0.5306	0.4912	0.4651	0.4498
DDMH [14]	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**

[Q2] It is claimed that the segmentation loss was used to avoid the SPDD problem and the classification loss was used to overcome the DPSD problem. However, it seems to be contradictory to the reported experimental results. Please double check it.

[R2] We thank the reviewer for careful polishing. We have checked and adjusted related results and analysis.

[Q3] What does “0.4 overlap” mean? (Line50, Column 2, Page 3)

[R3] Thanks for your detailed review. When sliding windows, the neighbor regions need to cover above 0.4 intersections.

[Q4] It would be better to briefly explain how to encode deep features into the hash code.

[R4] Thanks for your useful suggestion. We have improved the explanation of encoding deep feature into the hash code in Section 2, as follows.

“We apply feature aggregating to generate a k-bits hash-code from the learned convolutional feature maps X with C×H×W in the core node. The step of feature aggregation directly convolutes the size of C×H×W into the size of c×h×w. The three dimensions vectors further are squeezed into one dimension; its size is equal to the hash-code size of k-bits. Lastly, we apply the 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”

[Q5] Since the proposed model aims to address the problems of SPDD and DPSD, it would be necessary evaluate the model’s performance regarding to SPDD and DPSD.

[R5] Thanks for your useful suggestion. We have added a new paragraph to demonstrate our method's effectiveness in addressing the SPDD and DPSD problems in Section 4.3.2, and as follows.

“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."

[Q6] An ablation study is required to show the effectiveness of the FPN branch and R-MAC branch, respectively.

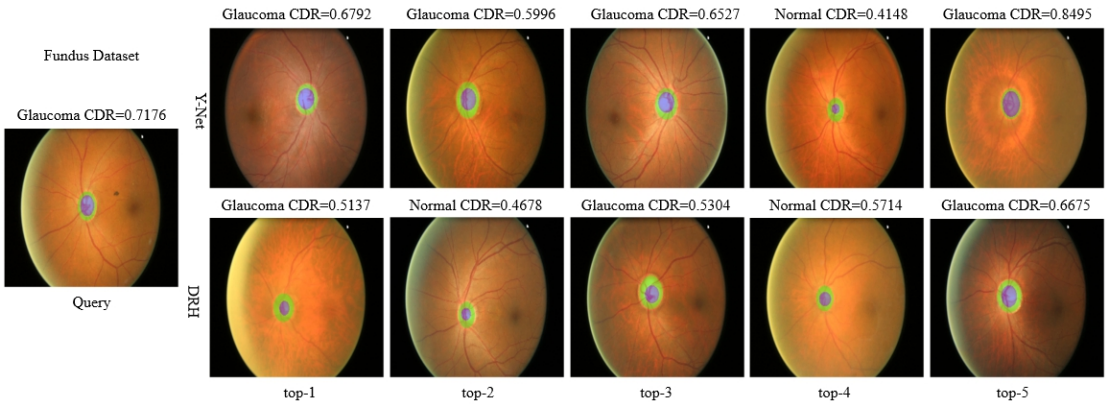
[R6] Thanks for your useful suggestion. We have improved the analysis of the ablation study. The FPN branch and R-MAC branch's effectiveness are shown in Table 3, and as follows.

"As shown in Table 3, 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 illustrated in below Figure, 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. "

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 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**



[Q7] When it comes to image retrieval, it is interesting to know how the retrieval method was evaluated. mAP is a good performance metric, but it cannot tell the whole story, since the retrieved cases are expected to be similar to the query case not only visually but also pathologically.

[R7] It is a challenging issue. Some works start to explore the evaluation problem of the CBIR system[18,27]. In terms of ranking evaluation, NDCG (Normalized Discounted Cumulative Gain) is a common metric. IG-CBIR[27] is proposed to explore the use of interpretability methods to improve medical image retrieval. This work can apply NDCG to evaluate retrieval performance by providing the ranking ground-truth. However, the current image dataset for retrieval tasks lacks a ranking label, thus we only rely on the mAP criteria. In this work, we try to evaluate the retrieval performance according to the pathologically pixel-wise label. We will follow this direction to explore the pathological region information to meet the evidence-based diagnosis in the future.

[Q8] If possible, please discuss the effectiveness of user feedback for this study.

[R8] We highly appreciate the reviewer for caring about the clinic feedback. When we explore CBIR methods, we also synchronously start to build the CBIR system. We have completed the data collection for the fundus image and will apply our method in Shenzhen People Hospital. This work was supported in part by The Science and Technology Innovation Committee of Shenzhen City (20200925174052004 and JCYJ20200109140820699). Besides, we have built the CBIR system for assisting chest X-ray screening in CVTE medical center (http://www.yibicom.com/). Now, our model is tested on the CVTE chest X-ray to deploy. The clinic feedback will come soon. We start to shift experiments from the public dataset to the private dataset that originates from the routine clinic.

Response to Reviewer #5:

This paper proposed a novel framework to compress images into compact hash codes for image retrieval. The method designed a multi-loss that unified the classification loss and pixel-wise segmentation loss to learn convolutional features. The results demonstrated that the method can avoid smoothing out visual features unique to pathologically abnormal regions. The method was evaluated on two image datasets and achieved state-of-the-art performance for image retrieval.
However, I have some major concerns about this work, which are listed below.

We thank the reviewer for the supportive comments. In the revised version, we have addressed all the concerns as follows.

[Q1] The main purpose of the propose Y-Net is to encode the image into hash-code (binary code). But the authors described very little the network structure for hash encoding, i.e. the main branch. In contrast, the segmentation branch and the classification branch, which are designed to assist in learning hash codes were described in great detail. The only sentences about hash encoding appear in lines 30-35 of the right column in page 3. However, these sentences are very subtle. I feel hard to figure out how the hashing is done.

[R1] Thanks for your careful review. We have improved the explanation of encoding deep feature into the hash code in Section 2, as follows.

"We apply feature aggregating to generate a k-bits hash-code from the learned convolutional feature maps X with C×H×W in the core node. The step of feature aggregation directly convolutes the size of C×H×W into the size of c×h×w. The three dimensions vectors further are squeezed into one dimension; its size is equal to the hash-code size of k-bits. Lastly, we apply the 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"

At the very least, the following problems need to be substantially clarified (correspondingly, the experiments may need to be redesigned):
It claims that the hashing is achieved by "feature aggregating" and then it is said "the feature aggregating convolutes the size of C×H×W into the size of c×h×w" and the hash-code of k-bits is equal to the size of c×h×w." So,
[Q1-1] What is the definition of the operation "convolute" ?

[R1-1] The convolutional features of C×H×W have learned the visual cues of pathological regions from the R-MAC branch and the FPN branch during model training, then encoded into a static convolutional feature of c×h×w by applying a convolutional layer.

[Q1-2] What is the exact setting of c, h, w? to meet k=36,64,128,256 (used in the experiments).

[R1-2] For example, the convolutional features of 256 × 8 × 8 are encoded into the static

convolutional feature of 1×8×8, which is equal to the hash-code size of 64.

[Q1-3] How to convert c×h×w into a vector of length k? If it is a direct flattening operation, the hash encoding implies a very strict restraint to the spatial structure of the image. I concern whether this restraint is positive for semantic retrieval? Then, it claims that "we do not introduce any weighting strategy on feature aggregation…". Here,

[R1-3] In fact, we can directly squeeze the convolutional features of C × H × W to generate hash-codes to avoid information loss. But the larger vector affects the retrieval efficiency. We make a trade-off between the performance and the efficiency. We use a convolutional layer to encode the convolutional features of C × H × W into the static convolutional feature of c×h×w, which is squeezed into one dimension vectors but not embedding. The information loss is in the process of the convolutional operation.

[Q1-4] Does the "convolute" operation not contain any weights? If not, how to compress the features from size C×H×W into the size of c×h×w?

[R1-4] The convolutional operation in the test stage can be viewed as feature aggregation across channels by the mean weighing strategy. Unlike the convolutional operation in the training stage, the feature aggregation cross channel or region are weighed with class labels or pixel-level masks.

[Q1-5] Why not involve a loss function to the k-bits output, to learn more powerful hashing networks? In my experience, a pair-wise or triplet loss with several regulation terms is very essential to achieve a considerable performance of hash retrieval. I'm not sure such a hash function without any constraint in the training works, although the authors perform classification constraint the segmentation constraint in the training.

[R1-5] Among the comparative methods, DPSH[22] is a classic hash-code learning framework by applying the pair-wise loss and residual block, and DSH[23] uses the triplet loss. If we directly train the hash-code by applying ranking loss in the main branch, our method only achieves fair performance to the DPSH and DSH. In the task of instance retrieval, the relevancy is mainly grounded on the visual similarity of an instance rather than the whole image, so the features of a region-wise instance residing in a retrieved image should be captured pointedly. Based on this motivation, we propose our Y-Net to combating the ambiguity for hash-code learning.

[Q1-6] More basic, what is the activation function for the k-bits output? How to convert the floating point outputs into binary codes?

[R1-6] Please forgive our negligence. we apply the tangent function to generate the value between -1 and 1, following by signed as binary hash-code.

[Q1-7] What is the similarity measurement for hashing retrieval? Is this hamming distance?

[R1-7] In our work, the indexing and similarity calculation for evaluation uses Faiss [24], a library

for efficient similarity search and clustering of dense vectors. For the similarity measurement function in the Faiss library, we select the hamming distance.

[Q2] All the methods for comparison are for natural sense images, among which the latest one is published in 2017. I concern whether the proposed method is really SOTA.

[R2] Thanks for your useful suggestions. Based on the prior state-of-the-art methods, we have added experiments on the two latest methods published in 2020 to demonstrate the effectiveness of our method. Two methods are introduced in Section 4.2, and the experimental results are shown in Table 2 of Section 4.3, as follows. DDMH is proposed for neuroimage search. The results show that our proposed method can achieve better performance than them.

"SOLAR-Local[16] focuses on second-order spatial information to learn local patch descriptors without extra supervision. Based on the feature weighting strategy, it combines the second-order spatial attention and the second-order descriptor loss to improve image features for retrieval and matching."

"DDMH[17] 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."

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 [42]	512	0.5223	0.4681	0.4471	0.4366	0.4993	0.4705	0.4396	0.4189
CAM [43]	2048	0.5917	0.5488	0.4982	0.4609	0.5611	0.5124	0.4497	0.4187
BLCF [45]	1000	0.4890	0.4793	0.4463	0.4216	0.4701	0.4356	0.4096	0.3903
SOLAR-Local [65]	1024	0.5701	0.5274	0.4766	0.4482	0.5443	0.4987	0.4264	0.4051
R-MAC [46]	512	0.5016	0.4884	0.4585	0.4528	0.4682	0.4191	0.3965	0.3812
R-MAC + RPN [66]	3072	0.5483	0.5024	0.4685	0.4446	0.4805	0.4461	0.4098	0.3951
Regional Attention [41]	2048	0.5674	0.5279	0.5070	0.4854	0.4984	0.4621	0.4289	0.4069
Deep Vision + SOLO [51]	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 [39]	64	0.5052	0.4882	0.4788	0.4734	0.5487	0.4921	0.4578	0.4332
DRH [40]	64	0.5712	0.5435	0.5322	0.5203	0.5306	0.4912	0.4651	0.4498
DDMH [14]	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**

[Q3] From the ablation study Page3, line53 "the spatially subtle differences … at different stages". The authors could elaborate why the FPN can get the same disease's subtle differences at different stages. I think it is not an obvious inference.

[R3] Thanks for your useful suggestions. We have improved the ablation study to interpret further the effectiveness of the R-MAC branch and the FPN branch, and as follows.

"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."

Minor comments:

[Q4-1] Several abbreviations are not explained before use, for example FPN, R-MAC, and ATH.

[R4-1] Thanks for your careful review. We have adjusted related writing.

[Q4-2] We usually do not use "On the one hand…On the other hand" to express coordinating relation.

[R4-2] Thanks for your careful review. We have adjusted related writing.

[Q4-3] Please give the mathematical definition of the metric mAP.

[R4-3] Thanks for your careful review. We have added the mathematical definition of the metric mAP in Section 4.2, and as follows.

"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 [25,26], and is calculated as follows:

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

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."

[Q4-4] The method names in the ablation study (Table III) are not friendly to read. I suggest to rename them as Y-Net w/o FPN, Y-Net w/o R-MAC, Y-Net w/o Circle loss, and Y-Net.

[R4-4] Thanks for your careful review. We have fixed related writing, and as follows.

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 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**

References

[1] Loyman, Mark, and Hayit Greenspan. "Semi-supervised lung nodule retrieval." *arXiv preprint arXiv:2005.01805* (2020).

[2] Faruque, Jessica, et al. "Content-based image retrieval in radiology: analysis of variability in human perception of similarity." *Journal of Medical Imaging* 2.2 (2015): 025501.

[3] Li, Zhongyu, et al. "Large-scale retrieval for medical image analytics: A comprehensive review." *Medical image analysis* 43 (2018): 66-84.

[4] Tolias, Giorgos, Yannis Avrithis, and Hervé Jégou. "To aggregate or not to aggregate: Selective match kernels for image search." *Proceedings of the IEEE International Conference on Computer Vision*. 2013.

[5] Tolias, Giorgos, Teddy Furon, and Hervé Jégou. "Orientation covariant aggregation of local descriptors with embeddings." *European Conference on Computer Vision*. Springer, Cham, 2014.

[6] Slaney, Malcolm, and Michael Casey. "Locality-sensitive hashing for finding nearest neighbors [lecture notes]." *IEEE Signal processing magazine* 25.2 (2008): 128-131.

[7] Raginsky, Maxim, and Svetlana Lazebnik. "Locality-sensitive binary codes from shift-invariant kernels." *Advances in neural information processing systems* 22 (2009): 1509-1517.

[8] Chen, Zhixiang, et al. "Order-sensitive deep hashing for multimorbidity medical image retrieval." *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer, Cham, 2018.

[9] Conjeti, Sailesh, et al. "Metric hashing forests." *Medical image analysis* 34 (2016): 13-29.

[10] Liu, Haomiao, et al. "Deep supervised hashing for fast image retrieval." *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2016.

[11] Öztürk, Şaban. "Image Inpainting based Compact Hash Code Learning using Modified U-Net." *2020 4th International Symposium on Multidisciplinary Studies and Innovative Technologies (ISMSIT)*. IEEE, 2020.

[12] Zhu, Han, et al. "Deep hashing network for efficient similarity retrieval." *Proceedings of the AAAI Conference on Artificial Intelligence*. Vol. 30. No. 1. 2016.

[13] Öztürk, Şaban. "Stacked auto-encoder based tagging with deep features for content-based medical image retrieval." *Expert Systems with Applications* 161 (2020): 113693.

[14] ÖZTÜRK, Şaban. "Two-Stage Sequential Losses based Automatic Hash Code Generation using Siamese Network." *Avrupa Bilim ve Teknoloji Dergisi*: 39-46.

[15] Kendall, Alex, Yarin Gal, and Roberto Cipolla. "Multi-task learning using uncertainty to weigh losses for scene geometry and semantics." *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2018.

[16] Ng, Tony, et al. "SOLAR: Second-Order Loss and Attention for Image Retrieval." *arXiv preprint arXiv:2001.08972* (2020).

[17] Yang, Erkun, et al. "Deep Disentangled Hashing with Momentum Triplets for Neuroimage Search." *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer, Cham, 2020.

[18] Loyman, Mark, and Hayit Greenspan. "Semi-supervised lung nodule retrieval." *arXiv preprint arXiv:2005.01805* (2020).

[19] Guan, Qingji, et al. "Diagnose like a radiologist: Attention guided convolutional neural network for thorax disease classification." *arXiv preprint arXiv:1801.09927* (2018).

[20] Shiraishi, Junji, et al. "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* 174.1 (2000): 71-74.

[21] Van Ginneken, Bram, Mikkel B. Stegmann, and Marco Loog. "Segmentation of anatomical structures in chest radiographs using supervised methods: a comparative study on a public database." *Medical image analysis* 10.1 (2006): 19-40.

[22] Li, Wu-Jun, Sheng Wang, and Wang-Cheng Kang. "Feature learning based deep supervised hashing with pairwise labels." *arXiv preprint arXiv:1511.03855* (2015).

[23] Wang, Xiaofang, Yi Shi, and Kris M. Kitani. "Deep supervised hashing with triplet labels." *Asian conference on computer vision*. Springer, Cham, 2016.

[24] Johnson, Jeff, Matthijs Douze, and Hervé Jégou. "Billion-scale similarity search with GPUs." *IEEE Transactions on Big Data* (2019).

[25] Zhou, Wengang, Houqiang Li, and Qi Tian. "Recent advance in content-based image retrieval: A literature survey." *arXiv preprint arXiv:1706.06064* (2017).

[26] Zheng, Liang, Yi Yang, and Qi Tian. "SIFT meets CNN: A decade survey of instance retrieval." *IEEE transactions on pattern analysis and machine intelligence* 40.5 (2017): 1224-1244.

[27] Silva, Wilson, et al. "Interpretability-Guided Content-Based Medical Image Retrieval." *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer, Cham, 2020.